Numer. Math. 32, 109–127 (1979)

Numerische Mathematik

An Efficient Degree-Computation Method for a Generalized Method of Bisection

Baker Kearfott

Department of Mathematics, University of Southwestern Louisiana, Lafayette, Louisiana 70504, USA

Summary. Let $\mathbf{P}$ be an n-dimensional polyhedron and let $b(\mathbf{P}) = \sum_{q=1}^{m} \langle X_1^q, \ldots, X_n^q \rangle$ be the oriented boundary of $\mathbf{P}$ in terms of the oriented $(n-1)$-simplexes $\mathbf{S}_q = \langle X_1^q, \ldots, X_n^q \rangle$, $q = 1, \ldots, m$. Let $F = (f_1, \ldots, f_n)$: $\mathbf{P} \to R^n$, and assume $F(X) \neq \theta$ for $X \in b(\mathbf{P})$. For each $\langle X_1^q, \ldots, X_n^q \rangle \in b(\mathbf{P})$ define a matrix $\mathscr{R}(\mathbf{S}_q, F)$ by setting the entry in the i-th row, j-th column of $\mathscr{R}(\mathbf{S}_q, F)$ equal to 1 if $\operatorname{sgn}(f_j(X_i^q)) \neq 1$ and 0 if $\operatorname{sgn}(f_j(X_i^q)) = -1$, where $\operatorname{sgn}(y) = 1$ if $y \geqq 0$, and $\operatorname{sgn}(y) = -1$ otherwise. To each such matrix $\mathscr{R}(\mathbf{S}_q, F)$ assign a number $\operatorname{Par}(\mathscr{R}(\mathbf{S}_q, F))$ in the following way: Set $\operatorname{Par}(\mathscr{R}(\mathbf{S}_q, F)) = +1$ if the entries on and below the main diagonal of $\mathscr{R}(\mathbf{S}_q, F)$ are 1 and the entries one row above the main diagonal are 0. Also set $\operatorname{Par}(\mathscr{R}(\mathbf{S}_q, F)) = 1$ if $\mathscr{R}(\mathbf{S}_q, F)$ can be put into this form by an even permutation of its rows, and set $\operatorname{Par}(\mathscr{R}(\mathbf{S}_q, F)) = -1$ if $\mathscr{R}(\mathbf{S}_q, F)$ can be put into form by an odd permutation of rows. Set $\operatorname{Par}(\mathscr{R}(\mathbf{S}_q, F)) = 0$ for all other matrices $\mathscr{R}(\mathbf{S}_q, F)$. Then, under rather general hypotheses and assuming diameter of each $\mathbf{S}_q \in b(\mathbf{P})$ is small, the topological degree of F at θ relative to $\mathbf{P}$ is given by:

$$d(F, \mathbf{P}, \theta) = \sum_{q=1}^{m} \operatorname{Par}(\mathscr{R}(\mathbf{S}_q, F)).$$

The assumptions are identical to those used by Stenger (Numer. Math. 25, 23–28).

Use of the characterization is illustrated, an algorithm for automatic computation is presented, and an application of this algorithm to finding roots of $F(X) = \theta$ is explained. The degree computation algorithm requires storage of a number of $(n-1)$-simplexes proportional to $\log n$, and $\operatorname{sgn}(f_j(\mathbf{S}_i^q))$ is evaluated once at most for each i, j, and q.

Subject Classifications. AMS(MOS): 65 H 10; CR: 5.15.

0029-599X/79/0032/0109/$03.80

1. Introduction

Suppose that $F: \mathscr{D} \to R^n$ is a differentiable function mapping some bounded domain $\mathscr{D} \subset R^n$ into R^n, and suppose that $F|b(\mathscr{D})$ does not vanish, where $b(\mathscr{D})$ is the boundary of $\mathscr{D}$. Suppose further that if $X \in \mathscr{D}$ is such that $F(X)=\theta$ then $J(F)(X)$, the Jacobian of F at X, is non-zero. Then the degree of F at θ relative to $\mathscr{D}$, written $d(F, \mathscr{D}, \theta)$, can be defined to be the number of points $X \in \mathscr{D}$ with $F(X)=\theta$ and $J(F)(X)>0$ minus the number of $X \in \mathscr{D}$ with $F(X)=\theta$ and $J(F)(X)<0$.

The above definition can be generalized when F is merely continuous ([4, 11]). Furthermore, when $d(F, \mathscr{D}, \theta)$ is thus defined for continuous F, Kronecker's theorem ([1, 4] p. 25, [11] p. 161) states that F has a root in $\mathscr{D}$ if $d(F, \mathscr{D}, \theta) \neq 0$. Moreover, if $\mathscr{D}=\mathscr{D}_1 \cup \mathscr{D}_2$ where $\mathscr{D}_1$ and $\mathscr{D}_2$ have disjoint interiors and $F(X) \neq \theta$ for all $X \in b(\mathscr{D}_1) \cup b(\mathscr{D}_2)$, then the degree is additive, i.e.:

$$d(F, \mathscr{D}, \theta)=d(F, \mathscr{D}_1, \theta)+d(F, \mathscr{D}_2, \theta).$$

Kronecker's theorem and the additivity of the topological degree form cornerstones of a generalized method of bisection based on computation of $d(F, \mathscr{D}, \theta)$.

To date, several characterizations of $d(F, \mathscr{D}, \theta)$ have been expounded and implemented in machine computations. Perhaps the oldest of these is the Kronecker integral ([1] p. 465, [7]). This integral has been evaluated with Gauss-Legendre quadrature ([10]).

A characterization involving calculation of determinants with entries ± 1 has also been proposed for computation of $d(F, \mathscr{D}, \theta)$ ([13]). This characterization has the advantage of requiring the components of F to be evaluated to sufficient accuracy only to determine their algebraic sign.

In Sect. 3 of this paper, a third characterization of $d(F, \mathscr{D}, \theta)$ will be presented and proven. This characterization also only requires the algebraic signs of the components of F to be evaluated, but it does not involve computation of determinants. The number of functional evaluations required for a given F and $\mathscr{D}$ is the same or less than that required for the determinant characterization.

In Sect. 4 of this paper, an algorithm to implement the new characterization is given. It seems to compare favorably with previously existing means of computing $d(F, \mathscr{D}, \theta)$.

In Sect. 5, the algorithm for computing $d(F, \mathscr{D}, \theta)$ is applied in a generalized method of bisection. This method of bisection does not require redundant calculation of $F(X)$ and only $1/(n+1)$ times the number of operations involved in initial computation of $d(F, \mathscr{D}, \theta)$ are usually needed per iteration of the bisection algorithm.

In Sect. 6, results of some preliminary degree-computation tests are given. In the last section, the scope and applicability of the method is discussed.

It is mentioned that objects considered in the characterization of $d(F, \mathscr{D}, \theta)$ happen to be Sperner simplexes for a labeling function determined by $n-1$ of the components of F (see [2, 3, 8] for an explanation of the labeling function). This relationship is discussed in [9].

2. Notation, Assumptions, and a Preliminary Lemma

Throughout, $\mathbf{S}$ will denote an oriented n-simplex in R^n, i.e. $\mathbf{S}$ will be the oriented closed convex hull of $n+1$ linearly independent points in R^n (see [4, 6], etc.). The equation:

$$\mathbf{S}=\langle X_0, X_1, \ldots, X_n\rangle$$

will read: $\mathbf{S}$ is the oriented n-simplex whose set of vertices is $\{X_0, \ldots, X_n\}$ and whose orientation is given by the order in which the vertices occur ([1, 4, 6], etc.). If $X_0, \ldots, X_m$ are $m+1$ linearly independent points in R^n where $m \leqq n$, then we will speak of the oriented m-simplex $\langle X_1, \ldots, X_m\rangle$ to refer to the oriented closed convex hull of $\{X_1, \ldots, X_m\}$ in R^n.

If $\mathbf{S}=\langle X_0, \ldots, X_n\rangle$ is an oriented n-simplex then the algebraic boundary of $\mathbf{S}$ will be given by:

$$b(\mathbf{S})=\sum_{i=0}^{n}(-1)^i\langle X_0, \ldots, \hat{X}_i, \ldots, X_n\rangle$$

where $\langle X_0, \ldots, \hat{X}_i, \ldots, X_n\rangle$ is the oriented $(n-1)$-simplex in R^n formed by deleting X_i from the list of vertices for $\mathbf{S}$ and $\sum$ has the usual meaning (ibid). The simplex $\mathbf{T}_i=\langle X_0, \ldots, \hat{X}_i, \ldots, X_n\rangle$ will be called the i-th facet of $\mathbf{S}$.

