MATHEMATICS OF COMPUTATION, VOLUME 33, NUMBER 146
APRIL 1979, PAGES 722–738

# A Three-Dimensional Analogue to the Method of Bisections for Solving Nonlinear Equations

By Krzysztof Sikorski

Abstract. This paper deals with a three-dimensional analogue to the method of bisections for solving a nonlinear system of equations $F(X) = \theta = (0, 0, 0)^T$, which does not require the evaluation of derivatives of $F$.

We divide the original parallelepiped (Figure 2.1) into 8 tetrahedra (Figure 2.2), and then bisect the tetrahedra to form an infinite sequence of tetrahedra, whose vertices converge to $Z \in R^3$ such that $F(Z) = \theta$. The process of bisecting a tetrahedron $\lhd\!\rhd\, E_1E_2E_3E_4$ with vertices $E_i$ is defined as follows. We first locate the longest edge $E_iE_j$, $i \neq j$, set $D = (E_i + E_j)/2$, and then define two new tetrahedra $\lhd\!\rhd\, E_iDE_kE_l$ and $\lhd\!\rhd\, DE_jE_kE_l$, where $j \neq l$, $l \neq i$, $i \neq k$, $k \neq j$ and $k \neq l$.

We give sufficient conditions for convergence of the algorithm. The results of our numerical experiments show that the required storage may be large in some cases.

**1. Introduction.** It is of interest to find a globally convergent numerical method for the solution of a nonlinear system of equations

$$(1.1) \qquad F(X) = \theta, \quad \text{where } F\colon D \subset R^n \longrightarrow R^n \text{ and } \theta = (0, \ldots, 0)^T \in R^n,$$

which does not require the evaluation of derivatives of $F$. In the scalar case $n = 1$ the method of bisections satisfies the above requirements. In this paper we deal with a three-dimensional analogue of the bisection method.

The main properties of the presented method are as follows:

(1) global convergence (i.e. starting polyhedron $P$ can be arbitrary large, see Section 2);

(2) the method requires only the evaluation of $F$;

(3) linear convergence of the method with a constant $q$ which is independent of $F$.

However, we must admit that the required storage may be large in some cases. Moreover, we will assume, that the topological degree $\deg(F, \operatorname{Int} P, \theta) \neq 0$. Then by Kronecker's theorem [2] the equation (1.1) has at least one solution in $\operatorname{Int} P$. The topological degree $\deg(F, \operatorname{Int} P, \theta)$ can be evaluated by means of the algorithm described in [3].

The method is a generalization of the two-dimensional bisection algorithm proposed by Harvey and Stenger; see [1]. We assume that the reader is more or less familiar with the work of Harvey and Stenger [1], since our ideas and proofs are similar to those presented in [1].

Received January 4, 1978.

*AMS (MOS) subject classifications* (1970). Primary 65H10; Secondary 50B30.

*Key words and phrases.* Bisection.

**2. Method of Bisection of Tetrahedra.** Let the points $E_i$, $i = 1, 2, 3, 4$, where $E_i = E_{i+4}$ be four noncollinear points in $R^3$. A tetrahedron $\triangleleft\!\triangleright\, E_1E_2E_3E_4$, where

$$\triangleleft\!\triangleright\, E_1E_2E_3E_4 = \left\{ X \in R^3 \colon X = \sum_{i=1}^{4} \lambda_i E_i,\ \lambda_i \geqslant 0,\ \sum_{i=1}^{4} \lambda_i = 1 \right\}$$

is bisected into two tetrahedra as follows. Find the longest edge $E_iE_j$, $i \neq j$. Next, set $D = (E_i + E_j)/2$ and define two new tetrahedra $\triangleleft\!\triangleright E_iDE_kE_l$ and $\triangleleft\!\triangleright\, DE_jE_kE_l$, where $j \neq l,\ l \neq i,\ i \neq k,\ k \neq j$ and $k \neq l$.

We now describe a simple test for determining whether or not the point $\theta$ belongs to $\triangleleft\!\triangleright\, E_1E_2E_3E_4$.

If $A = (a_1\ a_2\ a_3)^T$, $B = (b_1\ b_2\ b_3)^T$ and $C = (c_1\ c_2\ c_3)^T$ are three noncollinear points in the space, then define a linear form

$$\begin{aligned} L(A, B, C; X) = {} & (b_1 - a_1)(c_2 - a_2)(x_3 - a_3) + (b_2 - a_2)(c_3 - a_3)(x_1 - a_1) \\ & + (c_1 - a_1)(b_3 - a_3)(x_2 - a_2) - (b_3 - a_3)(c_2 - a_2)(x_1 - a_1) \\ & - (b_2 - a_2)(c_1 - a_1)(x_3 - a_3) - (c_3 - a_3)(b_1 - a_1)(x_2 - a_2), \end{aligned}$$

where $X = (x_1\ x_2\ x_3)^T$.

Let $L_{ABC} = \{X \in R^3 : L(A, B, C; X) = 0\}$ denote the plane containing the points $A$, $B$ and $C$. The plane $L_{ABC}$ divides the space into two regions $R_1$ and $R_2$ where

$$R_1 = \{X \in R^3 \colon L(A, B, C; X) \geqslant 0\}, \qquad R_2 = \{X \in R^3 : L(A, B, C; X) \leqslant 0\}.$$

Thus, the tetrahedron $\triangleleft\!\triangleright\, E_1E_2E_3E_4$ is given by

$$\triangleleft\!\triangleright\, E_1E_2E_3E_4 = \bigcap_{i=1}^{4} \{X \in R^3 \colon L(E_i, E_{i+1}, E_{i+2}; E_{i+3}) L\,(E_i, E_{i+1}, E_{i+2}; X) \geqslant 0\}.$$

The point $\theta$ belongs to $\triangleleft\!\triangleright\, E_1E_2E_3E_4$ if and only if

$$(2.1) \qquad L(E_i, E_{i+1}, E_{i+2}; E_{i+3})\, L(E_i, E_{i+1}, E_{i+2}; \theta) \geqslant 0 \quad \text{for } i = 1, 2, 3, 4.$$

It is easy to observe that

$$(2.2) \qquad L(A, B, C; X) = -\,L(X, A, B; C) = L(C, X, A; B) = -L(B, C, X; A),$$

$$(2.3) \quad L(A, B, C; \theta) - L(B, C, X; \theta) + L(C, X, A; \theta) - L(X, A, B; \theta) = L(A, B, C; X).$$

These equations simplify the verification whether or not $\theta$ belongs to $\triangleleft\!\triangleright\, E_1E_2E_3E_4$.

Finally, let us set $T_F \triangleleft\!\triangleright\, E_1E_2E_3E_4 = \triangleleft\!\triangleright\, F(E_1)F(E_2)F(E_3)F(E_4)$. Note that all these concepts can be naturally generalized to the case $n > 3$.

*Forming a Starting Parallelepiped.* A starting parallelepiped is similar to the rectanglar parallelepiped shown in Figure 2.1.

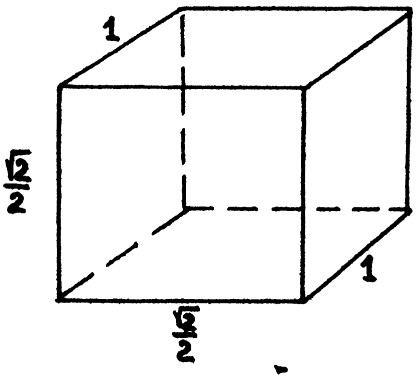


FIGURE 2.1

We divide this polyhedron into 8 tetrahedra as in Figure 2.2.

