Complexity Analysis of Algorithms in Algebraic Computation

by

Vikram Sharma

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

Courant Institute of Mathematical Sciences

New York University

January 2007

Approved: _____

Research Advisor: Prof. Chee K. Yap

To my parents.

Acknowledgments

My foremost gratitude is to my thesis advisor, Professor Chee Yap, for his guidance and motivation; without them I would not have reached the milestones where I stand now. He was always there when I needed his support and his sage advice has been helpful in the past four years and I am sure will be even more so in the future. I am also in debt to Professor Richard Pollack for his wise words and who through his own example has given me a deeper understanding of the aesthetics of research. Apart from them, I am obliged to Professor Michael Overton, Professor Margaret Wright, and Professor Victor Shoup who generously consented to be members of my thesis defence committee and provided me with invaluable feedback on my dissertation. I am also grateful to my collaborators Zilin Du and Arno Eigenwillig for the insightful discussions that I exchanged with them and for the knowledge that I acquired in the process. I would also like to express my gratitude towards the staff at the Department of Computer Science, NYU, and especially to Rosemary Amico and Anina Karmen for their efficient handling of the bureaucratic hurdles that I faced in my stay at the department.

Many people have contributed in other ways to my five years of doctoral study. My friends, who have always challenged, in a jocular manner I suppose, my ability to successfully complete these years and forced me to dig deep within myself to change their convictions, and who have gifted me memorable moments during these years. My final debt, which I cannot clear in a lifetime, is to my parents and brother for having faith in me even when I had lost all hope and courage.

Abstract

Numerical computations with real algebraic numbers require algorithms for approximating and isolating real roots of polynomials. A classical choice for root approximation is Newton's method. For an analytic function on a Banach space, Smale introduced the concept of approximate zeros, i.e., points from which Newton's method for the function converges quadratically. To identify these approximate zeros he gave computationally verifiable convergence criteria called point estimates. However, in developing these results Smale assumed that Newton's method is computed exactly. For a system of n homogeneous polynomials in $n+1$ variables, Malajovich developed point estimates for a different definition of approximate zero, assuming that all operations in Newton's method are computed with fixed precision. In the first half of this dissertation, we develop point estimates for these two different definitions of approximate zeros of an analytic function on a Banach space, but assume the strong bigfloat computational model of Brent, i.e., where all operations involve bigfloats with varying precision. In this model, we derive a uniform complexity bound for approximating a root of a zero-dimensional system of n integer polynomials in n variables. We also derive a non-asymptotic bound, in terms of the condition number of the system, on the precision required to implement the robust Newton method.

The second part of the dissertation analyses the worst-case complexity of two algorithms for isolating real roots of a square-free polynomial with real coefficients: The Descartes method and Akritas' continued fractions algorithm. The analysis of both algorithms is based upon amortization bounds such as the Davenport-Mahler bound. For the Descartes method, we give a unified framework that encompasses both the power basis and the Bernstein basis variant of the method; we derive an $O(n(L+\log n))$ bound on the size of the recursion tree obtained by applying the method to a square-free polynomial of degree n with integer coefficients of bit-length L, the bound is tight for $L=\Omega(\log n)$; based upon this result we readily obtain the best known bit-complexity bound of $\widetilde{O}(n^4L^2)$ for the Descartes method, where $\widetilde{O}$ means we ignore logarithmic factors. Similar worst case bounds on the bit-complexity of Akritas' algorithm were not known in the literature. We provide the first such bound, $\widetilde{O}(n^8L^3)$, for a square-free integer polynomial of degree n with coefficients of bit-length L.

Table of Contents

List of Figures

List of Tables

List of Appendices

Introduction

Numerical non-robustness is a recurring phenomenon in scientific computing. It is primarily caused by numerical errors arising because of fixed precision arithmetic. Most of these errors can be considered harmless, but occasionally there are "catastrophic" errors in the computation that cause non-robust behaviour such as crashing of the program, or infinite loops. Geometric algorithms are especially vulnerable to such non-robust behaviour. To illustrate the vulnerability, we consider the case of computing convex hull of four points in the plane, as shown in Figure 1(a). We proceed by forming the convex hull of any three non-colinear points, say p, q, r, and try to identify whether the fourth point s is inside or outside the convex hull of p, q, r (see Figure 1(b)). For the purpose of this identification we need the orientation predicate

$$orientation(p, q, r) := sign((q_x - p_x)(r_y - p_y) - (q_y - p_y)(r_x - p_x))$$

that takes as input three points p, q, r and returns $+1, -1$ or 0 based upon the following:

- $orientation(p, q, r) = +1$ iff the polyline (p, q, r) is a left turn.
- $orientation(p, q, r) = -1$ iff the polyline (p, q, r) is a right turn.
- $orientation(p, q, r) = 0$ iff the three points p, q, r are collinear.

Thus for the point s to lie outside the triangle determined by the points p, q, r there must be one edge of the triangle such that the point s is to the right of it. We next check the orientation of s relative to the edges of the triangle defined by points p, q, r, where the edges are considered in counter-clockwise direction. To check the orientation, however, we use a fixed precision implementation $float\_orient(p, q, r)$ of the orientation predicate, i.e., where each operation is done with fixed relative precision. Let the output of $float\_orient(p, q, s)$ and $float\_orient(q, r, s)$ be $+1$ as expected. Now we compute $float\_orient(r, p, s)$. But it *may* so happen that the output is $+1$ instead of -1, because of numerical errors, see Figure 1(c). Thus we incorrectly identify the point s to be inside the triangle determined by the points p, q, r and hence output the incorrect convex hull, as shown in Figure 1(d).

In general, geometric algorithms consist of two parts: a *combinatorial structure* characterizing the discrete relations between various geometric objects, and a *numerical representation* of the geometric objects; in our illustration, the combinatorial part was the positioning of the points relative to various line segments, and the numerical part was the coordinates of the four points. Geometric algorithms characterize the combinatorial structure by computing the discrete relations

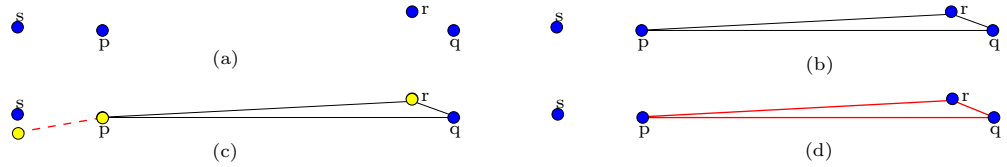

Figure 1: (a) Four points in the plane; (b) Is s outside the triangle Δpqr?; (c) Incorrect orientation of s relative to the line segment (r, p); (d) The wrong convex hull.

between geometric objects using numerical computations; for instance, using $float\_orient(p, q, r)$ to identify the orientation of r w.r.t. the line segment with endpoints p, q. However, numerical errors may yield us an incorrect characterization of the combinatorial structure which leads to non-robustness in the algorithm [Hof89, Yap97a, YM01]; in our example above, incorrect output of $float\_orient(r, p, s)$, resulting from numerical errors, resulted in computing the wrong convex hull.

In computational geometry, many approaches have been proposed to overcome the issue of non-robustness in geometric algorithms (see [Hof89, For96, Sch99, Yap04] for surveys). There is one approach which has become significant in the past years: Exact Geometric Computation (EGC) [Yap97b]. The principle of this approach is to *ensure that a geometric algorithm computes the correct combinatorial structure*. This principle is guaranteed by ensuring correct evaluation of all geometric predicates; to achieve this guarantee, the EGC approach does not necessarily require "exact arithmetic", but relies on guaranteed precision arithmetic, i.e., arithmetic where operations are done with sufficient precision to guarantee *a priori* precision requirements on the output of operations. The most alluring aspect of the EGC approach is its applicability to a large domain of problems. Moreover, the development of Filter Techniques [FvW93] has made this approach practically viable, which is evident from the success of `CGAL` and `LEDA`, two general libraries based upon the EGC model.

Most geometric predicates are determined by evaluating the sign of a polynomial at a real algebraic number. For instance, to determine whether a point $p \in \mathbb{R}^2$ is to the left, right, or on the line $\{(x, y) : \ell(x, y) = 0\}$, we can compute the sign of $\ell(p)$. In general, correct evaluation of geometric predicates requires the ability to represent and compute with algebraic numbers. There are two general approaches for representing algebraic numbers: *Algebraic* and *Numeric*.

The Algebraic approach represents real algebraic numbers based upon their algebraic properties. For example, Thom's encoding [CR88] of a real algebraic number α is given by the sequence

of signs obtained by evaluating all the derivatives of a polynomial that vanishes at α; comparing two algebraic numbers is straightforward in this representation (see [BPR03]). Other representations are the minimal polynomial representation, the standard representation, the matrix representation and the conjugate vector representation; of these four, the last three representations assume that the algebraic number is in some algebraic field (see [Coh93] for details). Arithmetic of algebraic numbers, in most of these representations, amounts to computing with polynomials. Even though robust and exact, this approach is inefficient.

The Numerical approach represents real algebraic numbers by *expressions*. An expression is a directed acyclic graph whose internal nodes are operators such as $\{+, -, *, /, \sqrt[k]{}\}$ and whose leaves are integers. To encompass all the real algebraic numbers, however, we need an operator that constructs an expression representing the real root of an arbitrary polynomial in $\mathbb{Z}[X]$. This operator is called the RootOf operator in `Core Library`[YLP+04], and the diamond operator in `LEDA Real`[Sch05]; the latter operator is more general since it can construct an expression representing the root of a polynomial whose coefficients are real algebraic numbers. The RootOf operator constructs a special expression that has two components: first, an integer polynomial $A(X)$ which has the real algebraic number as its root and the multiplicity of all its roots is one, and second an interval with rational endpoints, called the isolating interval, that contains the number inside it and excludes all other roots of $A(X)$. This representation is the standard isolating interval representation of algebraic numbers [Yap00]. To compare two expressions we use *constructive zero bounds* [BFM+01, MS00, LY01, PY03, Sek98]. A zero bound is a lower bound on the absolute value of a non-zero real number. A constructive zero bound is a function that takes as input an expression representing a real algebraic number and constructs a lower bound on its absolute value; the lower bound is valid only if the the real algebraic number is non-zero. Now to compare two real algebraic numbers α and β, represented as expressions, we first use the constructive zero bound function on the expression representing their difference $\alpha - \beta$ to get a lower bound of the form $|\alpha - \beta| > 2^{-k}$; then compute a numerical approximations $\widetilde{\alpha}$ and $\widetilde{\beta}$ to α and β such that $|\widetilde{\alpha} - \alpha|, |\widetilde{\beta} - \beta| < 2^{-k-3}$; then $|\widetilde{\alpha} - \widetilde{\beta}| < 2^{-k-1}$ iff $\alpha = \beta$. Thus comparison of two algebraic numbers represented as expressions, or even evaluating an expression that has an algebraic number at its leaves, requires the ability to approximate the number to any desired precision. Moreover, constructing the RootOf operator entails the capability to get an isolating interval for a real root of an integer polynomial. These two problems, namely approximating a real algebraic number to any desired precision and getting an isolating interval for a real root of

a polynomial, are the focus of this dissertation.

The problem of approximating a unique root of a polynomial in an isolating interval is a special instance of the general problem of approximating the root of a continuous function which has different signs at the endpoints of an interval that contains a unique root of the function. There is a rich literature of algorithms solving this general problem [Dek67, Ost60, Bre73, Kea90, Abb06]; for a detailed list see *Bracketing Methods* in [McN93]. One straightforward method is to bisect the interval until the width of the interval is smaller than the desired precision and the function has different signs at the endpoints of the interval. This procedure only yields one extra bit of precision at each bisection; nonetheless, it guarantees the desired precision, assuming the sign of the evaluated is computed correctly. Ideally, one would like to choose a point in the interval and iteratively apply methods that have super-linear convergence, such as Newton's method or Secant method. The problem with this approach is that these methods are not guaranteed to converge from an arbitrary point. In light of this drawback, methods have been proposed [Dek67, Bre73, Abb06] that combine the bisection method with super-linearly convergent methods and yield methods that are both fast in practice and have guaranteed convergence. Usually, these hybrid methods keep testing whether a bisection is required even when the interval is contained in an area around the root where super-linear convergence is guaranteed. To stop these unnecessary tests, we need the ability to identify whether a point is in a neighbourhood of a root where super-linear convergence sets in, i.e., the point is an approximate zero. The usual convergence criteria [KA64, OR70, Ost73], however, are based upon bounds on a neighbourhood of the root and hence are not easy to verify computationally. Smale [Sma81b, Sma85], on the other hand, developed computational convergence criteria called point estimates that depend only upon the knowledge of the point and are easily verifiable. The first chapter of the dissertation studies the notion of approximate zeros and point estimates under three computational models, which depend upon the amount of numerical precision one uses to compute various operations.

The problem of real root isolation is to assign an enclosing interval to each real root of a polynomial such that distinct roots are assigned disjoint intervals. It is a special case of the classical problem of root isolation, which has a rich literature of algorithms associated with it (see [Pan97] for a survey and [McN93] for a detailed bibliography). We may classify these algorithms as follows:

1. Bisection based approaches: Algorithms in this approach are generalizations of the bisection method used in root-approximation. The idea is to start with a region that encompasses all

the roots of the polynomial and then divide it into smaller parts until we reach a region that contains just one root. Clearly, we need the ability to identify whether a region contains a root or not, and for termination we need to further know when there is exactly one root in the region. For isolating complex roots, one can either use Sturm sequences in one or two dimensions [Ped91, Mil92], or use Turan's proximity test [Tur84]; the former has been used in [Wil78], and the latter in [Wey24] and its modifications mentioned in [Pan97]. For isolating only real roots, in addition to the methods mentioned, one can use the Descartes' rule of signs [Wan04]. In practice, real root isolation using the Descartes' rule of signs is faster [Joh98, RZ04] than using Sturm sequences. For polynomials with only real roots, [BOT90] gives a specialized version of Sturm sequences. Algorithms based upon the Descartes' rule of signs come in two variants depending upon the basis used to represent the polynomial: power basis [CA76, Kra95, Joh98] or Bernstein basis [LR81, Spe94, MRR04, MRR05]; the latter approach is preferred in CAGD [Far90], because of its stability and conditioning properties [FR87, FR88]. Another choice for determining whether an interval contains a root or not is based upon interval arithmetic [Moo66, AH83, Kea87, Mit91]. All the algorithms in this approach can be used to isolate a subset of roots in a specific region, unlike the approaches mentioned below (with the exception of the continued fractions approach) which necessarily isolate all the roots; another distinction is that all the operations in this approach are exact as long as extended precision is available.

2. Companion Matrix and Eigenvalue based approach: It is well known [Wil63] that the problem of root-approximation is ill-conditioned, i.e., small perturbations in the coefficients can distort the distribution of roots by a large amount. But one can reduce the problem of root isolation to finding the eigenvalues of the corresponding Frobenius companion matrix [EM95]; this is the preferred approach in numerical analysis, and root isolation in `MATLAB` is based upon this approach. The benefit of this reduction is that there are many stable algorithms for eigenvalue approximation [TB97]. A straightforward implementation, however, has poor accuracy for polynomials with high degree; algorithms have been proposed in [ACGR01] that overcome this drawback. Other forms of companion matrices have been used in [MV95], such as Fiedler's companion matrix which uses root approximations obtained from previous stages; a partial complexity analysis of the algorithm is provided in [For01]. An extensive survey of these approaches can be found in [PMR$^+$06]. Problems with this approach are that the algorithms consume $O(n^2)$ space, where n is the degree

of the polynomial whose roots we are trying to find, and each iteration could potentially cost $O(n^2)$, though improvements have been made [DAB04, DAB05] utilizing the special structure of the companion matrices.

3. The Durand-Kerner or Weierstrass method: The methods under this approach are Durand-Kerner [Dur60, Ker66], Aberth's method [Abe73] and Werner's method [Wer82]. These methods are iterative in nature and each iteration takes $O(n^2)$ time, where n is the degree of the polynomial. In [Pan02], a general derivation of the convergence behaviour of such algorithms is given; also, the computation has been modified to be carried out in single precision, a feature which was missing in the original approach. A key drawback of these algorithms is that their convergence is not guaranteed from any set of initial estimates. In practice, however, these methods are very efficient. `MPSolve` [BF00] is a multiple-precision library that provides an implementation of an algorithm for real root isolation based upon Aberth's method.

4. Factorization based approach: This approach is also called the Divide-and-Conquer approach or the splitting circle method. The algorithms (for example Schönhage's [Sch82]) produce a factorization of the polynomial into non-constant factors of lower degree , i.e., given an input precision ϵ these algorithm produce an approximation x_i to the roots of a degree n polynomial $A(X)$ such that

$$\|A(X) - \prod_{i=1}^{n}(X - x_i)\| \leq \epsilon \|A(X)\|.$$

   To isolate roots of an integer polynomial $A(X)$ using this approach, we require $\epsilon = 2^{\Omega(-n^2 L)}$, where L is the bit-length of the coefficients $A(X)$. One would expect that $\epsilon = 2^{\Omega(-nL)}$ would suffice, but the extra precision is needed to ensure that the distance between x_i and the closest root x_i^* of $A(X)$ is smaller than the root separation bound. Schönhage had showed that the bit-complexity of root isolation using his algorithm is $\widetilde{O}(n^3 L)$. The arithmetic complexity of his algorithm has been subsequently improved (see [NR94, Pan96]), but the bit-complexity has not improved substantially. Pan [Pan96] has provided optimal algorithms under this approach. These algorithms have the best known complexity bounds, but in practice they have not been as promising as the algorithms in the Durand-Kerner approach.

5. Homotopy or Path lifting: This a common method in numerical analysis. The method starts

with a polynomial whose zeros are known and constructs a sequence of polynomials, each of whose zeros it has approximated, that converges to the input polynomial whose roots we want to approximate. Algorithms in this category are [Sma85, KS94, HSS01]. These algorithms have been extended (see [SS93a, SS96, BCSS98, Mal93]) to approximating roots of systems of polynomials and their complexity in terms of condition numbers is also well studied.

6. Continued Fractions: The algorithms under this category are Vincent's algorithm [Vin36], its modification by Akritas [Akr78b], and Uspensky's algorithm [Usp48]. These algorithms also rely on the Descartes' rule of signs. The algorithms construct a continued fraction approximation to the real roots of the polynomial; we may subsequently apply Lagrange's method [Yap00, p. 470] to approximate the isolated root to any desired accuracy.

The above classification does not purport to cover all algorithms in the literature, but only gives a perspective on some of the interesting algorithms. Other perspectives have also been suggested, such as the partitioning between algorithms that are iterative in nature and algorithms that are exact.

The last two chapters of this dissertation focus on studying the worst-case complexity of two algorithms: the Descartes method, in its two equal formulations, i.e., when the polynomials are represented either in the power basis or in the Bernstein basis, and the continued fractions algorithm by Akritas. Both Akritas' algorithm [Akr78b] and the power basis variant of the Descartes method by Collins and Akritas [CA76] were proposed to improve the exponential running time of Vincent's algorithm [Vin36] for real root isolation. The algorithm proposed by Akritas is a simple modification of Vincent's algorithm and preserves its spirit. The algorithm by Collins and Akritas, however, is substantially different from Vincent's; their algorithm first constrains all the roots of the input polynomial in an interval and then sub-divides this interval into two equal parts and searches for roots in each of these halves. Both the algorithms rely on the Descartes' rule of signs for termination and have comparable running times.

The second chapter of this dissertation gives a uniform framework that encompasses both the power and Bernstein basis variant of the Descartes method. We derive tight bounds on the size of the recursion tree of the algorithm; based upon this we obtain the best known bit-complexity bounds for the algorithm. In the third chapter of this dissertation, we derive a polynomial bound on the worst-case running time of Akritas' algorithm.

Acknowledgments. The results in the first chapter are a generalization of my joint work with Zilin Du and Chee Yap [SDY05]. The results in the second chapter are from my joint work with Arno Eigenwillig and Chee Yap [ESY06].

1
Robust Approximate Zeros

Over the centuries Newton's method has been studied in many settings; for a history of the method see [Caj11, Ypm95]. We will focus on studying convergence criteria for Newton's method, and specifically those criteria which are based upon "point data", i.e., that depend only upon the "knowledge" of a given point. Moreover, we will only be interested in the case of analytic functions on Banach space (see Appendix A for a brief overview). Let $f : E \to F$ be an analytic map between two Banach spaces E and F. For a point $z \in E$ such that the Fréchet derivative $Df(z) : E \to F$ of f at z is non-singular we can define the **Newton map** $N_f : E \to E$ as

$$N_f(z) = z - Df(z)^{-1} f(z). \tag{1.1}$$

The *sequence* (z_i) *of Newton iterates starting from a point* $z_0 \in E$ is defined by the recurrence $z_i := N_f(z_{i-1})$. This sequence is said to be *well-defined* if $Df(z_i)$ is non-singular for all $i \geq 0$.

A classic convergence criteria [Ost73] for Newton's method states the following: Let $\tau^* \in E$ be a fixed point of N_f, i.e., $N_f(\tau^*) = \tau^*$, such that $|DN_f(\tau^*)| < 1$; then there exists an open convex neighbourhood $S \subseteq E$ of τ^* such that N_f is closed (i.e., $N_f(S) \subseteq S$) and differentiable on S, and for all $\tau \in S$ the sequence $(N_f^n(\tau))_{n \geq 1}$ converges to τ^*. This criterion depends upon the knowledge of the fixed point of N_f, or equivalently the zero of f, and hence is not useful in practice.

Kantorovich [Kan52] developed a convergence criterion that does not depend upon knowing the zero. He stated that if there are constants A, B such that for a point z (1) $\|Df(z)^{-1}\| \leq A$, (2) $\|Df(z)^{-1} f(z)\| \leq B$, (3) there exists an open convex set S containing z such that $Df(z)$ is Lipschitz on S with constant C, i.e., $\|Df(x) - Df(w)\| \leq C\|x - w\|$ for all $x, w \in S$, and (4) $ABC \leq \frac{1}{2}$, then there exists a unique zero $z^* \in S$ of f such that the sequence of Newton iterates starting from z converge to it. A proof based upon majorant sequences can be found in [KA64, Ort68, GT74]. Yamamoto [Yam85, Yam86] gives sharp bounds on the size of S. The drawback of Kantarovich's result is that the third constraint depends upon bounding the derivative in a region and hence is not easy to verify computationally.

Smale [Sma81a, Sma85] gave a computationally verifiable criterion based upon point data, which he later called **point estimate**. An example of a point estimate is the following: For a polynomial $f(z) \in \mathbb{C}[z]$, if a point $z \in \mathbb{C}$ is such that $9|f(z)| \leq \min(|f(\theta)| : f'(\theta) = 0)$ then the Newton iterates starting from z converge to a zero z^* of f; at first it may appear that this is not computationally verifiable, but assuming that $f(z)$ has roots with multiplicity 1,

we can derive (see [Yap00, p. 183]) lower bounds on $|f(\theta)|$, where θ is a critical point of $f(z)$. Shub and Smale [SS85, SS86] generalized this result to iterative methods with higher order of convergence, such as the k-th Euler incremental algorithm that has an order of convergence $k+1$. The proof in these results relied on Bieberbach conjecture from the theory of Schlicht functions [Dur77]. This dependency was later removed by Smale in [Sma86] by changing the definition of approximate zeros so that it depends upon the behaviour of the iterates in the domain space (see Definition 1.1). Moreover, in the same paper he developed results for analytic functions in Banach spaces; this assumption of analytic functions is stronger than Kantorovich's requirement of differentiability in a neighbourhood. In particular, Smale showed that if $\alpha(f, z) < 0.13$ then z is an approximate zero for an analytic function $f : E \to F$, where E and F are Banach spaces and $\alpha(f, z)$ is a standard function in the theory of point estimates (see (1.4) below). Similar point estimates were shown by Kim [Kim86, Kim88]. For the same definition of approximate zeros as in [Sma86], Curry [Cur87] has developed point estimates for the k-th Euler incremental algorithm. The constant (0.13 above) involved in the point estimate by Smale was improved to $3 - 2\sqrt{2} \simeq 0.17\ldots$ by Wang and Zhao [DF95] using Kantorovich's approach of majorant sequences; they also developed point estimates for the Weierstrass method [Dur60, Ker66], as was done by Petkovic et al. [PCT95, PHI98] and Batra [Bat98]. Shub and Smale [SS93a] have derived point estimates for the special case of a system of multivariate polynomials in the affine and projective space. Chen [Che94] has developed point estimates for any quadratically convergent algorithm.

All the above results assume that $N_f(z)$ is computed using exact arithmetic. This is hardly possible in practice, and even when it is possible, such as the case of polynomials with rational coefficients, it is undesirable because of inefficiency. In practice, the iterates are represented by floating-point numbers. In this chapter we compute with **bigfloats**, i.e., rational numbers of the form $n2^m$, for integers n, m (see Appendix C). Bigfloat arithmetic is basically the multiple-precision arithmetic of Brent [Bre76a, Bre76b]. There are two ways of computing with bigfloats: the **weak model** where all the operations are done to a fixed precision, similar to the IEEE floating-point arithmetic [IEE85]; and the **strong model** where all the operations can be done with varying precision.

Malajovich [Mal93] has developed point estimates in the weak model, whereas Sharma, Du and Yap [SDY05] have developed point estimates in the strong model; however, unlike Malajovich their result was developed for the restricted case of analytic functions in the complex plane.

This chapter extends the results in [SDY05]: We develop point estimates in the weak (Theorem 1.5) and the strong (Theorem 1.7) model for analytic functions on Banach spaces; we derive the complexity (Theorem 1.16) of approximating a common root of a zero-dimensional system of polynomials when the computations are done in the strong bigfloat model. In developing the complexity result we give a non-asymptotic worst-case bound (Lemma 1.21) on the precision needed to implement robust Newton iteration, Algorithm RN in §1.5.

The complexity estimates in this chapter are based upon Schönhage's pointer machine model [Sch80], rather than the standard multi-tape Turing machines, because the latter introduce unwanted complications in our complexity estimates involving unbounded bigfloats (i.e., bigfloats with arbitrary large exponents); for instance, if a bigfloat $n2^m$ is represented in the obvious way on a Turing tape (say m followed by n and the tape head on m), we cannot read n without scanning m, which unnecessarily distorts the complexity of basic operations such as truncation.

Functions used in error analysis. Let $f : E \to F$ be an analytic map as earlier and $z \in E$ such that $Df(z)$ is non-singular. Following Smale [Sma86] we can define the following functions:

- The **beta** function
$$\beta(f, z) := \|N_f(z) - z\| = \|Df(z)^{-1} f(z)\|. \tag{1.2}$$
- The **gamma** function
$$\gamma(f, z) := \sup_{k>1} \left(\frac{1}{k!} \|Df(z)^{-1} D^k f(z)\| \right)^{1/k-1}. \tag{1.3}$$
- The **alpha** function
$$\alpha(f, z) := \beta(f, z)\gamma(f, z). \tag{1.4}$$
- For $z, w \in E$ define
$$u(z, w) := \gamma(f, z)\|z - w\|. \tag{1.5}$$
For the special case when z is a zero of f, we use the succinct notation u_w.

If $Df(z)$ is singular the first three functions are defined to be ∞. We will shorten the three functions to $\beta(z)$, $\gamma(z)$ and $\alpha(z)$ if f is clear from the context. We always use z^* to represent a zero of f and γ_* to denote $\gamma(f, z^*)$. For $z \in E$ and $r \in \mathbb{R}_{\geq 0}$ let

$$\overline{B}(z, r) := \{w \in E : \|z - w\| \leq r\}. \tag{1.6}$$

In addition to the above, we define the following polynomial which will be useful in our analysis:

$$\psi(x) := 1 - 4x + 2x^2. \tag{1.7}$$

Remark 1.1. *The least positive zero of this polynomial is* $1 - 1/\sqrt{2}$*. Moreover, the polynomial is monotonically decreasing from left to right in the interval* $[0, 1 - 1/\sqrt{2}]$*.*

In what follows, unless stated otherwise, we take $f : E \to F$ to be an analytic map between Banach spaces, and z to be a point in E.

Error Notation. We borrow two convenient notations for error bounds from [SDY05]: we shall write

$$[z]_t \qquad (\text{resp., } \langle z \rangle_t) \tag{1.8}$$

for *any* relative (resp., absolute) t-bit approximation of z.

The following meta-notation is convenient: whenever we write "$z = \widetilde{z} \pm \epsilon$" it means "$z = \widetilde{z} + \theta\epsilon$" for some $\theta \in [-1, 1]$. More generally, the sequence "$\pm h$" is always to be rewritten as "$+\theta h$" where θ is an implicit real variable satisfying $|\theta| \leq 1$. Unless the context dictates otherwise, different occurrences of $\pm$ will introduce different θ-variables. E.g., $x(1 \pm u)(1 \pm v)$ means $x(1 + \theta u)(1 + \theta' v)$ for some $\theta, \theta' \in [-1, 1]$. The effect of this notation is to replace inequalities by equalities, and to remove the use of absolute values.

1.1 Approximate Zeros

Let (z_i) be the sequence of Newton iterates for f starting from a point z_0. Suppose that the sequence converges to a root z^* of f. Intuitively, z_0 is called an approximate zero if the sequence converges "quadratically" to z^*. We may quantify the rate of convergence in two ways: first, in the range space by using the value of the residual $\|f(z_i)\|$, or second, in the domain space by using the value $\|z_i - z^*\|$, or even $\|z_i - z_{i-1}\|$. Based upon these two ways to measure the rate of convergence, we may broadly classify the different definitions of approximate zeros in the literature. We only focus on definitions of the second type, i.e., those that measure the rate of convergence in the domain space. In this setting, one possible definition for an approximate zero is that the sequence converges quadratically in the standard sense, i.e., if $\|z_i - z^*\| \leq C\|z_{i-1} - z^*\|^2$, for some constant $C \in \mathbb{R}_{>0}$. However, this definition is too strong for our purposes, because it

is hard to guarantee in the presence of errors in the computation. Following the nomenclature suggested by Smale [Sma86], we have the following.

Definition 1.1. Let $z_0 \in E$ be such that the sequence of Newton iterates (z_i), given by the recurrence $z_i := N_f(z_{i-1})$, is well defined. Then

- z_0 is an **approximate zero of the first kind** if there is a unique zero $z^* \in E$ of f such that for all $i \in \mathbb{N}_{\geq 1}$
$$\|z_i - z_{i-1}\| \leq 2^{1-2^{i-1}}\|z_1 - z_0\|;$$
- z_0 is an **approximate zero of the second kind** if there is a unique zero $z^* \in E$ of f such that for all $i \in \mathbb{N}_{\geq 0}$
$$\|z_i - z^*\| \leq 2^{1-2^i}\|z_0 - z^*\|;$$
- z_0 is an **approximate zero of the third kind** if there is a unique zero $z^* \in E$ of f such that for all $i \in \mathbb{N}_{\geq 0}$
$$\|z_i - z^*\| \leq 2^{-2^i};$$
- z_0 is an **approximate zero of the fourth kind** if there is a unique zero $z^* \in E$ of f such that for all $i \in \mathbb{N}_{\geq 0}$
$$\frac{\|z_i - z^*\|}{\|z_i\|} \leq 2^{1-2^i}.$$

We call z^* the **associated zero** of z_0.

The first two definitions are by Smale [Sma86]; the third definition is by Kantorovich [Kan52]; the fourth definition is by Malajovich [Mal93]. We next clarify the relations amongst these definitions.

1.1.1 Relation amongst various approximate zeros

Consider the definitions of approximate zeros of the first and second kind. We show that the two definitions are almost equivalent. If z_0 is an approximate zero of the first kind then for any $N > i$ we know
$$\|z_N - z_i\| \leq \sum_{j=i+1}^{N} \|z_j - z_{j-1}\| \leq \|z_1 - z_0\| \sum_{j=i+1}^{N} 2^{1-2^{j-1}}.$$
Letting N tend to ∞ we get
$$\|z_i - z^*\| \leq 2^{2-2^i}\|z_1 - z_0\| \lesssim 2^{2-2^i}\|z_0 - z^*\| \tag{1.9}$$

since from Lemma 1.12 below we know that $\|z_1 - z_0\| \sim \|z_0 - z^*\|$. Assuming that z_0 is an approximate zero of the second kind we get for any $i \geq 1$,

$$\begin{aligned}
\|z_i - z_{i-1}\| &\leq \|z_i - z^*\| + \|z_{i-1} - z^*\| \\
&\leq 2^{1-2^{i-1}}(2^{-2^{i-1}} + 1)\|z_0 - z^*\| \\
&\leq 2^{2-2^{i-1}}\|z_0 - z^*\| \\
&\lesssim 2^{2-2^{i-1}}\|z_1 - z_0\|.
\end{aligned}$$

Overlooking the additional constant factors in the result above and (1.9), we get the desired equivalence of approximate zeros of the first and second kind.

The relation between approximate zeros of the second and third kind is trivial: the former implies the latter if $\|z_0 - z^*\| \leq \frac{1}{2}$, which is quite likely to hold in practice, and the latter implies the former only if $\|z_0 - z^*\| = \frac{1}{2}$, which is unlikely to hold.

The definition of approximate zeros of the fourth kind was proposed by Malajovich [Mal93]. It is obvious that the definition holds only if $z_i \neq 0$. This assumption is justified when the points z_i and the zero z^* are elements in the projective space $\mathbb{P}^n(\mathbb{C})$, which is the original setting of the definition as proposed by Malajovich. To accommodate the presence of $\|z_i\|$ in the definition, Malajovich has used different definitions of the three functions $\alpha(f, z), \beta(f, z), \gamma(f, z)$ (see §1.3); in the same section, we will show that point estimates for approximate zeros of the fourth kind can be derived from the results for approximate zeros of the second kind.

For each kind of approximate zero above there are three computational models to consider, namely the exact, the weak and the strong model. For each of these models two results can be developed: a point estimate and the complexity of approximating the root of a zero-dimensional system of polynomials in terms of the condition number of the system. In this categorization, we can reconsider the literature on point estimates.

Smale [Sma86] developed point estimates in the exact model, for approximate zeros of the first kind; later Shub and Smale [SS93a] derived complexity results for approximate zeros of the first kind; Blum et al. [BCSS98] derived point estimates and complexity results for approximate zeros of the second kind. Malajovich [Mal93] developed both point estimates and complexity for approximate zeros of the fourth kind in the exact and the weak model. Sharma et al. [SDY05] have developed point estimates and complexity for approximate zeros of the second kind in the strong model. There have been no explicit point estimates for approximate zeros of the third kind, though from our observation earlier we can easily derive them from point estimates of the

second kind.

The aim of the rest of the chapter is to derive point estimates and complexity for approximate zeros of the second kind in the weak and the strong model. For the sake of understanding, however, we will first re-derive the results in the exact model. From now on an approximate zero will always mean an approximate zero of the second kind. We will also re-derive the point estimates for approximate zeros of the fourth kind in the exact and the weak model, following an approach different from Malajovich's, and extend the result to the strong model.

1.2 Point Estimate of Second Kind

In this section we start by re-deriving a result of Smale [Sma86] that identifies a set of approximate zeros in the neighbourhood of a root, and the subsequent point estimate given in [BCSS98]. Based upon this result we will derive a point estimate in the weak model, which readily yields us the point estimate in the strong model. All the derivations proceed in two steps similar to [BCSS98]:

- We first identify a closed set $\overline{B}(z^*, R_1)$ around a simple zero z^* such that all points in this set are approximate zeros.
- Then we identify a criterion such that if any point z satisfies it then there is a zero z^* of f in $\overline{B}(z, R_2)$. Thus for z to be an approximate zero we additionally want $R_2 \leq R_1$.

The following property of bounded linear maps will be useful later on:

Lemma 1.1. *Let $M : E \to F$ be a bounded linear map such that $\|M\| < 1$. Then*

1. $(I - M)^{-1} = \sum_{i=0}^{\infty} M^i$ *and*
2. $\|(I - M)^{-1}\| < \frac{1}{1-\|M\|}$.

Proof. It is not hard to verify that $(I - M)\sum_{i=0}^{\infty} M^i = I$. From the first property we know that $\|(I - M)^{-1}\| \leq \sum_{i=0}^{\infty} \|M\|^i$, and since $\|M\| < 1$ we get the desired result. □

As a consequence of this lemma we have the following ([Sma86, Lem. 1])

Lemma 1.2. *Let $A, B : E \to F$ be bounded linear maps such that A is invertible and $c := \|A^{-1}B - I\| < 1$. Then B is invertible and $\|B^{-1}A\| < \frac{1}{1-c}$.*

This follows by choosing $M = I - A^{-1}B$ in Lemma 1.1.

1.2.1 The Exact Model

Let $z' := N_f(z)$, be well defined, i.e., $Df(z)$ be non-singular. Let z^* be a zero of f such that $Df(z^*)$ is non-singular. Consider

$$\begin{aligned}\|z' - z^*\| &= \|z - z^* - Df(z)^{-1}f(z)\| \\ &= \|Df(z)^{-1}(Df(z)(z - z^*) - f(z))\|.\end{aligned}$$

Consider the term $Df(z)(z-z^*)-f(z)$. By writing $Df(z)$ and $f(z)$ in terms of Taylor's expansion around z^* we get

$$Df(z)(z - z^*) - f(z) = \sum_{k=2}^{\infty} \frac{1}{k!} D^k f(z^*)(z - z^*)^k.$$

Thus

$$z - z^* - Df(z)^{-1}f(z) = Df(z)^{-1}Df(z^*) \sum_{k=2}^{\infty} \frac{1}{k!} Df(z^*)^{-1} D^k f(z^*)(z - z^*)^k.$$

Taking norms on both sides we obtain

$$\begin{aligned}\|z' - z^*\| &\leq \|Df(z)^{-1}Df(z^*)\| \sum_{k=2}^{\infty} \frac{1}{k!} \|Df(z^*)^{-1} D^k f(z^*)\| \|z - z^*\|^k \\ &\leq \|Df(z)^{-1}Df(z^*)\| \|z - z^*\| \sum_{k=2}^{\infty} (k-1)(\gamma_* \|z - z^*\|)^{k-1} \\ &\leq \|Df(z)^{-1}Df(z^*)\| \|z - z^*\| \sum_{k=2}^{\infty} (k-1) u_z^{k-1}.\end{aligned}$$

Assuming $u_z < 1$ we obtain

$$\|z' - z^*\| \leq \|Df(z)^{-1}Df(z^*)\| \|z - z^*\| \frac{u_z}{(1 - u_z)^2}. \tag{1.10}$$

We next bound $\|Df(z)^{-1}Df(z^*)\|$. The following lemma gives us the desired bound.

Lemma 1.3. *If $z, w \in E$ are such that $u(z, w) < 1 - 1/\sqrt{2}$ then*

$$\|Df(w)^{-1}Df(z)\| < \frac{(1 - u(z, w))^2}{\psi(u(z, w))}.$$

Proof. Let $u := u(z, w)$. Then the Taylor expansion of $Df(w)$ about z gives us

$$Df(w) = \sum_{k=0}^{\infty} \frac{1}{k!} D^{k+1} f(z)(w - z)^k.$$

Multiplying across by $Df(z)^{-1}$ we obtain

$$Df(z)^{-1}Df(w) = I - \sum_{k=1}^{\infty} \frac{k+1}{(k+1)!} Df(z)^{-1} D^{k+1} f(z)(w - z)^k$$

and hence

$$\begin{aligned}\|Df(z)^{-1}Df(w) - I\| &\leq \sum_{k=1}^{\infty} \frac{k+1}{(k+1)!}\|Df(z)^{-1}D^{k+1}f(z)\|\|w-z\|^k \\ &\leq \sum_{k=1}^{\infty}(k+1)(\gamma(z)\|w-z\|)^k \\ &= (1-u)^{-2}-1\end{aligned}$$

since by assumption $u(z,w) < 1$. Since $u < 1 - 1/\sqrt{2}$ we know that $(1-u)^{-2} - 1 < 1$ and hence we can apply Lemma 1.2 to obtain

$$\|Df(w)^{-1}Df(z)\| \leq \frac{(1-u)^2}{\psi(u)}.$$

□

The lemma above along with (1.10) gives us

Lemma 1.4. *If $z \in E$ is such that $u_z < 1 - 1/\sqrt{2}$ then*

$$\|N_f(z) - z^*\| \leq \frac{u_z}{\psi(u_z)}\|z - z^*\|.$$

Again, let $z' := N_f(z)$. Then from the lemma above we know that $u_{z'} \leq \frac{u_z}{\psi(u_z)}u_z$. If $z \in E$ be such that $u_z < \frac{5-\sqrt{17}}{4}$ then we get $u_{z'} \leq u_z$ and hence $\psi(u_{z'}) \geq \psi(u_z)$ (see Remark 1.1).

Based upon these results we can inductively show the following:

Lemma 1.5. *Let z_0 be such that $u_{z_0} < \frac{5-\sqrt{17}}{4}$. Then the sequence of Newton iterates z_i starting from z_0 satisfy*

$$\|z_i - z^*\| \leq \left(\frac{u_{z_0}}{\psi(u_{z_0})}\right)^{2^i-1}\|z_0 - z^*\|.$$

Proof. For sake of succinctness let $u_i := u_{z_i}$. The proof is inductive; the base case is trivial. Suppose the hypothesis holds for $i-1$, i.e.,

$$\|z_{i-1} - z^*\| \leq \left(\frac{u_0}{\psi(u_0)}\right)^{2^{i-1}-1}\|z_0 - z^*\|.$$

The we know that $u_{i-1} < u_0 < \frac{5-\sqrt{17}}{4}$ and hence from Lemma 1.4 we obtain

$$\|z_i - z^*\| \leq \frac{u_{i-1}}{\psi(u_{i-1})}\|z_i - z^*\| \leq \frac{\gamma_*}{\psi(u_{i-1})}\|z_i - z^*\|^2.$$

From Remark 1.1 we further know that $\psi(u_{i-1}) > \psi(u_0)$. Thus

$$\|z_i - z^*\| \leq \frac{\gamma_*}{\psi(u_0)}\|z_i - z^*\|^2.$$

Applying the inductive hypothesis we obtain

$$\|z_i - z^*\| \leq \frac{\gamma_*}{\psi(u_0)} \left(\frac{u_0}{\psi(u_0)}\right)^{2^i-2} \|z_0 - z^*\|^2 = \left(\frac{u_0}{\psi(u_0)}\right)^{2^i-1} \|z_0 - z^*\|.$$

□

Furthermore, if we choose z_0 such that $\frac{u_{z_0}}{\psi(u_{z_0})} \leq \frac{1}{2}$ then we have shown that z_0 is an approximate zero of f with associated zero z^*. But this follows if $u_{z_0} \leq \frac{3-\sqrt{7}}{2}$. Thus we have the following result:

Theorem 1.2 ([Sma86, Thm. C])**.** *Let z^* be a simple zero of f. If $z \in E$ is such that*

$$\|z - z^*\| \leq \frac{3-\sqrt{7}}{2\gamma(z^*)}$$

then z is an approximate zero of f with z^* as the associated zero.

This result corresponds to the first part mentioned in the beginning of §1.2. To obtain the second part, we will need the concept of a contracting operator: A map $\Gamma : \mathcal{X} \subset E \to \mathcal{X}$ is called a **contracting operator** if there exists a $\kappa < 1$, called the **contraction bound** of Γ, such that for all $z, w \in \mathcal{X}$ we have

$$\|\Gamma(z) - \Gamma(w)\| \leq \kappa \|z - w\|.$$

The Banach principle of contracting operator is that if $\mathcal{X}$ is complete then there is a $z^* \in \mathcal{X}$ such that $\Gamma(z^*) = z^*$, i.e., there is a unique fixed point of Γ in $\mathcal{X}$. Moreover, for any point $z \in \mathcal{X}$ the sequence $(\Gamma^n(z))$, $n \geq 0$, converges to z^*. Also, for such a Γ we can show that (see [Ost73, Thm. 32.1])

$$\frac{\|\Gamma(z) - z\|}{1+\kappa} \leq \|z - z^*\| \leq \frac{\|\Gamma(z) - z\|}{1-\kappa}.$$

Suppose that $\mathcal{X}$ is convex and Γ is differentiable over $\mathcal{X}$. If for all $z \in \mathcal{X}$,

$$\|D\Gamma(z)\| \leq C < 1,$$

then by the mean value theorem we know that C is a contraction bound for Γ.

Given the results above, we need to determine for what points z is the Newton map N_f a contracting operator. To do this, we will bound $\|DN_f(w)\|$, where w is a point in some neighbourhood of z. What is $DN_f(w)$? From the definition of the Newton operator we know that

$$Df(w)N_f(w) = Df(w)w - f(w).$$

Differentiating both sides and moving the term $D^2f(w)N_f(w)$ to the right we obtain

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

Thus we have

$$DN_f(w) = Df(w)^{-1}D^2f(w)Df(w)^{-1}f(w)$$

and hence

$$\|DN_f(w)\| \leq \|Df(w)^{-1}D^2f(w)\|\|Df(w)^{-1}f(w)\| \leq 2\gamma(w)\beta(w) = 2\alpha(w). \tag{1.11}$$

To derive bounds on $\|DN_f(w)\|$, we need to bound $\alpha(w)$. This will be done by first expressing $\alpha(w)$ in terms of $\alpha(z)$. The following lemma will be useful in deriving this relation:

Lemma 1.6. *For $0 \leq x < 1$ and $k \in \mathbb{N}$ we have*

$$\sum_{i=0}^{\infty} \binom{k+i}{i} x^i = \frac{1}{(1-x)^{k+1}}.$$

Proof. Notice the right hand side is just $\prod_{l=1}^{k+1} \sum_{j_l=0}^{\infty} x^{j_l}$. Thus we need to show that the coefficient of x^i in this product is $\binom{k+i}{i}$. This is the same as the number of choices of j_l, $l = 1, \ldots, k+1$, such that $\sum_{l=1}^{k+1} j_l = i$. Clearly, there are $\binom{i+k}{k}$ such options; since the coefficient of each x^{j_l} is one, the coefficient of x^i is $\binom{k+i}{i}$. □

The following lemma gives us the relation amongst the three functions for z and a point w in a neighbourhood of z.

Lemma 1.7. *Let $w \in E$ be such that $u := u(z,w) < 1 - 1/\sqrt{2}$. Then*

- $\beta(w) \leq \frac{1-u}{\phi(u)}((1-u)\beta(z) + \|z-w\|)$,
- $\gamma(w) \leq \frac{\gamma(z)}{(1-u)\phi(u)}$, *and*
- $\alpha(w) \leq \frac{\alpha(z)+u}{\phi(u)^2}$.

Proof. From (1.2) we know that

$$\beta(w) = \|Df(w)^{-1}f(w)\| \leq \|Df(w)^{-1}Df(z)\|\|Df(z)^{-1}f(w)\|.$$

Since $u < 1 - 1/\sqrt{2}$, from Lemma 1.3 we obtain

$$\begin{aligned}
\beta(w) &\leq \frac{(1-u)^2}{\psi(u)}\|Df(z)^{-1}f(w)\| \\
&= \frac{(1-u)^2}{\psi(u)}\|Df(z)^{-1}f(z) + \sum_{k=1}^{\infty}\frac{1}{k!}Df(z)^{-1}D^k f(z)(w-z)^k\| \\
&\leq \frac{(1-u)^2}{\psi(u)}\left(\|Df(z)^{-1}f(z)\| + \sum_{k=1}^{\infty}\frac{1}{k!}\|Df(z)^{-1}D^k f(z)\|\|w-z\|^k\right) \\
&\leq \frac{(1-u)^2}{\psi(u)}\left(\|Df(z)^{-1}f(z)\| + \|w-z\|\sum_{k=2}^{\infty}\frac{\|Df(z)^{-1}D^k f(z)\|}{k!}\|w-z\|^k\right) \\
&\leq \frac{(1-u)^2}{\psi(u)}\left(\beta(z) + \|w-z\|\sum_{k=1}^{\infty}u^{k-1}\right) \\
&= \frac{1-u}{\phi(u)}((1-u)\beta(z) + \|z-w\|).
\end{aligned}$$

From (1.3) we know that

$$\gamma(w)\sup_{k>1}\left(\frac{1}{k!}\|Df(w)^{-1}D^k f(w)\|\right)^{k-1}. \tag{1.12}$$

Consider the following term

$$\begin{aligned}
\frac{1}{k!}\|Df(w)^{-1}D^k f(w)\| &= \frac{1}{k!}\|Df(w)^{-1}\sum_{i=0}^{\infty}\frac{1}{i!}D^{k+i}f(z)(w-z)^i\| \\
&\leq \|Df(w)^{-1}Df(z)\|\|\sum_{i=0}^{\infty}\frac{1}{k!i!}Df(z)^{-1}D^{k+i}f(z)(w-z)^i\|.
\end{aligned}$$

Applying Lemma 1.3 we obtain

$$\begin{aligned}
\frac{1}{k!}\|Df(w)^{-1}D^k f(w)\| &\leq \frac{(1-u)^2}{\psi(u)}\sum_{i=0}^{\infty}\frac{1}{k!i!}\|Df(z)^{-1}D^{k+i}f(z)(w-z)^i\| \\
&\leq \frac{(1-u)^2}{\psi(u)}\sum_{i=0}^{\infty}\frac{(k+i)!}{k!i!}\frac{\|Df(z)^{-1}D^{k+i}f(z)\|}{(k+i)!}\|w-z\|^i \\
&\leq \frac{(1-u)^2}{\psi(u)}\sum_{i=0}^{\infty}\frac{(k+i)!}{k!i!}\gamma(z)^{k+i-1}\|w-z\|^i \\
&\leq \frac{(1-u)^2}{\psi(u)}\gamma(z)^{k-1}\sum_{i=0}^{\infty}\frac{(k+i)!}{k!i!}u^i \\
&= \frac{1}{\psi(u)}\left(\frac{\gamma(z)}{1-u}\right)^{k-1}
\end{aligned}$$

where the last step follows from Lemma 1.6. Applying this bound in (1.12) we obtain

$$\gamma(w) \leq \frac{\gamma(z)}{1-u}(\sup_{k>1}\psi(u)^{-1/(k-1)}) \leq \frac{\gamma(z)}{(1-u)\psi(u)}$$

since for $0 < u < 1 - 1/\sqrt{2}$, $\psi(u) < 1$.