If $\mathbf{P}=\bigcup_{i=1}^{q}\mathbf{S}_i$ where each $\mathbf{S}_i$ is an n-simplex and the $\mathbf{S}_i$ have pairwise-disjoint interiors, then we refer to $\mathbf{P}$ as an n-polygon. We write:

$$\mathbf{P}=\sum_{i=1}^{q}\mathbf{S}_i$$

and:

$$b(\mathbf{P})=\sum_{i=1}^{q}b(\mathbf{S}_i).$$

The results in this paper are stated for polygons or for simplexes. However, $d(F, \mathscr{D}, \theta)$ can be calculated for general regions $\mathscr{D}$ by either approximating $\mathscr{D}$ by a polygon or mapping a polygon onto $\mathscr{D}$.

The characterization of $d(F, \mathbf{P}, \theta)$ proven in this paper depends upon $b(\mathbf{P})$ being written as a sum of $(n-1)$-simplexes, all of whose diameters are small. This property of $b(\mathbf{P})$ is clarified in the following two definitions, which are slight modifications of those in [13].

Definition. Suppose $\mathbf{P}=[a, b]\subset R$ and $f: P\to R$ is continuous. Then $b(\mathbf{P})=b-a$ is *sufficiently refined* relative to $\operatorname{sgn}(f)$ if and only if $f(b)\neq 0$ and $f(a)\neq 0$.

Definition. Suppose $\mathbf{P}$ is an n-polygon and $F=(f_1, \ldots, f_n)$: $\mathbf{P}\to R^n$. Then we say $b(\mathbf{P})$ is sufficiently refined relative to $\operatorname{sgn}(F)$ if and only if $b(\mathbf{P})$ is written as the union of a finite number of $(n-1)$-polygons $\beta_{n-1}^1, \ldots, \beta_{n-1}^k$ which have the following three properties:

(a) The interiors of the β_{n-1}^r are disjoint and each β_{n-1}^r is connected.

(b) At least one of the components of F does not vanish on β_{n-1}^r for each r between 1 and k.

(c) If $f_{i_r} \neq 0$ on β_{n-1}^r, then β_{n-1}^r is sufficiently refined relative to $\text{sgn}(F_{i_r})$, where

$$F_{i_r} = (f_1, \ldots, f_{i_r-1}, f_{i_r+1}, \ldots, f_n).$$

It is not difficult to show that the above definition can replace Definition 4.4 in [13] without altering the truth of any theorems ([14]).

A recursion formula can now be stated. This formula relates $d(F, \mathbf{P}, \theta)$, where $\mathbf{P}$ is an n-polygon and F: $\mathbf{P} \to R^n$, to degrees of a truncated mapping F defined on $(n-1)$-polygons obtained from $\mathbf{P}$.

Lemma. (The Recursion Formula.) Suppose $\mathbf{P}$ is an n-polygon, suppose $F = (f_1, \ldots, f_n)$: $\mathbf{P} \to R^n$ is continuous, and suppose $b(\mathbf{P})$ is sufficiently refined relative to $\text{sgn}(F)$. Let $\beta_{n-1}^1, \ldots, \beta_{n-1}^k$ be as in the definition of sufficient refinement, and let J be the set of indices such that $i \in J$ if and only if $f_1 > 0$ on β_{n-1}^1. Then the following formula holds:

$$d(F, \mathbf{P}, \theta) = \sum_{i \in J} d(F_1, \beta_{n-1}^i, \theta),$$

where $F_1 = (f_2, \ldots, f_n)$: $b(\mathbf{P}) \to R^{n-1}$.
A general version of the above lemma appears as Formula (4.15) in [13]. See also [9] for a proof.

3. A Useful Characterization of $d(F, \mathbf{P}, \theta)$

The following concepts will be used to state the main characterization.

Definition. Let $\mathbf{S} = \langle X_1, \ldots, X_n \rangle$ be an $(n-1)$-simplex in R^n and let $F = (f_1, \ldots, f_n)$: $\mathbf{S} \to R^n$. Then the *range simplex associated with* $\mathbf{S}$ *and* F, denoted $\mathscr{R}(\mathbf{S}, F)$, is the $n \times n$ matrix whose entry in the j-th column and i-th row is 1 if $f_j(X_i) > 0$ and whose entry in the j-th column and i-th row is 0 if $f_j(X_i) < 0$.

Definition. Suppose $\mathbf{S}$, F, and $\mathscr{R}(\mathbf{S}, F)$ are as in the preceding definition. Then $\mathscr{R}(\mathbf{S}, F)$ is termed *useable* if one of the following two conditions holds:

(a) The entries of $\mathscr{R}(\mathbf{S}, F)$ on and below the main diagonal are 1, and the entries in each column of $\mathscr{R}(\mathbf{S}, F)$ in the row immediately above the diagonal entry are all 0.

(b) $\mathscr{R}(\mathbf{S}, F)$ can be put in to the form indicated in (a) by a permutation of its rows.

Definition. If $\mathscr{R}(\mathbf{S}, F)$ is useable, then $\text{Par}(\mathscr{R}(\mathbf{S}, F))$ is defined to be 1 if the permutation required to put $\mathscr{R}(\mathbf{S}, F)$ into the form (a) of the definition of useable is even; if that permutation is odd, then $\text{Par}(\mathscr{R}(\mathbf{S}, F))$ is defined to be -1. If $\mathscr{R}(\mathbf{S}, F)$ is not useable, we set $\text{Par}(\mathscr{R}(\mathbf{S}, F)) = 0$. We refer to $\text{Par}(\mathscr{R}(\mathbf{S}, F))$ as the *parity* of the matrix $\mathscr{R}(\mathbf{S}, F)$.

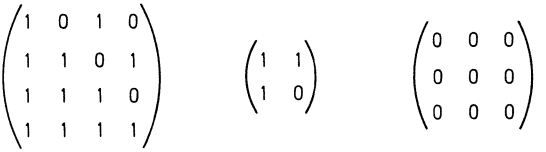

Fig. 1a–c (a) (b) (c)

For example, if $\mathscr{R}(\mathbf{S}, F)$ is as in Fig. 1a then $\mathrm{Par}(\mathscr{R}(\mathbf{S}, F)) = 1$, while in Fig. 1b and Fig. 1c $\mathrm{Par}(\mathscr{R}(\mathbf{S}, F)) = -1$ and $\mathrm{Par}(\mathscr{R}(\mathbf{S}, F)) = 0$, respectively.

Remark. It is often less cumbersome to formulate the definition of useable $\mathscr{R}(\mathbf{S}, F)$ and of $\mathrm{Par}(\mathscr{R}(\mathbf{S}, F))$ in terms of labelings. To the k-th row of $\mathscr{R}(\mathbf{S}, F)$ (for $k \in \{1, \ldots, n\}$) we assign a label $l_k \in \{1, \ldots, n\}$ by setting $l_k = l - 1$ if the first 0 (reading from left to right) in that row occurs in the l-th column; if there are no zeros in the k-th row, we set $l_k = n$. It is then easy to see that $\mathscr{R}(\mathbf{S}, F)$ is useable if and only if $\{l_1, \ldots, l_n\} = \{1, \ldots, n\}$. Furthermore, supposing $\mathscr{R}(\mathbf{S}, F)$ is useable, then $\mathrm{Par}(\mathscr{R}(\mathbf{S}, F)) = 1$ iff the permutation required to put these sequence $l_1, \ldots, l_n$ in natural order is even and $\mathrm{Par}(\mathscr{R}(\mathbf{S}, F)) = -1$ iff that permutation is odd.