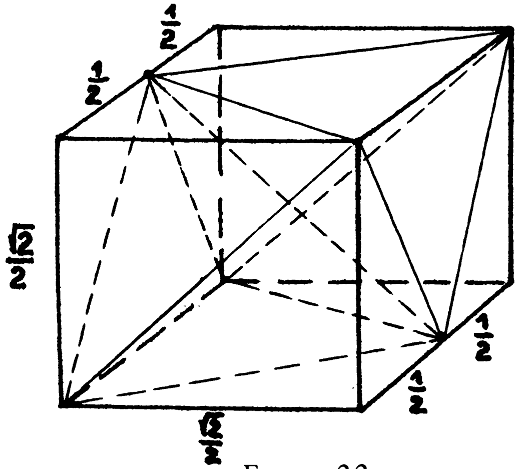
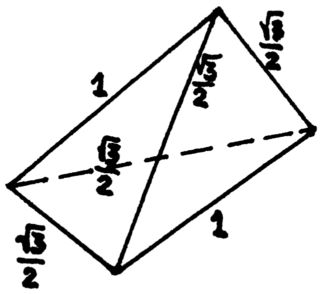


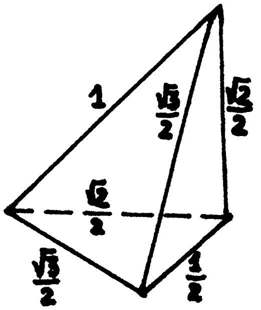
FIGURE 2.2

There are four tetrahedra congruent to the tetrahedron A and four congruent or plane symmetrical to the tetrahedron B.

A.



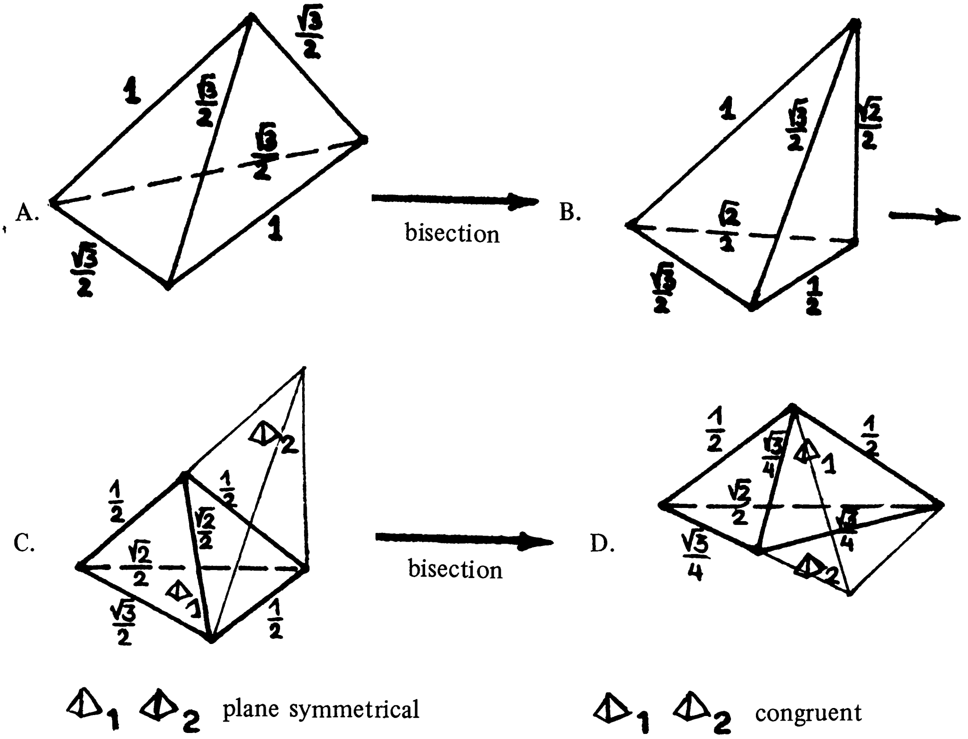
B.



It is evident that the tetrahedron B can be obtained by bisecting the tetrahedron A.

Let $S$ be a family of tetrahedra with the property that each member of the family is either similar to a particular member of that family, or else it is plane symmetrical to a member of the family.

We shall see that bisecting of the tetrahedron A yields only four distinct families $S_i$; $i = 1, 2, 3, 4$. From this we can easily compute the minimal angle of the considered tetrahedra. The bisection of the tetrahedron A yields the following four tetrahedra:



In the sequel the bisection of D yields two tetrahedra belonging to the family of B; see Figure 2.3.

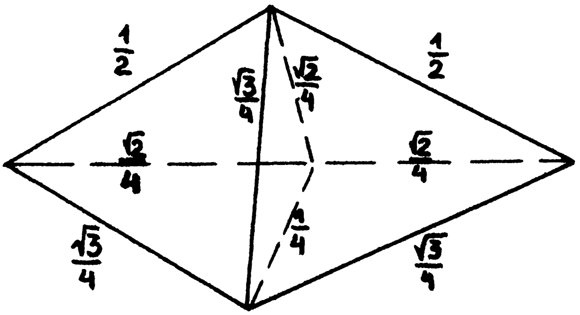


FIGURE 2.3

By simple evaluations we have

(2.4) The minimal angle between edges of the tetrahedra A, B, C and D is equal to $\gamma_1 \approx 35°16'$, $\sin \gamma_1 = \sqrt{3}/3$.

(2.5) The minimal angle between the edges and faces of the tetrahedra A, B, C and D is equal to $\gamma_2 = 30°$.

(2.6) The angles between faces whose intersection is the longest edge in A, B, C and D are equal to 90°, 45°, 60° and 90°.

(2.7) The angles between faces in A, B, C and D are equal to 45°, 60°, 90°, or 120°.

(2.8) *Bisection Algorithm.* In this section we describe an algorithm which is a generalization of the bisection method to the three-dimensional case.

```
(1) Divide $P$ into tetrahedra $\lhd\!\rhd_i$, $i = 1, \ldots, 8$; (see Figure 2.2)
    $M := 8$;
(2) $I := 1$;
(3) Is $I \leqslant M$
    (No) Go to (7)
    (Yes) $\theta \in T_F \lhd\!\rhd_I$?
        (No) $I := I + 1$; Go to (3)
        (Yes) Go to (5)
(4) Bisect $\lhd\!\rhd_I$ into $\lhd\!\rhd_1$ and $\lhd\!\rhd_2$
      $\lhd\!\rhd_I := \lhd\!\rhd_1$;
    for $J := M$ step $-1$ until $I + 1$ do
      $\lhd\!\rhd_{J+1} := \lhd\!\rhd_J$;
      $\lhd\!\rhd_{I+1} := \lhd\!\rhd_2$;
    $M := M + 1$;
    $\theta \in T_F \lhd\!\rhd_I$?
    (Yes) Go to (5)
    (No) $I := I + 1$;
    $\theta \in T_F \lhd\!\rhd_I$?
    (Yes) Go to (5)
    (No) Go to (6)
(5) $g_i$ := the length of the longest edge of $\lhd\!\rhd_I$.
    Is $g_i \leqslant \epsilon$?
    (Yes) Print $g_i$, the vertices of $\lhd\!\rhd_I$ and (STOP)
    (No) Go to (4).
(6) Let $\lhd\!\rhd_I = \lhd\!\rhd A_1A_2E_1E_2$, where $A_1A_2$ is the longest edge of $\lhd\!\rhd_I$
    (Figure 2.4) and let $D = (A_1 + A_2)/2$.
```

