A Numerical Algorithm for Zero Counting. I: Complexity and Accuracy

Felipe Cucker *
Dept. of Mathematics
City University of Hong Kong
HONG KONG
e-mail: macucker@cityu.edu.hk

Teresa Krick †
Departamento de Matemática
Univ. de Buenos Aires & CONICET
ARGENTINA
e-mail: krick@dm.uba.ar

Gregorio Malajovich‡
Depto. de Matemática Aplicada
Univ. Federal do Rio de Janeiro
BRASIL
e-mail: gregorio@ufrj.br

Mario Wschebor
Centro de Matemática
Universidad de la República
URUGUAY
e-mail: wschebor@cmat.edu.uy

Abstract. We describe an algorithm to count the number of distinct real zeros of a polynomial (square) system f. The algorithm performs $\mathcal{O}(\log(n\mathbf{D}\kappa(f)))$ iterations (grid refinements) where n is the number of polynomials (as well as the dimension of the ambient space), $\mathbf{D}$ is a bound on the polynomials' degree, and $\kappa(f)$ is a condition number for the system. Each iteration uses an exponential number of operations. The algorithm uses finite-precision arithmetic and a major feature in our results is a bound for the precision required to ensure the returned output is correct which is polynomial in n and $\mathbf{D}$ and logarithmic in $\kappa(f)$. The algorithm parallelizes well in the sense that each iteration can be computed in parallel time polynomial in n, $\log\mathbf{D}$ and $\log(\kappa(f))$.

1 Introduction

In recent years considerable attention was put on the complexity of counting problems over the reals. The counting complexity class $\#\mathrm{P}_{\mathbb{R}}$ was introduced [20] and completeness results for $\#\mathrm{P}_{\mathbb{R}}$ were established [3] for natural geometric problems notably, for the computation of the Euler characteristic of semialgebraic sets. As one could expect, the "basic" $\#\mathrm{P}_{\mathbb{R}}$-complete problem consists of counting the real zeros of a system of polynomial equations.

Algorithms for counting real zeros have existed since long. One such algorithm follows from the work of Tarski [25] on quantifier elimination for the theory of the reals. Its complexity is hyperexponential. Algorithms with improved complexity (doubly exponential) were devised in the 70s by Collins [5] and Wütrich [27]. A breakthrough was reached a decade

*Partially supported by City University SRG grant 7002106.
†Partially supported by grants UBACyT X112/06-09, CONICET PIP 2461/00 and ANPCyT 33671/05.
‡Partially supported by CNPq grants 304504/2004-1, 472486/2004-7, 470031/2007-7, 303565/2007-1, and by FAPERJ grant E26/170.734/2004.

later with the introduction of the critical points method by Grigoriev and Vorobjov [13, 12] which uses exponential time. Algorithms counting connected components (and hence, in the zero-dimensional case, solutions) based on this method can be found in [14, 16], and in the straight-line program model of computation in [1]. These algorithms parallelise well in the sense that one can devise versions of them working in parallel polynomial time when an exponential number of processors is available. The $\#P_{\mathbb{R}}$-completeness of the problem strongly indicates that this is the best we can hope for.

All the algorithms mentioned above are "symbolic algorithms." They have been devised upon the premise that no perturbation or round-off error is present. Were this not the case, it is not difficult to see that errors would accumulate quite badly. Roughly speaking, these algorithms construct some object of exponential size on which some basic computation (e.g., linear algebra) is eventually performed. A question is posed, can one devise "numerical algorithms" (maybe iterative, which need not terminate for ill-posed inputs) with a better behavior viz the accumulation of round-off errors? For the problem of deciding the existence of (or computing) a zero of a polynomial system such algorithms were given in [8, 6, 18]. The goal of this article is to describe and analyze a numerical algorithm for zero counting. We will do so by developping appropriate versions of the tools used in [8, 6].

Let $d_1, \dots, d_n \in \mathbb{N}$ and $\mathbf{d} = (d_1, \dots, d_n)$. We will denote by $\mathcal{H}_{\mathbf{d}}$ the space of polynomial systems $f = (f_1, \dots, f_n)$ with $f_i \in \mathbb{R}[X_0, \dots, X_n]$ homogeneous of degree d_i.

Zero rays of polynomial systems $f \in \mathcal{H}_{\mathbf{d}}$ are associated to pairs of zeros $(-\zeta, \zeta)$ of the restriction $f_{|S^n}$ of f to the n-dimensional unit sphere $S^n \subset \mathbb{R}^{n+1}$. Thus, it will be convenient to consider a system $f \in \mathcal{H}_{\mathbf{d}}$ as a (central symmetric, analytic) mapping of S^n into $\mathbb{R}^n$. If we denote by $Z(f) = \{\zeta \in S^n : f(\zeta) = 0\}$ the zero-set of f in S^n then the number $\#_{\mathbb{R}}(f)$ of zero rays of the system f is half the cardinality of $Z(f)$.

In this paper we describe a finite-precision algorithm computing $\#_{\mathbb{R}}(f)$, given $f \in \mathcal{H}_{\mathbf{d}}$. To analyze its complexity and accuracy, besides the number n of polynomials, we will rely on two more additional parameters. One is $\mathbf{D} = \max_{i \le n} d_i$. The other is a condition measure $\kappa(f)$ for the system f. We will describe this measure in detail in Section 2 below. We will also let $S = \max S_i$ where S_i is the number of non-zero coefficients of f_i. Note that S is bounded by a simple expression in terms of n and $\mathbf{D}$, namely, $S = \binom{n+\mathbf{D}}{\mathbf{D}}$. Yet, we will express dependancy on S since this may be relevant for the case of sparse systems of polynomials. Our main result is the following.

Theorem 1.1. *There exists an iterative algorithm which, with input $f \in \mathcal{H}_{\mathbf{d}}$,*

(1) *Returns $\#_{\mathbb{R}}(f)$.*

(2) *Performs $\mathcal{O}(\log(n\mathbf{D}\kappa(f)))$ iterations and has a total cost (number of arithmetic operations) of*

$$\mathcal{O}\left(\log(n\mathbf{D}\kappa(f))(n+1)^2\left(\frac{2(n+1)\mathbf{D}^2\kappa(f)^2}{\alpha_*}\right)^{2n}\right),$$

where $\alpha_* \approx 0.0384629388\ldots$ is a universal constant.

(3) *Can be well-parallelized in the sense that it admits a parallel version running in time*

$$\mathcal{O}(n^2\ln(n\mathbf{D}\kappa(f))(\ln(n\mathbf{D}\kappa(f))^2 + \ln(\alpha_*)^2))$$

with a number of processors exponential in this quantity.

(4) *Can be implemented with finite precision (both versions, sequential and parallel). The running time remains the same (with α_* replaced by $\alpha_\bullet \approx 0.028268\cdots$) and the returned value is $\#_{\mathbb{R}}(f)$ as long as the machine precision (i.e., the round-off unit) u satisfies*

$$u \leq \frac{1}{\mathcal{O}\left(\mathbf{D}^2 n^{5/2}\kappa(f)^3(\log S + n^{3/2}\mathbf{D}^2\kappa(f)^2)\right)}.$$

(5) *It can be modified to return, in addition and for each real zero $\zeta \in S^n$ of f, an approximate zero x of f in the sense that Newton's iteration, starting at x, converges to ζ quadratically fast.*

Remark 1.2. (i) A system f for which arbitrarily small perturbations may change the value $\#_{\mathbb{R}}(f)$ is considered *ill-posed* in our context since for arbitrarily small machine precisions finite precision algorithms may return an incorrect value. Consequently, the condition number $\kappa(f)$ is infinite in these cases (and only then). This happens when f has multiple real zeros and, in particular, when f has infinitely many real zeros. In these cases the algorithm of Theorem 1.1 may not halt.

(ii) Numerical algorithms compute functions φ on real data. Error analysis for algorithms computing (vectors of) real numbers —i.e. for which the image of φ has non-empty interior— are usually expressed in terms of bounds for the relative error of the computed quantities. That is, for data d, bounds in

$$\frac{\|\varphi(d) - \mathtt{fl}(\varphi(d)\|}{\|\varphi(d)\|}$$

where $\mathtt{fl}(\varphi(d))$ is the vector actually computed with finite precision. This relative error varies continuously with d and depends on the condition of d and on the precision u. Such a form of analysis, however, becomes meaningless when computing quantities taking a finite number of values. Indeed, if R_a denotes the set of input data d for which $\varphi(d) = a$ the following happens. When d is in the interior of R_a we have that the relative error above is 0 for sufficiently small u. In contrast, when d is on the boundary of R_a, that error may remain constant for all $u > 0$. Because of this, error analysis for this kind of discrete-valued problems has a different form, as in Theorem 1.1. One bounds how small u needs to be to guarantee a correct answer. Such a bound, needless to say, also depends on the condition of the data d. Examples of this type of analysis can be found in [4, 6, 7, 8]. In each of these references a condition number for the problem at hand occurs in the error analysis. We note that the one in [6] is essentially our $\kappa(f)$.

The rest of the paper is organized as follows. In Section 2 we describe the basic objects we will deal with as well as fixing the notation. In Sections 3 and 4 we prove the two technical results our algorithm relies on. In Section5 we describe the algorithm under the assumption of infinite precision and we prove parts (1), (2), and (3) of Theorem 1.1. The geometric ideas making the algorithm work are best seen in this context. Section 6 then describes the necessary modifications to make the algorithm work as well under finite precision. These modifications are simple and can be summarized by saying that we relax a bit the inequalities tested in the algorithms to make room for the finite-precision errors to fit in.

2 Preliminaries

Denote by $\mathcal{H}_d$ the subspace of $\mathbb{R}[X_0, \ldots, X_n]$ of homogeneous polynomials of degree d. Then, $\mathcal{H}_{\mathbf{d}} = \mathcal{H}_{d_1} \times \cdots \times \mathcal{H}_{d_n}$.

If $g \in \mathcal{H}_d$ we write

$$g(X) = \sum_J g_J X^J$$

where $J = (J_0, \dots, J_n)$ is assumed to range over all multi-indices such that $|J| = \sum_{k=0}^n J_k = d$, $X^J = X_0^{J_0} X_1^{J_1} \dots X_n^{J_n}$ and $g_J \in \mathbb{R}$. Multinomial coefficients are defined by:

$$\binom{d}{J} = \frac{d!}{J_0! J_1! \cdots J_n!}.$$

The space $\mathcal{H}_d$ is endowed with the inner product

$$\langle g, h \rangle = \sum_{|J|=d} \frac{g_J h_J}{\binom{d}{J}}$$

which gives rise to the norm $\|g\| = \sqrt{\langle g, g \rangle}$. These norms, for $d_1, \dots, d_n$, induce a norm in $\mathcal{H}_{\mathbf{d}}$ by taking for $f = (f_1, \dots, f_n) \in \mathcal{H}_{\mathbf{d}}$:

$$\|f\| = \|(f_1, \dots, f_n)\| = \max_{1 \le i \le n} \|f_i\|.$$

Let $O(n+1)$ be the orthogonal group. The inner product above is known to be $O(n+1)$-invariant: for all $Q \in O(n+1)$ and all $g, h \in \mathcal{H}_d$,

$$\langle g \circ Q, h \circ Q \rangle = \langle g, h \rangle.$$

(This is a direct consequence of [26, III-7] or [2, Theorem 1 p. 218], by considering $O(n+1)$ as subgroup of $U(n+1)$). The associated norm $\|f\|$ on $\mathcal{H}_{\mathbf{d}}$ is therefore also $O(n+1)$-invariant. We will use this norm on $\mathcal{H}_{\mathbf{d}}$ all along this paper. For $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ we recall that $\|x\|_2 = (x_1^2 + \cdots + x_n^2)^{1/2}$ and $\|x\|_\infty = \max\{|x_1|, \dots, |x_n|\}$. We will often denote $\|x\|_2$ simply by $\|x\|$.

For $f \in \mathcal{H}_{\mathbf{d}}$ and $x \in S^n$ define

$$\mu_{\text{norm}}(f, x) = \|f\| \sqrt{n} \left\| Df(x)_{|T_x S^n}^{-1} \begin{bmatrix} \sqrt{d_1} & & & \\ & \sqrt{d_2} & & \\ & & \ddots & \\ & & & \sqrt{d_n} \end{bmatrix} \right\| \tag{1}$$

where $Df(x)_{|T_x S^n}$ is the restriction to the tangent space of x at S^n of the derivative of f at x and the norm is the spectral norm, i.e. the operator norm with respect to $\| \; \|_2$. We now define the *condition number* $\kappa(f)$ of $f \in \mathcal{H}_{\mathbf{d}}$:

$$\kappa(f) = \max_{x \in S^n} \min \left\{ \mu_{\text{norm}}(f, x), \frac{\|f\|}{\|f(x)\|_\infty} \right\}.$$

Remark 2.1. The quantity $\kappa(f)$ is closely related to other condition numbers for similar problems.

A version of the quantity $\mu_{\text{norm}}(f, \zeta)$ was introduced in [21, 22, 23] (see also [2, Chapter 12]) for a complex polynomial system f and a zero ζ of f in the complex unit sphere $S^n_{\mathbb{C}} \subset \mathbb{C}^{n+1}$. The *normalized condition number* of such a system f was then defined to be

$$\mu_{\text{norm}}(f) := \max_{\zeta \in S^n_{\mathbb{C}} \mid f(\zeta)=0} \mu_{\text{norm}}(f, \zeta). \tag{2}$$

Actually, the version of $\mu_{\rm norm}(f,\zeta)$ introduced in [21, 22, 23] differs from (1) in the fact that $\|f\|$ is defined as $(\sum \|f_i\|^2)^{1/2}$ (and there is no $\sqrt{n}$ factor). It is bounded above by the expression in (1).

