MATHEMATICS OF COMPUTATION
VOLUME 49, NUMBER 179
JULY 1987, PAGES 187–202

Abstract Generalized Bisection and a Cost Bound

By R. Baker Kearfott

Abstract. The purpose of this paper is to study desirable properties of binary search algorithms for isolating all solutions to nonlinear systems of equations $F(X) = 0$ within a given compact domain $\mathbf{D} \in \mathbf{R}^n$. We devise a general framework such that any algorithm fitting into the general framework will always isolate all solutions $Z \in \mathbf{D}$ such that $F(Z) = 0$; this framework contains a new idea for handling the occurrence of roots on boundaries. We then present and prove a bound on the total amount of computation which is valid for any algorithm in the class.

Finally, we define a specific prototypical algorithm valid for F satisfying certain natural smoothness properties; we show that it satisfies the hypotheses for the general framework. This algorithm is based on "bisection" of generalized rectangles, the Kantorovich theorem, and second-order Taylor type models for F. It is meant to provide further guidelines for the development of effective heuristics, etc., for actual implementations.

1. Motivation, Purpose, and Scope. Consider the problem:

(1.1) Find, *with certainty*, *approximations to all* solutions of $F(X) = 0$ within a bounded, closed set $\mathbf{D} \in \mathbf{R}^n$, where F: $\mathbf{D} \to \mathbf{R}^n$, $F = (f_1, f_2, \ldots, f_n)$.

This problem can sometimes be solved by one of the following:

(i) random search;

(ii) placing a grid $\{X_i\}_{i=1}^{k}$ on $\mathbf{D}$ and evaluating $F(X_i)$, $1 \leqslant i \leqslant k$, for a small enough mesh.

(iii) a homotopy continuation method (cf. e.g., [1], [5], [16], [20], etc.);

(iv) Newton's method with repeated random starts;

(v) more specialized methods (such as the method of bisection with sign changes, methods for roots of polynomials, etc.) which take advantage of specific properties (cf. [4], [16], etc.).

Methods (i) and (ii) involve large amounts of work, and are especially impractical for n large. Methods (i), (ii), and (iv) often are employed without assurance that their execution will find approximations to all roots. (This is the nature of random starts, and it may not be possible to determine how small a mesh is small enough to guarantee that all roots are captured.) Though the methods in class (iii) are relatively successful, problems can still arise both in the underlying theory and in practical computations on finite machines.

Received November 18, 1985; revised June 19, 1986 and October 6, 1986.

1980 *Mathematics Subject Classification*. Primary 65H10; Secondary 65K99, 68B99, 68C25, 68E99.

Key words and phrases. Nonlinear algebraic systems, global optimization, analysis of algorithms, generalized bisection, Kantorovich theorem.

0025-5718/87 $1.00 + $.25 per page

In short, there is ample room for further analysis and algorithmic development to solve the general problem (1.1).

It is not in general easy to determine a priori how many, if any, roots F has in $\mathbf{D}$ prior to computation, even if derivative information is available. To be able to computationally do so with less work than a grid search, then to determine a point for each root within the domain of attraction of Newton's method, is thus of value. This paper addresses the latter task.

Ideas here stem from a number of previously published methods, all termed "generalized bisection." In these methods, a region in $\mathbf{R}^n$ is subdivided into two or more subregions; one or more numerical-analytic techniques then determine whether the subregions contain roots. In [10], triangles in $\mathbf{R}^2$ are bisected by bisecting the longest edge, and in [3] simplices in $\mathbf{R}^n$ are similarly bisected; determining whether roots exist was done heuristically using a linearization. Sikorski presented a related approach for $\mathbf{R}^3$ in [18]. Stenger proposed the topological degree to determine the existence of roots, using the generalized triangle bisection process on simplices in $\mathbf{R}^n$ (cf. [19]). These ideas were further developed in [12], [13], etc., and are related to simplicial homotopy methods surveyed in [1], [5], [20] and elsewhere.

In parallel with the above, Moore and others defined generalized bisection algorithms for generalized rectangles by bisecting one of the axes (cf. [6], [7], [8], [15], etc.). There, determining whether a generalized rectangle contained a root relied heavily on methods of interval mathematics, such as interval inclusion of the zero vector and the Krawczyk method.

The topological degree and generalized bisection of simplices have been touted as methods appropriate when F is not smooth or when F cannot be evaluated accurately. In contrast, the interval mathematics approach was claimed to be very robust, and appropriate for finding starting points for Newton's method when more than one root exists. In fact, depending on the particular problem, such methods can be both more efficient and more reliable than others, which, in theory, will also solve it.

For example, consider the problem of finding all *real* roots to the following polynomial system of equations:

$$f_1(x_1, x_2, x_2) = 5x_1^9 - 6x_1^5x_2^2 + x_1x_2^4 + 2x_1x_3,$$
$$f_2(x_1, x_2, x_3) = -2x_1^6x_2 + 2x_1^2x_2^3 + 2x_2x_3,$$
$$f_3(x_1, x_2, x_3) = x_1^2 + x_2^2 - .265625.$$

Theory dictates that this system can be expected to have $9 \cdot 7 \cdot 2 = 126$ roots in the complex domain (cf. [4]), and that these roots may be obtained with a homotopy method. In [14], we used a standard homotopy method and the recognized continuation method software from [17] with mixed success. We encountered problems because: (i) we needed to trace all 126 complex solution paths, even though there are just 6 real roots of interest; and (ii) we encountered ill-conditioning and path-jumping due to near bifurcation. Thus (in the absence of special analysis, preprocessing, and scaling) we required hundreds of thousands of evaluations of F to solve the problem, the algorithm required significant tuning, and human interpretation of the results was necessary. In contrast, a method employing generalized bisection of

rectangles and an interval-arithmetic test for roots was able to automatically compute the 6 desired solutions to machine accuracy in approximately 1300 function evaluations ([9] and unpublished experiments).

This paper contributes the following: First, an abstract framework with points in common to actual successful generalized bisection algorithms is presented; desirable properties of such methods are thus highlighted. Within this scheme, a new mechanism is introduced to handle problems which occur when roots happen to be near the boundary of one of the subregions. (These problems lead to excessive computation or redundant listing of roots; their resolution is particularly important in order to generalize the techniques to larger, sparse systems of equations.) Also, a new algorithm for $\mathbf{R}^n$ is presented and is shown to satisfy the hypotheses of the abstract framework.

In Section 2 the abstract framework is defined. We show there that any algorithm satisfying the hypotheses in the abstract framework will produce subregions of an initial region $\mathbf{D}$ such that: (i) the subregions are disjoint and each subregion contains a unique root of F in $\mathbf{D}$; (ii) each root of F in $\mathbf{D}$ is in one of the subregions; and (iii) no root is listed in more than one subregion. A bound on the total amount of work for this computation is presented. In Section 3 the specific generalized bisection method is introduced and is shown to satisfy all of the hypotheses of the abstract framework.

We hope this presentation will facilitate design of practical algorithms and codes.

2. Abstract Generalized Bisection. Suppose we are given a closed region $\mathbf{D} \in \mathbf{R}^n$ and a function F: $\mathbf{D} \to \mathbf{R}^n$. (The region $\mathbf{D}$ may be a simplex, a generalized rectangle or other polygonal figure, etc., with an appropriate structure.) Then

Definition 2.1. *Abstract generalized bisection* consists of

(i) a *geometrical bisection process* $\mathscr{B}$;

(ii) a *root inclusion test* $\mathscr{T}$; and

(iii) a *binary search algorithm* based on $\mathscr{B}$ and $\mathscr{T}$,

where these three terms are defined below.