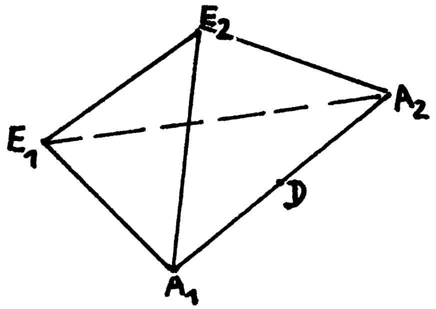


FIGURE 2.4

```
    Compute the angle $\alpha$ between the planes of $A_1A_2E_1$ and $A_1A_2E_2$.
    $N := 2\pi/\alpha$; From Eq. (2.6), $2\pi/\alpha$ is equal to 4, 6 or 8.
    $K := 2$;
```

```
(6.1) Compute $E_{K+1}$ as the symmetrical point to $E_{K-1}$ with respect to the
      plane of $A_1A_2E_K$.
     Is $E_{K+1}$ in $P$?
     (No) Go to (7)
     (Yes) $\Diamond_1 := \Diamond A_1E_{K+1}DE_K$
           $\Diamond_2 := \Diamond DE_{K+1}A_2E_K$
     for $J := M$ step $-1$ until $I + 1$ do
        $\Diamond_{J+2} := \Diamond_J$;
        $\Diamond_{I+1} := \Diamond_1$;
        $\Diamond_{I+2} := \Diamond_2$;
     $M := M + 2$;
     $I := I + 1$;
     $\theta \in T_F \Diamond_I$?
     (Yes) Go to (5)
     (No) $I := I + 1$;
     $\theta \in T_F \Diamond_I$?
     (Yes) Go to (5)
     (No) $K := K + 1$;
     $K \leqslant N$?
     (Yes)  Go to (6.1)
     (No) Go to (7)
   (7) $I := M$;
 (7.1) Bisect $\Diamond_I$ into $\Diamond_1$ and $\Diamond_2$
        $\Diamond_{2I} := \Diamond_1$;
        $\Diamond_{2I-1} := \Diamond_2$;
     $I := I - 1$;
     $I \geqslant 1$?
     (Yes) Go to (7.1)
     (No) $M := 2M$;
     Go to (2)
```

**3. Convergence.** In this section we give sufficient conditions for convergence of the algorithm (2.8).

3.1. *Assumptions.*

1. $F = (f_1\, f_2\, f_3)^T$, $F: D \subset R^3 \longrightarrow R^3$, $F \in C^2(D)$, where $P \subset D$ and $D$ is an open set;

2. $j = \min_{X \in P} |\det F'(X)| > 0$;

3. $d = \min_{X \in \partial P} \|F(X)\|_2 > 0$;

4. $\deg(F, \operatorname{Int} P, \theta) \neq 0$;

5. $h < h_5 = \min(1, h_1, h_2, h_3, \sqrt{d/(3r)}, d/(2p))$, where $h$ is the longest edge of all tetrahedra in $P$, where

$$h_1 = 1/(rH_1H_2^2),$$

$$h_2 = 1/r\{[(E + 2p)^2 + 2j/M \sin \gamma_1]^{1/2} - (E + 2p)\},$$

$$h_3 = 1/r\{[(½JH_2 + EH_1)^2 + 2j \sin \gamma_1 \sin \gamma_2 (E + H_1 + r/2)]^{1/2}$$

$$- (½JH_2 + EH_1)\}\, 1/(E + H_1 + r/2),$$

and where

$$E = \max_i e_i, \qquad e_i = \max_P \left( \sum_{k=1}^{3} (f^2_{i+1,x_k} + f^2_{i+2,x_k}) \right)^{1/2}, \quad i = 1, 2, 3,$$

where $f_{i+3} = f_i$;

$$p = \max_{X \in P} \|F'(X)\|_E, \quad \text{where } \|F'(X)\|_E = \left( \sum_{i,j=1}^{3} (f_{i,x_j}(X))^2 \right)^{1/2};$$

$$M = \max_{X \in P} \|F'(X)\|_\infty; \qquad J = \max_{X \in P} |\det F'(X)|;$$

$$H_1 = \max_{X \in P} \|F'(X)\|_2; \qquad H_2 = \max_{X \in P} \|F'(X)^{-1}\|_2;$$

$$r = \left( \sum_{l=1}^{3} q_l^2 \right)^{1/2}, \quad \text{where } q_l = \max_P \left( \sum_{i,j=1}^{3} \left( \frac{\partial^2 f_1}{\partial x_i \partial x_j} \right)^2 \right)^{1/2}$$

and $\gamma_1$ and $\gamma_2$ are the smallest positive angles defined by $\sin \gamma_1 = \sqrt{3}/3$ and $\sin \gamma_2 = ½$ (see (2.4) and (2.5)).

6. Let $0 < \epsilon < h_5$, where $\epsilon$ appears in step (5) of the algorithm.

Note that all numbers $E$, $p$, $M$, $J$, $H_1$, $H_2$ and $r$ are finite since the considered functions are continuous and $P$ is compact.

THEOREM 3.2. *Suppose that all assumptions* 3.1 *except the fifth hold. Then the algorithm finds the points* $A_i$, $i = 1, 2, 3, 4$ *such that* $\|A_i - Z\|_2 \leqslant 2\epsilon$, *where* $Z$ *is a solution of* $F(Z) = \theta$.

*Proof.* We shall split the proof of this theorem into statements and proofs of a series of lemmas.

For given positive $\beta$, $g$ and $r$ consider a spindle-shaped region

$$S_\beta = \{X \in R^3,\ X = (x_1\ x_2\ x_3)^T : 0 \leqslant x_1 \leqslant \beta \text{ and}$$

$$\|(0\ x_2\ x_3)^T\|_2 \leqslant 1/(\beta^2 \sqrt{1 - (g^2 r/(2\beta))^2})\, ½\, g^2 r x_1 (\beta - x_1)\}.$$

It is well defined whenever

$$g^2 r/(2\beta) < 1. \tag{3.1}$$

Suppose that $r$ is as in the assumptions 3.1, $\beta = \|F(A) - F(B)\|_2$ and $g = \|A - B\|_2$, where $A, B \in P$. We want to show that $g < h_1$ implies that $g^2 r < \beta$ and, of course, (3.1) holds.

Indeed, the inequality $g < h_1$ is equivalent to $\|A - B\|_2 H_1 H_2^2 < 1/r$. Since $\|A - B\|_2 H_1 \geqslant \|F(A) - F(B)\|_2$ and $H_2 \|F(A) - F(B)\|_2 \geqslant \|A - B\|_2$, then $g^2 r < \beta$ and (3.1) holds.

Let $L_1$ be the linear transformation which transforms the vector $(\beta, 0, 0)^T$ onto the vector $F(B) - F(A)$. Let $L_2(X) = X + F(A)$ be a translation. We now define the region $S_{AB}$ by: $S_{AB} = \{Y \in R^3\colon Y = L_2(L_1 X) \text{ for } X \in S_\beta\}$. The sets $S_\beta$ and $S_{AB}$ are ilustrated in Figure 3.1.

