QUARTERLY OF APPLIED MATHEMATICS
JANUARY, 1976

A TWO-DIMENSIONAL ANALOGUE TO THE METHOD OF BISECTIONS FOR SOLVING NONLINEAR EQUATIONS*

BY

CHARLES HARVEY (*Dickinson College, Carlisle, Pa.*)

AND

FRANK STENGER (*University of Utah*)

1. Introduction and summary. There has been considerable interest in the development of numerical methods for solving nonlinear systems of equations $F(X) = \theta$, which do not require the evaluation of partial derivatives of F. The development of such methods has been slow, since proofs of convergence for well-known one-dimensional methods such as the method of bisections or the method of false position are not easily extended to higher dimensions. Nevertheless, there is an effective analogue to the method of false position due to Kincaid [1] and several higher-dimensional "Newton-like" methods (see Rheinboldt [3] and Dennis [4]), as well as some minimization methods (see Powell [7]). A direct extension of the method of false position is the method of Gauss [2, p. 234] for which no proof of convergence appears to exist.

In the present paper we develop a two-dimensional method for obtaining an approximate solution of the system of equations

$$F(X) \equiv F(x, y) \equiv (f(x, y), g(x, y)) = \theta \equiv (0, 0) \tag{1.1}$$

which resembles the one-dimensional method of bisections. Let us motivate the two-dimensional method by a brief description of the one-dimensional method.

At the outset, we want to ensure that we can in fact apply the method of bisections to determine an approximate solution ξ of the one-dimensional problem

$$Q(x) = 0. \tag{1.2}$$

There are simple sufficient conditions for this to be the case, namely, if Q is continuous and real on a finite interval $[a, b]$ and if we can find two points x_1 and $x_2 (x_1 < x_2)$ on $[a, b]$ such that $Q(x_1)Q(x_2) < 0$. At the outset, then, we can search for two such points x_1 and x_2 by evaluating Q at

$$a, b, \qquad a + \frac{2k - 1}{2n}(b - a), \qquad k = 1, 2, \cdots, n.$$

If we can find two points x_1 and x_2 or $[a, b]$ such that $Q(x_1)Q(x_2) \leq 0$, then there exists a point ξ on the interval $[x_1, x_2]$ such that $q(\xi) = 0$. We can then compute an approximate value of ξ by the method of bisections, i.e., by means of the following algorithm:

(i) Set $c = (x_1 + x_2)/2$;

* Received September 16, 1972; revised version received August 5, 1974.

(ii) If $f(x_1)f(c) \leq 0$, set $x_2 = c$ and return to (i); If $f(x_1)f(c) > 0$, set $x_1 = c$ and return to (i).

Using this algorithm we halve the interval $[x_1, x_2]$ at each step. If after n steps we approximate ξ by either x_1 or x_2, the error is at most $x_2 - x_1 \leq (b - a)/2^n$.

In Sec. 2 we describe a simple test to determine whether the system (1.1) has a solution in a polygonal domain $\mathfrak{D}$ with boundary points at $X^1, \cdots, X^N$, which one meets consecutively as one traverses $\mathfrak{D}$ in a counter-clockwise manner. Our test is based on the use of the formula

$$\delta_N(F, \mathfrak{D}, \theta) \equiv \frac{1}{8} \sum_{i=1}^{N} \begin{vmatrix} \operatorname{sgn} f(X^i) & \operatorname{sgn} g(X^i) \\ \operatorname{sgn} f(X^{i+1}) & \operatorname{sgn} g(X^{i+1}) \end{vmatrix} \tag{1.3}$$

where $X^{N+1} = X^1$. This formula yields $\delta_N(F, \mathfrak{D}, \theta) = d(F, \mathfrak{D}, \theta)$, which is the topological degree of F at θ relative to $\mathfrak{D}$, provided that f and g are real and continuous in $\bar{\mathfrak{D}}$, the closure of $\mathfrak{D}$, $F \neq \theta$ on the boundary of $\mathfrak{D}$, and fg has at most one sign change on each line segment $\overline{X^iX^{i+1}}$ joining the points X^i and X^{i+1}. By Kronecker's theorem [9, p. 161], if $d(F, \mathfrak{D}, \theta) \neq 0$ then the system (1.1) has at least one solution in $\mathfrak{D}$.

In Sec. 3 we devise a simple test for determining whether or not the point θ is contained in a triangle,

$$\Delta A^1A^2A^3 \equiv \left\{X: X = \sum_{i=1}^{3} \lambda_i A^i, \lambda_i \geq 0, \sum_{i=1}^{3} \lambda_i = 1\right\},$$

where the points A^i are three non-collinear points in the plane. If $A = (a_1, a_2)$ and $B = (b_1, b_2)$ are two distinct points in the plane, then we can define a linear form

$$L(A, B, X) \equiv (b_2 - a_2)(x - a_1) - (b_1 - a_1)(y - a_2). \tag{1.4}$$

We show that $\theta \in \Delta A^1A^2A^3$ if and only if

$$L(A^i, A^{i+1}, \theta)L(A^i, A^{i+1}, A^{i+2}) \geq 0 \quad \text{for} \quad i = 1, 2, 3. \tag{1.5}$$

where $A^{i+3} = A^i$. We then describe an algorithm (Algorithm 3.1) for finding an approximate solution of (1.1) which combines the results of Sec. 2 and uses the relations (1.5) as well as the idea of *bisecting triangles*. We bisect $\Delta A^1A^2A^3$ by first locating the longest edge $\overline{A^iA^{i+1}}$, setting $D = (A^i + A^{i+1})/2$, and then forming two triangles, ΔA^iDA^{i+2} and $\Delta DA^{i+1}A^{i+2}$.

In Sec. 4 we state conditions which enable us to prove the convergence of Algorithm 3.1.

Algorithm 3.1 has been tested in applications. In [11] a Fortran program has been written which begins with a rectangular region $\mathfrak{D}$, and which computes $d(F, \mathfrak{D}, \theta)$ using (1.3). If $d(F, \mathfrak{D}, \theta) = 0$, the program requests a new rectangle; if $d(F, \mathfrak{D}, \theta) \neq 0$, the program branches to the "triangulation stage," which starts by bisecting $\mathfrak{D}$ into two triangles. Indeed, for the problem

$$\begin{aligned} f(x, y) &\equiv .15 - \frac{3M_2 - 6M_1M_0 + 2M_0^3}{(2M_1 - M_0^2)^{3/2}} = 0 \\ g(x, y) &\equiv 3.2 - \frac{4M_3 - 12M_2M_0 + 12M_1M_0^2 - 3M_0^4}{(2M_1 - M_0^2)^{3/2}} = 0 \end{aligned} \tag{1.6}$$

in which $M_i = \Gamma((j + 1)/x)\Gamma(y - (j + 1)/x)/(x\Gamma(y))$, and which arose in statistical

applications, Newton's method, Powell's method [7] the method of steepest descents, Box's complex algorithm [12] and the flexiplex algorithm [13] all failed to produce a solution, whereas the program of [11] enabled us to solve (1.6) to 3 dec. accuracy.

2. Location of a region containing a root. Let P be a polygon in the $X = xy$-plane, with N vertices $X^1, X^2, \cdots, X^N$, which one meets consecutively as one traverses P in a counter-clockwise manner. Let the polygon P form the boundary of a simply connected and bounded domain $\mathfrak{D}$. With reference to (1.1), let f, g, f_x, f_y, g_x and g_y be real, continuous and bounded on $\bar{\mathfrak{D}}$, where $\bar{\mathfrak{D}}$ denotes the closure of $\mathfrak{D}(\mathfrak{D} = \bar{\mathfrak{D}} - P)$, and let $f^2 + g^2 \neq 0$ on P.

