COMPUTATION OF TOPOLOGICAL DEGREE
USING INTERVAL ARITHMETIC, AND APPLICATIONS

OLIVER ABERTH

ABSTRACT. A method is described for computing the topological degree of a mapping from R^n into R^n defined by n functions of n variables on a region specified as a product of n intervals, a generalized box B. The method is an adaptation of Kearfott's method to boxes, and begins by checking the signs of the n functions on the boundary of B with interval arithmetic. On the basis of this check, a portion, $B^{(1)}$, of the boundary of B is designated for further investigation, and one of the n functions defining the mapping is dropped. The signs of the remaining functions are checked on the boundary of $B^{(1)}$. Again a portion, $B^{(2)}$, of the boundary of $B^{(1)}$ is designated for further investigation, and another of the functions is dropped. On the nth cycle of the process, the topological degree finally is evaluated by determining the signs of a single function on a collection of isolated points, comprising the boundary of a region $B^{(n-1)}$.

When the topological degree is nonzero, there is at least one point inside B where the n functions are simultaneously zero. To locate such a point, the familiar bisection method for functions $f(x)$ defined over an interval $[a, b]$, using sign changes of $f(x)$, is easily generalized to apply to n functions defined over boxes, using the topological degree. For this application we actually use the topological degree mod 2, the *crossing parity*, because its computation is easier. If the n functions have all partial derivatives in the box B, with a nonzero Jacobian at any point where the functions are simultaneously zero, then all such points inside B can be located by another method, which also uses the crossing parity.

1. INTRODUCTION

Let B be a closed region of R^n defined as a product of the n intervals

$$(1) \qquad a_i \leq x_i \leq b_i, \qquad i = 1, 2, \ldots, n.$$

When n is 3, the region B is a box, and for arbitrary n we will use this term for the region B. Defined on the box B is the mapping F from R^n into R^n given by the equations

$$y_i = f_i(x_1, \ldots, x_n), \qquad i = 1, 2, \ldots, n,$$

where the functions f_i are all real and continuous on B. A point P_0 in B where all functions f_i are simultaneously zero is usually called a *zero* of the

functions. The computational problem of finding a zero can be solved by a general bisection method [10] using the topological degree to detect the presence of zeros inside subregions of B. Various methods of computing the topological degree have been proposed. O'Neal and Thomas [14] suggest using quadrature methods to evaluate the Kronecker integral formula for topological degree. Stenger [16] describes a method, derived from the Kronecker integral formula, which involves determining only the signs of the functions f_i over portions of the boundary of a polyhedral region of interest. This method was used to develop bisection methods [7, 15] which reduce the region of search to progressively smaller simplexes in R^n. Kearfott describes two methods, one [8] that is related to Stenger's method, and another [9] that is recursive, repeatedly reducing by one the number of functions to be considered. This second Kearfott method is the basis of our method, which is restricted to boxes. Here, interval arithmetic [3, 11, 12, 13] is used to determine the signs of the functions f_i on various k-dimensional subboxes contained in the boundary of the starting box B.

In the next two sections the method is described, and an example of its application is given. The general bisection method we employed is described in §4. When all n functions have continuous first partial derivatives in B, with a nonzero Jacobian at each zero, then all zeros inside B can be found, and a method for doing this is described in §5.

2. A METHOD FOR COMPUTING THE TOPOLOGICAL DEGREE

In this section we use the terminology and notation of combinatorial topology [4, 6]. The term "Jacobian" will denote the determinant of the n-square matrix with elements $\partial f_i / \partial x_j$ evaluated at some specific point. Let θ be the origin in R^n, and suppose that θ does not lie on $F(\partial B)$, the image of the boundary of B. The topological degree $d(F, B, \theta)$ equals $i(C^n, \theta)$, the intersection number of C^n with θ, where C^n is a polyhedral complex with boundary Z^{n-1} such that C^n approximates $F(B)$ to sufficient accuracy [6, pp. 2–55].