Multiplying the bounds on $\beta(w)$ and $\gamma(w)$ above we obtain

$$\alpha(w) \leq \frac{(1-u)\alpha(z)+u}{\psi(u)^2} \leq \frac{\alpha(z)+u}{\psi(u)^2}$$

since u is positive. □

The above lemma along with (1.11) yields us: if w is such that $u(z,w) < 1 - 1/\sqrt{2}$ then

$$\|DN_f(w)\| \leq 2\frac{\alpha(z)+u(z,w)}{\psi(u(z,w))^2}.$$

Notice that if we show that the RHS in the above inequality is smaller than one then we know that N_f is a contraction map on the set $\overline{B}(z, \frac{u_0}{\gamma(z)})$, where u_0 is a constant smaller than $1-1/\sqrt{2}$. Let α_0 be a constant such that $\alpha(z) < \alpha_0$. Define $C_0 := 2\frac{\alpha_0+u_0}{\psi(u_0)^2}$. Then from the result above we know that for all $w \in \overline{B}(z, \frac{u_0}{\gamma(z)})$,

$$\|DN_f(w)\| \leq C_0.$$

Thus to show that N_f is a contracting operator on $\overline{B}(z, \frac{u_0}{\gamma(z)})$ it suffices to choose constants α_0 and $u_0 < 1 - 1/\sqrt{2}$ such that

- $C_0 < 1$ and
- N_f is closed on the set $\overline{B}(z, \frac{u_0}{\gamma(z)})$.

The second condition follows if for all $w \in \overline{B}(z, \frac{u_0}{\gamma(z)})$ we have

$$\|N_f(w) - z\| \leq \frac{u_0}{\gamma(z)}.$$

This would follow if

$$\|N_f(w) - N_f(z)\| + \beta(z) \leq \frac{u_0}{\gamma(z)}$$

or in other words if $\beta(z) \leq (1-C_0)\frac{u_0}{\gamma(z)}$, i.e., if $\alpha_0 \leq (1-C_0)u_0$. Thus we have shown the following:

Lemma 1.8. *Suppose there exist constants α_0, u_0 and $C_0 := \frac{2(\alpha_0+u_0)}{\phi(u_0)^2}$ which satisfy the following criteria:*

1. $0 \leq u_0 < 1 - 1/\sqrt{2}$,
2. $C_0 < 1$, *and*
3. $\alpha_0 \leq (1-C_0)u_0$.

Then for any z such that $\alpha(z) < \alpha_0$ N_f, is a contracting operator on $\overline{B}(z, \frac{u_0}{\gamma(z)})$ with contraction bound C_0.

Thus we know that there is a zero z^* of f in $\overline{B}(z, \frac{u_0}{\gamma(f,z)})$, and all the Newton iterates starting from z stay within this neighbourhood. To show that z is indeed an approximate zero, it suffices (from theorem 1.2) to show that

$$\|z - z^*\| \leq \frac{3-\sqrt{7}}{2\gamma_*}.$$

This would follow if

$$\frac{u_0}{\gamma(z)} \leq \frac{3-\sqrt{7}}{2\gamma_*}.$$

Since $z^* \in \overline{B}(z, \frac{u_0}{\gamma(f,z)})$ we know that $u(z^*, z) < u_0$. Thus we can apply the second result in Lemma 1.7 to obtain the following:

Lemma 1.9. *Suppose there exist constants α_0, u_0 and $C_0 := \frac{2(\alpha_0+u_0)}{\phi(u_0)^2}$ which satisfy the following criteria:*

1. $0 \leq u_0 < 1 - 1/\sqrt{2}$,
2. $C_0 < 1$,
3. $\alpha_0 \leq (1 - C_0)u_0$, *and*
4. $\frac{u_0}{(1-u_0)\psi(u_0)} \leq \frac{3-\sqrt{7}}{2}$.

If $z \in E$ is such that $\alpha(z) < \alpha_0$ then we have the following:
(a) N_f is a contracting operator on $\overline{B}(z, \frac{u_0}{\gamma(z)})$ with contraction bound C_0.
(b) z is an approximate zero of f, with the associated zero $z^* \in \overline{B}(z, \frac{u_0}{\gamma(z)})$.

One choice of constants is $u_0 = 0.1$ and $\alpha_0 = 0.03$. Thus we have the following point estimate:

Theorem 1.3 ([BCSS98, Thm. 2,p. 260])**.** *Any $z \in E$ such that $\alpha(f, z) < 0.03$ is an approximate zero of f, with the associated zero $z^* \in \overline{B}(z, \frac{0.1}{\gamma(f,z)})$.*

We next derive similar results for the weak model. Before we do that we derive some tight estimates on $\beta(z)$ and $\|z - z^*\|$, when z is an approximate zero.

Some tight estimates

We will later need a criterion for terminating Newton iteration starting from an approximate zero such that in the end we have approximated the associated zero to the desired precision. The criterion we use depends upon the value of $\beta(z) = \|Df(z)^{-1}f(z)\|$. There are two advantages of choosing the value of $\beta(z)$: first, it is computed in the course of the algorithm and hence is easily available at no extra cost; and second it is tightly related to $\|z - z^*\|$, as we will show shortly.

Lemma 1.10. *Let $z, w \in E$ and $u := \gamma(z)\|z - w\| < 1 - \frac{1}{\sqrt{2}}$. Then we have*

$$\frac{\psi(u)}{(1-u)^2} \leq \|Df(z)^{-1}Df(w)\| \leq (1-u)^{-2}.$$

Proof. Consider the upper bound first:

$$\begin{aligned}
\|Df(z)^{-1}Df(w)\| &= \|I + \sum_{k=1}^{\infty} \frac{1}{k!} Df(z)^{-1}D^{k+1}f(z)(w-z)^k\| \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{1}{k!}\|Df(z)^{-1}D^{k+1}f(z)\|\|w-z\|^k \\
&= 1 + \sum_{k=1}^{\infty}(k+1)u^k \\
&= (1-u)^{-2},
\end{aligned}$$

since $u < 1$ the last step follows from Lemma 1.6. For the lower bound, we proceed in a similar manner:

$$\begin{aligned}
\|Df(z)^{-1}Df(w)\| &= \|I + \sum_{k=1}^{\infty} \frac{1}{k!} Df(z)^{-1}D^{k+1}f(z)(w-z)^k\| \\
&\geq 1 - \sum_{k=1}^{\infty} \frac{1}{k!}\|Df(z)^{-1}D^{k+1}f(z)\|\|w-z\|^k \\
&\geq 1 - \sum_{k=1}^{\infty}(k+1)u^k \\
&= \frac{\psi(u)}{(1-u)^2},
\end{aligned}$$

again the last step follows Lemma 1.6. □

In the neighbourhood of a simple zero we have the following:

Lemma 1.11. *Let $z \in E$ be such that $u = \gamma(z^*)\|z - z^*\| < 1$, where $z^* \in E$ is a simple zero of f. Then*

$$\frac{\|z - z^*\|(1-2u)}{1-u} \leq \|Df(z^*)^{-1}f(z)\| \leq \frac{\|z - z^*\|}{1-u}.$$

Proof. Consider the upper bound first:

$$\begin{aligned}\|Df(z^*)^{-1}f(z)\| &= \|(z-z^*)+\sum_{k=2}^{\infty}\frac{1}{k!}Df(z^*)^{-1}D^kf(z^*)(z-z^*)^k\| \\ &\leq \|z-z^*\|(1+\sum_{k=1}^{\infty}u^k) \\ &= \frac{\|z-z^*\|}{1-u},\end{aligned}$$

where the last step holds since $u<1$. The lower bound can be shown in a manner similar to the way it was obtained in Lemma 1.10. □

Based on the two lemmas above we have the following tight relation between $\|z-z^*\|$ and $\|Df(z)^{-1}f(z)\|$:

Lemma 1.12. *If $z\in E$ is such that $u:=\gamma(z^*)\|z-z^*\|<1-\frac{1}{\sqrt{2}}$, where $z^*\in E$ is a simple zero of f, then*

$$\|z-z^*\|(1-2u)(1-u)\leq\|Df(z)^{-1}f(z)\|\leq\|z-z^*\|\frac{1-u}{\psi(u)}.$$

Proof. We only prove the upper bound:

$$\|Df(z)^{-1}f(z)\|\leq\|Df(z)^{-1}Df(z^*)\|\|Df(z^*)^{-1}f(z)\|\leq\|z-z^*\|\frac{1-u}{\psi(u)},$$

where the last step follows from the upper bound in Lemma 1.11, and the lower bound in Lemma 1.10 along with Lemma 1.2. The lower bound can be shown similarly. □

1.2.2 The Weak Model

We first adapt our definitions of Newton iteration and approximate zeros for the weak model.

For any $z_0\in E$ and some $0\leq\delta\leq 1$ define the **robust Newton sequence relative to** δ as a sequence $(\widetilde{z}_i)_{i\geq 0}$ such that $\widetilde{z}_0:=z_0$ and for all $i\geq 1$,

$$\widetilde{z}_{i+1}:=N_f(\widetilde{z}_i)\pm\delta, \tag{1.13}$$

where $\|\widetilde{z}_i\|$ is bounded. Recalling our error notation, (1.13) means $\|\widetilde{z}_{i+1}-N_f(\widetilde{z}_i)\|\leq\delta$. Note that the definition assumes that the robust Newton sequence relative to δ is well-defined.

A $z_0\in E$ is called a **weak approximate zero of f relative to** δ if there exists a zero z^* of f such that *any* robust Newton sequence $(\widetilde{z}_i)$ of z_0 relative to δ satisfies

$$\|\widetilde{z}_i-z^*\|\leq\max(2^{1-2^i}\|z_0-z^*\|,\kappa\delta), \tag{1.14}$$

for some constant $\kappa > 0$; the zero z^* is called the **associated zero** of z_0.

We will proceed by developing the two parts, corresponding to those mentioned in the beginning of §1.2. For the first part we have the following analog of Theorem 1.2:

Theorem 1.4. *Let z^* be a simple zero of f. Then any $z_0 \in E$ such that*

$$\gamma(z^*)\|z_0 - z^*\| \leq \frac{4-\sqrt{14}}{2} \text{ and } \gamma(z^*)\delta \leq \frac{4-\sqrt{14}}{2}$$

is a weak approximate zero of f relative to δ with z^* as its associated zero. That is, the robust Newton sequence $(\widetilde{z}_i)$ relative to δ, defined in (1.13), satisfies

$$\|\widetilde{z}_i - z^*\| \leq \max(2^{1-2^i}\|z_0 - z^*\|, 2\delta).$$

Proof. Our proof is by induction on i; the base case holds trivially. For sake of succinctness let $u_i := u_{\widetilde{z}_i}$, $\beta_i := \beta(\widetilde{z}_i)$ and $\gamma_i := \gamma(\widetilde{z}_i)$. Assuming the hypothesis holds for i, we consider two cases.

- **Case 1: $2\delta \leq 2^{1-2^{i+1}}\|z_0 - z^*\|$.**

  Our induction hypothesis in this case is

  $$\|\widetilde{z}_i - z^*\| \leq 2^{1-2^i}\|z_0 - z^*\|. \tag{1.15}$$

  In particular, this implies $u_i \leq u_0 < 1 - 1/\sqrt{2}$, for $i \geq 0$, and hence

  $$\|\widetilde{z}_{i+1} - z^*\| \leq \|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)\| + \|N_f(\widetilde{z}_i) - z^*\| \leq \delta + \frac{\gamma_*}{\psi(u_i)}\|\widetilde{z}_i - z^*\|^2, \tag{1.16}$$

  where the last step follows from (1.13) and Lemma 1.4. Furthermore, from Remark 1.1 we know that $\psi(u_i) \geq \psi(u_0)$, and hence we get

  $$\|\widetilde{z}_{i+1} - z^*\| \leq \delta + \frac{\gamma_*}{\psi(u_0)}\|\widetilde{z}_i - z^*\|^2.$$

  Applying the inductive hypothesis (1.15) and the condition of this case to this equation we get

  $$\|\widetilde{z}_{i+1} - z^*\| \leq 2^{-2^{i+1}}\|z_0 - z^*\| + \frac{u_0}{\psi(u_0)}2^{2-2^{i+1}}\|z_0 - z^*\|.$$

  Since $u_0 < \frac{4-\sqrt{14}}{2}$ we know $\frac{u_0}{\psi(u_0)} \leq \frac{1}{4}$. Thus

  $$\|\widetilde{z}_{i+1} - z^*\| \leq 2^{1-2^{i+1}}\|z_0 - z^*\|$$

  which proves the inductive step.

- **Case 2: $2\delta > 2^{1-2^{i+1}}\|\mathbf{z_0} - \mathbf{z}^*\|$.**

  In this case our induction hypothesis is

  $$\|\widetilde{z}_i - z^*\| \leq \max(2^{1-2^i}\|z_0 - z^*\|, 2\delta). \tag{1.17}$$

  This gives us

  $$u_i \leq 2\max(2^{-2^i}u_0, \delta\gamma_*) < 1 - 1/\sqrt{2}, \tag{1.18}$$

  since $2\delta\gamma_*, 2u_0 \leq 4 - \sqrt{14} < 1 - 1/\sqrt{2}$; thus (1.16) still holds. From (1.17) we also know that

  $$\|\widetilde{z}_i - z^*\|^2 \leq \max(2^{2-2^{i+1}}\|z_0 - z^*\|^2, 4\delta^2) \leq \max(4\delta\|z_0 - z^*\|, 4\delta^2),$$

  where the last step holds from the condition of the case. Now we consider two sub-cases:

  1. If $\delta \leq \|z_0 - z^*\|$ then from (1.18) we know $u_i \leq u_0$ and hence from (1.16) it follows that

     $$\|\widetilde{z}_{i+1} - z^*\| \leq \delta + \frac{4\delta\gamma_*}{\psi(u_0)}\|z_0 - z^*\| = \delta + 4\delta\frac{u_0}{\psi(u_0)} \leq 2\delta$$

     since $u_0 \leq \frac{4-\sqrt{14}}{2}$ implies $4\frac{u_0}{\psi(u_0)} \leq 1$.

  2. If $\delta > \|z_0 - z^*\|$ then from (1.18) we know $u_i \leq 2\delta$. Hence from (1.16) we get

     $$\|\widetilde{z}_{i+1} - z^*\| \leq \delta + 4\delta^2\frac{\gamma_*}{\psi(u_i)}.$$

     Since $u_i \leq 2\delta\gamma_* < 1 - 1/\sqrt{2}$, we know that $\psi(u_i) \geq \psi(2\delta\gamma_*)$. Thus

     $$\|\widetilde{z}_{i+1} - z^*\| \leq \delta + 4\delta^2\frac{\gamma_*}{\psi(2\delta\gamma_*)}.$$

     But $\delta\gamma_* \leq \frac{4-\sqrt{14}}{2}$ implies $4\delta\frac{\gamma_*}{\psi(2\delta\gamma_*)} \leq 1$, and hence

     $$\|\widetilde{z}_{i+1} - z^*\| \leq 2\delta.$$

  In both sub-cases we have proved the inductive step.

□

Based upon the above result we now derive the point estimate in the weak model. To achieve this we first prove the following analog of Lemma 1.9:

Lemma 1.13. *Suppose there exist constants α_0, u_0 and $C_0 := \frac{2(\alpha_0+u_0)}{\psi(u_0)^2}$ which satisfy the following criteria:*

1. $0 \leq u_0 < 1 - 1/\sqrt{2}$,
2. $C_0 < 1$,
3. $\alpha_0 \leq (1 - C_0)u_0$, *and*
4. $\frac{u_0}{(1-u_0)\psi(u_0)} \leq \frac{4-\sqrt{14}}{2}$.

If $z_0 \in E$ is such that $\alpha(z_0) + \gamma(z_0)\delta < \alpha_0$ and $\gamma(z_0)\delta \leq \frac{4-\sqrt{14}}{2}$ then we have the following:
(a) N_f is a contracting operator on $\overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$ with contraction bound C_0 and
(b) z_0 is a weak approximate zero of f relative to δ, with the associated zero $z^* \in \overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$.

Proof. The first part follows as a direct consequence of Lemma 1.9; thus we know that there is a zero z^* of f in $\overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$.

We will next show that all the iterates $\widetilde{z}_i$ are contained in the set $\overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$. This will be done using induction; the base case trivially holds. Inductively assume $\widetilde{z}_i \in \overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$. The distance between $\widetilde{z}_{i+1}$ and z_0 is

$$\begin{aligned} \|\widetilde{z}_{i+1} - z_0\| &\leq \|\widetilde{z}_{i+1} - N_f(z_i)\| + \|N_f(z_i) - N_f(z_0)\| + \|N_f(z_0) - z_0\| \\ &\leq \delta + C_0\|z_i - z_0\| + \beta(z_0) \end{aligned}$$

where the last step follows from (1.13) and the fact that N_f is a contracting operator. Moreover, from our inductive assumption we obtain

$$\|\widetilde{z}_{i+1} - z_0\| \leq \delta + C_0\frac{u_0}{\gamma(z_0)} + \beta(z_0).$$

Thus $\widetilde{z}_{i+1} \in \overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$ if

$$\delta + C_0\frac{u_0}{\gamma(z_0)} + \beta(z_0) \leq \frac{u_0}{\gamma(z_0)}$$

or if

$$\alpha(z_0) + \gamma(z_0)\delta \leq (1 - C_0)u_0$$

which is true by the conditions of the lemma.

To show that z_0 is a weak approximate w.r.t. δ with the associated zero z^*, we need to show, in addition to the above, that

$$\|z_0 - z^*\| \leq \frac{u_0}{\gamma(z_0)} \leq \frac{4 - \sqrt{14}}{2\gamma(z^*)}.$$

From Lemma 1.7 this follows if

$$\frac{u_0}{(1 - u_0)\psi(u_0)} \leq \frac{4 - \sqrt{14}}{2}.$$

□

By choosing $u_0 = 0.07$ and $\alpha_0 = 0.03$ we obtain the following point estimate in the weak model:

Theorem 1.5 (**Weak Point Estimate**). *Any $z_0 \in E$ such that*

$$\alpha(f, z_0) + \gamma(f, z_0)\delta < 0.03 \text{ and } \gamma(f, z_0)\delta \leq \frac{4 - \sqrt{14}}{2}$$

is a weak approximate zero of f relative to δ, with the associated zero $z^* \in \overline{B}(z_0, \frac{0.07}{\gamma(f,z_0)})$.

Deriving the point estimates in the strong model is straightforward, given the results in the weak model. This is our objective in the next section.

1.2.3 The Strong Model

We begin with adapting the definition of approximate zero in the weak model to our current setting.

For any $z_0 \in \mathbb{C}$ and $C \in \mathbb{R}$, a **robust iteration sequence of z_0 (relative to C and f)** is an infinite sequence $(\widetilde{z}_i)_{i\geq 0}$ such that $\widetilde{z}_0 = z_0$, and for all $i \geq 1$,

$$\widetilde{z}_i = \langle N_f(\widetilde{z}_{i-1})\rangle_{2^i+C}\,, \tag{1.19}$$

where $\|\widetilde{z}_i\|$ is bounded.

Our key definition is as follows: z_0 is a **robust approximate zero** of f if there exists a zero z^* of f such that for all C satisfying

$$2^{-C} \leq \|z_0 - z^*\|, \tag{1.20}$$

any robust iteration sequence $(\widetilde{z}_i)_{i\geq 0}$ of z_0 (relative to C and f) is such that for all $i \geq 0$,

$$\|\widetilde{z}_i - z^*\| \leq 2^{1-2^i}\|z_0 - z^*\|. \tag{1.21}$$

Call z^* the **associated zero** of z_0.

We have the following theorem as a direct consequence of Theorem 1.4:

Theorem 1.6. *Let z^* be a simple zero of f. Then any $z_0 \in E$ such that*

$$\gamma(z^*)\|z_0 - z^*\| \leq \frac{4 - \sqrt{14}}{2}$$

is a robust approximate zero of f with z^* as the associated zero.

Proof. It is straightforward to see that a robust iteration sequence of z_0 relative to C (and f) is a robust iteration sequence of z_0 relative to δ where $\delta = 2^{-1-C}$. Thus we only need to verify that $\gamma_* 2^{-1-C} \leq \frac{4-\sqrt{14}}{2}$. But this holds since

$$\gamma_* 2^{-1-C} \leq \frac{1}{2}\gamma_* \|z_0 - z^*\| < \frac{4-\sqrt{14}}{2}.$$

□

Similarly, we have the following analog of Lemma 1.13:

Lemma 1.14. *Suppose there exist constants α_0, u_0 and $C_0 := \frac{2(\alpha_0+u_0)}{\psi(u_0)^2}$ which satisfy the following criteria:*

1. $0 \leq u_0 < 1 - 1/\sqrt{2}$,
2. $C_0 < \frac{3}{4}$,
3. $\alpha_0 \leq (\frac{3}{4} - C_0)u_0$, *and*
4. $\frac{u_0}{(1-u_0)\psi(u_0)} \leq \frac{4-\sqrt{14}}{2}$.

If $z_0 \in E$ is such that $\alpha(z_0) < \alpha_0$ then we have the following:
(a) N_f is a contracting operator on $\overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$ with contraction bound C_0 and
(b) z_0 is a robust approximate zero of f with the associated zero $z^* \in \overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$.

Proof. We will show that robust Newton iterates $\widetilde{z}_i$ are contained in $\overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$. Proceeding in the same way as in Lemma 1.13, i.e., assuming that $\widetilde{z}_i \in \overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$, $\widetilde{z}_{i+1}$ will also be in $\overline{B}(z_0, \frac{u_0}{\gamma(z_0)})$ if

$$\alpha(z_0) + \gamma(z_0)2^{-2^{i+1}-C} \leq u_0(1 - C_0),$$

or if

$$\alpha(z_0) + \gamma(z_0)2^{-2-C} \leq u_0(1 - C_0),$$

since $i \geq 0$. Since $2^{-C} \leq \|z_0 - z^*\|$, the above follows if

$$\alpha(z_0) \leq (\frac{3}{4} - C_0)u_0$$

since $i \geq 1$. But this is true from the constraints of the lemma. □

Choosing $u_0 = 0.07$ and $\alpha_0 = .02$ we obtain the following generalization of [SDY05, Thm. 2]:

Theorem 1.7 (Robust Point Estimate). *Any $z_0 \in E$ such that $\alpha(f, z_0) < 0.02$ is a robust approximate zero of f, with the associated zero $z^* \in \overline{B}(z_0, \frac{0.07}{\gamma(f,z_0)})$.*

In the next section we develop similar results for approximate zeros of fourth kind.

1.3 Point Estimate of Fourth Kind

Malajovich [Mal93] was the first to derive point estimates for approximate zeros of the fourth kind. However, from his proofs the uniqueness of the associated zero is not clear. We give an alternative derivation of point estimates for approximate zeros of the fourth kind. Our approach follows the two steps mentioned in the beginning of §1.2. In order to achieve this, we propose an alternative definition of approximate zeros of the fourth kind which is stronger than the original definition, in the same sense as the definition of approximate zeros of the second kind is stronger than the definition of approximate zeros of the third kind.

Definition 1.8. Let $z_0 \in E$ be such that the sequence of Newton iterates (z_i), given by the recurrence $z_i := N_f(z_{i-1})$, is well defined. Then z_0 is called an **approximate zero of the fourth kind** if there is a unique zero z^* of f such that for all $i \in \mathbb{N}_{\geq 0}$

$$\frac{\|z_i - z^*\|}{\|z_i\|} \leq 2^{1-2^i} \frac{\|z_0 - z^*\|}{\|z_0\|}. \tag{1.22}$$

We call z^* the **associated zero** of z.

It is clear that if $\frac{\|z_0 - z^*\|}{\|z_0\|} \leq 1$ then the original definition follows. Our definition also helps us to utilize the results developed for approximate zeros of the second kind. To achieve this reuse, we modify the definitions of the three functions $\alpha(z)$, $\beta(z)$ and $\gamma(z)$ to take into account the presence of $\|z\|$ in the denominator of (1.22) above; this modification is originally given by Malajovich. Let the three functions be $\alpha'(z)$, $\beta'(z)$ and $\gamma'(z)$. Then the relation between these new definitions and the old ones is as follows:

$$\beta'(z) := \frac{\beta(z)}{\|z\|}, \tag{1.23}$$

$$\gamma'(z) := \max\{1, \|z\|\gamma(z)\}, \text{ and} \tag{1.24}$$

$$\alpha'(z) := \beta'(z)\gamma'(z) \geq \alpha(z). \tag{1.25}$$

Corresponding to $u(z, w)$ in (1.5), define the function

$$u'(z, w) := \frac{\gamma'(z)}{\|z\|}\|z - w\|. \tag{1.26}$$

For the special case when $z = z^*$ is a root of f we often use the shorthand γ'_* for $\gamma'(z^*)$ and u'_z for $u'(z^*, w)$; if z^* is a simple root of f, i.e., $Df(z^*)$ is non-singular, then these are well-defined quantities. Since $\gamma'(z^*) \geq \gamma(z^*)\|z^*\|$, we have

$$u'_z \geq \|z - z^*\|\gamma(z^*) = u_z. \tag{1.27}$$

Moreover, since $\gamma'(z) \geq 1$ we also have

$$u'(z,w) \geq \frac{\|z-w\|}{\|z\|}$$

, which gives us

$$1 - u'(z,w) \leq \frac{\|z\|}{\|w\|} \leq 1 + u'(z,w); \tag{1.28}$$

from this inequality it follows that if $0 \leq u'(z,w) < 1$ then both $\|z\|$ and $\|w\|$ are bounded away from zero. We start with deriving point estimates in the exact model for the definition of approximate zeros given above.

1.3.1 The Exact Model

The results here are derived from the analogous results in §1.2.1. We begin with deriving the analog of Lemma 1.4; this will instead depend upon the following lemma which follows in a straightforward manner from (1.27), Lemma 1.4 and Remark 1.1:

Lemma 1.15. *If $z \in E$ is such that $u'_z < 1 - 1/\sqrt{2}$ then*

$$\|N_f(z) - z^*\| \leq \frac{u'_z}{\psi(u'_z)}\|z - z^*\|.$$

Let $z' := N_f(z)$. From this lemma we know that

$$\begin{aligned} \frac{\|z' - z^*\|}{\|z'\|} &\leq \frac{u'_z}{\|z'\|\psi(u'_z)}\|z - z^*\| \\ &= \frac{\|z\|^2}{\|z'\|\|z^*\|}\frac{\gamma'_*}{\psi(u'_z)}\left(\frac{\|z - z^*\|}{\|z\|}\right)^2. \end{aligned}$$

Using the upper bound from (1.28) we obtain

$$\frac{\|z' - z^*\|}{\|z'\|} \leq \frac{\|z\|}{\|z'\|}\frac{(1+u'_z)\gamma'_*}{\psi(u'_z)}\left(\frac{\|z - z^*\|}{\|z\|}\right)^2. \tag{1.29}$$

Since $z' = N_f(z)$ we have

$$\frac{\|z'\|}{\|z\|} \geq \frac{\|z\| - \|Df(z)^{-1}\|\|f(z)\|}{\|z\|} \geq 1 - \beta'(z). \tag{1.30}$$

Using (1.27) we also have the following tight relation between $\beta'(z)$ and $\frac{\|z-z^*\|}{\|z^*\|}$:

Lemma 1.16. *If $z \in E$ is such that $u'_z < 1 - \frac{1}{\sqrt{2}}$, where $z^* \in E$ is a simple zero of f, then*

$$\frac{\|z - z^*\|}{\|z^*\|}(1 - 2u'_z)(1 - u'_z)(1 + u'_z)^{-1} \leq \beta'(z) \leq \frac{\|z - z^*\|}{\|z^*\|\psi(u'_z)}.$$

Proof. From Lemma 1.12 and (1.27) it follows that

$$\frac{\|z-z^*\|}{\|z^*\|}(1-2u'_z)(1-u'_z) \leq \beta'(z)\frac{\|z\|}{\|z^*\|} \leq \|z-z^*\|\frac{1-u'_z}{\psi(u'_z)}.$$

Applying the inequality from (1.28) gives us the desired result. □

Based upon the these results we have the following analog of Lemma 1.4:

Lemma 1.17. *If $z \in E$ is such that $u'_z < \frac{5-\sqrt{17}}{4}$, where $z^* \in E$ is a simple zero of f, then*

$$\frac{\|N_f(z)-z^*\|}{\|N_f(z)\|} \leq \frac{(1+u'_z)}{\phi(u'_z)}\gamma'_*\left(\frac{\|z-z^*\|}{\|z\|}\right)^2,$$

where

$$\phi(x) := \psi(x) - x = 1 - 5x + 2x^2. \tag{1.31}$$

Proof. From (1.29) and (1.30) we know that

$$\frac{\|z'-z^*\|}{\|z'\|} \leq \frac{(1+u'_z)\gamma'_*}{(1-\beta'(z))\psi(u'_z)}\left(\frac{\|z-z^*\|}{\|z\|}\right)^2.$$

Since $\gamma'_* \geq 1$, from the upper bound in Lemma 1.16 we know that

$$1-\beta'(z) \geq 1-\beta'(z)\gamma'_* \geq 1-\frac{u'_z}{\psi(u'_z)} \geq \frac{\phi(u'_z)}{\psi(u'_z)}.$$

Thus

$$\frac{\|z'-z^*\|}{\|z'\|} \leq \frac{(1+u'_z)\gamma'_*}{\phi(u'_z)}\left(\frac{\|z-z^*\|}{\|z\|}\right)^2.$$

Moreover, the right hand side is well defined since $u'_z < \frac{5-\sqrt{17}}{4}$ implies $\phi(u'_z) > 0$. □

Remark 1.2. *The smallest positive root of $\phi(x)$ is $\frac{5-\sqrt{17}}{4}$. Moreover, $\phi(x)$ is strictly decreasing from left to right in the interval $[0, \frac{5-\sqrt{17}}{4}]$.*

Based upon the results above we have the following analog to Theorem 1.2:

Theorem 1.9. *Let $z^* \in E$ be a simple zero of f. If $z \in E$ is such that*

$$\frac{\|z-z^*\|}{\|z^*\|}\gamma'(z^*) \leq 0.13$$

then z is an approximate zero fourth kind of f with z^* as the associated zero.

Proof. Let $u'_i := \frac{\|z_i-z^*\|}{\|z^*\|}\gamma'_*$. The proof is by induction. The base case $i = 0$ trivially holds. Suppose the hypothesis is true for i, i.e.,

$$\frac{\|z_i-z^*\|}{\|z_i\|} \leq 2^{1-2^i}\frac{\|z_0-z^*\|}{\|z_0\|}.$$

For $i \geq 1$ this implies $u'_i \frac{\|z^*\|}{\|z_i\|} \leq \frac{1}{2} u'_0 \frac{\|z^*\|}{\|z_0\|}$, or from (1.28) that $\frac{u'_i}{1+u'_i} \leq \frac{1}{2}\frac{u'_0}{1-u'_0} \leq u'_0$; from this we deduce that $u'_i \leq u'_0/(1-u'_0)$, which is smaller than $\frac{5-\sqrt{17}}{4}$ since $u'_0 \leq 0.13$. Thus from Lemma 1.17, along with the observation from Remark 1.2 that $\phi(u'_i) \geq \phi(u'_0)$, we obtain

$$\frac{\|z_{i+1} - z^*\|}{\|z_{i+1}\|} \leq \frac{(1+u'_0)}{\phi(u'_0)} \gamma'_* \left(\frac{\|z_i - z^*\|}{\|z\|} \right)^2.$$

Applying the inductive hypothesis we further get

$$\begin{aligned} \frac{\|z_{i+1} - z^*\|}{\|z_{i+1}\|} &\leq \frac{(1+u'_0)}{\phi(u'_0)} \gamma'_* \left(\frac{(1+u'_0)u'_0}{(1-u'_0)\phi(u'_0)} \right)^{2^{i+1}-2} \left(\frac{\|z_0 - z^*\|}{\|z_0\|} \right)^2 \\ &\leq \frac{(1+u'_0)u'_0}{\phi(u'_0)} \frac{\|z^*\|}{\|z_0\|} \left(\frac{(1+u'_0)u'_0}{(1-u'_0)\phi(u'_0)} \right)^{2^{i+1}-2} \frac{\|z_0 - z^*\|}{\|z_0\|} \\ &\leq \left(\frac{(1+u'_0)u'_0}{(1-u'_0)\phi(u'_0)} \right)^{2^{i+1}-1} \frac{\|z_0 - z^*\|}{\|z_0\|} \end{aligned}$$

where the last step follows from (1.28). This proves the inductive step. Furthermore, we can verify that $u'_{z_0} \leq 0.13$ implies $\frac{(1+u'_0)u'_0}{(1-u'_0)\phi(u'_0)} \leq \frac{1}{2}$. □

Based upon this theorem, we have the following analog of Lemma 1.9:

Lemma 1.18. *Suppose there exist constants α_0, u_0 and $C_0 := \frac{2(\alpha_0+u_0)}{\phi(u_0)^2}$ which satisfy the following criteria:*

1. $0 \leq u_0 < 1 - 1/\sqrt{2}$,
2. $C_0 < 1$,
3. $\alpha_0 \leq (1-C_0)u_0$, *and*
4. $\frac{u_0}{(1-u_0)\psi(u_0)} \leq 0.13$.

If $z \in E$ is such that $\alpha'(z) < \alpha_0$ then we have the following:
(a) N_f is a contracting operator on $\overline{B}(z, \frac{u_0\|z\|}{\gamma'(z)})$ with contraction bound C_0.
(b) z is an approximate zero of fourth kind of f, with the associated zero $z^* \in \overline{B}(z, \frac{u_0\|z\|}{\gamma'(z)})$.

Proof. The first part is a straightforward consequence of Lemma 1.9 and (1.24); thus we know that there is a root z^* of f in $\overline{B}(z, \frac{u_0\|z\|}{\gamma'(z)})$. For the second part observe that from Theorem 1.9 we know that z is an approximate zero of fourth kind with z^* as the associated zero if $\|z - z^*\| \leq \frac{0.13\|z^*\|}{\gamma'_*}$; since $z^* \in \overline{B}(z, \frac{u_0\|z\|}{\gamma'(z)})$ this follows if

$$\frac{u_0\|z\|\gamma'_*}{\gamma'(z)\|z^*\|} \leq 0.13. \tag{1.32}$$

For any $w \in \overline{B}(z, \frac{u_0\|z\|}{\gamma'(z)})$, let $u := u'(w,z) > \|z-w\|\gamma(z)$. If $u < 1 - 1/\sqrt{2}$, then from (1.24), Lemma 1.7 and Remark 1.1 we get that

$$\gamma'(w) \leq \max\left\{1, \frac{\|w\|}{\|z\|}\frac{\gamma'(z)}{(1-u)\psi(u)}\right\}.$$

But by definition we know that $\gamma'(z) \geq 1$; moreover, from (1.28) we know that $\frac{\|w\|}{\|z\|(1-u)} \geq 1$. Thus we have

$$\gamma'(w) \leq \frac{\|w\|}{\|z\|}\frac{\gamma'(z)}{(1-u)\psi(u)}. \tag{1.33}$$

Thus (1.32) follows from the fourth assumption of the lemma. □

One choice of constants is $u_0 = 0.06$ and $\alpha_0 = 0.03$. Thus we have the following point estimate for approximate zeros of fourth kind:

Theorem 1.10 (Point Estimate of Fourth Kind)**.** *Any $z \in E$ such that $\alpha'(f,z) < 0.03$ is an approximate zero of f, with the associated zero $z^* \in \overline{B}(z, \frac{0.06\|z\|}{\gamma'(f,z)})$.*

We now proceed to develop the same results in the weak model.

1.3.2 The Weak Model

We first modify the definition of a robust Newton sequence relative to δ from §1.2.2: For any $z_0 \in E$ and some $0 \leq \delta \leq 1$ define the **robust Newton sequence** $(\widetilde{z}_i)$ **relative to** δ as a sequence $(\widetilde{z}_i)_{i\geq 0}$ such that $\widetilde{z}_0 := z_0$, and for all $i \geq 1$

$$\frac{\|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_i\|} \leq \delta, \tag{1.34}$$

where $\|\widetilde{z}_i\|$ is bounded. Based upon this we have the following: A $z_0 \in E$ is called a **weak approximate zero of fourth kind of f relative to** δ if there exists a zero z^* of f such that *any* robust Newton sequence $(\widetilde{z}_i)$ of z_0 relative to δ satisfies

$$\frac{\|\widetilde{z}_i - z^*\|}{\|\widetilde{z}_i\|} \leq \max\left(2^{1-2^i}\frac{\|\widetilde{z}_0 - z^*\|}{\|\widetilde{z}_0\|}, \kappa\delta\right) \tag{1.35}$$

for some constant $\kappa > 0$; z^* is called the associated zero of z_0.

We have our first main result, an analog of Theorem 1.4:

Theorem 1.11. *Let $z^* \in E$ be a simple zero of f. Then any $z_0 \in E$ such that*

$$\gamma'(z^*)\frac{\|z_0 - z^*\|}{\|z^*\|} \leq 0.05 \text{ and } \gamma'(z^*)\delta \leq 0.01$$

is a weak approximate zero of fourth kind of f relative to δ, and z^* is the associated zero. That is, the robust Newton sequence $(\widetilde{z}_i)$, defined in (1.13), relative to δ satisfies

$$\frac{\|\widetilde{z}_i - z^*\|}{\|\widetilde{z}_i\|} \leq \max\left(2^{1-2^i}\frac{\|\widetilde{z}_0 - z^*\|}{\|\widetilde{z}_0\|}, 3\delta\right).$$

Proof. Our proof is by induction on i; the base case trivially holds. For sake of succinctness let $u'_i := u'_{\widetilde{z}_i}$, $\beta'_i := \beta'(\widetilde{z}_i)$ and $\gamma'_i := \gamma'(\widetilde{z}_i)$. Assume the hypothesis holds for i. We consider two cases.

- **Case 1: $3\delta \leq 2^{-2^{i+1}}\frac{\|\mathbf{z_0}-\mathbf{z}^*\|}{\|\mathbf{z_0}\|}$.**

  Our induction hypothesis in this case is

  $$\frac{\|\widetilde{z}_i - z^*\|}{\|\widetilde{z}_i\|} \leq 2^{1-2^i}\frac{\|\widetilde{z}_0 - z^*\|}{\|\widetilde{z}_0\|}. \tag{1.36}$$

  In particular, for $i \geq 1$ this implies $\frac{u'_i}{\|\widetilde{z}_i\|} \leq \frac{1}{2}\frac{u'_0}{\|\widetilde{z}_0\|}$, or from (1.28) that $\frac{u'_i}{1+u'_i} \leq \frac{1}{2}\frac{u'_0}{1-u'_0} \leq u'_0 < \frac{5-\sqrt{17}}{4}$. Thus we have

  $$\begin{aligned}\frac{\|\widetilde{z}_{i+1} - z^*\|}{\|\widetilde{z}_{i+1}\|} &\leq \frac{\|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_{i+1}\|} + \frac{\|N_f(\widetilde{z}_i) - z^*\|}{\|\widetilde{z}_{i+1}\|} \\ &\leq \delta\frac{\|\widetilde{z}_i\|}{\|\widetilde{z}_{i+1}\|} + \frac{\|N_f(\widetilde{z}_i) - z^*\|}{\|\widetilde{z}_{i+1}\|},\end{aligned}$$

  where the last step follows from (1.34). Furthermore, from Lemma 1.17 and the fact that $u'_i \leq u'_0$ implies $\phi(u'_i) \geq \phi(u'_0)$ (see Remark 1.2), we get

  $$\frac{\|\widetilde{z}_{i+1} - z^*\|}{\|\widetilde{z}_{i+1}\|} \leq \delta\frac{\|\widetilde{z}_i\|}{\|\widetilde{z}_{i+1}\|} + \frac{\|N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_{i+1}\|}\frac{(1+u'_0)\gamma'_*}{\phi(u'_0)}\left(\frac{\|\widetilde{z}_i - z^*\|}{\|\widetilde{z}_i\|}\right)^2, \tag{1.37}$$

  Now we derive upper bound on the fractions $\frac{\|N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_{i+1}\|}$ and $\frac{\|\widetilde{z}_i\|}{\|\widetilde{z}_{i+1}\|}$. Since $u'_i \leq u'_0 < 1-1/\sqrt{2}$, from Lemma 1.16 we know that

  $$\beta'_i \leq \frac{u'_i}{\psi(u'_i)(1-u'_i)} \leq \frac{u'_0}{\psi(u'_0)(1-u'_0)} \leq \frac{1}{4};$$

  and since $\delta\gamma'_* \leq 0.01$, we also know that $\delta < \frac{1}{4}$. Thus $\beta'_i + \delta < \frac{1}{2}$, and we have the following:

  1. $\frac{\|\widetilde{z}_i\|}{\|\widetilde{z}_{i+1}\|} < 2$. Since from (1.34) we have

     $$\begin{aligned} & \frac{\|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_i\|} && \leq && \delta \\ \Rightarrow\ & \frac{\|N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_i\|} - \delta && \leq && \frac{\|\widetilde{z}_{i+1}\|}{\|\widetilde{z}_i\|} \\ \Rightarrow\ & 1 - \beta'_i - \delta && \leq && \frac{\|\widetilde{z}_{i+1}\|}{\|\widetilde{z}_i\|} \\ \Rightarrow\ & \frac{1}{2} && < && \frac{\|\widetilde{z}_{i+1}\|}{\|\widetilde{z}_i\|}. \end{aligned}$$

2. $\frac{\|N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_{i+1}\|} < 3$. Since

$$\begin{aligned}\frac{\|N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_{i+1}\|} &\leq \frac{\|N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_i\|}\frac{\|\widetilde{z}_i|}{\|\widetilde{z}_{i+1}\|} \\ &= \frac{\|\widetilde{z}_i - Df(\widetilde{z}_i)^{-1}f(\widetilde{z}_i)\|}{\|\widetilde{z}_i\|}\frac{\|\widetilde{z}_i|}{\|\widetilde{z}_{i+1}\|} \\ &\leq \frac{1+\beta_i'}{1-(\beta_i'+\delta)} \\ &< 3.\end{aligned}$$

Plugging these bounds in (1.37) we infer that

$$\begin{aligned}\frac{\|\widetilde{z}_{i+1}-z^*\|}{\|\widetilde{z}_{i+1}\|} &< 2\delta + 3\frac{(1+u_0')\gamma_*'}{\phi(u_0')}\left(\frac{\|\widetilde{z}_i-z^*\|}{\|\widetilde{z}_i\|}\right)^2 \\ &< 2\delta + 3\frac{u_0'(1+u_0')}{\phi(u_0')}2^{2-2^{i+1}}\frac{\|z_0-z^*\|}{\|z_0\|} \quad \text{(from (1.36))}.\end{aligned}$$

Since $3\delta \leq 2^{2-2^{i+1}}\frac{\|z_0-z^*\|}{\|z^*\|}$ we further obtain

$$\begin{aligned}\frac{\|\widetilde{z}_{i+1}-z^*\|}{\|\widetilde{z}_{i+1}\|} &< 2^{-2^{i+1}}\frac{\|z_0-z^*\|}{\|z^*\|} + 3\frac{u_0'(1+u_0')}{\phi(u_0')}2^{2-2^{i+1}}\frac{\|z_0-z^*\|}{\|z_0\|} \\ &< 2^{1-2^{i+1}}\frac{\|z_0-z^*\|}{\|z_0\|},\end{aligned}$$

where the last step follows from the fact that $u_0' \leq 0.05$ implies $12\frac{u_0'(1+u_0')}{\phi(u_0')} \leq 1$.

- **Case 2: $3\delta > 2^{2-2^{i+1}}\frac{\|\mathbf{z_0}-\mathbf{z}^*\|}{\|\mathbf{z}^*\|}$.**

Our induction hypothesis in this case is

$$\frac{\|\widetilde{z}_i - z^*\|}{\|\widetilde{z}_i\|} \leq \max\left\{3\delta, 2^{1-2^i}\frac{\|z_0-z^*\|}{\|z_0\|}\right\}. \tag{1.38}$$

Multiplying throughout by γ_*', and applying (1.28) gives us for $i \geq 1$

$$\frac{u_i'}{1+u_i'} \leq \max\{3\gamma_*'\delta, u_0'\},$$

and since $u_0' \leq 0.05$ and $3\gamma_*'\delta \leq 0.03$, we know that $u_i' < \frac{5-\sqrt{17}}{4}$. Thus we still have

$$\frac{\|\widetilde{z}_{i+1} - z^*\|}{\|\widetilde{z}_{i+1}\|} \leq 2\delta + 3\frac{(1+u_0')\gamma_*'}{\phi(u_0')}\left(\frac{\|\widetilde{z}_i - z^*\|}{\|\widetilde{z}_i\|}\right)^2. \tag{1.39}$$

Moreover, from (1.38) we obtain

$$\begin{aligned}\left(\frac{\|\widetilde{z}_i - z^*\|}{\|\widetilde{z}_i\|}\right)^2 &\leq \max\left\{9\delta^2, 2^{2-2^{i+1}}\left(\frac{\|z_0-z^*\|}{\|z_0\|}\right)^2\right\} \\ &\leq 3\delta\max\left\{3\delta, \frac{\|z_0-z^*\|}{\|z_0\|}\right\},\end{aligned}$$

where the last step follows from the condition of the case. Now if $3\delta \leq \frac{\|z_0 - z^*\|}{\|z_0\|}$, then from (1.39) we get

$$\frac{\|\widetilde{z}_{i+1} - z^*\|}{\|\widetilde{z}_{i+1}\|} \leq 2\delta + 9\delta \frac{(1+u_0')u_0'}{\phi(u_0')} \leq 3\delta,$$

since $u_0' \leq 0.05$ implies $9\frac{(1+u_0')u_0'}{\phi(u_0')} \leq 1$. On the other hand if $3\delta > \frac{\|z_0 - z^*\|}{\|z_0\|}$, then from (1.39) we obtain

$$\frac{\|\widetilde{z}_{i+1} - z^*\|}{\|\widetilde{z}_{i+1}\|} \leq 2\delta + 3\frac{(1+u_0')\gamma_*'}{\phi(u_0')} \cdot 9\delta^2 \leq 3\delta,$$

since $27\gamma_*'\delta\frac{1+u_0'}{\phi(u_0')} \leq 1$ for $\gamma_*'\delta \leq 0.01$ and $u_0' \leq 0.05$. In either situation the inductive step holds.

□

We now derive the point estimate for the weak model.

Lemma 1.19. *Suppose there exists constants α_0, β_0, $C_0 := \frac{2\alpha_0 + u_0}{\psi(u_0)^2}$, and $0 \leq \delta \leq 1$ be such that*

1. $0 \leq u_0 < 1 - \frac{1}{\sqrt{2}}$,
2. $C_0 < \frac{1}{2}$, $\delta < \frac{1}{2}$,
3. $\alpha_0 \leq (1 - C_0 - \delta)u_0$, *and*
4. $\frac{u_0}{(1-u_0)\psi(u_0)} \leq 0.13$.

If $z_0 \in E$ is such that $\alpha'(z_0) + \gamma'(z_0)\delta < \alpha_0$ and $\gamma'(z_0)\delta \leq 0.01$ then we have the following:
(a) N_f is a contracting operator on $\overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$ with contraction bound C_0.
(b) z_0 is a weak approximate zero of the fourth kind of f relative to δ, with the associated zero $z^* \in \overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$.

Proof. The first part follows directly from Lemma 1.18. We now show that the iterates $\widetilde{z}_i$ are contained in $\overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$. We prove this by induction on i. Clearly, $z_0 \in \overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$. Assume that

$$\|\widetilde{z}_i - z_0\| \leq \frac{u_0\|z_0\|}{\gamma'(z_0)}. \tag{1.40}$$

Then $\|\widetilde{z}_{i+1} - z_0\| \leq \frac{u_0\|z_0\|}{\gamma'(z_0)}$ if

$$\|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)\| + \|N_f(\widetilde{z}_i) - N_f(z_0)\| + \|N_f(z_0) - z_0\| \leq \frac{u_0\|z_0\|}{\gamma'(z_0)}.$$

Applying (1.34), and the fact that N_f is a contraction map on $\overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$, the inequality above follows if

$$|\widetilde{z}_i\|\delta + C_0\|\widetilde{z}_i - z_0\| + \|z_0\|\beta'(z_0) \leq \frac{u_0\|z_0\|}{\gamma'(z_0)}.$$

From (1.40), we know that this inequality follows if

$$\|z_0\|(1 + \frac{u_0}{\gamma'(z_0)})\delta + C_0 u_0 \frac{\|z_0\|}{\gamma'(z_0)} + \|z_0\|\beta'(z_0) \leq \frac{u_0\|z_0\|}{\gamma'(z_0)},$$

or equivalently if

$$\alpha'(z_0) + \gamma'(z_0)\delta \leq u_0(1 - C_0 - \delta).$$

But this is true since by assumption $\alpha'(z_0) + \gamma'(z_0)\delta < \alpha_0 \leq u_0(1 - C_0 - \delta)$.

To show that z_0 is a weak approximate zero of the fourth kind with z^* as the associated zero, it suffices to show that

$$\frac{u_0\|z_0\|}{\gamma'(z_0)} \leq \frac{0.13\|z*\|}{\gamma'_*} \text{ and } \gamma'_*\delta \leq 0.01.$$

From (1.33), the first condition follows if $\frac{u_0}{(1-u_0)\psi(u_0)} \leq 0.13$; the second condition is true by assumption. □

One choice of constants is $u_0 = 0.06$ and $\alpha_0 = 0.01$. Thus our point estimate is

Theorem 1.12. *If $z_0 \in E$ is such that $\alpha'(z_0) + \gamma'(z_0)\delta < 0.01$ and $\gamma'(z_0)\delta \leq 0.01$ then z_0 is a weak approximate zero and the associated zero $z^* \in \overline{B}(z, \frac{0.06\|z_0\|}{\gamma'(z_0)})$.*

1.3.3 The Strong Model

We now derive point estimates in the strong bigfloat model. Deriving these results is straightforward given the results in the weak model above. Again, we first modify the definition of a robust iteration sequence relative to C and f as given in §1.2.3: For any $z_0 \in \mathbb{C}$ and $C \in \mathbb{R}$, a **robust iteration sequence of** z_0 (relative to C and f) is an infinite sequence $(\widetilde{z}_i)_{i\geq 0}$ such that $\widetilde{z}_0 = z_0$, and for all $i \geq 1$,

$$\frac{\|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)\|}{\|\widetilde{z}_i\|} \leq \delta_{i+1} := 2^{-2^{i+1}-C}, \tag{1.41}$$

where $\|\widetilde{z}_i\|$ is bounded. It is not hard to see that the definition is similar to (1.34), except δ varies with i now.