Suppose now that $\mathbf{P}$ is an n-polygon and that $F: \mathbf{P} \to R^n$ is continuous and does not vanish on $b(\mathbf{P})$. Then one need only examine the useable simplexes produced from a sufficient refinement of $b(\mathbf{P})$ in order to determine $d(F, \mathbf{P}, \theta)$. The following theorem (our main characterization) makes this fact explicit:

Theorem (The Parity Theorem). Suppose $\mathbf{P}$ is an n-polygon contained in R^n for some $n \geqq 2$. Suppose further that $\{\mathbf{S}_i\}_{i=1}^m$ is a finite set of $(n-1)$-simplexes such that $\sum_{i=1}^{m} \mathbf{S}_i = b(\mathbf{P})$, the members of $\{\mathbf{S}_i\}_{i=1}^m$ have disjoint interiors, and the simplexes $\mathbf{S}_i$ make $b(\mathbf{P})$ suffiently refined relative to $\mathrm{sgn}(F)$. Then:

$$d(F, \mathbf{P}, \theta) = \sum_{i=1}^{m} \mathrm{Par}(\mathscr{R}(\mathbf{S}_i, F)).$$

The Parity Theorem deals with the matrices $\mathscr{R}(\mathbf{S}, F)$. To enable us to use induction to prove the Parity Theorem we will define a number associated with submatrices of such $\mathscr{R}(\mathbf{S}, F)$: Suppose that the first column of $\mathscr{R}(\mathbf{S}, F)$ has only $+1$'s in it and set $F_1 = (f_2, \ldots, f_n)$. Also, consider $b(\mathbf{S}) = \sum_{k=1}^{n} (-1)^{k-1} \mathbf{T}_k$, where $\mathbf{T}_k = \langle X_1, \ldots, \hat{X}_k, \ldots, X_n \rangle$. We see that the range simplexes $\mathscr{R}(\mathbf{T}_k, F_1)$ can be obtained from $\mathscr{R}(\mathbf{S}, F)$ by deleting the first row and k-th column of $\mathscr{R}(\mathbf{S}, F)$. However, $\mathbf{T}_k$ occurs in the sum for $b(\mathbf{S})$ with an orientation of $(-1)^{k-1}$, and changing the orientation of any $(n-2)$-simplex $\mathbf{T}$ changes the sign of $\mathrm{Par}(\mathscr{R}(\mathbf{T}, F_1))$. With this in mind we make the definition:

Definition. The *net sum* $\sigma(\mathbf{S})$ of the parities of the range simplexes associated with simplexes from $b(\mathbf{S})$ and F_1 is given by:

$$\sigma(\mathbf{S}) = \sum_{k=1}^{n} (-1)^{k-1} \mathrm{Par}(\mathscr{R}(\mathbf{T}_k, F_1)).$$

For example, if:

$$\mathscr{R}(\mathbf{S}, F) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

then the net sum is:

$$\sigma(\mathbf{S}) = (-1)^{1-1} \operatorname{Par}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + (-1)^{2-1} \operatorname{Par}\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} + (-1)^{3-1} \operatorname{Par}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$
$$= 0 + (-1)(-1) + 0 = 1.$$

Dealing with the above notions, the following lemma will be used to prove the Parity Theorem.

Lemma. Suppose $\mathbf{S}, F$, $\mathscr{R}(\mathbf{S}, F)$, $\sigma(\mathbf{S})$, and n are as in the preceeding definitions, and $n \geqq 3$. If the entries in the second column of $\mathscr{R}(\mathbf{S}, F)$ are all 1, then $\sigma(\mathbf{S}) = 0$.

Proof. We will think of $\operatorname{Par}(\mathscr{R}(\mathbf{S}, F))$ in terms of the labels $l_1, \ldots, l_n$ assigned to the rows of $\mathscr{R}(\mathbf{S}, F)$. In order for $\sigma(\mathbf{S})$ to be non-zero, $\operatorname{Par}(\mathscr{R}(\mathbf{T}_k, F_1)) \neq 0$ for some k. But this can happen if and only if $\{2, \ldots, n\} \subseteq \{l_1, \ldots, l_n\}$, from the definition of $\operatorname{Par}(\mathscr{R}(\mathbf{T}_k, F_1))$, so assume $\{2, \ldots, \mathrm{n}\} \subset \{l_1, \ldots, l_n\}$. But $1 \notin \{l_1, \ldots, l_n\}$ by the assumption on the second row of $\mathscr{R}(\mathbf{S}, F)$. Hence there are a k and m such that $l_k = l_m$, and such that all of the other l_j's are distinct. From this we deduce:

$$\sigma(\mathbf{S}) = (-1)^{k-1} \operatorname{Par}(\mathscr{R}(\mathbf{T}_k, F_1)) + (-1)^{m-1} \operatorname{Par}(\mathscr{R}(\mathbf{T}_m, F_1)).$$

It will now be shown that $(-1)^{k-1} \operatorname{Par}(\mathscr{R}(\mathbf{T}_k, F_1)) + (-1)^{m-1} \operatorname{Par}(\mathscr{R}(\mathbf{T}_m, F_1)) = 0$. Consider these two possibilities for k and m:

(a) k and m are both odd or both even.
(b) k and m have opposite parities.

If case (a) occurs, then $(-1)^{k-1} = (-1)^{m-1}$. Suppose without loss of generality that $k < m$. Since $l_k = l_m$, however, we can get the sequence $l_1, \ldots, l_k, \ldots, l_{m-1}, l_{m+1}, \ldots, l_n$ from the sequence $l_1, \ldots, l_{k-1}, l_{k+1}, \ldots, l_m, \ldots, l_n$ by the $(k-m)$-cycle: $(l_{k+1}, l_{k+2}, \ldots, l_{m-1}, l_m)$. If $(k-m)$ is even then the parity of that permutation is odd, and $\operatorname{Par}(\mathscr{R}(\mathbf{T}_k, F_1)) = -\operatorname{Par}(\mathscr{R}(\mathbf{T}_m, F_1))$. Hence, when k and m have the same parity we have:

$$\sigma(\mathbf{S}) = (-1)^{k-1} \{\operatorname{Par}(\mathscr{R}(\mathbf{T}_k, F_1)) + \operatorname{Par}(\mathscr{R}(\mathbf{T}_m, F_1))\} = 0.$$

If we follow the same argument when k and m have opposite parities we get:

$$\sigma(\mathbf{S}) = \{(-1)^{k-1} + (-1)^{m-1}\} \operatorname{Par}(\mathscr{R}(\mathbf{T}_k, F_1)) = 0.$$

Thus the lemma is proven □

Proof of the Parity Theorem. The proof will proceed by induction on n. First, assume that $\mathbf{P}$ is a 2-polygon, that $F = (f_1, f_2)$: $\mathbf{P} \to R^2$ is continuous, and that $\{\mathbf{S}_i\}_{i=1}^m$ is a set of 1-simplexes whose union is $b(\mathbf{P})$ and such that $\{\mathbf{S}_i\}_{i=1}^m$ causes

$b(\mathbf{P})$ to be sufficiently refined relative to $\mathrm{Sgn}(F)$. Then the recursion formula in Sect. 2 holds with $n=2$, so that:

$$d(F,\mathbf{P},\theta)=\sum_{i\in J} d(f_2,\beta_1^i,0) \tag{3.1}$$

where the β_1^i are as in the definition of sufficient refinement. However, when $n=2$ each β_1^i is a sum $\sum_{j=1}^{m}\mathbf{S}_{j,i}$ of $\mathbf{S}_{j,i}\in\{\mathbf{S}_i\}_{i=1}^m$, where $\mathbf{S}_{j,i}=\langle A_{j,i},B_{j,i}\rangle$ is a line segment and $1\leqq j\leqq m_i$ for some m_i. Hence, ([13] p. 24):

$$d(f_2,\beta_1,0)=\sum_{j=1}^{m_i}[\mathrm{sgn}(f_2(B_{j,i}))-\mathrm{sgn}(f_2(A_{j,i}))]. \tag{3.2}$$

Combining (3.1) and (3.2) now gives:

$$d(F,\mathbf{P},\theta)=\sum_{i\in J}\sum_{j=1}^{m_i}[\mathrm{sgn}(f_2(B_{j,i}))-\mathrm{sgn}(f_2(A_{j,i}))]. \tag{3.3}$$

It will be shown that each $[\mathrm{sgn}(f_2(B_{j,i}))-\mathrm{sgn}(f_2(A_{j,i}))]$ in (3.3) can be replaced by $\mathrm{Par}(\mathscr{R}(\mathbf{S}_{j,i},F))$ without affecting the sum. Hereafter, unless subscripts are important, $\mathbf{S}$ will be written for $\mathbf{S}_{j,i}$. First it will be shown that without loss of generality $\mathbf{S}\subseteq\beta_1^i$ for some $i\in J$ if and only if the first column of $\mathscr{R}(\mathbf{S},F)$ is $(1,1)^t$. Clearly, if $\mathbf{S}\in\beta_1^i$ for some $i\in J$, then the first column of $\mathscr{R}(\mathbf{S},F)$ is $(1,1)^t$. For the converse, suppose that the first column of $\mathscr{R}(\mathbf{S},F)$ is $(1,1)^t$. Then by the sufficient refinement hypothesis $f_2>0$ or $f_2<0$ on $\mathbf{S}$, so that

$$[\mathrm{sgn}(f_2(B_{j,i}))-\mathrm{sgn}(f_2(A_{j,i}))]=\mathrm{Par}(\mathscr{R}(\mathbf{S},F))=0.$$

In this case, $\mathbf{S}$ may be included in the sum in the Parity Theorem.

There are four matrices $\mathscr{R}(\mathbf{S},F)$ whose first column is $(1,1)^t$. These are:

(a) $\begin{pmatrix}1&1\\1&1\end{pmatrix}$, (b) $\begin{pmatrix}1&0\\1&0\end{pmatrix}$, (c) $\begin{pmatrix}1&0\\1&1\end{pmatrix}$, (d) $\begin{pmatrix}1&1\\0&1\end{pmatrix}$.