For the case where $n \geq 2$, if the functions f_i have all partial derivatives defined and continuous on B, with their Jacobian nonzero at all zeros,

$$i(C^n, \theta) = \sum \operatorname{sgn} \left(\det \frac{\partial (f_1, f_2, \ldots, f_n)}{\partial (x_1, x_2, \ldots, x_n)} \right),$$

where the summation is over the zeros of F. (The function $\operatorname{sgn}(u)$ is $+1$ if u is positive and -1 if u is negative.) Of course, $i(C^n, \theta)$ is defined regardless of whether Jacobians are nonzero or even whether the partial derivatives are defined. The determination of the intersection number can be made by examination of the polyhedral complex used to approximate $F(B)$. To simplify the discussion, we will assume that partial derivatives (or ordinary derivatives) are defined and that the various Jacobians listed are nonzero.

There is the general relation [4, pp. 419–420]

$$i(C^n, \theta) = i(H^1, Z^{n-1}),$$

where H^1 is any one-dimensional ray from θ that is in general position with respect to the boundary cycle Z^{n-1}. To carry out the computation of topological degree as $i(H^1, Z^{n-1})$, we form a list L and load it initially with

designations for the $2n$ sides of B. This can be done efficiently by arranging that a list element contain the fields s_i, $i = 1, \ldots, n$, each field being flagged either as a "point" holding a single number a, or as an "interval" holding a number pair a, b, with $a < b$. We distinguish the two cases by writings $s_i = a$ or $s_i = [a, b]$. Thus the list is initially loaded by constructing, for $j = 1, 2, \ldots, n$, the two list elements:

$$(2) \qquad s_i = \begin{cases} [a_i, b_i] & \text{for } i \neq j, \\ a_j & \text{for } i = j, \end{cases}$$

$$(3) \qquad s_i = \begin{cases} [a_i, b_i] & \text{for } i \neq j, \\ b_j & \text{for } i = j. \end{cases}$$

A list element has one additional field, "orientation", which can assume one of two values, $+1$ or -1. We assign orientation as follows: Let O be the point of B which would become the origin if B were translated so that its edges lie along the positive coordinate axes, and let X_i be the point of B which after the translation would lie furthest along the x_i-axis. The positively oriented simplex $(OX_1X_2\cdots X_n)$ is used to assign an orientation to any side of B that contains the point O. Thus, the side which has $x_j = a_j$ corresponds to the simplex $(OX_1 \cdots X_{j-1}X_{j+1} \cdots X_n)$ with orientation $\sigma = (-1)^j$. The opposite side of B which has $x_j = b_j$ must take the opposite orientation, or $\sigma = (-1)^{j+1}$. Accordingly, the orientation fields of the list elements (2) and (3) are set to these values.

We take H^1 to be the ray

$$y_i = \begin{cases} t \geq 0 & \text{for } i = 1, \\ 0 & \text{for } i > 1. \end{cases}$$

A search is made to find the points where this ray meets the boundary Z^{n-1}. We do this as follows: Taking the first list element, for $i = 1, \ldots, n$, we set the computation variable x_i equal to either the interval or the point value given by the field s_i. Then with the functions $f_2, f_3, \ldots, f_n$, and using interval arithmetic, we generate interval values for $y_2, y_3, \ldots, y_n$. If any of these intervals does not contain the zero point, the ray H^1 cannot intersect this part of the boundary, and the list element is discarded. If all these intervals contain the zero point, then we form an interval for y_1 using f_1. If this interval lies to the left of the zero point, again the ray H^1 cannot intersect this part of the boundary, and the list element is discarded. If the interval lies to the right of the zero point, it is likely the ray H^1 meets this portion of the boundary, and the list element is removed from L and added to a second list L_1 (initially empty) for later attention. In the remaining case where the zero point is contained in the y_1-interval, whether or not the ray H^1 intersects this part of the boundary is uncertain, and we proceed as follows: From the list element we construct all possible list elements that can be formed by replacing interval fields s_i by either their right half or left half subintervals, copying point fields s_i, and copying the orientation value. These 2^{n-1} list elements together then replace the first element of the list L. The whole process described can now be repeated with the new first element of L. (An alternative plan would be to form only two new elements, subdividing only the widest interval, and using these to replace the first element of L.)