In practice, the geometrical bisection process and the root inclusion test depend adaptively on the function F; this abstract study should reflect this fact. Thus, in order to define $\mathscr{B}$ and $\mathscr{T}$ (and also in order to obtain an operations count), we must first delineate an appropriate class of F.

Assumption 2.2. The roots of F are isolated. Thus, for any compact domain $\mathbf{D}$, and any norm $\|\cdot\|_s$, there exists an $E > 0$ such that the distance in $\|\cdot\|_s$ between roots of F is at least E. (Here, the subscript "s" on the norm is used to indicate that it can be an adaptively scaled norm, as in Definition 3.3 below.)

Definition 2.3. The *geometrical bisection process* $\mathscr{B}$ is a function from the set of domains $\mathscr{D}$ to the set of pairs of such domains such that, for $\mathbf{D} \in \mathscr{D}$:

$$\mathscr{B}(\mathbf{D}) = \{\mathbf{D}_1, \mathbf{D}_2\},$$

where

(i) $\mathscr{B}$ is completely determined by the set of domains $\mathscr{D}$ and F.

(ii) $\mathbf{D}_1 \cup \mathbf{D}_2 = \mathbf{D}$.

(iii) $\mathbf{D}_1 \cap \mathbf{D}_2 \subset \partial\mathbf{D}_1 \cup \partial\mathbf{D}_2$, where $\partial\mathbf{D}$ is the boundary of a region $\mathbf{D}$.

(iv) $\mathbf{D}_1$ and $\mathbf{D}_2$ have the same structure as $\mathbf{D}$ in the sense that $\mathscr{B}$ can be applied to both $\mathbf{D}_1$ and $\mathbf{D}_2$.

(v) If $\mathscr{B}$ is applied repeatedly to its results $\mathbf{D}_1$ and $\mathbf{D}_2$, then the diameters of the resulting regions tend to zero. Furthermore, the rate is independent of $\mathbf{D}$ in the sense that there is a number K such that if $\tilde{\mathbf{D}}$ is the final result of applying $\mathscr{B}$ K times, starting with $\mathbf{D}$, then $d(\tilde{\mathbf{D}}) \leqslant d(\mathbf{D})/2$ (i.e., if the total number of applications of $\mathscr{B}$ to get $\overline{\mathbf{D}}$ is M, then, asymptotically, $d(\overline{\mathbf{D}}) \leqslant d(\mathbf{D})/2^{(M/K)}$). Here, $d(\mathbf{D})$ is the diameter of $\mathbf{D}$ with respect to a selected norm $\|\cdot\|_s$; this norm is dependent in general on $\mathbf{D}$, but well-behaved in the sense that there is a c independent of $\mathbf{D}$ with $\|X\|_s \geqslant c\|X\|_\infty$ for every $X \in \mathbf{R}^n$.

(vi) There is a number $0 < R \leqslant 1/2$ such that if $\overline{\mathbf{D}}$ is any region produced by applying $\mathscr{B}$ any number of times to $\mathbf{D}$, then $\overline{\mathbf{D}}$ contains an open ball $\mathbf{B}$ with center C and radius $Rd(\overline{\mathbf{D}})$, where distances are measured with respect to $\|\cdot\|_s$.

Condition (v) of Definition 2.3 bounds above the rate at which the diameters of the regions to which $\mathscr{B}$ is applied go to zero. This allows us to quantify how many bisections are required for certain local models to become valid within a region. (For example, an interval-arithmetic model for the range of F as X ranges over $\mathbf{D}$ may consist of linear terms added to intervals representing ranges of the coordinates of X in $\mathbf{D}$ times bounds on the second partial derivatives of F. The diameter of the range set so computed would then tend to zero as $d(\mathbf{D})$ tends to zero.) It may be possible to choose $\|\cdot\|_s$ so that the corresponding $\mathbf{D}$ is a ball with respect to $\|\cdot\|_s$; this simplifies analysis related to the distance of roots of F from $\partial\mathbf{D}$ in specific cases.

Condition (vi) insures that the regions $\overline{\mathbf{D}}$ do not become excessively thin when repeated bisection is applied. This allows us to replace $\overline{\mathbf{D}}$ by a region $\overline{\mathbf{D}}_b$ when there is a root near $\partial\overline{\mathbf{D}}$, such that $\overline{\mathbf{D}} \subset \overline{\mathbf{D}}_b$, the root is relatively far from $\partial\overline{\mathbf{D}}_b$, but $d(\overline{\mathbf{D}}_b)$ is not excessively larger than $d(\overline{\mathbf{D}})$.

Definition 2.4. A *root-inclusion test* $\mathscr{T}_F$ is any mapping from the set of domains $\mathscr{D}$ to {'true', 'false', 'unknown'} such that

(i) $\mathscr{T}_F(\mathbf{D}) =$ 'true' implies there is a unique $X \in \text{int}(\mathbf{D})$ with $d(X, \partial\mathbf{D}) \geqslant (R^2/4)d(\mathbf{D})$ such that $F(X) = 0$. Here, $\text{int}(\mathbf{D}) = \mathbf{D} \setminus \partial\mathbf{D}$ is the interior of $\mathbf{D}$, $d(X, \mathbf{S})$ is the distance of the point X from the set $\mathbf{S}$ in $\|\cdot\|_s$, R is as in Definition 2.3, and 0 denotes the zero vector.

(ii) $\mathscr{T}_F(\mathbf{D}) =$ 'false' implies there are no $X \in \mathbf{D}$ with $F(X) = 0$.

(iii) There is a number $\varepsilon' > 0$ such that, if $\mathbf{D}$ is any region with

(a) $d(\mathbf{D}) < \varepsilon'$; and

(b) If X is such that $F(X) = 0$, then $d(X, \partial\mathbf{D}) \geqslant (R/3)d(\mathbf{D})$, where R is as in Definition 2.3(vi),

then $\mathscr{T}_F(\mathbf{D}) =$ 'true' or $\mathscr{T}_F(\mathbf{D}) =$ 'false'.

Definition 2.4(iii) is so stated because zeros of local models of F will be used to approximate zeros of F itself. In such models, the approximation error will be bounded; the model will then be able to distinguish roots inside and outside of a region $\mathbf{D}$ when roots of F lie relatively far from $\partial\mathbf{D}$.

The following definition delineates the interplay between the class of functions defined in Assumption 2.2 and the conditions which define $\mathscr{B}$ and $\mathscr{T}_F$.

Definition 2.5. Choose $\varepsilon = \min\{\varepsilon', E/R\}$, where ε' is as in Definition 2.4(iii), E is as in Assumption 2.2, and R is as in Definition 2.3(vi). Thus, distinct roots of F are

of distance at least $R\varepsilon$ in $\|\cdot\|_s$. Furthermore, we may replace ε' by ε in Definition 2.4(iii) for simplicity of notation.