Based upon the above we have the following: z_0 is a **robust approximate zero of fourth kind** of f if there exists a zero z^* of f such that for all C satisfying

$$2^{-C} \leq \frac{\|z_0 - z^*\|}{\|z_0\|}, \tag{1.42}$$

any robust iteration sequence $(\widetilde{z}_i)_{i\geq 0}$ of z_0 (relative to C and f) is such that for all $i \geq 0$,

$$\frac{\|\widetilde{z}_i - z^*\|}{\|\widetilde{z}_i\|} \leq 2^{1-2^i} \frac{\|z_0 - z^*\|}{\|z_0\|}. \tag{1.43}$$

Call z^* the **associated zero** of z_0.

As a direct consequence of Theorem 1.11, we have

Theorem 1.13. *Let $z^* \in E$ be a simple root of f. Then any z_0 such that*

$$\frac{\|z_0 - z^*\|}{\|z^*\|}\gamma'_* \leq 0.01$$

is a robust approximate zero of fourth kind of f.

Proof. We only need to show that under the above constrains $\gamma'_*\delta_0 \leq 0.01$. But from (1.41) we know that

$$\gamma'_*\delta_0 = \gamma'_* 2^{-1-C} \leq \frac{1}{2}\gamma'_* \frac{\|z_0 - z^*\|}{\|z_0\|} \leq \frac{1}{2}\frac{u'_{z_0}}{1-u'_{z_0}} < u'_{z_0} \leq 0.01.$$

□

Similarly, corresponding to Lemma 1.19 we have:

Lemma 1.20. *Suppose there exists constants α_0, β_0, and $C_0 := \frac{2\alpha_0+u_0}{\psi(u_0)^2}$ such that*

1. $0 \leq u_0 < 1 - \frac{1}{\sqrt{2}}$,
2. $C_0 < \frac{1}{2}$,
3. $\alpha_0 \leq u_0(1 - C_0 - \frac{u_0+1}{4})$, *and*
4. $\frac{u_0}{(1-u_0)\psi(u_0)} \leq 0.01$.

If $z_0 \in E$ is such that $\alpha'(z_0) < \alpha_0$ then we have the following:
(a) N_f is a contracting operator on $\overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$ with contraction bound C_0.
(b) z_0 is a robust approximate zero of the fourth kind with the associated zero $z^* \in \overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$.

Proof. We will show that under the above constraints the robust Newton iterates as defined in (1.41) are contained in $\overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$. Following a line of argument similar to Lemma 1.19, i.e., assuming $\widetilde{z}_i \in \overline{B}(z_0, \frac{u_0\|z_0\|}{\gamma'(z_0)})$, $\widetilde{z}_{i+1}$ is also in the same region if

$$\alpha'(z_0) + \gamma'(z_0)2^{-2^{i+1}-C} \leq u_0(1 - C_0 - 2^{-2^{i+1}-C}).$$

Since $i \geq 0$, and $2^{-C} \leq \frac{\|z_0 - z^*\|}{\|z_0\|}$, the inequality above follows if we show that

$$\alpha'(z_0) + \frac{u_0}{4} \leq u_0(1 - C_0 - \frac{u_0}{4}).$$

But this is straightforward from the condition that $\alpha'(z_0) < \alpha_0 \leq u_0(1 - C_0 - \frac{u_0+1}{4})$.

To show that z_0 is a robust approximate zero of fourth kind, from Theorem 1.13 it suffices to show that $\|z_0 - z^*\| \leq \frac{0.01\|z^*\|}{\gamma'(z^*)}$, or that $\frac{u_0\gamma'(z^*)\|z_0\|}{\gamma'(z_0)\|z^*\|} \leq 0.01$; but this follows from (1.33) since $\frac{u_0}{\psi(u_0)(1-u_0)} \leq 0.01$. □

A judicious choice of the constants is $u_0 = 0.009$ and $\alpha_0 = 0.006$. This gives us a point estimate in the strong setting:

Theorem 1.14. *If $z_0 \in E$ is such that $\alpha'(z_0) < 0.006$ then z_0 is a robust approximate zero of fourth kind of f, and the associated zero $z^* \in \overline{B}(z_0, \frac{0.009\|z_0\|}{\gamma'(z_0)})$.*

Given the dependency of the results for approximate zeros of fourth kind on those of the second kind, from now on we only focus on the latter kind of zeros.

1.4 One Step of Robust Newton

In this section we give the details of how to implement one step of the robust Newton method when f is a system of multivariate polynomials, i.e., given the $(i-1)$-th iterate $\widetilde{z}_{i-1}$ how to obtain $\widetilde{z}_i$ such that

$$\widetilde{z}_i = \langle N_f(\widetilde{z}_{i-1}) \rangle_{2^i + C},$$

for some $C \geq 0$. We give the details to compute $N_f(z) \pm \delta$ for any z and $\delta \geq 0$, since by choosing $\delta := 2^{-2^i - C}$ we will get the desired result.

We start with some definitions that will be used subsequently:

Definition 1.15.

1. Let $\mathcal{F} : \mathbb{C}^n \to \mathbb{C}^n$ be a zero-dimensional system of n integer polynomials $F_1, \ldots, F_n \in \mathbb{Z}[Z_1, \ldots, Z_n]$, i.e., the system has only finitely many common roots.
2. Let D_i be the degree of F_i and $\mathcal{D} := \max(D_1, \ldots, D_n)$.
3. Let $S(F_i)$ be the number of non-zero coefficients in F_i, and $S(\mathcal{F})$ be the number of non-zero coefficients in the whole system.
4. Let $J_{\mathcal{F}}(\mathbf{Z})$ be the Jacobian matrix of $\mathcal{F}$ at the point $\mathbf{Z} \in \mathbb{C}^n$, i.e.,

$$J_{\mathcal{F}}(\mathbf{Z}) := \left[\frac{\partial F_i}{\partial Z_j}(\mathbf{Z})\right]_{i,j}, \text{ for } 1 \leq i, j \leq n.$$

5. Let $\hat{\mathcal{F}} : \mathbb{C}^{n+1} \to \mathbb{C}^n$ represent the homogenized version of $\mathcal{F}$, i.e., the polynomials F_i are homogenized to $\hat{F}_i \in \mathbb{Z}[Z_0, Z_1, \ldots, Z_n]$ by introducing a new variable Z_0 such that degree of $\hat{F}_i$ is D_i.

6. The norm $\|\cdot\|$ is the max-norm, i.e., $\|\mathbf{Z}\| = \max(|Z_1|, \ldots, |Z_n|)$; the matrix norm is the corresponding operator norms.

7. The Newton operator is $N_{\mathcal{F}}(\mathbf{Z}) := \mathbf{Z} - J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})$, where $\mathbf{Z} \in \mathbb{C}^n$.

We make the following assumption:

Our input point $\mathbf{Z} \in \mathbb{F}^n$ is a robust approximate zero, such that $\alpha(\mathcal{F}, \mathbf{Z}) < 0.02$, with the associated root $\mathbf{Z}^*$. Moreover, $\mathbf{Z}$ and $\mathbf{Z}^*$ are such that

$$\|\mathbf{Z}\|, \|\mathbf{Z}^*\| \leq B(\mathcal{F}), \tag{1.44}$$

where

$$\begin{aligned} B(\mathcal{F}) :=& (2^{1.5}NK)^{\mathcal{D}'} 2^{(n+1)D_1 \cdots D_n}, \\ N :=& \binom{1 + \sum D_i}{n}, \\ K :=& \max(\sqrt{n}H(\mathcal{F})) \end{aligned} \tag{1.45}$$

and

$$\mathcal{D}' := (1 + \sum D_i^{-1}) \prod D_j.$$

This is a reasonable assumption, because from [Yap00, Cor. 11.49,p. 355] we know that the norm $\|\mathbf{Z}^*\| \leq B(\mathcal{F})$, so without loss of generality we may assume that $\|\mathbf{Z}\|$ satisfies the same.

The algorithm to compute one step of the Newton method is fairly standard [Tis01, Hig96]; it takes as input $\mathcal{F}$ and $\mathbf{Z} \in \mathbb{F}^n$ such that $\alpha(\mathcal{F}, \mathbf{Z}) < 0.02$, and produces an output $\mathbf{Z}' \in \mathbb{F}^n$ such that

$$\|\mathbf{Z}' - N_{\mathcal{F}}(\mathbf{Z})\| \leq \delta. \tag{1.46}$$

The algorithm is as follows:

1. Compute the vector $\mathcal{F}(\mathbf{Z})$ and the matrix $J_{\mathcal{F}}(\mathbf{Z})$ exactly.
2. Compute matrices P_1, P_2, $\hat{L}$, and $\hat{U}$ using Gaussian elimination with partial pivoting such that $P_1 J_{\mathcal{F}}(\mathbf{Z}) P_2 = \hat{L}\hat{U}$.
3. Compute $\widetilde{w} = \hat{L}^{-1}\mathcal{F}(\mathbf{Z})$ by forward substitution.

4. Compute $\widetilde{v} = \hat{U}^{-1}\widetilde{w}$ by backward substitution.

5. Return $\mathbf{Z}' := \mathbf{Z} - \widetilde{v}$.

Since ring operations are exact and $\mathcal{F}$ is a system of integer polynomials, we know that the first and the last step have no errors. For steps 2,3 and 4 we use the weak model of computation where all the operations (including ring operations) are done to a fixed precision ϵ. The advantage of doing this is not in the implementation, but in the analysis since the three sub-routines (Gaussian elimination, forward and backward substitution) are very well studied in the weak model. Our aim is to bound ϵ such that $\mathbf{Z}'$ satisfies (1.46).

From the definition of $N_{\mathcal{F}}(\mathbf{Z})$ we can verify that

$$\|\mathbf{Z}' - N_{\mathcal{F}}(\mathbf{Z})\| = \|\widetilde{v} - J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\|.$$

Using backwardly stable algorithms for Gaussian elimination, forward substitution and backward substitution we know from [Hig96, p. 177] that

$$(J_{\mathcal{F}}(\mathbf{Z}) + \Delta)\widetilde{v} = \mathcal{F}(\mathbf{Z})$$

where

$$\|\Delta\| \leq n^3 2^{n+2}\|J_{\mathcal{F}}(\mathbf{Z})\|\frac{\epsilon}{1-3n\epsilon}. \tag{1.47}$$

Thus

$$\begin{aligned}
\|\widetilde{v} - J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\| &= \|(J_{\mathcal{F}}(\mathbf{Z}) + \Delta)^{-1}\mathcal{F}(\mathbf{Z}) - J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\| \\
&= \|(I + J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta)^{-1}J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z}) - J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\| \\
&\leq \|(I + J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta)^{-1} - I\|\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\|.
\end{aligned}$$

Choose ϵ such that $\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta\| \leq \frac{1}{2}$. Then from Lemma 1.1 we know that

$$(I + J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta)^{-1} - I = \sum_{i=1}^{\infty}(-J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta)^i$$

and hence

$$\begin{aligned}
\|(I + J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta)^{-1} - I\| &\leq \frac{\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta\|}{1 - \|J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta\|} \\
&\leq 2\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\Delta\| \\
&\leq 2\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\|\|\Delta\|.
\end{aligned}$$

This yields

$$\|\widetilde{v} - J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\| \leq 2\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\|\|\Delta\|\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\|.$$

Plugging in the upper bound from (1.47) we get that

$$\|\widetilde{v} - J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\| \leq n^3 2^{n+3}\kappa(J_{\mathcal{F}}(\mathbf{Z}))\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathcal{F}(\mathbf{Z})\|\frac{\epsilon}{1-3n\epsilon}, \tag{1.48}$$

where $\kappa(M)$ represents the condition number of a matrix M. Suppose that $3n\epsilon \leq \frac{1}{2}$, in addition to the earlier restriction on ϵ, then along with the definition of $\beta(\mathcal{F}, \mathbf{Z})$ (Equation (1.2)) we have

$$\|\mathbf{Z}' - N_{\mathcal{F}}(\mathbf{Z})\| \leq n^3 2^{n+4}\kappa(J_{\mathcal{F}}(\mathbf{Z}))\beta(\mathcal{F}, \mathbf{Z})\epsilon \leq n^3 2^{n+7}\kappa(J_{\mathcal{F}}(\mathbf{Z}))B(\mathcal{F})\epsilon,$$

since from our assumption $\alpha(\mathcal{F}, \mathbf{Z}) < 0.02$ we know from Theorem 1.7 that $u = 0.07$, and hence from Lemma 1.12 we obtain

$$\frac{1}{2}\|\mathbf{Z} - \mathbf{Z}^*\| \leq \beta(\mathcal{F}, \mathbf{Z}) \leq 4\|\mathbf{Z} - \mathbf{Z}^*\|; \tag{1.49}$$

and from (1.44) we know that $\|\mathbf{Z} - \mathbf{Z}^*\| \leq 2B(\mathcal{F})$.

Thus we have the following:

Lemma 1.21. *Let $\mathcal{F}$ be a system of integer polynomials and $\mathbf{Z} \in \mathbb{F}^n$ be a robust approximate zero such that $\alpha(\mathcal{F}, \mathbf{Z}) < 0.02$ with the associated zero $\mathbf{Z}^*$; moreover, assume $\mathbf{Z}, \mathbf{Z}^*$ satisfy (1.44). The amount of "machine precision" ϵ with which we need to do Gaussian elimination, forward and backward substitution in one step of robust Newton applied to $\mathbf{Z}$ such that the computed output $\mathbf{Z}'$ satisfies (1.46) is bounded by the maximum of*

$$\delta\left(n^3 2^{n+7}\kappa(J_{\mathcal{F}}(\mathbf{Z}))B(\mathcal{F})\right)^{-1} \text{ and } (6n)^{-1},$$

where $B(\mathcal{F})$ is defined as in (1.45).

Based upon the above lemma we next give the robust Newton iteration that takes as input a robust approximate zero and approximates the associated zero to any precision.

1.5 Robust Newton Iteration

In this section we generalize the algorithm in [SDY05] to a system of integer polynomials. The algorithm will take as input a system $\mathcal{F}$ and a robust approximate zero $\mathbf{Z}_0 \in \mathbb{F}^n$ such that

$\alpha(\mathcal{F}, \mathbf{Z}_0) < 0.02$, and the associated zero $\mathbf{Z}^*$ and $\mathbf{Z}_0$ satisfy (1.44); it will construct a robust iteration sequence $(\widetilde{\mathbf{Z}}_i)$, relative to C, such that

$$\widetilde{\mathbf{Z}}_i = \left\langle N_{\mathcal{F}}(\widetilde{\mathbf{Z}}_{i-1}) \right\rangle_{2^i+C}.$$

In order to do so, we first need to determine a C such that $2^{-C} \leq \|\mathbf{Z}_0 - \mathbf{Z}^*\|$, where $\mathbf{Z}^* \in \mathbb{R}^n$ is the associated zero of $\mathbf{Z}_0$.

NOTE: We restrict ourselves to the case when the zero is in $\mathbb{R}^n$. To handle the case when the zero is in $\mathbb{C}^n$, we need to compute with **Gaussian bigfloats** instead of bigfloats, i.e., members of the ring $\mathbb{F}[i]$, $i^2 = -1$, .

1.5.1 Distance between an approximate zero and its associated zero

Since $\mathcal{F}$ is a system of integer polynomials we can compute $\mathcal{F}(\mathbf{Z}_0)$ and $J_{\mathcal{F}}(\mathbf{Z}_0)$ exactly. Now compute $\widetilde{v}_0$ such that $J_{\mathcal{F}}(\mathbf{Z}_0)\widetilde{v}_0 = \mathcal{F}(\mathbf{Z}_0)$, where the precision used in solving the system is such that it satisfies Lemma 1.21 with $\delta = 1/4$. Then from (1.48) we obtain

$$\frac{\|\widetilde{v}_0 - J_{\mathcal{F}}(\mathbf{Z}_0)^{-1}\mathcal{F}(\mathbf{Z}_0)\|}{\|J_{\mathcal{F}}(\mathbf{Z}_0)^{-1}\mathcal{F}(\mathbf{Z}_0)\|} \leq \frac{1}{4}. \tag{1.50}$$

Thus $\widetilde{v}_0$ is a relative approximation to $J_{\mathcal{F}}(\mathbf{Z}_0)^{-1}\mathcal{F}(\mathbf{Z}_0)$; this also implies

$$3\beta(\mathcal{F}, \mathbf{Z}_0) \leq 4\|\widetilde{v}_0\| \leq 5\beta(\mathcal{F}, \mathbf{Z}_0). \tag{1.51}$$

Since $\alpha(\mathcal{F}, \mathbf{Z}_0) < 0.02$ we know from Theorem 1.7 that $u = 0.07$, and hence from Lemma 1.12 we obtain

$$\frac{1}{2}\|\mathbf{Z}_0 - \mathbf{Z}^*\| \leq \beta(\mathcal{F}, \mathbf{Z}_0) \leq 4\|\mathbf{Z}_0 - \mathbf{Z}^*\|. \tag{1.52}$$

Combining this with (1.51) we get

$$\frac{3}{8}\|\mathbf{Z}_0 - \mathbf{Z}^*\| \leq \|\widetilde{v}_0\| \leq 5\|\mathbf{Z}_0 - \mathbf{Z}^*\|.$$

Thus we have the following lemma on computing C:

Lemma 1.22. *Let $\mathbf{Z}_0$ be a robust approximate zero such that $\alpha(\mathcal{F}, \mathbf{Z}_0) < 0.02$ and $\mathbf{Z}^*$ be its associated zero; moreover, assume $\mathbf{Z}_0$ and $\mathbf{Z}^*$ satisfy (1.44). Let $\widetilde{v}_0$ be the solution to the linear system $J_{\mathcal{F}}(\mathbf{Z}_0)\mathbf{X} = \mathcal{F}(\mathbf{Z}_0)$, where the precision used in solving the system is*

$$\left(n^3 2^{n+9} \kappa(J_{\mathcal{F}}(\mathbf{Z}_0)) B(\mathcal{F})\right)^{-1},$$

$B(\mathcal{F})$ is defined as in (1.45). Then $C := 3 - \lfloor \log \|\widetilde{v}_0\| \rfloor$ satisfies

$$\frac{1}{8}\|\mathbf{Z}_0 - \mathbf{Z}^*\| < 2^{-C} < \|\mathbf{Z}_0 - \mathbf{Z}^*\|.$$

The complete robust Newton iteration is the following:

Algorithm RN

INPUT: A zero-dimensional system of multi-variate integer polynomials $\mathcal{F}$, precision $p \in \mathbb{N}_{\geq 0}$ and a $\mathbf{Z}_0 \in \mathbb{F}^n$ such that $\alpha(\mathcal{F}, \mathbf{Z}_0) < 0.02$.

OUTPUT: $\langle \mathbf{Z}^* \rangle_p$, where $\mathbf{Z}^*$ is the associated zero of $\mathbf{Z}_0$.

1. Let $\widetilde{v}_0$ be as described in Lemma 1.22. Let $C := 3 - \lfloor \log \|\widetilde{v}_0\| \rfloor$.
2. Assign $\widetilde{\mathbf{Z}}_0 := \left\langle \widetilde{\mathbf{Z}}_0 \right\rangle_{C+2}$.
3. do
   Compute $\mathcal{F}(\widetilde{\mathbf{Z}}_i)$, $J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)$ and an upper bound $\widetilde{\kappa}$ on $\kappa(J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i))$.
   Compute solution $\widetilde{v}_i$ to $J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)\mathbf{X} = \mathcal{F}(\widetilde{\mathbf{Z}}_i)$ using Gaussian elimination with partial pivoting where all operations are done to precision $\epsilon_i = 2^{-2^i} 2^{-C} (n^3 2^{n+7} B(\mathcal{F}) \widetilde{\kappa})^{-1}$.
   Let $\widetilde{\mathbf{Z}}_{i+1} := \widetilde{\mathbf{Z}}_i - \widetilde{v}_i$.
   while($\widetilde{v}_i \neq 0$ and $\|\widetilde{v}_i\| \geq 2^{-p-2}$)
4. Return $\widetilde{\mathbf{Z}}_i$.

Correctness of termination. Similar to (1.52) we can show that

$$\|\widetilde{\mathbf{Z}}_i - \mathbf{Z}^*\| \leq 2\beta(\mathcal{F}, \widetilde{\mathbf{Z}}_i).$$

But $\beta(\mathcal{F}, \widetilde{\mathbf{Z}}_i) \leq 2\|\widetilde{v}_i\|$, since the precision needed to compute $\widetilde{v}_i$ satisfies Lemma 1.21 with $\delta < \frac{1}{2}$ and hence from (1.48) we know that

$$\|\widetilde{v}_i - J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)^{-1}\mathcal{F}(\widetilde{\mathbf{Z}}_i)\| \leq \frac{1}{2}\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)^{-1}\mathcal{F}(\widetilde{\mathbf{Z}}_i)\|.$$

Thus

$$\|\widetilde{\mathbf{Z}}_i - \mathbf{Z}^*\| \leq 4\|\widetilde{v}_i\|. \tag{1.53}$$

So if $\|\widetilde{v}_i\| < 2^{-p-2}$ then $\|\widetilde{\mathbf{Z}}_i - \mathbf{Z}^*\| < 2^{-p}$.

We next bound the complexity of the algorithm. This bound will be expressed in terms of the condition number of the system of polynomials. Our next section gives the necessary details of this concept.

1.6 Uniform Complexity of Robust Newton

In this section, we bound the complexity of approximating, to a given absolute precision, a root of a zero-dimensional system of integer polynomials using Algorithm RN.

The algorithm takes as input the system of polynomials $\mathcal{F}$, precision $p \in \mathbb{N}_{\geq 0}$ and a $\mathbf{Z}_0 \in \mathbb{F}^n$ such that $\alpha(\mathcal{F}, \mathbf{Z}_0) < 0.02$, and produces an output $\langle \mathbf{Z}^* \rangle_p$, where $\mathbf{Z}^*$ is the associate zero of $\mathbf{Z}_0$. We further assume that $\mathbf{Z}_0$ and $\mathbf{Z}^*$ satisfy (1.44).

We now show that the robust Newton iterates $\|\widetilde{\mathbf{Z}}_i\|$ satisfy a bound similar to (1.44). Since $\mathbf{Z}_0$ is a robust approximate zero, from Theorem 1.7 we know that $\|\widetilde{\mathbf{Z}}_i - \mathbf{Z}_0\| \leq \frac{0.07}{\gamma(z_0)}$. From the second inequality in Lemma 1.7 we further get

$$\|\widetilde{\mathbf{Z}}_i - \mathbf{Z}_0\| \leq \frac{0.07}{\psi(u_0)(1-u_0)\gamma_*} \leq \gamma_*^{-1} \tag{1.54}$$

since $u_0 = 0.07$. We next derive a lower bound on γ_*. This will be done by generalizing a result of Kalantari [Kal05, Thm. 3.2] from analytic functions on the complex plane to our general setting of analytic functions on Banach spaces.

Lemma 1.23. *Let f be an analytic function between two Banach spaces E and F. The distance from any root $z^* \in E$ of f such that $Df(z^*)$ is non-singular to any other root of f is at least $\frac{1}{2\gamma(f,z^*)}$.*

Proof. Let τ be any other root of f distinct from z^*. The result follows easily from the claim that if $z \in E$ is such that $u := \|z - \tau\|\gamma(z) < 1$ then

$$\|N_f(z) - \tau\| \leq \frac{\|z-\tau\|u}{1-u}. \tag{1.55}$$

The reason is that if we choose $z := z^*$, and suppose $u := \|z^* - \tau\|\gamma(z^*) < 1$, then we get

$$\|N_f(z^*) - \tau\| \leq \frac{\|z^*-\tau\|u}{1-u}.$$

But both τ and z^* are fixed points of N_f, and hence we obtain

$$\frac{\|z^*-\tau\|\gamma(z^*)}{1-\|z^*-\tau\|\gamma(z^*)} \geq 1,$$

which implies that

$$\|z^* - \tau\| \geq \frac{1}{2\gamma(z^*)};$$

On the other hand if $\|z^* - \tau\|\gamma(z^*) \geq 1$ then the lower bound on $\|z^* - \tau\|$ trivially holds. We now prove (1.55). We have

$$\begin{aligned}
\|N_f(z) - \tau\| &= \|z - \tau - Df(z)^{-1}f(z)\| \\
&= \|z - \tau - Df(z)^{-1}f(z) + Df(z)^{-1}f(\tau)\| \\
&= \|z - \tau - Df(z)^{-1}f(z) + \sum_{k=0}^{\infty} \frac{Df(z)^{-1}D^k f(z)}{k!}(\tau - z)^k\| \\
&= \|\sum_{k=2}^{\infty} \frac{1}{k!} Df(z)^{-1}D^k f(z)(\tau - z)^k\| \\
&\leq \|z - \tau\| \sum_{k=2}^{\infty} (\gamma(z)\|z - \tau\|)^{k-1} \\
&\leq \frac{\|z - \tau\|u}{1 - u},
\end{aligned}$$

where the last step follows from the assumption that $u < 1$. □

Remark 1.3. *Following a line of argument as above, it seems possible to generalize the other results of Kalantari [Kal05] to our setting of analytic functions on Banach spaces.*

Applying the lemma above to (1.54) we get that $\|\widetilde{\mathbf{Z}}_i - \mathbf{Z}_0\|$ is smaller than twice the separation between the roots of $\mathcal{F}$, but from (1.44) we know that the maximum separation between any two roots of $\mathcal{F}$ is $2B(\mathcal{F})$, and hence we obtain $\|\widetilde{\mathbf{Z}}_i\| \leq \|\mathbf{Z}_0\| + 2B(\mathcal{F}) \leq 3B(\mathcal{F})$.

To bound the worst-case complexity of the algorithm we will first bound the number of iterative steps needed by the algorithm, then we will bound the worst-case precision required at each iteration; this latter bound will depend upon deriving a worst-case bound on C, as defined in Lemma 1.22, and $\kappa(J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i))$.

1.6.1 Bound on the number of iterative steps.

From (1.21) and (1.53) it is clear that the algorithm needs at most $2 + \log(p + 1 + \|\mathbf{Z}_0 - \mathbf{Z}^*\|)$ to compute $\langle \mathbf{Z}^* \rangle_p$. Moreover, since $\mathbf{Z}_0$ and $\mathbf{Z}^*$ satisfy (1.44) we know that $\|\mathbf{Z}_0 - \mathbf{Z}^*\| \leq 2B(\mathcal{F})$. Thus the number of iterative steps needed by the algorithm is

$$O(\log(p + 1 + B(\mathcal{F}))). \tag{1.56}$$

We next give an upper bound on the condition number $\kappa(J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i))$ by deriving upper bounds on $\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)^{-1}\|$ and $\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)\|$, starting with the former.

1.6.2 An upper bound on $\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)^{-1}\|$

Since $u := u(\widetilde{\mathbf{Z}}_i, \mathbf{Z}^*) \leq \frac{4-\sqrt{14}}{2}$, from (B.5) we know that

$$\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)^{-1}\| \leq 2(1 + \|\mathbf{Z}^*\|^2)\frac{\mu(\hat{\mathcal{F}}, (1, \widetilde{\mathbf{Z}}_i))}{\|\hat{\mathcal{F}}\|_k}.$$

From [Mal93, Lem. 31,p. 75] we further get that,

$$\mu(\hat{\mathcal{F}}, (1, \widetilde{\mathbf{Z}}_i)) \leq \frac{(1-u)^2}{\psi(u)}\mu(\hat{\mathcal{F}}, (1, \mathbf{Z}^*)) \leq 2\mu(\hat{\mathcal{F}}, (1, \mathbf{Z}^*)) \leq 2\mu(\hat{\mathcal{F}}).$$

Applying the bound (B.3) on $\mu(\hat{\mathcal{F}})$ we obtain

$$\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)^{-1}\| \leq \frac{4}{\|\hat{\mathcal{F}}\|_k}(1 + \|\mathbf{Z}^*\|^2)\mu(\Sigma)H(\hat{\mathcal{F}})^{d(\Sigma)}. \tag{1.57}$$

Thus we need a lower bound on $\|\hat{\mathcal{F}}\|_k$, which amounts to a lower bound on $\|\hat{F}_i\|_k$. But

$$\|\hat{F}_i\|_k \geq (\sum_{|J|=D_i} |\hat{F}_{iJ}|^2 (D_i!)^{-1})^{1/2} \geq (\sum_{|J|=D_i} D_i!^{-1})^{1/2}$$

since the coefficients of $\hat{F}_i$ are integers. Moreover, there are $S(\hat{F}_i) \geq 1$ terms in $\hat{F}_i$ thus

$$\|\hat{F}_i\|_k \geq \sqrt{S(\hat{F}_i)/D_i!} \geq (\mathcal{D}!)^{-1/2}.$$

Therefore

$$\|\hat{\mathcal{F}}\|_k \geq \sqrt{n}(\mathcal{D}!)^{-1/4}.$$

Applying this bound in (1.57), along with the upper bound on $\|\mathbf{Z}^*\|$ from (1.44), we obtain

$$\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)^{-1}\| \leq \frac{4}{\sqrt{n}}(\mathcal{D}!)^{1/4}B(\mathcal{F})^2\mu(\Sigma)H(\hat{\mathcal{F}})^{d(\Sigma)}. \tag{1.58}$$

1.6.3 An upper bound on $\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)\|$.

It is straightforward to show that for $1 \leq i, j \leq n$,

$$\left|\frac{\partial F_i}{\partial Z_j}(\widetilde{\mathbf{Z}}_i)\right| \leq D_i S(F_i)H(F_i)\|\widetilde{\mathbf{Z}}_i\|^{D_i}.$$

Thus

$$\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)\| \leq n\mathcal{D}S(\mathcal{F})H(\mathcal{F})\|\widetilde{\mathbf{Z}}_i\|^{\mathcal{D}}.$$

Since $\|\widetilde{\mathbf{Z}}_i\| = O(B(\mathcal{F}))$ we further get

$$\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)\| \leq n\mathcal{D}S(\mathcal{F})H(\mathcal{F})B(\mathcal{F})^{\mathcal{D}}. \tag{1.59}$$

Combining this with the bound on $\|J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)^{-1}\|$ in (1.58) gives us the bound

$$\kappa(J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)) \leq 4\sqrt{n}\mathcal{D}^{\mathcal{D}+1}\mu(\Sigma)H(\mathcal{F})^{1+d(\Sigma)}S(\mathcal{F})B(\mathcal{F})^{\mathcal{D}+2}. \tag{1.60}$$

1.6.4 Worst case lower bound on the distance to a zero

In §1.5.1 we had given a computational method to give a tight estimate on the distance $\|\mathbf{Z}_0 - \mathbf{Z}^*\|$ by computing a C as in Lemma 1.22. Here we derive a worst-case bound on C. From the lemma just mentioned we know that

$$2^{-C} = \Theta(\|J_{\mathcal{F}}(\mathbf{Z}_0)^{-1}\mathcal{F}(Z_0)\|),$$

thus it suffices to derive a lower bound on $\|J_{\mathcal{F}}(\mathbf{Z}_0)^{-1}\mathcal{F}(Z_0)\|$. Since the matrix $J_{\mathcal{F}}(\mathbf{Z}_0)$ is a non-singular square matrix we know that

$$\|\mathcal{F}(\mathbf{Z}_0)\| = \|J_{\mathcal{F}}(\mathbf{Z}_0)J_{\mathcal{F}}(\mathbf{Z}_0)^{-1}\mathcal{F}(\mathbf{Z}_0)\| \leq \|J_{\mathcal{F}}(\mathbf{Z}_0)\|\|J_{\mathcal{F}}(\mathbf{Z}_0)^{-1}\mathcal{F}(\mathbf{Z}_0)\|.$$

Thus to derive a lower bound on $\|J_{\mathcal{F}}(\mathbf{Z}_0)^{-1}\mathcal{F}(\mathbf{Z}_0)\|$ it suffices to derive a lower bound on $\|\mathcal{F}(\mathbf{Z}_0)\|$ and an upper bound on $\|J_{\mathcal{F}}(\mathbf{Z}_0)\|$; it can be shown that the latter bound is similar to the bound in (1.59), so we only focus on deriving the lower bound on $\|\mathcal{F}(\mathbf{Z}_0)\|$.

We know that $\mathbf{Z}_0$ is not a zero of the system, thus there must be some polynomial F_i in $\mathcal{F}$ such that $|F_i(\mathbf{Z}_0)| > 0$. Let L_0 be a bound on the bit-size of the coordinates (which are bigfloats) of $\mathbf{Z}_0$; this means that if we treat these coordinates as rational numbers then their denominator has at most bit-length L_0. Since the coefficients of F_i are integers it is not hard to see that $|F_i(\mathbf{Z}_0)|$ is a rational number whose denominator is at most $2^{D_i L_0} \leq 2^{\mathcal{D}L_0}$, and the numerator is at least one. Thus $|F_i(\mathbf{Z}_0)| \geq 2^{-\mathcal{D}L_0}$. This lower bound combined with the upper bound in (1.59) gives us

$$C = O(\mathcal{D}L_0 + \log(n\mathcal{D}H(\mathcal{F})S(\mathcal{F})B(\mathcal{F})^{\mathcal{D}})). \tag{1.61}$$

1.6.5 Worst-case complexity

The two most expensive steps in the loop of Algorithm RN are computing the Jacobian matrix $J_{\mathcal{F}}(\mathbf{Z})$ and solving the linear system of equations using Gaussian elimination. The precision used at the i-th iteration of this loop is bounded by

$$O(2^i + C + n + \log \kappa(J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i))).$$

Plugging in the bounds from (1.61) and (1.60) we know that this is bounded by

$$O(2^i + \mathcal{D}L_0 + \log T(\mathcal{F})),$$

where

$$T(\mathcal{F}) := O(\sqrt{n}\mathcal{D}^{\mathcal{D}+1}\mu(\Sigma)H(\hat{\mathcal{F}})^{d(\Sigma)}S(\mathcal{F})H(\mathcal{F})B(\mathcal{F})^{\mathcal{D}}). \tag{1.62}$$

The bound above is also a bound on the precision of the coordinates of $\widetilde{\mathbf{Z}}_i$. Thus from Appendix C, we know that the complexity of computing $J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)$ is

$$O(n^2\mathcal{D}^n M(2^i + \mathcal{D}L_0 \log T(\mathcal{F}))).$$

The cost of computing the solution to the system of linear equations $J_{\mathcal{F}}(\widetilde{\mathbf{Z}}_i)\mathbf{X} = \mathcal{F}(\widetilde{\mathbf{Z}}_i)$ is

$$O(n^3 M(2^i + \mathcal{D}L_0 + \log T(\mathcal{F}))),$$

since we require $O(n^3)$ operations each with the precision mentioned above. From these two bounds we conclude that the cost of the i-th iteration in Algorithm RN is bounded by

$$O(n^3\mathcal{D}^n M(2^i + \mathcal{D}L_0 + \log T(\mathcal{F}))),$$

and hence the total cost of Algorithm RN is

$$\begin{aligned} O(n^3\mathcal{D}^n &\sum_{i=0}^{\log(p+B(\mathcal{F}))} M(2^i)) \\ &+ O(n^3\mathcal{D}^{n+1}L_0 \log(p + B(\mathcal{F})) + n^3\mathcal{D}^n \log(p + B(\mathcal{F}))M(\log T(\mathcal{F})). \end{aligned}$$

Since $M(n)$ satisfies the weak regularity condition, i.e., $M(an) \leq bM(n)$ for $a, b \in (0, 1)$ and sufficiently large n, we know that $\sum_{i=0}^{k} M(2^i) = O(M(2^k))$ (see [Bre76a, Lem. 2.1]). Thus the cost of Algorithm RN is bounded by

$$O[n^3\mathcal{D}^n(M(p + B(\mathcal{F})) + \mathcal{D}L_0 \log(p + B(\mathcal{F})) + \log(p + B(\mathcal{F}))M(\log T(\mathcal{F})))].$$

It is not hard to see that the complexity of approximating a root in the weak model is

$$\begin{aligned} O[n^3\mathcal{D}^n &\log(p + B(\mathcal{F}))M(p + B(\mathcal{F})) \\ &+ n^3\mathcal{D}^{n+1}L_0 \log(p + B(\mathcal{F})) \\ &+ n^3\mathcal{D}^n \log(p + B(\mathcal{F}))M(\log T(\mathcal{F}))]. \end{aligned}$$

Assuming the system of polynomials $\mathcal{F}$ is fixed, this complexity exceeds the complexity in the strong model by a factor of $\log p$.

Theorem 1.16. *Let $\mathcal{F}$ be a zero-dimensional system of n integer polynomials in n variables. Given a robust approximate zero $\mathbf{Z}_0 \in \mathbb{R}^n$, such that $\alpha(\mathcal{F}, \mathbf{Z}_0) < 0.02$, we can compute $\langle \mathbf{Z}^* \rangle_n \in \mathbb{R}^n$, where $\mathbf{Z}^*$ is the associated root of $\mathbf{Z}_0$, in time*

$$O[n^3\mathcal{D}^n(M(p + B(\mathcal{F})) + \mathcal{D}L_0 \log(p + B(\mathcal{F})) + \log(p + B(\mathcal{F}))M(\log T(\mathcal{F})))],$$

where L_0 is an upper bound on the bit-size of the coordinate of $\mathbf{Z}_0$, $\mathcal{D}$ is the maximum amongst the degrees of the polynomials in $\mathcal{F}$, $B(\mathcal{F})$ is defined as in (1.45), and $T(\mathcal{F})$ is defined as in (1.62).

1.7 Experiments

In this section we provide experimental results on the running time of Algorithm RN for the special case when $\mathcal{F}$ is a single univariate integer polynomial $f(z)$ not identically zero. Such an algorithm has been provided by [SDY05], but we further simplify their algorithm by making use of the fact that the polynomials have integer coefficients.

The Jacobian $J_{\mathcal{F}}$ in this setting is just the derivative f', thus from Lemma 1.22 we deduce that if $C := 3 - \lfloor \log |\widetilde{v}_0| \rfloor$, where $\widetilde{v}_0 := \left\lfloor \left| \frac{f(z_0)}{f'(z_0)} \right| \right\rfloor_2$, then

$$\frac{|z_0 - z^*|}{8} < 2^{-C} < |z_0 - z^*|. \tag{1.63}$$

Similarly if we define $\widetilde{v}_i := \left\lfloor \frac{f(\widetilde{z}_i)}{f'(\widetilde{z}_i)} \right\rfloor_{2^i+5} \in \mathbb{F}$ we have the following:

Lemma 1.24. *Let $f(z) \in \mathbb{Z}[z]$ and z_0 be any bigfloat such that $\alpha(f, z_0) < 0.02$. Recursively define $\widetilde{z}_{i+1} := \widetilde{z}_i - \widetilde{v}_i$. Then $\widetilde{z}_{i+1} = \langle N_f(\widetilde{z}_i) \rangle_{2^{i+1}+C}$, for $i \geq 0$.*

Proof. Our definition of $\widetilde{z}_{i+1}$ implies

$$|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)| \leq \left| \frac{f(\widetilde{z}_i)}{f'(\widetilde{z}_i)} \right| 2^{-2^i-5}.$$

Applying the upper bound from Lemma 1.12 we obtain

$$|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)| < \frac{|\widetilde{z}_i - z^*|}{\psi(u_{\widetilde{z}_i})} 2^{-2^i-5}.$$

Since $\alpha(f, z_0) < 0.02$, we know from Theorem 1.7 that $u_{\widetilde{z}_i} \leq \frac{4-\sqrt{14}}{2}$ and hence $\psi(u_{\widetilde{z}_i}) \geq \frac{1}{2}$. Thus we get

$$|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)| < |\widetilde{z}_i - z^*| 2^{-2^i-4}.$$

Since $\widetilde{z}_i$ satisfies (1.21) we further obtain

$$|\widetilde{z}_{i+1} - N_f(\widetilde{z}_i)| < 2^{-2^{i+1}} |z_0 - z^*| 2^{-3} < 2^{-2^{i+1}-C},$$

where the last step follows from the lower bound in (1.63). □

It can be verified that our termination criterion for Algorithm RN still holds. Thus we have the following simplification of Algorithm RN:

Algorithm A

Input: $f(z) \in \mathbb{Z}[z]$, $p \geq 0$, and $z_0 \in \mathbb{F}$ where $\alpha(f, z_0) < 0.02$

Output: $\langle z^* \rangle_p$, z^* is the associated root of z_0

1 Compute $C := 3 - \left\lfloor \log \left[\left| \frac{f(z_0)}{f'(z_0)} \right| \right]_2 \right\rfloor$.

Let $\widetilde{z}_0 := \langle z_0 \rangle_{C+3}$, $i = 0$.

2 do

$\widetilde{v}_i := \left[\frac{f(\widetilde{z}_i)}{f'(\widetilde{z}_i)} \right]_{2^i+5}$.

$\widetilde{z}_{i+1} := \widetilde{z}_i - \widetilde{v}_i$.

$i := i + 1$.

while($\widetilde{v}_i \neq 0$ and $|\widetilde{v}_i| \geq 2^{-n-2}$).

3 Return $\widetilde{z}_i$.

Algorithm A assumes the strong model. The corresponding algorithm in the weak model is obtained by computing $\widetilde{v}_i := \left[\frac{f(\widetilde{z}_i)}{f'(\widetilde{z}_i)} \right]_p$. From now on, we call this modification the **full version**, and call Algorithm A the **robust version**. The results below compare the running times of these two versions.

For each of the polynomials in the tables below we will approximate a fixed root of that polynomial to precision $n = 1000, 5000, 10000, 20000, 40000$ using the two versions above. The starting point is chosen such that empirically it is both a robust approximate zero and guarantees the quadratic convergence of the full version. We did not apply the point estimate mentioned in Theorem 1.7, because these polynomials have very large $\gamma(f, z)$, which forces a very high accuracy for the starting point. This shows that there is still a gap between the theory of point estimates and their use in practice.

The initial approximation, around 20 digits of accuracy, for each of the roots was obtained using the MPSolve package of Bini and Fiorentino [BF00]; the polynomials are also taken from the same resource. We briefly describe the polynomials that we used for our test.

The polynomials chebyshev40 and chebyshev80 are the Chebyshev polynomials of the first kind of degree 40 and 80 respectively; the degree $n+1$ Chebyshev polynomial of first kind $T_{n+1}(X)$ satisfies the recurrence $T_{n+1}(X) = 2xT_n(X) - T_{n-1}(X)$, where $T_0(X) = 1$ and $T_1(X) = X$. The polynomials hermite40 and hermite80 are the physicists Hermite polynomials of degree 40 and degree 80 ; the degree $n + 1$ physicist Hermite polynomial $H_{n+1}(X)$ satisfies the recurrence $H_{n+1}(X) = 2XH_n(X) - 2nH_{n-1}(X)$, where $H_0(X) = 1$ and $H_1(X) = 2X$. Similarly

laguerre40 and laguerre80 represent the Laguerre polynomials; the degree $n+1$ Laguerre polynomial $L_{n+1}(X)$ satisfies the recurrence $(n+1)L_{n+1}(X) = (2n+1-X)L_n(X) - nL_{n-1}(X)$, where $L_0(X) = 1$ and $L_1(X) - 1 - X$. The polynomials mand31 and mand63 are the Mandelbrot polynomials of degree 31 and 63; the degree $n+1$ Mandelbrot polynomial $M_{n+1}(X)$ is defined as $M_{n+1}(X) = XM_n(X)^2 + 1$, where $M_0(X) = 1$; the roots of these polynomials lie on a fractal. The last polynomial wilk40 is the Wilkinson polynomial of degree 40; the degree n Wilkinson polynomial is defined as $\prod_{k=1}^{n}(X-k)$.

Tables 1.1 and 1.2 show the time in seconds taken by the two versions. Note that for wilk40 the robust version always takes the same time, because rounding produces the exact root after a fixed number of steps. The last column shows the relative running times of the two algorithms: theoretically, this should grow as $\log(p)$; although this ratio is increasing with p, it seems to be smaller than expected.

The implementations were done using the bigfloat package of Core Library [KLPY99]. The code and the sequence of tests are available under `progs/newton` in the files `newton.h` and `test.h`, respectively. Our implementation exploits a particular property of the bigfloat package in Core Library, viz., the ring operations $(+, -, \times)$ are error-free. This is in contrast to certain bigfloat packages, such as `gmp`'s `mpfr`, where each operation is guaranteed up to some arbitrarily specified precision. The workstation is Sun Blade 1000, 2x750 MHz UltraSPARC III CPU, 8 MB Cache each, with 2 GB of RAM.

1.8 Future Work

The difficulty with the point estimates presented above is that they are expensive to compute. For instance, a naive computation of $\alpha(f, z)$, for a univariate polynomial $f(z)$ of degree n, requires evaluating all the derivatives of $f(z)$, which takes $O(n^2)$ algebraic operations; this can be improved to $O(n)$, see [Sma86, Thm. B]. This prohibitive cost makes them ineffective in the context of our original motivation, which was to prevent the unnecessary checks performed by bracketing methods even when we have entered a region of super-linear convergence; the effectiveness will only become evident if we want to approximate the root to a very high precision. Thus the desirable aim of making point estimates practically useful is far from achieved.

Another direction to pursue is to develop robust point estimates for other iterative methods, such as the secant method; this method is of special interest, because it does not depend upon the derivative of the function. One can also extend the point estimate by Curry [Cur87] in the

exact model for the k-th Euler incremental algorithm to the weak and the strong model.

As was done in [DF95], one may possibly improve the constants involved in the point estimates above by developing the results using the majorant sequence approach.

Polynomial	Initial Approximation	n	Time by Robust (t)	Time by Full (T)	T/t
chebyshev40	-0.99922903624072293	1000	0.09	0.26	2.89
		5000	1.27	3.00	2.76
		10000	3.64	11.39	3.12
		20000	9.59	34.00	3.56
		40000	27.39	107.00	3.92
chebyshev80	-0.862734385977791819	1000	0.33	0.75	2.27
		5000	5.14	15.17	2.95
		10000	14.64	46.00	3.18
		20000	38.49	151.00	3.93
		40000	112.22	444.00	3.96
hermite40	-8.098761139250850052	1000	0.1	0.18	1.8
		5000	1.32	3.00	2.68
		10000	3.64	11.40	3.13
		20000	9.56	35.00	3.68
		40000	27.31	107.00	3.94
hermite80	-1.364377457054006838	1000	0.32	0.70	2.18
		5000	5.11	14.68	2.87
		10000	14.75	46.00	3.16
		20000	39.37	148.00	3.76
		40000	110.68	447.03	4.04

Table 1.1: A comparison of weak and robust Newton iteration I

Polynomial	Initial Approximation	n	Time by Robust (t)	Time by Full (T)	T/t
laguerre40	0.0357003943088883851	1000	0.09	0.18	2
		5000	1.38	3.00	2.56
		10000	3.61	11.35	3.14
		20000	9.87	35.00	3.64
		40000	27.47	109.00	3.98
laguerre80	0.0179604233006983654	1000	0.34	0.70	2.06
		5000	5.32	14.75	2.77
		10000	14.68	46.00	3.17
		20000	38.68	143.00	3.72
		40000	112.56	445.00	3.96
mand31	-1.996376137711193750	1000	0.06	0.11	1.83
		5000	0.80	2.05	2.56
		10000	2.15	6.73	3.13
		20000	5.57	21.00	3.78
		40000	16.28	64.00	3.99
mand63	-1.999095682327018473	1000	0.20	0.43	2.15
		5000	3.19	9.25	2.90
		10000	8.86	29.00	3.30
		20000	23.99	88.00	3.68
wlik40	11.232223434543512321	1000	0.03	0.40	13.67
		5000	0.03	5.00	166.67

Table 1.2: A comparison of weak and robust Newton iteration II

2 Real Root Isolation: The Descartes Method

Let $A(X)$ be a polynomial of degree $n > 1$ with real coefficients. A classic approach to real root isolation starts from an open interval I_0 containing all real roots of $A(X)$ and bisects it recursively as follows: Given an interval J, test for the number $\#(J)$ of real roots in it. If $\#(J) = 0$ is known, stop. If $\#(J) = 1$ is known, report J as an isolating interval and stop. Otherwise, subdivide $J = (c, d)$ at its midpoint $m = (c + d)/2$; report $[m, m]$ if $f(m) = 0$; recur on (c, m) and (m, d).

To carry out this approach, we need a method for estimating the number of roots in an interval. Two possible choices are **Sturm sequences** (e.g., [Yap00, chap. 7]), which give an exact count of distinct real roots in an interval, and

Descartes' rule of signs (e.g., Proposition 2.1 below), which counts real roots with multiplicity and may overestimate this number by an even positive integer. Despite the apparent inferiority of Descartes' rule as compared to Sturm sequences, there is considerable recent interest in the Descartes approach because of its excellent performance in practice [Joh98, RZ01, MRR05, RZ04].

This chapter[1] shows that the asymptotic worst case bound on recursion tree size for the Descartes method (Theorem 2.2) is no worse than the best known bound for Sturm's method (Theorem 6 of [DSY05]). For the particular case of polynomials with L-bit integer coefficients, the recursion tree is $O(n(L+\log n))$ both for Sturm's method [Dav85, DSY05] and the Descartes method (Corollary 2.1); and the work at each node of this tree can be done with $\widetilde{O}(n^3L)$ bit operations (using asymptotically fast basic operations), where $\widetilde{O}$ indicates that we are omitting logarithmic factors (see [Rei97, LR01, DSY05] or Theorem 2.5, respectively).

The connection between root isolation in the power basis using the Descartes method, and root isolation in the Bernstein basis using de Casteljau's algorithm and the variation-diminishing property of Bézier curves was already pointed out by Lane and Riesenfeld [LR81], but this connection is often unclear in the literature. In Section 2.1, we provide a general framework for viewing both as a form of the Descartes method. In Section 2.2, we present the main result, which is a new upper bound on the size of the recursion tree in the Descartes method. Up to that point, our analysis holds for all square-free polynomials with real coefficients. We then restrict to the case of integer polynomials with coefficients of bit-length L to show that this new

[1] The results in this chapter are from joint work with Arno Eigenwillig, and Chee Yap [ESY06].

bound on tree size is optimal under the assumption $L = \Omega(\log n)$ (Section 2.2.3) and allows a straightforward derivation of the best known bit complexity bound (Section 2.3).

