ON THE SOLUTION OF AN IMPLICIT EQUATION

BY

SMBAT ABIAN AND ARTHUR B. BROWN

In this paper a method of solving an implicit equation $g(x_1, \cdots, x_n; y) = 0$ by successive substitutions is given. The customary hypotheses that the function g be differentiable and that $g = 0$ at a given point are replaced by weaker hypotheses.

Appraisals of the remainder error are given, as well as a method of minimizing one of them. Three of the appraisals are valid regardless of miscalculations at earlier stages of the work.

It is also proved that if, in addition, the function g satisfies a Lipschitz condition in a subset of the x_i's, then the solution $y = Y(x_1, \cdots, x_n)$ will also satisfy a Lipschitz condition in the same subset.

In the statements and proofs which follow, unless otherwise specified the index i runs from 1 to n; $(x) \equiv (x_1, \cdots, x_n)$ and $(x; y) \equiv (x_1, \cdots, x_n; y)$. All independent variables and functions mentioned are understood to be real, and the functions single valued.

THEOREM 1. *Let $g(x_1, \cdots, x_n; y) \equiv g(x; y)$ bs a continuous function defined on the closed region $N_1 \subset E^{n+1}$ determined by*

$$|x_i - a_i| \leqq \alpha_{i1}, \qquad |y - b| \leqq \beta, \tag{1}$$

where α_{i1} and β are positive constants, and let there be constants C and D such that

$$|g(a; b)| < C\beta, \tag{2}$$

and, if $(x; u)$ and $(x; v)$ are any two distinct points of N_1,

$$0 < C \leqq \frac{g(x; u) - g(x; v)}{u - v} \leqq D. \tag{3}$$

Then there exist n positive constants $\alpha_i \leqq \alpha_{i1}$ and a continuous function $Y(x)$ such that, if T is the closed region determined by $|x_i - a_i| \leqq \alpha_i$, the locus of the equation $y = Y(x)$ for $x \in T$ is the same as that of $g(x; y) = 0$ for $(x; y) \in N$, where $N \subset N_1$ is the closed region determined by

$$|x_i - a_i| \leqq \alpha_i, \qquad |y - b| \leqq \beta.$$

We shall prove Theorem 1 simultaneously with Theorem 2.

THEOREM 2. *The constants α_i of Theorem 1 can be chosen subject only to the condition*

$$|g(x; b)| < C\beta, \qquad |x_i - a_i| \leqq \alpha_i. \tag{4}$$

Received February 11, 1958; received in revised form May 1, 1958.

If furthermore a constant k is chosen so that

$$0 < k[D\beta + |\, g(x; b)|] < 2\beta, \qquad x \in T, \tag{5}$$

and if we introduce

$$f(x; y) \equiv y - kg(x; y), \qquad (x; y) \in N_1, \tag{6}$$

and take $Y_0(x)$ as a function, not necessarily continuous, satisfying

$$|\, Y_0(x) - b\,| \leqq \beta, \qquad x \in T, \tag{7}$$

then

$$Y_{m+1}(x) = f[x; Y_m(x)] = Y_m(x) - kg[x; Y_m(x)], \qquad m \geqq 0, \tag{8}$$

is well defined for $x \in T$, and

$$Y(x) = \lim_{m\to\infty} Y_m(x). \tag{9}$$

Proof of Theorems 1 *and* 2. Since $g(x; y)$ is continuous, we see from (1) and (2) that there exist n positive constants $\alpha_i \leqq \alpha_{i1}$ such that (4) is valid for $|\, x_i - a_i\,| \leqq \alpha_i$, i.e., for $x \in T$. Since $D\beta > 0$, from (4) we see that relation (5) is satisfied by a suitably chosen constant k.

