Sign Methods for Enumerating Solutions of Nonlinear Algebraic Systems

Ioannis Z. Emiris[1], Bernard Mourrain[2], Michael N. Vrahatis[3]

Abstract— **We implement the concept of topological degree to isolate and compute all zeros of systems of nonlinear algebraic equations when the only computable information required is the algebraic signs of the components of the function. The basic theorems of Kronecker–Picard theory relate the number of roots to the topological degree. They are combined with grid methods in order to compute the topological degree by using the minimum possible information, namely the sign of a function at some value. Recent fast methods, which work over fixed precision, are applied to determine the sign of algebraic systems.**

Keywords—**Polynomial system solving, zero isolation, bisection, characteristic polytope, topological degree, sign determination.**

I. Introduction

ALGEBRAIC equations are instrumental in studying and solving problems on geometric, kinematic, and other constraints in a variety of fields including modeling and graphics, robotics, vision, molecular biology, signal processing, and computational economics; see [3, 5, 9, 10] and the references thereof. Most of the algorithms of this article apply to the wider class of analytic functions.

We exploit methods based on the minimum possible information, namely the algebraic sign of the (algebraic) function. We rely on the concept of topological degree in order to isolate and compute all zeros of systems of nonlinear algebraic equations given only the sign. Such methods can be applied to problems with imprecise function values. This is significant for problems where the function values follows as a result of an infinite series such as Bessel or Airy functions, because the sign stabilizes after a relatively small number of terms [17]. Topological degree methods can also be extended to counting and computing the extrema of systems of equations.

This paper is organized as follows. The next section offers a background on topological degree. Then we explain bisection methods for solving systems of equations, in particular the notion of characteristic polytopes. Section IV discusses algorithms for computing the topological degree and thus counting and isolating roots. In section V we analyze fast and accurate methods for computing the sign of algebraic functions over fixed precision.

[1] INRIA, B.P. 93, Sophia-Antipolis 06902 France, Ioannis.Emiris@inria.fr

[2] INRIA, B.P. 93, Sophia-Antipolis 06902 France, Bernard.Mourrain@inria.fr

[3] Department of Mathematics, University of Patras Artificial Intelligence Research Center (UPAIRC), University of Patras, GR-26110 Patras, Greece, vrahatis@math.upatras.gr

II. Topological degree basics

We briefly exploit topological degree theory and Picard's extension [6] for determining the exact number of real roots (and extrema) of a system of nonlinear algebraic equations. Isolation and approximation of all real roots as well as computation of the topological degree are all discussed in the next sections.

Suppose that the function $F_n = (f_1, \ldots, f_n)$: $\overline{\mathcal{D}}_n \subset \mathbb{R}^n \to \mathbb{R}^n$ is defined and two times continuously differentiable on the closure of an open and bounded domain $\mathcal{D}_n$ of $\mathbb{R}^n$ with boundary $\mathrm{b}(\mathcal{D}_n)$. Suppose further that the roots of the equation $F_n(x) = \Theta_n$, ($\Theta_n = (0, \ldots, 0)$ denotes the origin of $\mathbb{R}^n$), are not located on the boundary $\mathrm{b}(\mathcal{D}_n)$ of $\mathcal{D}_n$ and they are simple i.e. the determinant $\det J_{F_n}$ of the Jacobian J_{F_n} of F_n at these roots is nonzero.

Definition II.1: If $F_n|_{\overline{\mathcal{D}}_n} \in C^2$ and all roots in $\mathcal{D}_n \subset \mathbb{R}$ are simple and do not lie on $\mathrm{b}(\mathcal{D}_n)$, then the *topological degree of F_n at Θ_n relative to $\mathcal{D}_n$* is denoted by $\deg[F_n, \mathcal{D}_n, \Theta_n]$ and can be defined by the following sum:

$$\deg[F_n, \mathcal{D}_n, \Theta_n] = \sum_{x \in \mathcal{D}_n : F_n(x) = \Theta_n} \operatorname{sgn} \det J_{F_n}(x),$$

where sgn denotes the sign function.