Over the reals, the right-hand side in (2) may not be well-defined since the zero set of f may be empty. In [8] real systems were considered (as in the present paper) and an algorithm deciding feasibility of f (i.e., whether f has a real zero) was proposed. Its complexity was analyzed in terms of a condition number which, using our notation and modulo minor details, is defined as follows

$$\begin{cases} \displaystyle\min_{\zeta\in S^n \mid f(\zeta)=0} \mu_{\rm norm}(f,\zeta) & \text{if } f \text{ is feasible} \\ \displaystyle\max_{\zeta\in S^n} \frac{\|f\|}{\|f(\zeta)\|_\infty} & \text{if } f \text{ is infeasible.} \end{cases}$$

Note the use of min (instead of max) in the first line above. This is due to the fact that the time needed for the algorithm in [8] to detect the existence of a zero depends on the best conditioned zero of f. The existence of other, poorly conditioned (or even singular), zeros of f is irrelevant.

Shortly after, the algorithm in [8] was extended to an algorithm which would, in addition and if f is feasible, return a zero of f [6]. The complexity of this extension was studied in terms of a condition number (denoted $\varrho(f)$ in [6]) which, essentially, coincides with our $\kappa(f)$.

Proposition 2.2. *For all $f\in\mathcal{H}_{\mathbf{d}}$, $\kappa(f)\geq 1$.*

PROOF. Let $x\in S^n$. Because of orthogonal invariance, we may assume without loss of generality that $x=e_0:=(1,0,\dots,0)$.

It is then immediate that $\|f(x)\|_\infty\leq\|f\|$. This shows that the second expression in the definition of κ is at least 1.

For the first expression, i.e., $\mu_{\rm norm}(f,x)$, define $g=(g_1,\dots,g_n)\in\mathcal{H}_{\mathbf{d}}$ by $g_i(X)=f_i(X)-f_i(e_0)X_0^{d_i}$. Then $g(e_0)=0$ and [2, Corollary 3 p. 234], $\mu_{\rm norm}(g,e_0)\geq 1$ (this is shown for the version of $\mu_{\rm norm}$ with the 2-norm for $\|f\|$, which is bounded above by the expression (1)). Since $Df(e_0)=Dg(e_0)$ and $\|g\|\leq\|f\|$, we can conclude $\mu_{\rm norm}(f,e_0)\geq\mu_{\rm norm}(g,e_0)\geq 1$. $\square$

3 The exclusion Lemma

In this article, $d(\ ,\ )$ denotes the Riemannian (angular) distance in S^n (which satisfies $0\leq d(x,y)\leq\pi$, $\forall\, x,y\in S^n$) and for $x\in S^n, r>0$, we set $B(x,r):=\{y\in S^n : d(y,x)<r\}$ and $\overline{B}(x,r):=\{y\in S^n : d(y,x)\leq r\}$.

The following result can be used to support an exclusion test.

Lemma 3.1. *Let $f\in\mathcal{H}_{\mathbf{d}}$ and let $x,y\in S^n$ such that $d(x,y)\leq\sqrt{2}$. Then,*

$$\|f(x)-f(y)\|_\infty\leq\|f\|\sqrt{\mathbf{D}}\ d(x,y)$$

In particular, if $f(x)\neq 0$, there is no zero of f in $B(x,\min\{\|f(x)\|_\infty/(\|f\|\sqrt{\mathbf{D}}),\sqrt{2}\})$.

PROOF. An immediate consequence of the definition of the $O(n+1)$-invariant inner product is that $\mathcal{H}_d$ endowed with this inner product is a reproducing kernel Hilbert space [9, Prop. 2.21]. This implies that, for all $g \in \mathcal{H}_d$ and $x \in \mathbb{R}^{n+1}$,

$$g(x) = \langle g(X), (x^T X)^{\deg g} \rangle. \tag{3}$$

Because of orthogonal invariance, we can assume that $x = \mathrm{e}_0$ and $y = \mathrm{e}_0 \cos\theta + \mathrm{e}_1 \sin\theta$, where $\theta = d(x,y)$. Equation (3) implies that

$$\begin{aligned} f_i(x) - f_i(y) &= \langle f_i(X), (x^T X)^{d_i} \rangle - \langle f_i(X), (y^T X)^{d_i} \rangle = \langle f_i(X), (x^T X)^{d_i} - (y^T X)^{d_i} \rangle \\ &= \langle f_i(X), X_0^{d_i} - (X_0 \cos\theta + X_1 \sin\theta)^{d_i} \rangle. \end{aligned}$$

Hence, Cauchy-Schwarz-Bunyakowsky implies:

$$|f_i(x) - f_i(y)| \le \|f_i\| \, \|X_0^{d_i} - (X_0 \cos\theta + X_1 \sin\theta)^{d_i}\|.$$

Since

$$X_0^{d_i} - (X_0 \cos\theta + X_1 \sin\theta)^{d_i} = X_0^{d_i}(1 - (\cos\theta)^{d_i}) + \sum_{k=1}^{d_i} \binom{d_i}{k} (\cos\theta)^{d_i - k} (\sin\theta)^k X_0^{d_i - k} X_1^k,$$

we have:

$$\begin{aligned} \|X_0^{d_i} - (X_0 \cos\theta + X_1 \sin\theta)^{d_i}\|^2 &= (1 - (\cos\theta)^{d_i})^2 + \sum_{k=1}^{d_i} \binom{d_i}{k} (\cos\theta)^{2(d_i - k)} (\sin\theta)^{2k} \\ &= (1 - (\cos\theta)^{d_i})^2 + 1 - (\cos\theta)^{2d_i} \\ &= 2(1 - (\cos\theta)^{d_i}) \\ &\le 2(1 - (1 - \frac{\theta^2}{2})^{d_i}) \qquad (4) \\ &\le 2(1 - (1 - d_i \frac{\theta^2}{2})) \qquad (5) \\ &\le d_i \theta^2, \end{aligned}$$

where the inequality in line (4) is obtained from Taylor expanding $\cos\theta$ around 0, and the inequality in line (5) is due to the fact that $(1-a)^d \ge 1 - da$ for $a \le 1$.
We conclude that

$$|f_i(x) - f_i(y)| \le \|f_i\| \, \theta \sqrt{d_i}$$

and hence

$$\|f(x) - f(y)\|_\infty \le \|f\| \, \theta \sqrt{\max_i d_i}.$$

For the second assertion, we have

$$\begin{aligned} \|f(y)\|_\infty &\ge \|f(x)\|_\infty - \|f(x) - f(y)\|_\infty \\ &\ge \|f(x)\|_\infty - \|f\| \sqrt{\mathbf{D}} \, d(x,y) \qquad \text{since } d(x,y) \le \sqrt{2} \\ &> \|f(x)\|_\infty - \|f\| \sqrt{\mathbf{D}} \, \|f(x)\|_\infty / (\|f\| \sqrt{\mathbf{D}}) \quad = \quad 0. \end{aligned}$$

$\square$

4 The proximity Theorem

4.1 Newton and Smale

Newton iteration on the sphere S^n is defined by

$$\begin{array}{rccl} N_f: & S^n & \to & S^n \\ & x & \mapsto & N_f(x) = \exp_x\left(-Df(x)^{-1}_{|T_x S^n} f(x)\right) \end{array}$$

where $\exp_x$ is the exponential map at x,

$$\exp_x h = \cos(\|h\|)x + \frac{\sin(\|h\|)}{\|h\|} h.$$

Furthermore, the standard invariants of α-theory, introduced by Smale in [24], can be defined as:

$$\begin{aligned} \beta(f,x) &= \left\| Df(x)^{-1}_{|T_x S^n} f(x) \right\|, \\ \gamma(f,x) &= \sup_{k\geq 2} \left\| \frac{Df(x)^{-1}_{|T_x S^n} D^k f(x)_{|(T_x S^n)^k}}{k!} \right\|^{1/(k-1)}, \\ \alpha(f,x) &= \beta(f,x)\gamma(f,x). \end{aligned}$$

Remark 4.1.

(i) It is easy to see that $\beta(f,x) = d(x, N_f(x))$.

(ii) We will not use Newton's method in our algorithm. We are instead interested in its alpha theory which guarantees existence of zeros near points x with $\alpha(f,x)$ small enough.

(iii) The Newton iteration presented above is not the iteration known as 'projective Newton'. There is an alpha theory for that method, available in [19].

Here we use slight modifications of the quantities α, β and γ, more adapted to our purposes. We set

$$\begin{aligned} \overline{\beta}(f,x) &:= \mu_{\mathrm{norm}}(f,x)\frac{\|f(x)\|_\infty}{\|f\|} \\ \overline{\gamma}(f,x) &:= \frac{\mathbf{D}^{3/2}}{2}\mu_{\mathrm{norm}}(f,x) \\ \overline{\alpha}(f,x) &:= \overline{\beta}(f,x)\overline{\gamma}(f,x). \end{aligned}$$

The definition of $\overline{\gamma}$ is motivated by the estimate of γ [2, Theorem 2 p. 267].

$$\gamma(f,x) \leq \overline{\gamma}(f,x).$$

which yields the lower bound

$$\kappa(f) \geq \max_{\zeta | f(\xi)=0} 2\mathbf{D}^{-3/2}\gamma(f,\zeta). \tag{6}$$

We also observe that $\overline{\gamma}(f,x) \geq \frac{\mathbf{D}^{3/2}}{2}$ since $\mu_{\rm norm}(f,x) \geq 1$ and that $\beta(f,x) \leq \overline{\beta}(f,x)$ since

$$\beta(f,x) = \left\| Df(x)^{-1}_{|T_xS^n} f(x) \right\| \leq \sqrt{n}\|f(x)\|_\infty \left\| Df(x)^{-1}_{|T_xS^n} \right\| \leq \mu_{\rm norm}(f,x)\frac{\|f(x)\|_\infty}{\|f\|} = \overline{\beta}(f,x).$$

Therefore $\alpha(f,x) \leq \overline{\alpha}(f,x)$.

4.2 Proximity and unicity from data at a point

Definition 4.2. We say that $x \in S^n$ is an *approximate zero* for f if and only if the Newton sequence $\{x_k\}_{k\in\mathbb{N}}$, where $x_0 := x$ and $x_{k+1} := N_f(x_k)$, is defined for all k and moreover

$$d(x_k, x_{k+1}) \leq \left(\frac{1}{2}\right)^{2^k-1} d(x_0, x_1).$$

The limit point $\zeta = \lim_{k\to\infty} x_k$ is a fixed point for Newton iteration and a zero of f. It is called the *associated zero* to x.

In what follows we denote $\sigma := \sum_{k\geq 0} 2^{-2^k+1} = 1.632843018\ldots$ and we set

$$\overline{B}_f(x) := \{y \in S^n \mid d(x,y) \leq \sigma\overline{\beta}(f,x)\}.$$

The main technical tool in our algorithm is provided by the following result.

Theorem 4.3. *There exists an universal constant $\alpha_* := 0.0384629388\ldots$ such that for all $x \in S^n$, if $\overline{\alpha}(f,x) < \alpha_*$, then*

(i) *x is an approximate zero of f.*

(ii) *If ζ denotes its associated zero then $\zeta \in \overline{B}_f(x)$.*

(iii) *Furthermore, for each point z in $\overline{B}_f(x)$ the Newton sequence starting at z converges to ζ.*

4.3 Background material

Theorem 4.3 is a consequence of the following two results, which are restatements of results proved in [10]. While [10] deals with Newton iteration on arbitrary complete real analytic Riemannian manifolds, here we reword them in terms of Newton iteration on the unit sphere S^n (Example 1 in [10]). The γ-Theorem for mappings [10, Theorem 1.3] becomes the following.

Theorem 4.4. *Let $f : S^n \to \mathbb{R}^n$ be analytic. Suppose that $f(\zeta) = 0$ and $Df(\zeta)$ is an isomorphism. Let*

$$R(f,\zeta) := \min\left\{\pi, \frac{3-\sqrt{7}}{2\gamma(f,\zeta)}\right\}.$$

If $d(x,\zeta) \leq R(f,\zeta)$, then the Newton sequence $x_k = N_f^k(x)$ is defined for all $k \geq 0$ and $d(x_k,\zeta) \leq \left(\frac{1}{2}\right)^{2^k-1} d(x,\zeta)$. In particular, $\{x_k\}$ converges to ζ.

Now let $\alpha_0 := 0.130716944\ldots$ denote the smallest positive root of the polynomial $\psi(u)^2 - 2u$, and

$$s_0 := \frac{1}{\sigma + \frac{(1-\sigma\alpha_0)^2}{\psi(\sigma\alpha_0)}\left(1 + \frac{\sigma}{1-\sigma\alpha_0}\right)} = 0.103621842\ldots$$

We state the α-Theorem for mappings [10, Theorem 1.4] for the sphere S^n.