It can be shown [6, p. 321], that

$$N_+ - N_- = \frac{1}{2\pi} \int_P \frac{f\,dg - g\,df}{f^2 + g^2} \tag{2.1}$$

where $N_+(N_-)$ denotes the number of solutions of (3.1) in $\mathfrak{D}$ at which $f_x g_y - f_y g_x > 0 (< 0)$, provided that $(f_x g_y - f_y g_x)(Y) \neq 0$ for all $Y \in \mathfrak{D}$ such that $F(Y) = \theta$. In general, if $F = (f, g) \neq \theta \equiv (0, 0)$ on P, then the right-hand side of (2.1) is called the topological degree of F at θ relative to $\mathfrak{D}$ and is denoted by $d(F, \mathfrak{D}, \theta)$. If $d(F, \mathfrak{D}, \theta) \neq 0$, then by Kronecker's theorem (see [9, p. 161]) there exists at least one point $X \in D$ such that $F(X) = \theta$. Since the integrand in (2.1) is just $d(\arctan(g/f))$, if the maximum distance between X^i and X^{i+1} is sufficiently small, then (2.1) can be evaluated by means of the formula

$$d(F, \mathfrak{D}, \theta) = \frac{1}{2\pi} \sum_{i=1}^{N} \arctan \frac{g(X)}{f(X)} \Bigg|_{X^i}^{X^{i+1}} \tag{2.2}$$

where $X^{N+1} = X^1$.

The above sum can be very simply evaluated. To this end, we introduce a simple notion of a sign change of (f, g). Let $\overline{AB}$ denote the closed line segment in the plane, joining the points A and B. *A point $X \in \overline{AB}$ is called a sign change of (f, g) on $\overline{AB}$ if fg changes sign at X.*

We shall assume that the vertices X^i of P are chosen such that (f, g) has at most one sign change on each segment $\overline{X^i X^{i+1}}$, $i = 1, 2, \cdots, N$. We then replace the coordinates $(f(X^i), g(X^i))$ by (u_i, v_i), where $u_i = \operatorname{sgn} f(X^i)$(sgn $a = 1$ if $a > 0$, 0 if $a = 0$, and -1 if $a < 0$) and $v_i = \operatorname{sgn} g(X^i)$. We thus get a "graph" as in Fig. 2.1, where the u_i and v_i are either ± 1 or 0.

Let us assign the numerical value b_i to the line segment joining (u_i, v_i) and (u_{i+1}, v_{i+1}), where

$$b_i = \frac{1}{8} \begin{vmatrix} u_i & v_i \\ u_{i+1} & v_{i+1} \end{vmatrix} = \frac{u_i v_{i+1} - u_{i+1} v_i}{8} \tag{2.3}$$

and where $(u_{N+1}, v_{N+1}) = (u_1, v_1)$.

The following result is then established in [8].

THEOREM 2.1. *Let the polygon P be defined as above, where $f^2 + g^2 \neq 0$ on P, and such that fg has at most one sign change on each of the line segments $\overline{X^i X^{i+1}}$, $i = 1, 2, \cdots, N$. Then the number $d(F, \mathfrak{D}, \theta)$ on the right-hand side of (2.1) is also given by*

$$d(f, \mathfrak{D}, \theta) = \sum_{i=1}^{N} b_i \tag{2.4}$$

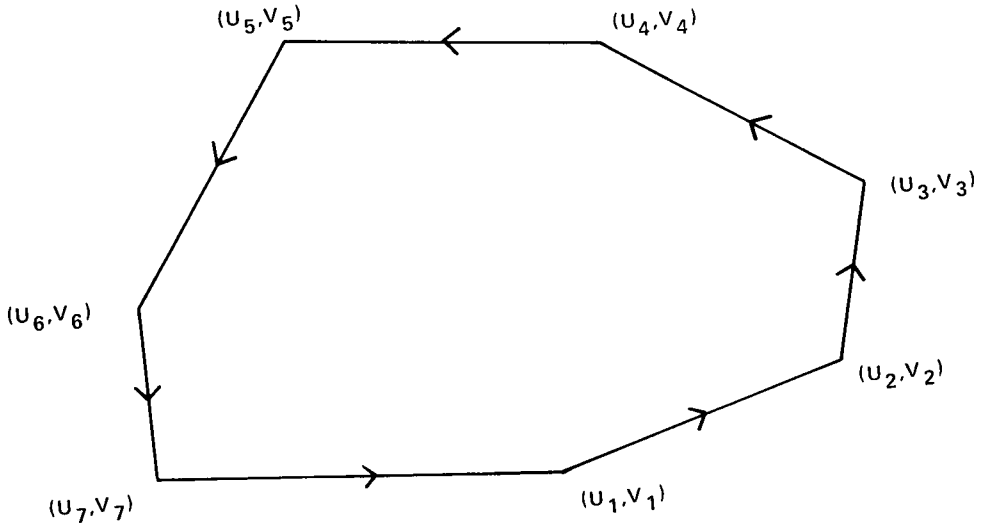

FIG. 2.1. Graph of the sign changes of F on P.

where the b_i are defined in (2.3).

Ex. 2.1. Let us apply the above procedure to show that the system of equations

$$f(x, y) \equiv x^2 - 4y = 0, \qquad g(x, y) \equiv y^2 - 2x + 4y = 0 \tag{2.5}$$

has at least one solution in the domain

$$\mathfrak{D} = \{(x, y): |x| < 2, |y| < 1/4\}. \tag{2.6}$$

Notice that the system $(f, g) = (0, 0)$ has the solution $(x, y) = (0, 0)$ in $\mathfrak{D}$.

We enclose the domain $\mathfrak{D}$ by a polygon P, being careful to choose the vertices of P in consecutive and counter-clockwise order so that (f, g) has at most one sign change on each of the segments $\overline{X^i X^{i+1}}$. Seven points were thus chosen, to yield the results in Table 2.1. Evaluating the sum of the b_i, we get

$$d(F, \mathfrak{D}, \theta) = \sum_{i=1}^{7} b_i = -\tfrac{1}{4} + 0 - \tfrac{1}{4} + 0 + 0 - \tfrac{1}{4} - \tfrac{1}{4} = -1.$$

Hence the system of equations $(f, g) = (0, 0)$ has at least one solution in $\mathfrak{D}$. Notice that it does not suffice to take only the points X^2, X^3, X^5 and X^6, i.e. the corner points of the rectangular region $\mathfrak{D}$, although the points X^3, X^4 and X^5 could, for example, have been dropped.

TABLE 2.1. Zeros in $\mathfrak{D}$ of $(f, g) = (0, 0)$.

i	x_i	y_i	$f(x_i, y_i)$	$g(x_i, y_i)$	u_i	v_i	b_i
1	− .5	.25	− .75	2.0625	−1	1	−1/4
2	−2.0	.25	3.0	5.0625	1	1	0
3	−2.0	−.25	5.0	3.0625	1	1	−1/4
4	.75	−.25	1.5625	−2.4375	1	−1	0
5	2.0	−.25	5.0	−4.9375	1	−1	0
6	2.0	.25	3.0	−2.9375	1	−1	−1/4
7	.75	.25	− .4375	− .4375	−1	−1	−1/4

In practice, it will often be worth while to start with a rectangular domain and to continually add points at each mid-point between every pair of consecutive points on the boundary, until the sum (2.4) remains a fixed integer.

3. Method of bisection of triangles. In this section we describe an algorithm for solving the system

$$F(X) = F(x, y) = (f(x, y), \quad g(x, y)) = \theta = (0, 0). \tag{3.1}$$

At the outset, we introduce some fundamental definitions which simplify the description of the algorithm. We then give a brief algorithmic statement of the algorithm, and we follow up each step by a detailed discussion. A proof of convergence is given in Sec. 4.

Let $A = (a_1, a_2)$ and $B = (b_1, b_2)$ denote two distinct points in the plane. We denote by $L(A; B; X) = L(A; B; (x, y))$ the linear form

$$L(A; B; X) \equiv (b_2 - a_2)(x - a_1) - (b_1 - a_1)(y - a_2). \tag{3.2}$$

Let

$$l_{AB} \equiv \{X: L(A; B; X) = 0\} \tag{3.3}$$

denote the straight line through the points A and B. The line l_{AB} divides the plane into two regions $R_{AB}{}^1$ and $R_{AB}{}^2$ where $R_{AB}{}^1 = \{X: L(A; B; X) \geq 0\}$, $R_{AB}{}^2 = \{X: L(A; B; X) \leq 0\}$. Let $B^i = (b_1{}^i, b_2{}^i)$ $(i = 1, 2, 3)$ denote three non-collinear points in the plane, and let us further set $B^{i+3} = B^i$, $i = 1, 2, 3$. We denote a triangle with vertices at B^1, B^2 and B^3 by

$$\Delta B^1B^2B^3 \equiv \bigcap_{i=1}^{3} \{X: L(B^i; B^{i+1}; X)L(B^i; B^{i+1}; B^{i+2}) \geq 0\}. \tag{3.4}$$