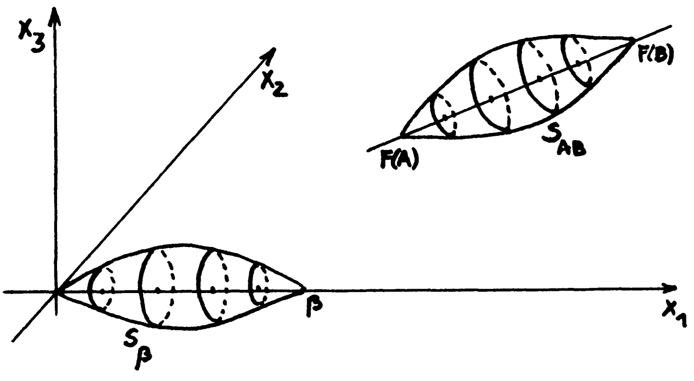


FIGURE 3.1

We can now formulate the first lemma.

LEMMA 3.3. *If $A, B \in P$ and $g < h_1$, then the curve*

$$Z = \{Z(t) \in R^3\colon Z(t) = F(tB + (1-t)A) \text{ for } t \in [0, 1]\}$$

*lies in $S_{AB}$.*

*Proof.* Let $A = (a_1\ a_2\ a_3)^T$, $B = (b_1\ b_2\ b_3)^T$ and $X(t) = tB + (1-t)A$ for $t \in [0, 1]$. We know that

$$\frac{d^2}{dt^2} f_l(X(t)) = (b_1 - a_1, b_2 - a_2, b_3 - a_3) \left(\frac{\partial^2 f_l}{\partial x_i \partial x_j}\right)_{i,j=1,2,3} \begin{pmatrix} b_1 - a_1 \\ b_2 - a_2 \\ b_3 - a_3 \end{pmatrix}$$

By Schwarz's inequality we get

$$\left|\frac{d^2}{dt^2} f_l(X(t))\right| \leq \|B - A\|_2^2 \max_P \left(\sum_{i,j=1}^{3} \left(\frac{\partial^2 f_l}{\partial x_i \partial x_j}\right)^2\right)^{1/2} = g^2 q_l.$$

Define $Y(t) = tF(B) + (1-t)F(A)$ for $t \in [0, 1]$. Then by the use of Lagrange interpolation we get

$$F(X(t)) - Y(t) = \frac{t(t-1)}{2}\left(\frac{d^2}{dt^2} f_1(X(t))|_{t=t_1}, \frac{d^2}{dt^2} f_2(X(t))|_{t=t_2}, \frac{d^2}{dt^2} f_3(X(t))|_{t=t_3}\right)^T,$$

where $t_j \in [0, 1]$, $j = 1, 2, 3$. Combining these formulas, we get

$$\|F(X(t)) - Y(t)\|_2 \leq \tfrac{1}{2} t(1-t) g^2 r. \tag{3.2}$$

For $t \in [0, 1]$, let $W(t)$ denote an envelope of the family of balls with centers $Y(t)$ and radii $\frac{1}{2}t(1-t)g^2 r$; see Figure 3.2. Then (3.2) states that the curve $F(X(t))$ lies in a bounded and closed set $W$ whose boundary is $W(t)$. One can simply verify that $W \subset S_{AB}$, which completes the proof of Lemma 3.3. □

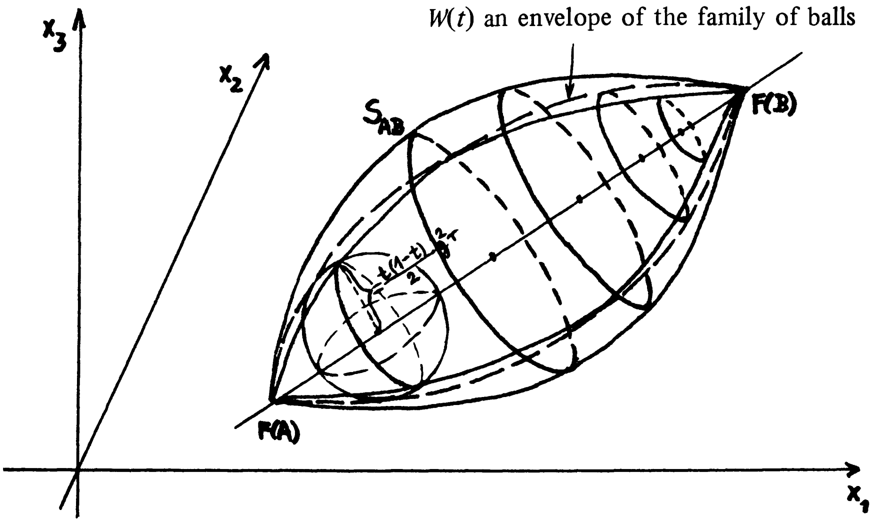


FIGURE 3.2

LEMMA 3.4. *Let $\triangleleft\!\triangleright A_1A_2E_1E_2$ be a tetrahedron lying in P, with the edges of length $\leqslant \min(h_1 h_2)$ and the face angles $\geqslant \gamma_1$. Then the intersection of the regions $S_{A_1A_2}$, $S_{A_1E_1}$ and $S_{A_1E_2}$ has exactly one point $F(A)$.*

*Proof.* Let $l_0$ denote the length of the longest edge of $\triangleleft\!\triangleright A_1A_2E_1E_2$. Note that $l_0 \leqslant h_1$ implies that the spindle-shaped regions $S_{A_1A_2}$, $S_{A_1E_1}$ and $S_{A_1E_2}$ are well defined.

We first wish to get a lower bound on the modulus of the sine of the angle $\varphi$ between $F(A_2) - F(A_1)$ and $F(E_1) - F(A_1)$; see Figure 3.3.

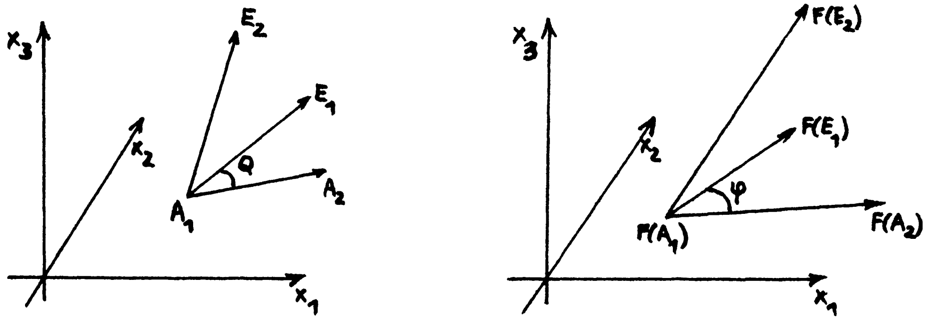


FIGURE 3.3

For this purpose set

$$E_j - A_1 = l_j \vec{s_j} : \|\vec{s_j}\|_2 = 1, \quad j = 0, 1, 2, \text{ where } E_0 = A_2$$

and $\|F(E_j) - F(A_1)\|_2 = \delta_j$. By Taylor's formula we get

$$F(E_j) - F(A_1) = l_j F'(A_1)\vec{s_j} + \vec{\epsilon_j}, \quad \text{where } |\epsilon_{j,k}| \leqslant \tfrac{1}{2} l_j^2 q_k \tag{3.3}$$

for $k = 1, 2, 3$. Finally,

$$\begin{aligned}|\sin\varphi| = \{&[(f_2(A_2)-f_2(A_1))(f_3(E_1)-f_3(A_1))\\ &\qquad -(f_3(A_2)-f_3(A_1))(f_2(E_1)-f_2(A_1))]^2\\ &+[(f_3(A_2)-f_3(A_1))(f_1(E_1)-f_1(A_1))\\ &\qquad -(f_1(A_2)-f_1(A_1))(f_3(E_1)-f_3(A_1))]^2\\ &+[(f_1(A_2)-f_1(A_1))(f_2(E_1)-f_2(A_1))\\ &\qquad -(f_2(A_2)-f_2(A_1))(f_1(E_1)-f_1(A_1))]^2\}^{1/2}\cdot 1/(\delta_0\delta_1).\end{aligned}$$