Theorem 4.5. *Let $f : S^n \to \mathbb{R}^n$ be analytic. Let $x \in S^n$ be such that $\beta(f,x) \leq s_0\pi$ and $\alpha(f,x) \leq \alpha_0$. Then the Newton sequence $x_k = N_f^k(x)$ is defined for all $k \geq 0$ and converges to a zero ζ of f. Moreover,*

$$d(x_k, x_{k+1}) \leq \left(\frac{1}{2}\right)^{2^k - 1} \beta(f,x)$$

and

$$d(x_k, \zeta) \leq \sigma\beta(f,x).$$

Finally we introduce $\psi(u) := 1-4u+2u^2$, which is positive and decreasing for $0 < u < 1-\frac{\sqrt{2}}{2}$, and state [10, Lemma 4.3]:

Lemma 4.6. *Let $x, y \in S^n$ with $d(x,y) < \pi$. Suppose that $Df(x)$ is nonsingular and*

$$\nu := d(x,y)\gamma(f,x) < 1 - \frac{\sqrt{2}}{2}.$$

Then

$$\gamma(f,y) \leq \frac{\gamma(f,x)}{(1-\nu)\psi(\nu)}.$$

4.4 Proof of Theorem 4.3

Set $\nu_* := 0.0628039411\ldots$ for the only real root of the polynomial

$$\Psi(u) := (3-\sqrt{7})(1-u)\psi(u) - 4u,$$

and $\alpha_* := \frac{\nu_*}{\sigma} = 0.0384629388\ldots$. Note that $\alpha_* \leq \min\{\alpha_0, s_0\pi\}$.

Since $\overline{\gamma}(f,x) \geq \frac{\mathbf{D}^{3/2}}{2}$, the hypothesis of Theorem 4.5 hold from $\alpha(f,x) \leq \overline{\alpha}(f,x) < \alpha_* \leq \alpha_0$ and $\beta(f,x) \leq \overline{\beta}(f,x) \leq \frac{2\overline{\alpha}(f,x)}{\mathbf{D}^{3/2}} < \frac{2\alpha_*}{\mathbf{D}^{3/2}} < s_0\pi$.
Using Remark 4.1(i) it follows that x is an approximate zero of f, and that the associated zero ζ satisfies:

$$d(x,\zeta) \leq \sigma\beta(f,x) \leq \sigma\overline{\beta}(f,x).$$

This already proves Parts (i) and (ii) of Theorem 4.3.
We show (iii). Since $d(x,\zeta) \leq \sigma\overline{\beta}(f,x) < \sigma s_0\pi < \pi$,

$$\nu = d(x,\zeta)\gamma(f,x) \leq \sigma\overline{\beta}(f,x)\gamma(f,x) \leq \sigma\overline{\alpha}(f,x) \leq \sigma\alpha_* = \nu_* < 1 - \frac{\sqrt{2}}{2},$$

and we can apply Lemma 4.6. Therefore

$$4\sigma\overline{\beta}(f,x)\gamma(f,\zeta) \leq 4\sigma\overline{\beta}(f,x)\gamma(f,x)\frac{1}{(1-\nu)\psi(\nu)} \leq 4\nu_*\frac{1}{(1-\nu_*)\psi(\nu_*)} = 3-\sqrt{7},$$

because $(1-u)\psi(u)$ decreases for $0 < u < 1-\frac{\sqrt{2}}{2}$, and ν_* is a zero of $(3-\sqrt{7})(1-u)\psi(u)-4u$. This shows, since $2\sigma\overline{\beta}(f,x) \leq \pi$, that

$$2\sigma\beta(f,x) \leq R(f,\zeta) = \min\left\{\pi, \frac{3-\sqrt{7}}{2\gamma(f,\zeta)}\right\}.$$

We conclude applying Theorem 4.4 to $z \in \overline{B}_f(x)$, since

$$d(z,\zeta) \leq d(z,x) + d(x,\zeta) \leq 2\sigma\overline{\beta}(f,x) \leq R(f,\zeta).$$

It follows that the Newton sequence $\{z_k\}_{k\in\mathbb{N}}$ starting at z converges to ζ.

Remark 4.7. The hypothesis on the radius of injectivity in [10] was recently found to be redundant.

5 Infinite precision

5.1 Grids and Graphs

Our algorithm works on a grid on S^n. We easily construct one by projecting onto S^n a grid on the cube $C^n = \{y \mid \|y\|_\infty = 1\}$. We make use of the (easy to compute) bijections $\phi : C^n \to S^n$ and $\phi^{-1} : S^n \to C^n$ given by $\phi(y) = \frac{y}{\|y\|}$ and $\phi^{-1}(x) = \frac{x}{\|x\|_\infty}$.

Given $\eta := 2^{-k}$ for some $k \geq 1$, we consider the uniform grid $\mathcal{U}_\eta$ of mesh η on C^n. This is the set of points in C^n whose coordinates are of the form $i2^{-k}$ for $i \in \{-2^k, -2^k+1, \ldots, 2^k\}$, with at least one coordinate equal to 1 or -1. We denote by $\mathcal{G}_\eta$ its image by ϕ in S^n. Note that, for $y_1, y_2 \in C^n$,

$$d(\phi(y_1), \phi(y_2)) \leq \frac{\pi}{2}\|y_1 - y_2\|_2 \leq \frac{\pi}{2}\sqrt{n+1}\,\|y_1 - y_2\|_\infty. \tag{7}$$

Given η as above we associate to it a graph G_η as follows. We set $A(f) := \{x \in S^n \mid \overline{\alpha}(f,x) < \alpha_*\}$. The vertices of the graph are the points in $\mathcal{G}_\eta \cap A(f)$. Two vertices $x, y \in \mathcal{G}_\eta$ are joined by an edge if and only if $\overline{B}_f(x) \cap \overline{B}_f(y) \neq \emptyset$.

Note that as a simple consequence of Theorem 4.3 we obtain the following lemma.

Lemma 5.1.

(i) *For each $x \in A(f)$ there exists $\zeta_x \in Z(f)$ such that $\zeta_x \in \overline{B}_f(x)$. Moreover for each point z in $\overline{B}_f(x)$, the Newton sequence starting at z converges to ζ_x.*

(ii) *Let $x, y \in A(f)$. Then $\zeta_x = \zeta_y \iff \overline{B}_f(x) \cap \overline{B}_f(y) \neq \emptyset$.* □

We define $Z(G_\eta) := \bigcup_{x\in G_\eta} \overline{B}_f(x) \subset S^n$ where $x \in G_\eta$ has to be understood as x running over all the vertices of G_η. Similarly, for a connected component U of G_η, we define

$$Z(U) := \bigcup_{x\in U} \overline{B}_f(x).$$

Lemma 5.2.

(i) *For each component U of G_η, there is a unique zero $\zeta_U \in Z(f)$ such that $\zeta_U \in Z(U)$. Moreover, $\zeta_U \in \cap_{x\in U}\overline{B}_f(x)$.*

(ii) *If U and V are different components of G_η, then $\zeta_U \neq \zeta_V$.*

PROOF. (i) Let $x \in U$. Since $x \in A(f)$, by Lemma 5.1 (i) there exists a zero ζ_x of f in $\overline{B}_f(x) \subseteq Z(U)$. This shows the existence. For the second assertion and the uniqueness, assume that there exist ζ and ξ zeros of f in $Z(U)$. Let $x, y \in U$ be such that $\zeta \in \overline{B}_f(x)$, and $\xi \in \overline{B}_f(y)$. Since U is connected, there exist $x_0 = x, x_1, \ldots, x_{k-1}, x_k := y$ in $A(f)$ such that (x_i, x_{i+1}) is an edge of G_η for $i = 0, \ldots, k-1$, that is, $\overline{B}_f(x_i) \cap \overline{B}_f(x_{i+1}) \neq \emptyset$. If ζ_i and ζ_{i+1} are the associated zeros of x_i and x_{i+1} in $Z(f)$ respectively, then by Lemma 5.1(ii) we have $\zeta_i = \zeta_{i+1}$, and thus $\zeta = \xi \in \overline{B}_f(x) \cap \overline{B}_f(y)$.
(ii) Assume $\zeta_U = \zeta_V \in \overline{B}_f(x) \cap \overline{B}_f(y) \subset Z(U) \cap Z(V)$, then x and y are joined by an edge and belong to the same connected component. $\square$

5.2 The infinite precision algorithm

```
Count_Roots_1(f)
let η := 2√2 / (π√(n+1))
(1) let U_1, ..., U_r be the connected components of G_η
if
   (i) for 1 ≤ i < j ≤ r
      for all x_i ∈ U_i and all x_j ∈ U_j, d(x_i, x_j) > πη√(n+1)
and
   (ii) for all x ∈ 𝒢_η \ A(f), ‖f(x)‖_∞ > (π/2)η√((n+1)D)‖f‖
then HALT and return r/2
else η := η/2
   go to (1)
```

5.3 Proof of Theorem 1.1(1–3)

Proof of Part (1) This proof requires some arguments of convexity. We can naturally define spherical convex hulls for sets of points in H^n, an open half-sphere in S^n. If $x_1, \ldots, x_q \in H^n$ we define

$$\mathsf{SCH}(x_1, \ldots, x_q) := \mathsf{Cone}(x_1, \ldots, x_q) \cap S^n$$

where $\mathsf{Cone}(x_1, \ldots, x_q)$ is the smallest convex cone with vertex at the origin and containing the points $x_1, \ldots, x_q$. Alternatively, we have,

$$\mathsf{SCH}(x_1, \ldots, x_q) = \left\{ \frac{\lambda_1 x_1 + \cdots + \lambda_q x_q}{\|\lambda_1 x_1 + \cdots + \lambda_q x_q\|} \mid \lambda_1, \ldots, \lambda_q \geq 0, \sum \lambda_i = 1 \right\}.$$

We will use the following fact.

Lemma 5.3. *Let $x_1, ..., x_q \in H^n \subset \mathbb{R}^{n+1}$. If $\bigcap_{i=1}^q \overline{B}(x_i, r_i) \neq \emptyset$, then $\mathsf{SCH}(x_1, \ldots, x_q) \subset \bigcup_{i=1}^q \overline{B}(x_i, r_i)$.*

PROOF. Let $x \in \mathsf{SCH}(x_1, \ldots, x_q)$ and $y \in \bigcap_{i=1}^q \overline{B}(x_i, r_i)$. We will prove that $x \in \overline{B}(x_i, r_i)$ for some i.

If $x = y$, this is obvious.

If $x \neq y$, let H be the half-space

$$H := \left\{ z \in \mathbb{R}^{n+1} : \langle z, y - x \rangle < 0 \right\}.$$

Since $\|x\| = \|y\| = 1$, we have $\langle x+y, y-x\rangle = 0$, and we note that in this case, $x+y$ determines the mid-line between x and y. Moreover, since $x \neq y$, we have $x \in H$ since $\langle x, y-x\rangle = \langle x,y\rangle - \|x\|^2 < \|x\|\,\|y\| - \|x\|^2 = 0$. Therefore the half-space H is the set of points z in $\mathbb{R}^{n+1}$ such that the Euclidean distance $\|z-x\| < \|z-y\|$.

On the other hand, H must contain at least one point of the set $\{x_1, ..., x_q\}$ since if this were not the case, the convex set $\mathsf{Cone}(\mathsf{CH}(x_1,\dots,x_q))$ would be contained in $\{z : \langle z, y-x\rangle \geq 0\}$, contradicting $x \in \mathsf{SCH}(x_1,\dots,x_q)$. Let, therefore, $x_i \in H$. It follows that

$$\|x - x_i\| < \|y - x_i\|$$

which implies

$$d(x, x_i) < d(y, x_i) \leq r_i. \qquad \square$$

We can now proceed. Assume the algorithm halts, we want to show that if r equals the number of connected components of G_η, then $\#_{\mathbb{R}}(f) = \#Z(f)/2 = r/2$. We already know by Lemma 5.2 that each connected component U of G_η determines uniquely a zero $\zeta_U \in Z(f)$. Thus it is enough to prove that $Z(f) \subset Z(G_\eta)$.

Assume that there is a zero ζ of f in S^n such that ζ is not in $Z(G_\eta)$. Let $B_\infty(\phi^{-1}(\zeta), \eta) := \{y \in \mathcal{U}_\eta \mid \|y - \phi^{-1}(\zeta)\|_\infty \leq \eta\} = \{y_1, \dots, y_q\}$, the set of all neighbors of $\phi^{-1}(\zeta)$ in $\mathcal{U}_\eta$, and let $x_i = \phi(y_i)$, $i = 1, \dots, q$. Clearly, $\phi^{-1}(\zeta)$ is in the cone spanned by $\{y_1, \dots, y_q\}$ and hence $\zeta \in \mathsf{SCH}(x_1, \dots, x_q)$.

We claim that there exists $j \leq q$ such that $x_j \notin A(f)$. Indeed, assume this is not the case. We consider two cases.

(a) All the x_i belong to the same connected component U of G_η. By Lemma 5.2 there exists a unique zero $\zeta_U \in S^n$ of f in $Z(U)$ and $\zeta_U \in \cap_i \overline{B}_f(x_i)$. We may apply Lemma 5.3 to deduce that

$$\mathsf{SCH}(x_1, \dots, x_q) \subseteq \bigcup \overline{B}_f(x_i).$$

It follows that, for some $i \in \{1, \dots, q\}$, $\zeta \in \overline{B}_f(x_i) \subseteq Z(U)$, contradicting that $\zeta \notin Z(G_\eta)$.