That is, $\Delta B^1B^2B^3$ is the region common to the three half planes, $R_{B^iB^{i+1}}{}^{j_i}$, where j_i is either 1 or 2, and $R_{B^iB^{i+1}}{}^{j_i}$ is that half plane defined above by the points B^i and B^{i+1} which contains the point B^{i+2}. Thus *the point* $\theta = (0, 0)$ *is in* $\Delta B^1B^2B^3$ *if and only if*

$$L(B^i; B^{i+1}; \theta)L(B^i; B^{i+1}; B^{i+2}) \geq 0, \qquad i = 1, 2, 3. \tag{3.5}$$

Notice that $L(B^i; B^{i+1}; B^{i+2})$ is independent of i and has the same value, plus or minus twice the area of $\Delta B^1B^2B^3$, for $i = 1, 2, 3$, and that

$$L(B^1; B^2; \theta) + L(B^2; B^3; \theta) + L(B^3; B^1; \theta) = L(B^1; B^2; B^3). \tag{3.6}$$

It will be convenient to let

$$T_F B^1B^2B^3 \equiv \Delta F(B^1)F(B^2)F(B^3). \tag{3.7}$$

The process of *bisecting a triangle* $\Delta A^1A^2A^3$ *into two triangles* is defined as follows. We *first locate the longest side* $\overline{A^iA^{i+1}}$ *of* $\Delta A^1A^2A^3$, *where* $A^{i+3} = A^i$. *Next, we set* $D = (A^i + A^{i+1})/2$, *to get two new triangles,* ΔA^iDA^{i+2} *and* $\Delta DA^{i+1}A^{i+2}$.

ALGORITHM 3.1.

1. Form the polygon P and evaluate $d(F, \mathfrak{D}, \theta)$.
2. Does $\mathfrak{D}$ contain a solution of (1.1)?
   (Yes) Go to 3.
   (No) Go to 1.

A detailed description of Steps 1 and 2 is given in Sec. 2. The number $d(F, \mathfrak{D}, \theta)$ is computed using (2.4).

3. Triangulate $\bar{\mathfrak{D}}$ with M suitable triangles Δ_I, $I = 1, 2, \cdots, M$, such that all interior angles of the triangles are $\geq \alpha$, where $0 < \alpha \leq \pi/3$, such that

$$\bigcup_{I=1}^{M} \Delta_I = \bar{\mathfrak{D}},$$

and such that the intersection of any two of the triangles has zero area.

Upon arriving at Step 3, we have found a polygon P for which $d(F, \mathfrak{D}, \theta) \neq 0$. We triangulate $\bar{\mathfrak{D}} = \mathfrak{D} \cup P$ by adding points in the interior of $\mathfrak{D}$, if necessary, such that the size of each interior angle of the triangle is at least α, where $0 < \alpha \leq \pi/3$. Here α is arbitrary, although the convergence of the algorithm may be more rapid for a larger value of α. We also index each of the triangles and then proceed to Step 4.

4. $I = 1$.
5. Is $I \leq M$?
   (No) Go to 10.
   (Yes) $\theta \in T_F\, \Delta_I$?
   (No) $I = I + 1$. Go to 5.
   (Yes) Go to 7.

   In steps 4 and 5 we systematically test each of the triangles Δ_I in $\bar{\mathfrak{D}}$ in order to find a triangle $T_F\, \Delta_I$ (see Eq. (3.7)) which contains the point $\theta = (0, 0)$, by means of (3.5). If we find such a triangle Δ_I, we proceed to Step 7. If we do not find such a triangle after all of the M triangles have been tested, we proceed to Step 10.
6. Bisect $\Delta_I = \Delta A^1 A^2 A^3$ into $\Delta^{(1)} = \Delta A^i D A^{i+2}$ and $\Delta^{(2)} = \Delta D A^{i+1} A^{i+2}$

   $\Delta \leftarrow \Delta_I$

   $\Delta_I \leftarrow \Delta^{(1)}$

   $\Delta_{J+1} \leftarrow \Delta_J$, $J = M, M - 1, \cdots, I + 1$

   $\Delta_{I+1} \leftarrow \Delta^{(2)}$

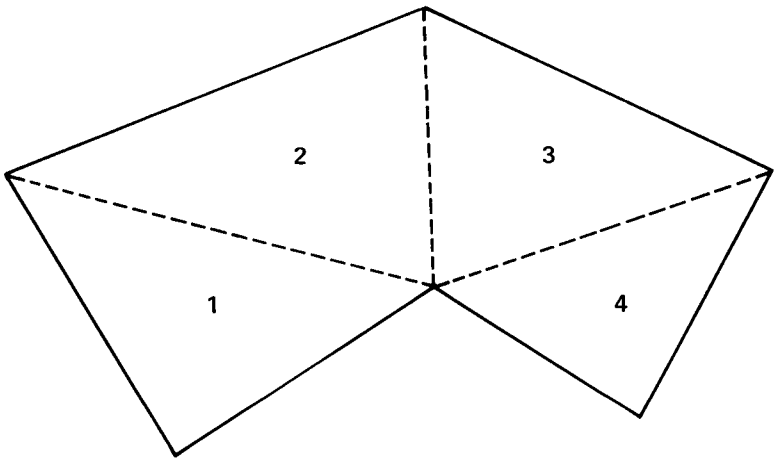

FIG. 3.1. The region $\bar{\mathfrak{D}}$ and its triangulation.

$M \leftarrow M + 1$

$\theta \in T_F\, \Delta_I$?

(Yes) Go to 7.

(No) $I \leftarrow I + 1$

$\theta \in T_F\, \Delta_I$?

(Yes) Go to 7.

(No) Go to 8.

Upon arriving at Step 6, $\theta \in T_F\, \Delta_I$. We bisect $\Delta_I = \Delta A^1A^2A^3$ into $\Delta^{(1)} = \Delta A^iDA^{i+2}$ and $\Delta^{(2)} = \Delta DA^{i+1}A^{i+2}$, where $\overline{A^iA^{i+1}}$ is the longest edge of Δ_I , and $D = (A^i + A^{i+1})/2$. The triangles $\Delta^{(1)}$ and $\Delta^{(2)}$ then become Δ_I and Δ_{I+1} ; the remaining triangles Δ_I in the array $(J = I + 1, \cdots, M)$ are moved up one position to make room for the newly defined triangles. If θ is in one of $T_F\, \Delta_I$ or $T_F\, \Delta_{I+1}$, we proceed to Step 7; if θ is in neither of these, we go to Step 8. Note that we require $F(D)$ in order to carry out the test $\theta \in T_F\, \Delta_I$.

7. h_I = length of longest side of $\Delta A^1A^2A^3$.

Is $h_I \leq \epsilon$?

(Yes) Print h_I , A^1, A^2, A^3 and stop.

(No) Go to 6.

8. Let $\Delta = \Delta A^1A^2A^3$, where $\overline{A^iA^{i+1}}$ is the longest side of Δ. Set $E = A^i + A^{i+1} - A^{i+2}$.

Is $\Delta A^i EA^{i+1}$ wholly in $\mathfrak{D}$?

(No) Go to 10.

(Yes) $\Delta^{(1)} \leftarrow \Delta A^iED$

$\Delta^{(2)} \leftarrow \Delta DEA^{i+1}$

$\Delta_{J+2} \leftarrow \Delta_J$, $J = M, M - 1, \cdots, I + 1$

$\Delta_{I+1} \leftarrow \Delta^{(1)}$

$\Delta_{I+2} \leftarrow \Delta^{(2)}$

$M \leftarrow M + 2$. Go to 9.