By Schwarz's inequality and (3.3) we get

$$\tag{3.4} \begin{aligned}|\sin\varphi| \geqslant 1/(\sqrt{3}\,\delta_0\delta_1)\Bigg\{&\left|l_0l_1\det\begin{pmatrix} a_1 & a_2 & a_3\\ f_{2,x_1} & f_{2,x_2} & f_{2,x_3}\\ f_{3,x_1} & f_{3,x_2} & f_{3,x_3}\end{pmatrix}+\eta_1+\xi_1\right|\\ &+\left|l_0l_1\det\begin{pmatrix} f_{1,x_1} & f_{1,x_2} & f_{1,x_3}\\ a_1 & a_2 & a_3\\ f_{3,x_1} & f_{3,x_2} & f_{3,x_3}\end{pmatrix}+\eta_2+\xi_2\right|\\ &+\left|l_0l_1\det\begin{pmatrix} f_{1,x_1} & f_{1,x_2} & f_{1,x_3}\\ f_{2,x_1} & f_{2,x_2} & f_{2,x_3}\\ a_1 & a_2 & a_3\end{pmatrix}+\eta_3+\xi_3\right|\Bigg\},\end{aligned}$$

where the all derivatives are evaluated at $A_1$ and

$$\xi_k = \epsilon_{0,k+1}\cdot\epsilon_{1,k+2}-\epsilon_{0,k+2}\cdot\epsilon_{1,k+2},\qquad k=1,2,3,$$

$$\begin{aligned}\eta_k = l_0&\left[\epsilon_{1,k+2}\sum_{l=1}^{3}(f_{k+1,x_l}s_{0,l})-\epsilon_{1,k+1}\sum_{l=1}^{3}(f_{k+2,x_l}s_{0,l})\right]\\ +l_1&\left[\epsilon_{0,k+1}\sum_{l=1}^{3}(f_{k+2,x_l}s_{1,l})-\epsilon_{0,k+2}\sum_{l=1}^{3}(f_{k+1,x_l}s_{1,l})\right],\end{aligned}$$

where $\epsilon_{j,k+3}=\epsilon_{j,k}$, $j=0,1$, and

$$a_1 = s_{0,2}s_{1,3}-s_{0,3}s_{1,2},\quad a_2 = s_{0,3}s_{1,1}-s_{0,1}s_{1,3},\quad a_3 = s_{0,1}s_{1,2}-s_{0,2}s_{1,1},$$

$$\tag{3.5} \begin{aligned}|\xi_k| &\leqslant \tfrac{1}{2}l_0l_1^2(q_{k+1}^2+q_{k+2}^2),\\ |\eta_k| &\leqslant \tfrac{1}{2}l_0l_1(l_0+l_1)(q_{k+1}^2+q_{k+2}^2)^{1/2}e_k,\end{aligned}$$

where $q_{k+3} = q_k$.

From Cramer's formula, (3.5) and the equation $\sin Q = (a_1^2 + a_2^2 + a_3^2)^{½}$ we finally get

$$|\sin \varphi| \geqslant 1/(\sqrt{3}\,\delta_0\delta_1)\left[l_0 l_1 j/M \sin \gamma_1 - \frac{l_0 l_1}{2}(l_0 + l_1)rE - \frac{1}{2} l_0 l_1 r^2\right].$$

The angle $\varphi$ belongs to $(0, \pi)$, so

$$\tan(\varphi/2) = \frac{\sin \varphi}{1 + \cos \varphi} > \frac{1}{2}\sin \varphi.$$

The region $S_{A_1A_2}$, (rsp. $S_{A_1E_1}$) lies entirely in a cone $C_{A_1A_2}$, (rsp. $C_{A_1E_1}$) with vertex at $F(A_1)$ opening in the direction $F(A_2) - F(A_1)$, (rsp. $F(E_1) - F(A_1)$) and with interior angle $2u_0$, (rsp. $2u_1$), where

$$\tan u_j = \frac{l_j r}{2\delta_j\sqrt{1 - (l_j^2 r/(2\delta_j))^2}}, \qquad j = 0, 1.$$

The regions $S_{A_1A_2}$ and $S_{A_1E_1}$ have only one common point $F(A_1)$ whenever the corresponding cones $C_{A_1A_2}$ and $C_{A_1E_1}$ have only one common point $F(A_1)$, that is if

$$(3.6) \qquad \tan u_0 \leqslant ½ \sin \varphi,$$

$$(3.7) \qquad \tan u_1 \leqslant ½ \sin \varphi.$$

We will first prove (3.6). Denote $X(t) = A_1 + l_1 t\vec{s_1}$ for $t \in [0, 1]$. Then

$$\delta_1 = \left\|\, l_1 \vec{s_1}^{\,T} \int_0^1 F'(X(t))\,dt \,\right\|_2 \leqslant l_1 p$$

and similarly $\delta_0 \leqslant l_0 p$. Since $0 < l_1 \leqslant l_0$ (3.6) holds whenever

$$(3.8) \qquad \frac{l_0^2 r}{2\delta_0\sqrt{1 - (l_0^2 r/(2\delta_0))^2}} \leqslant \frac{1}{2\sqrt{3}}\,[l_0 j/M \sin \gamma_1 - l_0^2 rE - ½ l_0^3 r^2]\,\frac{1}{\delta_0\delta_1}.$$

Note that (3.8) is a quadratic inequality for $l_0$. This has a solution whenever

$$(3.9) \qquad 0 < l_0 \leqslant h_2 = 1/r\{[(E + 2p)^2 + 2j/M \sin \gamma_1]^{½} - (E + 2p)\}.$$

It is easy to show that (3.9) implies (3.7). Thus, if $0 < l_0 \leqslant \min(h_1, h_2)$, then $S_{A_1A_2} \cap S_{A_1E_1} = \{F(A_1)\}$. By an analogous argument we get $S_{A_1A_2} \cap S_{A_1E_2} = \{F(A_1)\}$ and $S_{A_1E_1} \cap S_{A_1E_2} = \{F(A_1)\}$ for $0 < l_0 \leqslant \min(h_1, h_2)$. These equations imply that

$$S_{A_1A_2} \cap S_{A_1E_1} \cap S_{A_1E_2} = \{F(A_1)\}. \quad \square$$

LEMMA 3.5. *Let the tetrahedron* $\lhd\!\rhd\, A_1A_2E_1E_2$ *lie in P and let it belong to the family generated by* A, B, C, *or* D; *see Section* 2. *Let* $l_0$ *be the length of its longest edge, such that* $l_0 < h_4$, *where* $h_4 = \min(1, h_1, h_2, h_3)$. *If* $\theta \in T_F \lhd\!\rhd\, A_1A_2E_1E_2$, *then* $\theta \in T_F \lhd\!\rhd_i$, *where* $\lhd\!\rhd_i$ *is one of the tetrahedra constructed in step* (6) *of the algorithm.*

*Proof.* Let us illustrate the tetrahedra constructed in step (6) (see Figure 3.4). We shall prove that for $l_0 < h_4$ and for all pairs of the indices $m, j$ where $m \in \{i + 1, \dots, i + N/2 - 1\}, j \in \{i - 1, \dots, i - N/2 + 1\}, i = 1, \dots, N$, the points $F(E_m)$ and $F(E_j)$, $(E_{i \pm N} = E_i)$ lie on the opposite sides of the plane defined by the points $F(A_1)$, $F(A_2)$ and $F(E_i)$. For $N = 6$ we get Figure 3.4.