If the point θ does not lie on $F(\partial B)$, then the process described will eventually exhaust the list L. The list L_1 may now be examined. In general it contains elements which together define a subset $B^{(1)}$ of the boundary of B. Consider an element on this list with orientation σ defining a point set S lying in a side of B with x_j fixed. The ray H^1 may meet the F image of S at one or more points, and for each of these there will be a point of S where the functions $f_2, f_3, \ldots, f_n$ are simultaneously zero. The contribution made by these points to $i(H^1, Z^{n-1})$ is

$$(4) \qquad \sigma \sum \operatorname{sgn} \left(\det \frac{\partial(f_2, f_3, \ldots, f_n)}{\partial(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)} \right)$$

or σ times the topological degree of the R^{n-1} to R^{n-1} mapping defined by $f_2, f_3, \ldots, f_n$ on the $(n-1)$-dimensional region S (embedded in R^n). We need to sum the contributions (4) over all elements in the list L_1 in order to get $d(F, B, \theta)$.

For this computation we proceed in a similar manner as with B and form boundary elements. For each element of L_1 with orientation σ, we form $2(n-1)$ boundary elements for a new list L by copying all fields s_i except for one interval field which is converted to a point field, using an endpoint of the interval. If the interval field converted was the jth interval field, counting from s_1 toward s_n, then the orientation assigned is σ times $(-1)^j$ or times $(-1)^{j+1}$ according as the left or the right endpoint is assigned. This time, however, regions defined by two elements on the list L may be identical. These elements always have opposite orientations and both elements must be discarded. It is also possible, because of the subdivision process, for the region of one element to be contained within that of a second element. Here the list must be corrected by discarding these elements and replacing them with new elements to represent the unduplicated region. After the list L is corrected, it is certain that the functions $f_2, f_3, \ldots, f_n$ are never simultaneously zero on the region defined by any element. This is because an element appears on the list L only if it is part of the boundary of some element on the list L_1 (so f_1 is positive) and simultaneously is part of the boundary of some element not on the list L_1 (so some function other than f_1 is positive or is negative).

The computation with the list L is similar to that described before, except that the number of functions evaluated is reduced by one. This time the ray H^1 is

$$y_i = \begin{cases} t \geq 0 & \text{for } i = 2, \\ 0 & \text{for } i > 2. \end{cases}$$

The process will exhaust the list L and again yield a list L_1 requiring further examination.

Each time the process is carried out, the dimension of the point sets examined and the number of functions treated is reduced by one. The general description given of the process for the first and second cycles can be revised to cover all instances. For example, the first cycle may be viewed as the case where L_1 has a single list element with orientation $+1$ defining the starting box B. On the nth iteration, after the list L_1 is converted to a list L, all interval fields s_i of an L-element have been replaced by endpoints, so the element designates a point P with an orientation σ. At the end of this iteration, only the point elements with positive f_n-values are on the new list L_1. For this case, the

relevant equation (4) has the number 1 in place of the Jacobian, so summing the orientations of elements on this last L_1-list finally yields $d(F, B, \theta)$.

With this method only signs need to be determined, but the precision of computation necessary depends on the complexity of the functions f_i. If the precision is too low, the intervals obtained may be too wide, enclosing the zero point when they should not. This can cause the subdivision of list elements to continue indefinitely. We programmed the computation using range arithmetic [2], so that, if necessary, the precision could be increased during the computation.