We now study an abstract version of the binary search algorithm. Suppose $\mathscr{B}$ is as in Definition 2.3 and $\mathscr{T}_F$ is as in Definition 2.4. Also suppose R and ε are as in Definition 2.5 and are known a priori, and that C can be computed for each $\mathbf{D}$, where C is the center as in Definition 2.3(vi). Then we have

ALGORITHM 2.6. (Abstract generalized bisection)

1. (Initialization phase)
   (a) Set: $k \leftarrow 1$.
   (b) Set: $\mathbf{D}^1 \leftarrow \mathbf{D}$.
2. (Subdivision phase)
   (a) Form $\{\mathbf{D}_1^k, \mathbf{D}_2^k\} = \mathscr{B}(\mathbf{D}^k)$.
   (b) $\mathbf{D}^{k+1} \leftarrow \mathbf{D}_1^k$.
   (c) $k \leftarrow k + 1$.
3. (Test phase and storage of roots)
   (a) Compute $d(\mathbf{D}^k)$.
   (b) If $d(\mathbf{D}^k) < (R/2)\varepsilon$ and $\mathbf{D}^k$ has nonnull intersection with a region which has been expanded in Step 4, then go to Step 5. Otherwise, continue to Step 3(c).
   (c) Compute $\mathscr{T}_F(\mathbf{D}^k)$.
   (d) If $\mathscr{T}_F(\mathbf{D}^k) =$ 'unknown' and $d(\mathbf{D}^k) \geqslant (R^2/4)\varepsilon$, then return to Step 2.
   (e) If $\mathscr{T}_F(\mathbf{D}^k) =$ 'false', then go to Step 5.
   (f) If $\mathscr{T}_F(\mathbf{D}^k) =$ 'true' then:
      (i) If $\mathbf{D}^k$ has null intersection with every region which has been expanded in Step 4, then store $\mathbf{D}^k$ in list $\mathscr{L}$.
      (ii) Go to Step 5.
4. (Adjustment step for roots on a boundary: in this case $\mathscr{T}_F(\mathbf{D}^k) =$ 'unknown' and $d(\mathbf{D}^k) < (R^2/4)\varepsilon$.)
   (a) Replace $\mathbf{D}^k$ by the geometrically similar region $\mathbf{D}_b^k$ obtained by replacing every $X \in \mathbf{D}^k$ by $\overline{X}$, where

$$\overline{X} = C + (2/R)(X - C).$$

   (Here, C is the center as in Definition 2.3(vi).)
   (b) Delete from $\mathscr{L}$ all $\mathbf{D}' \in \mathscr{L}$ for which $\mathbf{D}' \cap \mathbf{D}_b^k$ is nonempty and $d(\mathbf{D}') < (R/2)\varepsilon$.
   (c) Store D_b^k in list $\mathscr{L}$.
5. (Backtrack to less subdivided regions)
   (a) If $k = 1$, then exit with the list $\mathscr{L}$.
   (b) If $\mathbf{D}^k$ was $\mathbf{D}_1^{k-1}$, then:
      (i) Set: $\mathbf{D}^k \leftarrow \mathbf{D}_2^k$.
      (ii) Go to Step 3.
   (c) $k \leftarrow k - 1$.
   (d) Return to Step 5(a). □

In the remainder of this section, we will prove desired properties of Algorithm 2.6.

LEMMA 2.7. *Suppose* $\mathbf{D}^k$ *and* $\mathbf{D}_b^k$ *are as in Step* 4 *of Algorithm* 2.6. *Then*
(i) $d(\mathbf{D}_b^k) = (2/R)d(\mathbf{D}^k) < (R/2)\varepsilon$;
(ii) $\mathbf{D}_b^k$ *contains a ball* $\mathbf{B}_b$ *of radius* $2d(\mathbf{D}^k)$ *in its interior*;
(iii) $\mathbf{D}^k \subset \mathbf{B}_b$. *Furthermore*,

$$d(\partial\mathbf{B}_b, \partial\mathbf{D}^k) \geqslant d(\mathbf{D}^k).$$

Here, the norm used to define $d(\mathbf{D}_b^k)$ *is the same as the norm used to define* $d(\mathbf{D}^k)$.

Proof of Lemma 2.7. For (i), let X and Y be extreme points of $\mathbf{D}^k$; i.e., $d(X, Y) = d(\mathbf{D}^k)$; let $\overline{X}$ and $\overline{Y}$ be the points in $\mathbf{D}_b^k$ obtained from X and Y by the construction of $\mathbf{D}_b^k$. Then, by constructing similar triangles with vertex sets $\{X, Y, C\}$ and $\{\overline{X}, \overline{Y}, C\}$, it is clear that $d(\overline{X}, \overline{Y}) = d(\mathbf{D}_b^k) \geqslant (2/R)d(\mathbf{D}^k)$. (Note that, no matter what norm is used, similar triangles have side lengths which are proportional, provided corresponding sides of the triangles make the same angles with the coordinate axes.) However, if $\overline{Z}$ and $\overline{W}$ are points in $\mathbf{D}_b^k$ with $d(\overline{Z}, \overline{W}) > (2/R)d(\mathbf{D}^k)$, then an analogous construction of similar triangles leads to the contradiction $d(\mathbf{D}^k) > d(\mathbf{D}^k)$. Therefore, (i) is true.

For (ii), simply take $\mathbf{B}^b$ to be the images of the points in the ball $\mathbf{B}$ in Definition 2.3(vi).

We now prove (iii). First observe that $\mathbf{D}^k \subset \mathbf{B}_b$ since the radius of $\mathbf{B}_b$ is $2d(\mathbf{D}^k)$, because of the construction in Step 4 of Algorithm 2.6, and because $C \in \mathbf{D}^k$. Now suppose $A \in \partial\mathbf{D}^k$ and $B \in \partial\mathbf{B}_b$ are such that $d(A, B) = d(\partial\mathbf{D}^k, \partial\mathbf{B}_b)$, and let C be as above. Then $d(B, C) = 2d(\mathbf{D}^k)$, but $d(A, C) \leqslant d(\mathbf{D}^k)$. Thus, the triangle inequality gives $d(B, A) > 2d(\mathbf{D}^k) - d(\mathbf{D}^k) = d(\mathbf{D}^k)$; combining this with (i) gives (iii). □

LEMMA 2.8. *Suppose* $\mathbf{D}^k$ *is a region for which Algorithm* 2.6 *has entered Step* 4. *Then*:
(i) *there is a root* Z *of* F *such that* $d(Z, \partial\mathbf{D}^k) < (R/3)d(\mathbf{D}^k)$;
(ii) $Z \in \mathbf{D}_b^k$;
(iii) $d(Z, \partial\mathbf{D}_b^k) \geqslant (R/3)d(\mathbf{D}_b^k)$.

Proof of Lemma 2.8. Since $d(\mathbf{D}^k) < \varepsilon$ and $\mathcal{T}_F(\mathbf{D}^k) =$ 'unknown', Definition 2.4(iii) implies (i). The assertion (ii) follows from (i), Lemma 2.7(ii) and (iii), and the triangle inequality. In particular, let C be as above, let $A \in \partial\mathbf{D}^k$ be such that $d(A, Z) = d(\partial\mathbf{D}^k, Z)$, and let $B \in \partial\mathbf{B}_b$ be such that $d(Z, B) = d(Z, \partial\mathbf{B}_b)$. Then (i) implies

$$d(A, Z) < (R/3)d(\mathbf{D}^k), \tag{2.1}$$

while

$$d(C, A) \leqslant d(\mathbf{D}^k). \tag{2.2}$$

The triangle inequality then implies

$$\begin{aligned} d(C, Z) &\leqslant d(C, A) + d(A, Z) \\ &< [(1 + (R/3))]\, d(\mathbf{D}^k) < 2d(\mathbf{D}^k). \end{aligned} \tag{2.3}$$

Formula (2.3) combined with Lemma 2.7(ii) gives (ii).