When F_n is at least continuous, then $F_n(x) = \Theta_n$ has at least one root in $\mathcal{D}_n$ if $\deg[F_n, \mathcal{D}_n, \Theta_n] \neq 0$. Furthermore, if $\mathcal{D}_n = \mathcal{D}_n^1 \cup \mathcal{D}_n^2$, where $\mathcal{D}_n^1$ and $\mathcal{D}_n^2$ have disjoint interiors, then $\deg[F_n, \mathcal{D}_n, \Theta_n] = \deg[F_n, \mathcal{D}_n^1, \Theta_n] + \deg[F_n, \mathcal{D}_n^2, \Theta_n]$. We may extend this theory to counting complex roots by doubling the dimension. For instance, the total number of simple complex zeros of an analytic function $f : \mathcal{D}_2 \subset \mathbb{C} \to \mathbb{C}$ in an open bounded region $\mathcal{D}_2$ is equal to $\deg[F_2, \mathcal{D}_2, \Theta_2]$, where $F_2 = (f_1, f_2)$, with $f_1 = \Re\{f(x_1 + ix_2)\}$, $f_2 = \Im\{(x_1 + ix_2)\}$.

Picard considered the extension $F_{n+1} = (f_1, \ldots, f_n, f_{n+1}) : \mathcal{D}_{n+1} \subset \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$, where $f_{n+1} = y \det J_{F_n}$, y is the $(n+1)$-st variable, and $\mathcal{D}_{n+1} = \mathcal{D}_n \times I$, with I being an arbitrary interval of the y–axis containing the point $y = 0$. Then the roots of the following system of equations:

$$f_i(x_1, x_2, \ldots, x_n) = 0, \quad i = 1, \ldots, n,$$
$$y \det J_{F_n}(x_1, x_2, \ldots, x_n) = 0,$$

are the same simple roots of $F_n(x) = \Theta_n$ provided $y = 0$. On the other hand it is easily seen that the determinant of the Jacobian of F_{n+1} is equal to $[J_{F_n}(x)]^2$.

Theorem II.1: The total number of real roots of $F_n(x) = \Theta_n$ equals $\deg[F_{n+1},\ \mathcal{D}_{n+1},\ \Theta_{n+1}]$, provided that F_n is two times continuously differentiable and that all roots are simple and lie in the strict interior of $\mathcal{D}_{n+1}$.

III. Bisection methods for computing the roots

This section uses bisection methods, based on the topological degree, for solving arbitrary systems of analytic functions. Our focus is on a generalized bisection method using characteristic polytopes.

Let us concentrate on one-dimensional problems for simplicity. For a continuous function f, it is well known that a solution of $f(x) = 0$ is guaranteed to exist in some interval $[a, b]$ where $f(a)f(b) \neq 0$ if $f(a)\, f(b) \leqslant 0$. This criterion is known as Bolzano's existence criterion and, essentially, transfers all information regarding the roots to the boundary of the given region. We can use Bolzano's criterion or topological degree to calculate a solution of f by bisecting the interval $I_0 = (a, b)$ into two intervals $(a, c], [c, b)$ where $c = (a + b)/2$ so that we always keep a solution within a smaller interval. This is called the *bisection method* and can be generalized to higher dimensions [16].