(b) There exist $\ell \neq s$ and $1 \leq i < j \leq r$ such that $x_\ell \in U_i$ and $x_s \in U_j$. Since condition (i) in the algorithm is satisfied, $d(x_\ell, x_s) > \pi\eta\sqrt{n+1}$. But, by (7),

$$d(x_\ell, x_s) \leq \frac{\pi}{2}\sqrt{n+1}\|y_\ell - y_s\|_\infty \leq \frac{\pi}{2}\sqrt{n+1}\left(\|y_\ell - \phi^{-1}(\zeta)\|_\infty + \|\phi^{-1}(\zeta) - y_s\|_\infty\right) \leq \pi\eta\sqrt{n+1},$$

a contradiction.

We have thus proved the claim. Let then $1 \leq j \leq q$ be such that $x_j \notin A(f)$. Since condition (ii) in the algorithm is satisfied $\|f(x_j)\|_\infty > \frac{\pi}{2}\eta\sqrt{(n+1)\mathbf{D}}\|f\|$. It follows from the inequality $d(x_j, \zeta) \leq \frac{\pi}{2}\sqrt{n+1}\eta$ and Lemma 3.1 that $\|f(\zeta)\|_\infty > 0$, a contradiction.

Proof of Part (2) We need a few lemmas.

Lemma 5.4. *If $\zeta_1 \neq \zeta_2 \in Z(f)$ then*

$$d(\zeta_1, \zeta_2) \geq \frac{2(3-\sqrt{7})\mathbf{D}^{-3/2}}{\kappa(f)}.$$

Proof. For $i=1,2$, using (6) and Proposition 2.2,

$$R(f,\zeta_i)=\min\left\{\pi,\frac{3-\sqrt{7}}{2\gamma(f,\zeta_i)}\right\}\geq\min\left\{\pi,\frac{(3-\sqrt{7})\mathbf{D}^{-3/2}}{\kappa(f)}\right\}=\frac{(3-\sqrt{7})\mathbf{D}^{-3/2}}{\kappa(f)}.$$

Now suppose that $d(\zeta_1,\zeta_2)<R(f,\zeta_1)+R(f,\zeta_2)$ and choose $x\in S^n$ such that $d(x,\zeta_1)<R(f,\zeta_1)$ and $d(x,\zeta_2)<R(f,\zeta_2)$. Then Theorem 4.4 implies that $\zeta_1=\zeta_2$, a contradiction. □

Lemma 5.5. *Let $x_1,x_2\in G_\eta$ with associated zeros $\zeta_1\neq\zeta_2$. If $\eta\leq\frac{2(3-\sqrt{7})\mathbf{D}^{-3/2}}{3\pi\kappa(f)\sqrt{n+1}}$ then $d(x_1,x_2)>\pi\eta\sqrt{n+1}$.*

Proof. Assume $d(x_1,x_2)\leq\pi\eta\sqrt{n+1}$. Since $x_2\notin\overline{B}_f(x_1)$, $d(x_1,x_2)>\sigma\overline{\beta}(f,x_1)$. Consequently,

$$d(x_1,\zeta_1)\leq\sigma\overline{\beta}(f,x_1)<d(x_1,x_2)\leq\pi\eta\sqrt{n+1}$$

and, similarly, $d(x_2,\zeta_2)<\pi\eta\sqrt{n+1}$. But then,

$$d(\zeta_1,\zeta_2)\leq d(\zeta_1,x_1)+d(x_1,x_2)+d(x_2,\zeta_2)<3\pi\eta\sqrt{n+1}\leq\frac{2(3-\sqrt{7})\mathbf{D}^{-3/2}}{\kappa(f)}$$

contradicting Lemma 5.4. □

Lemma 5.6. *Let $x\in S^n$ such that $x\notin A(f)$. If $\eta\leq\frac{\alpha_*}{(n+1)\mathbf{D}^2\kappa(f)^2}$ then $\|f(x)\|_\infty>\frac{\pi}{2}\eta\sqrt{(n+1)\mathbf{D}}\|f\|$.*

Proof. Since $x\notin A(f)$ we have $\overline{\alpha}(f,x)\geq\alpha_*$. We divide the proof in two cases.

$$\boxed{\textbf{Case I.}\ \min\left\{\mu_{\rm norm}(f,x),\frac{\|f\|}{\|f(x)\|_\infty}\right\}=\frac{\|f\|}{\|f(x)\|_\infty}}$$

In this case

$$\eta\leq\frac{\alpha_*}{(n+1)\mathbf{D}^2\kappa(f)^2}\leq\frac{\alpha_*\|f(x)\|_\infty^2}{(n+1)\mathbf{D}^2\|f\|^2}$$

which implies, since $\eta\leq\frac{1}{2}<\frac{4\mathbf{D}}{\pi^2\alpha_*}$,

$$\|f(x)\|_\infty\geq\frac{\sqrt{\eta}\sqrt{n+1}\mathbf{D}\|f\|}{\sqrt{\alpha_*}}>\frac{\pi}{2}\eta\sqrt{(n+1)\mathbf{D}}\|f\|.$$

$$\boxed{\textbf{Case II.}\ \min\left\{\mu_{\rm norm}(f,x),\frac{\|f\|}{\|f(x)\|_\infty}\right\}=\mu_{\rm norm}(f,x)}$$

In this case

$$\eta\leq\frac{\alpha_*}{(n+1)\mathbf{D}^2\kappa(f)^2}\leq\frac{\alpha_*}{(n+1)\mathbf{D}^2\mu_{\rm norm}(f,x)^2}$$

which implies $\alpha_*\geq\eta(n+1)\mathbf{D}^2\mu_{\rm norm}(f,x)^2$. Also,

$$\alpha_*\leq\overline{\alpha}(f,x)=\frac{1}{2}\overline{\beta}(f,x)\mu_{\rm norm}(f,x)\mathbf{D}^{3/2}\leq\frac{1}{2\|f\|}\mu_{\rm norm}(f,x)^2\mathbf{D}^{3/2}\|f(x)\|_\infty.$$

Putting both inequalities together we obtain

$$\eta(n+1)\mathbf{D}^2\mu_{\text{norm}}(f,x)^2 \le \frac{1}{2\|f\|}\mu_{\text{norm}}(f,x)^2\mathbf{D}^{3/2}\|f(x)\|_\infty$$

or yet,

$$\|f(x)\|_\infty \ge 2\eta(n+1)\mathbf{D}^{1/2}\|f\| > \frac{\pi}{2}\eta\sqrt{(n+1)\mathbf{D}}\|f\|. \qquad \square$$

We can now conclude the proof of Part (2). Assume $\eta \le \frac{\alpha_*}{(n+1)\mathbf{D}^2\kappa(f)^2}$. Then the hypotheses of Lemmas 5.5 and 5.6 hold. The first of these lemmas ensures that condition (i) in the algorithm is satisfied. The second, that condition (ii) is so. Therefore, the algorithm halts as soon as $\frac{\alpha_*}{2(n+1)\mathbf{D}^2\kappa(f)^2} < \eta \le \frac{\alpha_*}{(n+1)\mathbf{D}^2\kappa(f)^2}$. This gives a bound of $\mathcal{O}(\ln(n\mathbf{D}\kappa(f)))$ for the number of iterations. Since the number of grid points considered at this iteration ($\eta = \frac{\alpha_*}{(n+1)\mathbf{D}^2\kappa(f)^2}$) is at most $2(n+1)\left(\frac{2(n+1)\mathbf{D}^2\kappa(f)^2}{\alpha_*}\right)^n$, the bound for the total complexity follows.

Proof of Parts (3) and (5) We have already seen that the number of iterations is bounded by $\mathcal{O}(\ln(n\mathbf{D}\kappa(f)))$. At each of these iterations, we need to perform a number of computations on the (at most) $2(n+1)\left(\frac{2(n+1)\mathbf{D}^2\kappa(f)^2}{\alpha_*}\right)^n$ grid points to decide whether they are in $A(f)$. These can be done independently. Then, we need to compute the number of connected components of G_η. This can be done (see, e.g., [15]) in parallel time $\mathcal{O}(\ln(|V_\eta|))^2$ where $|V_\eta|$ denotes the number of vertices of G_η and therefore, in parallel time at most $\mathcal{O}(n^2(\ln(n\mathbf{D}\kappa(f))^2 + \ln(\alpha_*)^2))$. Since this is the dominant step in the computation at a given iteration, it follows that the total parallel time consumed by the algorithm is at most $\mathcal{O}(n^2\ln(n\mathbf{D}\kappa(f))(\ln(n\mathbf{D}\kappa(f))^2 + \ln(\alpha_*)^2))$. This shows part (3). For part (5), just note that, for $i = 1, \ldots, r$, any vertex x_i of U_i is an approximate zero of the only zero of f in $Z(U_i)$. $\square$

6 Finite Precision

6.1 Making room to allow errors

Our finite precision algorithm will be a variation of Algorithm `Count_Roots_1`. But since finite precision computations will be affected by errors, we need to make room in the infinite precision algorithm to allow them. For this aim, we state the corresponding version of Theorem 4.3.

Theorem 6.1. *There exist a universal constant $\alpha_\bullet = 0.028268\cdots$ such that, for all $x \in S^n$, if $\overline{\alpha}(f,x) < \alpha_\bullet$, then*

(i) *x is an approximate zero of f.*

(ii) *If ζ denotes its associated zero then $\zeta \in \overline{B}_f(x)$.*

(iii) *Furthermore, for each point z s.t. $d(x,z) \le 2\sigma\overline{\beta}(f,x)$ the Newton sequence starting at z converges to ζ.*

PROOF. Parts (i) and (ii) follow from Theorem 4.3 and the fact that $\alpha_\bullet < \alpha_*$. Part (iii) is proved by taking $\nu_\bullet = 0.046158\cdots$ to be the only real root of the polynomial $\Psi(u) := (3-\sqrt{7})(1-u)\psi(u) - 6u$, and $\alpha_\bullet = \frac{\nu_\bullet}{\sigma} = 0.028268$. Then, one proves as in Theorem 4.3 that $3\sigma\overline{\beta}(f,x) \leq R(f,\zeta)$ from which it follows that, for all z s.t. $d(x,z) \leq 2\sigma\overline{\beta}(f,x)$,

$$d(z,\zeta) \leq d(z,x) + d(x,\zeta) \leq 3\sigma\overline{\beta}(f,x) \leq R(f,\zeta)$$

and hence, that the Newton sequence $\{z_k\}_{k\in\mathbb{N}}$ starting at z converges to ζ. □

The proofs of Lemmas 5.5 and 5.6 yield, *mutatis mutandis*, the following results.

Lemma 6.2. *Let $x_1, x_2 \in G_\eta$ with associated zeros ξ_1 and ξ_2, $\xi_1 \neq \xi_2$. If $\eta \leq \frac{(3-\sqrt{7})\mathbf{D}^{-3/2}}{3\pi\kappa(f)\sqrt{n+1}}$ then $d(x_1,x_2) > 2\pi\eta\sqrt{n+1}$.* □

Lemma 6.3. *Let $x \in S^n$ such that $\overline{\alpha}(f,x) > \frac{\alpha_\bullet}{3}$. If $\eta \leq \frac{\alpha_\bullet}{4\mathbf{D}^2(n+1)\kappa(f)^2}$ then $\|f(x)\|_\infty > \pi\eta\sqrt{(n+1)\mathbf{D}}\|f\|$.* □

6.2 Basic facts

We recall the basics of a floating-point arithmetic which idealizes the usual IEEE standard arithmetic. This system is defined by a set $\mathbb{F} \subset \mathbb{R}$ containing 0 (the *floating-point numbers*), a transformation $r : \mathbb{R} \to \mathbb{F}$ (the *rounding map*), and a constant $u \in \mathbb{R}$ (the *round-off unit*) satisfying $0 < u < 1$. The properties we require for such a system are the following:

(i) For any $x \in \mathbb{F}$, $r(x) = x$. In particular, $r(0) = 0$.

(ii) For any $x \in \mathbb{R}$, $r(x) = x(1+\delta)$ with $|\delta| \leq u$.

We also define on $\mathbb{F}$ arithmetic operations following the classical scheme

$$x \tilde{\circ} y = r(x \circ y)$$

for any $x, y \in \mathbb{F}$ and $\circ \in \{+,-,\times,/\}$, so that

$$\tilde{\circ} : \mathbb{F} \times \mathbb{F} \to \mathbb{F}.$$

The following is an immediate consequence of property (ii) above.

Proposition 6.4. *For any $x, y \in \mathbb{F}$ we have*

$$x \tilde{\circ} y = (x \circ y)(1+\delta), \qquad |\delta| \leq u.$$ □

When combining many operations in floating-point arithmetic, quantities such as $\prod_{i=1}^n (1+\delta_i)^{\rho_i}$ naturally appear. Our round-off analysis uses the notations and ideas in Chapter 3 of [17], from where we quote the following results:

Proposition 6.5. *If $|\delta_i| \leq u$, $\rho_i \in \{-1,1\}$, and $nu < 1$, then*

$$\prod_{i=1}^n (1+\delta_i)^{\rho_i} = 1 + \theta_n,$$

where

$$|\theta_n| \leq \gamma_n = \frac{nu}{1-nu}.$$ □

Proposition 6.6. *For any positive integer k such that $ku < 1$, let θ_k, θ_j be any quantities satisfying*

$$|\theta_k| \leq \gamma_k = \frac{ku}{1-ku} \qquad |\theta_j| \leq \gamma_j = \frac{ju}{1-ju}.$$

The following relations hold.