3. An example

Consider the mapping from R^3 into R^3 defined by

$$f_1 = x_1^2 + x_2^2 - x_3, \quad f_2 = x_2^2 + x_3^2 - x_1, \quad f_3 = x_3^2 + x_1^2 - x_2.$$

These functions have a zero at $(0, 0, 0)$ and at $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. We compute the degree for a box B enclosing the origin, the product of the interval $[-\frac{1}{4}, \frac{1}{4}]$ for each variable x_i. For this example, no subdivision of boundary regions is needed. In our calculation, for the sake of simplicity, we always give the true interval values. An actual implementation of interval arithmetic may yield somewhat wider intervals, depending on the particular system used.

There are initially six elements defining sides of B on the list L, and only the element given below is moved to the list L_1 when it is tested:

$$(5) \qquad s_1 = [-\tfrac{1}{4}, \tfrac{1}{4}], \quad s_2 = [-\tfrac{1}{4}, \tfrac{1}{4}], \quad s_3 = -\tfrac{1}{4}; \qquad \sigma = -1.$$

For this domain, the interval for f_1 is $[\frac{1}{4}, \frac{3}{8}]$, to the right of 0, as required, and the intervals for f_2 and f_3 are $[-\frac{3}{16}, \frac{3}{8}]$, containing 0, as required. All of the other five L-elements are discarded, either because their f_2- or f_3-interval does not contain 0, or because their f_1-interval is to the left of 0.

The region $B^{(1)}$ is defined by the single element (5), and so the list L on the second iteration has four elements defining its sides. Of these only the one given below is transferred to L_1 during testing:

$$(6) \qquad s_1 = -\tfrac{1}{4}, \quad s_2 = [-\tfrac{1}{4}, \tfrac{1}{4}], \quad s_3 = -\tfrac{1}{4}; \qquad \sigma = 1.$$

For this element, the interval for f_2 is $[\frac{5}{16}, \frac{3}{8}]$, to the right of 0, as required, and the interval for f_3 is $[-\frac{1}{8}, \frac{3}{8}]$, containing the zero point, as required. The other three elements either have an f_3-interval that does not contain 0, or have an f_2-interval to the left of 0.

The region $B^{(2)}$ is defined by the single element (6), and so the list L on the last iteration has two elements defining its sides, which now are points. Only the element

$$s_1 = -\tfrac{1}{4}, \quad s_2 = -\tfrac{1}{4}, \quad s_3 = -\tfrac{1}{4}; \qquad \sigma = -1$$

yields a positive sign for f_3. Accordingly, the degree equals -1, the orientation value of this single element.

4. A bisection method

When n is one, the topological degree, when it is defined, can equal only $+1, -1$, or 0, while when n is greater than one, the topological degree can equal any integer. On the other hand, the topological degree mod 2, or the

crossing parity, is more consistent and has a value of zero or one for all n. A nonzero crossing parity for F over a box B also indicates a zero inside B. The crossing parity is considerably easier to calculate than the topological degree, since in mod 2 arithmetic $+1$ and -1 are the same, so the orientation field can be dropped from the representation of a list element. At the same time the useful property of the topological degree, additivity over disjoint regions, is preserved. We used the crossing parity for our generalization of the bisection method.

The computation description in §2 serves also for the crossing parity, except that all orientation computation is ignored. The crossing parity then is the number mod 2 of elements on the list L_1 after the nth iteration is complete.

The procedure described for topological degree or crossing parity never terminates if the boundary of B contains a zero, so some changes are needed to allow a general program to handle this case appropriately. Unending subdivision can only occur during the first iteration with the list L, when the boundary of B is examined. Our arrangement for the subdivision of the lead element of L is as follows: An x_i-interval is divided into two only if its width is over a preset minimum W. If it turns out that no x_i-interval is divided, this is the signal for halting the procedure, but first the precision of computation is checked. If it is not adequate, precision is increased and the computation is reentered with the lead element unchanged. If precision is adequate, the lead element is preserved as the "terminating element".