These four matrices are checked individually to prove the theorem for $n=2$.

Suppose now that the theorem is true with $(n-1)$ replacing n, where $n>2$; also let $\mathbf{P}$ be an n-polygon, let $F=(f_1,\ldots,f_n)\colon \mathbf{P}\to R^n$ be continuous, and let $b(\mathbf{P})=\sum_{i=1}^{m}\mathbf{S}_i$ be sufficiently refined relative to $\mathrm{sgn}(F)$. Then the Recursion Formula holds and:

$$d(F,\mathbf{P},\theta)=\sum_{i\in J} d(F_1,\beta_{n-1}^i,\theta) \tag{3.4}$$

where the b_{n-1}^i are as in the definition of sufficient refinement. But by the induction hypothesis, the inductive part of the definition of sufficient refinement, and the definition of $\sigma(\mathbf{S})$ we have:

$$d(F_1,\beta_{n-1}^i,\theta)=\sum_{\mathbf{T}\subset b(\beta_{n-1}^i)}\mathrm{Par}(\mathscr{R}(\mathbf{T},F_1))=\sum_{\mathbf{S}\subset\beta_{n-1}^i}\sigma(\mathbf{S})\quad\text{for } i\in J. \tag{3.5}$$

Combining (3.4) and (3.5) gives:

$$d(F,\mathbf{P},\theta)=\sum_{i\in J}\sum_{\mathbf{S}\subseteq\beta^i_{n-1}}\sigma(\mathbf{S}). \tag{3.6}$$

It is first verified (infra) that $\mathbf{S}$ need be included in the right member of (3.6) if and only if $\mathscr{R}(\mathbf{S},F)$ is useable. Then we apply the fact that, if $\mathscr{R}(\mathbf{S},F)$ is useable, then $\sigma(\mathbf{S})=\mathrm{Par}(\mathscr{R}(\mathbf{S},F))$ (infra). Hence:

$$\sum_{i\in J}\sum_{\mathbf{S}\subseteq\beta^i_{n-1}}\sigma(\mathbf{S})=\sum_{\mathbf{S}\,\text{useable}}\mathrm{Par}(\mathscr{R}(\mathbf{S},F))=\sum_{i=1}^{m}\mathrm{Par}(\mathscr{R}(\mathbf{S}_i,F)). \tag{3.7}$$

Combining (3.6) and (3.7) gives the result of the theorem. □

The following are auxiliary lemmas used in the proof of the Parity Theorem.

Lemma. Suppose $b(\mathbf{P})$ is sufficiently refined relative to $\mathrm{Sgn}(F)$, $\mathbf{S}\subset b(\mathbf{P})$, and $\mathscr{R}(\mathbf{S},F)$ is useable. Then $\mathbf{S}\subseteq\beta^i_{n-1}$ for some $i\in J$. Also, suppose $\mathbf{S}\subseteq\beta^i_{n-1}$ for some $i\subset J$ and $\mathscr{R}(\mathbf{S},F)$ is not useable. Then $\sigma(\mathbf{S})=0$.

Proof. If $\mathscr{R}(\mathbf{S},F)$ is useable, then the first column of $\mathscr{R}(\mathbf{S},F)$ is the only column which does not contain both 1's and 0's. But, since at least one component of F does not vanish on $\mathbf{S}$ by the sufficient refinement hypotheses, $\mathbf{S}\subseteq\beta^i_{n-1}$ for some $i\in J$.

Now suppose $\mathbf{S}\subseteq\beta^i_{n-1}$ for some $i\in J$ and $\mathscr{R}(\mathbf{S},F)$ is not useable. Then there are more than one, precisely one, or no zeros in the second column of $\mathscr{R}(\mathbf{S},F)$. If there are no zeros in the second column of $\mathscr{R}(\mathbf{S},F)$, then $\sigma(\mathbf{S})=0$ by the lemma preceeding the Parity Theorem. Finally, if there are one or more zeros in the second column of $\mathscr{R}(\mathbf{S},F)$ and $\mathscr{R}(\mathbf{S},F)$ is not useable, then 1 appears at least once in the list $\{l_1,\ldots,l_n\}$, so there is a $j\in\{2,3,\ldots,n\}$ such that $j\notin\{l_1,\ldots,l_n\}$. In that case $\mathrm{Par}(\mathscr{R}(\mathbf{T}_k,F_1))=0$ for every $k\in\{1,\ldots,n\}$ so $\sigma(\mathbf{S})=0$. □

Lemma. Suppose $\mathscr{R}(\mathbf{S},F)$ is useable. Then $\sigma(\mathbf{S})=\mathrm{Par}(\mathscr{R}(\mathbf{S},F))$.

Proof. Suppose $\mathscr{R}(\mathbf{S},F)$ is useable and set $\mathrm{b}(\mathbf{S})-\sum_{k=1}^{n}(-1)^{k-1}\mathbf{T}_k$ as in the definition of $\sigma(\mathbf{S})$. Then, since $\mathscr{R}(\mathbf{T}_k,F_1)$ would not have any 0's in its first column if it were useable, $\mathrm{Par}(\mathscr{R}(\mathbf{T}_k,F_1))=0$ for all but one k, *e.g.* k_0. Let P_1 be the permutation on $1,\ldots,n$ required to put the k_0-th row of $\mathscr{R}(\mathbf{S},F)=\mathscr{R}(\langle X_1,\ldots,X_n\rangle,F)$ into the first position, leaving the other rows fixed relative to each other; also, let P_2 be the permutation on $1,2,\ldots,k_0-1,k_0+1,\ldots,n$ required to put $\mathscr{R}(\langle X_1,\ldots,\hat{X}_k,\ldots,X_n\rangle,F_1)$ into form (a) of the definition of useable $\mathscr{R}(\mathbf{S},F)$. Let P be the permutation required to put $\mathscr{R}(\mathbf{S},F)$ into form (a) of the definition of useable. Then, thinking of P_2 as a permutation on n objects which leaves the first object fixed, we have:

$$P_3=P_2P_1. \tag{3.8}$$

However, $P_1=(1,2,\ldots,k_0-1,k_0)$ is a k_0-cycle whose parity is the parity of the integer k_0-1. Hence:

$$\mathrm{Par}(\mathscr{R}(\mathbf{S},F))=(-1)^{k_0-1}\,\mathrm{Par}(\mathscr{R}(\mathbf{T}_{k_0},F_1)). \tag{3.9}$$

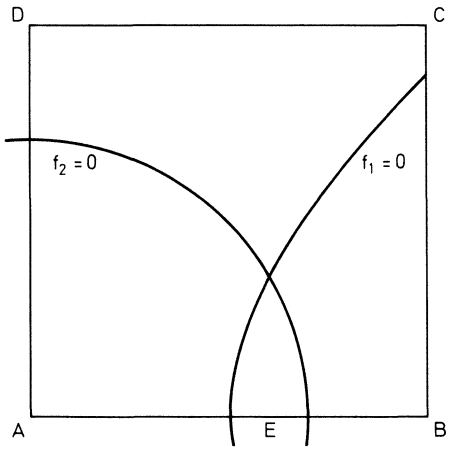

Fig. 2

However, since $(-1)^{k_0-1}\,\mathrm{Par}(\mathscr{R}(\mathbf{T}_{k_0}, F_1))$ was the only non-zero term in the sum for $\sigma(\mathbf{S})$, the conclusion of the lemma follows from (3.9). □

Before going on to investigate automatic computation of $d(F, \mathbf{P}, \theta)$, an example of the Parity Theorem will be given.

Example. Suppose that $n=2$, and $F(X)=(f_1(X), f_2(X))$, where $X=(x, y)$, $f_1(X)=x^2-y^2-1$, and $f_2(X)=x^2+y^2-2$. Suppose that $\mathbf{P}$ is the rectangle $\{0\leqq x\leqq 2; 0\leqq y\leqq 2\}$. Then compute $d(F, \mathbf{P}, \theta)$ via the Parity Theorem.