1. $(1+\theta_k)(1+\theta_j) = 1+\theta_{k+j}$ *for some* $|\theta_{k+j}| \leq \gamma_{k+j}$.
2. $$\frac{1+\theta_k}{1+\theta_j} = \begin{cases} 1+\theta_{k+j} & \text{if } j \leq k, \\ 1+\theta_{k+2j} & \text{if } j > k. \end{cases}$$
   for some $|\theta_{k+j}| \leq \gamma_{k+j}$ *or some* $|\theta_{k+2j}| \leq \gamma_{k+2j}$.
3. *If* $ku, ju \leq 1/2$, *then* $\gamma_k\gamma_j \leq \gamma_{\min\{k,j\}}$.
4. $i\gamma_k \leq \gamma_{ik}$.
5. $\gamma_k + u \leq \gamma_{k+1}$.
6. $\gamma_k + \gamma_j + \gamma_k\gamma_j \leq \gamma_{k+j}$. □

From now on, whenever we write an expression containing θ_k we mean that the same expression is true for some θ_k, with $|\theta_k| \leq \gamma_k$.

When computing an arithmetic expression q with a round-off algorithm, errors will accumulate and we will obtain another quantity which we will denote by $\mathtt{fl}(q)$. We write $\mathtt{Error}\,(q) = |q - \mathtt{fl}(q)|$.

An example of round-off analysis which will be useful in what follows is given in the next proposition, the proof of which can be found in Section 3.1 of [17].

Proposition 6.7. *There is a round-off algorithm which, with input $x, y \in \mathbb{R}^n$, computes the dot product of x and y. The computed value $\mathtt{fl}(\langle x,y\rangle)$ satisfies*

$$\mathtt{fl}(\langle x,y\rangle) = \langle x,y\rangle + \theta_{\lceil \log_2 n\rceil+1}\langle |x|,|y|\rangle,$$

where $|x| = (|x_1|,\ldots,|x_n|)$. In particular, if $x = y$, the algorithm computes $\mathtt{fl}(\|x\|^2)$ satisfying

$$\mathtt{fl}(\|x\|^2) = \|x\|^2(1+\theta_{\lceil \log_2 n\rceil+1}).$$ □

We will also have to deal with square roots and arccosinus. The following result will help us to do so.

Lemma 6.8. **(i)** *Let $\theta \in \mathbb{R}$ such that $|\theta| \leq 1/2$. Then, $\sqrt{1-\theta} = 1-\theta'$ with $|\theta'| \leq |\theta|$.*

(ii) *Let $0 < a \leq 1$ and $\varepsilon \in \mathbb{R}$ such that $0 < a+\varepsilon < 1$. Then,* $\arccos(a+\varepsilon) = \arccos(a) + \upsilon\frac{1}{\sqrt{1-(a+\varepsilon)^2}}$ *with $|\upsilon| \leq |\varepsilon|$.*

Proof. Assume $\theta > 0$ (if $\theta < 0$ it is done similarly). By the intermediate value theorem we have that $1-\sqrt{1-\theta} = \theta(\sqrt{\xi})'$ with $\xi \in (1-\theta, 1)$. But

$$(\sqrt{\xi})' = \frac{1}{2\sqrt{\xi}} \leq \frac{1}{\sqrt{2}},$$

the last since $\xi \geq 1/2$. This proves (i).

Part (ii) is shown similarly. Again, assume for simplicity that $\varepsilon > 0$. Then, for some $\xi \in (a, a+\varepsilon)$,

$$\arccos(a+\varepsilon) - \arccos(a) = \varepsilon \arccos'(\xi) = \varepsilon \frac{1}{\sqrt{1-\xi^2}} = \frac{\upsilon}{\sqrt{1-(a+\varepsilon)^2}}. \qquad \square$$

We assume that, besides the four basic arithmetic operations, we are allowed to compute square roots and arccosinus with finite precision. That is, if `op` denotes any of these two operators, we compute $\widetilde{\mathtt{op}}$ such that

$$\widetilde{\mathtt{op}}(x) = \mathtt{op}(x)(1+\delta), \quad |\delta| \leq u.$$

From Lemma 6.8(i) it follows that, for all $a > 0$,

$$\widetilde{\sqrt{a(1+\theta_k)}} = \sqrt{a}(1+\theta_{k+1}).$$

Remark 6.9. Our choice of the precision u in Theorem 1.1(4) guarantees that $ku < 1/2$ holds whenever we encounter θ_k in what follows, and consequently, $|\theta_k| \leq \gamma_k \leq 2ku$. This implies that in all what follows we have $\gamma_g = \mathcal{O}(ug)$ for all the expressions g we will encounter.

According to the previous remark we will introduce a further notation that will considerably simplify our exposition. For all expression g, we will write

$$[\![g]\!] := \mathcal{O}(ug).$$

This notation will avoid we burden ourselves with the consideration of multiplicative constants.

6.3 The finite precision algorithm

Our finite precision algorithm is a variation of Algorithm `Count_Roots_1` in Section 5.3. Given $x \in S^n$ we define below $\mathtt{fl}(A'(f))$ and $\mathtt{fl}(\overline{B}'_f(x))$, which are convenient floating versions of the sets $A'(f) = \left\{x \in S^n \mid \overline{\alpha}(f,x) < \frac{1}{2}\alpha_\bullet\right\}$ and $\overline{B}'_f(y) = \{z \in S^n \mid d(x,y) \leq \frac{3}{2}\sigma\overline{\beta}(f,x)\}$ respectively.

Given $f \in \mathcal{H}_{\mathbf{d}}$ and $x \in S^n$, we let $M \in \mathbb{R}^{n \times n}$ be a matrix representing

$$\begin{bmatrix} \frac{1}{\sqrt{d_1}} & & & \\ & \frac{1}{\sqrt{d_2}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{d_n}} \end{bmatrix} Df(x)_{|T_xS^n}.$$

and we set $\sigma_{\min}(M) = \|M^{-1}\|^{-1}$. Therefore

$$\begin{aligned}
\mu_{\text{norm}}(f,x) &= \|f\|\sqrt{n}\,\|M^{-1}\| = \|f\|\sqrt{n}\,\sigma_{\min}(M)^{-1}, \\
\overline{\beta}(f,x) &= \mu_{\text{norm}}(f,x)\frac{\|f(x)\|_\infty}{\|f\|} = \sqrt{n}\,\sigma_{\min}(M)^{-1}\|f(x)\|_\infty, \\
\overline{\alpha}(f,x) &= \overline{\beta}(f,x)\mu_{\text{norm}}(f,x)\frac{\mathbf{D}^{3/2}}{2} = \|f\|\,n\,\sigma_{\min}(M)^{-2}\|f(x)\|_\infty\frac{\mathbf{D}^{3/2}}{2}.
\end{aligned}$$

This implies that

$$\begin{array}{lclcl} y \in \overline{B}'_f(x) & \iff & d(x,y) \le \frac{3}{2}\sigma\overline{\beta}(f,x) & \iff & \sigma_{\min}(M)d(x,y) \le \frac{3}{2}\sigma\sqrt{n}\|f(x)\|_\infty, \\ x \in A'(f) & \iff & \overline{\alpha}(f,x) < \frac{\alpha_\bullet}{2} & \iff & \|f\|n\|f(x)\|_\infty \mathbf{D}^{3/2} < \alpha_\bullet \sigma_{\min}(M)^2. \end{array}$$

These statements are equivalent under infinite precision, but the expressions at the right-hand side are more convenient to handle when working with finite precision. This motivates our definitions of

$$\begin{array}{rcl} \mathtt{fl}(\overline{B}'_f(x)) & := & \left\{ y \in S^n \mid \mathtt{fl}(\sigma_{\min}(M)d(x,y)) \le \mathtt{fl}(\dfrac{3}{2}\sigma\sqrt{n}\|f(x)\|_\infty) \right\} \\ \mathtt{fl}(A'(f)) & := & \left\{ x \in S^n \mid \mathtt{fl}(\|f\|\, n\, \|f(x)\|_\infty \mathbf{D}^{3/2}) < \mathtt{fl}(\alpha_\bullet \sigma_{\min}(M)^2) \right\} \end{array}$$

We also define accordingly the graph $\mathtt{fl}(G'_\eta)$ whose vertices are the points in $G_\eta \cap \mathtt{fl}(A'(f))$, and with two vertices x, y joined by an edge if and only if $\mathtt{fl}(\overline{B}'_f(x)) \cap \mathtt{fl}(\overline{B}'_f(x)) \neq \emptyset$. Its connected components are denoted by $\mathtt{fl}(U)$.

Our algorithm is the following:

> $\mathtt{Count\_Roots\_2}(f)$
> let $\eta := \frac{2\sqrt{2}}{\pi\sqrt{n+1}}$
> (1) let $\mathtt{fl}(U_1), \dots, \mathtt{fl}(U_r)$ be the connected components of $\mathtt{fl}(\overline{G}_\eta)$
> if
> (i) for $1 \le i < j \le r$
> for all $x_i \in \mathtt{fl}(U_i)$ and all $x_j \in \mathtt{fl}(U_j)$, $\mathtt{fl}(d(x_i, x_j)) > \mathtt{fl}(\frac{3}{2}\pi\eta\sqrt{n+1})$
> and
> (ii) for all $x \in \mathcal{G}_\eta \setminus \mathtt{fl}(A'(f))$, $\mathtt{fl}(\|f(x)\|_\infty) > \mathtt{fl}(\frac{\sqrt{2}}{2}\pi\eta\sqrt{(n+1)\mathbf{D}}\|f\|)$
> then HALT and return $r/2$
> else $\eta := \eta/2$
> go to (1)

In the rest of the section we will see that, when the precision u satisfies $u \le \frac{1}{\mathcal{O}\left(\mathbf{D}^2 n^{5/2}\kappa(f)^3(\log S + n^{3/2}\mathbf{D}^2\kappa(f)^2)\right)}$, this algorithm is correct and halts as soon as $\eta \le \frac{\alpha_\bullet}{4\mathbf{D}^2(n+1)\kappa(f)^2}$.

6.4 Bounding errors for elementary computations

The goal of this subsection is to exhibit bounds for the accumulated error in the main computations of $\mathtt{Count\_Roots\_2}$. We will rely on the basic notations and results described in §6.2.

To simplify notation, and without loss of generality, in all what follows we assume that $\|f\| = 1$. We denote by $S(\mathcal{H}_\mathbf{d})$ the sphere of such systems. Also, we do not discuss in what follows the accumulated error in the computation of $\phi : C^n \to S^n$. This is a minor detail which can be taken care of using Lemma 6.8(i).

Proposition 6.10. *Given $f \in S(\mathcal{H}_\mathbf{d})$ and $x \in S^n$, we can compute $\|f(x)\|_\infty$ with finite precision u such that*

$$\mathtt{Error}\,(\|f(x)\|_\infty) = [\![\mathbf{D} + \log S]\!]$$

where S is a bound on the number of coefficients of each f_i.

PROOF. Let $f = (f_1, \dots, f_n)$. For $i \le n$ write $f_i = \sum c_J X^J$ and let S be the number of coefficients of f_i. To compute $f(x)$ one computes each monomial $c_J x^J$ with $\mathtt{fl}(c_J x^J) = c_J x^J(1+\theta_D)$. Then, one computes $f_i(x)$ to get

$$\begin{aligned}
\mathtt{fl}(f_i(x)) &= \mathtt{fl}(\sum \mathtt{fl}(c_J x^J)) \\
&= \mathtt{fl}(\sum c_J x^J (1+\theta_{\mathbf{D}}^{(J)})) \\
&= \sum c_J x^J (1+\theta_{\mathbf{D}}^{(J)}) + \theta_{\log S} \sum |c_J x^J| (1+\theta_{\mathbf{D}}^{(J)}) \\
&= f_i(x) + \sum c_J x^J \theta_{\mathbf{D}}^{(J)} + \theta_{\log S} \sum |c_J x^J| (1+\theta_{\mathbf{D}}^{(J)})
\end{aligned}$$

where in the third line we reasoned as in the proof of Proposition 6.7. Therefore

$$\begin{aligned}
\mathtt{Error}\,(\|f(x)\|_\infty) &\le \left| \sum c_J x^J \theta_{\mathbf{D}}^{(J)} + \theta_{\log S} \sum |c_J x^J| (1+\theta_{\mathbf{D}}^{(J)}) \right| \\
&\le \sum |c_J|\, \|x^J\| (\gamma_{\mathbf{D}} + \gamma_{\log S} + \gamma_{\mathbf{D}} \gamma_{\log S}) \\
&\le \gamma_{\mathbf{D} + \log S}
\end{aligned}$$

where we used that for any $x \in S^n$, $|\sum |c_J| x^J| \le \|\sum |c_J| x^J\| = \|f_i\| \le \|f\| = 1$ and Proposition 6.6 (6). The conclusion follows from Remark 6.9. $\square$

Proposition 6.11. *Given $f \in S(\mathcal{H}_{\mathbf{d}})$ and $x \in S^n$, let $M \in \mathbb{R}^{n\times n}$ be a matrix representing*

$$\begin{bmatrix} \frac{1}{\sqrt{d_1}} & & & \\ & \frac{1}{\sqrt{d_2}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{d_n}} \end{bmatrix} Df(x)_{|T_x S^n}$$

in some orthonormal basis of $T_x S^n$. Then $\|M\| \le \sqrt{n}$. In addition, we can compute such a matrix M with finite precision u such that

$$\|\mathtt{Error}\,(M)\|_F = [\![n(\log S + \mathbf{D} + \log n)]\!].$$

PROOF. **Step 1:** Let $y = \frac{x - \mathrm{e}_{n+1}}{\|x - \mathrm{e}_{n+1}\|}$. The *Householder symmetry*

$$H_y = I_{n+1} - 2yy^t$$

swaps vectors e_{n+1} and x, and fixes $y^\perp$. The first n columns of H_y are therefore an orthonormal basis of $T_x S^n$, while the last column is x. Let $H \in \mathbb{R}^{(n+1)\times n}$ denote the submatrix obtained from the first n columns of H_y. With that notation, we set

$$M = \begin{bmatrix} \frac{1}{\sqrt{d_1}} & & & \\ & \frac{1}{\sqrt{d_2}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{d_n}} \end{bmatrix} Df(x) H.$$

Step 2: We claim that $P_{i,x} : \mathcal{H}_{d_i} \to \mathbb{R}^n$, $f_i \mapsto \frac{1}{\sqrt{d_i}} Df_i(x)_{|T_x S^n}$ is an orthogonal projection, in the sense that for any fixed x, the map $(P_{i,x})_{|\ker(P_{i,x})^\perp}$ is an isometry.