Instead of Bolzano's criterion we may use the value $\deg[f, (a, b), 0]$ which equals $\frac{1}{2}\left(\operatorname{sgn} f(b) - \operatorname{sgn} f(a)\right)$. If this value is not zero we know with certainty that there is at least one solution in (a, b). It also gives additional information concerning the behavior of the solutions of f in (a, b) relative to the slopes of f and can be generalized to higher dimensions [14, 15]. If $f\colon [a, b] \subset \mathbb{R} \to \mathbb{R}$ is continuous, a simplified version of the bisection method leads to the formula $x_{i+1} = x_i + \operatorname{sgn} f(x_0) \operatorname{sgn} f(x_i)\, (b-a)/2^{i+1}$, for $i = 0, 1, \ldots$, with $x_0 = a$. This sequence converges to a root $r \in (a, b)$ if for some x_i, $\quad i = 1, 2, \ldots$, $\operatorname{sgn} f(x_0) \operatorname{sgn} f(x_i) = -1$. An analogous formula holds otherwise [14, 15]. The number of iterations for obtaining an approximate root r^* such that $|r - r^*| \leqslant \varepsilon$ for some $\varepsilon \in (0, 1)$ is given by $\lceil \log_2((b-a)\, \varepsilon^{-1}) \rceil$. The method always converges within the given interval and is a global convergence method. Moreover it has the great advantage that it possesses asymptotically the best possible rate of convergence. Additionally, it is easy to have beforehand the number of iterations that are required for the attainment of an approximate root to a predetermined accuracy. Finally, it requires only the algebraic signs of the functions values to be computed.

The rest of the section examines systems of n equations in n variables and describes characteristic polytopes. We implement topological degree theory to provide a criterion for the existence of a zero of $F_n(x) = \Theta_n$ within a given region. Once a zero has been isolated, this procedure approximates its value to any desired accuracy [14, 15]. Suppose that $\mathcal{P}_n = \langle V_1, V_2, \ldots, V_{2^n} \rangle$ is an oriented n–dimensional polytope with 2^n vertices, $V_i \in \mathbb{R}^n$, (i.e. an orientation has been assigned to its vertices), and let $F_n = (f_1, f_2, \ldots, f_n)\colon \mathcal{P}_n \subset \mathbb{R}^n \to \mathbb{R}^n$ be a continuous mapping. Let $\mathcal{S}(F_n, \mathcal{P}_n)$ be the $2^n \times n$ matrix whose entries in the k-th row are the corresponding coordinates of vector $\Big(\operatorname{sgn} f_1(V_k), \operatorname{sgn} f_2(V_k), \ldots, \operatorname{sgn} f_n(V_k)\Big)$. This will be called *matrix of signs* associated with F_n and $\mathcal{P}_n$. Then, $\mathcal{P}_n$ is called a *characteristic* n–polytope relative to F_n, if $\mathcal{S}(F_n; \mathcal{P}_n)$ is identical, possibly

after some permutations of its rows, with a $2^n \times n$ matrix whose rows are formed by all possible combinations of $-1,\ 1$.

Theorem III.1: [13]. Let $\Pi = \{\Pi_i\}_{i=1}^{2n}$ be the set of facets of characteristic polytope $\mathcal{P}_n$. Let $S = \{S_{i,j}\}_{i=1,j=1}^{2n\ \ j_i}$, where $j_i \geqslant 1$ denotes the number of simplices corresponding to facet Π_i, be a set of $(n-1)$-simplices which lie on $\mathrm{b}(\mathcal{P}_n)$. Suppose that:
(a) $\mathrm{b}(\mathcal{P}_n) = \sum_{i=1}^{2n} \sum_{j=1}^{j_i} S_{i,j}$,
(b) the interiors of the members of S are disjoint,
(c) these simplices make $\mathrm{b}(\mathcal{P}_n)$ sufficiently refined relative to sgn F_n, and
(d) the extreme points of simplices $S_{i,j}$ for any j are vertices of Π_i.
Then $\deg[F_n, \mathcal{P}_n, \Theta_n] = \pm 1 \neq 0$.

When the topological degree is nonzero, this implies that there is at least one root inside $\mathcal{P}_n$, although the converse is not always true. The condition on the sufficient refinement guarantees, informally, that no more than one function changes sign between two successive critical points of the boundary. These points are precisely the points where the algebraic signs are computed. Section IV formalizes and quantifies this discussion; see also [7, 11, 12].