The polygon $\mathbf{P}$ is drawn in Fig. 2. Set $E=1.25$. Then $b(\mathbf{P})$ is sufficiently refined relative to F provided we write:

$$b(\mathbf{P})=\langle A, E\rangle+\langle E, B\rangle+\langle B, C\rangle+\langle C, D\rangle+\langle D, A\rangle. \tag{3.11}$$

We then have:

$$\begin{aligned} d(F, \mathbf{P}, \theta)=&\,\mathrm{Par}(\mathscr{R}(\langle A, E\rangle, F))+\mathrm{Par}(\mathscr{R}(\langle E, B\rangle, F)) \\ &+\mathrm{Par}(\mathscr{R}(\langle B, C\rangle, F))+\mathrm{Par}(\mathscr{R}(\langle C, D\rangle, F)) \\ &+\mathrm{Par}(\mathscr{R}(\langle D, A\rangle, F)). \end{aligned} \tag{3.12}$$

We immediately see:

$$\mathscr{R}(\langle A, E\rangle, F)=\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

$$\mathscr{R}(\langle E, B\rangle, F)=\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

$$\mathscr{R}(\langle B, C\rangle, F)=\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$\mathscr{R}(\langle C, D\rangle, F) = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and}$$

$$\mathscr{R}(\langle D, A\rangle, F) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Hence

$$\begin{aligned} \mathrm{Par}(\mathscr{R}(\langle A, E\rangle, F)) &= \mathrm{Par}(\mathscr{R}(\langle B, C\rangle, F)) = \mathrm{Par}(\mathscr{R}(\langle C, D\rangle, F)) \\ &= \mathrm{Par}(\mathscr{R}(\langle D, A\rangle, F) = 0, \end{aligned}$$

and $\mathrm{Par}(\mathscr{R}(\langle E, B\rangle, F) = 1$. Therefore, $d(F, \mathbf{P}, \theta) = 1$ by the Parity Theorem.

It is easy to see that $X = (\sqrt{\frac{3}{2}}, \sqrt{\frac{1}{2}})$ is the only solution in $\mathbf{P}$ of $F(X) = \theta$, and that $J(F)(X) > 0$ for this X, where $J(F)$ is the Jacobian of F, thus illustrating the validity of the Parity Theorem for our example.

4. An Algorithm to Compute $d(F, \mathbf{P}, \theta)$

The result of the Parity Theorem indicates that we can easily compute $d(F, \mathbf{P}, \theta)$ in terms of simplexes comprising $b(\mathbf{P})$, provided the diameters of those simplexes are "small enough" (i.e., so sufficient refinement is attained). Below, an iterative procedure of subdividing the $(n-1)$-simplexes of $b(\mathbf{P})$ so that the diameters of the resulting $(n-1)$-simplexes become small is described.

This procedure is a generalization of bisection of line segments in R^1. To carry out the procedure on a simplex $\mathbf{S}$, we find the longest line segment to be formed by taking the vertices of $\mathbf{S}$ two at a time, then use the midpoint of the longest segment to form two new simplexes from $\mathbf{S}$. This is illustrated for $(n-1) = 2$ in Fig. 3a and we formally define bisections below.

When bisections are carried out, too many $(n-1)$-simplexes may be produced to be stored effectively in the machine. Furthermore, a casual approach easily results in repeated evaluation of F at same points ([9]). Simplistic approaches may also result in extra work to produce a uniformly fine subdivision of $b(\mathbf{P})$, while only some of the simplexes in the subdivision need have small diameters to assure sufficient refinement. For these reasons we will introduce an address scheme for labeling simplexes in subdivisions of $b(\mathbf{P})$-produced by bisection. This address scheme depends on a one-to one cor-

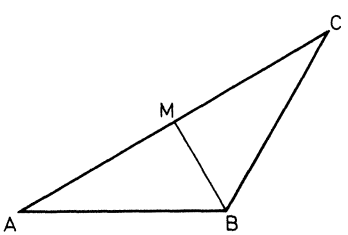

Fig. 3a. A simple illustration of bisection when $n-1=2$

respondence between the simplexes in the subdivisions and the nodes in a binary search tree, where each node is labeled by an ordered pair of integers.

This correspondence between bisection and binary search trees and the address scheme is presented in this section following the formal definition of bisection. The remainder of the section is devoted to the actual presentation of an efficient algorithm to compute $d(F, \mathbf{P}, \theta)$, where $\mathbf{P}$ is a polygon.

Definition. Suppose $\mathbf{S}=\langle X_1, \ldots, X_n\rangle$ is an $(n-1)$-simplex. Then a *simple subdivision* of $\mathbf{S}$ is any ordered pair of simplexes $\{\mathbf{S}_1, \mathbf{S}_2\}$ such that for some k and m between 1 and n and some $A \in \langle X_k, X_m\rangle$:

$$\mathbf{S}_1=\langle X_1, \ldots, X_{k-1}, A, X_{k+1}, \ldots, X_m, \ldots, X_n\rangle, \quad \text{and}$$

$$\mathbf{S}_2=\langle X_1, \ldots, X_k, \ldots, X_{m-1}, A, X_{m+1}, \ldots, X_n\rangle.$$

Remark. It follows from the definition that $\mathbf{S}=\mathbf{S}_1+\mathbf{S}_2$ if $\{\mathbf{S}_1, S_2\}$ is a simple subdivision of $\mathbf{S}$.

Definition. Suppose $\mathscr{S}_1=\{\mathbf{S}\}$, $\mathscr{S}_2, \ldots, \mathscr{S}_\kappa$ are sets of simplexes such that for each simplex $\mathbf{T} \in \mathscr{S}_i$ with $i<\kappa$, either $\mathbf{T} \in \mathscr{S}_{i+1}$ or there is a pair $\{\mathbf{U}, \mathbf{V}\} \subseteq \mathscr{S}_{i+1}$ such that $\{\mathbf{U}, \mathbf{V}\}$ is a simple subdivision of $\mathbf{T}$. Then $\mathscr{S}_i$ will be called a subdivision of $\mathscr{S}_j$ for $i<j \leqq \kappa$. Any process of generating subdivisions of $\mathscr{S}_1$ will also be termed subdivision of $\mathbf{S}$.

Definition. If $b(\mathbf{P})=\sum_{i=1}^{m} \mathbf{S}_i$ is the boundary of an $n-1$ polygon, then subdivisions of $b(\mathbf{P})$ are defined in the natural way as unions of subdivisions of the component simplexes of $b(\mathbf{P})$.

Definition. Let $\mathbf{S}$, k, m, A, and $\{\mathbf{S}_1, \mathbf{S}_2\}$ be as in the definition of simple subdivision. Suppose $\langle X_k, X_m\rangle$ is the longest one-dimensional side to be formed from the vertices of $\mathbf{S}$, where the length of $\langle X_i, X_j\rangle$ is taken to be $\|X_i-X_j\|_2^2$, and suppose $A=(X_k+X_m)/2$. Then the pair $\{\mathbf{S}_1, \mathbf{S}_2\}$ is called the *bisection* of $\mathbf{S}$.

We will henceforth consider only subdivisions of $b(\mathbf{P})$ formed from bisections, and proceed to form the correspondence between such subdivisions and binary search trees. We use special terminology to define a particular kind of binary search tree suitable to our purposes.

Definition. A simplex tree $\mathscr{T}$ is a finite, partially ordered set of points with the following four properties:

(a) Each point of $\mathscr{T}$ precedes no points or is the immediate predecessor of precisely two points.

(b) There is a unique point $\mathbf{S}_0 \in \mathscr{T}$, designated the *original point*, such that no point of $\mathscr{T}$ precedes $\mathbf{S}_0$, and all other points of $\mathscr{T}$ follow $\mathbf{S}_0$.

(c) Each point of $\mathscr{T}$ other than the original point is preceded by precisely 1 point.

(d) If $\mathbf{S}$ is the immediate predecessor of $\mathbf{S}_1$ and $\mathbf{S}_2$, then the pair $\{\mathbf{S}_1, \mathbf{S}_2\}$ is ordered. The first element of the pair will be called the *lower point from* $\mathbf{S}$ and the second element will be called the *upper point from* $\mathbf{S}$.

If $\{\mathscr{S}_i\}_{i=1}^{\kappa}$ is a sequence of sets of simplexes such that $\mathscr{S}_{i+1}$ is a subdivision of

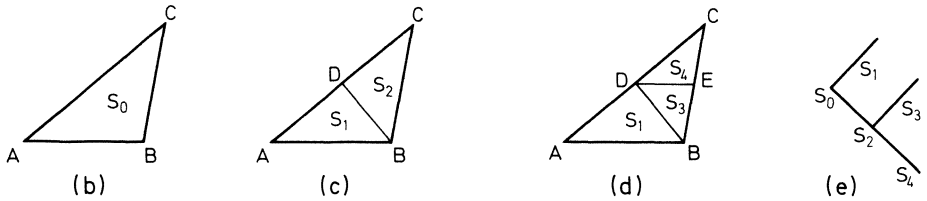

Fig. 3b–e

$\mathscr{S}_i$ for $i<\kappa$, and $\mathscr{S}_1=\{\mathbf{S}_0\}$, we can order the elements of $\bigcup_{i=1}^{\kappa} \mathscr{S}_i$ in a simplex tree as indicated in the following definition.