In Step 8 we locate a point E by forming a parallelogram $\amalg$ whose vertices are A^i, E, A^{i+1} and A^{i+2}, and such that A^{i+2} and E are opposite corners of $\amalg$ (see Fig. 3.2). Thus $E = A^i + A^{i+1} - A^{i+2}$. We then check whether or not the newly formed triangle ΔA^iEA^{i+1}

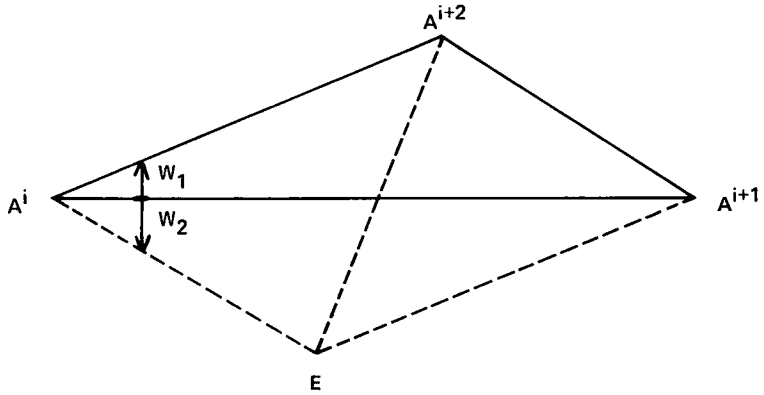

FIG. 3.2. Location of the point E.

is in $\mathfrak{D}$. If the original polygonal region $\mathfrak{D}$ has many vertices, this may be a difficult test to perform. If for example, $\mathfrak{D}$ is a rectangle with vertices at (x_1, y_1), (x_2, y_1), (x_2, y_2), and (x_1, y_2), where $x_1 < x_2$ and $y_1 < y_2$, then we need only check to ensure that $x_1 \leq x_E \leq x_2$, $y_1 \leq y_E \leq y_2$, where $E = (x_E, y_E)$. If $\Delta A^i E A^{i+1}$ is not wholly in $\mathfrak{D}$, we go to Step 10. Otherwise we bisect $\Delta A^i E A^{i+1}$ into $\Delta^{(1)} = \Delta A^i E D$ and $\Delta^{(2)} = \Delta D E A^{i+1}$, where $D = (A^i + A^{i+1})/2$, and we move the triangles Δ_J, $J = M, M-1, \cdots, I+1$ up two positions to make room for two newly defined triangles $\Delta_{I+1} = \Delta^{(1)}$ and $\Delta_{I+2} = \Delta^{(2)}$. Notice that these newly formed triangles may overlap some of the triangles that are already in the array.

9. $I \leftarrow I + 1$

   $\theta \in T_F\, \Delta_I$?

   (Yes) Go to 7.

   (No) I $\leftarrow I + 1$

   $\theta \in T_F\, \Delta_I$?

   (Yes) Go to 7.

   (No) Go to 10.

   In Step 9 we check whether or not $\theta \in T_F\, \Delta_I$ or $\theta \in T_F\, \Delta_{I+1}$, where Δ_I and Δ_{I+1} are defined in Step 8. If so we return to Step 7; if not, we go to Step 10.

10. $I \leftarrow M$

    Bisect Δ_I into $\Delta^{(1)}$ and $\Delta^{(2)}$.

    $\Delta_{2I} \leftarrow \Delta^{(1)}$

    $\Delta_{2I-1} \leftarrow \Delta^{(2)}$

    $I \leftarrow I - 1$

    Is $I \geq 1$?

    (Yes).

    (No) $M \leftarrow 2M$. Go to 4.

    In Step 10 we bisect every triangle Δ_I, $I = M, M-1, \cdots, 1$, to create $2M$ new triangles. We then return to Step 4. It could happen that M is a very large number. Then the problems of storage and an overhead (or combinatory) cost of one iterative step appear. A method of circumventing this difficulty has been implemented in [11].

Ex. 3.1. Let us apply the above algorithm to obtain an approximate solution of the problem of Ex. 2.1, namely

$$f(x, y) \equiv x^2 - 4y = 0, \qquad g(x, y) \equiv y^2 - 2x + 4y = 0. \tag{3.8}$$

We shall describe what happens in each step of Algorithm 3.1. The vertices of the polygon P and the corresponding values of (f, g) are given in Table 2.1. The vertices of the successive triangles Δ_I such that $\theta \in T_F\, \Delta_I$ are tabulated in Table 3.1.

Steps 1 and 2. These have already been carried out in Ex. 2.1, where it was shown that the system (3.8) has a solution in $\mathfrak{D}$.

Step 3. We triangulate $\mathfrak{D}$ into 5 triangles as in Fig. 3.3:
$\Delta_1 = \Delta X^1 X^2 X^3$, $\Delta_2 = \Delta X^1 X^3 X^4$, $\Delta_3 = \Delta X^1 X^4 X^7$, $\Delta_4 = \Delta X^7 X^4 X^5$, $\Delta_5 = \Delta X^5 X^6 X^7$. The points X^1 to X^7 in Fig. 3.3 are the same as those in Table 2.1. We arbitrarily take $\alpha = .08$, $\epsilon = .2$.

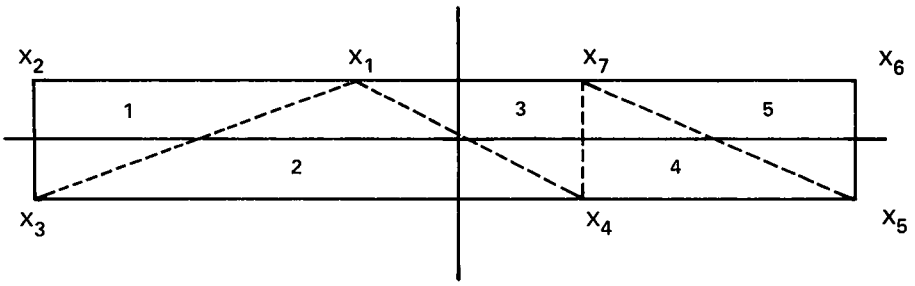

FIG. 3.3. Triangulation of the region $\bar{\mathfrak{D}}$.

Step 5. The following tests are carried out in Step 5. $\theta \in T_F\,\Delta_1$? (No); $\theta \in T_F\,\Delta_2$? (No); $\theta \in T_F\,\Delta_3$? (Yes). For example, to test whether or not $\theta \in T_F\,\Delta_3$, we evaluate

$$L(F(X^1); F(X^4); F(X^7)) = (-2.4375 - 2.0625)(-.4375 - (-.75)) - (1.5625 - (-.75))(-.4375 - 2.0625) \simeq 4.37$$

$$L(F(X^2); F(X^4); \theta) = (-2.4375 - 2.0625)(0 - (-.75)) - (1.5625 - (-.75))(0 - 2.0625) \simeq 1.39$$

$$L(F(X^4); F(X^7); \theta) = (-4.375 - 1 - 2.4375))(0 - 1.5625) - (.4375 - 1.5625)(0 - (-2.4375)) \simeq 1.75$$

$$L(F(X^7; F(X^1); \theta) = 4.37 - 1.39 - 1.75 = 1.23.$$

Since $4.37 \times 1.39 \geq 0$, $4.37 \times 1.75 \geq 0$, and $4.37 \times 1.23 \geq 0$, $\theta = (0, 0) \in T_F\,\Delta_3 = T_F\,X^1X^4X^7$. Notice that the system (3.8) has the solution $(x, y) = (0, 0) \in \Delta_2$, whereas $(0, 0) \notin T_F\,\Delta_2$.

Step 7. h_3, the longest side of Δ_3, is $||X_1 - X_4|| \simeq 1.34 > .2$, and so we go to Step 6.

Step 6. Here we bisect $\Delta_3 = \Delta X^1X^4X^7$. Since the longest side of Δ_3 is $\overline{X^1X^4}$ $D = (X^1 + X^4)/2 = (.125, 0)$. Thus we set $\Delta_6 \leftarrow \Delta_5$, $\Delta_5 \leftarrow \Delta_4$, $\Delta_4 \leftarrow \Delta^{(2)}$, $\Delta_3 \leftarrow \Delta^{(1)}$, where $\Delta^{(1)} = \Delta X^1DX^7$, $\Delta^{(2)} = \Delta DX^4X^7$. We then make the tests $\theta \in T_F\,\Delta_3$? (No); $\theta \in T_F\,\Delta_4$? (No). We thus proceed to Step 8.

Step 8. Here we locate the point

$$E = X^1 + X^4 - X^7 = (-.5, .25) + (.75, -.25) - (.75, .25) = (-.5, -.25).$$

Since $-2 \leq -.5 \leq 2$, and $-.25 \leq -.25 \leq .25$, the newly formed triangle, ΔX^1EX^4 clearly lies wholly in $\bar{\mathfrak{D}}$. We thus bisect ΔX^1EX^4 into ΔX^1ED and ΔDEX^4, where $D = (X^1 + X^4)/2 = (.125, 0)$. The triangles ΔX^1ED and ΔDEX^4 become Δ_5 and Δ_6 respectively; the triangles #6 and #7 now become Δ_7 and Δ_8 respectively. Notice that the triangles Δ_5 and Δ_6 overlap with triangle #2. We now proceed to Step 9.