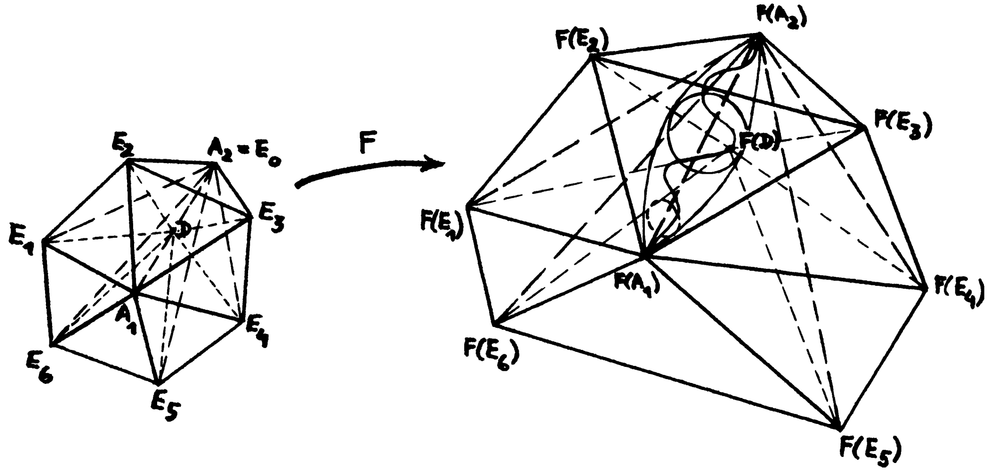


FIGURE 3.4

We first want to get the cosine of the angle between the vectors $F(E_j) - F(A_1)$ and $\vec{w}_i$ for $j = 1, \dots, N, j \neq i$, where $\vec{w}_i$ is a normal to the plane defined by the points $F(A_1)$, $F(A_2)$ and $F(E_i)$. For this purpose set

$$E_i - A_1 = l_i \vec{s}_i : \| \vec{s}_i \|_2 = 1, \qquad i = 0, \dots, N, \; E_0 = A_2,$$

$$\|F(E_i) - F(A_1)\|_2 = \delta_i, \qquad \sphericalangle(A_2 - A_1, E_i - A_1) = Q_i,$$

$$\sphericalangle(\vec{w}_{1,i}, E_t - A_1) = \omega_{t,i}, \qquad i, t = 1, \dots, N,$$

where $\vec{w}_{1,i}$ is a normal to the plane defined by the points $A_1$, $A_2$ and $E_i$. $\sphericalangle(\vec{w}_i, F(E_t) - F(A_1)) = \varphi_{t,i}$ and $\delta_{i \pm N} = \delta_i$, $l_{i \pm N} = l_i$, $\omega_{t \pm N, i} = \omega_{t,i}$, $\varphi_{t \pm N, i} = \varphi_{t,i}$, $Q_{i \pm N} = Q_i$. Taylor's formula then yields

$$F(E_i) - F(A_1) = l_i F'(A_1)\, \vec{s}_i + \vec{\epsilon}_i, \qquad i = 0, \dots, N,$$

where $|\epsilon_{i,k}| \leqslant \frac{1}{2} l_i^2 q_k$, $k = 1, 2, 3$.

From the formula for the cosine between two vectors, the equation $\vec{a}_i^T \cdot \vec{s}_t = \cos \omega_{t,i} \sin Q_i$, $i, t = 1, \dots, N$, where

$$\vec{a}_i = \begin{pmatrix} s_{0,2}s_{i,3} - s_{0,3}s_{i,2} \\ s_{0,3}s_{i,1} - s_{0,1}s_{i,3} \\ s_{0,1}s_{i,2} - s_{0,2}s_{i,1} \end{pmatrix}$$

and from the considerations similar to those used in the proof of Lemma 3.4 we get

$$
\begin{aligned}
\cos \varphi_{t,i} = \frac{1}{\| \vec{w}_i \|_2 \delta_t} \{ & \det F'(A_1) l_t l_i l_0 \cos \omega_{t,i} \sin Q_t \\
& + \det F'(A_1) \vec{a}_i^{\,T} F'(A_1)^{-1} \cdot \vec{\epsilon}_t \\
& + l_t (\vec{\eta}_i^{\,T} + \vec{\xi}_i^{\,T}) F'(A_1) \vec{s}_t + (\vec{\eta}_i^{\,T} + \vec{\xi}_i^{\,T}) \vec{\epsilon}_t \}, \qquad t \neq i,
\end{aligned}
\tag{3.9}
$$

where

$$
\begin{aligned}
\eta_{i,k} = {} & l_0 \left[ \epsilon_{i,k+2} \sum_{l=1}^{3} (f_{k+1,x_l} s_{0,l}) - \epsilon_{i,k+1} \sum_{l=1}^{3} (f_{k+2,x_l} s_{0,l}) \right] \\
& + l_i \left[ \epsilon_{0,k+1} \sum_{l=1}^{3} (f_{k+2,x_l} s_{i,l}) - \epsilon_{0,k+2} \sum_{l=1}^{3} (f_{k+1,x_l} s_{i,l}) \right], \\
\xi_{i,k} = {} & \epsilon_{0,k+1} \cdot \epsilon_{i,k+2} - \epsilon_{0,k+2} \cdot \epsilon_{i,k+1}, \qquad k = 1, 2, 3,
\end{aligned}
$$

and

$$
\epsilon_{i,k+3} = \epsilon_{i,k}, \qquad f_{k+3} = f_k .
$$

Schwarz's inequality yields

$$
\begin{aligned}
& |\det F'(A_1)\, \vec{a}_i^{\,T} F'(A_1)^{-1}\, \vec{\epsilon}_t \,| \leqslant J \|F'(A_1)^{-1}\|_2 \, ½ l_t^2 l_i l_0 r, \\
& |(\vec{\eta}_i + \vec{\xi}_i)^T F'(A_1) \vec{s}_t \,| \leqslant \|F'(A_1)\|_2 (½ l_i^2 l_0^2 r^2 + ½ l_i l_0 (l_i + l_0) r E), \\
& |(\vec{\eta}_i + \vec{\xi}_i)^T \vec{\epsilon}_t| \leqslant (½ l_i^2 l_0^2 r^2 + ½ l_i l_0 (l_i + l_0) r E) \cdot ½ l_t^2 r.
\end{aligned}
\tag{3.10}
$$

From (2.4), (2.5), the assumption $l_i \leqslant l_0 \leqslant 1$, and (3.10) we find

$$
\begin{aligned}
& |\cos \varphi_{t,i}| \, \|\vec{w}_i\|_2 \\
& \geqslant 1/\delta_t l_i l_0 (j \sin \gamma_1 \sin \gamma_2 - l_0^2 (½ H_1 r^2 + ½ E r^2 + ¼ r^3) - l_0 (½ J H_2 r + E H_1 r)) \\
& = PR,
\end{aligned}
\tag{3.11}
$$

where $t \neq i$. Furthermore, the inequalities $\delta_i \leqslant l_i p$ and a simple transformation yield

$$
PR > 0 \quad \text{for } 0 < l_0 < h_3 .
\tag{3.12}
$$

Next, (3.9) and (3.10) imply that

$$
|\cos \varphi_{t,i} \|\vec{w}_i\|_2 - L_{t,i}| \leqslant P_{t,i},
\tag{3.13}
$$

where

$$
\begin{aligned}
L_{t,i} = {} & 1/\delta_t \det F'(A_1) l_t l_i l_0 \cos \omega_{t,i} \sin Q_t, \\
P_{t,i} = {} & 1/\delta_t [½ l_t^2 l_i l_0 J H_2 r + ½ l_t l_i^2 l_0^2 H_1 r^2 + ½ l_t l_i l_0 (l_i + l_0) r E H_1 \\
& + ½ l_t^2 r (½ l_i^2 l_0^2 r^2 + ½ l_i l_0 (l_i + l_0) r E)] .
\end{aligned}
$$