Definition. Let k, m, A, $\mathbf{S}$, $\mathbf{S}_1$, and $\mathbf{S}_2$ be as in the definition of simple subdivision. We will say $\mathbf{S}<\mathbf{S}_1$ and $\mathbf{S}<\mathbf{S}_2$; we will furthermore distinguish $\mathbf{S}_1$ and $\mathbf{S}_2$ by calling $\mathbf{S}_1$ the *lower simplex* and calling $\mathbf{S}_2$ the *upper simplex* of the subdivision $\{\mathbf{S}_1, \mathbf{S}_2\}$ of $\mathbf{S}$.

Figures 3b, c and 3d show successive subdivisions of $\mathbf{S}_0=\langle A, B, C\rangle$ and Fig. 3e shows the corresponding simplex tree. (The flow of paths in the simplex tree is implicity given from left to right).

If $\mathscr{T}=\{\mathbf{S}_i\}_{i=0}^{q}$ is such a simplex tree, each point $\mathbf{S}_j$ of $\mathscr{T}$ will be labeled with an ordered pair of integers (n_1, n_2) as follows: We let n_2 be the order of the set $\mathscr{B}=\{\mathbf{S}\in\mathscr{T}, \mathbf{S}<\mathbf{S}_j\}$, i.e., n_2 will be the number of points of $\mathscr{T}$ which precede $\mathbf{S}_j$. To define n_1, we observe $\mathscr{B}\cup\{\mathbf{S}_j\}$ consists of all points on the path in $\mathscr{T}$ between the original point and $\mathbf{S}_j$; we re-index the elements of $\mathscr{B}\cup\{\mathbf{S}_j\}$ so that $\mathbf{S}_0$ is the original point in this path, $\mathbf{S}_1$ is the first point, $\mathbf{S}_i$ is the i-th point for $1<i<n_2$, and $\mathbf{S}_{n_2}$ is the point previously called $\mathbf{S}_j$. Then, for $i\in\{1,\ldots,n_2\}$, $\mathbf{S}_i$ is either an upper point or a lower point in $\mathscr{T}$. We form the unique integer n_1 from this information by setting the i-th digit (eg. counting from the right) in the binary expansion of n_1 equal to 1 if $\mathbf{S}_i$ is an upper simplex and otherwise setting the i-th digit of the binary expansion of n_1 equal to 0 (It is assumed that n_1 has only n_2 digits in its binary expansion, thus assuring the uniqueness of n_1).

When we define n_1 and n_2 as above, each point in $\mathscr{T}$ is uniquely labeled by the ordered pair (n_1, n_2). We illustrate the labeling of the tree in Fig. 3e with Fig. 3f.

Definition. The pair (n_1, n_2) will be called the pair of *location numbers* for the corresponding point of $\mathscr{T}$. If $\mathscr{T}$ corresponds to a set of simplexes, we will refer to (n_1, n_2) as the pair of *location numbers* for the corresponding simplex.

With introduction of binary trees and the labeling scheme, we are equipped to present an efficient algorithm to compute $d(F, \mathbf{P}, \theta)$. The following concepts aid the exposition.

Definition. A *bough* of a simplex tree is a maximal linearly-ordered set of the tree.

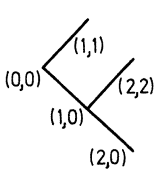

Fig. 3f

Definition. A *leaf* of a simplex tree is any point which preceeds no points in the tree.

Remark. If $\mathscr{S}_1, \ldots, \mathscr{S}_\kappa$ are sets of simplexes such that $\mathscr{S}_1 = \{\mathbf{S}\}$ and $\mathscr{S}_j$ is a subdivision of $\mathscr{S}_i$ if $j > i$, then there is a simplex tree corresponding to $\bigcup_{i=1}^{\kappa} \mathscr{S}_i$. The leaves of this tree correspond to elements of $\mathscr{S}_\kappa$.

If $b(\mathbf{P}) = \sum_{i=1}^{m} \mathbf{S}_i$, we will subdivide each $\mathbf{S}_i$ one at a time to get $\mathbf{S}_i = \sum_{j=1}^{\kappa_i} \mathbf{S}_{i,j}$, so that $\sum_{j=1}^{\kappa_i} \mathrm{Par}(\mathscr{R}(\mathbf{S}_{i,j}) F))$ is the *contribution* of $\mathbf{S}_i$ to the sum for $d(F, \mathbf{P}, \theta)$ in the Parity Theorem. The algorithm will follow one bough of the simplex tree at a time, proceeding until all leaves $\mathbf{S}_{i,j}$, have been considered.

An iteration of the procedure to follow a given bough consists of: (1) producing a bisection $\{\mathbf{S}_1, \mathbf{S}_2\}$; (2) discarding of one of the elements of the bisection and keeping of the other, e.g. $\mathbf{S}_j$ ($j = 1$ or 2) for subsequent iterations; (3) Calculation of $\mathscr{R}(\mathbf{S}_j, F)$; and (4) Comparison of $\mathscr{R}(\mathbf{S}_j, F)$ with the range simplex of the previous iteration.

The comparison of range simplexes is carried out to determine whether more iterations need be performed to assure sufficient refinement; the comparison includes a test for sufficient refinement so that iteration is interrupted if the test is positive. $\mathrm{Par}(\mathscr{R}(\mathbf{S}_j, F))$ is determined in the course of the stopping test, provided the test is positive.

We outline one possible stopping test: We specify a priori a parameter p. We then observe that the range simplex in a given iteration of the algorithm differs from the range simplex of the previous iteration only in a single row, which is designated the *new row*. If the new row is identical to a row of the previous range simplex, we will say there is an *agreement*. The result of the stopping test will be positive iff there has been an agreement in p consecutive iterations.

If the stopping test is positive, we store $\mathrm{Par}(\mathscr{R}(\mathbf{S}, F))$ in an element of an array D. Iteration then starts again and a different bough is followed. The whole process repeats itself until all boughs (and hence, all leaves) have been checked. The elements of D are added together to get the contribution $\sum_{j=1}^{\kappa_i} \mathrm{Par}(\mathscr{R}(\mathbf{S}_{i,j} F))$.

The boughs are checked sequentially by examination and adjustment of the parameters n_1' and n_2. Between each sequence of iterations of the bisection-taking process, n_1' and n_2 are adjusted to determine what bough will be followed next. During the bisection-taking process, n_1', and n_2 are examined to determine which element of each bisection is to be retained; if $\mathbf{S}_j$ is the element to be retained, then the location numbers of $\mathbf{S}_j$ in the simplex tree are (n_1', n_2).

An array of simplexes SA and an array of range simplexes RA are stored to enable us to avoid computational redundancy. These arrays contain only simplexes and range simplexes from the previous bough considered, and have relatively few elements. Some additional information may also be stored for the stopping test.

An outline of the algorithm is presented below. Further details can be found in ([9]) and future publications. These details include stopping tests, calculation of $\mathscr{R}(\mathbf{S}, F)$, modifications, etc. (Calculations of contributions of $d(F, \mathbf{P}, \theta)$.)

4.1. Algorithm

(1) Read in the boundary simplex to be subdivided and store in in S_1. Also specify a stopping parameter p.

(2) $n_1' \leftarrow 0$. $n_2 \leftarrow 0$, $k \leftarrow 1$, $q \leftarrow 0$

(3) $R_2 \leftarrow \mathscr{R}(S_1, F)$

(4) $RA_1 \leftarrow R_2$, $SA_1 \leftarrow S_1$

(5) $n_2 \leftarrow n_2 + 1$

(6) Find the bisection of S_1, storing the lower simplex back in S_1 and storing the upper simplex in both S_2 and SA_{n_2+1}.

(7) $RA_{n_2+1} \leftarrow R(S_2, F)$,

(8) Examine the n_2-th binary digit of n_1'. If the digit equals 1, $S_1 \leftarrow S_2$.

(9) $R_1 \leftarrow R_2$

(10) $R_2 \leftarrow \mathscr{R}(S_1, F)$.

(11) (Comparison of Range Simplexes). Compare the new row of $\mathscr{R}(\mathbf{S}, F)$ to the rows of the previous $\mathscr{R}(\mathbf{S}, F)$. If each entry of the new row equals the corresponding entry of one of the previous rows, then $q \leftarrow q+1$; otherwise, $q \leftarrow 0$. *Also*, $QA_{n_2+1} \leftarrow q$.

(12) Perform the stopping test ($q \geqq p$?). If negative, return to step (5), and proceed if positive.

(13) $D_k \leftarrow \mathrm{Par}(R_2)$

(14) $k \leftarrow k+1$

(15) $n_2 \leftarrow n_2 - 1$

(16) If the n_2+1st digit of n_1' does not equal 1, go to step (18)

(17) If $n_2 = 0$, find $\sum_{j=1}^{k-1} D_j$ and stop; otherwise, return to step (15).