Step 9. Here we make the test $\theta \in T_F\,\Delta_5$? (Yes). Hence we go to Step 7. From this point onward the algorithm does not return to Steps 8 and 9, but remains in Steps 7 and 6.

4. Convergence. In this section we obtain sufficient conditions for the convergence of Algorithm 3.1.

TABLE 3.1. Tabulation of the successive vertices of $\Delta_I = \Delta A^1A^2A^3$ such that $\theta \in T_P\Delta_I$.

A^1		A^2		A^3	
−.5	.25	.75	−.25	.75	.25
−.5	.25	−.5	−.25	.125	0
−.1875	.125	−.5	−.25	.125	0
−.1875	.125	−.1875	−.125	.125	0
−.03125	.0625	−.1875	−.125	.125	0
−.03125	.0625	−.03125	−.0625	.125	0

Assumptions 4.1.

(i) f, g, f_x, f_y, g_x, g_y, f_{xx}, f_{xy}, f_{yy}, g_{xx}, g_{xy} and g_{yy} are real, continuous and bounded in $\bar{\mathfrak{D}}$;

(ii) $j \equiv \min_{\bar{\mathfrak{D}}} |f_x g_y - f_y g_x| > 0$;

(iii) $p \equiv \max_{\bar{\mathfrak{D}}} (f_x^2 + f_y^2 + g_x^2 + g_y^2)^{1/2} < \infty$;

(iv) $q_1 \equiv \max_{\bar{\mathfrak{D}}} (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)^{1/2} < \infty$

$q_2 \equiv \max_{\bar{\mathfrak{D}}} (g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2)^{1/2} < \infty$;

(v) $d \equiv \min_{P} (f^2 + g^2)^{1/2} > 0$;

(vi) $d(F, \mathfrak{D}, \theta) \equiv \dfrac{1}{2\pi} \displaystyle\int_P \frac{f\,dg - g\,df}{f^2 + g^2} \neq 0$;

(vii) $h \leq h^* \equiv \min \{d/(2p), [(16p^2 + 4j \sin(\alpha/2))^{1/2} - 4p]/r, (7d/r)^{1/2}\}$ where h is the longest side of any triangle in $\bar{\mathfrak{D}}$ and $r = (q_1^2 + q_2^2)^{1/2}$.

THEOREM 4.2. *Let* $0 < \epsilon < h^*$, *where* ϵ *appears in* Step 7 *of* Algorithm 3.1 *and where* h^* *is defined in* Assumptions 4.1 (vii). *If* Assumptions 4.1 (i)–(vi) *are satisfied, then* Algorithm 3.1 *prints* h_I, A^1, A^2 *and* A^3 *where* h_I *is the longest edge of* $\Delta A^1A^2A^3$ *and where each* A^i *is within* 2ϵ *of a solution* (ξ, η) *in* $\mathfrak{D}$ *of* Eq. (1.1).

Notice that we do *not assume* a sufficiently small distance between X^i and X^{i+1} in the definition of P such that the conditions of Theorem 2.1 are satisfied.

It is convenient to split the proof of this theorem into statements and proofs of a series of lemmas.

Let $\Delta_{11} = \Delta ABC$ be a triangle having all of its interior angles $\geq \alpha > 0$. Let us bisect Δ_{11} to form two triangles Δ_{2i}, $i = 1, 2$, then bisect each of the triangles Δ_{2i} to obtain four triangles, Δ_{3i}, $i = 1, 2, 3, 4$, and so on, to form a family T of triangles. The following result is established in [5].

LEMMA 4.3. *If* $\Delta \in T$ *and* θ *is an interior angle of* Δ, *then* $\theta \geq \alpha/2$.

We next establish several interpolation results.

If $A = (a_1, a_2)$ and $B = (b_1, b_2)$ we denote the distance $[(a_1 - a_2)^2 + (b_1 - b_2)^2]^{1/2}$ by $||B - A||$.

For given positive β, h and r, consider the lens-shaped region

$$S_\beta = \{X = (x, y): 0 \leq x \leq \beta, |y| \leq \beta^{-2}x(\beta - x)h^2r/2\}. \tag{4.1}$$

Suppose r is as defined in Assumption 4.1 (vii), $h = ||B - A||$, and $\beta = ||F(B) - F(A)||$. Define c and s by

$$(c, s) \equiv [F(B) - F(A)]/||F(B) - F(A)|| . \tag{4.2}$$

We may then define a region S_{AB} by

$$S_{AB} = \left\{U = (u, v): U = F(A) + (x, y)\begin{pmatrix} c & s \\ -s & c \end{pmatrix}, (x, y) \in S_\beta\right\}. \tag{4.3}$$

The lens-shaped region S_{AB} is illustrated in Fig. 4.1.

LEMMA 4.4. *If the line segment*

$$\{X = (x, y): X = X(t) \equiv tB + (1 - t)A, 0 \leq t \leq 1\} \tag{4.4}$$

lies in $\mathfrak{D}$, *then the image curve*

$$\{Z = (x, y): Z = Z(t) \equiv F(tB + (1 - t)A), 0 \leq t \leq 1\} \tag{4.5}$$

lies in S_{AB} .

Proof of Lemma 4.4. Let $X(t)$ be defined as in (4.4) and let us define $Y(t)$ by

$$Y(t) = tF(B) + (1 - t)F(A), 0 \leq t \leq 1. \tag{4.6}$$

Then by use of Lagrange interpolation with error,

$$F(X(t)) - Y(t) = \frac{t(t - 1)}{2}\left(\frac{d^2}{dt^2} f(X(t))\Big|_{t=\xi}, \frac{d^2}{dt^2} g(X(t))\Big|_{t=\eta}\right) \tag{4.7}$$

for some numbers ξ and η between 0 and 1. Now if $A = (a_1 , a_2)$, $B = (b_1 , b_2)$, then $X(t)$ takes the form

$$X(t) = t(b_1 , b_2) + (1 - t)(a_1 , a_2), 0 \leq t \leq 1,$$

and so

$$\frac{d^2}{dt^2} f(X(t)) = (b_1 - a_1 , b_2 - a_2)\begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}\begin{pmatrix} b_1 - a_1 \\ b_2 - a_2 \end{pmatrix}.$$

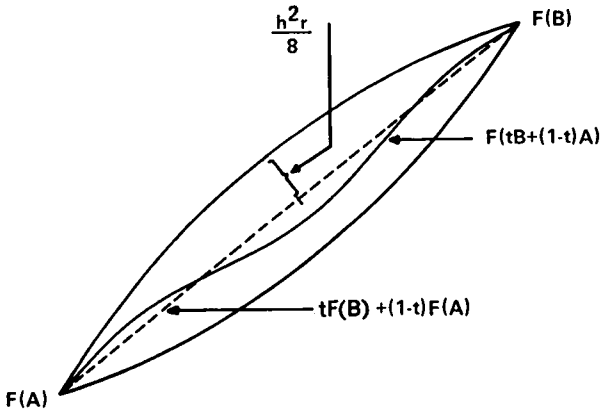

FIG. 4.1. The lens-shaped region S_{AB} .

By applying Schwarz's inequality we get

$$\left|\frac{d^2}{dt^2} f(X(t))\right| \leq q_1 h^2, \qquad \left|\frac{d^2}{dt^2} g(X(t))\right| \leq q_2 h^2 \tag{4.8}$$

where $h = ||B - A||$, and q_1 and q_2 are defined in Assumptions 4.1 (iv). Combining (4.7) and (4.8), we get

$$||F(X(t)) - Y(t)|| \leq \frac{t(1-t)}{2} h^2 r \tag{4.9}$$

for $0 \leq t \leq 1$, where $r = (q_1^2 + q_2^2)^{1/2}$.

Let us first assume that $F(A) = (0, 0)$, $F(B) = (\beta, 0)$. The inequality (4.9) then states that the curve $F(X(t))$ lies in the region bounded by the two parabolas, $y = \pm\beta^{-2}x(\beta - x)h^2 r/2$, where $0 \leq x \leq \beta$ and where $\beta = ||F(B) - F(A)||$. If we make the transformation

$$Y = F(A) + X\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

where $(c, s) = [F(B) - F(A)]/||F(B) - F(A)||$ to transform the vector $(t\beta, 0)$ in the X-plane onto the vector $tF(B) + (1-t)F(A)$ in the Y-plane, the statement of Lemma 4.4 follows.