2.0.1 Previous work

Root isolation using Descartes' rule of signs was cast into its modern form by Collins and Akritas [CA76], using a representation of polynomials in the usual power basis. Rouillier and Zimmermann [RZ04] summarize various improvements of this method until 2004.

The algorithm's equivalent formulation using the Bernstein basis was first described by Lane and Riesenfeld [LR81] and more recently by Mourrain, Rouillier and Roy [MRR05] and Mourrain, Vrahatis and Yakoubsohn [MVY02]; see also [BPR03, §10.2]. Johnson et al. [JKL$^+$06] have developed an architecture aware implementation of the Bernstein basis variant of the Descartes method that automatically generates architecture-aware high-level code and leaves further optimizations to the compiler.

The crucial tool for our bound on the size of the recursion tree is Davenport's generalization [Dav85] of Mahler's bound [Mah64] on root separation. Davenport used his bound for an analysis of Sturm's method (see [DSY05]). He mentioned a relation to the Descartes method but did not work it out. This was done later by Johnson [Joh98] and, filling a gap in Johnson's argument, by Krandick [Kra95]. However, they bound the number of internal nodes at each level of the recursion tree separately. This leads to bounds that imply[2] a tree size of $O(n \log n\,(\log n + L))$ and a bit complexity of $O(n^5(\log n + L)^2)$ for a polynomial of degree n with L-bit integer coefficients. Their arguments use a termination criterion for the Descartes method due to Collins and Johnson [CJ89].

Krandick and Mehlhorn [KM06] employ a theorem by Ostrowski [Ost50] that yields a sharper termination criterion. However, they just use it to improve on the constants of the bounds in [Kra95][3]. We will show that Ostrowski's result allows an immediate bound on the number of *all* internal nodes of the recursion tree. This bound is better by a factor of $\log n$ and leads to the same bit complexity bound in a simpler fashion.

[2] Personal communication, Krandick and Mehlhorn.

[3] This potential use of Ostrowski's result is mentioned but not carried out in the Ph.D. thesis of P. Batra [Bat99].

2.1 The Descartes Method

2.1.1 A Basis-free Framework

The Descartes method is based on the following theorem about sign variations. A **sign variation** in a sequence $(a_0, \ldots, a_n)$ of real numbers is a pair $i < j$ of indices such that $a_i a_j < 0$ and $a_{i+1} = \cdots = a_{j-1} = 0$. The number of sign variations in a sequence $(a_0, \ldots, a_n)$ is denoted $\mathrm{Var}(a_0, \ldots, a_n)$.

Proposition 2.1. [Descartes' rule of signs]
Let $A(X) = \sum_{i=0}^{n} a_i X^i$ be a polynomial with real coefficients that has exactly p positive real roots, counted with multiplicities. Let $v = \mathrm{Var}(a_0, \ldots, a_n)$ be the number of sign variations in its coefficient sequence. Then $v \geq p$, and $v - p$ is even.

See [KM06] for a proof with careful historic references. Already Jacobi [Jac35, IV] made the observation that this extends to estimating the number of real roots of a real polynomial $A(X)$ of degree n over an arbitrary open interval (c, d) by applying Descartes' rule to $(X+1)^n A((cX + d)/(X + 1)) = \sum_{i=0}^{n} a_i^* X^i$, because the Möbius transformation $X \mapsto (cX + d)/(X + 1)$ maps $(0, \infty)$ in one-to-one and onto correspondence with (c, d). So we define $\mathit{DescartesTest}(A, (c, d)) := \mathrm{Var}(a_0^*, \ldots, a_n^*)$. Since $v - p$ is non-negative and even, the Descartes test yields the exact number of roots whenever its result is 0 or 1.

The Descartes method for isolating the real roots of an input polynomial $A_{\mathrm{in}}(X)$ in an open interval J consists of a recursive procedure $\mathit{Descartes}(A, J)$ operating on a polynomial $A(X)$ and an interval J where the roots of $A(X)$ in $(0, 1)$ correspond to the roots of $A_{\mathrm{in}}(X)$ in J as follows:

(*) There is a constant $\lambda \neq 0$ and an affine transformation $\phi\colon \mathbb{R} \to \mathbb{R}$ such that $J = \phi((0, 1))$ and $\lambda A = A_{\mathrm{in}} \circ \phi$.

To isolate all the roots of $A_{\mathrm{in}}(X)$, we choose an interval $I_0 = (-B_1, +B_2)$ enclosing all real roots of A_{in} (see, e.g., [Yap00, §6.2]). The recursion begins with $\mathit{Descartes}(A, I_0)$, where $A(X) := A_{\mathrm{in}}((B_1 + B_2)X - B_1)$; thus initially the roots of $A(X)$ in $(0, 1)$ correspond to the real roots of $A_{\mathrm{in}}(X)$ in I_0 via the affine transformation $\phi(X) = (B_1 + B_2)X - B_1$. The procedure goes as follows:

procedure *Descartes*$(A,(c,d))$

{Assert: Invariant (*) holds with $J=(c,d)$.}

1. $v := DescartesTest(A,(0,1))$.
2. **if** $v=0$ **then return**.
3. **if** $v=1$ **then** report (c,d) and **return**.
4. $m := (c+d)/2$.
5. $(A_L, A_R) := (H(A), TH(A))$.
6. **if** $A_R(0)=0$ **then** report $[m,m]$.
7. Call *Descartes*$(A_L,(c,m))$, *Descartes*$(A_R,(m,d))$ and **return**.

The polynomials $A_L(X)$ and $A_R(X)$ are constructed using the homothetic transformation $H(A)(X) := 2^n A(X/2)$ and the translation transformation $T(A)(X) := A(X+1)$. For later use, we also introduce the reversal transformation $R(A)(X) := X^n A(1/X)$. Given two such transformations Φ, Ψ and a polynomial $A(X)$, let $\Phi\Psi(A) := \Phi(\Psi(A))$.

Note that in the initial invocation of *Descartes*$(A,(c,d))$, one has

$$DescartesTest(A,(0,1)) = DescartesTest(A_{\text{in}},(c,d)).$$

In its recursive calls, one has

$$DescartesTest(A_L,(0,1)) = DescartesTest(A_{\text{in}},(c,m))$$

and

$$DescartesTest(A_R,(0,1)) = DescartesTest(A_{\text{in}},(m,d)),$$

and so on.

The above description of *Descartes*() does not refer to any basis in the vector space of polynomials of degree at most n. However, an implementation needs to represent polynomials by coefficients with respect to some specific basis.

The classical choice of basis for *Descartes*() is the usual power basis

$$(1, X, X^2, \ldots, X^n).$$

The transformations H, T and R are carried out literally. $DescartesTest(A,(0,1))$ consists in counting the number of sign changes in the coefficient sequence of $TR(A)$. The test whether

$A_R(0) = 0$ amounts to inspection of the constant term. We call the resulting algorithm the *power basis variant* of the Descartes method.

An alternative choice of basis is the $[0,1]$-Bernstein basis

$$(B_0^n(X), B_1^n(X), \ldots, B_n^n(X)),$$

with $B_i^n(X) := B_i^n[0,1](X)$ where

$$B_i^n[c,d](X) := \binom{n}{i} \frac{(X-c)^i (d-X)^{n-i}}{(d-c)^n}, \quad 0 \le i \le n.$$

Its usefulness for the Descartes method lies in the following: Since

$$TR(B_i^n)(X) = \binom{n}{i} X^{n-i}, \tag{2.1}$$

for $A(X) = \sum_{i=0}^n b_i B_i^n(X)$ one has that

$$\mathit{DescartesTest}(A,(0,1)) = \mathrm{Var}(b_0, \ldots, b_n),$$

without any additional transformation.

To obtain A_L and A_R from $A(X) = \sum_{i=0}^n b_i B_i^n(X)$, we use a fraction-free variant of de Casteljau's algorithm [PBP02]: For $0 \le i \le n$ set $b_{0,i} := b_i$. For $1 \le j \le n$ and $0 \le i \le n-j$ set $b_{j,i} := b_{j-1,i} + b_{j-1,i+1}$. From this, one obtains coefficients of $2^n A(X) = \sum_{i=0}^n b'_i B_i^n[0,\frac{1}{2}](X) = \sum_{i=0}^n b''_i B_i^n[\frac{1}{2},1](X)$ by setting $b'_i := 2^{n-i} b_{i,0}$ and $b''_i := 2^i b_{n-i,i}$. Since

$$\begin{aligned} H(2^{-n} B_i^n[0,\tfrac{1}{2}])(X) &= B_i^n[0,1] \\ TH(2^{-n} B_i^n[\tfrac{1}{2},1])(X) &= B_i^n[0,1], \end{aligned}$$

one has

$$A_L(X) = H(A)(X) = \sum_{i=0}^n b'_i B_i^n(X)$$

and

$$A_R(X) = TH(A)(X) = \sum_{i=0}^n b''_i B_i^n(X).$$

Finally, the test whether $A_R(0) = 0$ amounts to inspection of b''_0, since $B_i^n(0) = 0$ for $i > 0$. We call the resulting algorithm the *Bernstein basis variant* of the Descartes method.

For consistency with the power basis variant, we have described the Bernstein basis variant as passing transformed polynomials A_L and A_R expressed in a globally fixed basis $(B_n^i[0,1])_i$ in recursive calls. Equivalently, one can think of it as passing (a constant multiple of) the

same polynomial all the time, but converting it to the Bernstein basis w.r.t. the interval under consideration.

Both variants of the Descartes method as presented above work for polynomials with arbitrary real coefficients. However, if the initial coefficients are integers, then integrality is preserved. If this is not needed, one can leave out the factor 2^n in the definition of $H(A)$ and, for the Bernstein basis variant, apply the ordinary instead of the fraction-free de Casteljau algorithm.

2.1.2 Termination

Since the Descartes test only gives an upper bound on the number of real roots in an interval, an extra argument is needed that each path in the recursion tree of the Descartes method eventually reaches an interval for which it counts 0 or 1 and thus terminates. We use a result from Krandick and Mehlhorn [KM06] based on a theorem by Ostrowski [Ost50]. To describe this result, following [KM06], we associate three open discs in the complex plane with an open interval $J = (c, d)$. The disc C_J is bounded by the circle that has centre $(c + d)/2$, and radius $(d - c)/2$; the disc $\overline{C}_J$ is bounded by the circle that has centre $(c+d)/2+i(\sqrt{3}/6)(d-c)/2$, and passes through the end-points of J; and the disc $\underline{C}_J$ is bounded by the circle that has centre $(c+d)/2-i(\sqrt{3}/6)(d-c)/2$, and passes through the end-points of J. The three discs are illustrated in Figure 2.1.

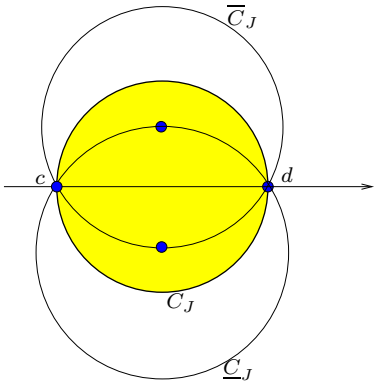

Figure 2.1: Three discs associated with the interval $J = (c, d)$.

Consider a real polynomial $A(X)$ and its roots in the complex plane. Let $J = (c, d)$ be an open interval, and let $v = \mathit{DescartesTest}(A, J)$.

Proposition 2.2. [One-Circle Theorem] *If the open disc C_J does not contain any root of $A(X)$, then $v = 0$.*

Proposition 2.3. [Two-Circle Theorem] *If the union of the open discs $\underline{C}_J$ and $\overline{C}_J$ contains precisely one simple root of $A(X)$ (which is then necessarily a real root), then $v = 1$.*

See [KM06] for proofs. In the sequel, we call the union of discs $\underline{C}_J$ and $\overline{C}_J$ the **two-circles figure** around interval J. Notice that the two-circles figure contains the disc C_J.

2.2 The Size of the Recursion Tree

2.2.1 The Davenport-Mahler Bound

The Davenport-Mahler theorem gives a lower bound on the product of differences of certain pairs of roots of a polynomial $A(X) = a_n \prod_{i=1}^{n}(X - \alpha_i)$ in terms of its **discriminant** $\mathrm{discr}(A) = a_n^{2n-2} \prod_{1 \le i < j \le n}(\alpha_i - \alpha_j)^2$ and **Mahler measure**

$$\mathrm{M}(A) = |a_n| \prod_{i=1}^{n} \max\{1, |\alpha_i|\}; \tag{2.2}$$

see [Yap00, §6.6, §4.5] [Mc99, §1.5, §2.1]. This theorem appears in the literature in several variants that all use the same proof but formulate different conditions on how roots may be paired so that the proof works. We give the most general condition supported by the proof. It is equivalent to Johnson's formulation [Joh98] and generalizes Davenport's original formulation [Dav85, Prop. I.5.8].

Theorem 2.1. *Let $A(X) = a_n \prod_{i=1}^{n}(X - \alpha_i)$ be a square-free complex polynomial of degree n. Let $G = (V, E)$ be a directed graph whose nodes $\{v_1, \ldots, v_k\}$ are a subset of the roots of $A(X)$ such that*

1. *If $(v_i, v_j) \in E$ then $|v_i| \le |v_j|$.*
2. *G is acyclic.*
3. *The in-degree of any node is at most 1.*

If exactly m of the nodes have in-degree 1, then

$$\prod_{(v_i, v_j) \in E} |v_i - v_j| \ge \sqrt{|\mathrm{discr}(A)|} \cdot \mathrm{M}(A)^{-(n-1)} \cdot (n/\sqrt{3})^{-m} \cdot n^{-n/2}.$$

Proof. This proof is not self-contained, but refers to the standard argument from Davenport [Dav85, Yap00]. Let $(v_1, \ldots, v_k)$ be the topologically sorted list of the vertices of G, where

$(v_i, v_j) \in E$ implies $j < i$. Given such an ordering we modify the $n \times n$ Vandermonde matrix $W_A = (\alpha_i^{j-1})_{j,i}$ as follows: For $j = 1$ to k in turn, we process v_j. If there exists an $i > j$ such that $(v_i, v_j) \in E$ then in W_A we subtract the column of v_i from the column of v_j; if no such i exists then the column of v_j remains unchanged. This finally yields a transformed matrix M such that $\det W_A = \det M$. Note that exactly m columns of M are modified from W_A. Moreover, $\det M = \prod_{(v_i,v_j)\in E}(v_j - v_i) \cdot \det M'$, where M' is a matrix similar to the one in [Yap00, Theorem 6.28, Eqn. (19)]. As in the proof in [Yap00], we conclude:

$$|\det(W_A)| \leq \left(\prod_{(v_i,v_j)\in E} |v_i - v_j| \right) \cdot \mathrm{M}(A)^{(n-1)} \left(\frac{n}{\sqrt{3}} \right)^m n^{n/2}.$$

But $\sqrt{|\mathrm{discr}(A)|} = |\det W_A|$, thus giving us the desired result. □

Remark 2.1. *The bound in Theorem 2.1 is invariant under replacing $A(X)$ by a non-zero scalar multiple $\lambda A(X)$.*

Remark 2.2. *A bound similar to Theorem 2.1 appears in [Mig95]. Instead of $\mathrm{M}(A)^{n-1}$, it uses a product of root magnitudes with varying exponents of $n-1$ or less.*

Remark 2.3. *Let $\mathrm{sep}(A)$ be the minimum distance between two distinct roots of $A(X)$. Then we have*

$$\mathrm{sep}(A) \geq \sqrt{\mathrm{discr}(A)}\mathrm{M}(A)^{-(n-1)} \cdot (n/\sqrt{3}) \cdot n^{-n/2}.$$

2.2.2 The Recursion Tree

Our application of the Davenport-Mahler theorem rests on the following lemma. It reflects an important structural advantage of Proposition 2.3 over the weaker two-circle theorem by Collins and Johnson [CJ89]: An intersection of the two-circles figures of two non-overlapping intervals can only occur if the intervals are adjacent, even if they reside on very different levels of the recursion tree.

Lemma 2.1. *Let J_0 and J_1 be any two open intervals appearing in the recursive subdivision of some initial interval I_0. If the two-circles figures of Proposition 2.3 around J_0 and J_1 intersect, then J_0 and J_1 overlap or have a common endpoint.*

Proof. We show that non-overlapping intervals with intersecting two-circles figures have a common endpoint. Let us choose indices such that $w(J_0) \geq w(J_1)$. Assume J_0 lies to the left of

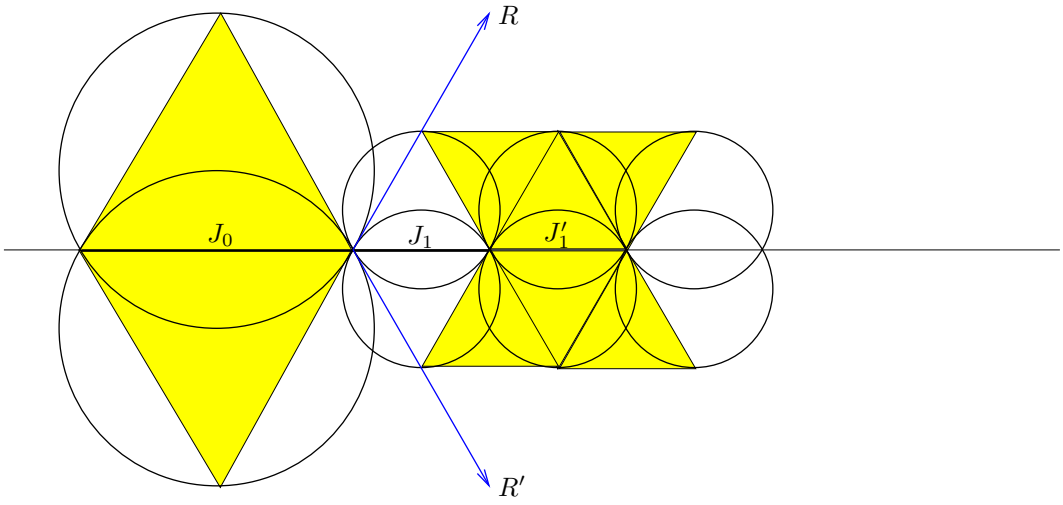

Figure 2.2: The two-circles figure around J_0 can overlap with that of J_1 but not with any two-circles figure further right.

J_1 (the opposite case is symmetric). All intervals right of J_0 that have width $w(J_1)$ and appear in the recursive subdivision of I_0 have distance $k \cdot w(J_1)$ from J_0 for a non-negative integer k. They are depicted in Figure 2.2. The interval with $k = 0$ has a two-circles figure intersecting the two-circles figure of J_0. For $k > 0$, we claim that the two-circles figure of J_0 is disjoint from the two-circles figure of J_1. To see this, consider the convex cone delimited by the two tangent rays (R, R') of the two-circles figure of J_0 at its right endpoint. The two-circles figure of J_0 lies outside that cone, but if $k > 0$, then the two-circles figure of J_1 lies inside the cone. Figure 2.2 illustrates this for the case $k = 1$: the corresponding interval is J_1', and the two-circles figure of J_1' is covered by six equilateral triangles. Since the rays R, R' meet the x-axis at $60°$, this shows that the six equilateral triangles lie within the cone. Hence there is no intersection. □

The recursion tree T of the Descartes method in Section 2.1 is a binary tree. With each node $u \in T$ we can associate an interval I_u; the root is associated with I_0. A leaf u of T is said to be of **type-i** if the open interval I_u contains exactly i real roots; the termination condition of the algorithm implies i is either 0 or 1.

Our aim is to bound the number of nodes in T, denoted by $\#(T)$. We next introduce a subtree T' of T by pruning certain leaves from T:

- If a leaf u has a sibling that is a non-leaf, we prune u.

- If u, v are both leaves and siblings of each other, then we prune exactly one of them; the choice to prune can be arbitrary except that we prefer to prune a type-0 leaf over a type-1.

Clearly, $\#(T) < 2\#(T')$; hence it is enough to bound $\#(T')$. Let U be the set of leaves in T'. Then the number of nodes along the path from any $u \in U$ to the root of T' is exactly $\log \frac{w(I_0)}{w(I_u)}$. Thus

$$\#(T') \leq \sum_{u \in U} \log \frac{w(I_0)}{w(I_u)}. \tag{2.3}$$

Our next goal is to reduce this bound to the Davenport-Mahler type bound in Theorem 2.1.

Let u be a leaf of T', and v be its parent. We will define two roots α_u, β_u such that the number of nodes along the path from u to the root is

$$O\left(\log \frac{w(I_0)}{|\alpha_u - \beta_u|}\right).$$

Furthermore, we will show that if u, u' are two leaves of the same type (both type-0 or both type-1), then $\{\alpha_u, \beta_u\}$ and $\{\alpha_{u'}, \beta_{u'}\}$ are disjoint.

1. If u is type-1 then its interval I_u contains a real root α. Consider its parent v. By Proposition 2.3, $\overline{C}_{I_v} \cup \underline{C}_{I_v}$ must contain a root apart from α_u; let β_u be any root in this region. Then it follows that

$$|\alpha_u - \beta_u| < \frac{2}{\sqrt{3}} w(I_v) = \frac{4}{\sqrt{3}} w(I_u). \tag{2.4}$$

   Thus the number of nodes in the path from u to the root of T' is

$$\log \frac{w(I_0)}{w(I_u)} < \log \frac{4w(I_0)}{\sqrt{3}|\alpha_u - \beta_u|}. \tag{2.5}$$

   Let u' be another type-1 leaf different from u. Clearly, $\alpha_u \neq \alpha_{u'}$. We claim that β_u and $\beta_{u'}$ can be chosen such that $\beta_u \neq \beta_{u'}$. From Lemma 2.1 it is clear that we only need to consider the case when I_v and $I_{v'}$ are adjacent to each other. Moreover, assume β_u and $\overline{\beta_u}$ are the only non-real roots in $\overline{C}_{I_v} \cup \underline{C}_{I_v}$ and $\overline{C}_{I_{v'}} \cup \underline{C}_{I_{v'}}$. Then it must be that either $\beta_u \in \overline{C}_{I_v} \cap \overline{C}_{I_{v'}}$ or $\beta_u \in \underline{C}_{I_v} \cap \underline{C}_{I_{v'}}$. In either case we can choose $\beta_{u'} = \overline{\beta_u}$ distinct from β_u.

2. If u is type-0, it had a type-0 sibling that was pruned. Consider their parent node v and let I_v be the interval associated with it. There are two cases to consider:

- I_v does not contain a real root. Thus Proposition 2.2 implies that C_{I_v} must contain some non-real root α_u and its conjugate $\beta_u := \overline{\alpha_u}$. Moreover,

$$|\alpha_u - \beta_u| \leq w(I_v) = 2w(I_u). \tag{2.6}$$

- The midpoint of I_v is a real root, say α. Since $DescartesTest(A, I_v) \geq 2$, there is a pair of non-real roots $(\beta, \overline{\beta})$ in $\overline{C}_{I_v} \cup \underline{C}_{I_v}$. If $\beta \in C_{I_v}$ then let $\alpha_u := \beta$ and $\beta_u := \overline{\beta}$; otherwise, let $\alpha_u = \alpha$ and $\beta_u = \beta$. It can be verified that (2.6) still holds.

Hence the number of nodes on the path from u to root of T' is

$$\log \frac{w(I_0)}{w(I_u)} \leq \log \frac{2w(I_0)}{|\alpha_u - \beta_u|}. \tag{2.7}$$

Again, if u' is another type-0 leaf different from u, then $\alpha_u \neq \alpha_{u'}$, since $\alpha_u \in C_{I_u}$, $\alpha_{u'} \in C_{I'_u}$ and $C_{I_u} \cap C_{I_{u'}} = \varnothing$. Furthermore, we can choose β_u and $\beta_{u'}$ such that $\beta_u \neq \beta_{u'}$. This is clear if both α_u and $\alpha_{u'}$ are not real, since then $\beta_w = \overline{\alpha_w}$, $w = u, u'$; if both are real then β_u and $\beta_{u'}$ can be chosen as in the argument of type-1 leaves; otherwise, say α_u is real and $\alpha_{u'}$ is not, we can choose $\beta_u = \alpha_{u'}$ and $\beta_{u'} = \overline{\alpha_{u'}}$ without affecting (2.7).

Let $U_0 \subseteq U$ and $U_1 \subseteq U$ denote the set of type-0 and type-1 leaves respectively. Then substituting (2.5) and (2.7) in (2.3) we get

$$\#(T') \leq \sum_{u \in U_0} \log \frac{2w(I_0)}{|\alpha_u - \beta_u|} + \sum_{u \in U_1} \log \frac{4w(I_0)}{\sqrt{3}|\alpha_u - \beta_u|}. \tag{2.8}$$

We obtain a bound on the number of type-0 and type-1 leaves:

Lemma 2.2. *For U_0 and U_1 defined as above we have:*

1. *$|U_0|$ is at most the number of non-real roots of $A(X)$.*

2. *$|U_1|$ is at most the number of real roots of $A(X)$.*

Proof. As shown above, with each $u \in U_0$ we can associate a unique pair of roots (α_u, β_u), where at least one of them is complex and uniquely chosen, thus implying the upper bound on $|U_0|$.

Again by the arguments given earlier, for each $u \in U_1$ we can associate a unique real root α_u, and hence the upper bound on $|U_1|$. □

Now we can show our main result:

Theorem 2.2. *Let $A(X) \in \mathbb{R}[X]$ be a square-free polynomial of degree n. Let T be the recursion tree of the Descartes method run on (A, I_0). Then the number of nodes in T is*

$$O(\log(\frac{1}{|\mathrm{discr}(A)|}) + n(\log \mathrm{M}(A) + \log n + \log w(I_0))).$$

Proof. From (2.8), we know that the number of nodes in T' is bounded by

$$\#(T') \leq |U| \log 4w(I_0) - \sum_{u \in U} \log(|\alpha_u - \beta_u|). \tag{2.9}$$

Consider the graph G whose edge set is $E_1 \cup E_0$, where

$$E_0 := \{(\alpha_u, \beta_u) | u \in U_0\} \text{ and } E_1 := \{(\alpha_u, \beta_u) | u \in U_1\}.$$

We want to show that G satisfies the conditions of Theorem 2.1. First of all, for any $u \in U$ we can reorder the pair (α_u, β_u) to ensure that $|\alpha_u| \leq |\beta_u|$ without affecting (2.8).

Now we show that the in-degree of G may be assumed to be at most one. Clearly, the edge sets E_0 and E_1 have in-degree one. However, in $E_0 \cup E_1$ cases like that illustrated in Figure 2.3 may occur. But we can reduce the in-degree of β_u to one in both cases: in (a), we can always re-order the edge $(\alpha_{u'}, \beta_{u'})$ to $(\beta_{u'}, \alpha_{u'})$, since $\beta_{u'} = \overline{\alpha_{u'}}$; in (b), we can choose $\beta_{u'} = \overline{\beta_u}$.

Applying Theorem 2.1 to G we get:

$$\prod_{u \in U} |\alpha_u - \beta_u| \geq \sqrt{|\mathrm{discr}(A)|} \cdot \mathrm{M}(A)^{-(n-1)} \cdot \left(\frac{n}{\sqrt{3}}\right)^{-|U|} n^{-n/2}. \tag{2.10}$$

Taking logarithms on both sides yields:

$$\begin{aligned} \sum_{u \in U} \log |\alpha_u - \beta_u| \geq & \frac{1}{2} \log(|\mathrm{discr}(A)|) - (n-1) \log \mathrm{M}(A) \\ & - n \log \frac{n}{\sqrt{3}} - \frac{n}{2} \log n; \end{aligned} \tag{2.11}$$

since $|U| \leq n$ (by Lemma 2.2). Plugging this into (2.9) gives us:

$$\begin{aligned} \#(T') \quad \leq \quad & |U| \log w(I_0) + 2|U| + n \log \mathrm{M}(A) \\ & + \frac{1}{2} \log \frac{1}{|\mathrm{discr}(A)|} + 2n \log n. \end{aligned}$$

Using $|U| \leq n$ again, the claim follows. □

Remark 2.4. *(i) There exist an interval I_0 enclosing all real roots of $A(X)$ such that $w(I_0) \leq 2\mathrm{M}(A)/|a_n|$, because $\mathrm{M}(A)/|a_n|$ is an upper bound on the magnitude of all roots.*
(ii) Landau's inequality $\mathrm{M}(A) \leq \|A\|_2$ (e.g., [Yap00, Lem. 4.14(i)]) and the obvious estimate $\|A\|_2 \leq \sqrt{n+1}\|A\|_\infty$ immediately yield bounds on the number of nodes in T in terms of these norms of $A(X)$.

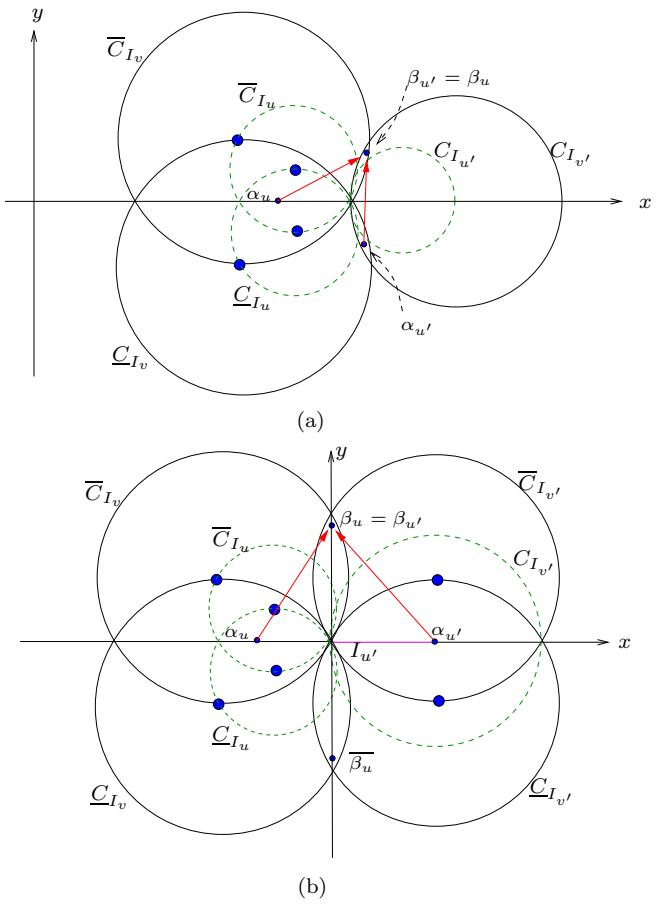

Figure 2.3: A type-0 and type-1 leaf sharing the same root.

Corollary 2.1. *Let $A(X)$ be a square-free polynomial of degree n with integer coefficients of magnitude less than 2^L. Let I_0 be an open interval enclosing all real roots of $A(X)$ such that $\log w(I_0) = O(L)$. Let T be the recursion tree of the Descartes method run on (A, I_0). Then the number of nodes in T is $O(n(L + \log n))$.*

Proof. Since $A(X)$ is a square-free integer polynomial, $|\mathrm{discr}(A)|$ is at least one. From the remark above, we have $\mathrm{M}(A) < 2^L\sqrt{n+1}$. Finally, $\log w(I_0) \leq L + 1$. □

The condition $\log w(I_0) = O(L)$ is no restriction, as 2^L is an upper bound on the absolute value of all roots of $A(X)$ (e.g., [Yap00, Cor. 6.8]).

2.2.3 Almost Tight Lower Bound

We show that for integer polynomials our tree size bound $O(n(L + \log n))$ is optimal under the assumption $L = \Omega(\log n)$. To do so, we construct a family of inputs of unbounded degree n and coefficient length L for which the height of the recursion tree is $\Omega(nL)$.

Mignotte [Mig81] gave a family of polynomials $P(X) = X^n - 2(aX - 1)^2$ parametrized by integers $n \geq 3$ and $a \geq 3$. By Eisenstein's criterion, $P(X)$ is irreducible (use the prime number 2). Let $h = a^{-n/2-1}$. Since $P(a^{-1}) > 0$ and $P(a^{-1} \pm h) = (a^{-1} \pm h)^n - 2a^{-n} < 0$, there exist two distinct roots α and β of $P(X)$ in $(a^{-1} - h,\, a^{-1} + h)$. Clearly, $|\alpha - \beta| < 2h$. In the sequel, we shall restrict to the case that the degree n is even. This allows us to conclude that any interval I_0 enclosing all roots of $P(X)$ is a superset of $(0, 1)$, because the sign of $P(X)$ is positive for $X \to \pm\infty$ but negative for $X = 0$ and $X = 1$.

If one is willing to accept certain assumptions on the choice of the initial interval $I_0 = (-B_1, +B_2)$, such as integrality of B_1 and B_2, the input $P(X)$ can be used to demonstrate the necessity of $\Omega(nL)$ bisections before α and β are separated. However, less customary choices of I_0 could cause some bisection to separate α and β much earlier.

We shall avoid this problem. Let us consider the closely related polynomial $P_2(X) = X^n - (aX - 1)^2$ which appears in a later work of Mignotte [Mig95] on complex roots. Again, we see that $P_2(a^{-1}) > 0$, and furthermore $P_2(a^{-1} - h) = (a^{-1} - h)^n - a^{-n} < 0$. Hence there is a root γ of $P_2(X)$ in $(a^{-1} - h,\, a^{-1})$. By irreducibility of $P(X)$, the product $Q(X) = P(X) \cdot P_2(X)$ is square free and has three distinct roots α, β, and γ in $(a^{-1} - h,\, a^{-1} + h)$.

Theorem 2.3. *Let $a \geq 3$ be an L-bit integer and $n \geq 4$ be an even integer. Consider the square-free polynomial $Q(X) = P(X) \cdot P_2(X)$ of degree $2n$. Its coefficients are integers of at most $O(L)$ bits. The Descartes method executed for $Q(X)$ and any initial interval I_0 enclosing all roots of $Q(X)$ has a recursion tree of height $\Omega(nL)$.*

Proof. As discussed above, I_0 is a superset of $(0, 1)$ and thus has width $w(I_0) > 1$. Let I_1 be the isolating interval reported by the Descartes method for the median of $\alpha, \beta, \gamma \in (a^{-1} - h,\, a^{-1} + h)$. Clearly, $w(I_1) < 2h$. The number of bisections needed to obtain I_1 from I_0 is $\log w(I_0)/w(I_1) > \log(1/2h) \geq (n/2 + 1)(L - 1) - 1 = \Omega(nL)$. □

Clearly, the same argument applies to any form of root isolation by repeated bisection, including Sturm's method.

2.3 The Bit Complexity

We derive the bit complexity of the Descartes method for a square-free polynomial $A_{\text{in}}(X)$ with integer coefficients of magnitude less than 2^L in the power basis. We can enclose all its real roots in an interval $(-B_1, +B_2)$ such that B_1 and B_2 are positive integers of magnitude less than 2^{L+1} (e.g., [Yap00, Cor. 6.8]).

We discuss the bit complexity of the power basis and Bernstein basis variants of the Descartes method applied to the scaled polynomial

$$A(X) := \sum_{i=0}^{n} a_i X^i := A_{\text{in}}((B_1 + B_2)X - B_1).$$

We can bound the bit length of its coefficients as follows. The power basis coefficients a_i of $A(X)$ have bit lengths $O(nL)$. For conversion from power basis to Bernstein basis, one has [PBP02, §2.8]

$$n!A(X) = \sum_{i=0}^{n} B_i^n(X) \sum_{k=0}^{i} i(i-1)\cdots(i-k+1)(n-k)!a_k. \tag{2.12}$$

To avoid fractions, we use $n!A(X)$ for the Bernstein basis variant. Observe that $l(l-1)\cdots(l-k+1)(n-k)! \leq n! \leq n^n$, so that the Bernstein coefficients of $n!A(X)$ have bit length $O(nL+n\log n)$.

From Corollary 2.1 we know that the size of the recursion tree is $O(n(L+\log n))$. Note that the transformation from $A_{\text{in}}(X)$ to $A(X)$ does not affect the size of the recursion tree, i.e., the size does not increase to $O(n(L'+\log n))$ where L' bounds the bit size of the coefficients of $A(X)$ or $n!A(X)$.

Let us now bound coefficient length at depth $h > 0$. For the power basis variant, we start with coefficients of length $O(nL)$. Both the H and TH transformations increase the length of the coefficients by $O(n)$ bits on each level. It is known that we can perform the T-transformation in $O(n^2)$ additions [Kra95, JKR05, vzGG99]; the H-transformation needs $O(n)$ shift operations. Hence a node at recursion depth h has bit cost $O(n^2(nL+nh))$ for the power basis. In the Bernstein basis, we need $O(n^2)$ additions and $O(n)$ shifts for the fraction-free de Casteljau algorithm, which also increases the length of the coefficients by $O(n)$ bits on each level. This gives us a bit cost of $O(n^2(nL+n\log n+nh))$. Since $h = O(n(L+\log n))$, the worst-case cost in any node is $O(n^4(L+\log n))$ for both variants. Multiplied with the tree size, this yields an overall bit complexity of $O(n^5(L+\log n)^2)$, cf. [Joh98, Thm. 13] [Kra95, Thm. 50]. To summarize:

Theorem 2.4. *Let $A(X)$ be a square-free polynomial of degree n with integer coefficients of magnitude less than 2^L. Then the bit complexity of isolating all real roots of $A(X)$ using the*

Descartes method (in either the power basis or the Bernstein basis variant) is $O(n^5(L+\log n)^2)$ using classical Taylor shifts in the power basis and de Casteljau's algorithm in the Bernstein basis. Except for the initial transformation, only additions and shifts are used.

For the Bernstein basis variant, this result is an improvement by a factor of n on the result in [MRR05]. For the power basis variant, this bound was already achieved by Krandick [Kra95]. Theorem 2.4 can be improved using a fast Taylor shift algorithm [vzGG99, Method F]:

Theorem 2.5. *Let $A(X)$ be a square-free polynomial of degree n with integer coefficients of magnitude less than 2^L. Then the bit complexity of isolating the real roots of $A(X)$ using the Descartes method in the power basis with a fast Taylor shift is $O(nM(n^3(L+\log n))(L+\log n))$. Here, $M(n)$ is the bit complexity of multiplying two n-bit integers.*

Proof. The work at a node at depth h of the recursion tree has bit cost

$$O(M(n^2\log n + n^2L + n^2h)$$

[vzGG99]. Substituting $h = O(n(L+\log n))$, we get the bound $O(M(n^3(L+\log n))$. Multiplied by the tree size $O(n(L+\log n))$, we obtain the desired result. □

To obtain a similar speedup for the Bernstein basis variant, Emiris, Mourrain, and Tsigaridas [EMT06] describe the following approach: Suppose the vector $(b_i)_i$ of Bernstein coefficients of $A(X) = \sum_{i=0}^n b_i B_i^n(X)$ is given and the Bernstein coefficients $(b'_i)_i$ of

$$A_L(X) = H(A)(X) = \sum_{i=0}^n b'_i B_i^n(X)$$

are wanted. Define the auxiliary polynomial

$$Q(X) = TR(A(X)) = \sum_{i=0}^n b_{n-i}\binom{n}{i}X^i$$

and transform it by substituting $2X+1$ for X. It is straightforward to verify that if $Q_L(X) := H_2T(Q)$, where $H_2(Q) = Q(2X)$, then $Q_L(X) = Q(2X+1) = \sum_{i=0}^n b'_{n-i}\binom{n}{i}X^i$. *Thus one can compute the Bernstein coefficients of $A_L(X)$ from the Bernstein coefficients of $A(X)$ using one asymptotically fast Taylor shift and scalings of coefficients.* By symmetry, the same holds for the Bernstein coefficients of $A_R(X)$. More precisely, if $Q_R(X) := RH_2TR(Q)$, then $Q_R(X) = (2+X)^nQ(X/(2+X)) = \sum_{i=0}^n b''_{n-i}\binom{n}{i}X^i$. Thus we get b''_i's, the Bernstein coefficients of $A_R(X)$. Together with bounds on the size of the recursion tree (Cor. 2.1) and the lengths of coefficients,

this leads [EMT06] to a bit complexity of $\widetilde{O}(n^4L^2)$ for the Bernstein basis variant of the Descartes method.

However, repeatedly putting in and taking out the extra factor $\binom{n}{i}$ in the i-th coefficient is an unnecessary artifact of insisting on the Bernstein basis. A more natural formulation of this approach avoids this extra scaling and the reversal of the coefficient sequence by representing polynomials in the scaled and reversed Bernstein basis $\widetilde{B}_i^n(X) = \binom{n}{i}^{-1} B_{n-i}^n(X) = (1-X)^i X^{n-i}$. Now the steps from $A(X)$ to $Q(X)$ and back from $Q(2X+1)$ to $A_L(X)$ are purely conceptual: reinterpret the coefficients of $\widetilde{B}_i^n(X)$ as coefficients of X^i and vice versa. The resulting algorithm is the **scaled Bernstein basis variant** of the Descartes method.

An alternative view on this variant is to regard it as an optimization of the power basis variant: By Eq. (2.1), the reinterpretation of coefficients is equivalent to the transformation TR. Recall that each recursive invocation of the power basis variant handles four polynomials: $A(X)$ is received from the parent, the Descartes test constructs $TR(A)(X)$, and subdivision computes $A_L(X)$ and $A_R(X)$. In these terms, the scaled Bernstein basis variant receives $TR(A)(X)$ instead of $A(X)$, eliminating the need for a separate transformation in the Descartes test, and it subdivides $TR(A)(X)$ into $TR(A_L)(X)$ and $TR(A_R)(X)$ directly, without explicitly constructing $A_L(X)$ and $A_R(X)$. Over the entire recursion tree, this saves one third of the T transformations in the power basis formulation.

2.4 Conclusion and Future Work

The aim of this chapter was to achieve the best possible complexity bounds for the Descartes method (using either the power basis or the Bernstein basis), and to match similar bounds for Sturm's method. We achieved matching bounds for two measures: (1) the size of the recursion tree, and (2) the bit complexity of the overall algorithm. Moreover, we showed that the tree size bound is the best possible under the assumption that $L = \Omega(\log n)$. It would be of some interest to completely resolve this optimality question.

Another direction of interest is to extend these algorithms and results to the case of polynomials that are not square-free. The standard way to achieve such extensions is to apply the above results to the square-free part $A/\gcd(A, A')$ of a given polynomial A (see, e.g., [BPR03, Algo. 10.41] [EMT06]) – but the real challenge is to provide an algorithm based on the Descartes method that works directly on polynomials that are not square-free.

A desirable aim would be to match the best know complexity bound of $\widetilde{O}(n^3L)$ which is at-

tained by Schönhage's algorithm [Sch82]. However, this calls for more insight and understanding of the Descartes method.

3
Real Root Isolation: Continued Fractions

In his paper, Sur la résolution des équations numériques [Vin36], Vincent made the following observation: Given a square-free polynomial $A(X) \in \mathbb{R}[X]$ of degree n, recursively define $A_i(X) := X^n A_{i-1}(a_i + X^{-1})$, where $A_0(X) := A(X)$ and a_i's are arbitrary positive integers, then for i sufficiently large $A_i(X)$ will have at most one sign variation; for Vincent's original proof and extensive historic information on related work, see [AG98]. Based upon this observation he suggested an algorithm that isolates the real roots of a polynomial. The algorithm can be described as a recursive procedure that at each recursion step takes as input a polynomial $A(X)$ and a Möbius transformation $M(X) = \frac{pX+q}{rX+s}$ such that $A(X) = (rX+s)^n A_{\text{in}}(M(X))$, where $A_{\text{in}}(X)$ is the original input polynomial of degree n; initially $A(X) = A_{\text{in}}(X)$ and $M(X)$ is X or $-X$ depending upon whether we want to isolate the positive or the negative roots of $A_{\text{in}}(X)$. It then constructs two polynomials $A_R(X) := A(X+1)$, $A_L(X) := (X+1)^n A(1/(X+1))$ and two Möbius transformations $M_R(X) := M(X+1)$, $M_L(X) := M(\frac{1}{X+1})$; the polynomial $A_L(X)$ and the transformation $M_L(X)$ are constructed only if the number of sign variations (i.e., the number of changes from positive to negative and vice versa) in the coefficients of $A_R(X)$ are less than those in $A(X)$. It can be verified that the roots of $A(X)$ greater than one correspond to the positive roots of $A_R(X)$ and the roots of $A(X)$ in the unit interval correspond to the positive roots of $A_L(X)$; the transformations $M_R(X)$ and $M_L(X)$ respectively define these correspondence. The algorithm then proceeds recursively on $(A_R(X), M_R(X))$, and on $(A_L(X), M_L(X))$ if $A_L(X)$ was constructed. The algorithm stops whenever it counts zero or one sign variation in the coefficients of $A(X)$; in the latter case, it outputs the interval with endpoints $M(0)$ and $M(\infty)$.

The search tree generated by the algorithm is a binary tree, and with each node of this tree we can associate a pair $(A(X), M(X))$. At a node along any path in this tree we transform the polynomial associated with the node by either the **inverse transformation** $X \to (X+1)^{-1}$ or the Taylor shift $X \to (X+1)$. Uspensky [Usp48] gave an explicit bound on the number of inverse transformations along any path of the tree. In particular, it can be deduced from his result that for a square-free integer polynomial of degree n with coefficients of bit-length L, the number of inverse transformations along any path in the search tree is $\widetilde{O}(nL)$ (see [Yap00, Cor. 14.6,p. 477]

or Equation (3.36) below), where $\widetilde{O}()$ means we omit logarithmic terms. The number of Taylor shifts, however, can be exponential; the length of the right most path is at least the floor of the smallest positive root of $A_{\text{in}}(X)$, which may be exponentially large.

Akritas [Akr78b] gave a modification of Vincent's algorithm, which maintains the spirit of the original, but tries to avoid exponential behaviour. Unlike Vincent, who tries to approximate the floor of the least positive root of the polynomial with shifts of unit length, Akritas shifts by a lower bound on the least positive root, thus speeding the process of approximation. Assuming the **ideal Positive Lower Bound (**PLB**) function**, i.e., a function which returns the floor of the least positive root of the polynomial and can identify if there is no such root, Akritas showed that his algorithm has a worst case complexity of $\widetilde{O}(n^5L^3)$ if the Taylor shifts are done in the classical way using $O(n^2)$ operations; but, as mentioned in [ET06], his analysis does not account for the increased coefficient size after performing the Taylor shifts. Again assuming the ideal PLB function, Emiris and Tsigaridas [ET06] have derived an expected bound of $\widetilde{O}(n^4L^2)$ on Akritas' algorithm using bounds by Khinchin [Khi97] on the expected bit-size of the partial quotients appearing in the continued fraction approximation of a real number. In practice, however, we never use the ideal PLB function because of its prohibitive cost (intuitively it is almost equivalent to doing real root isolation), but instead use functions that are based upon upper bounds on the absolute value of the roots of a polynomial, such as Cauchy's bound, Zassenhaus' bound (see §3.1 for details). The advantage of the latter bounds are that they are easy to implement and give us a good lower bound on the least positive root. Thus the complexity analysis of Akritas' algorithm in the current literature does not correspond with the actual implementation of the algorithm. Moreover, given the similarity between Akritas' and Vincent's algorithm, it may appear that the former is also exponential in the worst case, though we do not know of any example where this is the situation.

In this chapter, we derive a worst case bound of $\widetilde{O}(n^8L^3)$ on Akritas' algorithm *without assuming the ideal* PLB *function*. But if we are allowed to make this assumption then we can improve this worst case bound to $\widetilde{O}(n^5L^2)$.

Since the key distinction between Vincent's and Akritas' algorithm is the use of lower bounds on the positive real roots of the polynomial, it is clear that the worst case analysis of the latter algorithm has to hinge upon the tightness of these lower bounds. This is the subject that we treat next.

3.1 Tight Bounds on Roots

Given a polynomial $A(X) = \sum_{i=0}^{n} a_i X^i \in \mathbb{Z}[X]$, $a_0 \neq 0$, let PLB(A) be any procedure that returns a lower bound on the least positive root of $A(X)$. If $\mu(A)$ denotes the largest absolute value over all the roots of a polynomial $A(X)$, then PLB(A) is usually computed by taking the inverse of an upper bound on $\mu(R(A))$, where $R(A)(X) := X^n A(1/X)$. Thus in order to get a tight lower bound on the positive roots of a polynomial we need a tight upper bound on $\mu(A)$. One way of getting such an upper bound is to use the function:

$$S(A) := 2 \max_{i=1,\ldots,n} \left| \frac{a_{n-i}}{a_n} \right|^{1/i}.$$

Van der Sluis [vdS70] showed that $S(A)$ is at most twice the optimal bound amongst all bounds based solely upon the absolute value of the coefficients; he also showed the following:

$$\mu(A) < S(A) \leq 2n\mu(A). \tag{3.1}$$

The proof for the lower bound on $S(A)$ can be found, for instance, in [Yap00, Lem. 6.5, p. 147]. The upper bound on $S(A)$ will follow if we show for all $0 < i \leq n$ that $\left| \frac{a_{n-i}}{a_n} \right|^{1/i} \leq n\mu(A)$. Let $\alpha_1, \ldots, \alpha_n$ be the roots of $A(X)$. Then we know that

$$\left| \frac{a_{n-i}}{a_n} \right| \leq \sum_{1 \leq j_1 < \cdots < j_i \leq n} |\alpha_{j_1} \cdots \alpha_{j_i}| \leq \binom{n}{i} \mu(A)^i.$$

Taking the i-th root on both sides, along with the observation that for $1 \leq i \leq n$, $\binom{n}{i}^{1/i} \leq \left(\frac{n^i}{i!} \right)^{1/i} \leq n$, shows the upper bound in (3.1).

Clearly, $S(A)$ cannot be computed exactly in general. Instead, we use a procedure $U(A)$, similar to that suggested by Akritas [Akr89, p. 350], which computes an upper bound on $\mu(A)$ when $A(X) \in \mathbb{R}[X]$.

Procedure U(A)
INPUT: An integer polynomial $A(X) = \sum_{i=0}^{n} a_i x^i$, $a_i \in \mathbb{R}$, $a_n > 0$.
OUTPUT: A power of two that is an upper bound on the roots of $A(X)$.
1. Let m be the number of negative coefficients of $A(X)$.
2. **If** $n = 0$ or $m = 0$ **then return**.
3. $q' := -\infty$.
4. **For** i **from** 1 **to** n do the following:
   $p := \lfloor \log |a_{n-i}| \rfloor - \lfloor \log |a_n| \rfloor - 1$.
   Let $q = \lfloor p/i \rfloor$.
   $q' := \max(q', q+2)$.
5. **Return** $2^{q'}$.