(18) Set the n_2+1st digit of n_1' equal to 1 and set all subsequent digits of n_1' equal to 0.

(19) $R_2 \leftarrow RA_{n_2+2}$, $S_1 \leftarrow SA_{n_2+2}$, $q \leftarrow QA_{n_2+2}$, $n_2 \leftarrow n_2+1$

(20) Return to step (12).

A Fortran program for Algorithm 4.1 was implemented, and some test results appear in Sect. 6 below.

5. A Generalized Method of Bisection in R^n

Algorithm 4.1 can easily be modified for inclusion in a root-finding algorithm for determining $X \in \mathbf{P}$ such that $F(X) = \theta$, where $F: \mathbf{P} \to R^n$ and $\mathbf{P}$ is an n-polygon.

The root-finding algorithm is based on subdivision of each n-simplex of $\mathbf{P}$ into n-simplexes with smaller diameters. It is determined in which of these smaller simplexes a root lies. The subdivision process is then repeated with the smaller simplex.

The method of subdivision of the n-simplexes is analogous to production of bisections of $(n-1)$-simplexes. If $\mathbf{S}$ is one of the n-simplexes, the existence of a root $X \in \mathbf{S}$ of $F(X) = \theta$ is ascertained by computation of $d(F, \mathbf{S}, \theta)$. If $d(F, \mathbf{S}, \theta) \neq 0$ for one of the original n-simplexes comprising $\mathbf{P}$, then we are assured ([4] p. 32) that the degree relative to one of the smaller simplexes is also non-zero.

Computation of the degrees relative to simplexes in bisections of an original n-simplex usually requires $1/(n+1)$ times the number of computations needed to

calculate the degree relative to an original n-simplex. This is due to relationships between the elements of the bisection of an n-simplex $\mathbf{S}$ and the elements of the bisections of the $n+1$ $(n-1)$-simplexes comprising $b(\mathbf{S})$.

We begin by defining bisections of n-simplexes and delineating the relationship between such and bisections of $(n-1)$-simplexes.

Definition. Let $\mathbf{S}=\langle X_0, X_1, \ldots, X_n\rangle$ be an n-simplex in R^n, let

$$\|X_k - X_m\|_2 = \max_{0 \leqq i < j \leqq n} \|X_j - X_i\|_2,$$

and set $A=(X_k+X_m)/2$. Then if we set:

$$\mathbf{S}_1 = \langle X_0, \ldots, X_{k-1}, A, X_{k+1}, \ldots, X_m, \ldots, X_n\rangle$$

and set:

$$\mathbf{S}_2 = \langle X_0, \ldots, X_k, \ldots, X_{m-1}, A, X_{m+1}, \ldots, X_n\rangle$$

we have $\mathbf{S}=\mathbf{S}_1+\mathbf{S}_2$. We will say that the ordered pair $(\mathbf{S}_1, \mathbf{S}_2)$ is the bisection of the n-simplex $\mathbf{S}$. We say $\mathbf{S}_1$ is the lower n-simplex for $\mathbf{S}$, and $\mathbf{S}_2$ is the upper n-simplex for $\mathbf{S}$.

5.1. Lemma. Suppose $\mathbf{S}, X_k, X_m, A, \mathbf{S}_1$, and $\mathbf{S}_2$ are as in the definition of bisection of n-simplexes. Then if $i \neq k$ and $i \neq m$, the i-th facet of $\mathbf{S}_1$ is the lower simplex for the i-th facet of $\mathbf{S}$; likewise, the i-th facet of $\mathbf{S}_2$ is the upper simplex for the i-th facet of $\mathbf{S}$.

5.2. Lemma. Let $\mathbf{S}_1$, X_k, X_m, $\mathbf{S}_1$, and $\mathbf{S}_2$ be as in the definition of bisection of n-simplexes. Then the *k-th facet of* $\mathbf{S}_1$ is equal to the k-th facet of $\mathbf{S}$, while the m-th facet of $\mathbf{S}_1$ is equal to the m-th facet of $\mathbf{S}$. The m-th facet of $\mathbf{S}_1$ and the k-th facet of $\mathbf{S}_2$ have no interior points in common with any facets of $\mathbf{S}$ but the m-th facet of $\mathbf{S}_1$ equals $-(-1)^{m-k}$ times the k-th facet of $\mathbf{S}_2$. Hence if $\mathbf{U}_1^k$ is $(-1)^k$ times the k-th facet of $\mathbf{S}_1$ and if $\mathbf{U}_2^m$ is $(-1)^m$ times the m-th facet of $\mathbf{S}_2$ then $\mathbf{U}_1^k = -\mathbf{U}_2^m$.

The proofs of Lemma 5.1 and Lemma 5.2 are straightforward and appear in [9].

Figure 5.1 illustrates Lemma 5.1 and Lemma 5.2. There, $n=3$, $k=1$, $m=3$. $\mathbf{S}_1 = \langle X_0, A, X_2, X_3\rangle$ lies to the right and rear, and $\mathbf{S}_2 = \langle X_0, X_1, X_2, A\rangle$ lies to the left and front. The k-th facet of $\mathbf{S}_1$ is $\langle X_0, X_2, X_3\rangle$, which equals the k-th facet of $\mathbf{S}$. The m-th facet of $\mathbf{S}_2$ is $\langle X_0, X_1, X_2\rangle$, which equals the m-th facet of $\mathbf{S}$. We see in Fig. 5.1 that $\langle X_0, A, X_2\rangle$ is both the m-th facet of $\mathbf{S}_1$ and the k-th facet of $\mathbf{S}_2$. We also see how the bisection of $\mathbf{S}$ induces bisections of the 0-th and second facets of $\mathbf{S}$.

The root-finding algorithm begins by calculating $d(F, \mathbf{S}, \theta)$, where $\mathbf{S}$ is a given n-simplex. If $d(F, \mathbf{S}, \theta) \neq 0$, then $d(F, \mathbf{S}_1, \theta) \neq 0$ or $d(F, \mathbf{S}_2, \theta) \neq 0$, where $\{\mathbf{S}_1, \mathbf{S}_2\}$ is the bisection of $\mathbf{S}$.

The algorithm then iterates by: (1) taking the bisection of $\mathbf{S}$; (2) selecting one of the elements $\mathbf{S}_i$ of the bisection for which $d(F, \mathbf{S}_i, \theta) \neq 0$; (3) repeating step (1) with the element selected from step (2). We take advantage of the relationships presented in Lemma 5.1 and Lemma 5.2 to simplify computation of $d(F, \mathbf{S}_i, \theta)$ if $\mathbf{S}_i$ is an element of a bisection. This is done by storing and retrieving location numbers and "contributions" for simplexes in the trees for the facets of $\mathbf{S}$. The "contribution" of a simplex $\mathbf{S}$ in the tree will be $\sum_{i=1}^{m_{\mathbf{S}}} \mathrm{Par}(R(\mathbf{S}_{j,i}, F))$, *where* $\{\mathbf{S}_{j,i}\}_{i=1}^{m_{\mathbf{S}}}$ is the set of leaves following $\mathbf{S}$. (These contributions are calculated in the course of performing steps

(12)–(18) of Algorithm 4.1; see [9]). The algorithm is designed so that $F(X)$ is never evaluated twice for the same X.

The root-finding algorithm is outlined below.

5.3. Algorithm. (A Generalized Method of Bisection)

(1) Determine a stopping diameter E.

(2) Read in the original n-simplex and store in $\mathbf{S}_1$.

(3) Calculate $d(F,\mathbf{S}_1,\theta)$ via Algorithm 4.1; return to step (2) if $d(F,\mathbf{S}_1,\theta)=0$.

(4) Calculate the bisection of $\mathbf{S}_1$, storing the lower simplex back in $\mathbf{S}_1$ and the upper simplex in $\mathbf{S}_2$.

(5) Determine $d(F,\mathbf{S}_1,\theta)$ and $d(F,\mathbf{S}_2,\theta)$ by retrieving appropriate stored information about the i-th facets of $\mathbf{S}_1$ and $\mathbf{S}_2$ ($i\neq k$ or $i\neq m$) and applying Algorithm 4.1 to the new facet (cf. Lemma 5.2).

(6) If $d(F,\mathbf{S}_1,\theta)=0$ then: $\mathbf{S}_1\leftarrow\mathbf{S}_2$.

(7) If the diameter of $\mathbf{S}_1$ is less than E, then print $\mathbf{S}_1$ and stop.

(8) Return to step (5).

Several problems can arise in step (5). For example, more bisections of n-simplexes may be taken then were taken for corresponding $(n-1)$-simplexes in the course of Algorithm 4.1. These problems are discussed in [9].