LEMMA 4.5. *Let ΔABC be a triangle lying in $\bar{\mathfrak{D}}$ with sides of length $\leq h$ and interior angles $\geq \alpha/2 > 0$. If* Assumption 4.1 (vii) *is satisfied, then the regions S_{AB} and S_{AC} defined in* (4.3) *have only the point $F(A)$ in common.*

Proof of Lemma 4.5. Let us first obtain a lower bound on the modulus of the sine of the angle φ between $F(B) - F(A)$ and $F(C) - F(A)$. For this purpose, set $B - A \equiv h(c, s)$, $C - A \equiv k(\gamma, \sigma)$, where $c = \cos w$, $s = \sin w$, $\gamma = \cos(w + \omega)$, $\sigma = \sin(w + \omega)$, and where $\alpha/2 \leq |\omega| \leq \pi - \alpha$. Taylor's formula then yields

$$\begin{aligned} f(B) - f(A) &= h[cf_x + sf_y] + \epsilon_f^{(1)} \\ f(C) - f(A) &= k[\gamma f_x + \sigma f_y] + \epsilon_f^{(2)} \\ g(B) - g(A) &= h[cg_x + sg_y] + \epsilon_g^{(1)} \\ g(C) - g(A) &= k[\gamma g_x + \sigma g_y] + \epsilon_g^{(2)}, \end{aligned} \tag{4.10}$$

where f_x, f_y, g_x and g_y are evaluated at A. The errors $\epsilon_f^{(i)}$ and $\epsilon_g^{(i)}$ $(i = 1, 2)$ are given by

$$\begin{aligned} \epsilon_f^{(1)} &= \frac{1}{2}\frac{d^2}{dt^2} f(X(t))\Big|_{t=\tau_1}, & \epsilon_f^{(2)} &= \frac{1}{2}\frac{d^2}{dt^2} f(Y(t))\Big|_{t=\tau_2} \\ \epsilon_g^{(1)} &= \frac{1}{2}\frac{d^2}{dt^2} g(X(t))\Big|_{t=\tau_3}, & \epsilon_g^{(2)} &= \frac{1}{2}\frac{d^2}{dt^2} g(Y(t))\Big|_{t=\tau_4} \end{aligned} \tag{4.11}$$

where $X(t) = A + th(c, s)$, $Y(t) = A + tk(\gamma, \sigma)$, $0 \leq t \leq 1$, and where $0 < \tau_i < 1$, $i = 1, 2, 3, 4$. By proceeding as for (4.7), we find that

$$\begin{aligned} |\epsilon_f^{(1)}| &\leqq \frac{h^2}{2} q_1, & |\epsilon_f^{(2)}| &\leqq \frac{k^2}{2} q_1 \\ |\epsilon_g^{(1)}| &\leqq \frac{h^2}{2} q_2, & |\epsilon_g^{(2)}| &\leqq \frac{k^2}{2} q_2. \end{aligned} \tag{4.12}$$

Now setting $\beta = ||F(B) - F(A)||$, $\delta = ||F(C) - F(A)||$, and using (4.10), it follows that

$$\begin{aligned}\sin\varphi &= \frac{[F(B) - F(A)] \times [F(C) - F(A)]}{\beta\delta} \\ &= \frac{[f(B) - f(A)][g(C) - g(A)] - [f(C) - f(A)][g(B) - g(A)]}{\beta\delta} \\ &= \frac{hk(f_x g_y - f_y g_x)\sin\omega + \eta_1 + \eta_2}{\beta\delta}\end{aligned} \tag{4.13}$$

where

$$\begin{aligned}\eta_1 &= \epsilon_f^{(1)} k(\gamma g_x + \sigma g_y) - \epsilon_g^{(1)} k(\gamma f_x + \sigma f_y) + \epsilon_g^{(2)} h(c f_x + s f_y) - \epsilon_f^{(2)} h(c g_x + s g_y), \\ \eta_2 &= \epsilon_f^{(1)}\epsilon_g^{(2)} - \epsilon_f^{(2)}\epsilon_g^{(1)}.\end{aligned} \tag{4.14}$$

Using Schwarz's inequality and (4.12), we find that

$$\begin{aligned}|\eta_1| &\le \frac{kh^2}{2}(q_1^2 + q_2^2)^{1/2} p + \frac{k^2 h}{2}(q_1^2 + q_2^2)^{1/2} p = \frac{kh}{2}(h + k) r p \\ |\eta_2| &\le [(\epsilon_f^{(1)})^2 + (\epsilon_g^{(1)})^2]^{1/2}[(\epsilon_f^{(2)})^2 + (\epsilon_g^{(2)})^2]^{1/2} \\ &\le \frac{h^2}{2}(q_1^2 + q_2^2)^{1/2}\frac{k^2}{2}(q_1^2 + q_2^2)^{1/2} = \frac{h^2 k^2}{4} r^2\end{aligned} \tag{4.15}$$

where p and r are defined in Assumptions 4.1.

We now combine (4.13), (4.14), (4.15) and Assumption 4.1 (ii) to get the lower bound

$$|\sin\varphi| \ge \frac{hkj\,|\sin\omega| - hk(h + k)pr/2 - h^2 k^2 r^2/4}{\beta\delta} \tag{4.16}$$

on the modulus of the sine of the angle between the vectors $F(B) - F(A)$ and $F(C) - F(A)$.

Let us assume without loss of generality that the angle φ is a positive acute angle, so that

$$\tan(\varphi/2) = \frac{\sin\varphi}{1 + \cos\varphi} > (1/2)\sin\varphi. \tag{4.17}$$

Using (4.1), we find that the region $S_{AB}(S_{AC})$ lies entirely in a cone $C_{AB}(C_{AC})$ with vertex at $F(A)$, opening in the direction $F(B) - F(A)(F(C) - F(A))$ and with interior angle $2u(2v)$, where

$$\tan u = h^2 r/2\beta, \qquad \tan v = k^2 r/2\delta \tag{4.18}$$

The regions S_{AB} and S_{AC} defined in (4.3) have only the point $F(A)$ in common if the corresponding cones C_{AB} and C_{AC} have only the point $F(A)$ in common, that is, if

$$\tan u \le (1/2)\sin\varphi, \qquad \tan v \le (1/2)\sin\varphi. \tag{4.19}$$

We shall first prove that $\tan u \le (1/2)\sin\varphi$. To this end, we insert the inequality

$$\delta = \left\| \int_0^1 k(\gamma - \sigma) \begin{bmatrix} f_x(Y(t)) & g_x(Y(t)) \\ f_y(Y(t)) & g_y(Y(t)) \end{bmatrix} dt \right\| \le kp \tag{4.20}$$

in (4.16). We furthermore assume without loss of generality that $0 < k \leq h$, and we also recall the inequality $\alpha/2 \leq |\omega| \leq \pi - \alpha$. The relation $\tan u \leq (1/2) \sin \varphi$ will thus be satisfied if

$$\frac{h^2 r}{2\beta} \leq \frac{h^2 j \sin(\alpha/2) - h^3 pr - h^4 r^2/4}{2h\beta p}, \tag{4.21}$$

that is, solving for h, if

$$0 < h \leq r^{-1}[(16p^2 + 4j \sin(\alpha/2))^{1/2} - 4p]. \tag{4.22}$$

The proof that if Assumption 4.1 (vii) is satisfied then $\tan v \leq (1/2) \sin \varphi$ is similar, and we omit it.

LEMMA 4.6. *With reference to* Fig. 3.2, *let* $\Delta A^1A^3A^3$ *and* ΔA^iEA^{i+1} *be triangles in* $\mathfrak{D}$, *constructed in* Steps 6 and 8 *of* Algorithm 3.1, *such that the sides of these triangles are* $\leq h$ *in length, where* h *satisfies* (4.22). *Set* $D = (A^i + A^{i+1})/2$. *If* $\theta \in T_F A^1A^2A^3$, *then* θ *is in one of the triangles* $T_F A^iDA^{i+2}$, $T_F DA^{i+2}A^{i+2}$, $T_F A^iED$ *or* $T_F DEA^{i+1}$.