For (iii), first observe that Lemma 2.7(ii) implies

$$d(C, B) = 2d(\mathbf{D}^k). \tag{2.4}$$

The triangle inequality then implies

$$\begin{aligned} d\big(Z, \partial \mathbf{D}_b^k\big) &\geqslant d(Z, B) \geqslant d(C, B) - d(C, Z) \\ &> 2d(\mathbf{D}^k) - \big[1 + (R/3)\big] d(\mathbf{D}^k) \\ &\geqslant (5/6) d(\mathbf{D}^k) > (R/3) d\big(\mathbf{D}_b^k\big). \end{aligned} \tag{2.5}$$

(The first inequality follows from the fact that any line segment drawn from Z to $\partial \mathbf{D}_b^k$ must intersect $\partial \mathbf{B}_b$.) Thus, Lemma 2.8 is proved. □

We now state and prove our two main results.

THEOREM 2.9. *Assume Algorithm* 2.6 *terminates. Then*:

(i) *If* $\mathbf{D}' \in \mathscr{L}$, *then* $\mathbf{D}'$ *contains a unique root* Z *of* F.

(ii) *If* $Z \in \mathbf{D}$ *has* $F(Z) = 0$, *then* $Z \in \mathbf{D}'$ *for some* $\mathbf{D}' \in \mathscr{L}$.

(iii) *If* $\mathbf{D}' \in \mathscr{L}$, *then* $\mathscr{T}_F(\mathbf{D}') =$ '*true*'.

(iv) *If* $Z \in \mathbf{D}$ *is a root of* F, *then* Z *is in only one* $\mathbf{D}' \in \mathscr{L}$.

Proof of Theorem 2.9. We first show (i). To this end, assume $\mathbf{D}' \in \mathscr{L}$. Then $\mathbf{D}'$ was placed in $\mathscr{L}$ in either Step 3(f) or Step (4). If $\mathbf{D}'$ came from Step 3(f), then Definition 2.4(i) implies that $\mathbf{D}'$ contains a unique root, so assume $\mathbf{D}'$ came from Step 4; let $\overline{\mathbf{D}}$ be the region which gave rise to $\mathbf{D}'$ by expansion. Then Lemma 2.8(ii) implies there is a root $Z \in \mathbf{D}'$. However, Lemma 2.7(i) implies that if $Z' \in \mathbf{D}'$ is also a root, then $d(Z, Z') \leqslant (R/2)\varepsilon$; Definition 2.5 then implies $Z = Z'$.

We now verify (ii). Assume that Z is a root which is not in any $\mathbf{D}' \in \mathscr{L}$. Then Z must be in some region $\mathbf{D}^k$ which was rejected in Step 3(b) or Step 4(b). (This is because all other $\mathbf{D}^k$ entering Step 3 will be bisected further, do not contain roots, have indeed been placed in $\mathscr{L}$, or have been placed in $\mathscr{L}$ after expansion.) Thus

$$d(\mathbf{D}^k) < (R/2)\varepsilon. \tag{2.6}$$

Furthermore, if $\mathbf{D}' \in \mathscr{L}$ is the region from Step 4 such that $\mathbf{D}^k \cap \mathbf{D}'$ is nonempty then Lemma 2.7(i) implies

$$d(\mathbf{D}') < (R/2)\varepsilon. \tag{2.7}$$

However, there is a $Z' \in \mathbf{D}'$ for which $F(Z') = 0$, by Lemma 2.8(i). Formulas (2.6) and (2.7) now imply

$$d(Z, Z') < (R/2)\varepsilon + (R/2)\varepsilon = R\varepsilon, \tag{2.8}$$

which contradicts Definition 2.5. Thus, (ii) is true.

To prove (iii), we must verify the hypotheses of Definition 2.4(iii) for an arbitrary $\mathbf{D}' \in \mathscr{L}$. If $\mathbf{D}'$ was placed in $\mathscr{L}$ in Step 3(f), then the conclusion is trivially true, so assume $\mathbf{D}'$ was placed in $\mathscr{L}$ in Step 4. Then Lemma 2.7(i) assures us that $d(\mathbf{D}') < \varepsilon$. Furthermore, if Z is the root of F in $\mathbf{D}'$, then Lemma 2.8(iii) assures us that $d(Z, \partial \mathbf{D}') > (R/3) d(\mathbf{D}')$; assume Z' is any other root, and let $A \in \partial \mathbf{D}'$ be such that

$$d(Z', A) = d(Z', \partial \mathbf{D}'). \tag{2.9}$$

Then

$$\begin{aligned} d(Z', \partial \mathbf{D}') &\geqslant d(Z', Z) - d(Z, A) \geqslant R\varepsilon - d(\mathbf{D}') \\ &\geqslant R\varepsilon - R(\varepsilon/2) = (R/2)\varepsilon > (R/3)\varepsilon. \end{aligned} \tag{2.10}$$

Thus, (iii) is proved.

We will prove (iv) by contradiction: Suppose $Z \in \mathbf{D} \in \mathscr{L}$ and $Z \in \mathbf{D}' \in \mathscr{L}$, so that $\mathbf{D} \cap \mathbf{D}'$ is nonnull. But both $\mathbf{D}$ and $\mathbf{D}'$ cannot be from Step 4(b), since an ancestor $\mathbf{D}_a$ of $\mathbf{D}$ or $\mathbf{D}'$ would then have had $d(\mathbf{D}_a) < (R/2)\varepsilon$ and would thus have been eliminated from $\mathscr{L}$ in Step 3(b). Thus, without loss of generality, assume $\mathbf{D}$ came from Step 3(f). Then $d(\mathbf{D}) \geqslant (R/2)\varepsilon$, since otherwise $\mathbf{D}$ would have been rejected in either Step 3(b) or Step 4(b). However, since the definition of $\mathscr{T}_F$ implies Z is not in $\partial\mathbf{D}$, the nature of $\mathscr{B}$ implies that $\mathbf{D}'$ is some $\mathbf{D}_b^k$ from Step 4. But then Lemma 2.8 and (iii) imply $d(Z, \partial\mathbf{D}) < (R/3)[(R^2/4)\varepsilon] = (R^3/12)\varepsilon$. Thus, $d(Z, \partial\mathbf{D}) < (R^2/4)d(\mathbf{D})$, which contradicts $\mathscr{T}_F(\mathbf{D}) =$ 'true'. Thus, (iv) is proven. □

Theorem 2.9(iii) is useful when $\mathscr{T}_F$ is defined so that $\mathscr{T}_F(\mathbf{D}') =$ 'true' implies that an associated iterative method will converge to the root $Z \in \mathbf{D}'$ for any initial guess $X \in \mathbf{D}'$.

THEOREM 2.10. *Algorithm* 2.6 *terminates after at most* $2(4M^K - 1)$ *applications of* $\mathscr{B}$, *where*

$$M = \left\lfloor d(\mathbf{D}) / \left[(R^2/4)\varepsilon \right] \right\rfloor,$$

and where K is as in Definition 2.3(v).

Proof of Theorem 2.10. We first note that bisection stops when the diameter of the region $\overline{\mathbf{D}}$ produced in Algorithm 2.6 is less than $(R^2/4)\varepsilon$, so that the reduction in diameter necessary to obtain such $\overline{\mathbf{D}}$ from $\mathbf{D}$ is by approximately a factor of M. However, by Definition 2.3(v), this reduction of diameter would require at most

$$L = \left\lfloor K(\log_2(\tilde{M}) + 1) \right\rfloor$$

successive subdivisions, where $\tilde{M} = d(\mathbf{D})/[(R^2/4)\varepsilon]$.