Remark 3.1. *If $A(X)$ is an integer polynomial with coefficients of bit-length L then the cost of computing $U(A)$ is $\widetilde{O}(nL)$. The reason is that most expensive operation in the loop on Line 4 is computing the floor of the coefficients which can be done in $O(L)$ time; since the loop runs n times we have the desired bound.*

We have the following relation between $U(A)$ and $S(A)$:

Lemma 3.1.

$$\frac{U(A)}{4} < S(A) < U(A).$$

Proof. Suppose $S(A) = 2\left(\frac{|a_{n-i}|}{|a_n|}\right)^{1/i}$. Let $p := \lfloor \log |a_{n-i}| \rfloor - \lfloor \log |a_n| \rfloor - 1$, $q = \lfloor p/i \rfloor$ and $r := p - q \cdot i$, $0 \le r < i$. Then we know that

$$2^p < \frac{|a_{n-i}|}{|a_n|} < 2^{p+2}.$$

Taking the i-th root we get

$$2^q < \left(\frac{|a_{n-i}|}{|a_n|}\right)^{1/i} < 2^{q+2},$$

since $q \le p/i$ and $(p+2)/i = q + (r+2)/i \le q+2$. But $U(A) = 2^{q+2}$, and hence we get our desired inequality. □

The lemma above along with (3.1) gives us

$$\mu(A) < U(A) < 8n\mu(A). \tag{3.2}$$

Define

$$\mathrm{PLB}(A) := \frac{1}{U(R(A))} \tag{3.3}$$

where $R(A) = X^n A(1/X)$. Then from (3.2) we know that

$$\frac{1}{8n\mu(R(A))} < \mathrm{PLB}(A) < \frac{1}{\mu(R(A))}.$$

Let $\kappa(A)$ denote the minimum of the absolute values of the roots of $A(X)$, assuming that zero is not a root of $A(X)$. Then we have

$$\frac{\kappa(A)}{8n} < \mathrm{PLB}(A) < \kappa(A), \tag{3.4}$$

since $\kappa(A) = \frac{1}{\mu(R(A))}$.

In practice one can use bounds from [Kio86, Şte05] that utilize the sign of the coefficients and give a better estimate than $S(A)$. For instance, Kioustelidis [Kio86] has shown that the bound

$$K(A) := 2 \max_{a_i<0} \left| \frac{a_{n-i}}{a_n} \right|^{1/i},$$

where a_i is a coefficient of $A(X)$, is an upper bound on the largest positive root of $A(X)$; by definition we have $K(A) \leq S(A)$. We do not use this bound in our analysis because we do not know a relation corresponding to (3.1) between $K(A)$ and the largest positive root of $A(X)$. But such a relation seems unlikely to hold. Consider the situation when there is only one negative root of $A(X)$, which has the largest absolute value amongst all the roots of $A(X)$. Then the summation in

$$\left| \frac{a_j}{a_n} \right| = \left| \sum_{1 \leq i_1 < \cdots < i_{n-j} \leq n} \alpha_{i_1} \cdots \alpha_{i_{n-j}} \right|,$$

where $\alpha_1, \ldots, \alpha_n$ are the roots of $A(X)$, will be dominated by the negative root. Thus, it appears, that the best we can say is $K(A) \leq 2n\mu(A)$. The same argument applies to the bound by Ştefănescu [Şte05].

Hong [Hon98] has given bounds on the positive roots of a polynomial in more than one variable. For univariate polynomials, his bound is

$$2 \max_{a_j<0} \min_{a_k>0, k>j} \left| \frac{a_j}{a_k} \right|^{1/(k-j)}.$$

It is clear that this bound is an improvement over the bound by Kioustelidis. However, again it is not obvious whether a tight relation similar to (3.1) holds between the largest positive root and the bound above. This is because Hong's bound is on the absolute positiveness of a polynomial,

i.e., a bound such that the evaluation of the polynomial and *all* its derivatives at any point larger than the bound is strictly positive. In case of univariate polynomials this means Hong's bound is an upper bound on the positive roots of the polynomial and its derivatives. The difficulty in obtaining a tight relation suggested above is that the real roots of the derivatives may be greater than the positive roots of the polynomial. For example, the derivative of the polynomial $3X^3 - 15X^2 + 11X - 7 = 3(X-1)(X-2+i/\sqrt{3})(X-2-i/\sqrt{3})$ has a real root $5/3$ which is greater than the real root (in this case one) of the polynomial. The desired tight relation may be derived, if we can find a relation between the largest positive root of the polynomial and the largest positive root of its derivative.

Now that we have a procedure to obtain lower bounds on the absolute values of the roots of the polynomial, we give the details of Akritas' algorithm for real root isolation.

3.2 The Continued Fraction Algorithm by Akritas

Definition 3.1. Given a polynomial $A(X) = a_n X^n + a_{n-1}X^{n-1} + \cdots + a_0$, let $\texttt{Var}(A)$ represent the number of sign changes (positive to negative and negative to positive) in the sequence $(a_n, a_{n-1}, \ldots, a_0)$.

The two crucial components of the algorithm are the procedure PLB(A), described above, and the Descartes' rule of signs:

Proposition 3.1. *Let $A(X) = a_n X^n + a_{n-1}X^{n-1} + \cdots + a_0$ be a polynomial with real coefficients, which has exactly p positive real roots counted with multiplicities. Then $\texttt{Var}(A) \geq p$, and $\texttt{Var}(A) - p$ is even.*

Akritas' algorithm for isolating the real roots of a square-free input polynomial $A_{\text{in}}(X)$ uses a recursive procedure CF(A, M) that takes as input a polynomial $A(X)$ and a Möbius transformation $M(X) = \frac{pX+q}{rX+s}$, where $p, q, r, s \in \mathbb{N}$ and $ps - rq \neq 0$. With the transformation $M(X)$ we can associate an interval I_M that has endpoints p/r and q/s. The relation among $A_{\text{in}}(X)$, $A(X)$ and $M(X)$ is

$$A(X) = (rX + s)^n A_{\text{in}}(M(X)). \tag{3.5}$$

From this relation it follows that 1) for every root $\alpha \in I_M$ of $A_{\text{in}}(X)$ there is a unique root $\beta > 0$ of $A(X)$ such that $M^{-1}(\alpha) = \beta$ and vice versa, i.e., for every root $\beta > 0$ of $A(X)$ there is a unique root $\alpha \in I_M$ of $A_{\text{in}}(X)$ such that $\alpha = M(\beta)$; and 2) that $A(X)$ is square-free. Given (3.5),

the procedure $\mathrm{CF}(A, M)$ returns a list of isolating intervals for the roots of $A_{\mathrm{in}}(X)$ in I_M. To isolate all the positive roots of $A_{\mathrm{in}}(X)$ initiate $\mathrm{CF}(A, M)$ with $A(X) = A_{\mathrm{in}}(X)$ and $M(X) = X$; to isolate the negative roots of $A_{\mathrm{in}}(X)$ initiate $\mathrm{CF}(A, M)$ on $A(X) := A_{\mathrm{in}}(-X)$ and $M(X) = X$ and flip the signs of the intervals returned.

The procedure $\mathrm{CF}(A, M)$ is as follows:

Procedure $\mathrm{CF}(A, M)$

Input: A square-free polynomial $A(X) \in \mathbb{R}[X]$ and a Möbius transformation $M(X)$ satisfying (3.5).

Output: A list of isolating intervals for the roots of $A_{\mathrm{in}}(X)$ in I_M.

1. **If** $A(0) = 0$ **then**
   Output interval $[M(0), M(0)]$.
   $A(X) := A(X)/X$; **return** $\mathrm{CF}(A, M)$.
2. **If** $\mathtt{Var}(A) = 0$ **then return**.
3. **If** $\mathtt{Var}(A) = 1$ **then**
   Output the interval I_M and **return**.
4. $b := \mathrm{PLB}(A)$.
5. **If** $b > 1$ **then** $A(X) := A(X + b)$ and $M(X) := M(X + b)$.
6. $A_R(X) := A(1 + X)$ and $M_R(X) := M(1 + X)$.
7. $\mathrm{CF}(A_R, M_R)$.
8. **If** $\mathtt{Var}(A_R) < \mathtt{Var}(A)$ **then**
9. $A_L(X) := (1 + X)^n A\left(\frac{1}{1+X}\right)$ and $M_L(X) := M\left(\frac{1}{1+X}\right)$, $n = \deg(A)$.
10. **If** $A_L(0) = 0$ **then** $A_L(X) := A_L(X)/X$.
11. $\mathrm{CF}(A_L, M_L)$.

Some remarks on the procedure:

- Removing lines 4 and 5 gives us the procedure proposed by Vincent for isolating positive roots.
- The positive roots of $A_R(X)$ are in bijective correspondence with the roots of $A_{\mathrm{in}}(X)$ in the interval I_{M_R} and those of $A_L(X)$ are in bijective correspondence with the roots of $A_{\mathrm{in}}(X)$ in the interval I_{M_L}.
- Line 8 avoids unnecessary computations of $A_L(X)$ since if $\mathtt{Var}(A_R) = \mathtt{Var}(A)$ then from

Budan's theorem [Akr82] we know that there are no real roots of $A(X)$ between 0 and $b+1$. This test is missing in Uspensky's formulation of the algorithm [Usp48, p. 128] and was pointed out in [Akr86].

- Line 10 is necessary to avoid recounting; since $A_L(0) = 0$ if and only if $A_R(0) = 0$, the root would have been reported in the recursive call at line 7.

- $A_L(X)$ can be computed from $A(X)$ by performing a Taylor shift by one on the reverse polynomial $X^n A(1/X)$.

To show that each path in the recursion tree of the above algorithm terminates we need to revise some basic relations between Möbius transformations and continued fractions; for details see [Yap00, Ch. 15].

3.3 Continued Fractions and Möbius Transformations

A continued fraction is a possible infinite expression of the form

$$q_0 + \cfrac{p_1}{q_1 + \cfrac{p_2}{q_2 + \cfrac{p_3}{q_3 + \cdots}}}$$

where p_i, q_i are in $\overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$; p_i is called the i-th partial numerator and q_i the i-th partial denominator. For ease of writing, we express the above continued fraction as

$$q_0 + \frac{p_1}{q_1+}\frac{p_2}{q_2+}\cdots .$$

The value P_i/Q_i of the continued fraction

$$q_0 + \frac{p_1}{q_1+}\frac{p_2}{q_2+}\cdots\frac{p_{i-1}}{q_{i-1}+}\frac{p_i}{q_i}$$

is called the i-th quotient of the continued fraction; it may be infinite if $P_i \neq 0$ and $Q_i = 0$, or indefinite if $P_i = Q_i = 0$. In particular, for $i = 0$ we have $P_0 = q_0$ and $Q_0 = 1$. If we choose $P_{-1} = 1$ and $Q_{-1} = 0$ then we have

$$P_i = p_i P_{i-2} + q_i P_{i-1} \text{ and } Q_i = p_i Q_{i-2} + q_i Q_{i-1}. \tag{3.6}$$

An **ordinary continued fraction** [1] is of the form

$$q_0 + \frac{1}{q_1+}\frac{1}{q_2+}\frac{1}{q_3+}\cdots,$$

i.e., a continued fraction all of whose partial numerators are one. Again, for the ease of writing we express it as

$$[q_0, q_1, q_2, \ldots].$$

For example, the ordinary continued fraction $[1, 1, 1, \ldots]$ is the golden ratio $(\sqrt{5}+1)/2$. With the finite ordinary continued fraction $[q_0, \ldots, q_m] = \frac{P_m}{Q_m}$ we can associate the Möbius transformation

$$M(X) := \frac{P_{m-1} + P_m X}{Q_{m-1} + Q_m X}.$$

We denote by I_M the interval with endpoints $M(\infty) = P_m/Q_m$ and $M(0) = P_{m-1}/Q_{m-1}$. Since $[q_0, q_1, \ldots, q_m]$ is an ordinary continued fraction, we know that [Yap00, p. 463]

$$|P_m Q_{m-1} - P_{m-1} Q_m| = 1; \tag{3.7}$$

thus the Möbius transformation associated with an ordinary continued fraction is unimodal. For any two numbers $\alpha, \eta \in \overline{\mathbb{C}}$, if $\alpha = M(\eta)$ then by applying the inverse transformation $M^{-1}(X)$ we get

$$\eta = -\frac{P_{m-1} - Q_{m-1}\alpha}{P_m - Q_m \alpha}. \tag{3.8}$$

For a complex number z let $\Re(z)$ represent its real part and $\Im(z)$ its imaginary part.

3.4 Termination

Consider the recursion tree of the procedure CF(A, M), described in §3.2, initiated with $A(X) = A_{\text{in}}(X) \in \mathbb{R}[X]$ and $M(X) = X$, for a square-free polynomial $A_{\text{in}}(X)$. The right child of any node in this tree corresponds to the Taylor shift $X \to X + \delta$, $\delta \geq 1$, and the left child of the node corresponds to the inverse transformation $X \to (X+1)^{-1}$. A sequence of Taylor shifts $X \to X_0 + \delta_0, X_0 \to X_1 + \delta_1, \ldots, X_{i-1} \to X_i + \delta_i$ can be thought of as a single Taylor shift $X \to X + q$, $q = \sum_{j=0}^{i} \delta_j$. Moreover, a sequence of Taylor shifts by a total amount q followed by an inverse transformation $X \to (X+1)^{-1}$ is the same as the transformation $X \to q + (1+X)^{-1}$. Thus with each node in the recursion tree we can associate an ordinary continued fraction $[q_0, q_1, \ldots, q_m] = P_m/Q_m$, for some $q_i \in \mathbb{N}$, and hence the Möbius transformation $(P_m X +$

[1] They are also called simple continued fractions, or regular continued fractions.

$P_{m-1})/(Q_mX + Q_{m-1})$; note that the nodes on the right most path of the recursion tree are associated with the continued fraction $[q_0]$, for some $q_0 \in \mathbb{N}$, and the Möbius transformation $X + q_0$, because there are no inverse transformations along the right most path. Based upon the Möbius transformation $(P_mX + P_{m-1})/(Q_mX + Q_{m-1})$ associated with a node in the recursion tree, we can further associate the polynomial $A_M(X) := (Q_mX + Q_{m-1})^n A_{\text{in}}(P_mX + P_{m-1})$ with the same node.

Vincent's observation, mentioned earlier, says that if m is large enough then $A_M(X)$ will exhibit at most one sign variation. Uspensky [Usp48, p. 298] quantified this by showing the following: Let $A_{\text{in}}(X) \in \mathbb{R}[X]$ be a square-free polynomial of degree n and Δ be the smallest distance between any pair of its roots. If m is such that

$$F_{m-1}\frac{\Delta}{2} > 1 \text{ and } F_{m-1}F_m\Delta > 1 + \epsilon_n^{-1} \tag{3.9}$$

where F_i is the i-th Fibonacci number and $\epsilon_n := (1 + 1/n)^{1/(n-1)} - 1$, then $A_M(X)$ exhibits at most one sign variation [2].

Ostrowski [Ost50] later improved and simplified Uspensky's criterion (3.9) to $F_mF_{m-1}\Delta \geq \sqrt{3}$. Similar criterion were derived by Alesina and Galuzzi [AG98, p. 246] and Yap [Yap00, Thm. 14.5,p. 476]. We next derive a termination criterion that depends on Δ_α, the shortest distance from α to another root of $A(X)$. To achieve this we recall from §2.1.2 the definitions of the three open discs, $\overline{C}_I$, $\underline{C}_I$ and C_I w.r.t. an interval I. Also, following [KM06] we define the **cone**

$$\mathcal{C} := \left\{ a + ib | a \leq 0 \text{ and } |b| \leq |a|\sqrt{3} \right\}.$$

We have the following key observation which is implicit in Ostrowski's proof and is also used by Alesina and Galuzzi in [AG98, p. 249]:

Lemma 3.2. *Let $a, b, c, d \in \mathbb{R}_{>0}$, I be an interval with unordered endpoints $\frac{a}{c}, \frac{b}{d}$, and define the Möbius transformation $M(z) := \frac{az+b}{cz+d}$. Then $M^{-1}(z)$ maps the closed region $\overline{\mathbb{C}} - (\overline{C}_{I_M} \cup \underline{C}_{I_M})$ bijectively on the cone $\mathcal{C}$, and maps the open disc C_{I_M} bijectively on the half plane $\Re(z) > 0$.*

Proof. Since Möbius transformations map circles to circles (lines are circles with infinite radius) we only show the effect of $M^{-1}(z)$ on three points of the circles.

Suppose $a/c < b/d$. Then it can be verified that $M^{-1}(b/d) = 0$, $M^{-1}(a/c) = \infty$, and $M^{-1}(z) = \frac{-d}{2c}(1 + i\sqrt{3})$, $M^{-1}(\overline{z}) = \frac{-d}{2c}(1 - i\sqrt{3})$ where

$$z := \frac{1}{2}(a/c + b/d) + i(b/d - a/c)\sqrt{3}/2.$$

[2] Uspensky's original proof incorrectly states $F_{m-1}\Delta > \frac{1}{2}$. This was later corrected by Akritas [Akr78a].

This proves the first bijection. The second bijection follows from the additional fact that for

$$w = \frac{1}{2}(a/c + b/d) + i\frac{1}{2}(b/d - a/c)$$

we have $M^{-1}(w) = -id/c$ and $M^{-1}(\overline{w}) = id/c$. This correspondence is illustrated in Figure 3.1.

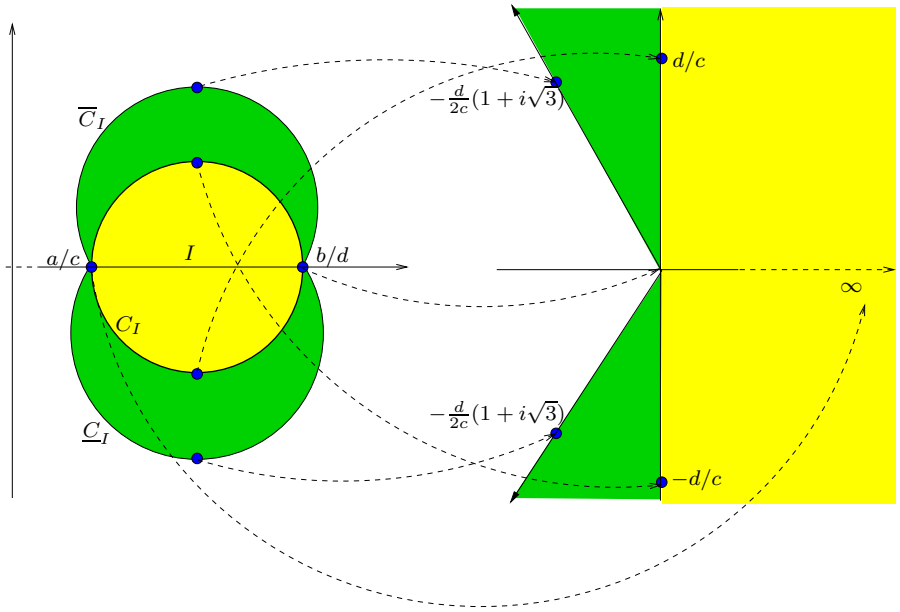

Figure 3.1: The effect of $M^{-1}(z)$ on the three circles

The lemma holds if $a/c > b/d$ because the point $\frac{1}{2}(a/c + b/d) + i(a/c - b/d)\sqrt{3}/2$ is the conjugate of the point z defined above, and similarly the point $\frac{1}{2}(a/c + b/d) + i\frac{1}{2}(a/c - b/d)$ is the conjugate of the point w. □

From Lemma 3.2 it follows that $\mathrm{Var}(A_M) = \mathit{DescartesTest}(A, I_M)$, and hence from Proposition 2.3 we know the following:

Theorem 3.2. *Let $A(X)$ be a square-free polynomial of degree n,*

$$M(X) := \frac{P_m X + P_{m-1}}{Q_m X + Q_{m-1}}$$

and $A_M(X) := (Q_m X + Q_{m-1})^n A(M(X))$. If α is the only simple root of $A(X)$ in the interval I_M and there are no other roots of $A(X)$ in $\overline{C}_{I_M} \cup \underline{C}_{I_M}$ then $\mathtt{Var}(A_M) = 1$.

As a corollary to the above we have the following termination criterion:

Corollary 3.1. *Let $A(X)$ be a square-free polynomial of degree n,*

$$M(X) := \frac{P_m X + P_{m-1}}{Q_m X + Q_{m-1}}$$

and $A_M(X) := (Q_m X + Q_{m-1})^n A(M(X))$. If I_M is an isolating interval for a real root α of $A(X)$ and m is such that $Q_m Q_{m-1} \Delta_\alpha \geq 2/\sqrt{3}$ then $\mathtt{Var}(A_M) = 1$.

Proof. We know from (3.7) that the width of I is $1/(Q_m Q_{m-1})$. By our assumption this is less than $\sqrt{3}\Delta_\alpha/2$. Thus all the roots of $A(X)$ except α are in $\overline{\mathbb{C}} - (\overline{C}_{I_M} \cup \underline{C}_{I_M})$. From Lemma 3.2 we know that all the roots of $A_M(X)$ except the one corresponding to α lie in the cone $\mathcal{C}$ and from Proposition 2.3 we get that $\mathtt{Var}(A_M) = 1$. □

Remark 3.2. *This result seems to be the one hinted in [ET06, p. 5] as an interesting problem. Their Remark 7, however, is equivalent to this result but cannot be derived directly by substituting Δ_α for Δ in the standard proofs by Uspensky, Ostrowski, Akritas and Yap, since all of them depend upon the fact that the imaginary parts of all the imaginary roots are at least $\Delta/2$, a criterion which may not hold in case of Δ_α. The result, however, is straightforward if we replace Δ by Δ_α in the proof of the termination criterion derived by Alesina and Galuzzi.*

The above theorem corresponds to the two-circle theorem in [KM06]. The corresponding one-circle theorem, which again is a direct consequence of Lemma 3.2 and Proposition 2.2, is the following :

Theorem 3.3. *Let $A(X)$ be a square-free polynomial of degree n,*

$$M(X) := \frac{P_m X + P_{m-1}}{Q_m X + Q_{m-1}}$$

and $A_M(X) := (Q_m X + Q_{m-1})^n A(M(X))$. If C_{I_M} does not contain any roots then $\mathtt{Var}(A_M) = 0$.

As a corollary to the above we have:

Corollary 3.2. *Let $A(X)$ be a square-free polynomial of degree n,*

$$M(X) := \frac{P_m X + P_{m-1}}{Q_m X + Q_{m-1}}$$

and $A_M(X) := (Q_m X + Q_{m-1})^n A(M(X))$. If C_{I_M} does not contain any roots and m is such that $Q_m Q_{m-1} \Delta \geq 2$ then $\mathtt{Var}(A_M) = 0$, where Δ is the root separation bound for $A(X)$.

Proof. Again the width of the interval I_M is less than $\Delta/2$ and hence the open disc C_{I_M} does not contain any roots. Thus all the roots of $A_M(X)$ have a negative real part and hence $\mathtt{Var}(A_M) = 0$. □

3.5 The Size of the Recursion Tree: Real Roots Only

In this section we will bound the size $\#(T)$ of the recursion tree T of the procedure $\mathrm{CF}(A_{\mathrm{in}}, X)$, where $A_{\mathrm{in}}(X) \in \mathbb{R}[X]$ is a square-free polynomial of degree n with *only real roots*. Consider a sub-tree T' of T obtained by pruning certain leaves from T:

- Prune all the leaves of T that declare the absence of roots.
- If two leaves are siblings of each other, prune the left one.

Clearly, $\#(T') < \#(T) < 2\#(T')$; thus it suffices to bound $\#(T')$.

Let U be the set of leaves in T'. Recall from §3.4 that with each node of T we can associate an ordinary continued fraction. In particular, for a leaf $u \in U$ let $[q_0, \ldots, q_{m+1}] = P_{m+1}/Q_{m+1}$ be the associated ordinary continued fraction. Then we can associate with u the Möbius transformation

$$M_u(X) := \frac{P_{m+1}X + P_m}{Q_{m+1}X + Q_m} \tag{3.10}$$

and the interval $I_u := I_{M_u}$. Moreover, we can also associate with u a unique pair (α_u, β_u), where α_u and β_u are roots of $A_{\mathrm{in}}(X)$. The way T' has been constructed we know that for all $u \in U$ there is a real root $\alpha_u \in I_u$. To define the root β_u we consider the interval associated with the parent v of u. Let the Möbius transformation $M_v(X)$ associated with v be

$$M_v(X) = \frac{P_m X + P_{m-1} + P_m \delta_v}{Q_m X + Q_{m-1} + Q_m \delta_v}, \tag{3.11}$$

for some $\delta_v \in \mathbb{N}$ such that $1 \leq \delta_v < q_{m+1}$, and the interval associated with v be $I_v := I_{M_v}$. Since $\mathtt{Var}(A_{M_v}) > 1$, from Theorem 3.2 we know that there must be a root apart from α_u in the interval I_v; choose $\beta_u \in I_v$ to be the root closest to α_u. It is clear that for two leaves $u, u' \in U$, the corresponding pairs (α_u, β_u) and $(\alpha_{u'}, \beta_{u'})$ are such that $\alpha_u \neq \alpha_{u'}$, even though β_u may be the same as $\beta_{u'}$; from this uniqueness it follows that the size $|U|$ of the set U is at most n, the degree of the polynomial $A_{\mathrm{in}}(X)$.

We will bound the length of the path from the root of T' to a leaf $u \in U$ by bounding the number of inverse transformations $X \to 1/(X+1)$ and Taylor shifts $X \to (X+b)$, $b \geq 1$, along the path. For the remaining part of this section, let $M_u(X)$, I_u, v, $M_v(X)$ and I_v be as defined above.

3.5.1 Bounding the Inverse Transformations

Since both α_u and β_u are in the interval I_v we get

$$(Q_m Q_{m-1})^{-1} \geq (Q_m(Q_{m-1} + Q_m\delta_v))^{-1} \geq |\alpha_u - \beta_u|. \tag{3.12}$$

But $Q_i \geq F_{i+1}$, the $(i+1)$-th Fibonacci number; this follows from the recurrence equation (3.6) and the fact that $p_i = 1$ and $q_i \geq 1$. Moreover, $F_{i+1} \geq \phi^i$, where $\phi = (\sqrt{5}+1)/2$. Thus $Q_i \geq \phi^i$ and hence from (3.12) we have

$$\phi^{2m-1} \leq |\alpha_u - \beta_u|^{-1},$$

which implies

$$m \leq \frac{1}{2}(1 - \log_\phi |\alpha_u - \beta_u|). \tag{3.13}$$

Thus the total number of inverse transformations in T' are bounded by

$$\sum_{u \in U} \frac{1}{2} + \frac{1}{2}\log_\phi \prod_{u \in U} |\alpha_u - \beta_u|^{-1}.$$

Since $|U| \leq n$, the above bound is smaller than

$$n + \frac{1}{2}\log_\phi \prod_{u \in U} |\alpha_u - \beta_u|^{-1}. \tag{3.14}$$

3.5.2 Bounding the Taylor Shifts

The purpose of the Taylor shifts in the procedure $\mathrm{CF}(A, M)$ was to compute the floor of the least positive root of a polynomial. Using property (3.4) of the $\mathrm{PLB}(A)$ function (defined in (3.3)) we will bound the number of Taylor shifts required to compute the floor of the least positive root of some polynomial $B(X) \in \mathbb{R}[X]$. We start with the simple case when $B(X)$ has only positive roots. We introduce the following notation for convenience: for any $x \in \mathbb{C}$ let

$$\mathrm{logm}\,(x) := \log \max(1, |x|).$$

Lemma 3.3. *Let $B_1(X) := B(X)$, and for $i > 1$ recursively define*

$$B_i(X) := \begin{cases} B_{i-1}(X + \mathrm{PLB}(B_{i-1}) + 1) & \text{if } \mathrm{PLB}(B_{i-1}) > 1 \\ B_{i-1}(X+1) & \text{otherwise.} \end{cases} \tag{3.15}$$

Let α_i denote the least positive real root of $B_i(X)$. Then $\alpha_i \leq 1$ if $i \geq 2 + 8n + \gamma_n \mathrm{logm}\,\alpha_1$ where

$$\gamma_n := (\log 8n - \log(8n-1))^{-1}.$$

Proof. Let $b_i := \mathrm{PLB}(B_i)$. Then from (3.4) we get that for all $i \geq 1$, $\frac{\alpha_i}{8n} < b_i < \alpha_i$. Let j be the index such that for all $i < j$, $\alpha_i > 8n$. Then for $i < j$ we know that $b_i > 1$. Thus

$$\alpha_i = \alpha_{i-1} - b_{i-1} - 1 < \alpha_{i-1}\left(1 - \frac{b_{i-1}}{\alpha_{i-1}}\right) < \alpha_{i-1}(1 - \frac{1}{8n})$$

and recursively we get $\alpha_i < \alpha_1(1 - \frac{1}{8n})^{i-1}$. So $\alpha_j \leq 8n$ if $\alpha_{j-1} \leq 8n$, but this follows if $\alpha_1(1 - \frac{1}{8n})^{j-2} \leq 1$ or if $j \geq 2 + \gamma_n \mathrm{logm}\, \alpha_1$.

Since α_i is monotonically decreasing, for $i \geq j$ we have $\alpha_i \leq 8n$. Hence we need $8n$ additional shifts, i.e., if $i - j > 8n$ then $\alpha_i < 1$. Combining this with the lower bound on j gives us our result. □

The following lemma extends this result to the case when $B(X)$ has both positive and negative roots but $B(0) \neq 0$. In order to do this we introduce some notation: Let $\mathrm{LP}(B)$ denote the **least positive** root of $B(X)$ and $\mathrm{LN}(B)$ denote the **largest negative** root of $B(X)$.

Lemma 3.4. *Let $B_1(X) := B(X)$, and recursively define $B_i(X)$ as in the above lemma. Let $\alpha_i := LP(B_i)$ and $\beta_i := LN(B_i)$. Then $\alpha_i \leq 1$ if*

$$i > 3 + 16n + \kappa_n \mathrm{logm}\left(\frac{\alpha_1}{\beta_1}\right) + \mathrm{logm}\, \alpha_1$$

where

$$\kappa_n := (\log(8n+1) - \log 8n)^{-1}. \tag{3.16}$$

Proof. We suppose $|\alpha_1| > |\beta_1|$, since otherwise the result trivially follows from Lemma 3.3. Let j be the first index such that for $i \geq j$, $|\beta_i| > |\alpha_i|$; once this holds we know that α_i, $i \geq j$, is the root with the smallest absolute value and hence from Lemma 3.3 it follows that if $i - j > 2 + 8n + \gamma_n \mathrm{logm}\, \alpha_i$ then $\alpha_i \leq 1$; but $\kappa_n > \gamma_n$ and hence if

$$i - j > 2 + 8n + \kappa_n \mathrm{logm}\, \alpha_i \tag{3.17}$$

then we are sure $\alpha_i \geq 1$. We next give a lower bound on j.

Again, let $b_i := \mathrm{PLB}(B_i)$. Since for $i \leq j$, β_i is the root with smallest absolute value, we know from (3.4) that $\frac{|\beta_j|}{8n} < b_j < |\beta_j|$. Assume that $|\beta_i| > 8n$. Then $b_i > 1$ and from the definition of $B_i(X)$ we know that

$$|\beta_i| = |\beta_{i-1}|\left(1 + \frac{1 + b_{i-1}}{|\beta_{i-1}|}\right) > |\beta_{i-1}|\left(1 + \frac{b_{i-1}}{|\beta_{i-1}|}\right) > |\beta_{i-1}|\left(1 + \frac{1}{8n}\right).$$

And recursively, $|\beta_i| > |\beta_1|(1 + \frac{1}{8n})^{i-1}$. Thus if

$$j > 1 + \kappa_n \mathrm{logm}\left(\frac{\alpha_1}{\beta_1}\right)$$

then we are sure $|\beta_j| \geq |\alpha_1| > |\alpha_j|$. In addition to these shifts, we initially need $8n$ shifts to ensure that $|\beta_i| > 8n$. Combining these additional shifts with the bound on j and the bound on $i - j$ in (3.17) we get that if

$$i > 3 + 16n + \kappa_n \text{logm} \left(\frac{\alpha_1}{\beta_1} \right) + \kappa_n \text{logm}\, \alpha_i$$

then $\alpha_i \leq 1$. Since $|\alpha_i| \leq |\alpha_1|$ we have our result. □

To bound the number of Taylor shifts along the path from the root of T' to the leaf u, we will bound the number of Taylor shifts that compose each q_i, $i = 0, \ldots, m+1$, in the continued fraction approximation of α_u.

Definition 3.4. For $0 \leq i \leq m+1$ define the following quantities:

1. $M_i(X) := [q_0, \ldots, q_i, 1 + X] = \frac{P_i X + P_{i-1} + P_i}{Q_i X + Q_{i-1} + Q_i}$;
2. $A_i(X) := (Q_i X + Q_{i-1} + Q_i)^n A_{\text{in}}(M_i(X))$, i.e., the polynomial obtained by performing the ith inverse transformation and on which we will perform a Taylor shift by the amount q_{i+1};
3. $\eta_i := M_i^{-1}(\alpha_u)$
4. $r_i := P_i/Q_i$, $s_i := \frac{P_i + P_{i-1}}{Q_i + Q_{i-1}}$ and
5. $J_i := I_{M_i}$, i.e., the interval with endpoints r_i and s_i.

Clearly, for $0 \leq i \leq m$, $I_v \subseteq J_i$ and hence we have

$$(Q_i Q_{i-1})^{-1} \geq |\alpha_u - \beta_u|. \tag{3.18}$$

The same cannot be said for Q_{m+1}, because $J_{m+1} \subseteq I_v$ is an isolating interval.

Since $\alpha_u \in J_i$ we know from (3.8) that $\eta_i > 0$. Based upon the above lemmas we derive an upper bound on the number of Taylor shifts needed to obtain q_{i+1}.

Let $B_1(X) := A_i(X)$ and recursively define $B_i(X)$ as in (3.15); we may safely assume that $A_i(0) \neq 0$ since the procedure $\text{CF}(A_i, M_i)$ would replace $A_i(X)$ by $A_i(X)/X$ otherwise. Define the sequence of indices

$$i_0 = 1 \leq i_1 < i_2 < \cdots < i_\ell, \tag{3.19}$$

where the index i_j is such that $\text{LP}(B_{i_j})$ is contained in the unit interval; if $i < m$ the last index i_ℓ is such that the root in $B_{i_\ell}(X)$ corresponding to η_i is in the unit interval; for $i = m$ the index

i_ℓ is such that the node that has $B_{i_\ell}(X)$ as the corresponding polynomial is the parent v of the leaf u; these constraints imply that $\ell \leq n$.

From Lemma 3.4 we get

$$i_{j+1} - i_j = O\left(n + \kappa_n \operatorname{logm} \frac{\mathrm{LP}(B_{1+i_j})}{\mathrm{LN}(B_{1+i_j})} + \kappa_n \operatorname{logm} \mathrm{LP}(B_{1+i_j})\right).$$

But $\mathrm{LP}(B_k)$, $k = 1, 2, \ldots, i_\ell$, corresponds to a positive root, say η_k, of $A_i(X)$; moreover, $\eta_k > \mathrm{LP}(B_k)$, thus

$$i_{j+1} - i_j = O\left(n + \kappa_n \operatorname{logm} \frac{\eta_{1+i_j}}{\mathrm{LN}(B_{1+i_j})} + \kappa_n \operatorname{logm} \eta_{1+i_j}\right).$$

Summing this inequality for $j = 0, \ldots, \ell - 1 < n$ we get that

$$i_\ell = O(n^2) + O\left(\sum_{j=0}^{\ell-1} \kappa_n \operatorname{logm} \frac{\eta_{1+i_j}}{\mathrm{LN}(B_{1+i_j})} + \kappa_n \operatorname{logm} \eta_{1+i_j}\right). \tag{3.20}$$

The last term in the summation above is smaller than

$$\kappa_n \left(\operatorname{logm} \frac{\eta_i}{\mathrm{LN}(B_{1+i_{\ell-1}})} + \operatorname{logm} \eta_i.\right) \tag{3.21}$$

We consider this term as the *contribution of* α_u *to* q_{i+1}, $0 \leq i \leq m$. Our aim now is to bound it primarily as a function of $\log |\alpha_u - \beta_u|^{-1}$; the advantage becomes evident when we try to sum the term over all $u \in U$, since then we can use the Davenport-Mahler bound to give an amortized bound on the sum $\sum_{u \in U} \log |\alpha_u - \beta_u|^{-1}$. The remaining terms in the summation above are the contributions of different $\alpha_{u'}$ to q_{i+1}, where $u' \in U - \{u\}$ is such that $\eta_{1+i_j} = M_i^{-1}(\alpha_{u'})$, and can be bounded in terms of $|\alpha_{u'} - \beta_{u'}|$. Note that the contribution of α_u to q_0 is not accounted for, but this will be taken care of later.

We next derive an upper bound on $|\eta_i|$ and a lower bound on $|\mathrm{LN}(B_{1+i_{\ell-1}})|$. In deriving these bounds we will often require lower bounds on $|\alpha - P/Q|$, where α is a root of a degree n polynomial $A(X)$ and P/Q is a fraction such that $0 < |\alpha - P/Q| \leq 1$ and $A(P/Q) \neq 0$. The lower bounds in the literature can be parametrized by some $N \in \mathbb{R}_{\geq 1}$ as follows:

$$|\alpha - P/Q| \geq C(A, N) \cdot Q^{-N}; \tag{3.22}$$

note that the lower bound holds for all conjugates of α. For example, in case of Liouville's inequality [Lio40] we have $N = n$; in case of Roth's theorem [Rot55] we have $N > 2$; for Thue's result [Thu09] we have $N > 1 + n/2$; and for Dyson's result [Dys47] we have $N > \sqrt{2n}$. However, explicit bounds on $C(A, N)$ are known only for Liouville's inequality.

In bounding $|\eta_i|$ and $|\mathrm{LN}(B_{1+i_{\ell-1}})|$, we need to consider two separate cases depending upon whether Q_i is zero or not. The situation $Q_i = 0$ occurs only on the right-most path of the tree T since there are no inverse transformations along this path. In bounding the length of the right-most path we will also account for the contribution of α_u to q_0, for all $u \in U$. The argument for bounding the length of the path is similar to the argument that was used to derive the bound in (3.20) above.

Define the polynomial $B_1(X) := A_{\mathrm{in}}(X)$ and the polynomials $B_i(X)$ as in (3.15). Since we only perform Taylor shifts, the roots of $B_i(X)$ are of the form $\alpha - \delta$, where $\alpha > 0$ is some root of $A_{\mathrm{in}}(X)$ and $\delta \in \mathbb{N}$. We define the sequence of indices (3.19) as was done earlier and follow the same line of argument used to obtain the bound in (3.20), where η_{i_j} now is some positive root of $A_{\mathrm{in}}(X)$. But $|\eta_{i_j}| \leq \mu(A_{\mathrm{in}})$, the largest absolute value amongst all the roots of $A_{\mathrm{in}}(X)$. To get a lower bound on $|\mathrm{LN}(B_{1+i_{\ell-1}})|$, we use the fact that $\mathrm{LN}(B_{1+i_{\ell-1}}) = \eta_{1+i_{\ell-1}} - \delta$, where $\eta_{1+i_{\ell-1}}$ is some positive root of $A_{\mathrm{in}}(X)$ and $\delta \in \mathbb{N}$. Thus from (3.22) we get that $|\mathrm{LN}(B_{1+i_{\ell-1}})| \geq C(A_{\mathrm{in}}, N)$, and hence the length of the right-most path in the recursion tree T is bounded by

$$\sum_{u \in U} \kappa_n (\mathrm{logm}\, \mu(A_{\mathrm{in}}) - \mathrm{logm}\, C(A_{\mathrm{in}}, N));$$

since $|U| \leq n$ this bound is smaller than

$$\kappa_n n (\mathrm{logm}\, \mu(A_{\mathrm{in}}) - \mathrm{logm}\, C(A_{\mathrm{in}}, N)). \tag{3.23}$$

We next derive bounds on $|\eta_i|$ and $|\mathrm{LN}(B_{1+i_{\ell-1}})|$ *assuming* that $Q_i \geq 1$.

An upper bound on $\log \eta_i$**.** Recall from Definition 3.4 that $\eta_i = M_i^{-1}(\alpha_u)$. Thus from (3.8) we get

$$\eta_i = \frac{Q_i + Q_{i-1}}{Q_i} \frac{|s_i - \alpha_u|}{|r_i - \alpha_u|}.$$

But $|s_i - \alpha_u| \leq (Q_i(Q_i + Q_{i-1}))^{-1}$ since $\alpha_u \in J_i$. Thus

$$\eta_i \leq Q_i^{-2} |r_i - \alpha_u|^{-1}.$$

Khinchin [Khi97, Thm. 13,p. 15] has shown that

$$|\alpha_u - r_i| \geq \frac{1}{Q_i(Q_i + Q_{i+1})};$$

the proof is based upon the observation that the fraction $(P_i + P_{i+1})/(Q_i + Q_{i+1})$ lies between the value α_u of the continued fraction and the i-th quotient P_i/Q_i. Using this result from Khinchin

we get

$$\eta_i \leq 2\left(\frac{Q_{i+1}}{Q_i}\right)^2.$$

For $i < m$, we can apply (3.18) and take the logarithm to obtain

$$\log \eta_i \leq 1 - \log|\alpha_u - \beta_u|.$$

For $i = m$ we cannot apply (3.18) because J_{m+1} is an isolating interval, but we will obtain a bound that is asymptotically the same as the bound on η_i for $i < m$. To obtain this result we will show that $Q_{m+1} = O(|\alpha_u - \beta_u|^{-1})$; this will follow if $q_{m+1} \leq |\alpha_u - \beta_u|^{-1}$, because we know from (3.6) that $Q_{m+1} = Q_{m-1} + q_{m+1}Q_m$, and from (3.18) that both Q_m and Q_{m-1} are bounded by $|\alpha_u - \beta_u|^{-1}$. Recall from the construction of T' from the starting of this section that the leaf u is the right child of its parent v; so $q_{m+1} = \delta_v + \delta + 1$, where δ_v is defined as in (3.11) and $\delta = \mathrm{PLB}(A_{M_v}) \in \mathbb{Z}_{\geq 0}$ is the amount of Taylor shift done at v. Since all the roots are real we know that $A_{M_v}(X + \delta)$ has at least two positive roots; so $\mathrm{Var}(A_{M_v}(X + \delta)) \geq 2$ and hence from Theorem 3.2 it follows that the interval with endpoints $M_v(\delta)$ and $M_v(\infty)$ must contain the roots α_u and β_u, which implies that

$$|M_v(\delta) - M_v(\infty)| = (Q_m(Q_{m-1} + Q_m(\delta_v + \delta)))^{-1} \geq |\alpha_u - \beta_u|$$

and hence it follows that

$$q_{m+1} = 1 + \delta_v + \delta \leq Q_{m-1} + Q_m(\delta_v + \delta) \leq |\alpha_u - \beta_u|^{-1}$$

as desired. Thus for $i \leq m$ we have

$$\log \eta_i = O(-\log|\alpha_u - \beta_u|). \tag{3.24}$$

A lower bound on $|\mathbf{LN}(B_{1+i_{\ell-1}})|$.

We may safely assume that $|\mathrm{LN}(B_{1+i_{\ell-1}})| \neq 0$ since if zero is a root of $B_{1+i_{\ell-1}}(X)$ then in the procedure $\mathrm{CF}(A, M)$ we always divide the polynomial by X and remove this degenerate case. We will consider two cases: first, when the root $\mathrm{LN}(B_{1+i_{\ell-1}})$ corresponds to a positive root of $A_i(X)$, and second when it corresponds to a negative root of $A_i(X)$. If γ is the root of $A_{\mathrm{in}}(X)$ that corresponds to $\mathrm{LN}(B_{1+i_{\ell-1}})$ then the first case is equivalent to the condition $\gamma \in J_i$ and the second to $\gamma \notin J_i$. We derive bounds on $|\mathrm{LN}(B_{1+i_{\ell-1}})|$ under these two conditions, starting with the first case.

1. In this case the polynomial $B_{1+i_{\ell-1}}(X) = A_i(X+\delta)$, where δ is defined as

$$\delta = 1 + \sum_{j=1}^{i_{\ell-1}} 1 + \begin{cases} \mathrm{PLB}(B_j) & \text{if } \mathrm{PLB}(B_j) > 1 \\ 0 & \text{otherwise;} \end{cases} \tag{3.25}$$

note that δ as defined is a natural number since $\mathrm{PLB}(B_j)$ is a natural number if it is greater than one. The transformation

$$M'(X) := \frac{P_i X + P_{i-1} + P_i \delta}{Q_i X + Q_{i-1} + Q_i \delta}$$

gives the bijective correspondence between the roots of $A_{\mathrm{in}}(X)$ and the roots of $B_{1+i_{\ell-1}}(X)$. In particular, it follows that

$$\gamma = M'(\mathrm{LN}(B_{1+i_{\ell-1}}))$$

and hence

$$\begin{aligned} |\mathrm{LN}(B_{1+i_{\ell-1}})| &= |M'^{-1}(\gamma)| \\ &= \left| \frac{P_{i-1} + P_i\delta - (Q_{i-1} + Q_i\delta)\gamma}{P_i - Q_i\gamma} \right| \\ &= \frac{\delta Q_i + Q_{i-1}}{|P_i - Q_i\gamma|} \left| \gamma - \frac{\delta P_i + P_{i-1}}{\delta Q_i + Q_{i-1}} \right| \end{aligned}$$

(observe that $\frac{\delta P_i + P_{i-1}}{\delta Q_i + Q_{i-1}} = M'(0)$). From (3.22) we get

$$|\mathrm{LN}(B_{1+i_{\ell-1}})| \geq \frac{C(A_{\mathrm{in}}, N)}{|P_i - Q_i\gamma|} (\delta Q_i + Q_{i-1})^{-(N-1)}.$$

Since $\delta Q_i > Q_{i-1}$ we further get

$$|\mathrm{LN}(B_{1+i_{\ell-1}})| > \frac{C(A_{\mathrm{in}}, N)}{|P_i - Q_i\gamma|} (2\delta Q_i)^{-(N-1)} > C(A_{\mathrm{in}}, N)(2\delta Q_i)^{-(N-1)},$$

where the last step follows from the fact that $|P_i - Q_i\gamma| \leq (Q_i Q_{i-1})^{-1} \leq 1$. But $\delta \leq q_{i+1} < Q_{i+1}$, for $i < m$, and for $i = m$, $\delta \leq \delta_v$, where δ_v is defined as in (3.11); along with (3.18) and (3.12) it follows that $\delta, Q_i < |\alpha_u - \beta_u|^{-1}$. Thus

$$|\mathrm{LN}(B_{1+i_{\ell-1}})| \geq C(A_{\mathrm{in}}, N)(2|\alpha_u - \beta_u|)^{2(N-1)}$$

and hence

$$-\log|\mathrm{LN}(B_{1+i_{\ell-1}})| = O(-N \log|\alpha_u - \beta_u| - \log C(A_{\mathrm{in}}, N)). \tag{3.26}$$

2. If $\mathrm{LN}(B_{1+i_{\ell-1}})$ corresponds with a negative root of $A_i(X)$ then $B_{1+i_{\ell-1}}(X) = A_i(X)$, because for $j \geq i_1$, $\mathrm{LN}(B_j)$ corresponds to a positive root of $A_i(X)$. Thus deriving a

lower bound on $|\mathrm{LN}(B_{1+i_{\ell-1}})|$ amounts to deriving a lower bound on $|\mathrm{LN}(A_i)|$. Also, $\gamma = M_i(\mathrm{LN}(A_i)) \notin J_i$. From (3.8) we obtain

$$|\mathrm{LN}(A_i)| = \frac{Q_i + Q_{i-1}}{Q_i}\frac{|s_i - \gamma|}{|r_i - \gamma|} \geq \frac{1}{Q_i|r_i - \gamma|}C(A_{\mathrm{in}}, N)(Q_i + Q_{i-1})^{1-N},$$

where the last step follows by applying (3.22) to $|s_i - \gamma|$. Since γ is outside J_i and α_u is inside J_i we have

$$|r_i - \gamma| \leq |\gamma - \alpha_u| + |\alpha_u - r_i| \leq |\gamma - \alpha_u| + (Q_iQ_{i-1})^{-1}.$$

Thus

$$\begin{aligned}
|\mathrm{LN}(A_i)| &\geq \frac{1}{Q_i|\gamma - \alpha_u| + Q_{i-1}^{-1}}C(A_{\mathrm{in}}, N)(Q_i + Q_{i-1})^{1-N} \\
&\geq \frac{1}{Q_i(1 + |\gamma - \alpha_u|)}C(A_{\mathrm{in}}, N)(Q_i + Q_{i-1})^{1-N} \\
&\geq \frac{1}{1 + |\gamma - \alpha_u|}C(A_{\mathrm{in}}, N)(Q_iQ_{i-1})^{1-N}.
\end{aligned}$$

Applying the bound from (3.18) we get

$$|\mathrm{LN}(A_i)| \geq \frac{1}{1 + |\gamma - \alpha_u|}C(A_{\mathrm{in}}, N)|\alpha_u - \beta_u|^N.$$

But from the definition of $\mu(A_{\mathrm{in}})$ we know that $|\gamma - \alpha_u| \leq 2\mu(A_{\mathrm{in}})$, and hence we have

$$|\mathrm{LN}(A_i)| \geq \frac{1}{1 + 2\mu(A_{\mathrm{in}})}C(A_{\mathrm{in}}, N)|\alpha_u - \beta_u|^N$$

which gives us

$$-\log|\mathrm{LN}(A_i)| = O(-N\log|\alpha_u - \beta_u| - \log C(A_{\mathrm{in}}, N) + \mathrm{logm}\,\mu(A_{\mathrm{in}})). \tag{3.27}$$

From (3.26) and (3.27) we safely conclude that in general

$$-\log|\mathrm{LN}(B_{1+i_{\ell-1}})| = O(-N\log|\alpha_u - \beta_u| - \log C(A_{\mathrm{in}}, N) + \mathrm{logm}\,\mu(A_{\mathrm{in}})). \tag{3.28}$$

Given the bound on $C(A_{\mathrm{in}}, N)$ in Lemma 3.5 and the fact that $|\alpha_u - \beta_u| < 1$ it follows that the bound in (3.28) dominates the bound on $\log|\eta_i|$ in (3.24) and hence asymptotically the term (3.21) is the same as the bound in (3.28) multiplied by κ_n. Thus the total contribution of α_u to each q_i, $i = 1, \ldots, m+1$, is bounded by

$$\sum_{i=1}^{m+1} \kappa_n O(-N\log|\alpha_u - \beta_u| - \log C(A_{\mathrm{in}}, N) + \mathrm{logm}\,\mu(A_{\mathrm{in}})), \tag{3.29}$$

where m satisfies (3.13); to show the dependency of the number inverse transformations along a path we write m as m_u, where $u \in U$ is the leaf we are interested in. Thus the total number of Taylor shifts along the path starting from the root of the tree T' and terminating at the leaf $u \in U$ is bounded by

$$\sum_{i=1}^{m_u+1} \sum_{u' \in U} \kappa_n O(-N \log |\alpha_{u'} - \beta_{u'}| - \log C(A_{\text{in}}, N) + \operatorname{logm} \mu(A_{\text{in}})),$$

where u' are the leaves that are to the left of u and that share an ancestor with u.