The above theorem yields an algorithmic procedure for the approximation of isolated roots. In two dimensions, the algorithm bisects the longest edge of the characteristic quadrilateral thus defining a new one, which has again vertices corresponding to all four combinations of two signs. In order to approximate the value of the root with accuracy ϵ, the algorithm takes $O(lgD/\epsilon)$ steps, where D is the initial diameter of the characteristic polygon.

IV. Computation of topological degree

Several analytic formulae for the computation of the topological degree have been proposed, particularly those based on the Kronecker integral. These methods count the number of roots in a given region and, therefore, lead to algorithms for the isolation of roots. Notice that root isolation is not straightforward with other methods such as those based on characteristic polytopes. Hence, topological degree computation can be the first step in a root isolation and approximation package; see, e.g. [11]. Here we focus on a method by Kearfott [7] which can be combined with the characteristic polytope algorithm and compares favourably to other methods in efficiency. The second part of the section briefly analyzes the complexity of a method for sufficiently refining the boundary.

Suppose that $S^{n-1} = \langle x_1, x_2, \ldots, x_n \rangle$ is an $(n-1)$–simplex in $\mathbb{R}^n$ and assume $F_n = (f_1, f_2, \ldots, f_n) : S^{n-1} \to \mathbb{R}^n$ is continuous. Then the *range simplex associated with* S^{n-1} *and* F_n, denoted by $\mathcal{R}(S^{n-1}, F_n)$, is an $n \times n$ matrix with elements $\varrho_{ij},\ 1 \leqslant i, j \leqslant n$ such that $\varrho_{ij} = 1$ if $f_j(x_i) \geqslant 0$ and $\varrho_{ij} = -1$ otherwise. $\mathcal{R}(S^{n-1}, F_n)$ is called *usable* if, after a permutation of its rows, the elements ϱ_{ij} of the matrix are: $\varrho_{ij} = 1$, if $i \geqslant j$, and $\varrho_{ij} = -1$, if $j = i + 1$.

When $\mathcal{R}(S^{n-1}, F_n)$ is usable, then its parity $\mathrm{Par}\Big(\mathcal{R}(S^{n-1}, F_n)\Big)$ is defined to be 1, if the number of the permutations of the rows required to put it into usable form is even. Otherwise the parity equals -1. For all other cases, we set the parity to zero. Suppose that $\mathcal{P}^n$ is an n–dimensional polytope for some $n \geqslant 2$; possibly a characteristic polytope. Furthermore, suppose that $\{S_i^{n-1}\}_{i=1}^m$ is a finite set of $(n-1)$–simplices with disjoint interiors such that $\sum_{i=1}^m S_i^{n-1} = \mathrm{b}(\mathcal{P}^n)$. If the boundary of S_i^{n-1} is sufficiently refined relative to the signs of F_n, then

$$\deg[F_n, \mathcal{P}^n, \Theta_n] = \sum_{i=1}^{m} \mathrm{Par}\Big(\mathcal{R}(S_i^{n-1}, F_n)\Big).$$

The rest of the section addresses the issue of ensuring sufficient refinement of the boundary of the examined region, which is the main factor in establishing the method's overall complexity. We base our approach on the deterministic algorithm of Stynes [12] in order to derive bounds on the number of sign determinations. This article's main algorithm in section 4 starts with a collection of k_0 simplices and successively subdivides them in such a way as to sufficiently refine the polytope they bound. Let k_i be the total number of simplices at step i, for $i = 0, \ldots, \mu$, and let d_i be the maximum edge length among all simplices at step i.