Algorithm 2.6 can be thought of as producing a binary tree whose nodes are the regions obtained through successive application of $\mathscr{B}$: The root of this tree is the node corresponding to the original region $\mathbf{D}$, and the number of edges (equal to the number of nodes less 1) is equal to the total number of applications of $\mathscr{B}$. The depth of the tree is equal to the maximum number of successive subdivisions, which is at most L. The maximum total number of subdivisions would occur if the tree were balanced, and would be equal to

$$\begin{aligned} 2^{L+1} - 2 = 2(2^L - 1) &\leqslant 2[2^{K \log_2(\tilde{M})+2} - 1] \\ &\leqslant 2(4\tilde{M}^K - 1) \leqslant 2(4M^K - 1) \end{aligned}$$

(cf., e.g., [21, p. 277]). This completes the proof. □

We note that the bound in Theorem 2.10 is a worst-case estimate for the class of functions defined by Assumption 2.2. Since this bound assumes a balanced tree, and since a good algorithm will produce a tree which is far from balanced, this bound is unrealistically large in many cases.

3. A Specific Generalized Bisection Algorithm. We now present a prototypical example of a generalized bisection algorithm.

Let F: $\mathbf{D} \to \mathbf{R}^n$ ($\mathbf{D}$ a closed domain in $\mathbf{R}^n$ as above) satisfy the following

Assumption 3.1. (i) F is twice differentiable, and its Jacobian matrix $J(X)$ and second derivative operator $H(X)$ are continuous at every X in $\mathbf{D}$.

(ii) There is an M such that $\|J(X)\| \leqslant M$ for every X in $\mathbf{D}$. (Here, the operator norm $\|\cdot\|$ depends on the scaled norms $\|\cdot\|_s$ of Definition 3.3 below, which in turn depend on the specific region $\overline{\mathbf{D}}$ produced by the algorithm; one M should hold for all such norms encountered during the course of the algorithm.)

(iii) There are quantities M_1 and an ε_0 such that $\|J^{-1}(X)\| \leqslant M_1$ for every X such that $\|Z - X\| \leqslant \varepsilon_0$ for some Z with $F(Z) = 0$. (As in (ii), the operator norm here will depend upon $\|\cdot\|_s$ defined below.)

(iv) There is a γ_0 such that $\|H_i(X)\| \leqslant \gamma_0$ for $1 \leqslant i \leqslant n$ and every $X \in \mathbf{D}$, where H_i is the Hessian matrix of f_i. Let

$$\gamma = \max_{|v_i| \leqslant \gamma_0} \|(v_1, v_2, \ldots, v_n)\|,$$

where the above norm is the norm on the range of F.

Modifications of Assumptions 3.1(ii), (iii), and (iv) can be verified with interval arithmetic as computation proceeds, provided interval arithmetic is available and F has an interval arithmetic extension: In such cases, sets containing $J(\mathbf{D})$, etc., are explicitly computed. Then, for example, if $\mathscr{J}$ is such a set corresponding to $J(\mathbf{D})$, then $\|\mathscr{J}\|$ can be explicitly computed and used as the bound M.

Definition 3.2. Let

$$\mathbf{D} = \prod_{i=1}^{n} [a_i, b_i]$$

be a generalized rectangle in $\mathbf{R}^n$. Define $\mathscr{B}_0(\mathbf{D})$ by $\mathscr{B}_0(\mathbf{D}) = \{\mathbf{D}_1, \mathbf{D}_2\}$, where

$$\mathbf{D}_1 = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [m, b_i] \times \cdots \times [a_n, b_n],$$

$$\mathbf{D}_2 = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_i, m] \times \cdots \times [a_n, b_n],$$

$$m = (a_i + b_i)/2, \quad \text{and} \quad |b_i - a_i| = \max_{1 \leqslant j \leqslant n} |b_j - a_j|.$$

The process $\mathscr{B}_0$ is used in [6], [7], [8], [15] and elsewhere. Its advantages include predictability and a separation of the coordinate directions. Modifications (choosing i other than as above) are possible to effect automatic scaling or for handling sparse systems.

Suppose $\overline{\mathbf{D}}$ is produced from $\mathbf{D}$ via repeated application of $\mathscr{B}_0$ as in Definition 3.2. Then we will use a local norm specific to $\overline{\mathbf{D}}$ as follows:

Definition 3.3. Let $\overline{\mathbf{D}} = \prod_{i=1}^{n} [a_i, b_i]$, let $\delta_i = |b_i - a_i|$, assume $\delta_i > 0$ for $1 \leqslant i \leqslant n$, set $\delta_j = \max_{1 \leqslant i \leqslant n} \delta_i$, and let $V = (v_1, v_2, \ldots, v_n)^T \in \mathbf{R}^n$. Then define the scaled norm

$$\|V\|_s = \max_{1 \leqslant i \leqslant n} \left\{ (\delta_j/\delta_i) |v_i| \right\}.$$

Define $d_s(\overline{\mathbf{D}})$, $d_s(X, Y)$, etc., with respect to $\|\cdot\|_s$ as in Section 2. Note that, in $\|\cdot\|_s$, $\overline{\mathbf{D}}$ is a ball of radius $\delta_j/2$, and $d_s(\overline{\mathbf{D}}) = d_\infty(\overline{\mathbf{D}}) = \delta_j$, where $d_\infty(\overline{\mathbf{D}})$ is the diameter of $\overline{\mathbf{D}}$ using the unscaled maximum norm.

From Definition 3.2 and Definition 3.3, we have

LEMMA 3.4. (i) *$\mathscr{B}_0$ is a geometrical bisection process as in Definition* 2.3.

(ii) *The number K in Definition* 2.3(v) *can be taken to be equal to n.*

(iii) *The number R in Definition* 2.3(vi) *can be taken to equal* $1/2$, *and*

$$C = ((a_1 + b_1)/2, (a_2 + b_2)/2, \ldots, (a_n + b_n)/2).$$

Proof of Lemma 3.4. Items (i) through (iv) of Definition 2.3 are immediately evident. For item (v), note that a single application of $\mathscr{B}_0$ halves the width of the generalized rectangle in one of the coordinate directions. If the original $\mathbf{D}$ is a hypercube, then the thickness in each coordinate direction is halved after n applications, and the resulting region $\overline{\mathbf{D}}$ is a hypercube with diameter equal to $d_s(\overline{\mathbf{D}}) = d_s(\mathbf{D})/2$. In any case, the maximum width δ_j of $\mathbf{D}$ is halved on the first application of $\mathscr{B}_0$; when $\mathscr{B}_0$ is then recursively applied, the jth width of the resulting generalized rectangle must then be halved at least once every n applications. (This can be proven by induction on n.) The fact that c in Definition 2.3(v) exists is clear from the fact that the ratio $\max(\delta_i)/\min(\delta_i)$ remains bounded as $\mathscr{B}_0$ is applied recursively. (This can also be proven by induction.) For Definition 2.3(vi) and assertion (iii) of the lemma, simply note that each $\overline{\mathbf{D}}$ is a ball in the norm presently being used. □

Arguments similar to those in the preceding proof can be used to show that the bound M in Assumption 3.1(ii) exists for all smooth functions on compact domains $\mathbf{D}$.

The root inclusion test in this example will use the Kantorovich theorem (cf., e.g., [2, p. 92]), which we state as

THEOREM 3.5 (KANTOROVICH). *Let $C \in \mathbf{D}$ be arbitrary and let $\|\cdot\|$ be any norm. Assume $\|J^{-1}(C)\| \leqslant \beta$, let γ be such that*

$$\|J(Y) - J(X)\| \leqslant \gamma \|Y - X\| \quad \text{for } X \in \mathbf{D} \text{ and } Y \in \mathbf{D},$$

assume $\|J^{-1}(C)F(C)\| \leqslant \eta$, and define $\alpha = \beta\gamma\eta$. Furthermore, assume F is continuously differentiable in the ball

$$\mathbf{B}_0 = \left\{ X \in \mathbf{D}, \|X - C\| < \frac{1 + \sqrt{1 - 2\alpha}}{\beta\gamma} \right\} \subset \mathbf{D}.$$

If $\alpha < 1/2$, then Newton's method, using C as starting point, converges to a root Z such that

$$\|Z - C\| \leqslant \frac{1 - \sqrt{1 - 2\alpha}}{\beta\gamma}.$$

Furthermore, Z is unique within $\mathbf{B}_0$.