Recall that this bound is valid for all the leaves $u \in U$ except the leaf corresponding to the right-most path in the tree. Thus the summation of the above bound for all $u \in U$, along with the bound in (3.23) on the length of the right-most path of the tree, gives us the following bound on the total number of Taylor shifts in the tree T'

$$\sum_{u \in U} \sum_{i=1}^{m_u+1} \sum_{u' \in U} \kappa_n O(-N \log |\alpha_{u'} - \beta_{u'}| - \log C(A_{\text{in}}, N) + \operatorname{logm} \mu(A_{\text{in}})). \tag{3.30}$$

Combined with the bound in (3.14) on the total number of inverse transformations we get the following bound on the size of the tree T'

$$\begin{aligned} \#(T') = O(n - \sum_{u \in U} \log_\phi |\alpha_u - \beta_u|) \\ + \sum_{u \in U} \sum_{i=1}^{m_u} \sum_{u' \in U} \kappa_n O(-N \log |\alpha_{u'} - \beta_{u'}| - \log C(A_{\text{in}}, N) + \operatorname{logm} \mu(A_{\text{in}})). \end{aligned} \tag{3.31}$$

3.5.3 Worst Case Size of the Tree

In order to derive a worst-case bound on the size of the tree T, from the bound given in (3.31), we need an upper bound on $\sum_{u \in U} \log |\alpha_u - \beta_u|^{-1}$. For this purpose we employ the Davenport-Mahler bound.

Consider the graph G whose edge set is $\{(\alpha_u, \beta_u) | u \in U\}$. Re-order the edge (α_u, β_u) in G such that $|\alpha_u| < |\beta_u|$. Since all the roots are real, it is not hard to see that the in-degree of G is at most one. Thus G satisfies the conditions of Theorem 2.1 and hence

$$-\sum_{u \in U} \log |\alpha_u - \beta_u| = O(B(A_{\text{in}})), \tag{3.32}$$

where

$$B(A_{\text{in}}) := O(n \log \mathrm{M}(A_{\text{in}}) + n \log n - \log \operatorname{discr}(A_{\text{in}})), \tag{3.33}$$

$\mathrm{M}(A_{\mathrm{in}})$ is the Mahler measure of $A_{\mathrm{in}}(X)$ (see (2.2) in §2.2.1), and $\mathrm{discr}(A_{\mathrm{in}})$ is its discriminant. Based upon this bound we have the following:

Theorem 3.5. *Let $A_{in}(X) \in \mathbb{R}[X]$ be a square-free polynomial of degree n which has only real roots. The size of the recursion tree of Akritas' algorithm applied to $A_{in}(X)$ is bounded by*

$$nO(NB(A_{in})^2 - nB(A_{in}) \log C(A_{in}, N) + nB(A_{in}) \log \mu(A_{in})),$$

where $B(A_{in})$ is defined as above, $C(A_{in}, N)$ is the constant involved in the inequality (3.22) and $\mu(A_{in})$ is the largest absolute value amongst all the roots of $A_{in}(X)$.

Proof. Applying the bound in (3.32), along with the observation that $|U| \leq n$, to (3.31) we get that the size of the tree is bounded by

$$O(n + B(A_{\mathrm{in}})) + \sum_{u \in U} \sum_{i=1}^{m_u} \kappa_n O(NB(A_{\mathrm{in}}) - n \log C(A_{\mathrm{in}}, N) + n\mathrm{logm}\, \mu(A_{\mathrm{in}})).$$

Note that the summation term dominates the first term in the bound, so we omit the latter term. From (3.14) we further get that the bound above is smaller than

$$\kappa_n O(NB(A_{\mathrm{in}}) - n \log C(A_{\mathrm{in}}, N) + n\mathrm{logm}\, \mu(A_{\mathrm{in}})) \sum_{u \in U} \frac{1}{2}(1 - \log_\phi |\alpha_u - \beta_u|).$$

Again applying (3.32) we get that the size of the tree is bounded by

$$\kappa_n O(NB(A_{\mathrm{in}})^2 - nB(A_{\mathrm{in}}) \log C(A_{\mathrm{in}}, N) + nB(A_{\mathrm{in}})\mathrm{logm}\, \mu(A_{\mathrm{in}})).$$

From the observation that $\kappa_n = \Theta(n)$ (see the definition in (3.16)), we get the desired result. □

We will next give a specialization of the above theorem for the case of integer polynomials, but for achieving this we need to derive bounds on the quantities N and $C(A, N)$ involved in (3.22).

Lemma 3.5. *Let α be a root of an integer polynomial $A(X)$ of degree n. Suppose $P/Q \in \mathbb{Q}$, $Q > 0$, is such that $0 < |\alpha - P/Q| \leq 1$ and $A(P/Q) \neq 0$, then $|\alpha - P/Q| \geq C(\alpha) \cdot Q^{-n}$ where*

$$C(\alpha) \geq 2^{-n - \log n - (n+1) \log \|A\|_\infty}. \tag{3.34}$$

Proof. From the mean value theorem we know that

$$|A(\alpha) - A(P/Q)| = |A'(\beta)||\alpha - P/Q|,$$

where $\beta = (1-t)\alpha + tP/Q$, $0 \le t \le 1$. But $|A(P/Q)| \ge Q^{-n}$, so

$$|\alpha - P/Q| = \left|\frac{A(P/Q)}{A'(\beta)}\right| \ge |A'(\beta)|^{-1} Q^{-n}.$$

Since $|\alpha - P/Q| \le 1$ we know that $|\beta| \le 1 + |\alpha|$. and hence it can be showed that $|A'(\beta)|$ is smaller than $n\|A\|_\infty(1+|\alpha|)^n$. Using Cauchy's upper bound [Yap00, Cor. 6.8,p. 149] on $|\alpha|$ we get the bound on the constant $C(\alpha)$ mentioned in the lemma. □

We now have the desired specialization of the theorem above.

Corollary 3.3. *Let $A(X)$ be a square-free polynomial of degree n with integer coefficients of magnitude less than 2^L. If $A(X)$ has only real roots then the number of nodes in the recursion tree of Akritas' algorithm run on $A(X)$ is $\widetilde{O}(n^4L^2)$.*

Proof. From Remark 2.4 we know that

$$\mathrm{M}(A) \le \|A\|_2 \le \sqrt{n+1}\|A\|_\infty < \sqrt{n+1}2^L.$$

Moreover, $|\mathrm{discr}(A)| \ge 1$ since $A(X)$ is square-free and its coefficients are integers. From these observations we conclude that $B(A) = \widetilde{O}(nL)$. Furthermore, from Cauchy's bound [Yap00, Cor. 6.8,p. 149] we know that $\mu(A) \le 2^L$. Plugging these bounds along with the bounds in Lemma 3.5 in the theorem above gives us the desired result. □

How good is this bound? The answer depends upon the tightness of the bounds derived on the number of inverse transformations and Taylor shifts. The bound derived on the former, in (3.14), is perhaps the best one can expect, considering that the same bound holds for root isolation using Sturm's method [Dav85, DSY05] and for the Descartes method [ESY06, EMT06]. Thus we may ask what is the best bound on the number of Taylor shifts. The answer depends upon the effectiveness of $\mathrm{PLB}(A)$. Consider the ideal positive lower bound function, i.e., one that returns the floor of the least positive root of $A(X)$ and -1 if none exists. Then the total number of Taylor shifts required along a path is proportional to the number of inverse transformations along the path, because between two consecutive inverse transformations we only perform a constant number of Taylor shifts. Thus the total number of Taylor shifts in the recursion tree is proportional to the number of inverse transformations in the tree and hence the best possible bound on the size of the recursion tree is $\widetilde{O}(nL)$. This shows that there is a huge gap to be overcome in the bound derived in Corollary 3.3.

In retrospect, one may ask why the bound on the number of inverse transformations is so good compared to the bound on the Taylor shifts. The answer lies in the relation of the width of the interval associated with a node and the width of the intervals associated with its left and its right child, which are respectively obtained by an inverse transformation and Taylor shift. Suppose the continued fraction associated with an internal node v in the recursion tree T is $[q_0, \ldots, q_i] = P_i/Q_i$, where we assume $Q_i > 0$; thus the interval associated with the node v has endpoints P_i/Q_i and P_{i-1}/Q_{i-1}, and hence its width is $(Q_iQ_{i-1})^{-1}$. Now the continued fraction associated with the left child of v is $[q_0, \ldots, q_i, 1] = P_{i+1}/Q_{i+1}$ that has width $(Q_iQ_{i+1})^{-1}$. The decrease in the width of the interval in going from the parent v to its left child is the ratio $Q_{i-1}/Q_{i+1} = 1/(1 + Q_i/Q_{i-1}) \leq \frac{1}{2}$ since $Q_i \geq Q_{i-1}$. This means that whenever we take a left path in the tree T the width of the interval decreases by at least half and hence the number of inverse transformations is bounded by the logarithm of the separation bound as stated in (3.13). The same relation does not always hold between a node and its right child. Suppose the right child w of the node v is obtained by a Taylor shift by an amount $\delta_1 > 0$. Then the interval associated with w has endpoints P_i/Q_i and $(P_{i-1} + P_i\delta_1)/(Q_{i-1} + Q_i\delta_1)$, and hence has width $(Q_i(Q_{i-1} + \delta_1 Q_i))^{-1}$. Again we can show, assuming that $Q_i, \delta > 0$, that the width of the interval associated with w is at most half the width of the interval associated with v. However, this relation does not necessarily hold between w and its right child. More precisely, let the right child of w be obtained by performing a Taylor shift by an amount $\delta_2 > 0$. Then the width of the interval associated with the right child of w is $\frac{1}{Q_i(Q_{i-1}+Q_i(\delta_1+\delta_2))}$. Thus the decrease in the width of the interval in going from w to its right child is

$$\frac{Q_{i-1} + \delta_1 Q_i}{Q_{i-1} + (\delta_1 + \delta_2)Q_i} = \left(1 + \frac{\delta_2 Q_i}{Q_{i-1} + \delta_1 Q_i}\right)^{-1} \leq \left(1 + \frac{\delta_2}{1 + \delta_1}\right)^{-1}$$

which is a fraction that is less than half if $\delta_2 > \delta_1 + 1$; however, this condition is less likely to hold since δ_2 is a lower bound on the absolute value of the root $\alpha - \delta_1$, where δ_1 was a lower bound on the root α. This implies that along a path where we only take the right branch at each node the width of the interval decreases, but not necessarily by half. This difference between the effect of inverse transformations and Taylor shifts on the width of the interval gives some insight to the difference between the bounds achieved above.

Following the same line of argument as in this section, we next derive a bound on the size of the recursion tree for isolating the real roots of a general polynomial.

3.6 The Size of the Recursion Tree: The General Case

In this section we bound the worst-case size of the recursion tree T of the procedure described in §3.2 initiated with a square-free polynomial $A_{\text{in}}(X) \in \mathbb{R}[X]$ of degree n and the Möbius transformation X, without assuming that all the roots of $A(X)$ are real.

We partition the leaves of T into two types: **type-0** leaves are those that declare the absence of a real root and **type-1** leaves are those that declare the presence of a real root.

Consider the tree T' obtained by pruning certain leaves from T: prune all the type-0 leaves that have either a non-leaf or a type-1 leaf as sibling; if two leaves are siblings of each other then arbitrarily prune one of them. Again, $\#(T') < \#(T) \leq 2\#(T')$. Thus we bound $\#(T')$.

Let U be the set of leaves of T'. Consider a leaf $u \in U$ and let v be its parent. As was supposed in §3.5, let $[q_0, \ldots, q_{m+1}] = P_{m+1}/Q_{m+1}$ be the continued fraction associated with the leaf u. Define the Möbius transformation $M_u(X)$ and $M_v(X)$ as in (3.10) and (3.11); moreover, let $I_u := I_{M_u}$ and $I_v := I_{M_v}$. Since v is not a leaf we know that $\texttt{Var}(A_{M_v}) > 1$. To each leaf $u \in U$ we assign a unique pair (α_u, β_u) of roots of $A_{\text{in}}(X)$ as follows:

1. If u is a type-1 leaf then there is a unique root $\alpha_u \in I_u$. Since $\texttt{Var}(A_{M_v}) > 1$, from Theorem 3.2 we know that there must be a root in $\overline{C}_{I_v} \cup \underline{C}_{I_v}$ apart from α_u; let β_u be one such root. Thus with each type-1 leaf we can associate a pair (α_u, β_u). Moreover, this can be done in a unique manner. Suppose u' is another type-1 leaf and v' is its parent then $\alpha_u \neq \alpha_{u'}$. From Lemma 2.1 it is clear that we only need to consider the case when I_v and $I_{v'}$ are adjacent to each other. Moreover, assume β_u and $\overline{\beta_u}$ are the only non-real roots in $\overline{C}_{I_v} \cup \underline{C}_{I_v}$ and $\overline{C}_{I_{v'}} \cup \underline{C}_{I_{v'}}$. Then it must be that either $\beta_u \in \overline{C}_{I_v} \cap \overline{C}_{I_{v'}}$ or $\beta_u \in \underline{C}_{I_v} \cap \underline{C}_{I_{v'}}$. In either case we can choose $\beta_{u'} = \overline{\beta_u}$ distinct from β_u.

2. If u is a type-0 leaf then we know that it had a type-0 leaf as its sibling in T. We consider two sub-cases:

   - If I_v does not contain a real root then we know from Theorem 3.3 that there must be a pair of complex conjugate roots in C_{I_v}. Let (α_u, β_u), $\beta_u := \overline{\alpha_u}$, be one such pair. The uniqueness of the pair is immediate since C_{I_v} does not overlap with $C_{I_{v'}}$ for the parent v' of any other type-0 leaf.
   - If I_v does contain a root then it must be the midpoint of the interval I_v; let α_u denote this root. From Theorem 3.2 we also know that there must be a pair of complex

conjugates $(\beta, \overline{\beta})$ in $\overline{C}_{I_v} \cup \underline{C}_{I_v}$; choose $\beta_u := \beta$. The pair is unique because α_u is unique.

Note the similarity of the above assignment with Section 2.2.2. As was done there, we will try to bound the size of the path terminating at the leaf u in terms of the distance $|\alpha_u - \beta_u|$. Before we proceed further, we have two observations: first, because of the uniqueness of the pair (α_u, β_u) it follows that the size of the set $|U| \leq n$; and second, since $\alpha_u, \beta_u \in \overline{C}_{I_v} \cap \underline{C}_{I_v}$ we know that

$$(Q_m(Q_{m-1} + \delta_v Q_m))^{-1} > \frac{\sqrt{3}}{2}|\alpha_u - \beta_u| > \frac{1}{2}|\alpha_u - \beta_u|, \tag{3.35}$$

where δ_v is defined as in (3.11). We again start with a bound on the transformations of the form $1/(X+1)$ along any path in T'.

3.6.1 Bounding the Inverse Transformations

From (3.35) it follows that

$$|\alpha_u - \beta_u| < 2(Q_m Q_{m-1})^{-1}.$$

Recall that $Q_i \geq F_{i+1}$, the $(i+1)$-th Fibonacci number, and $F_{i+1} \geq \phi^i$, where $\phi = (\sqrt{5}+1)/2$. Thus

$$\phi^{2m-1} \leq 2|\alpha_u - \beta_u|^{-1}$$

and hence

$$m \leq \frac{1}{2}(1 + \log_\phi 2 - \log_\phi |\alpha_u - \beta_u|). \tag{3.36}$$

So the total number of inverse transformations in T' are bounded by

$$\sum_{u \in U} \frac{1}{2}(1 + \log_\phi 2 - \log_\phi |\alpha_u - \beta_u|) \leq 2n + \sum_{u \in U} \log_\phi (|\alpha_u - \beta_u|)^{-1}. \tag{3.37}$$

3.6.2 Bounding the Taylor Shifts

In §3.5, the key components used to bound the number of Taylor shifts were Lemma 3.3 and Lemma 3.4. We derive similar results in the general case. Before we do so, we have the following observation on the effect of shifts in the complex plane:

Lemma 3.6. *If $\alpha, \beta \in \mathbb{C}$ are such that $|\alpha| \leq |\beta|$ and $|\alpha - \delta| \geq |\beta - \delta|$, for any positive real number δ, then $\Re(\beta) \geq \Re(\alpha)$.*

Proof. $|\alpha - \delta| \geq |\beta - \delta|$ implies

$$2\delta(\Re(\beta) - \Re(\alpha)) \geq |\beta|^2 - |\alpha|^2 \geq 0.$$

Since δ is positive we have our result. □

Intuitively, this lemma says if the origin is shifted to the right then only the complex numbers to the right of the number α and in absolute value greater than α can possibly become smaller than α in absolute value.

Let $B(X) \in \mathbb{R}[X]$ be a polynomial all of whose roots are in the open half plane $\Re(z) > 0$. Then we having the following analogue to Lemma 3.3:

Lemma 3.7. *Let $B_1(X) := B(X)$, and for $i > 1$ recursively define*

$$B_i(X) := B_{i-1}(X + \delta_{i-1})$$

where

$$\delta_{i-1} := \begin{cases} \mathrm{PLB}(B_{i-1}) + 1 & \text{if } \mathrm{PLB}(B_{i-1}) > 1 \\ 1 & \text{otherwise.} \end{cases}$$

Let α_1 denote a root of $B_1(X)$ with the smallest absolute value, and recursively let $\alpha_i = \alpha_{i-1} - \delta_{i-1}$. Then $\Re(\alpha_i) \leq 1$ if $i \geq 2 + 8n + \gamma_n \mathrm{logm}\, \Re(\alpha_1)$.

Proof. Let $b_i := \mathrm{PLB}(B_i)$, and β_i be the root of $B_i(X)$ with the smallest absolute value. The difficulty in this case, as compared to Lemma 3.3, is that β_i may not be the same as α_i, except initially. But there is still some relation between the two, namely $\Re(\alpha_i) \leq \Re(\beta_i)$, for $i \geq 1$. The proof is by induction; the base case holds by the definition of α_1.

Suppose inductively $\Re(\alpha_{i-1}) \leq \Re(\beta_{i-1})$. Let β be the root of $B_{i-1}(X)$ such that $\beta_i = \beta - \delta_{i-1}$. Then we know by the definition of β_{i-1} that $|\beta_{i-1}| \leq |\beta|$. However, we also have

$$|\beta_i| = |\beta - \delta_{i-1}| \leq |\beta_{i-1} - \delta_{i-1}|.$$

Thus from Lemma 3.6 we know that $\Re(\beta) \geq \Re(\beta_{i-1})$ and hence

$$\Re(\beta_i) = \Re(\beta) - \delta_{i-1} \geq \Re(\beta_{i-1}) - \delta_{i-1} \geq \Re(\alpha_{i-1}) - \delta_{i-1} = \Re(\alpha_i).$$

Since β_i is the root of $B_i(X)$ with the smallest absolute value, from (3.4) we know that $\frac{|\beta_i|}{8n} < b_i < |\beta_i|$. Moreover, because $\Re(\beta_i) \geq \Re(\alpha_i)$ we have $b_i > \Re(\alpha_i)/8n$. Let j be the index such that $\Re(\alpha_i) > 8n$ for $i < j$. Then for $i < j$ we know that $b_i > 1$. Thus $\Re(\alpha_i) =$

$\Re(\alpha_{i-1}) - b_{i-1} - 1 < \Re(\alpha_{i-1})(1 - \frac{1}{8n})$ and recursively $\Re(\alpha_i) < \Re(\alpha)(1 - \frac{1}{8n})^{i-1}$. So $\Re(\alpha_j) \leq 8n$ if $\Re(\alpha_{j-1}) \leq 8n$ or if

$$j \geq 2 + \gamma_n \operatorname{logm} \Re(\alpha_1).$$

For $i \geq j$ we know that $\Re(\alpha_i) \leq 8n$, because $\Re(\alpha_i)$ is monotonically decreasing. Thus if $i > j$ is such that $i - j \geq 8n$ then $\Re(\alpha_i) \leq 1$. Combining this lower bound on $i - j$ with the lower bound on j we get the result of the lemma. □

We can derive an analogous result to Lemma 3.4, but now we assume that the roots with the smallest absolute value can have negative real parts and $B(0) \neq 0$; note that this cannot be the case in the previous lemma because of Lemma 3.6. To derive the result we extend the definition of $\mathrm{LP}(B)$ and $\mathrm{LN}(B)$ in §3.5 as follows:

Definition 3.6. Let $\mathrm{LP}(B)$ denote the root of $B(X)$ in $\Re(z) > 0$ that has the smallest real part and the smallest absolute value, and $\mathrm{LN}(B)$ denote the root of $B(X)$ in $\Re(z) \leq 0$ that has the largest real part and the smallest absolute value.

Lemma 3.8. *Let $B_1(X) := B(X)$, and recursively define δ_i, and $B_i(X)$ as in the above lemma. Let $\alpha_1 := LP(B_1)$, $\beta_1 := LN(B_1)$ and recursively define $\alpha_i = \alpha_{i-1} - \delta_{i-1}$ and $\beta_i = \beta_{i-1} - \delta_{i-1}$. If*

$$i = \Omega\left(n + \kappa_n \operatorname{logm} \frac{|\alpha_1|}{|\beta_1|} + \kappa_n \operatorname{logm} |\alpha_1|\right)$$

then $\Re(\alpha_i) \leq 1$.

Proof. Let $b_i := \mathrm{PLB}(B_i)$. We assume that $|\beta_1| < |\alpha_1|$, otherwise the bound in the lemma trivially follows from the previous lemma. Let γ_i, denote the root of $B_i(X)$ with the smallest absolute value; by definition and our assumption that $|\beta_1| < |\alpha_1|$ we initially have $\gamma_1 = \beta_1$. Let j be the first index i such that $\gamma_i \neq \beta_i$. Then for $i > j$, $\Re(\gamma_i) \geq \Re(\alpha_i)$; this follows from the fact that $\alpha_1 = \mathrm{LP}(B_1)$ and from Lemma 3.6. Thus if $i > j$ is such that

$$i - j > 1 + 8n + \kappa_n \operatorname{logm} |\alpha_1| > 1 + 8n + \gamma_n \operatorname{logm} |\alpha_i|$$

then from Lemma 3.7 we are sure that $\Re(\alpha_i) \leq 1$. But if we choose j such that $|\beta_i| > |\alpha_i|$, for $i > j$, then $\gamma_i \neq \beta_i$ for $i > j$, because all the roots with negative real parts are to the left of β_i and in absolute value greater than $|\beta_i|$. We next give a lower bound on j.

For $i < j$ we have from (3.4)

$$b_i > \frac{|\beta_i|}{8n}. \tag{3.38}$$

Assume that $b_i > 1$, then we know that $\delta_i = b_i + 1$. Since $\beta_{i+1} = \beta_i - \delta_i$ it follows that $\Im(\beta_{i+1}) = \Im(\beta_i)$ and hence

$$\begin{aligned} |\beta_{i+1}| = |\beta_{i-1} - (b_i + 1)| = &= (((|\Re(\beta_{i-1})| + b_i + 1)^2 + \Im(\beta_{i-1})^2)^{\frac{1}{2}} \\ &> (|\Re(\beta_{i-1})|^2 + b_i^2 + 1 + \Im(\beta_{i-1})^2)^{\frac{1}{2}} \\ &= (|\beta_{i-1}|^2 + b_i^2 + 1)^{\frac{1}{2}}. \end{aligned}$$

Applying the bound from (3.38) we get

$$|\beta_{i+1}| > (|\beta_{i-1}|^2(1 + (8n)^{-2}) + 1)^{\frac{1}{2}} > |\beta_{i-1}|(1 + \frac{1}{8n}),$$

because $2 < 8n$ for $n \geq 1$, which is trivially true. Thus recursively we know that $|\beta_{i+1}| > |\beta_1|(1 + 1/8n)^i$. Hence if

$$j > 1 + 8n + \log \frac{|\alpha_1|}{|\beta_1|} \tag{3.39}$$

then $|\beta_i| > |\alpha_1| \geq |\alpha_i|$, for $i > j$. From (3.38) it is clear that we need $8n$ shifts initially to ensure $b_i > 1$. These additional shifts, along with (3.39) and the bound on $i - j$ above give us the desired lower bound on i which ensures that $\Re(\alpha_i) \leq 1$. □

Based upon the above two lemmas we will bound the number of Taylor shifts from the root of T' to the leaf u, with the associated continued fraction $[q_0, \ldots, q_{m+1}]$, by bounding the number of Taylor shifts that compose each q_i, $i = 0, \ldots, m+1$. Recall from the beginning of this section the definitions of the two Möbius transformation $M_u(X)$ and $M_v(X)$, the intervals I_u and I_v, and the pair (α_u, β_u) for a leaf $u \in U$. Following §3.5, we define the following:

Definition 3.7. For $0 \leq i \leq m + 1$ let

- $M_i(X) := [q_0, \ldots, q_i, X] = \frac{P_i X + P_{i-1} + P_i}{Q_i X + Q_{i-1} + Q_i}$,
- $A_i(X) := (Q_i X + Q_{i-1} + Q_i)^n A(M_i(X))$,
- $\eta_i := M_i^{-1}(\alpha_u)$,
- $r_i = P_i/Q_i$, $s_i := \frac{P_i + P_{i-1}}{Q_i + Q_{i-1}}$ and
- $J_i := I_{M_i}$.

By its definition J_i, for $0 \leq i \leq m$, contains I_u and hence it follows from (3.35) that for $0 \leq i \leq m$

$$(Q_i Q_{i-1})^{-1} \geq \frac{|\alpha_u - \beta_u|}{2}. \tag{3.40}$$

Based upon the above lemmas we can bound the total number of Taylor shifts needed to obtain q_{i+1}. Let $B_1(X) := A_i(X)$, and recursively define the polynomials $B_i(X)$ as in Lemma 3.7. Define the sequence of indices

$$1 = i_0 \leq i_1 < i_2 < \cdots < i_\ell, \tag{3.41}$$

where the index i_j is such that $\Re(\mathrm{LP}(B_{i_j}))$ is contained in the unit interval; if $i < m$ the last index i_ℓ is such that the real part of the root in $B_{i_\ell}(X)$ corresponding to η_i is in the unit interval; for $i = m$ the index i_ℓ is such that the node that has $B_{i_\ell}(X)$ as the corresponding polynomial is the parent v of the leaf u. Clearly, $\ell \leq n$.

From Lemma 3.8 we know that

$$i_{j+1} - i_j = O\left(n + \kappa_n \mathrm{logm}\frac{|\mathrm{LP}(B_{1+i_j})|}{|\mathrm{LN}(B_{1+i_j})|} + \kappa_n \mathrm{logm}\,|\mathrm{LP}(B_{1+i_j})|\right).$$

Summing this inequality for $j = 0, \ldots, \ell - 1 < n$ we get that if

$$i_\ell = O(n^2) + O\left(\sum_{j=0}^{\ell-1} \kappa_n \mathrm{logm}\frac{|\mathrm{LP}(B_{1+i_j})|}{|\mathrm{LN}(B_{1+i_j})|} + \kappa_n \mathrm{logm}\,|\mathrm{LP}(B_{1+i_j})|\right) \tag{3.42}$$

then the real part of the root in $B_{i_\ell}(X)$ that corresponds to η_i is in the unit interval, i.e., the number of Taylor shifts which constitute q_{i+1} are bounded by this bound.

The last term in the summation above is smaller than

$$\kappa_n \left(\mathrm{logm}\frac{|\eta_i|}{|\mathrm{LN}(B_{1+i_{\ell-1}})|} + \mathrm{logm}\,|\eta_i|\right), \tag{3.43}$$

because $|\mathrm{LP}(B_{1+i_{\ell-1}})|$ is smaller than $|\eta_i|$. We call this term the *contribution of α_u to q_{i+1}*. Again, we will derive an upper bound on this term by deriving an upper bound on $|\eta_i|$ and a lower bound on $|\mathrm{LN}(B_{1+i_{\ell-1}})|$ mainly in terms of $\log|\alpha_u - \beta_u|^{-1}$. Following §3.5 we separately consider the two cases $Q_i = 0$ and $Q_i > 0$.

The situation $Q_i = 0$ only occurs on the right-most path of the tree since there are no inverse transformations along this path. The argument for bounding the length of this path is the similar to the one used to derive the bound in (3.42) above. Let $B_1(X) := A_{\mathrm{in}}(X)$ and recursively define $B_i(X)$ as in Lemma 3.7. Define the sequence of indices as in (3.41) and follow the same line of argument used to obtain (3.42), except now we can replace $|\mathrm{LP}(B_{1+i_j})|$ by $|\eta_{i_j}|$, the absolute value of some root of $A_{\mathrm{in}}(X)$ in $\Re(z) > 0$. Moreover, we know that $|\eta_{i_j}| \leq \mu(A_{\mathrm{in}})$. To obtain a lower bound on $|\mathrm{LN}(B_{1+i_j})|$ we observe that $\mathrm{LN}(B_{1+i_j}) = \alpha - \delta$, where α is some root of $A_{\mathrm{in}}(X)$ and $\delta \in \mathbb{N}$ is such that $B_{1+i_j}(X) = A_{\mathrm{in}}(X + \delta)$, and hence from (3.22) we get

$|\mathrm{LN}(B_{1+i_j})| \geq C(A_{\mathrm{in}}, N)$. Thus the length of the right-most path in the tree T' is bounded by the same bound as in (3.23), namely

$$\kappa_n n(\mathrm{logm}\,\mu(A_{\mathrm{in}}) - \mathrm{logm}\,C(A_{\mathrm{in}}, N)). \tag{3.44}$$

We next derive an upper bound on $|\eta_i|$ and a lower bound on $|\mathrm{LN}(B_{1+i_{\ell-1}})|$ *assuming* that $Q_i \geq 1$.

Upper bound on $|\eta_i|$. Recall from Definition 3.7 that $\eta_i = M_i^{-1}(\alpha_u)$. Thus from (3.8) we get

$$\eta_i = \frac{Q_i + Q_{i-1}}{Q_i} \frac{|s_i - \alpha_u|}{|r_i - \alpha_u|} \leq \frac{Q_i + Q_{i-1}}{Q_i} \frac{|s_i - \alpha_u|}{C(A_{\mathrm{in}}, N)} Q_i^N,$$

where the second inequality follows from (3.22). But

$$|s_i - \alpha_u| \leq \sqrt{3}|J_i| < 2(Q_i Q_{i-1})^{-1}.$$

Thus

$$\eta_i < 2C(A_{\mathrm{in}}, N)^{-1} Q_i^N.$$

Moreover, from (3.40) we know that $Q_i \leq 2|\alpha_u - \beta_u|^{-1}$. Plugging this bound on Q_i into the bound on η_i we obtain

$$\eta_i < 2^{N+1}|\alpha_u - \beta_u|^{-N} C(A_{\mathrm{in}}, N)^{-1}.$$

Taking logarithm on both sides we get

$$\log \eta_i \leq -N \log|\alpha_u - \beta_u| - \log C(A_{\mathrm{in}}, N) + N + 1. \tag{3.45}$$

Remark 3.3. *Notice that this bound is not as good as the bound in (3.24) which was obtained under the assumption that all the roots of the polynomial were real. The reason is that now we cannot show $Q_{m+1} = O(-\log|\alpha_u - \beta_u|)$ and use the result from Khinchin for $i = m$, though for $i < m$ this bound still holds. To see why we cannot obtain the desired bound on Q_{m+1} we recall from the argument for deriving the bound (3.24) that we had showed that $\delta := \mathrm{PLB}(A_{M_v}) = O(-\log|\alpha_u - \beta_u|)$, where $A_{M_v}(X)$ is the polynomial associated with the parent v of u; this was possible because $\mathrm{Var}(A_{M_v}(X + \delta)) \geq 2$, but now a situation like that shown in Figure 3.2 may occur, namely the polynomial $A_{M_v}(X)$ has a pair $(\alpha, \overline{\alpha})$ of complex roots that have positive real part and also the smallest absolute value and a real root $\beta > \Re(\alpha)$; in this case δ is a lower bound on $|\alpha|$, but it is greater than $\Re(\alpha)$, so when we shift the origin to the right by δ the resulting polynomial $A_{M_v}(X + \delta)$ contains at most one sign variation and hence the two circle figure w.r.t. the interval with endpoints $M_v(\delta)$ and $M_v(0)$ does not contain two roots, which was necessary to obtain (3.24).*

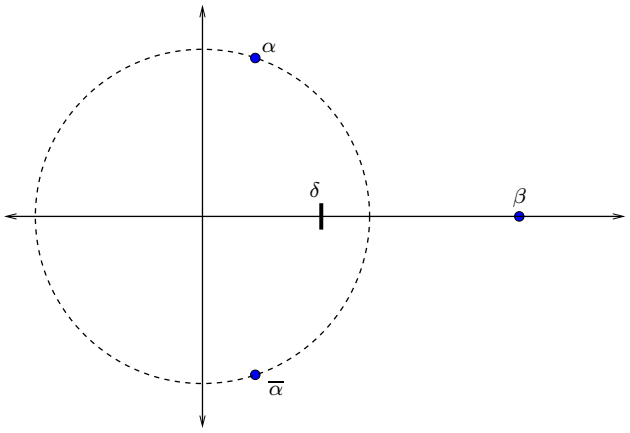

Figure 3.2: The roots of the polynomial $A_{M_v}(X)$ in $\mathbb{C}$.

A lower bound on $|\mathbf{LN}(B_{1+i_{\ell-1}})|$.

The approach is the same as the one used to obtain (3.28), except now we use (3.40) instead of (3.18); we may again safely assume that $\mathrm{LN}(B_{1+i_{\ell-1}}) \neq 0$. We derive lower bounds for two different cases: first, when the root $\mathrm{LN}(B_{1+i_{\ell-1}})$ corresponds to a root of $A_i(X)$ in $\Re(z) > 0$, and second when $\mathrm{LN}(B_{1+i_{\ell-1}})$ corresponds to a root of $A_i(X)$ in $\Re(z) \leq 0$. Let γ be the root of $A_{\mathrm{in}}(X)$ that corresponds to $\mathrm{LN}(B_{1+i_{\ell-1}})$. Then the first case is equivalent to saying that $\gamma \in C_{J_i}$ and the second to the condition that $\gamma \notin C_{J_i}$. We will derive bounds on $|\mathrm{LN}(B_{1+i_{\ell-1}})|$ under these two conditions, starting with the first.

1. Suppose the polynomial $B_{1+i_{\ell-1}}(X) = A_i(X+\delta)$, where δ is defined as in (3.25). Then the transformation

$$M'(X) := \frac{P_i X + P_{i-1} + P_i \delta}{Q_i X + Q_{i-1} + Q_i \delta}$$

gives the bijective correspondence between the roots of $A_{\mathrm{in}}(X)$ and the roots of $B_{1+i_{\ell-1}}(X)$. In particular, $\gamma = M'(\mathrm{LN}(B_{1+i_{\ell-1}}))$. Then

$$\begin{aligned} |\mathrm{LN}(B_{1+i_{\ell-1}})| &= |M'^{-1}(\gamma)| \\ &= \left| \frac{P_{i-1} + P_i\delta - (Q_{i-1} + Q_i\delta)\gamma}{P_i - Q_i\gamma} \right| \\ &= \frac{\delta Q_i + Q_{i-1}}{|P_i - Q_i\gamma|} \left| \gamma - \frac{\delta P_i + P_{i-1}}{\delta Q_i + Q_{i-1}} \right| \end{aligned}$$

(observe that $\frac{\delta P_i + P_{i-1}}{\delta Q_i + Q_{i-1}} = M'(0)$). From (3.22) we get

$$|\mathrm{LN}(B_{1+i_{\ell-1}})| \geq \frac{C(A_{\mathrm{in}}, N)}{|P_i - Q_i\gamma|}(\delta Q_i + Q_{i-1})^{-(N-1)}.$$

Since $\delta Q_i \geq Q_{i-1}$ we further get

$$|\mathrm{LN}(B_{1+i_{\ell-1}})| \geq \frac{C(A_{\mathrm{in}}, N)}{|P_i - Q_i\gamma|}(2\delta Q_i)^{-(N-1)} \geq C(A_{\mathrm{in}}, N)2^{-N}(\delta Q_i)^{-(N-1)},$$

where the last step follows from the fact that since $\gamma \in C_{J_i}$, $|P_i - Q_i\gamma| \leq (Q_i Q_{i-1})^{-1} \leq 1$. But $\delta \leq q_{i+1} < Q_{i+1}$, for $i < m$, and for $i = m$, $\delta \leq \delta_v$, where δ_v is defined as in (3.11); along with (3.40) and (3.35) it follows that $\delta, Q_i < 2|\alpha_u - \beta_u|^{-1}$. Thus

$$-\log|\mathrm{LN}(B_{1+i_{\ell-1}})| = O(-N\log|\alpha_u - \beta_u| - \log C(A_{\mathrm{in}}, N)). \tag{3.46}$$

2. If $\mathrm{LN}(B_{1+i_{\ell-1}})$ corresponds to a negative root of $A_i(X)$ then from Lemma 3.6 we know that $\mathrm{LN}(B_j)$, $j = 1, \ldots, 1 + i_{\ell-1}$, correspond to the same negative root of $A_i(X)$. Thus we derive a lower bound on $|\mathrm{LN}(B_1)| = |\mathrm{LN}(A_i)|$. From (3.8) we know that

$$|\mathrm{LN}(A_i)| = \frac{Q_i + Q_{i-1}}{Q_i}\frac{|s_i - \gamma|}{|r_i - \gamma|} \geq \frac{1}{Q_i|r_i - \gamma|}C(A_{\mathrm{in}}, N)(Q_i + Q_{i-1})^{1-N},$$

where the last step follows by applying (3.22) to $|s_i - \gamma|$. Since γ is outside C_{J_i} and $\alpha_u \in \overline{C}_{J_i} \cup \underline{C}_{J_i}$, we have

$$|r_i - \gamma| \leq |\gamma - \alpha_u| + |\alpha_u - r_i| \leq |\gamma - \alpha_u| + 2(Q_i Q_{i-1})^{-1}.$$

Thus

$$\begin{aligned}
|\mathrm{LN}(A_i)| &\geq \frac{Q_{i-1}}{Q_i|\gamma - \alpha_u| + 2Q_{i-1}^{-1}}C(A_{\mathrm{in}}, N)Q_{i-1}^{-N} \\
&\geq \frac{1}{Q_i(2 + |\gamma - \alpha_u|)}C(A_{\mathrm{in}}, N)Q_{i-1}^{-N} \\
&\geq \frac{1}{2 + |\gamma - \alpha_u|}C(A_{\mathrm{in}}, N)(Q_i Q_{i-1})^{-N}.
\end{aligned}$$

From (3.35), and the fact that $I_v \subseteq J_i$, we know that $Q_i Q_{i-1}|\alpha_u - \beta_u| \leq 2$. Thus

$$|\mathrm{LN}(A_i)| \geq \frac{1}{1 + |\gamma - \alpha_u|}C(A_{\mathrm{in}}, N)(2|\alpha_u - \beta_u|)^N.$$

But from the definition of $\mu(A_{\mathrm{in}})$ we know that $|\gamma - \alpha_u| \leq 2\mu(A_{\mathrm{in}})$, and hence we have

$$|\mathrm{LN}(A_i)| \geq \frac{1}{2 + 2\mu(A_{\mathrm{in}})}C(A_{\mathrm{in}}, N)(2|\alpha_u - \beta_u|)^N$$

which gives us

$$-\log|\mathrm{LN}(A_i)| = O(-N\log|\alpha_u - \beta_u| - \log C(A_{\mathrm{in}}, N) + \mathrm{logm}\,\mu(A_{\mathrm{in}})). \tag{3.47}$$

From (3.46) and (3.47) we may safely conclude that

$$-\log|\mathrm{LN}(B_{1+i_{\ell-1}})| = O(-N\log|\alpha_u-\beta_u| - \log C(A_{\mathrm{in}},N) + \log\mu(A_{\mathrm{in}})). \tag{3.48}$$

Moreover, since $|\alpha_u-\beta_u|, C(A_{\mathrm{in}},N) < 1$ this bound dominates the bound on $\log|\eta_i|$ in (3.45), and hence the term in (3.43) is bounded by

$$\kappa_n O(-N\log|\alpha_u-\beta_u| - \log C(A_{\mathrm{in}},N) + \log\mu(A_{\mathrm{in}})).$$

Thus the total contribution of α_u to each q_i, $i = 1,\ldots,m+1$, is bounded by the sum of this bound from $i = 1,\ldots,m+1$, i.e., by

$$\sum_{i=1}^{m} \kappa_n O(-N\log|\alpha_u-\beta_u| - \log C(A_{\mathrm{in}},N) + \log\mu(A_{\mathrm{in}})),$$

where m satisfies (3.36); to show the dependency of m on the choice of the leaf u, from now on we write m as m_u. Thus the total number of Taylor shifts along the path starting from the root of the tree T' and terminating at the leaf $u \in U$, except the leaf of the right-most path, is bounded by

$$\sum_{i=1}^{m_u}\sum_{u'\in U} \kappa_n O(-N\log|\alpha_{u'}-\beta_{u'}| - \log C(A_{\mathrm{in}},N) + \log\mu(A_{\mathrm{in}})),$$

where u' are the leaves to the left of u and that share a common ancestor with u. The total number of Taylor shifts in the tree T' is obtained by summing the above bound for all $u \in U$ and adding to it the bound in (3.44) on the length of the right-most path:

$$\sum_{u\in U}\sum_{i=1}^{m_u}\sum_{u'\in U} \kappa_n O(-N\log|\alpha_{u'}-\beta_{u'}| - \log C(A_{\mathrm{in}},N) + \log\mu(A_{\mathrm{in}})). \tag{3.49}$$

Combined with the bound in (3.37) on the total number of inverse transformations in the tree T', we get the following bound on the size of the tree T'

$$\begin{aligned}\#(T') = O(n &+ \sum_{u\in U}\log_\phi|\alpha_u-\beta_u|^{-1}) \\ &+ \sum_{u\in U}\sum_{i=1}^{m_u}\sum_{u'\in U} \kappa_n O(-N\log|\alpha_{u'}-\beta_{u'}| - \log C(A_{\mathrm{in}},N) + \log\mu(A_{\mathrm{in}})).\end{aligned} \tag{3.50}$$

3.6.3 Worst Case Size of the Tree

In order to derive a worst-case bound on the size of the tree T, from the bound given in (3.50), we need to derive an upper bound on $\sum_{u\in U} -\log|\alpha_u-\beta_u|$. For this purpose we resort to the Davenport-Mahler bound.

Consider the graph G whose edge set is $E_1 \cup E_0$, where $E_0 := \{(\alpha_u, \beta_u)\}$, u is a type-0 leaf and $E_1 := \{(\alpha_u, \beta_u)\}$, u is a type-1 leaf. We want to show that G satisfies the conditions of Theorem 2.1. First of all, for any $u \in U$ we can reorder the pair (α_u, β_u) to ensure that $|\alpha_u| \leq |\beta_u|$ without affecting the summation $\sum_{u \in U} -\log|\alpha_u - \beta_u|$. We note that the graph so obtained is similar to the graph described in §2.2.2; thus after properly reordering the edges as was mentioned there we may directly apply Theorem 2.1 to G to get

$$\sum_{u \in U} -\log|\alpha_u - \beta_u| = O(B(A_{\text{in}})), \tag{3.51}$$

where $B(A_{\text{in}})$ is as defined in (3.33). Based upon this bound we have the following:

Theorem 3.8. *Let $A_{in}(X) \in \mathbb{R}[X]$ be a square-free polynomial of degree n. Let T be the recursion tree of Akritas' algorithm applied to $A_{in}(X)$. The number of nodes in T is*

$$nO(NB(A_{in})^2 - nB(A_{in}) \log C(A_{in}, N) + nB(A_{in}) \log \mu(A_{in})),$$

where $B(A_{in})$ is defined in (3.33), $C(A_{in}, N)$ is the constant involved in the inequality (3.22), and $\mu(A_{in})$ is the largest absolute value amongst all the roots of $A(X)$.

Proof. Applying the bound in (3.51), along with the observation that $|U| \leq n$, to (3.50) we get that the size of the tree is bounded by

$$O(n + B(A_{\text{in}})) + \sum_{u \in U} \sum_{i=1}^{m_u} \kappa_n O(-NB(A_{\text{in}}) - n \log C(A_{\text{in}}, N) + n \log \mu(A_{\text{in}})).$$

From (3.36) we further get that the above bound is smaller than

$$\kappa_n O(-NB(A_{\text{in}}) - n \log C(A_{\text{in}}, N) + n \log \mu(A_{\text{in}})) \sum_{u \in U} \frac{1}{2}(1 + \log_\phi 2 - \log_\phi 2|\alpha_u - \beta_u|).$$

Again applying (3.51), we get that the size of the tree is bounded by

$$\kappa_n O(NB(A_{\text{in}})^2 - nB(A_{\text{in}}) \log C(A_{\text{in}}, N) + nB(A_{\text{in}}) \log \mu(A_{\text{in}})).$$

From the observation that $\kappa_n = \Theta(n)$, we get the desired result. □

Following the proof of Corollary 3.3, we get the following special case of the above theorem:

Corollary 3.4. *Let $A(X)$ be a square-free polynomial of degree n with integer coefficients of magnitude less than 2^L. The number of nodes in the recursion tree of Akritas' algorithm run on $A(X)$ is $\widetilde{O}(n^4 L^2)$.*

Remark 3.4. *Notice that asymptotically the bounds in (3.45) and (3.48) are the same, so even showing the tightness (as in (3.1)) of the bounds in [Kio86, Şte05, Hon98] does not improve upon the complexity result in the above corollary, though it will definitely lead to a simplification of the analysis.*

3.7 The Bit-Complexity

In this section we derive the bit-complexity of Akritas' algorithm for a square-free polynomial $A_{\text{in}}(X)$ such that $\|A_{\text{in}}\|_\infty < 2^L$. To do this we will bound the worst-case complexity at any node in the recursion tree; then along with Corollary 3.4 we have a bound on the bit-complexity of the algorithm.

Recall from starting of §3.6 the definitions of the set U, and of the pair (α_u, β_u) for any $u \in U$. Let

$$M_u(X) = [q_0, \ldots, q_{m+1}, X] = \frac{P_{m+1}X + P_m}{Q_{m+1}X + Q_m},$$
$$M_i(X) = [q_0, \ldots, q_i, X] = \frac{P_iX + P_{i-1}}{Q_iX + Q_{i-1}},$$

$A_i(X) = (Q_iX + Q_{i-1})^n A(M_i(X))$, L_i be such that $\|A_i\|_\infty < 2^{L_i}$, and b_i the bit-length of q_i for $i = 0, \ldots, m+1$.

To construct $A_{i+1}(X)$ from $A_i(X)$ we need to construct a sequence of polynomials $B_j(X)$, $1 \le j \le \ell$, such that $B_1(X) := A_i(X)$ and for $j > 1$, $B_j(X) := B_{j-1}(X + \delta_{j-1})$, where

$$\delta_j := 1 + \begin{cases} \text{PLB}(B_{j-1}) & \text{if PLB}(B_{j-1}) > 1 \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, $q_{i+1} = \sum_{j=1}^{\ell-1} \delta_j$. The two most important operations in computing $B_j(X)$ from $B_{j-1}(X)$ are computing $\text{PLB}(B_{j-1})$ and the Taylor shift by δ_j. We only focus on the latter operation since its cost dominates the cost of computing the former operation, which we know from Remark 3.1 is $\widetilde{O}(nL_i)$. Since $\delta_j \le q_{i+1}$, for $j < \ell$, the cost of computing each of Taylor shifts, i.e., the cost of computing $B_j(X)$ from $B_{j-1}(X)$ for all $j \le \ell$, is bounded by the cost of computing $A_i(X + q_{i+1})$; we bound this latter cost.

We know (see [Kra95]) that the computation of the Taylor shift can be arranged in a triangle of depth n; at each depth the multiplication by q_{i+1} increases the bit-length by b_{i+1}, so the bit-length of the coefficients of $A_i(X + q_{i+1})$ is bounded by $L_i + nb_{i+1}$. Moreover, using the classical approach, Taylor shifts can be performed in $O(n^2)$ additions [Kra95, JKR05, vzGG97]. Thus the

cost of computing $A_i(X+q_{i+1})$ is $O(n^2(L_i+nb_{i+1}))$. We further claim that $L_i = O(L+n\sum_{j=0}^{i} b_j)$; this is straightforward from the observation that $L_j \leq L_{j-1} + nb_{j-1}$. Thus the bit-complexity of computing $A_i(X + q_{i+1})$ is bounded by $O(n^2(L + n\sum_{j=0}^{i+1} b_j))$, if we use the classical Taylor shift. We next bound $\sum_{j=0}^{i+1} b_j$, $i \leq m$.