The general procedure of our bisection program is as follows: First each of the functions f_i is checked to some extent to expose difficulties in definition over the box B. Then an initial evaluation of crossing parity is made for B. If this fails, a point on the boundary of B, derived from the preserved terminating element, is displayed to indicate the source of difficulty. If the initial crossing parity is successful and indicates a zero inside B, the regular bisection routine can begin. The box B is divided into two subboxes by subdividing one x_i-interval, the choice made by rotation, except that no interval is subdivided further after its width becomes small enough to specify its variable to the desired number of decimal places. A successful crossing parity computation on either subbox results in the discard of one subbox, and the other one becomes the new box B. If the crossing parity can always be computed, eventually all dimensions of B become small enough to satisfy accuracy requirements, and the B-centerpoint can be displayed as the zero.

When the parity computation on a subbox fails, a likely reason is that by accident an isolated zero is on the common side of the two subboxes. To take advantage of this, we set the width W low enough so that the following can be done: A box B_0 contained in B is constructed to enclose the terminating boundary element, but of a size small enough to satisfy accuracy requirements. If the crossing parity computation on B_0 is successful and the parity is nonzero, then again the midpoint of this box can be displayed as the zero. When this is not the case, what procedure to use next is unclear. One can try various other strategems, like rearranging the subdivision of B into two subboxes, reducing the width W, etc. But it is not possible to program bisection to always succeed in locating a zero, because the functions may be zero or near zero over a wide region [1, pp. 45–47], an infrequent occurrence but a difficult one to treat successfully. At some point in the program there must be an "escape" exit where

no zero is claimed. Instead a point within the current box B can be indicated as a "near zero", perhaps along with its computed f_i-values.

5. A PROGRAM TO LOCATE ALL ZEROS IN B

If it is known that the functions f_i have all partial derivatives, and that their Jacobian is nonzero at each zero, then a structured search through B may be undertaken to locate the zeros, which cannot be infinite in number. The following detail is useful here. Suppose over the box B each partial derivative $\partial f_i/\partial x_j$ is bounded in an interval, and by further interval arithmetic it is determined that the functions' Jacobian over B is never zero, that is, its interval does not contain the zero point. In this case there is at most one zero in B. For if $P(\alpha_1, \ldots, \alpha_n)$ and $Q(\beta_1, \ldots, \beta_n)$ are two distinct zeros in B, then by the mean value theorem [5, pp. 268–269]

$$0 = f_i(P) - f_i(Q) = \sum_{j=1}^{n} \frac{\partial f_i(C_i)}{\partial x_j}(\alpha_j - \beta_j), \qquad i = 1, \ldots, n,$$

where C_i, $i = 1, \ldots, n$, are points along the line segment joining P and Q. Therefore,

$$\begin{bmatrix} \frac{\partial f_1(C_1)}{\partial x_1} & \frac{\partial f_1(C_1)}{\partial x_2} & \cdots & \frac{\partial f_1(C_1)}{\partial x_n} \\ \frac{\partial f_2(C_2)}{\partial x_1} & \frac{\partial f_2(C_2)}{\partial x_2} & \cdots & \frac{\partial f_2(C_2)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(C_n)}{\partial x_1} & \frac{\partial f_n(C_n)}{\partial x_2} & \cdots & \frac{\partial f_n(C_n)}{\partial x_n} \end{bmatrix} \begin{bmatrix} \alpha_1 - \beta_1 \\ \alpha_2 - \beta_2 \\ \vdots \\ \alpha_n - \beta_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The determinant of the matrix above must be zero, yet it may be viewed as a Jacobian with its elements chosen inside the various partial derivative intervals for B, and this contradicts the assumption that the Jacobian interval does not contain the zero point.