We use an orthogonal invariance argument. The special orthogonal group $SO(n+1)$ acts on $\mathcal{H}_{d_i}$ and on $\mathbb{R}^{n+1}$ isometrically as follows: to a given $Q \in SO(n+1)$, we associate respectively the following isometries:

$$x \mapsto Qx \quad , \quad f_i \mapsto f_i \circ Q^t.$$

We set $y = Qx$ and $g_i = f_i \circ Q^t$. Differentiating the equality $g_i(Qx) = f_i(x)$, we obtain:

$$Dg_i(y)Q = Df_i(x).$$

When x is fixed, we can set Q conveniently so that $y = \mathrm{e}_{n+1}$. Therefore

$$Dg_i(\mathrm{e}_{n+1})Q_{|T_xS^n} = Df_i(x)_{|T_xS^n}.$$

Since $Q(T_xS^n) = T_{\mathrm{e}_{n+1}}S^n$ we obtain

$$Dg_i(\mathrm{e}_{n+1})_{|T_{\mathrm{e}_{n+1}}S^n} = Df_i(x)_{|T_xS^n}.$$

This means that $P_{i,\mathrm{e}_{n+1}}(f_i \circ Q^t) = P_{i,x}(f_i)$. Thus, in order to prove our claim, it is enough to show that $P_{i,\mathrm{e}_{n+1}}$ is an orthogonal projection.

Since for $g = \sum_J g_J X^J$, $\frac{\partial g}{\partial X_j}(\mathrm{e}_{n+1}) = g_{(\mathrm{e}_j+(d-1)\mathrm{e}_{n+1})}$ and since $T_{e_{n+1}}S^n = \langle \mathrm{e}_1, \ldots, \mathrm{e}_n \rangle$, we have that for any $g_i \in \mathcal{H}_{d_i}$,

$$P_{i,\mathrm{e}_{n+1}}(g_i) = \frac{1}{\sqrt{d_i}} \big(g_{i(\mathrm{e}_1+(d_i-1)\mathrm{e}_{n+1})}, \ldots, g_{i(\mathrm{e}_n+(d_i-1)\mathrm{e}_{n+1})} \big) \; .$$

Hence, for any $g_i \in \ker(P_{i,\mathrm{e}_{n+1}})^\perp$, i.e. such that $g_{iJ} = 0$ for all $J \neq \mathrm{e}_j + (d_i - 1)\mathrm{e}_{n+1}$, $1 \leq j \leq n$, we have

$$\|g_i\|^2 = \sum_J \frac{g_{iJ}^2}{\binom{d_i}{J}} = \|P_{i,\mathrm{e}_{n+1}}(g_i)\|_2^2.$$

We conclude that $P_{i,x}$ is an orthogonal projection.

Step 3: From the previous step, for any $f_i \in \mathcal{H}_{d_i}$, using the orthogonal decomposition $f_i = f_i^\circ + f_i^\perp$ with $f_i^\circ \in \ker P_{i,x}$ and $f_i^\perp \in \ker P_{i,x}^\perp$, we have

$$\|P_{i,x}(f_i)\|_2^2 = \|P_{i,x}(f_i^\perp)\|_2^2 = \|f_i^\perp\|^2 \leq \|f_i\|^2.$$

It is now immediate from Step 1 and from the definition of $\|f\| = \max_i \|f_i\|$ that the Frobenius norm $\|M\|_F$ of the matrix M satisfies

$$\|M\|_F^2 = \sum_{i=1}^n \|P_{i,x}(f_i)\|_2^2 \leq \sum_{i=1}^n \|f_i\|^2 \leq n\|f\|^2 = n$$

and hence its spectral norm $\|M\|$ satisfies $\|M\| \leq \|M\|_F \leq \sqrt{n}$. This bound is independent of the choice of the basis for the space T_xS^n.

Step 4: We next present the algorithm to compute M, given f and x. This is a non-optimal algorithm, and can be significantly improved if more is known on the structure of the polynomial system f.

We can compute each entry m_{ij} of the matrix M as the scalar product of $\frac{1}{\sqrt{d_i}}Df_i(x)$ and the jth column $H_j := (h_{kj})_{1\leq k\leq n+1}$ of H.

Proceeding as in the proof of Proposition 6.10, we can compute $\frac{1}{\sqrt{d_i}}\frac{\partial f_i}{\partial X_k}(x)$ with

$$\mathtt{Error}\left(\frac{1}{\sqrt{d_i}}\frac{\partial f_i}{\partial X_k}(x)\right)=[\![\mathbf{D}+\log S]\!].$$

On the other hand, the vector $y=\frac{x-\mathrm{e}_{n+1}}{\|x-\mathrm{e}_{n+1}\|}$ can be computed using $2n+4$ operations, and clearly $\mathtt{Error}\,(y_j)=[\![\log(n)]\!]$ for all j. Hence, for all coefficients h_{kj} of H,

$$\mathtt{Error}\,(h_{kj})=[\![\log(n)]\!].$$

Applying Proposition 6.7 we conclude

$$\begin{aligned}\mathtt{Error}\,(m_{ij}) &= [\![\mathbf{D}+\log S+\log n]\!]\left\|\frac{1}{\sqrt{d_i}}Df_i(x))\right\|\,\|H_j\|\\ &= [\![\mathbf{D}+\log S+\log n]\!].\end{aligned}$$

The second equality holds because $\|H_j\|=1$ since H is unitary, and because, as in the proof of Step 2,

$$\left\|\frac{1}{\sqrt{d_i}}Df_i(x)\right\|^2=\left\|\frac{1}{\sqrt{d_i}}Dg_i(\mathrm{e}_{n+1})\right\|^2=\frac{1}{d_i}\|(g_{i(\mathrm{e}_1+(d_i-1)\mathrm{e}_{n+1})},\ldots,g_{i(d_i\mathrm{e}_{n+1})})\|^2\le\|g_i\|^2\le 1.$$

This implies

$$\|\mathtt{Error}\,(M)\|_F\le[\![n(\mathbf{D}+\log S+\log n)]\!].\qquad\square$$

Lemma 6.12. *Let $x\in S^n$ and M be as in Proposition 6.11. We can compute $\sigma_{\min}(M)=\|M^{-1}\|^{-1}$ satisfying*

$$\mathtt{Error}\,(\sigma_{\min}(M))=[\![n(\log S+\mathbf{D}+n^{3/2})]\!].$$

Proof. Let $E'=M-\mathtt{fl}(M)$. By Proposition 6.11,

$$\|E'\|\le\|E'\|_F\le[\![n(\log S+\mathbf{D}+\log n)]\!].$$

Let $\mathcal{M}=\mathtt{fl}(M)$. We compute $\sigma_{\min}(\mathcal{M})=\|M^{-1}\|^{-1}$ using a backward stable algorithm (e.g., QR factorization). Then the computed $\mathtt{fl}(\sigma_{\min}(\mathcal{M}))$ is the exact $\sigma_{\min}(\mathcal{M}+E'')$ for a matrix E'' with

$$\|E''\|\le cn^2u\|\mathcal{M}\|$$

for some universal constant c (see, e.g., [11, 17]). Thus,

$$\mathtt{fl}(\sigma_{\min}(M))=\mathtt{fl}(\sigma_{\min}(\mathcal{M}))=\sigma_{\min}(\mathcal{M}+E'')=\sigma_{\min}(M+E'+E'').$$

Write $E=E'+E''$. Then, using $\|M\|\le\sqrt{n}$,

$$\begin{aligned}\|E\| &\le \|E'\|+\|E''\|\le\|E'\|+cn^2u\|\mathcal{M}\|\le\|E'\|+cn^2u(\|M\|+\|E'\|)\\ &= [\![n(\log S+\mathbf{D}+\log n)]\!]+cn^2u(\sqrt{n}+[\![n(\log S+\mathbf{D}+\log n)]\!])\\ &= [\![n(\log S+\mathbf{D}+\log n)]\!]+cn^2u(\sqrt{n}+c'un(\log S+\mathbf{D}+n^{3/2}))\\ &= [\![n(\log S+\mathbf{D}+n^{3/2})]\!]\end{aligned}$$

since the hypothesis on u implies $c'un(\log S+\mathbf{D}+n^{3/2})$ is bounded by a constant term.

Therefore, $\mathtt{fl}(\sigma_{\min}(M))=\sigma_{\min}(M+E)$ which implies by [11, Corollary 8.3.2]:

$$\mathtt{Error}\,(\sigma_{\min}(M))\le\|E\|<[\![n(\log S+\mathbf{D}+n^{3/2})]\!].\qquad\square$$

Proposition 6.13. *Let $f \in S(\mathcal{H}_{\mathbf{d}})$). Assume $u \leq \frac{K}{\kappa(f)^2 n^2 \mathbf{D} \log S}$ for a small enough constant and let $x \in S^n$. Then*

(i) *If $x \notin \mathtt{fl}(A'(f))$ then $\overline{\alpha}(f,x) \geq \frac{1}{3}\alpha_\bullet$.*

(ii) *If $x \in \mathtt{fl}(A'(f))$ then $\overline{\alpha}(f,x) < \alpha_\bullet$.*

Proof. From Proposition 6.10

$$\begin{aligned}\mathtt{fl}(n\|f(x)\|_\infty \mathbf{D}^{3/2}) &= (\|f(x)\|_\infty + [\![\mathbf{D} + \log S]\!])(n\mathbf{D}^{3/2})(1+\theta_4)\\ &\leq n\mathbf{D}^{3/2}\|f(x)\|_\infty + [\![n\mathbf{D}^{3/2}(\mathbf{D}+\log S)]\!]\end{aligned}$$

. Also, from Lemma 6.12, using that $\sigma_{\min}(M) \leq \sqrt{n}$,

$$\begin{aligned}\mathtt{fl}(\alpha_\bullet \sigma_{\min}(M)^2) &= \alpha_\bullet \left(\sigma_{\min}(M) + [\![n(\log S + \mathbf{D} + n^{3/2})]\!]\right)^2 (1+\theta_2)\\ &\geq \alpha_\bullet \sigma_{\min}(M)^2 - 2\alpha_\bullet \sigma_{\min}(M)[\![n(\log S + \mathbf{D} + n^{3/2})]\!]\\ &\geq \alpha_\bullet \sigma_{\min}(M)^2 - [\![n^{3/2}(\log S + \mathbf{D} + n^{3/2})]\!].\end{aligned}$$

Therefore,

$$\begin{aligned}n\|f(x)\|_\infty \mathbf{D}^{3/2} + [\![n\mathbf{D}^{3/2}(\mathbf{D}+\log S)]\!] &\geq \mathtt{fl}(n\|f(x)\|_\infty \mathbf{D}^{3/2}) \geq \mathtt{fl}(\alpha_\bullet \sigma_{\min}^2)\\ &\geq \alpha_\bullet \sigma_{\min}^2 - [\![n^{3/2}(\log S + \mathbf{D} + n^{3/2})]\!]\end{aligned}$$

or yet,

$$\begin{aligned}n\|f(x)\|_\infty \mathbf{D}^{3/2} - \alpha_\bullet \sigma_{\min}^2 &\geq -([\![n\mathbf{D}^{3/2}(\mathbf{D}+\log S)]\!] + [\![n^{3/2}(\log S + \mathbf{D} + n^{3/2})]\!])\\ &\geq -[\![n^3\mathbf{D}^{5/2}\log S]\!].\end{aligned}$$

$\boxed{\textbf{Case I. } \min\left\{\mu_{\text{norm}}(f,x), \frac{1}{\|f(x)\|_\infty}\right\} = \frac{1}{\|f(x)\|_\infty}}$

In this case $\kappa(f) \geq \frac{1}{\|f(x)\|_\infty}$ and, therefore, using the hypothesis on u and the inequality $\kappa(f) \geq 1$,

$$\begin{aligned}[\![n^3\mathbf{D}^{5/2}\log S]\!] &= u\mathcal{O}(n^3\mathbf{D}^{5/2}\log S) \leq K\frac{\mathcal{O}(n^3\mathbf{D}^{5/2}\log S)}{\kappa(f)n^2\mathbf{D}\log S}\\ &\leq K\mathcal{O}(1)n\|f(x)\|_\infty \mathbf{D}^{3/2} \leq \frac{n\|f(x)\|_\infty \mathbf{D}^{3/2}}{2}\end{aligned}$$

the last by choosing K small enough. Hence, $n\|f(x)\|_\infty \mathbf{D}^{3/2} - \alpha_\bullet \sigma_{\min}^2 \geq -\left(\frac{n\|f(x)\|_\infty \mathbf{D}^{3/2}}{2}\right)$, which implies $\frac{3}{2}n\|f(x)\|_\infty \mathbf{D}^{3/2} \geq \alpha_\bullet \sigma_{\min}(M)^2$, i.e., $\overline{\alpha}(f,x) \geq \frac{\alpha_\bullet}{3}$.