We know that $Q_m = q_m Q_{m-1}+Q_{m-2}$; thus $Q_m \geq q_m Q_{m-1}$, and recursively we get that $Q_m \geq \prod_{j=1}^{m} q_j$. Moreover, from (3.40) and the worst-case separation bound (see Remark 2.3 in §2.2.1) we know that $\log Q_m = \widetilde{O}(nL)$. Thus $\sum_{j=0}^{m} b_j = \widetilde{O}(nL)$. The troublesome part is bounding q_{m+1}, since Q_{m+1} does not satisfy (3.40). However, we do know that $q_{m+1} \leq |M_m^{-1}(\alpha_u)|$, and from (3.45) that

$$\log |M_m^{-1}(\alpha_u)| = O(-N \log |\alpha_u - \beta_u| - \log C(A_{\text{in}}, N)).$$

But from Lemma 3.5 we have $N = n$ and $-\log C(A_{\text{in}}, N) = \widetilde{O}(nL)$, and from the separation bound it follows that $-\log |\alpha_u - \beta_u| = \widetilde{O}(nL)$. Thus $b_{m+1} = \widetilde{O}(n^2 L)$ and hence $\sum_{j=0}^{m+1} b_j = \widetilde{O}(n^2 L)$.

So the worst-case bit-complexity at any node is asymptotically the same as computing $A_m(X)$, which we know is $\widetilde{O}(n^2(L + n\sum_{j=0}^{i+1} b_j)) = \widetilde{O}(n^5 L)$, when we use classical Taylor shifts. Along with the result in Corollary 3.4 we get the following:

Theorem 3.9. *Let $A(X)$ be a square-free integer polynomial of degree n with integer coefficients of magnitude less than 2^L. Then the bit-complexity of isolating all the real roots of $A(X)$ using Akritas' algorithm based upon classical Taylor shift is $\widetilde{O}(n^9 L^3)$.*

We can improve on the above bound by a factor of n using the fast Taylor shift [vzGG97].

Theorem 3.10. *Let $A(X)$ be a square-free integer polynomial of degree n with integer coefficients of magnitude less than 2^L. Then the bit-complexity of isolating all the real roots of $A(X)$ using Akritas' algorithm based upon a fast Taylor shift is $\widetilde{O}(n^8 L^3)$.*

Proof. The cost of computing $A_i(X+q_{i+1})$ using the convolution method (method F in [vzGG97]) is $O(M(n^2 b_{i+1} + nL_i))$, where $M(n)$ is the complexity of multiplying two n-bit integers. From above we know that $L_i = O(L+n\sum_{j=0}^{i} b_j)$, thus the cost is $O(M(nL+n^2\sum_{j=0}^{i+1} b_j))$. Moreover, we also know that $\sum_{j=0}^{m+1} b_j = \widetilde{O}(n^2 L)$. Assuming the Schönhage-Strassen method, and the Turing machine model of computation, we have $M(n) = O(n \log(n) \log\log(n)) = \widetilde{O}(n)$. Hence the worst-case bit-complexity of a node is $\widetilde{O}(n^4 L)$. Multiplying with the bound $\widetilde{O}(n^4 L^2)$ (from Corollary 3.4) on the size of the tree we get the complexity as mentioned in the theorem. □

Remark 3.5. *If we were to use the ideal* PLB *function then the worst case bit complexity of Akritas' algorithm is* $\widetilde{O}(n^5L^2)$*, since in this case the size of the tree is* $\widetilde{O}(nL)$ *and we know that the worst case complexity of each node is* $\widetilde{O}(n^4L)$*.*

3.8 Conclusion and Future Work

The bound in Theorem 3.10 is not as impressive as the complexity of the Descartes method, which we know (see Theorem 2.5) is $\widetilde{O}(n^4L^2)$. One of the reasons for this difference is that in our analysis we have used Liouville's inequality instead of Roth's theorem. For the latter result we know $N = O(1)$, but we do not know any bounds on the constant $C(A, N)$. However, if we assume that the constant $C(A, N)$ for Roth's theorem is the same as that in Lemma 3.5 then it follows that the size of the recursion tree of Akritas' algorithm is $\widetilde{O}(n^3L^2)$ and the worst case complexity at a node in the recursion tree is $\widetilde{O}(n^3L)$, and hence the worst case complexity of the algorithm is $\widetilde{O}(n^6L^3)$; note that under this assumption $\log q_{m+1} = \widetilde{O}(nL)$ (instead of $\widetilde{O}(n^2L)$) as would be expected. Moreover, if we additionally assume the ideal PLB function then we would get a worst case complexity of $\widetilde{O}(n^4L^2)$, which matches the expected bound in [ET06] and also the worst case complexity of the Descartes method. The assumption that $C(A, N)$ satisfies the same bound as in Lemma 3.5 is reasonable since it is known that in (3.22) $C(A, N) = 1$ except for a finitely many rationals. Thus the bound $\widetilde{O}(n^6L^3)$ is a more accurate statement on the actual performance of Akritas' algorithm than the bound in Theorem 3.10, which is an artefact of our analysis.

Another possibility is to devise functions that compute an upper bound on the absolute value of the roots of a polynomial and that satisfy a tighter inequality compared to the inequality, (3.1), satisfied by Zassenhaus' bound; for instance, if in (3.1) the upper bound was off by a constant factor then our complexity estimate improves by a factor of n.

A likely direction to pursue is to modify Akritas' algorithm so that its complexity bound improves without affecting its efficiency in practice. One way to modify the algorithm is to ensure that at each recursive level the width of the interval decreases by half [3]; recall that this was not guaranteed if we do consecutive Taylor shifts. However, this direction is different from our pursuit in this chapter, namely, to understand the worst case behaviour of the original algorithm by Akritas.

[3] I am grateful to Bernard Mourrain for suggesting this modification.

Appendix A

Multilinear maps and Banach Space

This appendix aims to attain self sufficiency in understanding the tools and techniques used in first chapter. The rigorous details of these definitions can be found in [Kre74]; for a quick reference Ostrowski's book [Ost73] is recommended.

Matrices, Norms and Inverse. Let E and F be vector spaces. Then a **multilinear map** $M : E^k \to F$ is a mapping that is linear in each of the k coordinates. More precisely, for any $(z_1, \ldots, z_k) \in E^k$, $y \in E$ and scalar c we have

$$M(z_1, \ldots, cz_i + y, \ldots, z_k) = cM(z_1, \ldots, z_i, \ldots, z_k) + M(z_1, \ldots, \overbrace{y}^{i}, \ldots, z_k). \tag{A.1}$$

Let E and F be normed vector spaces and let $\|\cdot\|$ be a norm on them; even though the norms on E and F may be different, we represent them by the same symbol. The **induced matrix** norm for M is defined as

$$\|M\| := \sup\left\{\frac{\|M \cdot z^k\|}{\|z\|^k} : z \in E - \{0\}\right\}, \tag{A.2}$$

where z^k is used to represent the k-tuple $\overbrace{(z, \ldots, z)}^{k}$. M is said to be a **bounded** map if $\|M\|$ is bounded; it is not hard to show that this also implies that M is continuous. Clearly, the norm is **consistent**, i.e., for any $z \in E$, $\|M \cdot z^k\| \leq \|M\|\|z\|^k$. An equivalent version of the definition of norm is

$$\|M\| = \sup\{\|M \cdot z^k\| : z \in E \text{ and } \|z\| = 1\}. \tag{A.3}$$

We prove the equivalence of the two definitions. It is not hard to see that the norm defined in (A.3) is smaller than the norm defined previously; we will prove the converse. Consider (A.2), for any z

$$\begin{aligned} \frac{\|M \cdot z^k\|}{\|z\|^k} &= \|\frac{M \cdot z^k}{\|z\|^k}\| \\ &= \|M \cdot (\frac{z}{\|z\|}, \ldots, \frac{z}{\|z\|})\| \\ &= \|M \cdot y^k\| \end{aligned}$$

where $y = \frac{z}{\|z\|}$ is such that $\|y\| = 1$. The induced matrix norms are **submultiplicative**, i.e., for multilinear maps M, N we have $\|M \cdot N\| \leq \|M\|\|N\|$. Most of this can be found in [HJ85].

A multilinear map $M : E^k \to F$ is said to be **symmetric** if

$$M(z_1, \ldots, z_k) = M(z_{\sigma(1)}, \ldots, z_{\sigma(k)}),$$

where $z_1, \ldots, z_k \in E$, and σ is any permutation of indices $1, \ldots, k$. Thus for any $z \in E$, we write $M \cdot z$ to represent $M(z, \overbrace{1, \ldots, 1}^{k-1})$.

The **inverse** of a linear map $M : E \to F$ exists if and only if $Mz = 0$ implies $z = 0$; it is a map from the range of M to E. We will call M **non-singular** if M^{-1} exists. Moreover, M^{-1} is a linear isomorphism from the range space of M to E. Thus if E and F are finite dimensional, then this implies that the dimension of E is not larger than the dimension of F. In particular, if the dimension of E is m and that of F is n then $m \leq n$, and the size of M^{-1} is $m * n$, where $m \leq n$; thus $M^{-1}M$ is a square-matrix of size $m * m$.

We can generalize the above definition of inverse to the case when the dimension of E may be larger than F. Again let $M : E \to F$ be a linear map, and let $S \subseteq E$. Then the **inverse of** M **relative to the set** S is written as $M^{-1}_{|S}$ and satisfies

$$M^{-1}_{|S} Mz = z \text{ for all } z \in S.$$

This implies that the dimension of S is at most the dimension of F, and that $S \subseteq (\ker(M))^{\perp}$. The definition of the inverse given earlier is the case when $S = E$. If we choose $S = (\ker(M))^{\perp}$, i.e., the set of all elements in E that are not mapped to zero in F, then this definition of the inverse coincides with the Moore-Penrose pseudoinverse, which is represented as $M^{-1}_{|(\ker(M))^{\perp}}$. The norm $\|M^{-1}_{|S}\|$ of the inverse map $M^{-1}_{|S}$ is defined as

$$\|M^{-1}_{|S}\| := \sup \left\{ \frac{\|M^{-1}_{|S} w\|}{\|w\|} : w = Mz, z \in S \right\}.$$

Note that for defining the norm we can use the equivalent form as given in (A.3). We have the following relation:

$$\|M^{-1}_{|S}\| = (\inf_{\|z\|=1, z \in S} \|Mz\|)^{-1}. \tag{A.4}$$

The proof is as follows:

$$(\inf_{\|z\|=1, z \in S} \|Mz\|)^{-1} = \sup_{\|z\|=1, z \in S} \frac{1}{\|Mz\|} = \sup_{\|z\|=1, z \in S} \frac{\|z\|}{\|Mz\|};$$

if $w = Mz$ then we get

$$(\inf_{\|z\|=1} \|Mz\|)^{-1} = \sup_{w} \frac{\|M_{|S}^{-1} w\|}{\|w\|} = \|M_{|S}^{-1}\|.$$

The set of all bounded linear maps from E to F is represented as $L(E, F)$; it is a Banach space with the norm as the induced matrix norm [Kre74, Thm. 2.10-2].

Functions on Banach Spaces. A Banach space is a complete normed vector space. By its definition, a complete space contains the limits of all the Cauchy sequences defined on that space. An example of Banach space that will be of interest to us is the finite-dimensional vector space $\mathbb{C}^n$ for $n \geq 0$. Let E and F be two Banach spaces and let $\|\cdot\|$ be a norm on these spaces. Let $f : E \to F$ be an analytic map; thus there is a power series approximation in the neighbourhood of any point. For any $z \in E$ let $Df(z)$ denote the **Fréchet derivative** of f at z. By definition, it is a *bounded linear map* $M : E \to F$ such that

$$\lim_{h \to 0} \frac{\|f(z+h) - f(z) - M(z)\|}{\|h\|} = 0,$$

here the limit is taken over all sequences of non-zero elements in E which converge to zero. In case E and F are finite dimensional $Df(z)$ is the Jacobi matrix. The inverse of $Df(z)$, if it exists, is represented as $Df(z)^{-1}$, which is map from F to E. Since $Df(z)$ is bounded and $L(E, F)$ is a Banach space, we can recursively apply the definition above to obtain the higher derivatives. More precisely, the function f is $k+1$ times differentiable on E if it is k times differentiable on E and for all $z \in E$ there exists a continuous symmetric multilinear map A of $k+1$ arguments such that the limit

$$\lim_{h_{k+1} \to 0} \frac{\|D^k f(z + h_{k+1})(h_1, \ldots, h_k) - D^k f(z)(h_1, \ldots, h_k) - A(h_1, \ldots, h_k, h_{k+1})\|}{\|h_k\|} = 0$$

exists uniformly for all $h_1, \ldots, h_k$ in bounded sets in E. The multilinear map A is the $k+1$-th derivative of f at z and is represented as $D^{k+1} f(z) : E^{k+1} \to F$.

Now we can write the **Taylor's expansion** of f in the neighbourhood of a point z as follows:

$$f(z+h) = f(z) + \sum_{k=1}^{\infty} \frac{1}{k!} D^k f(z) \cdot h^k,$$

here the notation h^k is used as in (A.2).

Appendix B

The Condition Number

Let $\mathcal{F} : \mathbb{C}^n \to \mathbb{C}^n$ be a zero-dimensional system of n polynomials in n variables, $\hat{\mathcal{F}} : \mathbb{C}^{n+1} \to \mathbb{C}^n$ be its homogenized form, and $J_{\mathcal{F}}(\mathbf{Z})$ be the Jacobian matrix of $\mathcal{F}$ at a point $\mathbf{Z} \in \mathbb{C}^n$. Usually, the condition number of $\mathcal{F}$ at a point $\mathbf{Z} \in \mathbb{C}^n$ is $\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\|$, but this definition is not invariant under scalings of the form $\mathcal{F}$ to $\rho\mathcal{F}$ and $\mathbf{Z}$ to $\rho\mathbf{Z}$, for some $\rho \in \mathbb{C}$. Shub-Smale [SS93a] have proposed a definition that is invariant under such scalings. According to their definition the condition number of $\mathcal{F}$ at a point $\mathbf{Z}$ depends upon the condition number of $\hat{\mathcal{F}}$; this also explains the stress on invariance under scalings.

There are two key components in the definition of condition number by Shub-Smale. The first is a weighted two-norm called the **Kostlan norm** $\|\hat{\mathcal{F}}\|_k$ on $\hat{\mathcal{F}}$ and is defined as

$$\|\hat{\mathcal{F}}\|_k := \sqrt{\sum_{i=1}^{n} \|\hat{F}_i\|_k^2},$$

where

$$\|\hat{F}_i\|_k := \left[\sum_{|J|=D_i} |\hat{F}_{iJ}|^2 \binom{D_i}{J}^{-1} \right]^{1/2},$$

and $\binom{D_i}{J} := \frac{D_i!}{J_0!J_1!\cdots J_n!}$. The advantage of using Kostlan norm over the two-norm is that it is unitary invariant (see [Mal] for a proof); this property is useful in the complexity results by Shub-Smale, and Malajovich. The second key component is the set of vectors orthogonal to some vector $\hat{\mathbf{Z}} \in \mathbb{C}^{n+1}$

$$N_{\hat{\mathbf{Z}}} := \{\hat{\mathbf{Y}} \in \mathbb{C}^{n+1} | \langle \hat{\mathbf{Z}}, \hat{\mathbf{Y}} \rangle = 0\}$$

where $\langle X, Y \rangle = \sum_{i=0}^{n} \overline{Y}_i X_i$ denotes the Hermitian inner product. This set is the set of all vectors orthogonal to $\hat{\mathbf{Z}}$ and hence has dimension n.

The **condition number** $\mu(\hat{\mathcal{F}}, \hat{\mathbf{Z}})$ of a homogenized system $\hat{\mathcal{F}}$ at a point $\hat{\mathbf{Z}} \in \mathbb{P}^n(\mathbb{C})$ is defined as

$$\mu(\hat{\mathcal{F}}, \hat{\mathbf{Z}}) := \|\hat{\mathcal{F}}\|_k \|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}} \mathtt{diag}(\sqrt{D_i}\|\hat{\mathbf{Z}}\|^{D_i-1})\| \tag{B.1}$$

where $\mathtt{diag}(a_i)$ represents a matrix whose diagonal entries are a_i and remaining entries are zero. The definition is well-defined if the inverse $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}}$ is well-defined, or equivalently if the

matrix $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})$ has rank n. For a root $\hat{\mathbf{Z}}^*$ of $\hat{\mathcal{F}}$, the matrix $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)^{-1}_{|N_{\hat{\mathbf{Z}}^*}}$ is the Moore-Penrose pseudoinverse. This will follow if we show that $N_{\hat{\mathbf{Z}}^*} = (\ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*))^{\perp}$. Since $\hat{\mathcal{F}}$ is a system of homogeneous polynomials, we know that for all $\hat{\mathbf{Z}} \in \mathbb{C}^{n+1}$,

$$J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}) \cdot \hat{\mathbf{Z}} = [D_1\hat{F}_1(\hat{\mathbf{Z}}), D_2\hat{F}_2(\hat{\mathbf{Z}}), \ldots, D_n\hat{F}_n(\hat{\mathbf{Z}})]^T.$$

Thus the root $\hat{\mathbf{Z}}^*$ belongs in the kernel of $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)$; since this matrix has rank n, the kernel only contains $\hat{\mathbf{Z}}^*$, which implies the complement of the kernel is the set of all vectors orthogonal to $\hat{\mathbf{Z}}^*$, i.e., the set $N_{\hat{\mathbf{Z}}^*}$.

For various properties of $\mu(\hat{\mathcal{F}}, \hat{\mathbf{Z}})$ see [SS93b]. Intuitively, $\mu(\hat{\mathcal{F}}, \hat{\mathbf{Z}})$ is inversely proportional to the distance between $\hat{\mathcal{F}}$ and the nearest polynomial system that vanishes at $\hat{\mathbf{Z}}$ and whose Jacobian at $\hat{\mathbf{Z}}$ is non-singular, see [SS93a] for a proof; the metric used for measuring the distance is based is based upon the Kostlan norm. The condition number $\mu(\mathcal{F}, \mathbf{Z})$ of the system $\mathcal{F}$ at a point $\mathbf{Z} \in \mathbb{C}^n$ is defined as

$$\mu(\mathcal{F}, \mathbf{Z}) := \mu(\hat{\mathcal{F}}, (1, \mathbf{Z})). \tag{B.2}$$

Define the condition number of the system $\hat{\mathcal{F}}$ as

$$\mu(\hat{\mathcal{F}}) := \max_{\hat{\tau}:\hat{\mathcal{F}}(\hat{\tau})=0} \mu(\hat{\mathcal{F}}, \hat{\tau}).$$

Following (B.2) we define the condition number $\mu(\mathcal{F})$ of the system $\mathcal{F}$ as $\mu(\hat{\mathcal{F}})$, the condition number of the homogenized system $\hat{\mathcal{F}}$.

Malajovich [Mal93, Thm. 13,p. 50] has shown the following bound on the condition number of the system:

$$\mu(\hat{\mathcal{F}}) \leq \mu(\Sigma) H(\hat{\mathcal{F}})^{d(\Sigma)}, \tag{B.3}$$

where $\mathcal{D} = \max(D_1, \ldots, D_n)$,

$$d(\Sigma) := n \prod D_i \sum D_j,$$

$$\mu(\Sigma) := \sqrt{D!} d(\Sigma) \left[3(n-1)! n^2 (d(\Sigma) + \max S(\hat{F}_i)) 2^n (\sum D_j)^n \prod D_j \right]^{d(\Sigma)},$$

and $H(\hat{\mathcal{F}})$ is the maximum absolute value over all the coefficients of $\hat{F}_i$, $i = 1, \ldots, n$.

Let $\mathbf{Z}^*$ be any root of $\mathcal{F}$ such that $J_{\mathcal{F}}(\mathbf{Z}^*)$ is invertible. If $\mathbf{Z}$ is such that

$$u := \|\mathbf{Z} - \mathbf{Z}^*\| \gamma(\mathcal{F}, \mathbf{Z}^*) < 1 - 1/\sqrt{2}$$

then we know

$$\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\| \leq \|J_{\mathcal{F}}(\mathbf{Z})^{-1} J_{\mathcal{F}}(\mathbf{Z}^*)\| \|J_{\mathcal{F}}(\mathbf{Z}^*)^{-1}\|.$$

Applying the bound from Lemma 1.3 we get that

$$\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\| \leq \frac{(1-u)^2}{\psi(u)}\|J_{\mathcal{F}}(\mathbf{Z}^*)^{-1}\|.$$

We will next show that

$$\|J_{\mathcal{F}}(\mathbf{Z}^*)^{-1}\| \leq (1+\|\mathbf{Z}^*\|^2)\|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)^{-1}_{N_{\hat{\mathbf{Z}}^*}}\| \leq (1+\|\mathbf{Z}^*\|^2)\frac{\mu(\hat{\mathcal{F}},\hat{\mathbf{Z}}^*)}{\|\hat{\mathcal{F}}\|_k}, \tag{B.4}$$

where $\hat{\mathbf{Z}}^* := (1, \mathbf{Z}^*)$. This implies the following: if $\mathbf{Z}$ is such that

$$u := \|\mathbf{Z}-\mathbf{Z}^*\|\gamma(\mathcal{F},\mathbf{Z}^*) < 1-1/\sqrt{2}$$

then

$$\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\| \leq \frac{(1-u)^2}{\psi(u)}(1+\|\mathbf{Z}^*\|^2)\frac{\mu(\hat{\mathcal{F}},\hat{\mathbf{Z}}^*)}{\|\hat{\mathcal{F}}\|_k}. \tag{B.5}$$

We start with the second inequality in (B.4). From the definition of the condition number we know that

$$\frac{\mu(\hat{\mathcal{F}},\hat{\mathbf{Z}})}{\|\hat{\mathcal{F}}\|_k} = \|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}}\mathtt{diag}(\sqrt{D_i}\|\hat{\mathbf{Z}}\|^{D_i-1})\|.$$

Since $\hat{\mathbf{Z}} = (1,\mathbf{Z})$ we know that $\|\hat{\mathbf{Z}}\| \geq 1$, and hence all the diagonal entries in $\mathtt{diag}(\sqrt{D_i}\|\hat{\mathbf{Z}}\|^{D_i-1})$ are greater than one. Thus it remains to show that

$$\|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}}\mathtt{diag}(\sqrt{D_i}\|\hat{\mathbf{Z}}\|^{D_i-1})\| \geq \|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}}\|.$$

We prove this for the general setting of a linear map $M : \mathbb{C}^n \to \mathbb{C}^{n+1}$. More precisely we will show that $\|M\mathtt{diag}(c_i)\| \geq \|M\|$, where $\mathtt{diag}(c_i)$ is a diagonal matrix of dimension n and $c_i \geq 1$, for $i = 1,\ldots,n$. From the definition of norm we know that

$$\|M\,\mathtt{diag}(c_i)\| = \max_{\|y\|=1}\|M\,\mathtt{diag}(c_i)y\| = \max_{\|y\|=1}\|My'\|,$$

where $y' := \mathtt{diag}(c_i)y$. Since $c_i \geq 1$ it follows that $\|y'\| \geq \|y\| = 1$, thus

$$\|M\,\mathtt{diag}(c_i)\| = \max_{\|y\|=1}\|My'\| = \max_{\|y\|=1}\|y'\|\|M\frac{y'}{\|y'\|}\| \geq \max_{\|z\|=1}\|Mz\| = \|M\|,$$

where in the second last step we have $z := \frac{y'}{\|y'\|}$.

We now prove the first inequality in (B.4). We first look at the structure of the matrix $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)$. The dimensions of this matrix is $n*(n+1)$. We observe that the $n*n$ square-matrix that is obtained from this matrix by removing its first column, i.e., the column that corresponds to the partial derivatives with respect to the homogenizing variable, is $J_{\mathcal{F}}(\mathbf{Z}^*)$. So we can write

the matrix $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*) := [w|J_{\mathcal{F}}(\mathbf{Z}^*)]$, where $w := -J_{\mathcal{F}}(\mathbf{Z}^*)\mathbf{Z}^*$; this guarantees that $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)\hat{\mathbf{Z}}^* = 0$. The inverse matrix

$$J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}} = \frac{1}{1+\|\mathbf{Z}^*\|^2}\left[\frac{v^T}{J_{\mathcal{F}}(\mathbf{Z})^{-1}}\right],$$

where $v \in \mathbb{C}^n$ is defined as $(J_{\mathcal{F}}(\mathbf{Z}^*)^{-1})^T\overline{\mathbf{Z}^*}$; $\overline{\mathbf{Z}^*}$ means we take the conjugate of the elements in $\mathbf{Z}^*$. This implies that the $(n+1)*(n+1)$ square-matrix

$$M := J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}} J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})$$

has the following form

$$M = \frac{1}{1+\|\mathbf{Z}^*\|^2}\begin{bmatrix} \|\mathbf{Z}^*\|^2 & -\overline{\mathbf{Z}^*}^T \\ -\mathbf{Z}^* & I \end{bmatrix}$$

where I is the identity matrix of size $n*n$. Clearly, M is Hermitian as expected, since $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)^{-1}_{|N_{\hat{\mathbf{Z}}^*}}$ is the Moore-Penrose pseudoinverse. Moreover, it can be verified that $M \cdot (1, \mathbf{Z}^*) = 0$, and $M \cdot \hat{\mathbf{Y}} = \hat{\mathbf{Y}}$, for all $\hat{\mathbf{Y}} \in N_{\hat{\mathbf{Z}}^*}$; we note that $N_{\hat{\mathbf{Z}}^*}$ contains elements of the form $(0, \mathbf{Y})$ where $\mathbf{Y} \in \mathbb{C}^n$ is such that $\langle \mathbf{Z}^*, \mathbf{Y}\rangle = 0$, and the element $(1, \mathbf{Y})$ where $\langle \mathbf{Z}^*, \mathbf{Y}\rangle = -1$; thus to prove that M is the "identity" transformation on $N_{\hat{\mathbf{Z}}^*}$ it suffices to show that $M(0, \mathbf{Y}) = (0, \mathbf{Y})$, when $\langle \mathbf{Z}^*, \mathbf{Y}\rangle = 0$, and $M(1, \mathbf{Y}) = (1, \mathbf{Y})$, when $\langle \mathbf{Z}^*, \mathbf{Y}\rangle = -1$. In addition to these properties it can be easily shown that $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)M = J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)$ and $MJ_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)^{-1}_{|N_{\hat{\mathbf{Z}}^*}} = J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}}^*)^{-1}_{|N_{\hat{\mathbf{Z}}^*}}$.

From the above description of $J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}}$ it follows that

$$\begin{aligned} \|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}}\| &= \frac{1}{1+\|\mathbf{Z}^*\|^2}\max_{\|\mathbf{Y}\|=1}\|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}}\mathbf{Y}\| \\ &= \frac{1}{1+\|\mathbf{Z}^*\|^2}\max_{\|\mathbf{Y}\|=1}\|(v^T\mathbf{Y}, J_{\mathcal{F}}(\mathbf{Z})^{-1}\mathbf{Y})^t\| \end{aligned}$$

which implies

$$\|J_{\mathcal{F}}(\mathbf{Z})^{-1}\| \leq (1+\|\mathbf{Z}^*\|^2)\|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}_{|N_{\hat{\mathbf{Z}}}}\|.$$

Thus we have showed the first inequality in (B.4).

For the sake of completeness, we now show why the definition of the condition number as given in (B.1) is suitable instead of the usual one $\|J_{\hat{\mathcal{F}}}(\hat{\mathbf{Z}})^{-1}\|$. This proof specializes the proof of Dégot [Dég00], who has shown the result for an underdetermined system of polynomials, i.e., a system containing more polynomials than the number of variables.

Let $\mathcal{H}_{\mathcal{D}}$ denote the linear space of all homogeneous polynomial systems $\hat{\mathcal{F}} : \mathbb{C}^{n+1} \to \mathbb{C}^n$. For two homogeneous polynomials $F_i, Q_i : \mathbb{C}^{n+1} \to \mathbb{C}$ of degree D_i, **Bombieri's scalar product** is defined by

$$[\hat{F}_i, \hat{G}_i]_{(D_i)} = \sum_{|J|=D_i} a_J\overline{b_J}\binom{D_i}{J}^{-1}, \tag{B.6}$$

where $\hat{F}_i(\mathbf{Z}) = \sum_{|J|=D_i} a_J \hat{\mathbf{Z}}^J$ and $\hat{G}_i(\mathbf{Z}) = \sum_{|J|=D_i} b_J \hat{\mathbf{Z}}^J$. The norm of a homogeneous polynomial $\hat{F}$ of degree D is $\|\hat{F}\| := [\hat{F}, \hat{F}]_{(D)}^{1/2}$. We can show that the Bombieri's scalar product satisfies the Cauchy-Schwarz inequality:

$$|[\hat{F}, \hat{G}]_{(D)}| \leq \|\hat{F}\| \|\hat{G}\|.$$

The above scalar product induces a Hermitian inner product on $\mathcal{H}_{\mathcal{D}}$: for $\hat{\mathcal{F}}, \hat{\mathcal{G}} \in \mathcal{H}_{\mathcal{D}}$,

$$[\hat{\mathcal{F}}, \hat{\mathcal{G}}] = \sum_{i=1}^{n} [\hat{F}_i, \hat{G}_i]_{(D_i)}.$$

We denote by $\|\hat{\mathcal{F}}\|$ and $d(\hat{\mathcal{F}}, \hat{\mathcal{G}})$ the associated norm and the distance; it is not hard to see that the norm is the same as the Kostlan norm $\|\hat{\mathcal{F}}\|_k$. For any $\hat{\mathbf{X}} \in \mathbb{C}^{n+1}$, let $\delta_{\hat{\mathbf{X}}}$ be the homogeneous polynomial of degree one defined as

$$\delta_{\hat{\mathbf{X}}}(\mathbf{Z}) = \sum_{i=0}^{n} \overline{X_i} Z_i.$$

From this it follows that $[\delta_{\hat{\mathbf{X}}}, \delta_{\hat{\mathbf{Y}}}] = \langle \hat{\mathbf{Y}}, \hat{\mathbf{X}} \rangle$. In the following, all the vector norms are the Euclidean norm. Moreover, we will simply use $[\hat{F}, \hat{G}]$ to denote $[\hat{F}, \hat{G}]_{(D)}$ wherever the degree D of the two polynomials can be understood from the context.

Bombieri's scalar product gives us a convenient way to represent evaluation of homogeneous polynomials:

Lemma B.1. *If $\hat{F}$ is a homogeneous polynomial of degree D, then for all $\hat{\mathbf{X}} \in \mathbb{C}^{n+1}$ we have*

$$\hat{F}(\hat{\mathbf{X}}) = [\hat{F}, \delta_{\hat{\mathbf{X}}}^D]_{(D)}.$$

The proof easily follows from the observation that $\delta_{\hat{\mathbf{X}}}^D = \sum_{|J|=D} \binom{D}{J} (\overline{\hat{\mathbf{X}}} \hat{\mathbf{Z}})^J$.

Based upon this we bound $|\hat{F}(\hat{\mathbf{X}})|$. From the Cauchy-Schwarz inequality we know that

$$|[\hat{F}, \delta_{\hat{\mathbf{X}}}^D]_{(D)}| \leq \|\hat{F}\| \|\delta_{\hat{\mathbf{X}}}^D\|.$$

Moreover, we can verify that

$$\|\delta_{\hat{\mathbf{X}}}^D\| = \sum_{\|J\|=D} \binom{d}{J} (|X_0| \cdots |X_n|)^{2J} = \|\hat{\mathbf{X}}\|^{2D}.$$

Thus we have

$$|\hat{F}(\hat{\mathbf{X}})| \leq \|\hat{F}\| \|\hat{\mathbf{X}}\|^D. \tag{B.7}$$

Let $\hat{\mathbf{X}} \in \mathbb{C}^{n+1}$ be a zero of the polynomial system $\hat{\mathcal{F}}$. The condition number of the system should measure the *relative sensitivity* of the solution with respect to the change of the data.

More precisely, let $\Delta\hat{\mathcal{F}}$ be an infinitesimal perturbation in $\hat{\mathcal{F}}$, and let $\Delta\hat{\mathbf{X}}$ be the corresponding smallest first order perturbation in $\hat{\mathbf{X}}$ such that

$$(\hat{\mathcal{F}} + \Delta\hat{\mathcal{F}})(\hat{\mathbf{X}} + \Delta\hat{\mathbf{X}}) = 0.$$

This implies

$$\hat{\mathcal{F}}(\hat{\mathbf{X}} + \Delta\hat{\mathbf{X}}) + \Delta\hat{\mathcal{F}}(\hat{\mathbf{X}} + \Delta\hat{\mathbf{X}}) = 0.$$

Doing a first order analysis we get

$$\hat{\mathcal{F}}(\hat{\mathbf{X}}) + J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\Delta\hat{\mathbf{X}} + \Delta\hat{\mathcal{F}}(\hat{\mathbf{X}}) = 0.$$

Since $\hat{\mathcal{F}}(\hat{\mathbf{X}}) = 0$ we get

$$J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\Delta\hat{\mathbf{X}} + \Delta\hat{\mathcal{F}}(\hat{\mathbf{X}}) = 0.$$

Because the system is zero-dimensional we know that $\Delta\hat{\mathbf{X}} \in N_{\hat{\mathbf{X}}} = (\ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})))^{\perp}$, otherwise we would have $(\hat{\mathcal{F}} + \Delta\hat{\mathcal{F}})(\hat{\mathbf{X}}) = 0$ for all perturbations $\Delta\hat{\mathcal{F}}$. Thus we have

$$\Delta\hat{\mathbf{X}} = -J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}\Delta\hat{\mathcal{F}}(\hat{\mathbf{X}}),$$

where for convenience we write $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{-1}_{|N_{\hat{\mathbf{X}}}}$ as $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}$. This implies

$$\begin{aligned}
\|\Delta\hat{\mathbf{X}}\| &= \|J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}\Delta\hat{\mathcal{F}}(\hat{\mathbf{X}})\| \\
&= \|J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}\mathtt{diag}(\|\hat{\mathbf{X}}\|^{D_i})\mathtt{diag}(\|\hat{\mathbf{X}}\|^{-D_i})\Delta\hat{\mathcal{F}}(\hat{\mathbf{X}})\| \\
&\leq \|J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}\mathtt{diag}(\|\hat{\mathbf{X}}\|^{D_i})\|\|\mathtt{diag}(\|\hat{\mathbf{X}}\|^{-D_i})\Delta\hat{\mathcal{F}}(\hat{\mathbf{X}})\|.
\end{aligned}$$

We now bound the term

$$\|\mathtt{diag}(D_i^{-1/2}\|\hat{\mathbf{X}}\|^{-D_i})\Delta\hat{\mathcal{F}}(\hat{\mathbf{X}})\| = (\sum_{i=1}^{n}|\Delta\hat{F}_i(\hat{\mathbf{X}})|^2\|\hat{\mathbf{X}}\|^{-2D_i})^{1/2}.$$

From (B.7) we know that $|\Delta\hat{F}_i(\hat{\mathbf{X}})| \leq \|\Delta\hat{F}_i\|\|X\|^{D_i}$. Thus

$$\|\mathtt{diag}(\|\hat{\mathbf{X}}\|^{-D_i})\Delta\hat{\mathcal{F}}(\hat{\mathbf{X}})\| \leq (\sum_{i=1}^{n}|\Delta\hat{F}_i\|^2)^{1/2} = \|\Delta\hat{\mathcal{F}}\|.$$

From this we get that

$$\|\Delta\hat{\mathbf{X}}\| \leq \|J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}\mathtt{diag}(\|\hat{\mathbf{X}}\|^{D_i})\|\|\Delta\hat{\mathcal{F}}\|$$

and hence

$$\frac{\|\Delta\hat{\mathbf{X}}\|\|\hat{\mathcal{F}}\|}{\|\hat{\mathbf{X}}\|\|\Delta\hat{\mathcal{F}}\|} \leq \|\hat{\mathcal{F}}\|\|J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}\mathtt{diag}(\|\hat{\mathbf{X}}\|^{D_i-1})\|.$$

As compared to our condition number in (B.1), this condition number is off by a factor of D_i in the diagonal matrix on the right hand side. However, this factor can be introduced if we overestimate the upper bound in (B.7) by the factor $\sqrt{D}$. This overestimation was required by Shub-Smale to get better complexity results.

We next derive the condition number theorem. Intuitively the condition number $\mu(\hat{\mathcal{F}}, \hat{\mathbf{Z}})$ of the system $\hat{\mathcal{F}}$ at its root $\hat{\mathbf{X}}$ is the inverse of the distance to the nearest system $\hat{\mathcal{G}}$ that has $\hat{\mathbf{X}}$ as its root and the Jacobian of $\hat{\mathcal{G}}$ at $\hat{\mathbf{X}}$ is singular, or in other words, the distance to the nearest ill-posed problem. More precisely, define the set $\Sigma_{\hat{\mathbf{X}}}$ of all ill-posed problems at $\hat{\mathbf{X}}$ as

$$\Sigma_{\hat{\mathbf{X}}} := \{\hat{\mathcal{G}} \in \mathcal{H}_{\mathcal{D}} : \hat{\mathcal{G}}(\hat{\mathbf{X}}) = 0 \text{ and rank of } J_{\hat{\mathcal{G}}}(\hat{\mathbf{X}}) \text{ is less than } n\}. \tag{B.8}$$

Let $d(\hat{\mathcal{F}}, \Sigma_{\hat{\mathbf{X}}})$ denote the smallest distance $\|\hat{\mathcal{F}} - \hat{\mathcal{G}}\|$ for all $\hat{\mathcal{G}} \in \Sigma_{\hat{\mathbf{X}}}$. The condition number theorem states the following:

Theorem B.1. *Let $\hat{\mathcal{F}} : \mathbb{C}^{n+1} \to \mathbb{C}^n$ be a system of homogeneous polynomials, and let $\hat{\mathbf{X}} \in \mathbb{C}^{n+1}$ be such that $\hat{\mathcal{F}}(\hat{\mathbf{X}}) = 0$ and $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})$ has rank n. Then*

$$\mu(\hat{\mathcal{F}}, \hat{\mathbf{X}}) = \frac{1}{d(\hat{\mathcal{F}}, \Sigma_{\hat{\mathbf{X}}})}.$$

The proof is again obtained by specializing the proof of Dégot [Dég00]. The proof of the theorem depends upon the following properties of Bombieri's scalar product.

Lemma B.2. *If $\hat{F}$, $\hat{G}$, and $\hat{H}$ are three homogeneous polynomials with degrees D_1, D_2, and D_3 such that $D_1 + D_2 = D_3$ then*

$$[\hat{F}\hat{G}, \hat{H}] = \frac{D_2!}{D_3!}[\hat{G}, \overline{\hat{F}}(\partial)\hat{H}],$$

where $\hat{F}(\partial)$ is the differential operator defined as

$$\hat{F}(\partial) = \hat{F}(\frac{\partial}{\partial Z_0}, \frac{\partial}{\partial Z_1}, \ldots, \frac{\partial}{\partial Z_n}),$$

i.e., each monomial of the form $Z_0^{i_0} \cdots Z_n^{i_n}$ in $\hat{F}$ is replaced by $\frac{\partial^{i_0}}{\partial Z_0^{i_0}} \cdots \frac{\partial^{i_n}}{\partial Z_n^{i_n}}$.

Proof. Let $\hat{F} = \sum_{|J|=D_1} a_J \hat{\mathbf{Z}}^J$, $\hat{G} = \sum_{|J|=D_2} b_J \hat{\mathbf{Z}}^J$ and $\hat{H} = \sum_{|J|=D_3} c_J \hat{\mathbf{Z}}^J$. Then it follows that

$$\begin{aligned}
\overline{\hat{F}}(\partial)\hat{H} &= \sum_{|J|=D_1} \overline{a}_J \frac{\partial^{D_1}}{(\partial Z_0 \cdots \partial Z_n)^J} \sum_{|J''|=D_3} c_{J''} \hat{\mathbf{Z}}^{J''} \\
&= \sum_{|J|=D_1} \sum_{|J''|=D_3} \overline{a}_J c_{J''} \frac{\partial^{D_1}}{(\partial Z_0 \cdots \partial Z_n)^J} \hat{\mathbf{Z}}^{J''} \\
&= \sum_{|J|=D_1} \sum_{\substack{J'' \geq J \\ |J''|=D_3}} \overline{a}_J c_{J''} \frac{J''!}{(J''-J)!} \hat{\mathbf{Z}}^{J''-J} \\
&= \sum_{|J'|=D_2} \sum_{|J|=D_1} \overline{a}_J c_{J+J'} \frac{(J+J')!}{J'!} \hat{\mathbf{Z}}^{J'}.
\end{aligned}$$

Thus

$$[\hat{G}, \overline{\hat{F}}(\partial)\hat{H}] = \sum_{|J'|=D_2} \sum_{|J|=D_1} a_J \overline{c}_{J+J'} b_{J'} \frac{(J+J')!}{J'!} \frac{J'!}{D_2!}.$$

Cancelling the term $J'!$ we get that

$$[\hat{G}, \overline{\hat{F}}(\partial)\hat{H}] = \sum_{|J'|=D_2} \sum_{|J|=D_1} a_J \overline{c}_{J+J'} b_{J'} \frac{(J+J')!}{D_2!}. \tag{B.9}$$

Also, from the definition of the scalar product it is not hard to see that

$$[\hat{F}\hat{G}, \hat{H}] = \sum_{|J''|=D_3} \left(\sum_{\substack{|J|=D_1, |J'|=D_2 \\ J+J'=J''}} a_J b_{J'} \right) \overline{c}_{J''} \frac{J''!}{D_3!}.$$

By re-arranging the summation indices in the above equation we get

$$[\hat{F}\hat{G}, \hat{H}] = \sum_{|J'|=D_2} \sum_{|J|=D_1} a_J b_{J'} \overline{c}_{J+J'} \frac{(J+J')!}{D_3!}.$$

From this and (B.9) we get that

$$[\hat{F}\hat{G}, \hat{H}] = \frac{D_2!}{D_3!} [\hat{G}, \overline{\hat{F}}(\partial)\hat{H}],$$

□

Based upon the above lemma we can show the following:

Lemma B.3. *Let $\hat{F}_1, \ldots, \hat{F}_k$ and $\hat{G}_1, \ldots, \hat{G}_k$ be homogeneous polynomials of degree one. Then*

$$[\hat{F}_1 \cdots \hat{F}_k, \hat{G}_1 \cdots \hat{G}_k]_{(k)} = \frac{1}{k!} \sum_{\sigma} [\hat{F}_1, \hat{G}_{\sigma(1)}] \cdots [\hat{F}_k, \hat{G}_{\sigma(k)}],$$

where σ varies over the set of all permutations of $\{1, \ldots, k\}$.

Proof. The proof is by induction on k; the base case for $k = 1$ trivially holds. Suppose the hypothesis holds for two sets of $k-1$ polynomials. From Lemma B.2 we know that

$$\begin{aligned}
[\hat{F}_1 \cdots \hat{F}_k, \hat{G}_1 \cdots \hat{G}_k]_{(k)} &= \frac{1}{k}[\hat{F}_1 \cdots \hat{F}_{k-1}, \overline{\hat{F}}_k(\partial)\hat{G}_1 \cdots \hat{G}_k]_{(k-1)} \\
&= \frac{1}{k}[\hat{F}_1 \cdots \hat{F}_{k-1}, \sum_{j=1}^{k}(\hat{G}_1 \cdots \hat{G}_{j-1} \cdot \hat{G}_{j+1} \cdots \hat{G}_k)\overline{\hat{F}}_k(\partial)\hat{G}_j]_{(k-1)} \\
&= \frac{1}{k}\sum_{j=1}^{k}[\hat{F}_1 \cdots \hat{F}_{k-1}, (\hat{G}_1 \cdots \hat{G}_{j-1} \cdot \hat{G}_{j+1} \cdots \hat{G}_k)\overline{\hat{F}}_k(\partial)\hat{G}_j]_{(k-1)}.
\end{aligned}$$

But $\overline{\hat{F}}_k(\partial)\hat{G}_j = \overline{[\hat{F}_k, \hat{G}_j]}$. Thus

$$[\hat{F}_1 \cdots \hat{F}_k, \hat{G}_1 \cdots \hat{G}_k]_{(k)} = \frac{1}{k}\sum_{j=1}^{k}[\hat{F}_1 \cdots \hat{F}_{k-1}, (\hat{G}_1 \cdots \hat{G}_{j-1} \cdot \hat{G}_{j+1} \cdots \hat{G}_k)]_{(k-1)}[\hat{F}_k, \hat{G}_j].$$

Applying the induction hypothesis to the term

$$[\hat{F}_1 \cdots \hat{F}_{k-1}, (\hat{G}_1 \cdots \hat{G}_{j-1} \cdot \hat{G}_{j+1} \cdots \hat{G}_k)]_{(k-1)}$$

for $j = 1, \ldots, k$ we get

$$[\hat{F}_1 \cdots \hat{F}_k, \hat{G}_1 \cdots \hat{G}_k]_{(k)} = \frac{1}{k!}\sum_{j=1}^{k}\sum_{\sigma_j}[\hat{F}_1, \hat{G}_{\sigma_j(1)}] \cdots [\hat{F}_{k-1}, \hat{G}_{\sigma_j(k-1)}][\hat{F}_k, \hat{G}_j].$$

But we can define σ such that $\sigma(i) = \sigma_j(i)$, for $i = 1, \ldots, k-1$, and $\sigma(k) = j$; moreover, the set of all such σ's covers all the possible permutations on the set $\{1, \ldots, k\}$. Thus the above equation can be re-written as

$$[\hat{F}_1 \cdots \hat{F}_k, \hat{G}_1 \cdots \hat{G}_k]_{(k)} = \frac{1}{k!}\sum_{\sigma}[\hat{F}_1, \hat{G}_{\sigma(1)}] \cdots [\hat{F}_k, \hat{G}_{\sigma(k)}],$$

hence proving the lemma. □

We also have the following representation for the vector $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Y}}$, for $\hat{\mathbf{Y}} \in \mathbb{C}^{n+1}$.

Lemma B.4. *Let $\hat{\mathcal{F}} = (\hat{F}_1, \ldots, \hat{F}_n)$ be a system of homogeneous polynomials such that $\hat{F}_i$ has degree D_i. Then for $\hat{\mathbf{X}}, \hat{\mathbf{Y}} \in \mathbb{C}^{n+1}$ the i-th coordinate of the vector $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Y}}$ is $D_i[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Y}}}]$, for $i = 1, \ldots, n$.*

Proof. The i-th coordinate of $J_{\hat{\mathcal{F}}}\hat{\mathbf{Y}}$ is $\sum_{j=0}^{n} Y_j \frac{\partial \hat{F}_i}{\partial Z_j}(\hat{\mathbf{X}})$; but $\sum_{j=0}^{n} Y_j \frac{\partial \hat{F}_i}{\partial Z_j} = \delta_{\hat{\mathbf{Y}}}(\partial)\hat{F}_i$. Thus from Lemma B.1 we get that

$$\sum_{j=0}^{n} Y_j \frac{\partial \hat{F}_i}{\partial Z_j}(\hat{\mathbf{X}}) = [\delta_{\hat{\mathbf{Y}}}(\partial)\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}].$$

Applying Lemma B.2, we get the desired result. □

Now we proceed to prove the condition number theorem, Theorem B.1. Recall the definition of $d(\hat{\mathcal{F}}, \Sigma_{\hat{\mathbf{X}}})$ which was $\inf_{\hat{\mathcal{H}} \in \Sigma_{\hat{\mathbf{X}}}} \|\hat{\mathcal{F}} - \hat{\mathcal{H}}\|$; this is the same as the infimum of $\|\hat{\mathcal{G}}\|$ such that

1. $(\hat{\mathcal{F}} + \hat{\mathcal{G}})(\hat{\mathbf{X}}) = 0$, and
2. rank of $J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}})$ is less than n.

The proof consists of two parts: in the **first part**, we show that the norm of any $\hat{\mathcal{G}}$ that satisfies the above constraints is greater than $\mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$, thus implying $d(\hat{\mathcal{F}}, \Sigma_{\hat{\mathbf{X}}}) \geq \mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$; in the **second part** we construct a specific $\hat{\mathcal{G}}$ that satisfies the above constraints and show that its norm is less than $\mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$, which implies that $d(\hat{\mathcal{F}}, \Sigma_{\hat{\mathbf{X}}}) \leq \mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$. From these two parts it follows that $d(\hat{\mathcal{F}}, \Sigma_{\hat{\mathbf{X}}}) = \mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$.