We now present the inclusion test for this example.

Definition 3.6. Let F, M, ε_0, M_1, and γ be as in Assumption 3.1, let $\mathbf{D} = \overline{\mathbf{D}}$ and $\|\cdot\|_s$ be as in Definition 3.3, and let C be as in Lemma 3.4(iii). Define $\beta = \|J^{-1}(C)\|_s$ and $\eta = \|J^{-1}(C)F(C)\|_s$, where $\|J^{-1}(C)\|_s$ is the operator norm derived from $\|\cdot\|_s$ and from the norm on the range of F. Define

$$U(\mathbf{D}) = \left\{ L(F)(X) + \frac{\gamma_0 d_s^2(\mathbf{D})}{8} T,\ T \in [-1,1]^n,\ X \in \mathbf{D} \right\},$$

where $L(F)(X) = (L_1(F)(X), L_2(F)(X), \ldots, L_n(F)(X))$, and

$$L_i(F)(X) = f_i(X) + J_i(C)(X - C), \qquad J_i(C) = \text{row}_i[J(C)].$$

Then set $\mathscr{T}_{0,F}$ according to:

(i) $\mathscr{T}_{0,F}(\mathbf{D}) =$ "true" provided $\alpha = \beta\gamma\eta < 1/2$,

$$\frac{1 - \sqrt{1 - 2\alpha}}{\beta\gamma} < \frac{d_s(\mathbf{D})}{2},$$

and provided that $d(Z, \partial\mathbf{D}) \geqslant (R^2/4)d(\mathbf{D})$ where Z is the root to which Newton's method, starting with C, will converge (from Theorem 3.5) when the first two provisions are satisfied.

(ii) $\mathscr{T}_{0,F}(\mathbf{D}) =$ "false" provided $U(\mathbf{D})$ does not contain 0.

(iii) $\mathscr{T}_{0,F}(\mathbf{D}) =$ "unknown" provided the provisions in neither (i) nor (ii) hold.

Note that checking provision (ii) of Definition 3.6 is equivalent to checking that there is no solution $(X, T) \in \mathbf{D} \times [-1, 1]^n$ of

$$0 = L(F)(X) + \frac{\gamma_0 d_s^2(\mathbf{D})}{8} T. \tag{3.1}$$

If $\mathbf{D}$ is polygonal, then this is equivalent to checking for feasible solutions to a linear programming problem in $\mathbf{R}^{2n}$ with $5n$ inequality constraints: $\mathbf{D}$ is defined by $2n$ constraints, the range of the vector T is defined by $2n$ constraints, and (3.1) defines n constraints. (The objective function is undefined and irrelevant.) Specifically, checking (3.1) is equivalent to checking the following set of linear inequalities:

$$\begin{aligned} &\text{(i)} \quad X \in \mathbf{D}; \\ &\text{(ii)} \quad L(F)(X) \geqslant \frac{\gamma d_s(\mathbf{D})}{8} \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix}; \\ &\text{(iii)} \quad L(F)(X) \leqslant \frac{\gamma d_s(\mathbf{D})}{8} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \end{aligned} \tag{3.1}'$$

For example, if $\mathbf{D}$ is a box in $\mathbf{R}^n$, then the condition $X = (x_1, x_2, \ldots, x_n) \in \mathbf{D}$ reduces to

$$\begin{aligned} &x_i \geqslant a_i, \qquad i = 1, \ldots, n; \\ &x_i \leqslant b_i, \qquad i = 1, \ldots, n; \\ &\text{where } \mathbf{D} = \prod_{i=1}^{n} [a_i, b_i]. \end{aligned} \tag{3.1}'$$

The condition (3.1) may also be easily checked with interval arithmetic. In this case, a box $\mathbf{U}$ with $U(\mathbf{D}) \subset \mathbf{U}$ is explicitly computed by constructing a box $\mathbf{BD}$ with $\mathbf{D} \subset \mathbf{BD}$ and evaluating

$$\mathbf{U} = L(F)(\mathbf{BD}) + \frac{\gamma d_s(\mathbf{D})}{8} \begin{bmatrix} [-1, 1] \\ \vdots \\ [-1, 1] \end{bmatrix}.$$

However, a set $\mathbf{F}$ with $F(\mathbf{D}) \subset \mathbf{F}$ is computable by using interval arithmetic to evaluate F; in that case it is customary in practical algorithms to use a modification of Definition 3.6(ii) so that $\mathscr{T}_{0,F}(\mathbf{D}) =$ 'false' provided 0 is not contained in $\mathbf{F}$ (cf. [16], etc.).

THEOREM 3.7. *$\mathscr{T}_{0,F}$ is a root inclusion test.*

Proof of Theorem 3.7. We will check (i), (ii), and (iii) of Definition 2.4 in order.

First, assume $\mathscr{T}_{0,F}(\mathbf{D}) =$ "true". Then the Kantorovich theorem implies that there is a unique root of F in $\mathbf{D}$, so Definition 2.4(i) holds.

Second, assume $\mathcal{T}_{0,F}(\mathbf{D})$ = "false". If there is a Z in $\mathbf{D}$ such that $F(Z) = 0$, then

$$0 = L(F)(Z) + \frac{(Z - C)^{\mathrm{T}}\overline{H}(Z - C)}{2} \tag{3.2}$$

for some operator $\overline{H}$ such that the ith entry of $(Z - C)^{\mathrm{T}}\overline{H}(Z - C)$ is equal to $(Z - C)^{\mathrm{T}}H_i(\overline{X})(Z - C)$, where $H_i(\overline{X})$ is the Hessian matrix of f_i evaluated at some point $\overline{X}$ on the line segment joining C and Z. Assumption 3.1(iv) then implies

$$\frac{(Z - C)^{\mathrm{T}}\overline{H}(Z - T)}{2} = \frac{\gamma_0 d_s^2(\mathbf{D})}{8} T \tag{3.3}$$

for some vector $T = (t_1, t_2, \ldots, t_n)^{\mathrm{T}}$, $-1 \leqslant t_i \leqslant 1$ for $1 \leqslant i \leqslant n$. But (3.2) combined with (3.3) contradicts $\mathcal{T}_{0,F}(\mathbf{D})$ = "false", so (ii) of Definition 2.4 is verified.

Note that the above argument shows that the definition of $\mathcal{T}_{0,F}$ is consistent; i.e., $\mathcal{T}_{0,F}$ cannot simultaneously equal "true" and "false".