Our program begins by checking, with interval arithmetic, that the functions f_i and their partial derivatives are defined at all points in the initial box B, and that the crossing parity is defined on B. A list L_2 of boxes is constructed, initially containing a single list element defining the box B. Then the following iteration is performed until the list L_2 is empty. All functions f_i and their partial derivatives are bounded in intervals over the box defined by the first L_2-element. This list element is discarded if any function interval does not contain the zero point. If this is not the case, then an interval arithmetic check of the Jacobian is made. If it is not certain that the zero point is outside the determinant interval, the box is divided into two subboxes as described previously for bisection, and two list elements representing these subboxes replace the lead element of the list L_2. Otherwise, if it is certain that the Jacobian interval does not contain the zero point, then the crossing parity for this box is computed. If the parity is zero, it is safe to discard the list element, since as we have seen, there cannot be two or more zeros inside the element domain. If the parity is one, there is exactly one zero in the domain, and the element is removed and added to a second list L_3, initially empty.

It may happen that the crossing parity computation fails, and then, as with the bisection program, a terminating boundary element is obtained on which the functions f_i are zero or close to zero. In this case a list element is constructed

defining a small box B_0 which encloses the points of the boundary element, and lies entirely within two or more of the boxes on the list L_2. The B_0-element is placed at the head of the list L_2 after all L_2-list boxes intersecting B_0 are suitably dissected into a set of subboxes so that the B_0 region is not doubly represented.

After the list L_2 is exhausted, the zeros which have been individually box-enclosed by means of the list L_3 can be found by Newton's method, or a similar method.

Usually, for a box B it is not known in advance whether or not the Jacobian is nonzero at all zeros. Our program accepts any proposed box B, but terminates the search when the lead element of the list L_2 defines a box with dimensions smaller than a preset limit. A point within this box is displayed as a point where the Jacobian and the functions f_i are simultaneously "too close to zero".

A version of this program is contained in an electronically accessible file for range arithmetic [2].

BIBLIOGRAPHY

1. O. Aberth, *Precise numerical analysis*, Wm. C. Brown, Dubuque, Iowa, 1988.

2. O. Aberth and M. J. Schaefer, *Precise scientific computation with range arithmetic, via C++*, ACM Trans. Math. Software **18** (1992), 481–491.

3. G. Alefeld and J. Herzberger, *Introduction to interval computation*, Translated by Jon Rokne, Academic Press, New York, 1983.

4. P. Alexandroff and H. Hopf, *Topologie*, Chelsea, New York, 1935.

5. R. C. Buck, *Advanced calculus*, 2nd ed., McGraw-Hill, New York, 1965.

6. J. Cronin, *Fixed points and topolgical degree in nonlinear analysis*, Math Surveys, no. 11, Amer. Math. Soc., Providence, RI, 1964.

7. C. Harvey and F. Stenger, *A two-dimensional analogue to the method of bisections for solving nonlinear equations*, Quart. Appl. Math. **33** (1975), 351–368.

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

9. _____, *A summary of recent experiments to compute the topological degree*, Applied Nonlinear Analysis, Proceedings of an International Conference on Applied Nonlinear Analysis, University of Texas at Arlington, April 20–22, 1978 (V. Laskshmikantham, ed.), Academic Press, New York, 1979, pp. 627–633.

10. _____, *Abstract generalized bisection and a cost bound*, Math. Comp. **49** (1987), 187–202.

11. R. E. Moore, *Methods and applications of interval analysis*, SIAM Stud. Appl. Math., SIAM, Philadelphia, PA, 1979.

12. _____, *Interval analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

13. A. Neumaier, *Interval methods for systems of equations*, Cambridge Univ. Press, Cambridge, 1990.

14. T. O'Neal and J. Thomas, *The calculation of the topological degree by quadrature*, SIAM J. Numer. Anal. **12** (1975), 673–680.

15. K. Sikorski, *A three-dimensional analogue to the method of bisections for solving nonlinear equations*, Math. Comp. **33** (1979), 722–738.

16. F. Stenger, *Computing the topological degree of a mapping in R^n*, Numer. Math. **25** (1975), 23–38.

DEPARTMENT OF MATHEMATICS, TEXAS A&M UNIVERSITY, COLLEGE STATION, TEXAS 77843
E-mail address: aberth@math.tamu.edu