We begin with proving the first part. Since $(\hat{\mathcal{F}} + \hat{\mathcal{G}})(\hat{\mathbf{X}}) = 0$ it follows from Euler's identity that $\hat{\mathbf{X}} \in \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))$. We want to characterize the polynomial $\hat{\mathcal{G}}$ that satisfies the above two constraints and has the smallest norm. Consider the following set

$$S := \{\hat{\mathcal{H}} \in \mathcal{H}_{\mathcal{D}} : \hat{\mathcal{H}}(\hat{\mathbf{X}}) = 0 \text{ and } \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}})) \subseteq \ker(J_{\hat{\mathcal{H}}}(\hat{\mathbf{X}}))\}.$$

We claim that $\hat{\mathcal{G}}$ belongs to the complement of S; this is because the rank of $J_{\hat{\mathcal{G}}}(\hat{\mathbf{X}})$ is the same as the rank of $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})$, which is n, and since rank of $\ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))$ is less than n we know $\ker(J_{\hat{\mathcal{G}}}(\hat{\mathbf{X}})) \subset \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))$. Moreover, the complement of S is the set

$$\{(\hat{H}_1, \ldots, \hat{H}_n) : \hat{H}_i = \delta_{\hat{\mathbf{X}}}^{D_i - 1} \delta_{\hat{\mathbf{Y}}_i}, \hat{\mathbf{Y}}_i \in \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))\}.$$

The reason is that for all $\hat{\mathcal{H}} \in S$, $\ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}})) \subseteq \ker(J_{\hat{\mathcal{H}}}(\hat{\mathbf{X}}))$ implies $J_{\hat{\mathcal{H}}}(\hat{\mathbf{X}})\hat{\mathbf{Y}} = 0$ for all $\hat{\mathbf{Y}} \in \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))$; along with Lemma B.4 this implies that $[\hat{H}_i, \delta_{\hat{\mathbf{X}}}^{D_i - 1} \delta_{\hat{\mathbf{Y}}}] = 0$. Thus we have that the polynomial system $\hat{\mathcal{G}} = (\hat{G}_1, \ldots, \hat{G}_n)$ is such that

$$\hat{G}_i = \delta_{\hat{\mathbf{X}}}^{D_i - 1} \delta_{\hat{\mathbf{Y}}_i},$$

where $\hat{\mathbf{Y}}_i \in \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))$. Since $\hat{\mathcal{G}}(\hat{\mathbf{X}}) = 0$ we know from Lemma B.1 that

$$\hat{G}_i(\hat{\mathbf{X}}) = [\hat{\mathcal{G}}_i, \delta_{\hat{\mathbf{X}}}^{D_i - 1}] = [\delta_{\hat{\mathbf{X}}}^{D_i - 1} \delta_{\hat{\mathbf{Y}}_i}, \delta_{\hat{\mathbf{X}}}^{D_i}] = \|\hat{\mathbf{X}}\|^{2(D_i - 1)} \langle \hat{\mathbf{X}}, \hat{\mathbf{Y}}_i \rangle = 0,$$

or that $\langle \hat{\mathbf{X}}, \hat{\mathbf{Y}}_i \rangle = 0$, here the last step follows from Lemma B.3. We next derive a lower bound on $\|\hat{\mathcal{G}}\|$ and will show that it is greater than the inverse of the condition number. We know

$$\|\hat{\mathcal{G}}\|^2 = \sum_{i=1}^{n} [\hat{G}_i, \hat{G}_i] = \sum_{i=1}^{n} [\delta_{\hat{\mathbf{X}}}^{D_i - 1} \delta_{\hat{\mathbf{Y}}_i}, \delta_{\hat{\mathbf{X}}}^{D_i - 1} \delta_{\hat{\mathbf{Y}}_i}] = \sum_{i=1}^{n} \frac{1}{D_i} \|\hat{\mathbf{Y}}_i\|^2 \|\hat{\mathbf{X}}\|^{2(D_i - 1)},$$

where the last step follows from Lemma B.3 along with the fact that $\langle \hat{\mathbf{X}}, \hat{\mathbf{Y}}_i \rangle = 0$. We need to derive a lower bound on $\|\hat{\mathbf{Y}}_i\|$. Consider any $\hat{\mathbf{Z}} \in \ker(J_{\mathcal{F}}(\hat{\mathbf{X}}))^{\perp} \cup \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))$ such that $\|\hat{\mathbf{Z}}\| = 1$. Then we know that $J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}} = 0$, and hence from Lemma B.4 that for $i = 1, \ldots, n$

$$[\hat{F}_i + \hat{G}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}] = 0$$

or equivalently

$$[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}] = -[\hat{G}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}] = -[\delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Y}}_i}, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}] = -\frac{1}{D_i}\|\hat{\mathbf{X}}\|^{2(D_i-1)}\langle \hat{\mathbf{Z}}, \hat{\mathbf{Y}}_i \rangle,$$

where the last step follows from Lemma B.3 along with the fact that $\langle \hat{\mathbf{X}}, \hat{\mathbf{Y}}_i \rangle = 0$. This implies

$$|[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}]| \leq \frac{1}{D_i}\|\hat{\mathbf{X}}\|^{2(D_i-1)}\|\hat{\mathbf{Y}}_i\|,$$

because $\|\hat{\mathbf{Z}}\| = 1$ and $\langle \hat{\mathbf{Z}}, \hat{\mathbf{Y}}_i \rangle \leq \|\hat{\mathbf{Z}}\|\|\hat{\mathbf{Y}}_i\|$. Thus we have

$$\|\hat{\mathcal{G}}\|^2 \geq \sum_{i=1}^{n} D_i \|\hat{\mathbf{X}}\|^{-2(D_i-1)} |[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}]|^2.$$

But from Lemma B.4 we know that $[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}]$ is the i-th component of the vector $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}$. Thus the summation on the right hand side can be expressed as the square of norm of the vector

$$\mathtt{diag}(D_i^{-1/2}\|\hat{\mathbf{X}}\|^{-2(D_i-1)})J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}.$$

Thus

$$\|\hat{\mathcal{G}}\|^2 \geq \|\mathtt{diag}(D_i^{-1/2}\|\hat{\mathbf{X}}\|^{-2(D_i-1)})J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}\|^2.$$

But the right hand side is clearly greater than

$$\inf \|\mathtt{diag}(D_i^{-1/2}\|\hat{\mathbf{X}}\|^{-2(D_i-1)})J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}\|^2,$$

where the infimum is taken over all $\hat{\mathbf{Z}} \in \ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}}))^{\perp}$ such that $\|\hat{\mathbf{Z}}\| = 1$. Applying the relation (A.4) we get that

$$\|\hat{\mathcal{G}}\|^2 \geq (\|J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}\mathtt{diag}(D_i^{1/2}\|\hat{\mathbf{X}}\|^{2(D_i-1)})\|)^{-1} = \mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$$

as desired. This brings us to the conclusion of the first part mentioned earlier. Now we proceed to prove the second part.

Recall that in this part we construct a specific $\hat{\mathcal{G}}$ that satisfies the two constraints mentioned earlier and whose norm is less than $\mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$. Let $\hat{\mathbf{Z}}$ in $\ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}}))^{\perp}$ be such that $\|\hat{\mathbf{Z}}\| = 1$ and it attains the minimum norm for

$$\|\mathtt{diag}(D_i^{-1/2}\|\hat{\mathbf{X}}\|^{-2(D_i-1)})J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}\|$$

amongst all the elements in $\ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}}))^{\perp}$. Define the system of homogeneous polynomials $\hat{\mathcal{G}}$ as

$$\hat{G}_i := D_i^{-1}\|\hat{\mathbf{X}}\|^{-2(D_i-1)}|[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}]|\delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}, \tag{B.10}$$

for $i = 1, \ldots, n$. Then $\hat{\mathcal{G}}$ has the following properties:

1. $(\hat{\mathcal{F}} + \hat{\mathcal{G}})(\hat{\mathbf{X}}) = 0$. Consider the evaluation

$$\begin{aligned}\hat{G}_i(\hat{\mathbf{X}}) &= [G_i, \delta_{\hat{\mathbf{X}}}^{D_i}] \\ &= D_i^{-1}\|\hat{\mathbf{X}}\|^{-2(D_i-1)}|[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}]|[\delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}, \delta_{\hat{\mathbf{X}}}^{D_i}] \\ &= |[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}]|\langle\hat{\mathbf{Z}}, \hat{\mathbf{X}}\rangle,\end{aligned}$$

   where the last step follows from Lemma B.3. But $\langle\hat{\mathbf{Z}}, \hat{\mathbf{X}}\rangle = 0$, since $\hat{\mathbf{Z}} \in \ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}}))^{\perp}$ and $\hat{\mathbf{X}} \in \ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}}))$.

2. Rank of $J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}})$ is less than n. We already know that $\hat{\mathbf{X}} \in \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))$, but we claim that $\hat{\mathbf{Z}} \in \ker(J_{\hat{\mathcal{F}}+\hat{\mathcal{G}}}(\hat{\mathbf{X}}))$, and hence we will get the desired result. To prove the claim we will show that the sum of the i-th coordinate of the two vectors $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}$ and $J_{\hat{\mathcal{G}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}$ is zero. From Lemma B.4 we know that the i-th coordinate of $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}$ is $[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}]$; moreover, from the same lemma and (B.10) we also know that the i-th coordinate of $J_{\hat{\mathcal{G}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}$ is

$$-D_i\|\hat{\mathbf{X}}\|^{-2(D_i-1)}[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}][\delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}] = -[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}]$$

   since from Lemma B.3 we know that $[\delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}] = D_i^{-1}\|\hat{\mathbf{X}}\|^{2(D_i-1)}$. Thus the the sum of the i-th coordinate of the two vectors $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}$ and $J_{\hat{\mathcal{G}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}$ is indeed zero.

From the above two properties of $\hat{\mathcal{G}}$ it follows that $d(\hat{\mathcal{F}}, \Sigma_{\hat{\mathbf{X}}}) \leq \|\hat{\mathcal{G}}\|$. We now need to show that $\|\hat{\mathcal{G}}\| \leq \mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$. But

$$\begin{aligned}\|\hat{G}_i\|^2 = [\hat{G}_i, \hat{G}_i] &= D_i^2\|\hat{\mathbf{X}}\|^{-4(D_i-1)}[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}]^2[\delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}] \\ &= D_i\|\hat{\mathbf{X}}\|^{-2(D_i-1)}[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}]^2,\end{aligned}$$

where the last step follows from the fact that $\hat{\mathbf{Z}} \in \ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}}))^{\perp}$ and $\hat{\mathbf{X}} \in \ker(J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}}))$ implies $\langle\hat{\mathbf{X}}, \hat{\mathbf{Z}}\rangle = 0$, and hence from Lemma B.3 we get

$$[\delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}}] = D_i^{-1}\|\hat{\mathbf{X}}\|^{2(D_i-1)}.$$

Thus we have

$$\|\hat{\mathcal{G}}\|^2 = \sum_{i=1}^{n} D_i\|\hat{\mathbf{X}}\|^{-2(D_i-1)}|[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}]|^2.$$

But from Lemma B.4 we know that $[\hat{F}_i, \delta_{\hat{\mathbf{X}}}^{D_i-1}\delta_{\hat{\mathbf{Z}}_i}]$ is the i-th component of the vector $J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}$. Thus the summation on the right hand side can be expressed as the square of norm of the vector

$$\mathtt{diag}(D_i^{-1/2}\|\hat{\mathbf{X}}\|^{-2(D_i-1)})J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}.$$

Hence

$$\|\hat{\mathcal{G}}\| = \|\mathtt{diag}(D_i^{-1/2}\|\hat{\mathbf{X}}\|^{-2(D_i-1)})J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}\|.$$

But the way $\hat{\mathbf{Z}}$ was defined, we know from (A.4) that

$$\begin{aligned}\|\mathtt{diag}(D_i^{-1/2}\|\hat{\mathbf{X}}\|^{-2(D_i-1)})J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})\hat{\mathbf{Z}}\| &= (\|J_{\hat{\mathcal{F}}}(\hat{\mathbf{X}})^{+}\mathtt{diag}(D_i^{1/2}\|\hat{\mathbf{X}}\|^{2(D_i-1)})\|)^{-1} \\ &= \mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1},\end{aligned}$$

which implies that

$$d(\hat{\mathcal{F}}, \Sigma_{\hat{\mathbf{X}}}) \leq \|\hat{\mathcal{G}}\| = \mu(\hat{\mathcal{F}}, \hat{\mathbf{X}})^{-1}$$

as desired. Thus we have proved Theorem B.1.

Appendix C

BigFloat Computation

We review some basic facts about bigfloats. The name "bigfloat" serves to distinguish this from the usual programming concept of "floats" which has fixed precision. For a survey on bigfloat computation, see [YD95]. As in Brent [Bre76b, Bre76a], we use bigfloat numbers to approximate real or complex numbers. A (binary) **bigfloat** is a rational number of the form $x = n2^m$ where $n, m \in \mathbb{Z}$.

For an integer f, write $\langle f \rangle$ for the value $f2^{-\lfloor \lg |f| \rfloor}$. In the standard binary notation, $\langle f \rangle$ may be written as $\sigma(b_0.b_1b_2\cdots b_t)_2$, where $\sigma \in \{+,-\}$ and $f = \sigma \sum_{i=0}^{t} b_i 2^{t-i}$. We call $\langle f \rangle$ the "normalized value" of f. For example, $\langle 1 \rangle = \langle 2 \rangle = \langle 4 \rangle = 1$, $\langle 3 \rangle = \langle 6 \rangle = 1.5$, $\langle 5 \rangle = 1.25$, $\langle 7 \rangle = 1.75$, etc. In general, for $f \neq 0$, we have $|\langle f \rangle| \in [1, 2)$.

An alternative representation of bigfloats is $\langle e, f \rangle$, for integers e and f, that represents the bigfloat $f2^{e-\lfloor \lg f \rfloor} = \langle f \rangle 2^e$. E.g., the value of $\langle \lfloor \lg |f| \rfloor, f \rangle$ is f. We say $\langle e, f \rangle$ is **normalized** if $e = f = 0$ or if f is odd. Clearly every bigfloat has a unique normalized representation. We say $\langle e, f \rangle$ has **precision** t if $|f| < 2^t$. The advantage of this representation is that information about the magnitude is available in the exponent e, i.e., $2^e \leq \langle e, f \rangle < 2^{e+1}$, and is disjoint from the information about the precision which is available in f. A bigfloat is said to be **bounded** if $e = O(1)$. The **bit size** of $\langle e, f \rangle$ is the pair $(\lg(2 + |e|), \lg(2 + |f|))$.

Consider a bigfloat number

$$x = \langle e_x, f_x \rangle = f_x 2^{e_x - \lfloor \lg |f_x| \rfloor} = \langle f_x \rangle 2^{e_x}.$$

A restriction in Brent's complexity model is that all bigfloats x used in a given computation are **bounded**, i.e., $e_x = O(1)$ for any bigfloat $x = \langle e_x, f_x \rangle$. We are however interested in unbounded bigfloats. For unbounded bigfloats, we found it to be essential to adopt a more flexible computational model based on the Pointer machines of Schönhage [Sch80] rather than Turing machines.

Theorem C.1. *Let $x = \langle e_x, f_x \rangle, y = \langle e_y, f_y \rangle$ be unbounded bigfloats, and n be a positive natural number. Also, $f_x f_y \neq 0$.*

1. *We can compute $[x]_n$ in $C_0(n + \lg(2 + |e_x|))$ time.*

2. *We can compute* $[xy]_n$ *in* $C_0(M(n) + \lg(2 + |e_x e_y|))$ *time.*

3. *We can compute* $[x + y]_n$ *in* $C_0(n + \lg(2 + |e_x e_y|))$ *time provided* $xy \geq 0$ *or* $|x| > 2|y|$ *or* $|x| < |y|/2$. *In general, computing* $[x + y]_n$ *can be done in time* $O(\lg(2 + |f_x f_y e_x e_y|))$.

4. *An analogous statement holds for* $[x - y]_n$, *where we replace* $xy \geq 0$ *by* $xy \leq 0$.

C_0 *is a constant that is independent of* x *and* y.

Proof. 1. **Truncation**: To compute $[x]_n$ in $O(n + \lg(2 + |e_x|))$ time on a pointer model: given the input n in binary and $x = \langle e_x, f_x \rangle$, we simply treat n as a binary counter and count down to 0, it is well-known that this takes $O(n)$ steps; simultaneously, we output the most significant n-bits of f_x. In other words, this complexity does not depend on $\lg |f_x|$. We can also output e_x in $O(\lg(2 + |e_x|))$ time.

2. **Addition**: We can easily check that $xy \geq 0$ and $|x| > 2|y|$ or $2|x| \leq |y|$ in $O(2 + \lg |e_x e_y|))$ time. If so, we carry out

   (a) Compute $[x]_{n+2}$ and $[y]_{n+2}$. This takes time $O(n + \lg(2 + |e_x e_y|))$.

   (b) Compare e_x and e_y. This takes $O(\lg(2 + |e_x e_y|))$. Let $e_x \geq e_y$.

   (c) Compute $e_x - e_y$. This takes $O(\lg(2 + |e_x e_y|))$. Shift the decimal point of y by $\min\{e_x - e_y, n\}$ bits; this takes $O(n)$.

   (d) Add the two fractional parts; this takes $O(n)$. Since by assumption either both the fractional parts have the same sign, in which case no cancellation occurs, or the most significant bit of $x + y$ is to the right of x or y, depending upon whether $|x| \geq 2|y|$ or vice versa.

   Thus the total complexity is $O(n + \lg(2 + |e_x e_y|))$.

   In general, i.e., when the above assumptions fail, the complexity will be $O(\lg |f_x f_y| + \lg(2 + |e_x e_y|))$, because the fractional parts may be equal which would lead to catastrophic cancellation.

3. **Subtraction**: Has the same complexity as addition, except that the assumption $xy \geq 0$ should be $xy \leq 0$.

4. **Multiplication**: We carry these steps.

(a) Compute $[x]_{n+2}$ and $[y]_{n+2}$.

(b) Multiply the fractional parts of the truncations.

(c) Add the two exponents.

Thus the total complexity is $O(M(n) + \lg|e_x e_y|)$.

□

It is clear from these arguments that the constants in the preceding results are independent of the choice of x, y.

Evaluating a polynomial to absolute precision. Given $f(x) = \sum_{i=0}^{d} a_i x^i$, $a_i \in \mathbb{R}$, and $s \in \mathbb{Z}$, let $\widetilde{f}$ be the result of evaluating $f(x)$ at $x \in \mathbb{R}$ using Horner's rule where each operation is carried out with *relative* precision s. Given $n \in \mathbb{Z}$, we want to determine $s = s(n)$ such that $\langle f(x)\rangle_n = \widetilde{f}$. Here we assume that the coefficients a_i and x are "blackbox" numbers that output a desired approximation. Moreover, we assume that given a blackbox number α, we can compute $[\alpha]_n$, a bigfloat, in time $B(n)$. For instance, if α is a bigfloat then we know from the theorem above that $B(n) = C_0(n+\lg(2+|\lg\alpha|))$, where $C_0 > 0$ is independent α; in case α is an algebraic number $B(n) = O(M(n))$ – in fact this is what Brent has shown – however, the constant in O depends upon α.

Let e_i be such that $2^{e_i} \le a_i < 2^{e_i+1}$, e_x such that

$$2^{e_x} \le x < 2^{e_x+1} \tag{C.1}$$

and

$$e := \max(e_0, \ldots, e_d). \tag{C.2}$$

Similarly to Higham [Hig96, p. 105], we can show that

$$\begin{aligned} |\widetilde{f} - f(x)| &\le \gamma_{2d+1} \sum_{i=0}^{d} |a_i||x|^i \\ &\le \gamma_{2d+1} 2^{e+1} \sum_{i=0}^{d} |x|^i. \end{aligned}$$

where $\gamma_k := \frac{k2^{-s}}{1-k2^{-s}}$. We want to choose s such that the right hand side in the above inequality is less than 2^{-n}. To do so, we consider the following cases:

1. When $e_x \geq 0$, i.e., $|x| \geq 1$. Then we have

$$\begin{array}{lllll} & \gamma_{2d+1} \sum_{i=0}^{d} |a_i||x|^i & \leq & 2^{-n} & \\ \Leftarrow & 2^{e+3+d(e_x+1)} d(d+1)2^{-s} & \leq & 2^{-n} & \text{if } s \geq 2 + \lg d \\ \Leftarrow & s & \geq & n + e + d(e_x + 3) + 4 & (*). \end{array}$$

2. When $e_x < 0$, i.e, $|x| \leq 1$. Then

$$\begin{array}{lllll} & \gamma_{2d+1} \sum_{i=0}^{d} |a_i||x|^i & \leq & 2^{-n} & \\ \Leftarrow & \gamma_{2d+1} 2^{e+1}(d+1) & \leq & 2^{-n} & \\ \Leftarrow & 2^{e+3}(d+1)2^{-s} & \leq & 2^{-n} & \text{if } s \geq 2 + \lg(1+d) \\ \Leftarrow & s & \geq & n + e + \lg d + 4 & (**). \end{array}$$

The complexity of evaluation is evident from the following:

1. Compute $[a_i]_s$ for $i = 0, \ldots, d$ and $[x]_s$. This takes $O(dB(s))$ by our assumption on a_i and x.

2. Each addition and multiplication in the Horner's evaluation is between bigfloats of precision s. Thus the cost of each operation is $O(M(s))$. Since Horner's evaluation involves $O(d)$ steps, its cost is $O(dM(s))$.

Lemma C.1. *The complexity of evaluating $f(x) \in \mathbb{R}[x]$ at a point $x \in \mathbb{R}$ to absolute precision n is*

$$O(d[M(n + e + d \max\{1, e_x\}) + B(n + e + d \max\{1, e_x\})]),$$

where e and e_x are defined in (C.1) and (C.2) respectively.

NOTE: The complexity of computing $f'(x)$ is the same, since only the bit-size of the coefficients is increased to $e + \lg d$, which can be subsumed by $e + d \max\{1, e_x\}$.

For the special case of evaluating integer polynomials we have the following:

Lemma C.2. *Given $f(x) = \sum_{i=0}^{d} a_i x^i$, where $a_i \in \mathbb{Z}$ are L-bit integers, and x a bigfloat, we can evaluate $f(x)$ in time $O(dM(dL' + dL))$ where L' is the bit size of x.*

Proof. There are d algebraic operations involved in Horner's method. At each such operation the bit size increases by a factor of $O(L' + L)$ and hence the overall complexity is $\sum_{i=1}^{d} O(M(i(L' + L))) = O(dM(dL' + dL))$. □

Evaluating a system of multi-variate integer polynomials Let $\mathcal{F}$ be a system of polynomials F_i, $i = 1, \ldots, n$, where $F_i \in \mathbb{Z}[X_1, \ldots, X_n]$ is such that its coefficients have bit-length at most L and the total degree of F_i is D_i. We want to bound the complexity of evaluating $\mathcal{F}$ at a point $Y \in \mathbb{F}^n$, where the bit size of Y_i, $i = 1, \ldots, n$, is at most L'.

We start with analysing the complexity of evaluating a multi-variate polynomial $F(X_1, \ldots, X_n)$ with coefficients of bit-length L at a bigfloat that has bit size L'. To bound this complexity, we bound the number of algebraic operations needed to evaluate $F(X_1, \ldots, X_n)$ and the worst case bit size of the bigfloats appearing in the evaluation.

Let d_i be the maximum degree of X_i in $F(X_1, \ldots, X_n)$. The total number of monomials appearing in $F(X_1, \ldots, X_n)$ is bounded by $O(\prod_{i=1}^{n} d_i)$. Thus the number of algebraic operations needed to evaluate $F(X_1, \ldots, X_n)$ is $O(\prod_{i=1}^{n} d_i)$. To bound the worst-case bit size of the result of evaluation, we observe that the worst-case bit size of the monomial $X_1^{i_1} \cdots X_n^{i_n}$, where $0 \leq i_j \leq d_j$, $1 \leq j \leq n$, is $O(L' \sum_{i=1}^{n} d_i)$. Hence the worst-case bit size of a term $aX_1^{i_1} \cdots X_n^{i_n}$ appearing in $F(X_1, \ldots, X_n)$ is $O(L + L' \sum_{i=1}^{n} d_i)$. Since the number of algebraic operations needed to evaluate $F(X_1, \ldots, X_n)$ is $O(\prod_{i=1}^{n} d_i)$, the worst-case bit size of the result is $O(\prod_{i=1}^{n} d_i (L + L' \sum_{i=1}^{n} d_i))$.

Thus we have the following:

Lemma C.3. *Let $F(X_1, \ldots, X_n) \in \mathbb{Z}[X_1, \ldots, X_n]$ be a multi-variate polynomial with integer coefficients of bit-length L. Let d_i be the maximum degree of X_i in $F(X_1, \ldots, X_n)$. Then the worst case bit-complexity of evaluating F at a bigfloat of bit size L' is*

$$O[(\prod_{i=1}^{n} d_i) M((L + L' \sum_{i=1}^{n} d_i) d_1 \cdots d_n)],$$

where $M(\ell)$ is the complexity of multiplying two integers of bit-length ℓ.

From the above lemma we know that the complexity of evaluating F_i at a bigfloat Y of bit size L' is $O(D_i^n M((L + nD_iL')D_i^n)$ and hence the complexity of evaluating $\mathcal{F}$ at Y is $O(n\mathcal{D}^n M((L + n\mathcal{D}L')\mathcal{D}^n))$ where $\mathcal{D} := \max(D_1, \ldots, D_n)$. Moreover, it follows easily that the complexity of evaluating the Jacobian matrix $J_{\mathcal{F}}(\mathbf{Y})$ is $O(n^2\mathcal{D}^n M((L + n\mathcal{D}L')\mathcal{D}^n))$.

Bibliography

[Abb06] John Abbott. Quadratic Interval Refinement. Presented as a poster in ISSAC'06, 2006.

[Abe73] O. Aberth. Iteration methods for finding all zeros of a polynomial simultaneously. *Mathematics and Computation*, 27:339–344, 1973.

[ACGR01] G. S. Ammar, D. Calvetti, W. B. Gragg, and L. Reichel. Polynomial zero finders based on Szegő polynomials. *J. Computational and Applied Mathematics*, 127:1–16, 2001.

[AG98] Alberto Alesina and Massimo Galuzzi. A new proof of Vincent's theorem. *L'Enseignement Mathémathique*, 44:219–256, 1998.

[AH83] Göltz Alefeld and Jürgen Herzberger. *Introduction to Interval Computations*. Academic Press, New York, 1983.

[Akr78a] A.G. Akritas. A correction on a theorem by Uspensky. *Bull. Soc. Math. Gréce (N.S.)*, 19:278–285, 1978.

[Akr78b] A.G. Akritas. *Vincent's theorem in algebraic manipulation*. PhD thesis, Operations Research Program, North Carolina State University, Raleigh, North Carolina, 1978.

[Akr82] A.G. Akritas. Reflections on a pair of theorems by Budan and Fourier. *Mathematics Magazine*, 55(5):292–298, 1982.

[Akr86] A.G. Akritas. There is no "Uspensky's method". In *Proceedings of the 1986 Symposium on Symbolic and Algebraic Computation*, pages 88–90, Waterloo, Ontario, Canada, 1986.

[Akr89] Alkiviadis G. Akritas. *Elements of Computer Algebra with Applications*. John Wiley Interscience, New York, 1989.

[Bat98] Prashant Batra. Improvement of a convergence condition for the Durand-Kerner iteration. *J. of Comp. and Appl. Math.*, 96:117–125, 1998.

[Bat99] Prashant Batra. *Abschätzungen und Iterationsverfahren für Polynom-Nullstellen*. PhD thesis, Technical University Hamburg-Harburg, 1999.

[BCSS98] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale. *Complexity and Real Computation*. Springer-Verlag, New York, 1998.

[BF00] Dario Andrea Bini and Giuseppe Fiorentino. *Numerical Computation of Polynomial Roots Using MPSolve Version 2.2*. Dipartimento di Matematica, Universitaá di Pisa, Via Bonarroti 2, 56127 Pisa, January 2000. Manual for the Mpsolve package. Available at `ftp://ftp.dm.unipi.it/pub/mpsolve/MPSolve-2.2.tgz`.

[BFM+01] C. Burnikel, S. Funke, K. Mehlhorn, S. Schirra, and S. Schmitt. A separation bound for real algebraic expressions. In *9th ESA*, volume 2161 of *Lecture Notes in Computer Science*, pages 254–265. Springer, 2001. To appear, Algorithmica.

[BOT90] Michael Ben-Or and Prasoon Tiwari. Simple algorithm for approximating all roots of a polynomial with real roots. *J. Complexity*, 6:417–442, 1990.

[BPR03] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*. Algorithms and Computation in Mathematics. Springer, 2003.

[Bre73] Richard P. Brent. *Algorithms for minimization without derivatives*. Prentice Hall, Englewood Cliffs, NJ, 1973.

[Bre76a] Richard P. Brent. Fast multiple-precision evaluation of elementary functions. *J. of the ACM*, 23:242–251, 1976.

[Bre76b] Richard P. Brent. Multiple-precision zero-finding methods and the complexity of elementary function evaluation. In J. F. Traub, editor, *Proc. Symp. on Analytic Computational Complexity*, pages 151–176. Academic Press, 1976.

[CA76] George E. Collins and Alkiviadis G. Akritas. Polynomial real root isolation using Descartes' rule of signs. In R. D. Jenks, editor, *Proceedings of the 1976 ACM Symposium on Symbolic and Algebraic Computation*, pages 272–275. ACM Press, 1976.

[Caj11] Florian Cajori. Historical Note on the Newton-Raphson Method of Approximation. *The American Mathematical Monthly*, 18(2):29–32, February 1911.

[Che94] Pengyuan Chen. Approximate Zeros of Quadratically Convergent Algorithms. *Mathematics of Computation*, 63(207):247–270, July 1994.

[CJ89] George E. Collins and Jeremy R. Johnson. Quantifier elimination and the sign variation method for real root isolation. In *Proc. ACM-SIGSAM Symposium on Symbolic and Algebraic Computation*, pages 264–271, 1989.

[Coh93] Henri Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[CR88] M. Coste and M. F. Roy. Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. of Symbolic Computation*, 5:121–130, 1988.

[Cur87] James H. Curry. On Zero Finding Methods of Higher Order from Data at One Point. *J. of Complexity*, 5:219–237, 1987.

[DAB04] V. Y. Pan D. A. Bini, L. Gemignani. Improved Initialization of the Accelerated and Robust QR-like Polynomial Root-finding. *Electronic Transactions on Numerical Analysis*, 17:195–205, 2004.

[DAB05] V. Y. Pan D. A. Bini, L. Gemignani. Fast and Stable QR Eigenvalue Algorithms for Generalized Companion Matrices and Secular Equation. *Numerische Math.*, 3:373–408, 2005.

[Dav85] J. H. Davenport. Computer algebra for cylindrical algebraic decomposition. Technical report, The Royal Institute of Technology, Department of Numerical Analysis and Computing Science, S-100 44, Stockholm, Sweden, 1985. Reprinted as: Technical Report 88-10, School of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, England.

[Dég00] Jérôme Dégot. A condition number theorem for underdetermined polynomial systems. *Math. Comp.*, 40(233):329–335, 2000.

[Dek67] T.J. Dekker. Finding a Zero by Means of Successive Linear Interpolation. In Bruno Dejon and Peter Henrici, editors, *Constructive Aspects of the Fundamental Theorem of Algebra*, pages 37–48. Wiley Interscience, 1967.

[DF95] Wang Deren and Zhao Fengguang. The theory of Smale's point estimation and its applications. *J. of Comp. and Appl. Math.*, 60:253–269, 1995.

[DSY05] Zilin Du, Vikram Sharma, and Chee Yap. Amortized bounds for root isolation via Sturm sequences. In Dongming Wang and Lihong Zhi, editors, *Proc. Internat. Workshop on Symbolic-Numeric Computation*, pages 81–93, School of Science, Beihang University, Beijing, China, 2005. Int'l Workshop on Symbolic-Numeric Computation, Xi'an, China, Jul 19–21, 2005.

[Dur60] E. Durand. *Solutions Numériques des Équations Algébriques, Tome I: Equations du Type F(x) = 0*. Racines d'un Polynôme, Masson, Paris, 1960.

[Dur77] P. Duren. Coefficients of univalent functions. *Bull. Amer. Math. Soc.*, 83(5):891–911, 1977.

[Dys47] F. J. Dyson. The approximation to algebraic numbers by rationals. *Acta Math. 79*, 1947.

[EM95] Alan Edelman and H. Murakami. Polynomial roots from companion matrix eigenvalues. *Mathematics of Computation*, 64(210):763–776, 1995.

[EMT06] Ioannis Z. Emiris, Bernard Mourrain, and Elias P. Tsigaridas. Real algebraic numbers: Complexity analysis and experimentations. Research Report 5897, INRIA, April 2006. `http://www.inria.fr/rrrt/rr-5897.html`.

[ESY06] Arno Eigenwillig, Vikram Sharma, and Chee Yap. Almost tight complexity bounds for the Descartes method. In *Proc. Int'l Symp. Symbolic and Algebraic Computation (ISSAC'06)*, 2006. Genova, Italy. Jul 9-12, 2006.

[ET06] Ioannis Z. Emiris and Elias P. Tsigaridas. Univariate polynomial real root isolation: Continued fractions revisited. To appear in ESA 2006. Appeared on CS arxiv, Apr. 2006.

[Far90] Gerald Farin. *Curves and Surfaces for Computer Aided Geometric Design: A Practical Guide*. Academic Press, Inc, second edition, 1990.

[For96] Steven Fortune. Robustness Issues in Geometric Algorithms. In *WACG: 1st Workshop on Applied Computational Geometry: Towards Geometric Engineering*. LNCS, March 1996.

[For01] Steve Fortune. Polynomial root finding using an iterated eigenvalue computation. In *ISSAC*, pages 121–128, 2001.

[FR87] R.T. Farouki and V.T. Rajan. On the numerical condition of polynomials in Bernstein form. *Computer Aided Geometric Design*, 4:191–216, 1987.

[FR88] R.T. Farouki and V.T. Rajan. Algorithm for polynomials in Bernstein form. *Computer Aided Geometric Design*, 5:1–26, 1988.

[FvW93] Steven J. Fortune and Christopher J. van Wyk. Efficient exact arithmetic for computational geometry. In *Proc. 9th ACM Symp. on Computational Geom.*, pages 163–172, 1993.

[GT74] W.B. Gragg and R.A. Tapia. Optimal error bounds for the Newton-Kantorovich Theorem. *SIAM Journal of Numerical Analysis*, 11(1), March 1974.

[Hig96] Nicholas J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[HJ85] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[Hof89] Christoff M. Hoffmann. The problems of accuracy and robustness in geometric computation. *IEEE Computer*, 22(3):31–42, March 1989.

[Hon98] Hoon Hong. Bounds for absolute positiveness of multivariate polynomials. *J. of Symbolic Computation*, 25(5):571–585, 1998.

[HSS01] J. H. Hubbard, D. Schleicher, and Scott Sutherland. How to find all roots of complex polynomials by newton's method. *Inventiones Mathematicae*, 146(1):1–33, 2001.

[IEE85] IEEE. IEEE Standard 754-1985 for binary floating-point arithmetic, 1985. ANSI/IEEE Std 754-1985. From The Institute of Electrical and Electronic Engineers, Inc.

[Jac35] C. G. J. Jacobi. Observatiunculae ad theoriam aequationum pertinentes. *Journal für die reine und angewandte Mathematik*, 13:340–352, 1835. Available from `http://gdz.sub.uni-goettingen.de`.

[JKL+06] Jeremy R. Johnson, Werner Krandick, Kevin M. Lynch, David G. Richardson, and Anatole D. Ruslanov. High-performance implementations of the Descartes method.

In *ISSAC '06: Proceedings of the 2006 international symposium on Symbolic and algebraic computation*, pages 154–161, New York, NY, USA, 2006. ACM Press.

[JKR05] Jeremy R. Johnson, Werner Krandick, and Anatole D. Ruslanov. Architecture-aware classical Taylor shift by 1. In *Proc. 2005 International Symposium on Symbolic and Algebraic Computation (ISSAC 2005)*, pages 200–207. ACM, 2005.

[Joh98] J.R. Johnson. Algorithms for polynomial real root isolation. In B.F. Caviness and J.R. Johnson, editors, *Quantifier Elimination and Cylindrical Algebraic Decomposition*, Texts and monographs in Symbolic Computation, pages 269–299. Springer, 1998.

[KA64] L.V. Kantorovich and G.P. Akilov. *Functional Analysis in Normed Spaces*. New York, MacMillan, 1964.

[Kal05] Bahman Kalantari. An infinite family of bounds on zeros of analytic functions and relationship to Smale's bound. *Mathematics of Computation*, 74(250):841–852, 2005.

[Kan52] L.V. Kantorovich. Functional Analysis and Applied Mathematics. Technical Report 1509, National Bureau of Standards, 1952.

[Kea87] R. Baker Kearfott. Abstract Generalized Bisection and a Cost Bound. *Mathematics of Computation*, 49(179):187–202, July 1987.

[Kea90] R. Baker Kearfott. Interval Newton/generalized bisection when there are singularities near roots. *Annals of Operation Research*, 25:181–196, 1990.

[Ker66] I.O. Kerner. Ein Gesamtschrittverfahren zur Berechnung der Nullstellen von Polynomen. *Numer. Math.*, 8:290–294, 1966.

[Khi97] A. Ya. Khinchin. *Continued Fractions*. Dover Publications, 1997.

[Kim86] Myong-Hi Kim. *Computational Complexity of the Euler Type Algorithms for the Roots of polynomials*. PhD thesis, City University of New York, January 1986.

[Kim88] Myong-Hi Kim. On approximate zeroes and root finding algorithms for a complex polynomial. *Math. Comp.*, 51:707–719, 1988.

[Kio86] J. Kioustelidis. Bounds for the positive roots of the polynomials. *Journal of Computational and Applied Mathematics*, 16:241–244, 1986.

[KLPY99] V. Karamcheti, C. Li, I. Pechtchanski, and C. Yap. A Core library for robust numerical and geometric computation. In *15th ACM Symp. Computational Geometry*, pages 351–359, 1999.

[KM06] Werner Krandick and Kurt Mehlhorn. New bounds for the Descartes method. *J. Symbolic Computation*, 41(1):49–66, 2006.

[Kra95] Werner Krandick. Isolierung reeller Nullstellen von Polynomen. In J. Herzberger, editor, *Wissenschaftliches Rechnen*, pages 105–154. Akademie-Verlag, Berlin, 1995.

[Kre74] Erwin Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, 1974.

[KS94] Myong-Hi Kim and Scott Sutherland. Polynomial root-finding algorithms and branched covers. *SIAM J. Computing*, 23:415–436, 1994.

[Lio40] J. Liouville. Sur l'irrationalite du nombre e. *J. Math. Pures Appl.*, 1840.

[LR81] Jeffrey M. Lane and R. F. Riesenfeld. Bounds on a polynomial. *BIT*, 21:112–117, 1981.

[LR01] Thomas Lickteig and Marie-Françoise Roy. Sylvester-Habicht sequences and fast Cauchy index computation. *J. of Symbolic Computation*, 31:315–341, 2001.

[LY01] Chen Li and Chee Yap. A new constructive root bound for algebraic expressions. In *12th SODA*, pages 496–505, January 2001.

[Mah64] K. Mahler. An inequality for the discriminant of a polynomial. *The Michigan Mathematical Journal*, 11(3):257–262, 1964.

[Mal] Gregorio Malajovich. Unitary Invariance of the Kostlan Norm (Linear Algebra Proof). `http://citeseer.ist.psu.edu/206630.html`.

[Mal93] Gregorio Malajovich. *On the complexity of path-following Newton algorithms for solving systems of polynomial equations with integer coefficients*. PhD thesis, Berkeley, 1993.

[Mc99] Maurice Mignotte and Doru Ştefănescu. *Polynomials: An Algorithmic Approach*. Springer, Singapore, 1999.

[McN93] J.M. McNamee. A bibliography on roots of polynomials. *J. Comput. Appl. Math.*, 47:391–394, 1993. Available online at http://www.elsevier.com/homepage/sac/cam/mcnamee.

[Mig81] Maurice Mignotte. Some inequalities about univariate polynomials. In *Proc. 1981 ACM Symposium on Symbolic and Algebraic Computation (SYMSAC 1981)*, pages 195–199. ACM, 1981.

[Mig95] Maurice Mignotte. On the distance between the roots of a polynomial. *Applicable Algebra in Engineering, Commun., and Comput.*, 6:327–332, 1995.

[Mil92] Philip S. Milne. On the solutions of a set of polynomial equations. In B. R. Donald, D. Kapur, and J. L. Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 89–102. Academic Press, London, 1992.

[Mit91] D.P. Mitchell. Robust ray intersection with interval arithmetic. In *Graphics Interface*, pages 68–74. 1991.

[Moo66] Ramon E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1966.

[MRR04] Bernard Mourrain, Fabrice Rouillier, and Marie-Françoise Roy. Bernstein's basis and real root isolation. Rapport de recherche 5149, INRIA-Rocquencourt, March 2004. `http://www.inria.fr/rrrt/rr-5149.html`.

[MRR05] Bernard Mourrain, Fabrice Rouillier, and Marie-Françoise Roy. The Bernstein basis and real root isolation. In Jacob E. Goodman, János Pach, and Emo Welzl, editors, *Combinatorial and Computational Geometry*, number 52 in MSRI Publications, pages 459–478. Cambridge University Press, 2005.

[MS00] K. Mehlhorn and S. Schirra. A generalized and improved constructive separation bound for real algebraic expressions. Technical Report MPI-I-2000-004, Max-Planck-Institut für Informatik, November 2000.

[MV95] F. Malek and R. Vaillancourt. A composite polynomial zerofinding matrix algorithm. *Computers and Mathematics with Applications*, 30(2):37–47, July 1995.

[MVY02] B. Mourrain, M. N. Vrahatis, and J. C. Yakoubsohn. On the complexity of isolating real roots and computing with certainty the topological degree. *J. Complexity*, 18:612–640, 2002.

[NR94] C. Andrew Neff and John H. Reif. An $O(n^{1+\epsilon} \log b)$ algorithm for the complex roots problem. *IEEE Foundations of Computer Science*, 1994.

[OR70] J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, 1970.

[Ort68] J.M. Ortega. The Newton-Kantarovich Theorem. *The American Mathematical Monthly*, 75:658–660, June-July 1968.

[Ost50] A.M. Ostrowski. Note on Vincent's theorem. *The Annals of Mathematics*, 52(3):702–707, Nov 1950.

[Ost60] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, 1960.

[Ost73] A.M. Ostrowski. *Solution Of Equations In Euclidean And Banach Spaces*. Pure and Applied Mathematics. Academic Press, third edition, 1973.

[Pan96] Victor Y. Pan. Optimal and nearly optimal algorithms for approximating polynomial zeros. *Computers Mathematics and Applications*, 31(12):97–138, 1996.

[Pan97] Victor Y. Pan. Solving a polynomial equation: some history and recent progress. *SIAM Review*, 39(2):187–220, 1997.

[Pan02] Victor Y. Pan. Univariate polynomial root-finding with a lower computational precision and higher convergence rates, 2002.

[PBP02] Hartmut Prautzsch, Wolfgang Boehm, and Marco Paluszny. *Bézier and B-Spline Techniques*. Springer, 2002.

[PCT95] Miodrag S. Petković, Carsten Carstensen, and Miroslav Trajkovíc. Weierstrass formula and zero-finding methods. *Numer. Math.*, 69:353–372, 1995.

[Ped91] Paul Pedersen. *Counting Real Zeros*. PhD thesis, New York University, 1991. Also, Courant Institute Computer Science Technical Report 545 (Robotics Report R243).

[PHI98] Miodrag S. Petković, Dorde Herceg, and Snežana Ilić. Safe convergence of simultaneous methods for polynomial zeros. *Numerical Algorithms*, 17:313–331, 1998.

[PMR+06] Victor Y. Pan, Brian Murphy, Rhys Eric Rosholt, Dmitriy Ivolgin, and Yuqing Tang. Root-finding with Eigen-solving. Technical report, CUNY Ph.D. Program in Computer Science, 2006. http://www.cs.gc.cuny.edu/tr/techreport.php?id=185.

[PY03] Sylvain Pion and Chee Yap. Constructive root bound method for k-ary rational input numbers. In *19th SCG*, pages 256–263, San Diego, California., 2003. Accepted, Theoretical Computer Science (2006).

[Rei97] Daniel Reischert. Asymptotically fast computation of subresultants. In *ISSAC 97*, pages 233–240, 1997. Maui, Hawaii.

[Rot55] K.F. Roth. Rational approximations to algebraic numners. *Mathematika 2*, 1955.

[RZ01] Fabrice Rouillier and Paul Zimmermann. Efficient isolation of a polynomial['s] real roots. Rapport de Recherche 4113, INRIA, 2001. http://www.inria.fr/rrrt/rr-4113.html.

[RZ04] Fabrice Rouillier and Paul Zimmerman. Efficient isolation of [a] polynomial's real roots. *J. Computational and Applied Mathematics*, 162:33–50, 2004.

[Sch80] A. Schönhage. Storage modification machines. *SIAM J. Computing*, 9:490–508, 1980.

[Sch82] Arnold Schönhage. The fundamental theorem of algebra in terms of computational complexity, 1982. Manuscript, Department of Mathematics, University of Tübingen.

[Sch99] Stefan Schirra. Robustness and precision issues in geometric computation. In J.R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*. Elsevier Science Publishers, B.V. North-Holland, Amsterdam, 1999.

[Sch05] Susanne Schmitt. The diamond operator – implementation of exact real algebraic numbers. In V.G.Ganzha, E.W.Mayr, and E.V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing: 8th International Workshop (CASC 2005)*, pages 355–366, 2005. Kalamata, Greece. Sep 12-16, 2005.

[SDY05] Vikram Sharma, Zilin Du, and Chee Yap. Robust approximate zeros. In Gerth Stølting Brodal and Stefano Leonardi, editors, *Proc. 13th European Symp. on Algorithms (ESA)*, volume 3669 of *Lecture Notes in Computer Science*, pages 874–887. Springer-Verlag, April 2005. Palma de Mallorca, Spain, Oct 3-6, 2005.

[Sek98] Hiroshi Sekigawa. Using interval computation with the Mahler measure for zero determination of algebraic numbers. *Josai Information Sciences Researches*, 9(1):83–99, 1998.

[Sma81a] S. Smale. The fundamental theorem of algebra and complexity theory. *Bull. Amer. Math, Soc.*, 4:1–36, 1981.

[Sma81b] Steve Smale. The fundamental theorem of algebra and complexity theory. *Bull. Amer. Math. Soc. (N.S.)*, 4(1):1–36, 1981.

[Sma85] Steve Smale. On the efficiency of algorithms of analysis. *Bull. Amer. Math. Soc. (N.S.)*, 13(2):87–121, October 1985.

[Sma86] S. Smale. Newton's method estimates from data at one point. In R. Ewing, K. Gross, and C. Martin, editors, *The Merging of Disciplines: New Directions in Pure, Applied, and Computational Mathematics*. Springer-Verlag, 1986.

[Spe94] Melvin R. Spencer. *Polynomial Real Root Finding in Bernstein Form*. PhD thesis, Brigham Young University, 1994.

[SS85] Mike Shub and Steven Smale. Computational complexity. on the geometry of polynomials and a theory of cost. i. *Annales Scientifiques De L'É.N.S*, 18(1):107–142, 1985.

[SS86] M. Shub and S. Smale. Computational complexity: On the geometry of polynomials and a theory of cost: Ii. *SIAM J. of Comput.*, 15(1):145–161, February 1986.

[SS93a] Mike Shub and Steve Smale. Complexity of Bezout's Theorem I: Geometric aspects. *J. of Amer. Math. Soc.*, 6(2):459–501, 1993.

[SS93b] Mike Shub and Steve Smale. Complexity of Bezout's Theorem III: Condition number and packing. *J. of Complexity*, 9:4–14, 1993.

[SS96] Michael Shub and Steve Smale. Complexity of Bezout's theorem. IV. probability of success and extensions. *SIAM Journal on Numerical Analysis*, 33(1):128–148, 1996.

[Şte05] D. Ştefănescu. New bounds for the positive roots of polynomials. *Journal of Universal Computer Science*, 11(12):2132–2141, 2005.

[TB97] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

[Thu09] A. Thue. Uber Annaherungswerte algebraischer Zahlen. *J. Reine Angew. Math. 135*, 1909.

[Tis01] Françoise Tisseur. Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems. *SIAM J. on Matrix Anal. and Appl.*, 22(4):1038–1057, 2001.

[Tur84] P. Turan. *On a New Method of Analysis and its Applications*. Wiley, New Jersey, 1984.

[Usp48] J. V. Uspensky. *Theory of Equations*. McGraw-Hill, New York, 1948.

[vdS70] A. van der Sluis. Upper bounds for roots of polynomials. *Numer. Math.*, 15:250–262, 1970.

[Vin36] A.J.H. Vincent. Sur la résolution des équations numériques. *J. Math. Pures Appl.*, 1:341–372, 1836.

[vzGG97] Joachim von zur Gathen and Jürgen Gerhard. Fast algorithms for Taylor shifts and certain difference equations. In *Proc. 1997 International Symposium on Symbolic and Algebraic Computation (ISSAC 1997)*, pages 40–47. ACM, 1997.

[vzGG99] Joachim von zur Gathen and Jürgen Gerhard. *Modern computer algebra*. Cambridge University Press, Cambridge, 1999.

[Wan04] Xiaoshen Wang. A Simple Proof of Descartes' Rule of Signs. *The American Mathematical Monthly*, June-July 2004.

[Wer82] W. Werner. On the simultaneous determination of polynomial roots. Number 953 in Lecture Notes in Mathematics, pages 188–202. Springer-Verlag, Berlin, 1982.

[Wey24] H. Weyl. Randbemerkungen zu hauptproblemen der mathematik, ii, fundamentalsatz der algebra and grundalgen der mathematik. *Mathematics Zahlen*, 20:131–151, 1924.

[Wil63] J. H. Wilkinson. *Rounding Errors in Algebraic Processes*. Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963. Also published by Prentice-Hall, Englewood Cliffs, NJ, USA. Reprinted by Dover, New York, 1994.

[Wil78] H.S Wilf. A global bisection algorithm for computing the zeros of polynomials in the complex plane. *Journal of the ACM*, 25:415–420, 1978.

[Yam85] T. Yamamoto. A unified derivation of several error bounds for Newton's process. *Journal of Comp. and Appl. Mathematics*, 12-13:179–191, 1985.

[Yam86] T. Yamamoto. Error bounds for Newton's method under the Kantorovich assumptions. In R. Ewing, K. Gross, and C. Martin, editors, *The Merging of Disciplines: New Directions in Pure, Applied, and Computational Mathematics*. Springer-Verlag, 1986.

[Yap97a] Chee K. Yap. Robust geometric computation. In Jacob E. Goodman and Joseph O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, chapter 35, pages 653–668. CRC Press LLC, Boca Raton, FL, 1997.

[Yap97b] Chee K. Yap. Towards exact geometric computation. *Comput. Geometry: Theory and Appl.*, 7:3–23, 1997.

[Yap00] Chee K. Yap. *Fundamental Problems of Algorithmic Algebra*. Oxford University Press, 2000.

[Yap04] Chee K. Yap. Robust geometric computation. In Jacob E. Goodman and Joseph O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, chapter 41, pages 927–952. Chapmen & Hall/CRC, Boca Raton, FL, 2nd edition, 2004.

[YD95] Chee K. Yap and Thomas Dubé. The exact computation paradigm. In D.-Z. Du and F. K. Hwang, editors, *Computing in Euclidean Geometry*, pages 452–492. World Scientific Press, Singapore, 2nd edition, 1995.

[YLP+04] C. Yap, C. Li, S. Pion, Z. Du, and V. Sharma. Core library tutorial: a library for robust geometric computation, 1999–2004. Version 1.1 was released in Jan 1999. Latest Version 1.6 (Jun 2003). Download source and documents, `http://cs.nyu.edu/exact/`.

[YM01] Chee Yap and Kurt Mehlhorn. Towards robust geometric computation, 2001. Invited White Paper. CSTB-NSF Conference on Fundamentals of Computer Science, Washington DC, July 25-26, 2001. See Appendix, **Computer Science: Reflections on/from the Field**, The National Academies Press, Washington DC, 2004.

[Ypm95] Tjalling J. Ypma. Historical Development of the Netwon-Raphson method. *SIAM Review*, 37(4):531–551, December 1995.