To verify Definition 2.4(iii), define

$$\varepsilon = \min\left\{ d_0, \frac{3}{M_1\gamma}\left(\frac{\sqrt{5}}{2} - 1\right), \frac{8}{3M_1 + \gamma(8 + \tau)} \frac{1}{6\gamma[M_1/18 + 1/(4M)]}, 2\varepsilon_0 \right\}.$$

Here,

$$d_0 = \frac{-M/2 + \sqrt{(M/2)^2 + (\gamma\tau m_0/2)}}{(\gamma\tau/4)},$$

where

$$m_0 = \min_{X \in \mathbf{D}_0 \setminus \cup_{i=1}^{p} \mathbf{B}_i} \| F(X) \|,$$

where $\mathbf{B}_i$ is a ball of radius $1/(\gamma M_1)$ (in $\| \cdot \|_s$) about Z_i, with $\{Z_i\}_{i=1}^{p} \subset \mathbf{D}_0$, the initial region, being the set of roots of F. (Note that $\cup_{i=1}^{p} \mathbf{B}_i$ can be empty.) In the above,

$$\tau = \max_{T \text{ in } [-1,1]^n} \| T \|,$$

and M and M_1 are the constants in Assumption 3.1(ii) and Assumption 3.1(iii), respectively. (Throughout this proof, $\| \cdot \|$ is the norm used on the range of F.)

Note that the compactness of $\mathbf{D}_0$ implies $\varepsilon > 0$.

To complete the verification of Definition 2.4(iii), assume

$$d_s = d_s(\mathbf{D}) < \varepsilon,$$

and also assume that if $F(Z) = 0$ then

$$d_s(Z, \partial\mathbf{D}) \geqslant (R/3) d_s = d_s/6.$$

There are two subcases: (a) there are no $Z \in \mathbf{D}$ with $F(Z) = 0$; and (b) there is such a $Z \in \mathbf{D}$.

For case (a), we need to show $\mathcal{T}_{0,F}(\mathbf{D})$ = "false". Assume case (a) and $\mathcal{T}_{0,F}(\mathbf{D})$ is not "false"; then the Kantorovich theorem implies that $\mathcal{T}_{0,F}(\mathbf{D})$ = "unknown", so $0 \in U(\mathbf{D})$. Thus, there are an $X \in \mathbf{D}$ and a $T \in [-1, 1]^n$ such that

$$0 = F(C) + J(C)(X - C) + \frac{\gamma d_s^2}{8} T, \tag{3.4}$$

whence

$$\| J(C)(X - C) \| \geqslant m_0 - \frac{\gamma d_s^2}{8}\tau, \tag{3.5}$$

where m_0 and τ are as above. But

$$\|J(C)(X-C)\| \leqslant \frac{Md_s}{2}. \tag{3.6}$$

Combining (3.5) and (3.6) then gives

$$m_0 + \frac{M}{2}d_s - \frac{\gamma\tau}{8}d_s^2 \leqslant 0. \tag{3.7}$$

This is true only for

$$d_s \geqslant \frac{-M/2 + \sqrt{(M/2)^2 + (\gamma\tau m_0/2)}}{(\gamma\tau/4)}.$$

But this contradicts the definition of ε, so $\mathscr{T}_{0,F}(\mathbf{D})$ = 'false'. Now assume there are roots Z of F in $\mathbf{D}_0$ (i.e., assume case (b) and $\mathscr{T}_{0,F}(\mathbf{D})$ is not 'false'). Denote the closest such root (in $\|\cdot\|_s$) to $\mathbf{D}$ by $\bar{Z}$. If $\|\bar{Z} - C\|_s \geqslant (1/\gamma M_1)$, then formulas (3.5), (3.6), and (3.7) still hold, and the definition of ε is still contradicted; thus we may assume

$$\|\bar{Z} - C\|_s < (1/\gamma M_1).$$

In this case, we have

$$\frac{d_s}{2} + \frac{R}{3}d_s = \frac{2}{3}d_s \leqslant \|C - \bar{Z}\|_s \leqslant \frac{1}{\gamma M_1}. \tag{3.8}$$

Also,

$$F(X) = J(\bar{Z})(X - \bar{Z}) + \frac{(X-\bar{Z})^{\mathrm{T}}\hat{H}(X-\bar{Z})}{2}$$

for some linear second-order operator $\hat{H}$ as in (3.2); therefore,

$$\begin{aligned}\|F(X)\| &\geqslant \|J(\bar{Z})(X-\bar{Z})\| - \frac{\|X-\bar{Z}\|_s^2}{2}\gamma \\ &\geqslant \frac{1}{M_1}\|X-\bar{Z}\|_s - \frac{\gamma}{2}\|X-\bar{Z}\|_s^2.\end{aligned} \tag{3.9}$$

We note, however, that since $X \in \mathbf{D}$,

$$\frac{d_s}{3} \leqslant \|X - \bar{Z}\|_s \leqslant \frac{4}{3}d_s. \tag{3.10}$$

Combining (3.9) and (3.10) gives

$$\|F(X)\| \geqslant \min\left\{\frac{1}{3M_1}d_s - \frac{\gamma}{18}d_s^2, \frac{4}{3M_1}d_s - \frac{8\gamma}{9}d_s^2\right\}. \tag{3.11}$$

On the other hand,

$$F(X) = F(C) + J(C)(X-C) + \frac{(X-C)^{\mathrm{T}}\hat{H}(X-C)}{2} \tag{3.12}$$

for $\hat{H}$ analogous to $\bar{H}$ in (3.2) and $\hat{H}$ above. Combining (3.2) and (3.4) gives

$$F(X) = \frac{(X-C)^{\mathrm{T}}\hat{H}(X-C)}{2} - \frac{\gamma d_s^2}{8}T,$$

whence

$$\|F(X)\| \leqslant \frac{d_s^2}{8}\gamma(1+\tau). \tag{3.13}$$

Combining (3.11) and (3.13) gives

$$\min\left\{\frac{1}{3M_1}d_s - \frac{\gamma}{18}d_s^2,\ \frac{4}{3M_1}d_s - \frac{8\gamma}{9}d_s^2\right\} \leqslant \frac{d_s^2}{8}\gamma(1+\tau), \tag{3.14}$$

from which

$$\begin{aligned} d_s &\geqslant \min\left\{\frac{8}{3M_1\gamma[(13/9)+\tau]},\ \frac{8}{3M_1\gamma[8+\tau]}\right\} \\ &= \frac{8}{3M_1\gamma[8+\tau]}. \end{aligned} \tag{3.15}$$

This contradicts the definition of ε. Therefore, if $d_s(\mathbf{D}) < \varepsilon$, $d_s(Z, \mathbf{D}) \geqslant (R/3)d_s(\mathbf{D})$ for every $Z \in \mathbf{D}_0$ with $F(Z) = 0$, and there are no $Z \in \mathbf{D}$ with $F(Z) = 0$, then $\mathscr{T}_{0,F}(\mathbf{D}) =$ 'false'.

The final case to consider is then

$$\begin{aligned} &\exists Z \in \mathbf{D} \quad \text{with } F(Z) = 0, \\ &d_s(Z, \partial\mathbf{D}) \geqslant \frac{R}{3}d_s(\mathbf{D}) = \frac{1}{6}d_s(\mathbf{D}), \end{aligned} \tag{3.16}$$

so that

$$\|Z - C\|_s \leqslant \frac{1}{2}d_s - \frac{1}{6}d_s = \frac{1}{3}d_s, \tag{3.17}$$

where $d_s = d_s(\mathbf{D})$. (To see (3.17), note that a ray emanating from C and passing through Z passes through $\partial\mathbf{D}$ at a distance of $d_s/2$; combine this with (3.16) and the triangle inequality.) In this final case, we will check the conditions of the Kantorovich theorem.