Moreover, (3.12) implies that for $0 < l_0 < \min(1, h_3)$

$$|L_{t,i}| > P_{t,i}. \tag{3.14}$$

From the construction $\operatorname{sign}(\cos \omega_{m,i}) = -\operatorname{sign}(\cos \omega_{j,i})$. This equation, (3.13) and (3.14) give $\operatorname{sign}(\cos \varphi_{m,i}) = -\operatorname{sign}(\cos \varphi_{j,i})$, which we wanted to get. From this, Lemmas 3.3, 3.4 and a geometric interpretation (see Figure 3.4) we get Lemma 3.5. □

The next two lemmas will be proven in the $n$-dimensional case, where $n \geqslant 2$.

3.15. *Assumptions.*

1. Let $F: D \subset R^n \longrightarrow R^n$ belong to $C^2(D)$, where $D$ is an open set.

2. Let $P \subset D$ be the $n$-dimensional polyhedron, and $P = \bigcup_{i=1}^{M} S_i$, where $S_i$ are the $n$-dimensional simplexes such that $\operatorname{Int} S_i \cap \operatorname{Int} S_j = \emptyset$, $i \neq j$.

3. $\min_{X \in \partial P} \|F(X)\|_2 = d > 0$.

4. $\deg(F, \operatorname{Int} P, \theta) \neq 0$.

5. $g$ is the length of the longest edge (one-dimensional face) of any simplex $S_i$ in $P$.

6. $T_F S_i$ denotes a simplex with the vertices $F(s_{i,j})$, where $s_{i,j}$ are the vertices of a simplex $S_i$.

7. $r = (\Sigma_{l=1}^{n} q_l^2)^{1/2}$, where

$$q_l = \max_P \left( \sum_{i,j=1}^{n} \left( \frac{\partial^2 f_l}{\partial x_i \partial x_j} \right)^2 \right)^{1/2}, \quad l = 1, \ldots, n.$$

8. $p = \max_{X \in P} \|F'(X)\|_E$, where

$$\|F'(X)\|_E = \left( \sum_{i,j=1}^{n} (f_{i,x_j}(X))^2 \right)^{1/2}$$

We can now formulate

LEMMA 3.6. *If the assumptions* (3.15) (1–7) *are satisfied and* $g < \sqrt{d/(rn)}$, *then* $\theta \in T_F S_{i_0}$ *for some simplex* $S_{i_0}$ *in* $P$.

*Proof.* Let us define a linear transformation $G_i$ on every simplex $S_i$ such that $G(s_{i,j}) = F(s_{i,j})$. The resulting transformation $G$ is defined as $G(X) = G_i(X)$ if $X \in S_i$. Observe that $G$ is continuous on $P$.

Let $X$ be any point in $P$. Then there exists a simplex $S_i$ such that $X \in S_i$. Let us find an upper bound on $\|F(X) - G(X)\|_2$. Note that $X = t_1 s_{i,1} + (1 - t_1) t_2 s_{i,2} + \cdots + (1 - t_1) \cdots (1 - t_{n-1}) (t_n s_{i,n} + (1 - t_n) s_{i,n+1})$ for some $t_1, t_2, \ldots, t_n$, where $t_j \in [0, 1]$ for $j = 1, \ldots, n$. Then from the inequality (3.2), which also holds for $n \geqslant 2$, we obtain

$$\|F(X) - G(X)\|_2 \leqslant g^2 r/8 + (1 - t_1) g^2 r/8 + \cdots + (1 - t_1)(1 - t_2) \cdots (1 - t_{n-1}) \cdot g^2 r/8 \leqslant n g^2 r/8.$$

If $g < \sqrt{d/(rn)}$, then

$$\|F - G\|_P = \sup_{X \in P} \|F(X) - G(X)\|_2 \leqslant n g^2 r/8 < d/8.$$

Moreover, $0 < \min_{X \in \partial P} \|F(X) - \theta\|_2 = d > 7/8d$. These inequalities and Theorem

6.2.1 of [2] give

$$\deg(F, \operatorname{Int} P, \theta) = \deg(G, \operatorname{Int} P, \theta).$$

Since $\deg(F, \operatorname{Int} P, \theta) \neq 0$, we get $\deg(G, \operatorname{Int} P, \theta) \neq 0$; and by Kronecker's theorem (see [2]), there exists a point $Z \in P$ such that $G(Z) = \theta$. That is, $Z \in S_{i_0}$ for some $S_{i_0} \subset P$. This implies, however, by the linearity of $G$ on $S_{i_0}$ that $\theta \in T_G S_{i_0} = T_F S_{i_0}$. □

LEMMA 3.7. *Let all of the assumptions* (3.15) *be satisfied; let $S_{i_0} \subset P$ be a simplex such that $\theta \in T_F S_{i_0}$ and $g \leqslant d/(2p)$. Then*

$$\min_{X \in S_{i_0}; Y \in \partial P} \|X - Y\|_2 \geqslant g. \tag{3.16}$$

*Proof.* Every point $Z$ of $T_F S_{i_0}$ can be uniquely represented in the form $Z = \Sigma_{j=1}^{n+1} \alpha_j F(s_{i_0,j})$, where $\alpha_j \in [0, 1], j = 1, \ldots, n+1$, and $\Sigma_{j=1}^{n+1} \alpha_j = 1$. Let $X, Y \in P$ and $X(t) = Y + t(X - Y)$, $t \in [0, 1]$. Then

$$F(X) - F(Y) = (X - Y)^T \int_0^1 F'(X(t))\,dt \quad \text{and} \quad \|F(X) - F(Y)\|_2 \leqslant \|X - Y\|_2 p.$$

Now let $\|X - Y\|_2 = k$ and $X \in S_{i_0}$, $Y \in \partial P$ be such points that $\min_{V \in S_{i_0}; W \in \partial P} \|V - W\|_2 = \|X - Y\|_2$; these exist because the sets $P$ and $S_{i_0}$ are compact. By the triangle inequality,

$$\begin{aligned} \|Z\|_2 &= \left\| F(X) - \sum_{j=1}^{n+1} \alpha_j F(X) + \sum_{j=1}^{n+1} \alpha_j F(s_{i_0,j}) \right\|_2 \\ &\geqslant \|F(X)\|_2 - \sum_{j=1}^{n+1} \alpha_j \|F(X) - F(s_{i_0,j})\|_2. \end{aligned}$$

However, $\|X - s_{i_0,j}\|_2 \leqslant g$ for $j = 1, \ldots, n+1$, and $\|F(X)\|_2 \geqslant \|F(Y)\|_2 - \|F(X) - F(Y)\|_2$. From these inequalities we get

$$\|Z\|_2 \geqslant \|F(Y)\|_2 - kp - \sum_{j=1}^{n+1} \alpha_j pg = \|F(Y)\|_2 - (k+g)p \geqslant d - (k+g)p. \tag{3.17}$$

Since $Z$ is an arbitrary point of $T_F S_{i_0}$, then (3.17) and the assumption that $g \leqslant d/(2p)$ yields $\theta \notin T_F S_{i_0}$ whenever $k < g$. Hence, (3.16) holds. □

*Completion of Proof of Theorem* 3.2. a. Let us assume that all assumptions 3.1 are satisfied. We thus arrive at step (3) of the algorithm and by Lemma 3.6 we there find a tetrahedron $\triangleleft\!\!\triangleright_i$ such that $\theta \in T_F \triangleleft\!\!\triangleright_i$. We thus arrive at step (5) in which we check whether or not $g_i$, the length of the longest edge of $\triangleleft\!\!\triangleright_i$, is less than or equal to $\epsilon$. If so, we print out $g_i$ and the vertices $\triangleleft\!\!\triangleright_{i,k}$ of $\triangleleft\!\!\triangleright_i$ and go to (STOP). If no, we proceed to step (4).