$\boxed{\textbf{Case II. } \min\left\{\mu_{\text{norm}}(f,x), \frac{1}{\|f(x)\|_\infty}\right\} = \mu_{\text{norm}}(f,x)}$

In this case $\kappa(f) \geq \mu_{\text{norm}}(f,x) = \frac{\sqrt{n}}{\sigma_{\min}(M)}$. By the hypothesis on u,

$$\begin{aligned}[\![n^3\mathbf{D}^{5/2}\log S]\!] &= u\mathcal{O}(n^3\mathbf{D}^{5/2}\log S) \leq K\frac{\mathcal{O}(n^3\mathbf{D}^{5/2}\log S)}{\kappa(f)^2 n^2\mathbf{D}\log S}\\ &\leq K\mathcal{O}(1)\sigma_{\min}(M)^2\mathbf{D}^{3/2} \leq \frac{\alpha_\bullet \sigma_{\min}(M)^2}{3}\end{aligned}$$

the last by choosing K small enough. This implies $n\|f(x)\|_\infty \mathbf{D}^{3/2} - \alpha_\bullet \sigma_{\min}(M)^2 \geq -\frac{\alpha_\bullet \sigma_{\min}(M)^2}{3}$ or, equivalently, $\overline{\alpha}(f,x) \geq \frac{\alpha_\bullet}{3}$.

This shows part (i). For part (ii), one shows as above that

$$n\|f(x)\|_\infty \mathbf{D}^{3/2} - \alpha_\bullet \sigma_{\min}^2 \leq [\![n^3 \mathbf{D}^{5/2} \log S]\!].$$

Then, one proceeds as well by considering the two cases $\min\left\{\mu_{\text{norm}}(f,x), \frac{1}{\|f(x)\|_\infty}\right\} = \frac{1}{\|f(x)\|_\infty}$ and $\min\left\{\mu_{\text{norm}}(f,x), \frac{1}{\|f(x)\|_\infty}\right\} = \mu_{\text{norm}}(f,x)$. $\square$

Lemma 6.14. *Let $y_1, y_2 \in \mathcal{U}_\eta$ and let $x_i = \phi(y_i)$, $i = 1,2$. Then $d(x_1,x_2) \geq \frac{\eta}{2\sqrt{n+1}}$.*

Proof. The distance $d(x_1,x_2)$ is minimized at $y_1 = (1,\ldots,1,1)$ and $y_2 = (1,\ldots,1,1-\eta)$. Let $N = n+1$. Then

$$\begin{aligned}
\cos(d(x_1,x_2))^2 &= \frac{\langle y_1, y_2\rangle^2}{\|y_1\|^2\|y_2\|^2} \\
&= \frac{(N-\eta)^2}{N(N-2\eta+\eta^2)} \\
&= 1 - \frac{(N-1)\eta^2}{N^2 - 2N\eta + N\eta^2} \\
&\leq 1 - \eta^2 \frac{N-1}{N^2}.
\end{aligned}$$

Hence

$$d(x_1,x_2) \geq \arccos\left(\sqrt{1-\eta^2\frac{N-1}{N^2}}\right) = \arcsin\left(\frac{\eta}{N}\sqrt{N-1}\right) \geq \frac{\eta}{2\sqrt{N}}.$$

$\square$

Lemma 6.15. *Let $u < \frac{K\eta^2}{n\log n}$ for a small enough constant K. For $x_1, x_2 \in \mathcal{G}_\eta$ we can compute $d(x_1,x_2)$ such that*

$$\texttt{Error}\,(d(x_1,x_2)) \leq \left[\!\!\left[\frac{\sqrt{n}\log n}{\eta} \right]\!\!\right].$$

Proof. Let $y_i = \phi^{-1}(x_i)$, $i = 1,2$, and $a = \cos(d(x_1,x_2))$, i.e.,

$$a = \frac{\langle y_1, y_2\rangle}{\|y_1\|\|y_2\|}.$$

We have, using Proposition 6.7,

$$\texttt{fl}(\langle y_1,y_2\rangle) = \langle y_1,y_2\rangle + \theta_{\log n}\|y_1\|\|y_2\|$$

and $\texttt{fl}(\|y_1\|\|y_2\|) = \|y_1\|\|y_2\|(1+\theta_{\log n})$. Using now Propositions 6.4, 6.5, and 6.6, it follows that $\texttt{fl}(a) = a + \varepsilon$ with $\varepsilon = [\![\log n]\!]$.

By choosing K sufficiently small, $\varepsilon \leq \frac{\eta^2 n}{12(n+1)^2}$. Also, from the proof of Lemma 6.14,

$$a = \cos(d(x_1,x_2)) \leq \sqrt{1 - \frac{\eta^2 n}{(n+1)^2}}$$

and hence, using that $\sqrt{z}+y \leq \sqrt{z+3y}$ whenever $0<z,y\leq 1$, we obtain

$$a+\varepsilon \leq \sqrt{1-\frac{\eta^2 n}{(n+1)^2}}+\frac{\eta^2 n}{12(n+1)^2} \leq \sqrt{1-\frac{3\eta^2 n}{4(n+1)^2}} \leq \sqrt{1-\frac{\eta^2}{3(n+1)}}.$$

Using Lemma 6.8(ii) it follows that,

$$\begin{aligned} \arccos(a+\varepsilon) &= \arccos(a)+\varepsilon\left|\frac{1}{\sqrt{1-(a+\varepsilon)^2}}\right| \\ &= \arccos(a)+[\![\log n]\!]\left|\frac{\sqrt{3(n+1)}}{\eta}\right|. \end{aligned}$$

Therefore,

$$\mathtt{Error}\,(d(x_1,x_2)) \leq \left[\!\!\left[\frac{\sqrt{n}\log n}{\eta}\right]\!\!\right]. \qquad \square$$

Lemma 6.16. *Let $f \in S(\mathcal{H}_{\mathbf{d}})$. Assume that $\eta \geq \frac{\alpha_\bullet}{8\mathbf{D}^2(n+1)\kappa(f)^2}$ and $u \leq \frac{K}{\mathbf{D}^2 n^{5/2}\kappa(f)^3(\log S+n^{3/2}\mathbf{D}^2\kappa(f)^2)}$ with K small enough, and let $x,y \in \mathcal{G}_\eta$. Then*

(i) *If $y \in \mathtt{fl}(\overline{B}'_f(x))$ then $d(x,y) \leq 2\sigma\overline{\beta}(f,x)$.*

(ii) *If $y \notin \mathtt{fl}(\overline{B}'_f(x))$ then $d(x,y) > \sigma\overline{\beta}(f,x)$.*

PROOF. By Lemmas 6.12 and 6.15 (and using $\sigma_{\min}(M) \leq \sqrt{n}$ and the bound $d(x,y) \leq \frac{\pi}{2}\eta\sqrt{n+1}$ which follows from (7)),

$$\begin{aligned} \mathtt{Error}\,(\sigma_{\min}(M)d(x,y)) &= \mathcal{O}\big(d(x,y)\mathtt{Error}\,(\sigma_{\min}(M))+\sigma_{\min}(M)\mathtt{Error}\,(d(x,y))\big) \\ &= \eta\frac{\pi}{2}\sqrt{n+1}[\![n(\log S+\mathbf{D}+n^{3/2})]\!]+\sqrt{n}\left[\!\!\left[\frac{\sqrt{n}\log n}{\eta}\right]\!\!\right] \\ &= \eta[\![n^{3/2}(\log S+\mathbf{D}+n^{3/2})]\!]+\left[\!\!\left[\frac{n\log n}{\eta}\right]\!\!\right] \\ &\leq [\![n^{3/2}\log S+n^3\mathbf{D}^2\kappa(f)^2]\!] \end{aligned}$$

the last by the bounds on η. Also, using Proposition 6.10,

$$\mathtt{Error}\,\left(\frac{3}{2}\sigma\sqrt{n}\|f(x)\|_\infty\right) \leq [\![\sqrt{n}(\mathbf{D}+\log S)]\!].$$

Therefore, for part (i),

$$\begin{aligned} &\sigma_{\min}(M)d(x,y)-\frac{3}{2}\sigma\sqrt{n}\|f(x)\|_\infty \\ &\quad\leq \mathtt{fl}(\sigma_{\min}(M)d(x,y))-\mathtt{fl}\left(\frac{3}{2}\sigma\sqrt{n}\|f(x)\|_\infty\right)+[\![n^{3/2}\log S+n^3\mathbf{D}^2\kappa(f)^2]\!]+[\![\sqrt{n}(\mathbf{D}+\log S)]\!] \\ &\quad\leq [\![n^{3/2}\log S+n^3\mathbf{D}^2\kappa(f)^2]\!]+[\![\sqrt{n}(\mathbf{D}+\log S)]\!] \\ &\quad= [\![n^{3/2}\log S+n^3\mathbf{D}^2\kappa(f)^2]\!]. \end{aligned}$$

$$\boxed{\textbf{Case I.}\ \min\left\{\mu_{\rm norm}(f,x),\frac{1}{\|f(x)\|_\infty}\right\}=\frac{1}{\|f(x)\|_\infty}}$$

In this case $\kappa(f)\geq\frac{1}{\|f(x)\|_\infty}$ and, therefore, by the hypothesis on u,

$$\begin{aligned}
[\![n^{3/2}\log S+n^3\mathbf{D}^2\kappa(f)^2]\!] &= \mathcal{O}(n^{3/2}\log S+n^3\mathbf{D}^2\kappa(f)^2)\frac{K}{\kappa(f)n(\log S+n^{3/2}\mathbf{D}^2\kappa(f)^2)}\\
&\leq \frac{\sigma\sqrt{n}}{2\kappa(f)}\leq\frac{\sigma\sqrt{n}\|f(x)\|_\infty}{2}
\end{aligned}$$

the last line by taking K small enough. This implies that $\sigma_{\min}(M)d(x,y)\leq 2\sigma\sqrt{n}\|f(x)\|_\infty$, i.e., that $d(x,y)\leq 2\sigma\overline{\beta}(f,x)$.

$$\boxed{\textbf{Case II.}\ \min\left\{\mu_{\rm norm}(f,x),\frac{1}{\|f(x)\|_\infty}\right\}=\mu_{\rm norm}(f,x)}$$

In this case $\kappa(f)\geq\mu_{\rm norm}(f,x)=\frac{\sqrt{n}}{\sigma_{\min}(M)}$. By the hypothesis on u

$$\begin{aligned}
[\![n^{3/2}\log S+n^3\mathbf{D}^2\kappa(f)^2]\!] &= \mathcal{O}(n^{3/2}\log S+n^3\mathbf{D}^2\kappa(f)^2)\frac{K}{\mathbf{D}^2n^{5/2}\kappa(f)^3(\log S+n^{3/2}\mathbf{D}^2\kappa(f)^2)}\\
&\leq \frac{\sqrt{n}\alpha_\bullet}{48\mathbf{D}^2(n+1)^{3/2}\kappa(f)^3}\\
&\leq \frac{\sqrt{n}\eta}{8\sqrt{n+1}\kappa(f)}\leq\frac{\sqrt{n}d(x,y)}{4\kappa(f)}\leq\frac{\sigma_{\min}(M)d(x,y)}{4}
\end{aligned}$$

by taking K small enough and Lemma 6.14. This implies that $\frac{3}{4}\sigma_{\min}(M)d(x,y)\leq\frac{3}{2}\sigma\sqrt{n}\|f(x)\|_\infty$, i.e., that $d(x,y)\leq 2\sigma\overline{\beta}(f,x)$.