To estimate η, we expand F about C as in (3.2):

$$0 = F(C) + J(C)(Z - C) + \frac{(Z-C)^{\mathrm{T}}\bar{H}(Z-C)}{2}. \tag{3.18}$$

Furthermore, since $d_s < 2\varepsilon_0$, there follows $\|Z - C\|_s < \varepsilon_0$. Combining these latter facts with Assumption 3.1(iii) and (3.18) gives

$$\begin{aligned} \eta = \|J^{-1}(C)F(C)\|_s &\leqslant \|Z - C\|_s + M_1\gamma\frac{\|Z-C\|_s^2}{2} \\ &\leqslant \frac{1}{3}d_s + \frac{M_1\gamma}{18}d_s^2. \end{aligned} \tag{3.19}$$

Thus,

$$\beta\eta\gamma \leqslant M_1\gamma\left[\frac{1}{3}d_s + \frac{M_1\gamma}{18}d_s^2\right]. \tag{3.20}$$

The right member of (3.20) is bounded above by $(1/2)$ when

$$\left[\frac{M_1\gamma d_s}{3}\right]^2 + 2\left[\frac{M_1\gamma d_s}{3}\right] - 1 < 0, \tag{3.21}$$

which is true when

$$d_s < \frac{3}{M_1\gamma}\left[\frac{\sqrt{5}}{2} - 1\right]. \tag{3.22}$$

But (3.22) is true by the definition of ε, so the first condition in Definition (3.6)(i) is met. We check the second condition,

$$\frac{1 - \sqrt{1 - 2\alpha}}{\beta\gamma} < \frac{d_s(\mathbf{D})}{2}. \tag{3.23}$$

By rearranging (3.23), substituting $\alpha = \beta\gamma\eta$, and squaring to eliminate the radical, we see that (3.23) is true provided

$$\eta < \frac{1}{2}\left\{2\left[\frac{d_s}{2}\right] - \beta\gamma\left[\frac{d_s}{2}\right]^2\right\}. \tag{3.24}$$

Combining (3.19) and (3.24), we see that (3.24) is true provided

$$\frac{1}{3}d_s + \frac{M_1\gamma}{18}d_s^2 < \frac{1}{2}\left\{2\left[\frac{d_s}{2}\right] - \beta\gamma\left[\frac{d_s}{2}\right]^2\right\},$$

which reduces to

$$d_s\left[\frac{M_1\gamma}{18} + \frac{\beta\gamma}{4}\right] < \frac{1}{6}. \tag{3.25}$$

However, $\beta = \|J^{-1}(C)\| \geqslant (1/M)$, so (3.25) is true provided

$$d_s\left[\frac{M_1\gamma}{18} + \frac{\gamma}{4M}\right] < \frac{1}{6}. \tag{3.26}$$

Rearranging (3.25) gives

$$d_s < \frac{1}{6\gamma[(M_1/18) + (1/4M)]},$$

which is true by the assumption on ε. Thus, $\mathscr{T}_F(\mathbf{D}) =$ 'true' in the case defined by (3.16); this completes the proof of Theorem 3.7. □

Theorem 3.7 and the following lemma will allow us to apply Algorithm 2.6 and Theorem 2.10 in the case defined here.

LEMMA 3.8. *Assumption* 3.1 *implies* ε *is as in Definition* 2.5.

Proof of Lemma 3.8. The fact that J is continuous in $\mathbf{D}_0$ and invertible at $Z \in \mathbf{D}_0$ with $F(Z) = 0$ implies that the roots of F are isolated. More precisely, if $Z \in \mathbf{D}$ is such that $F(Z) = 0$, then (as in the formula preceding (3.9))

$$F(X) = J(\bar{Z})(X - \bar{Z}) + \frac{(X - \bar{Z})^{\mathrm{T}}\hat{H}(X - \bar{Z})}{2} \tag{3.27}$$

for some second-order linear operator $\hat{H}$. Thus, as in (3.9),

$$\|F(X)\| \geqslant \frac{1}{M_1}\|X - \bar{Z}\|_s - \frac{\gamma}{2}\|X - \bar{Z}\|_s^2$$

is true in this context. But the right member is positive for

$$0 < \|X - Z\|_s < \frac{2}{M_1\gamma}. \tag{3.28}$$

Thus, the distance between roots of F is at least

$$\frac{2}{M_1\gamma} > R\varepsilon = \frac{\varepsilon}{2},$$

where ε is as in Theorem 3.7. The latter formula, combined with Theorem 3.7, proves the assertion. □

Acknowledgment. I wish to thank the referees for their careful and thorough reading.

Department of Mathematics
The University of Southwestern Louisiana
P. O. Box 41010
Lafayette, Louisiana 70504-1010

1. E. L. ALLGOWER & K. GEORG, "Simplicial and continuation methods for approximating fixed points and solutions to systems of equations," *SIAM Rev.*, v. 22, 1980, pp. 28–85.

2. J. E. DENNIS & R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Least Squares*, Prentice-Hall, Englewood Cliffs, N. J., 1983.

3. A. EIGER, K. SIKORSKI & F. STENGER, "A method of bisections for solving n nonlinear equations," *ACM Trans. Math. Software*, v. 10, 1984, pp. 367–377.

4. C. B. GARCIA & T. Y. LI, "On the number of solutions to polynomial systems of equations," *SIAM J. Numer. Anal.*, v. 17, 1980, pp. 540–546.

5. C. B. GARCIA & W. I. ZANGWILL, *Pathways to Solutions, Fixed Points, and Equilibria*, Prentice-Hall, Englewood Cliffs, N. J., 1981.

6. E. HANSEN, "A globally convergent interval method for computing and bounding real roots," *BIT*, v. 18, 1978, pp. 415–424.

7. E. HANSEN, "Global optimization using interval analysis—The multidimensional case," *Numer. Math.*, v. 34, 1980, pp. 247–270.

8. E. HANSEN & S. SENGUPTA, "Bounding solutions of systems of equations using interval analysis," *BIT*, v. 21, 1981, pp. 203–211.

9. E. HANSEN, private communication, 1984.

10. C. HARVEY & F. STENGER, "A two-dimensional analogue to the method of bisections for solving nonlinear equations," *Quart. Appl. Math.*, v. 33, 1975, pp. 351–368.

11. F. B. HILDEBRAND, *Introduction to Numerical Analysis*, McGraw-Hill, New York, 1974.

12. R. B. KEARFOTT, *Computing the Degree of Maps and a Generalized Method of Bisection*, Ph.D. dissertation, University of Utah, 1977.

13. R. B. KEARFOTT, "An efficient degree-computation method for a generalized method of bisection," *Numer. Math.*, v. 32, 1979, pp. 109–127.

14. R. B. KEARFOTT, "On a general technique for finding directions proceeding from bifurcation points," in *Numerical Methods for Bifurcation Problems* (T. Küpper, H. D. Mittelmann, and H. Weber, eds.), Birkhäuser, Basel, 1984, pp. 210–218.

15. R. E. MOORE & S. T. JONES, "Safe starting regions for iterative methods," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 1051–1065.

16. A. P. MORGAN, "A method for computing all solutions to systems of polynomial equations," *ACM Trans. Math. Software*, v. 9, 1983, pp. 1–17.

17. W. C. RHEINBOLDT & J. V. BURKARDT, "A program for a locally-parametrized continuation process," *ACM Trans. Math. Software*, v. 9, 1983, pp. 215–235.

18. K. SIKORSKI, "A three-dimensional analogue to the method of bisections for solving nonlinear equations," *Math. Comp.*, v. 33, 1979, pp. 722–738.

19. F. STENGER, "An algorithm for the topological degree of a mapping in $\mathbf{R}^n$," *Numer. Math.*, v. 25, 1976, pp. 23–28.

20. M. J. TODD, *The Computation of Fixed Points and Applications*, Springer-Verlag, Berlin and New York, 1976.

21. A. TUCKER, *Applied Combinatorics*, Wiley, New York, 1980.