Proof of Lemma 4.6. Let $\theta \in T_F A^1A^2A^3$. If θ is in one of the triangles $T_F A^iDA^{i+2}$ or $T_F DA^{i+1}A^{i+2}$, there is nothing further to demonstrate. If θ is in neither of these triangles, then we recall by Lemma 4.5 that the interiors of the regions $S_{A^iA^{i+1}}$, $S_{A^{i+1}A^{i+2}}$ and $S_{A^iA^{i+2}}$ (see Fig. 4.2) are non-intersecting. Also the intoriors of the regions S_{A^iE}, $S_{A^iA^{i+1}}$ and $S_{EA^{i+1}}$ are non-intersecting.

Let us next show that if (4.22) is satisfied, then $F(A^{i+2})$ and $F(E)$ lie on opposite sides of the straight line through $F(A^i)$ and $F(A^{i+1})$. In (4.13), if we replace A by A^i, B by A^{i+1} and C first by A^{i+2} and then by E, we find, with reference to Figs. 3.2 and 4.2, that

$$\left|\sin \varphi_j - \frac{hk_j(f_x g_y - f_y g_x)(-1)^{j-1} \sin \omega_j}{\beta \delta_j}\right| \leq \left\{\frac{hk_j}{2}(h + k_j)pr + \frac{h^2 k_j^2 r^2}{4}\right\}/(\beta \delta_j), \qquad j = 1, 2, \tag{4.23}$$

where $h = ||A^{i+1} - A^i||$, $k_1 = ||A^{i+1} - A^i||$, $k_2 = ||E - A^i||$, $\beta = ||F(A^{i+1}) - F(A^i)||$, $\delta_1 = ||F(A^{i+2}) - F(A^{i+1})||$, $\delta_2 = ||F(E) - F(A^i)||$, ω_1 and ω_2 are defined as in Fig. 3.2, and $f_x g_y - f_y g_x$ is evaluated at A^i. Now $0 < k_j \leq h$, since $\overline{A^iA^{i+1}}$ is the longest edge of $\Delta A^1A^2A^3$. Thus if (4.22) is satisfied, then

$$hk_j\,|f_x g_y - f_y g_x|\,|\sin \omega_j| > \frac{hk_j}{2}(h + k_j)pr + \frac{h^2 k_j^2 r^2}{4}. \tag{4.24}$$

Consequently (4.23) implies that $\sin \varphi_1 > 0$, $\sin \varphi_2 < 0$, i.e. that $F(E)$ and $F(A^{i+2})$ are on opposite sides of the straight line through $F(A^i)$ and $F(A^{i+1})$.

Thus if $\theta \in T_F A^1A^2A^3$, but θ is in neither of $T_F A^iDA^{i+2}$ nor in $T_F DA^{i+1}A^{i+2}$, then we must have the situation in Fig. 4.2 where, by Lemma 4.5, $\theta \in S_{A^iA^{i+1}}$. Furthermore, $S_{A^iA^{i+1}}$ then lies in the interior of the two triangles $T_F A^iA^{i+1}A^{i+2}$ and $T_F A^iEA^{i+1}$. That is, θ is contained in one of the triangles $T_F A^iED$ or $T_F DEA^{i+1}$.

LEMMA 4.7. *If* Assumptions 4.1 (i)–(vii) *are satisfied, then* $\theta \in T_F\ \Delta_I$ *for some triangle* Δ_I *in* $\mathfrak{D}$.

Proof of Lemma 4.7. Let $X(t)$ and $Y(t)$ be defined as in (4.4) and (4.6) respectively, where $X(t) \in \mathfrak{D}$ for $0 \leq i \leq 1$. Since $t(1 - t) \leq 1/4$ for $0 \leq t \leq 1$, (4.7) yields

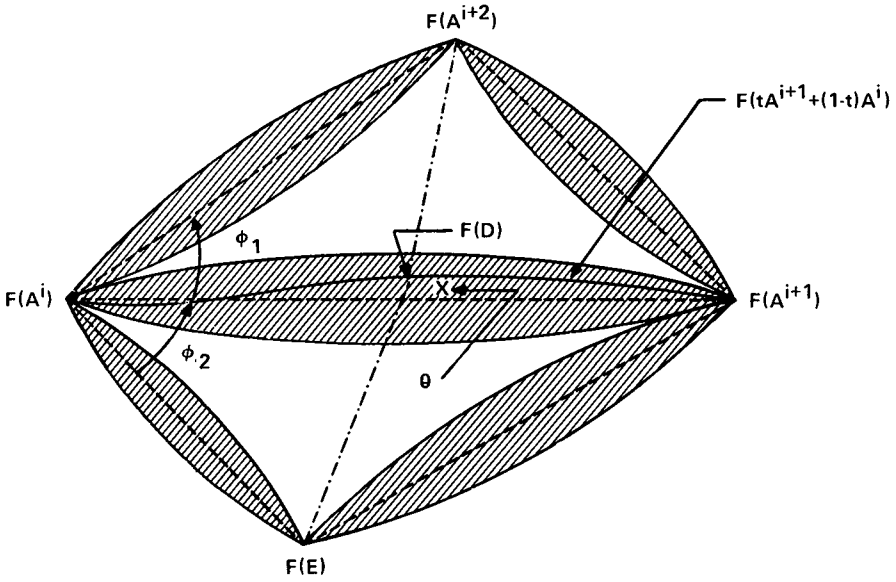

FIG. 4.2. The images of the boundaries of $\Delta A^1A^2A^3$ and of ΔA^iEA^{i+1}.

$$||F(X(t)) - Y(t)|| \leq \frac{h^2r}{8}. \tag{4.25}$$

Let us now consider only those Δ_I in $\bar{\mathfrak{D}}$ whose union is $\bar{\mathfrak{D}}$ and for which the intersection of any two distinct Δ_I has zero area. On every such $\Delta_I = \Delta ABC$ we define a linear map $G(X)$ such that $G(A) = F(A)$, $G(B) = F(B)$ and $G(C) = F(C)$. The resulting function G defined on all of $\bar{\mathfrak{D}}$ is continuous on $\bar{\mathfrak{D}}$. If h satisfies Assumption 4.1 (vii), namely, $h \leq (7d/r)^{1/2}$, then

$$||F(X) - G(X)|| \leq h^2r/8 < h^2r/7 \leq d \leq ||F(X) - \theta|| \tag{4.26}$$

for all $X \in P$. Hence by Roché's theorem (see e.g. [10]) $d(G, \mathfrak{D}, \theta) = d(F, \mathfrak{D}, \theta)$. However, by Assumption 4.1 (vi), $d(F, \mathfrak{D}, \theta) \neq 0$. Hence $d(G, \mathfrak{D}, \theta) \neq 0$. By Kronecker's theorem [9, p. 161], there exists a point $Z \in \mathfrak{D}$ such that $G(Z) = \theta$. That is, $Z \in \Delta_I$, for some $\Delta_I \subset \bar{\mathfrak{D}}$. This implies, however, by the linearity of G on Δ_I, that $\theta \in T_G\, \Delta_I = T_F\, \Delta_I$.

LEMMA 4.8. *Let the conditions of* Lemma 4.7 *be satisfied, and let* $\Delta_I \subset \bar{\mathfrak{D}}$ *be the triangle of* Lemma 4.7 *such that* $\theta \in T_F\, \Delta_I$. *Then*

$$\min_{X \in \Delta_I,\, Y \in P} ||X - Y|| \geqq h. \tag{4.27}$$

Proof of Lemma 4.8. Every point W of $T_F\, \Delta_I = T_F\, \Delta ABC$ may be uniquely represented in the form

$$W = \alpha F(A) + \beta F(B) + \gamma F(C), \tag{4.28}$$

where α, β and γ are nonnegative numbers such that $\alpha + \beta + \gamma = 1$.