This shows part (i). Part (ii) is shown in a similar way. □

Lemma 6.17. *Let $u\leq\frac{K\eta^2}{\log n}$ with K small enough and $x_1,x_2\in\mathcal{G}_\eta$.*

(i) *If $\mathtt{fl}(d(x_1,x_2))\leq\mathtt{fl}(\frac{3}{2}\pi\eta\sqrt{n+1})$ then $d(x_1,x_2)\leq 2\pi\eta\sqrt{n+1}$.*

(ii) *If $\mathtt{fl}(d(x_1,x_2))>\mathtt{fl}(\frac{3}{2}\pi\eta\sqrt{n+1})$ then $d(x_1,x_2)>\pi\eta\sqrt{n+1}$.*

Proof. By Lemma 6.15 and the hypothesis on u, we obtain

$$\mathtt{Error}\,(d(x_1,x_2))=\left[\!\!\left[\frac{\sqrt{n}\log n}{\eta}\right]\!\!\right]\leq\mathcal{O}\left(\frac{\sqrt{n}\log n}{\eta}\right)\frac{K\eta^2}{\log n}\leq\frac{\pi}{2}\eta\sqrt{n+1},$$

the last by taking K small enough. Also, $\mathtt{Error}\,(\frac{3}{2}\pi\eta\sqrt{n+1})\leq\frac{3}{2}\pi\eta\sqrt{n+1}\ \gamma_3$. The statement easily follows from these two bounds. □

Lemma 6.18. *Let $u\leq\frac{K\eta\sqrt{n\mathbf{D}}}{\mathbf{D}+\log S+\eta\sqrt{n\mathbf{D}}}$ with K small enough, $f\in S(\mathcal{H}_{\mathbf{d}})$ and $x\in S^n$.*

(i) *If $\mathtt{fl}(\|f(x)\|_\infty)\leq\mathtt{fl}(\frac{\sqrt{2}}{2}\pi\eta\sqrt{(n+1)\mathbf{D}})$ then $\|f(x)\|_\infty\leq\pi\eta\sqrt{(n+1)\mathbf{D}}$.*

(ii) *If $\mathtt{fl}(\|f(x)\|_\infty)>\mathtt{fl}(\frac{\sqrt{2}}{2}\pi\eta\sqrt{(n+1)\mathbf{D}})$ then $\|f(x)\|_\infty>\frac{\pi}{2}\eta\sqrt{(n+1)\mathbf{D}}$.*

Proof. For part (i), from Proposition 6.10,

$$\|f(x)\|_\infty \le \mathtt{fl}(\|f(x)\|_\infty) + [\![\mathbf{D} + \log S]\!].$$

Also,

$$\frac{\sqrt{2}}{2}\pi\eta\sqrt{(n+1)\mathbf{D}} \ge \mathtt{fl}(\frac{\sqrt{2}}{2}\pi\eta\sqrt{(n+1)\mathbf{D}}) - [\![\eta\sqrt{(n+1)\mathbf{D}}]\!].$$

Therefore,

$$\begin{aligned}
\|f(x)\|_\infty - \frac{\sqrt{2}}{2}\pi\eta\sqrt{(n+1)\mathbf{D}} &\le \mathtt{fl}(\|f(x)\|_\infty) - \mathtt{fl}(\frac{\sqrt{2}}{2}\pi\eta\sqrt{(n+1)\mathbf{D}}) + [\![\mathbf{D} + \log S + \eta\sqrt{(n+1)\mathbf{D}}]\!]\\
&\le [\![\mathbf{D} + \log S + \eta\sqrt{(n+1)\mathbf{D}}]\!]\\
&= \mathcal{O}\big(\mathbf{D} + \log S + \eta\sqrt{(n+1)\mathbf{D}}\big)\frac{K\eta\sqrt{n\mathbf{D}}}{\mathbf{D} + \log S + \eta\sqrt{n\mathbf{D}}}\\
&\le (1 - \frac{\sqrt{2}}{2})\eta\sqrt{(n+1)\mathbf{D}},
\end{aligned}$$

the last by taking K sufficiently small. It follows that $\|f(x)\|_\infty \le \pi\eta\sqrt{(n+1)\mathbf{D}}$ and hence, part (i) of the statement.

Part (ii) is proved similarly. □

6.5 Proof of Theorem 1.1(4): Correctness

We will show that, if $u \le \frac{1}{\mathcal{O}\big(\mathbf{D}^2 n^{5/2}\kappa(f)^3(\log S + n^{3/2}\mathbf{D}^2\kappa(f)^2)\big)}$, and the algorithm halts with $\eta \ge \frac{\alpha_\bullet}{8\mathbf{D}^2(n+1)\kappa(f)^2}$, then the value $r/2$ returned by the algorithm is $\#_{\mathbb{R}}(f)$. This is a consequence of the floating following versions of Lemmas 5.1 and 5.2.

Lemma 6.19. *Let $f \in S(\mathcal{H}_{\mathbf{d}})$, $\eta \ge \frac{\alpha_\bullet}{8\mathbf{D}^2(n+1)\kappa(f)^2}$ and $u \le \frac{1}{\mathcal{O}\big(\mathbf{D}^2 n^{5/2}\kappa(f)^3(\log S + n^{3/2}\mathbf{D}^2\kappa(f)^2)\big)}$.*

(i) *For each $x \in \mathtt{fl}(A'(f))$ there exists $\zeta_x \in Z(f)$ such that $\zeta_x \in \overline{B}_f(x)$. Moreover for each point $z \in \mathtt{fl}(\overline{B}'_f(x))$, the Newton sequence starting at z converges to ζ_x.*

(ii) *Let $x, y \in \mathtt{fl}(A'(f))$. Then $\zeta_x = \zeta_y \iff \mathtt{fl}(\overline{B}'_f(x)) \cap \mathtt{fl}(\overline{B}'_f(y)) \ne \emptyset$.*

Proof. (i) Applying Proposition 6.13(ii), $x \in \mathtt{fl}(A'(f))$ implies that $\overline{\alpha}(f,x) < \alpha_\bullet$. Therefore, by Theorem 6.1, there exists $\zeta_x \in Z(f)$ such that $\zeta_x \in \overline{B}_f(x)$. Moreover, if $z \in \mathtt{fl}(\overline{B}'_f(x))$, by Lemma 6.16(i), $d(x,z) \le 2\sigma\overline{\beta}(f,x)$ and the Newton sequence starting at z converges to ζ_x.
(ii) If $\zeta_x = \zeta_y$, then $\overline{B}_f(x) \cap \overline{B}_f(y) \ne \emptyset$ which implies by Lemma 6.16(ii) that there exists $z \in \mathtt{fl}(\overline{B}'_f(x)) \cap \mathtt{fl}(\overline{B}'_f(y))$. □

This immediately implies, using that $\overline{B}_f(x) \subset \mathtt{fl}(\overline{B}'_f(x))$ by Lemma 6.16(ii), the following corresponding floating version of Lemma 5.2.

Lemma 6.20. *Let $f \in S(\mathcal{H}_{\mathbf{d}})$, $\eta \ge \frac{\alpha_\bullet}{8\mathbf{D}^2(n+1)\kappa(f)^2}$ and $u \le \frac{1}{\mathcal{O}\big(\mathbf{D}^2 n^{5/2}\kappa(f)^3(\log S + n^{3/2}\mathbf{D}^2\kappa(f)^2)\big)}$.*

(i) *For each component $\mathtt{fl}(U)$ of $\mathtt{fl}(G'_\eta)$, there is a unique zero $\zeta_U \in Z(f)$ such that $\zeta_U \in Z(\mathtt{fl}(U))$. Moreover $\zeta_U \in \cap_{x\in\mathtt{fl}(U)}\overline{B}_f(x)$.*

(ii) *If* $\mathtt{fl}(U)$ *and* $\mathtt{fl}(V)$ *are different components of* $\mathtt{fl}(G'_\eta)$*, then* $\zeta_U \neq \zeta_V$. $\square$

In order to show the correctness of `Count_Roots_2`, we only need to prove that $Z(f) \subset Z(\mathtt{fl}(G'_\eta))$. This easily follows adapting the proof of Part (1) in Section 5.3 to this situation, making use of Lemma 6.20 and the facts that Condition (i), $\mathtt{fl}(d(x_i, x_j)) > \mathtt{fl}(\frac{3}{2}\pi\eta\sqrt{n+1})$, implies that $d(x_i, x_j) > \pi\eta\sqrt{n+1}$ (Lemma 6.17(ii)) and Condition (ii), $\mathtt{fl}(\|f(x)\|_\infty) > \mathtt{fl}(\frac{\sqrt{2}}{2}\pi\eta\sqrt{(n+1)\mathbf{D}})$, implies that $\|f(x)\|_\infty > \frac{\pi}{2}\eta\sqrt{(n+1)\mathbf{D}}$ (Lemma 6.18(ii)).

6.6 Proof of Theorem 1.1(4): Complexity

We want to show that if $\eta \leq \frac{\alpha_\bullet}{4\mathbf{D}^2(n+1)\kappa(f)^2}$ then `Count_Roots_2`(f) halts. Note that this means that

$$\frac{\alpha_\bullet}{8\mathbf{D}^2(n+1)\kappa(f)^2} < \eta \leq \frac{\alpha_\bullet}{4\mathbf{D}^2(n+1)\kappa(f)^2}$$

and hence, by § 6.5, that it correctly returns $\#_{\mathbb{R}}(f)$.

Because of the hypothesis on η, the hypotheses of Lemmas 6.2, and 6.3 are satisfied. Let $\mathtt{fl}(U) \neq \mathtt{fl}(V)$ be different components of $\mathtt{fl}(G'_\eta)$, and therefore, by Lemma 6.20, $\zeta_U \neq \zeta_V$, and for all $x \in \mathtt{fl}(U)$, $y \in \mathtt{fl}(V)$, by Lemma 6.2, $d(x,y) > 2\pi\eta\sqrt{n+1}$ holds. This implies, by Lemma 6.17(i), that Condition (i) in `Count_Roots_2` is satisfied.

Consider now $x \notin \mathtt{fl}(A'(f))$. By Proposition 6.13(i), $\overline{\alpha}(f,x) \geq \frac{\alpha_\bullet}{3}$. This implies, by Lemma 6.3, that $\|f(x)\|_\infty > \pi\eta\sqrt{(n+1)\mathbf{D}}$, which in turn, by Lemma 6.18(i), ensures that Condition (ii) in `Count_Roots_2` is satisfied. Hence, the algorithm halts.

Aknowledgement. We are grateful to André Galligo for a helpful discussion.

References

[1] B. Bank, M. Giusti, J. Heintz, and L. Pardo. Generalized polar varieties: geometry and algorithms. *J. Compl.*, 21:377–412, 2005.

[2] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation*. Springer-Verlag, 1998.

[3] P. Bürgisser and F. Cucker. Counting complexity classes for numeric computations II: Algebraic and semialgebraic sets. *J. Compl.*, 22:147–191, 2006.

[4] D. Cheung and F. Cucker. Solving linear programs with finite precision: II. Algorithms. *J. Compl.*, 22:305–335, 2006.

[5] G.E. Collins. *Quantifier elimination for real closed fields by cylindrical algebraic deccomposition*, volume 33 of *Lect. Notes in Comp. Sci.*, pages 134–183. Springer-Verlag, 1975.

[6] F. Cucker. Approximate zeros and condition numbers. *J. Compl.*, 15:214–226, 1999.

[7] F. Cucker and J. Peña. A primal-dual algorithm for solving polyhedral conic systems with a finite-precision machine. *SIAM J. Optim.*, 12:522–554, 2002.

[8] F. Cucker and S. Smale. Complexity estimates depending on condition and round-off error. *Journal of the ACM*, 46:113–184, 1999.

[9] F. Cucker and D.X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Univ. Press, 2007.

[10] J.-P. Dedieu, P. Priouret, and G. Malajovich. Newton method on Riemannian manifolds: Covariant alpha-theory. *IMA Journal of Numerical Analysis*, 23:395–419, 2003.

[11] G. Golub and C. Van Loan. *Matrix Computations*. John Hopkins Univ. Press, 3rd edition, 1996.

[12] D.Yu. Grigoriev. Complexity of deciding Tarski algebra. *Journal of Symbolic Computation*, 5:65–108, 1988.

[13] D.Yu. Grigoriev and N.N. Vorobjov. Solving systems of polynomial inequalities in subexponential time. *Journal of Symbolic Computation*, 5:37–64, 1988.

[14] D.Yu. Grigoriev and N.N. Vorobjov. Counting connected components of a semialgebraic set in subexponential time. *Computational Complexity*, 2:133–186, 1992.

[15] Y. Han and R.A. Wagner. An efficient and fast parallel-connected component algorithm. *Journal of the ACM*, 37(3):626–642, 1990.

[16] J. Heintz, M.-F. Roy, and P. Solerno. Single exponential path finding in semi-algebraic sets II: The general case. In C.L. Bajaj, editor, *Algebraic Geometry and its Applications*, pages 449–465. Springer-Verlag, 1994.

[17] N. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 1996.

[18] T.Y. Li. Numerical solution of polynomial systems by homotopy continuation methods. In P.G. Ciarlet and F. Cucker, editors, *Handbook of numerical analysis*, volume 11, pages 209–304. North-Holland, 2003.

[19] G. Malajovich. On generalized Newton algorithms: Quadratic convergence, path-following and error analysis. *Theoret. Comp. Sci.*, 133:65–84, 1994.

[20] K. Meer. Counting problems over the reals. *Theoret. Comp. Sci.*, 242:41–58, 2000.

[21] M. Shub and S. Smale. Complexity of Bézout's theorem I: geometric aspects. *Journal of the Amer. Math. Soc.*, 6:459–501, 1993.

[22] M. Shub and S. Smale. Complexity of Bézout's theorem III: condition number and packing. *Journal of Complexity*, 9:4–14, 1993.

[23] M. Shub and S. Smale. Complexity of Bézout's theorem IV: probability of success; extensions. *SIAM J. of Numer. Anal.*, 33:128–148, 1996.

[24] 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.

[25] A. Tarski. *A Decision Method for Elementary Algebra and Geometry*. University of California Press, 1951.

[26] H. Weyl. *The Theory of Groups and Quantum Mechanics*. Dover, 1932.

[27] H.R. Wüthrich. Ein Entscheidungsverfahren für die Theorie der reell-abgeschlossenen Körper. volume 43 of *Lect. Notes in Comp. Sci.*, pages 138–162. Springer-Verlag, 1976.