The method seems particularly suited to serve as a starting method when the function F is not smooth, when the original n-simplex has a very large diameter, or when it is difficult to evaluate the components of F accurately ([9]).

The method has features in common with the class of fixed point algorithms developed by Kuhn, Scarf, Eaves, Saigal, Allgower, Keller, Jeppson, etc. These methods involve labelings similar to the one given for an alternate definition of useable $\mathscr{R}(\mathbf{S},F)$.

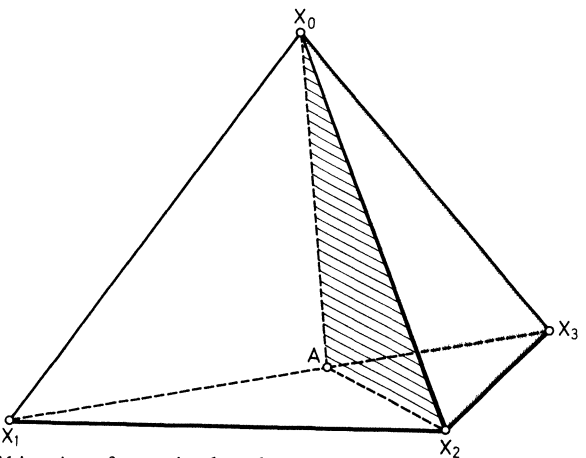

Fig. 4. Illustration of bisection of an n-simplex when $n=3$

6. Some Numerical Examples

In [9], some computational examples for $n=2$, $n=3$, and $n=5$ were presented; the algorithm was similar to Algorithm 4.1, except that the stopping test (step (11)) was more involved. Here, we give some additional examples.

The algorithm was programmed in Fortran on a Multics 68/80 system. The actual code differed from Algorithm 4.1 only in the manner of storage and retrieval of the simplexes and range simplexes. (It is only necessary to store the midpoints produced in bisection, instead of entire simplexes, in steps (4), (7), and (20).) The program was written so that there were no redundant function evaluations.

In two dimensions, we tried the analytic functions $F(z)=z^2$, $F(z)=z^3$, and $F(z)=z^4$, and we arbitrarily chose the simplex: $\langle(-4,-4),(4,-4),(0,4)\rangle$. The computed value of the degree, the maximum depth in the trees produced, and the total number of function evaluations are summarized in Tables 1(a), 1(b), and 1(c). We note that the algorithm behaved predictably and reliably.

In addition, several other simplexes not containing the root (0, 0) were tried. In all of those cases, the correct degree $d(F,\mathbf{S},\theta)=0$ was computed with a small number of function evaluations.

The program both required more function evaluations and was less reliable in higher dimensions, although it worked satisfactorily for $n=3$.

Four distinct functions were tried for $n=3$. The results for the identity function $F(x_1,x_2,x_3)=(x_1,x_2,x_3)$ and the function $F(x_1,x_2,x_3)=(x_1^5,x_2^5,x_3^5)$ were identical; these results are summarized in Table 2(a). The results for

$$F(x_1,x_2,x_3)=(x_1^2-x_2,x_2^2-x_3,x_3^2-x_1)$$

are summarized in Table 2(b).

Table 1. Experiments in two dimensions. $S=\langle(-4,-4),(4,-4),(0,4)\rangle$. "M.D." indicates maximum tree depth, "N.E." indicates the number of function evaluations, and "v" indicates the computed value for $d(F,S,\theta)$

p	M.D.	N.E.	v	p	M.D.	N.E.	v	p	M.D.	N.E.	v
1	3	12	2	1	4	18	1	1	4	12	1
2	5	34	2	2	5	54	3	2	6	62	4
3	6	68	2	3	6	108	3	3	7	124	4
4	7	136	2	4	7	216	3	4	8	248	4
5	8	272	2	5	8	432	3	5	9	496	4
6	9	544	2	6	9	864	3	6	10	992	4
(a) $F(z)=z^2$				(b) $F(z)=z^3$				(c) $F(z)=z^4$			

Table 2. Examples for $n=3$. $S=\langle(1,0,0),(0,1,0),(0,0,1),(-1,-1,-1)\rangle$. Note S has a negative orientation

p	M.D.	N.E.	v	p	M.D.	N.E.	v
1	4	32	−1	1	4	16	1
2	7	82	−1	2	7	38	1
3	9	206	−1	3	9	146	1
4	10	412	−1	4	11	412	1
5	11	824	−1	5	12	824	1
6	15	2,026	−1	6	14	2,152	1
(a) $F=(x_1,x_2,x_3)$				(b) $F=(x_1^2-x_2,x_2^2-x_3,x_3^2-x_1)$			

Table 3. Experiments for $F=(x_1^2-x_2^2, 2x_1x_2, x_3)$

S	p	M.D.	N.E.	v
$(1,0,0),(0,1,0),(0,0,1),(-1,-1,-1)$	1	5	18	0
$(1,0,0),(0,1,0),(0,0,1),(-1,-1,-1)$	2	131	1,022	−1
$(1.01,0.02,0.03),(0.04,1.05,0.06),(0.07,0.08,1.09),$ $(-1,-1,-1)$	2	11	182	−2
$(1.01,0.02,0.03),(0.04,1.05,0.06),(0.07,0.08,1.09),$ $(-1,-1,-1)$	3	16	558	−2
$(1.01,0.02,0.03),(0.04,1.05,0.06),(0.07,0.08,1.09),$ $(-1,-1,-1)$	4	–	–	–

Results for the fourth function $F(x_1,x_2,x_3)=(x_1^2-x_2^2, 2x_1x_2, x_3)$ are summarized in Table 3. For this F, the behaviour of the algorithm is sensitive to perturbations in the vertices of **S**, due to roots of components of F near the boundary of **S**. For $p=4$, exponent underflow in the function evaluations invalidated the results.

Functions and simplexes analogous to the above were tried for $n=4$, $n=5$, and $n=6$. Correct results were consistently given for $F(X)=(x_1^2-x_2,\ x_2^2-x_3,\ x_3^2-x_4,\ldots)$, but perturbations of the vertices of S needed to be tried with the other functions. In the cases that no roots of F lay within S, however, Algorithm 4.1 gave correct results for all values of p.

7. Summary and Assessment

The recursion formula, the parity theorem, and the characterization in terms of determinants [13] are in many cases useful in hand computations. Also, these formulae show promise, as they require only the signs to be correct in the function evaluations, and they hold for non-differentiable functions.

Automatic computation of the degree and use in root-finding at present seems to be intractible in large dimensions, and experiments have shown that Algorithm 4.1 has some other short-comings. However, Algorithm 4.1 seems to compare favorably to other methods ([5, 10, 13]) in efficiency; additionally, previous computations may be used in a generalized method of bisection, without redundancy. With an appropriate choice of stopping test (by modifying steps (11) and (12), Algorithm 4.1), and, prehaps, with schemes for automatically perturbing the vertices of **S**, Algorithm 4.1 is potentially very useful.

References

1. Alexandroff, P., Hopf, H.: Topologie, Chelsea, N.Y. 1935
2. Allgower, E.L., Jeppson, M.: The approximation of solutions of nonlinear elliptic boundary value problems having several solutions, Springer lecture notes **333**, 1–20 (1973)
3. Allgower, E.L., Keller, K.L.: A search routine for a sperner simplex, Computing **8**, 157–165 (1971)

4. Cronin, J.: Fixed points and topological degree in nonlinear analysis. Amer. Math. Soc. Surveys II, 1964
5. Erdelsky, P.J.: Computing the Brouwer degree in R^2, Math. Comp. **22**, 133–137, 1973
6. Greenberg, M.: Lectures on algebraic topology, W.A. Benjamin, N.Y. 1967
7. Hadamard, J.: Sur quelques applications de l'indice de Kronecker, Herman, Paris 1910
8. Jeppson, M.: A search for fixed points of a continuous mapping, Mathematical topics in economic theory and computation, 122–128, SIAM, Philadelphia 1972
9. Kearfott, R.B.: Computing the degree of maps and a generalized method of bisection, Ph.D. dissertation, University of Utah, S.L.C. 1977
10. O'Neil, T., Thomas, J.: The calculation of the topological degree by quadrature, SIAM J. Numer. Anal. **12**, 673–680 (1975)
11. Rheinboldt, W.C., Ortega, S.M.: Iterative solution of nonlinear equations in several variables, N.Y.: Academic Press 1970
12. Scarf, H.: The approximation of fixed points of a coninuous mapping, SIAM J. Appl. Math. **15**, 1328–1343 (1967)
13. Stenger, F.: Computing the topological degree of a mapping in R^2, Numer. Math. **25**, 23–38 (1975)
14. Stynes, M.: Ph.D. dissertation, Univ. of Oregon, Corvallis 1977

Received April 29, 1977/September 28, 1978