If we construct the tetrahedra around each edge of $\triangleleft\!\!\triangleright_i$ as in step (6) of the algorithm, then (2.7) and Lemma 3.5 imply that $T_F \triangleleft\!\!\triangleright_i \subset \bigcup_{j=1}^t F(\triangleleft\!\!\triangleright_j)$, where $\triangleleft\!\!\triangleright_j$ are the constructed tetrahedra. Since $\theta \in T_F \triangleleft\!\!\triangleright_i$, there exists a point $Z \in \triangleleft\!\!\triangleright_{j_0}$, $j_0 \in \{1, 2, \ldots, t\}$ such that $F(Z) = \theta$. From Lemma 3.7 each of these tetrahedra lies entirely in $P$. Moreover, from the construction $\|Z - \triangleleft\!\triangleright_{i,k}\|_2 \leqslant 2g_i \leqslant 2\epsilon$, $k = 1, 2, 3, 4$. Now suppose that $g_i > \epsilon$. Then, in step (4) we bisect $\triangleleft\!\triangleright_i$ and check which of the new tetrahedra $\triangleleft\!\triangleright_j$, $j = 1, 2$, satisfies $\theta \in T_F \triangleleft\!\triangleright_j$. If one of the $T_F \triangleleft\!\triangleright_j$ does contain $\theta$, we return to step (5). If no $T_F \triangleleft\!\triangleright_j$ contains $\theta$, we proceed to step (6). By Lemma 3.7 the newly formed tetrahedra in step (6) lie entirely in $P$. Moreover, by Lemma 3.5 there exists a $\triangleleft\!\triangleright_i$ such that $\theta \in T_F \triangleleft\!\triangleright_i$. We then return to step (5). In all cases we remain in steps (4), (5), (6). At every bisection the longest edge of a tetrahedron is halved–so after a finite number of returns to step (5) the inequality $g_i \leqslant \epsilon$ has to be satisfied.

b. Let us now assume that all of the assumptions 3.1 except 5 hold. In this case we either achieve convergence in steps (4), (5), (6); or we may pass to step (7) from step (3), because $\theta$ is not contained in any $T_F \triangleleft\!\triangleright_i$ for $i = 1, 2, \ldots, M$, or from step (6) because the newly constructed tetrahedron is not entirely in $P$ or $\theta \notin T_F \triangleleft\!\triangleright_j$ for all $\triangleleft\!\triangleright_j$ constructed in step (6). However, each time we arrive at step (7) the longest edge of each tetrahedron in $P$ is halved, so after a finite number of steps the assumption 3.1–5 becomes satisfied. □

*The Rate of Convergence.* If we traverse the route steps (4), (5), (4), (5) . . . , then the error after $n$ evaluations of $F$ is $O(2^{-n/3})$ as $n \longrightarrow \infty$. In the worst case, if we traverse steps (5), (4), (6), (5), (4), (6) . . . then the error after $n$ evaluations of $F$ is $O(2^{-n/15})$ as $n \longrightarrow \infty$. Numerical tests indicate that a bond $O(2^{-n/3})$ is rather practical when $h$ is sufficiently small.

**4. Numerical Tests of the Algorithm.** The algorithm (2.8) has been tested on the CDC 6000 computer for the following functions:

$$F(x, y, z) = \begin{pmatrix} (x - 0.5)^{10}(y - 0.5)z \\ \cos((y - 0.5)^2) - (x - 1.5)^2 \\ (y + z - 1.0)^2 \end{pmatrix}, \tag{4.1}$$

$$F(x, y, z) = \begin{pmatrix} \exp(\sin(x + y + 4z - 3)) - 1.0 \\ \cos((y - 0.5)^2) - (x - 1.5)^2 \\ (y + z - 1.0)^2 \end{pmatrix}, \tag{4.2}$$

$$F(x, y, z) = \begin{pmatrix} (x + 1.5)^4 - 16(y + z)^3 \\ \cos((y - 0.5)^2) - (x - 1.5)^2 \\ (y + z - 1.0)^2 \end{pmatrix}, \tag{4.3}$$

$$F(x, y, z) = \begin{pmatrix} (x - 0.5)^{10}(y - 0.5)z \\ \cos((y - 0.5)^2) - (x - 1.5)^2 \\ \cos(x + y + z - 1.5) \end{pmatrix}, \tag{4.4}$$

$$(4.5) \qquad F(x, y, z) = \begin{pmatrix} (x - 0.5)^{10}(y - 0.5)z \\ \cos((y - 0.5)^2) - (x - 1.5)^2 \\ (x - 0.5)^{10} + (y - 0.5)(z - 0.5) \end{pmatrix}.$$

In the case (4.1) we get convergence starting from the three parallelepipeds

$$P_1 = [-\sqrt{2}/2, \sqrt{2}/2] \times [-1, 1] \times [-\sqrt{2}/2, \sqrt{2}/2],$$

$$P_2 = [-\sqrt{2}, \sqrt{2}] \times [-2, 2] \times [-\sqrt{2}, \sqrt{2}],$$

$$P_3 = [-2\sqrt{2}, 2\sqrt{2}] \times [-4, 4] \times [-2\sqrt{2}, 2\sqrt{2}].$$

We solved (1.1) in 144 steps within an error of $10^{-14}$. For the functions (4.4) and (4.5) we obtained solution to $10^{-2}$ accuracy only, because there too little storage was reserved in the computer (we required over 36000 words). That is why we did not obtain solution for the functions (4.2) and (4.3). It seems to us that the amount of storage needed for the algorithm can sometimes be significant.

**5. Final Comments.** There are a number of interesting problems to study regarding this algorithm—for instance:

1. Can the algorithm be generalized to the $n$-dimensional case?

2. For the function (4.1) we obtained root $Z$ such that $\det F'(Z) = 0$. Does the algorithm always converge in this case?

3. How might this method be combined with one which converges more rapidly near a root?

4. How can one solve the problem of excessive storage requirements?

We shall study these problems in the future.

**Acknowledgments.** I am indebted to H. Woźniakowski for his comments during the preparation of this paper and to F. Stenger for his remarks on the manuscript.

Department of Mathematics
University of Warsaw
00–901 Warszawa PKiN 8p.
p. 850 Poland

1. CH. HARVEY & F. STENGER, "A two dimensional analogue to the method of bisections for solving nonlinear equations," *Quart. Appl. Math.*, v. 33, 1976, pp. 351–368.

2. J. M. ORTEGA & W. C. RHEINBOLDT, *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

3. F. STENGER, "Computing the topological degree of a mapping in $n$-space," *Numer. Math.*, v. 25, 1975, pp. 23–38.