In what follows we analyze the algorithm for the 2-dimensional case, where the polytope is a polygon and the simplices are its 1-dimensional edges. An analogous analysis holds in more than 2 dimensions. At every step, the algorithm splits all

edges whose length exceeds $2d_i/\sqrt{5}$, hence $d_i \leqslant d_0(2/\sqrt{5})^i$. On the 2-dimensional plane we can deduce that $k_i = -1 + 2k_{i-1}$, for $i \geqslant 1$, therefore $k_i = 1 + 2^i(k_0 - 1)$. The termination condition is that $d_\mu < \delta$, where δ depends on the modulus of continuity of the given (analytic) system. It suffices that $d_0(2/\sqrt{5})^\mu < \delta$, which is guaranteed for $\mu = \lceil \log_{\sqrt{5}/2}(d_0/\delta) \rceil$ (this is the worst case value of μ). Thus, we can bound the final number of simplices $k_\mu < 3k_0(d_0/\delta)^{6.3}$.

Bounding δ can be done in several ways, usually as a function of the infinite norm of the polynomials on the domain boundary and of the respective Lipschitz constants on the entire domain. These norms and constants may be bounded either by using analytic formulae or by interval analysis. These are recurrent and deep problems in topological degree estimation, (see e.g. [1]), so we do not pretend to offer original final solutions. However, we expect that our methods will be efficient enough in practice.

V. Sign determination of algebraic expressions

The complexity of sign computation for root counting and isolation is instrumental in establishing bounds on the overall complexity of our algorithms. This section adapts the methods of Brönnimann, Emiris, Pan and Pion [2] to the present context. We focus on multivariate polynomial expression with rational coefficients, which covers the case of rational expressions as well.

The computation has an exact part and a numeric part which are combined in order to guarantee the exactness of the sign computed. In the first part, exactness is achieved by using modular computation. The modular representation of a rational number includes a sufficient number of moduli, i.e. the projections of the number to the respective finite fields defined by a sequence of prime integers. Let $m_1, \ldots, m_k$ be k pairwise relatively prime integers and let $m = \prod_{i=1}^k m_i$. For any number x (not necessarily an integer), we let $x_i = x \bmod m_i$ be the only number in the range $\left[-\frac{m_i}{2}, \frac{m_i}{2}\right)$ such that $x_i - x$ is a multiple of m_i. Manipulating moduli in the finite field of m_i and computing the x_i, for all i, requires only fixed-precision operations. Now let k, b, $m_1, \ldots, m_k$ denote positive integers, $m_1, \ldots, m_k$ being pairwise relatively prime, such that $m_i \leqslant 2^{b/2+1}$. Let x be an integer whose magnitude is smaller than $\lfloor m/2 \rfloor$. Given the x_i, we have to compute the sign of x by using only floating-point arithmetic performed with b-bit precision.

The second part uses fixed-precision f.p. arithmetic for reconstructing the sign from the moduli of the value. This computation is approximative, but the error is bounded in such a way as to guarantee that the recovered sign is exact thus solving the above problem. One method, named after Lagrange, uses an iteration which stops when the computed value is large enough so that it has the proper sign, despite roundoff error. The algorithm then returns its sign. Intuitively, if x is large, this quantity will have the same sign as x for some large j. Otherwise, fewer moduli suffice to define x so we can consider a smaller j. An alternative method, named after Newton, is incremental and thus can be adapted to a probabilistic algorithm that does not require any bound on the magnitude of x.

Theorem V.1: [2] Assume f is a polynomial whose value in the domain of interest has modulus bounded by $\lfloor m/2 \rfloor$, where $m = \prod_{i=1}^k m_i$ and the m_i are fixed-precision relatively prime integers as above. In the worst case, the total bit complexity of both Lagrange and Newton methods for exact sign reconstruction is in $O(k^2)$.

By this theorem, these methods do not improve asymptotically over the standard multiprecision approach. The methods are simple, however, require little or no overhead, and their parallel time complexity is in $O(k \log k)$. To exploit modern day hardware, we exclusively rely on floating point (f.p.) numbers. In practice, our methods have complexity lower than that estimated above and compares favorably with known multiprecision methods, thus they are very well suited for implementation, as illustrated in [2].