Let X and Y be two points of $\bar{\mathfrak{D}}$ such that $||X - Y|| = k$. Let us set $X(t) = Y + tk(c, s)$, $0 \leq t \leq 1$, where $(c, s) = (X - Y)/||X - Y||$, and let us assume that the segment $\overline{XY}$ is in $\bar{\mathfrak{D}}$. Then

$$F(X) - F(Y) = k(c, s) \int_0^1 \begin{bmatrix} f_x(X(t)) & g_x(X(t)) \\ f_y(X(t)) & g_y(X(t)) \end{bmatrix} dt \tag{4.29}$$

so that, by taking the norm of each side,

$$||F(X) - F(Y)|| \leq kp. \tag{4.30}$$

Now let Z be a point on Δ_I which is nearest to P, and let Y be a point on P, such that

$$\min_{(X \in P)} ||Z - X|| = ||Z - Y|| = k. \tag{4.31}$$

By the triangle inequality,

$$||W|| = ||\alpha F(A) + \beta F(B) + \gamma F(C)|| \geq ||F(Z)|| - \{\alpha\, ||F(Z) - F(A)|| + \beta\, ||F(Z) - F(B)|| + \gamma\, ||F(Z) - F(C)||\}. \tag{4.32}$$

However, since Z, A, B and C are all on Δ_I and hence $||Z - A||$, $||Z - B||$, and $||Z - C||$ are bounded by h, we substitute (4.30) into (4.32) and use $||F(Z)|| \geq ||F(Y)|| - ||F(Z) - F(Y)|| \geq d - pk$, to get

$$||\alpha F(A) + \beta F(B) + \gamma F(C)|| \geq d - pk - (\alpha + \beta + \gamma)ph = d - p(h + k). \tag{4.33}$$

Since W given by (4.28) is an arbitrary point of $T_F\, \Delta_I$, it follows from (4.33) and the Assumption 4.1 (vii), $h \leq d/(2p)$, that if $k < h$, then $\theta \notin T_F\, \Delta_I$. Eq. (4.27) thus follows.

Completion of proof of Theorem 4.2. (a) Let us assume at the outset that Assumptions 4.1 (i)–(vii) are satisfied. We thus arrive at Step 5 of Algorithm 3.1 and by Lemma 4.7 we there find a triangle Δ_I such that $\theta \in T_F\, \Delta_I$. We thus arrive at Step 7.

We remark that since the interior angles of the initial triangles were $\geq \alpha$, then by Lemma 4.3, the interior angles of the resulting triangles obtained by repeated bisection are $\geq \alpha/2$.

In Step 7 we check whether or not the longest side of Δ_I is less than or equal to ϵ. If so, a printout of h_I, A, B and C follows, where $\Delta_I = \Delta ABC$. If not, we proceed to Step 6.

Now consider Fig. 4.3, in which the triangles ΔACF, ΔCBE, ΔADB and ΔABC are congruent. By Lemma 4.6 it follows that $T_F ABC \subset F(\Delta ABC) \cup F(\Delta ACF) \cup F(\Delta CBE) \cup F(\Delta ADB)$, where e.g. $F(\Delta ABC) = \{Y = F(X)\colon X \in \Delta ABC\}$. Since $\theta \in T_F ABC$, it follows that there exists a point Ξ in one of the four triangles in Fig. 4.3 such that $F(\Xi) = \theta$. It follows from Lemma 4.8 that each of these triangles lies wholly in $\bar{\mathfrak{D}}$; moreover, from our construction, $\max\{||\Xi - A||, ||\Xi - B||, ||\Xi - C||\} \leq 2h \leq 2\epsilon$.

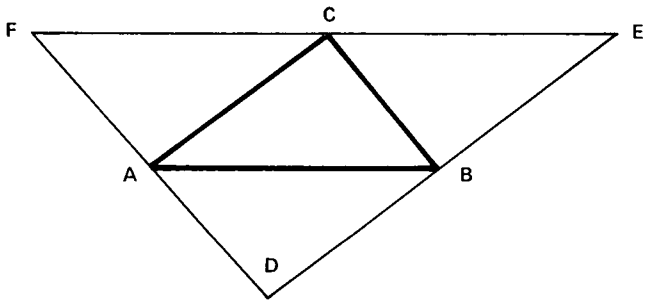

FIG. 4.3. ΔABC and its neighbors.

Let us now examine what happens in Step 6 in the case that printout did not occur in Step 7. Here we first bisect Δ_I and then check to see which of the new triangles $\Delta^{(i)}$ $(i = 1, 2)$ thus formed satisfies $\theta \in T_F \Delta^{(i)}$. If one of the range triangles $T_F \Delta^{(i)}$ $(i = 1, 2)$ does contain θ, we return to Step 7. If neither of these contains θ, we proceed to Step 8 and form two new triangles by locating the point E as described there. By Lemma 4.8 the two new triangles $\Delta^{(1)} = \Delta A^i ED$ and $\Delta^{(2)} = \Delta DEA^{i+1}$ formed in Step 8 lie wholly in $\mathfrak{D}$. Moreover, by Lemma 4.6, $\theta \in T_F \Delta^{(j)}$ for either $j = 1$ or $j = 2$. We thus proceed to Step 9 and then return to Step 7.

In all cases we therefore remain in Steps 6, 7, 8 and 9. At every bisection the longest side of a triangle is halved. Thus after a finite number of returns to Step 7, the test $h_I \leq \epsilon$ becomes satisfied, where h_I denotes the longest side of the triangle Δ_I such that $\theta \in T_F \Delta_I$.

(b) Let us now assume that only the Assumptions 4.1 (i)–(vi) are satisfied. In this case we either achieve convergence in Steps 6, 7, 8 and 9, or else we may branch to Step 10 from either Step 5, because θ is not contained in any $T_F \Delta_I$, from Step 8, because the new triangle, $\Delta A^i E A^{i+1}$ is not wholly in $\mathfrak{D}$, or from Step 9, because θ is in neither $T_F A^i ED$ nor in $T_F DEA^{i+1}$. However, each time we arrive at Step 10, the longest length h of the sides of each triangle in $\mathfrak{D}$ is halved, and since $\mathfrak{D}$,and hence p, d, j and r, are fixed (see Assumptions 4.1), unless convergence occurs first, the Assumption 4.1 (vii) becomes satisfied after arriving at Step 10 a finite number of times.

Remark 4.9. It is evident from the above proof, that after we reach Step 6 and h is sufficiently small, the number of times we need to evaluate F in Step 9 is small relative to the number of times we need to evaluate F in Step 6. If at the nth evaluation of F, we find that $\theta \in T_F A^1 A^2 A^3$, and we are still in Step 6 after two bisections of $\Delta A^1 A^2 A^3$ and two evaluations of F, the lengths of all the sides of the resulting triangle, $\Delta B^1 B^2 B^3$, such that $\theta \in T_F B^1 B^2 B^3$ are half of the lengths of those of $\Delta A^1 A^2 A^3$.

Hence if we traverse the route Steps 6–7–6–7–6:., the rate of convergence after n evaluations of F is $0(2^{-n/2})$ as $n \to \infty$. At worst, if we continually traverse the route Steps 6–8–9–7–6–8–9–7–6–8–9–7 etc. (an impossible occurrence, as is evident from the proofs of the preceding lemmas and theorem), the rate of convergence is $O(2^{-n/4})$ as $n \to \infty$.

References

[1] W. M. Kincaid, *A two-point method for the numerical solution of systems of simultaneous equations*, Quart. Appl. Math. **18**, 313–324 (1961)

[2] A. M. Ostrowski, *Solutions of equations and systems of equations*, Academic Press, N. Y. (1960)

[3] J. E. Dennis, *On the Kantorovich hypothesis for Newton's method*, SIAM J. Numer. Anal. **6**, 493–507 (1969)

[4] W. Rheinboldt, *Symposium on the numerical solution of nonlinear problems*, Philadelphia, Pa., 1968

[5] I. Rosenberg and F. Stenger, *A lower bound on the angles of triangles constructed by bisecting the longest side*, Math. Comp. **29**, 390–395 (1975)

[6] E. Goursat, *A course in mathematical analysis*, V. 1, *Applications to geometry, expansion in series, definite integrals derivatives and differentials.*

[7] M. J. D. Powell, *A method for minimizing a sum of squares of nonlinear functions without calculating derivatives*, Computer J. **7**, 303–307 (1965)

[8] F. Stenger, *Computing the topological degree of a mapping in n-space*, to appear in Numer. Math.

[9] J. M. Ortega and W. C. Rheinboldt, *Iterative solutions of nonlinear equations in several variables*, Academic Press, New York (1970)

[10] J. Mawhin, *Degré topologique et solutions périodiques des systèmes differentiels nonlinéaires*, Bull. Soc. Roy. Sciences de Liège **38**, 308–398 (1969)

[11] A. Eiger and F. Stenger, *A program of bisections for solving two nonlinear equations*, University of Utah, Department of Mathematics, to appear in Comm. ACM
[12] J. L. Kuester and J. H. Mize, *Optimization techniques with Fortran*, McGraw-Hill, 1973, 368–386
[13] D. A. Paviani, *A new method for the solution of a general nonlinear programming problem*, Ph. D. dissertation, The University of Texas, 1968