Remark V.1: In practice, Lagrange's technique performs $O(k)$ fixed-precision operations, which can be guaranteed when x is large, i.e. where (almost) all moduli are required in its representation. On the other hand, Newton's method has $O(k)$ total bit complexity when x is small, i.e. can be defined by a constant number of moduli.

Let us consider an arbitrary n-variate dense polynomial, with degree d in each variable. Assume that its sign must be determined on a grid of p^n points, defined by p distinct coordinates per dimension. Then the total bit complexity for all sign determinations is in $O(d^2p^n)$. Based on the bounds of the previous remark, this complexity becomes $O(dp^n)$. Moreover, the aggregate complexity may be reduced by exploiting the fact that the evaluation points usually lie on the axes after a suitable linear transformation. More sophisticated techniques can be applied if the polynomial has a certain form, e.g., it is expressed by a determinant.

References

[1] T. Boult and K. Sikorski, *An optimal complexity algorithm for computing the topological degree in two dimensions*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 686–698.

[2] H. Brönnimann, I.Z. Emiris, V. Pan, and S. Pion, *Sign determination in Residue Number Systems,* Theoretical Computer Science, Special Issue on Real Numbers and Computers, 210 (1999), pp. 173–197.

[3] P. Comon and B. Mourrain, *Decomposition of quantics in sums of powers*, In F. T. Luk, editor, *Proc. Adv. Signal Proc.: Algorithms, Architectures and Implementations*, vol. 2296, pp. 93–104, SPIE, 1994.

[4] J. Cronin, *Fixed points and topological degree in nonlinear analysis*, Mathematical Surveys No. 11, Amer. Math. Soc., Providence, Rhode Island, 1964.

[5] I.Z. Emiris, Symbolic-numeric algebra for polynomials, In *Encyclopedia of Computer Science and Technology*, vol. 39, pp. 261–281, Marcel Dekker, New York, 1998.

[6] B.J. Hoenders and C.H. Slump, *On the determination of the number and multiplicity of zeros of a function*, Computing, 47 (1992), pp. 323–336.

[7] R.B. Kearfott, *An efficient degree–computation method for a generalized method of bisection*, Numer. Math., 32 (1979), pp. 109–127.

[8] R.B. Kearfott, *Some tests of generalized bisection*, ACM Trans. Math. Software, 13 (1987), pp. 197–220.

[9] D. Manocha and J. Demmel. Algorithms for intersecting parametric and algebraic curves II: Multiple intersections. *Graphical Models and Image Proc.*, 57 (1995), pp. 81–100.

[10] J. Nielsen and B. Roth. Elimination methods for spatial synthesis. In J.-P. Merlet and B. Ravani, editors, *Computational Kinematics '95*, pp. 51–62. Kluwer, 1995.

[11] F. Stenger, *Computing the topological degree of a mapping in* $\mathbb{R}^n$, Numer. Math., 25 (1975), pp. 23–38.

[12] M. Stynes, *On the construction of sufficient refinements for computation of topological degree*, Numer. Math., 37 (1981), pp. 453–462.

[13] M.N. Vrahatis and K.I. Iordanidis, *A rapid generalized method of bisection for solving systems of non-linear equations*, Numer. Math., 49 (1986), pp. 123–138.

[14] M.N. Vrahatis, *Solving systems of nonlinear equations using the nonzero value of the topological degree*, ACM Trans. Math. Software, 14 (1988), pp. 312–329.

[15] M.N. Vrahatis, *CHABIS: A mathematical software package for locating and evaluating roots of systems of nonlinear equations*, ACM Trans. Math. Software, 14 (1988), pp. 330–336.

[16] M.N. Vrahatis, *A short proof and a generalization of Miranda's existence theorem*, Proc. Amer. Math. Soc., 107 (1989), pp. 701–703.

[17] M.N. Vrahatis, T.N. Grapsa, O. Ragos and F.A. Zafiropoulos, *On the localization and computation of zeros of Bessel functions*, Z. Angew. Math. Mech., 77 (1997), pp. 467–475.