Let

$$R = \max\,(|\, 1 - kD\,|, |\, 1 - kC\,|). \tag{10}$$

From (3) and (5) we obtain $0 < kC \leqq kD < 2$, and therefore

$$0 \leqq R < 1. \tag{11}$$

Since $C \leqq D$, we infer from (10) that

$$R = 1 - kC \quad \text{or} \quad R = kD - 1, \tag{12}$$

and correspondingly from (4) or (5), we have

$$k\,|\, g(x; b)\,| < (1 - R)\beta, \qquad x \in T. \tag{13}$$

From (3) and (6), for $(x; u)$ and $(x; v) \in N_1$, with $u \neq v$, we obtain

$$1 - kD \leqq \frac{f(x; u) - f(x; v)}{u - v} \leqq 1 - kC,$$

and hence, by (10),

$$|\, f(x; u) - f(x; v)\,| \leqq R\,|\, u - v\,|, \tag{14}$$

a relation which is true also when $u = v$.

Let $N \subset N_1$ be defined as in the statement of Theorem 1. We now introduce (8) and prove inductively that, for $m \geqq 0$,

$$|\, Y_m(x) - b\,| \leqq \beta, \qquad x \in T. \tag{15}$$

From (7) we see that (15) is true for $m = 0$. Now let us assume that (15) is true for $m = j$, so that for $x \in T$ the point $[x; Y_j(x)] \in N$. This, in view of

(8), implies that $Y_{j+1}(x)$ is well defined for $x \in T$. From (6) and (13) we see that

$$|f(x; b) - b| < (1 - R)\beta, \qquad x \in T. \tag{16}$$

From (8) we obtain

$$|Y_{j+1}(x) - b| \leqq |f[x; Y_j(x)] - f(x; b)| + |f(x; b) - b|,$$

a relation which, in view of (14), (15) with $m = j$, and (16), implies (15) with $m = j + 1$. Hence we infer that for $x \in T$ and $m \geqq 0$, $Y_m(x)$ is well defined, and (15) holds, so that the point $[x; Y_m(x)] \in N$.

From (8) and (14), if $m \geqq 1$, we have

$$|Y_{m+1}(x) - Y_m(x)| \leqq R\,|Y_m(x) - Y_{m-1}(x)|, \qquad x \in T. \tag{17}$$

By applying (17) with $m = 1, 2, \cdots, s$ and then replacing s by m, we obtain, for $m \geqq 1$,

$$|Y_{m+1}(x) - Y_m(x)| \leqq R^m\,|Y_1(x) - Y_0(x)|, \qquad x \in T. \tag{18}$$

By (15) for $m = 0$ and $m = 1$, we obtain $|Y_1(x) - Y_0(x)| \leqq 2\beta$. Hence we see from (11) and (18) that the sequence $\{Y_m(x)\}$ is uniformly convergent for $x \in T$. Therefore, $Y(x)$ for $x \in T$ is well defined by (9). Moreover, from (9) and (15) we conclude that for $x \in T$, $|Y(x) - b| \leqq \beta$, and therefore the locus of $y = Y(x)$ for $x \in T$ is contained in N.

From (9) and (8), in view of the continuity of $f(x; y)$ on N, we have

$$Y(x) \equiv f[x; Y(x)], \qquad x \in T, \tag{19}$$

and from (6) we infer that

$$g[x; Y(x)] \equiv 0, \qquad x \in T. \tag{20}$$

Since, by (3), $g(x; y)$ is a strictly monotonic function of y for fixed x, we conclude that $y = Y(x)$ given by (9), for $x \in T$, gives the complete locus of the equation $g(x; y) = 0$ for $(x; y) \in N$.

It remains to prove that $Y(x)$ is continuous. For this purpose, consider the special case in which we begin with $Y_0(x) = b$. Then $Y_0(x)$ is continuous, and (7) is satisfied. Examination of the proof shows that $Y_m(x)$ is continuous for $x \in T$ and $m \geqq 0$. Since the sequence $\{Y_m(x)\}$ converges uniformly for $x \in T$, its limit is a continuous function. But the limit must be the unique function $Y(x)$ already obtained. Hence $Y(x)$ is continuous, and the proof of Theorems 1 and 2 is complete.

We now give some appraisals of the remainder error.

THEOREM 3. *For $x \in T$ and $m \geqq 1$,*

$$|Y_m(x) - Y(x)| \leqq \frac{R^m}{1 - R}\,|Y_1(x) - Y_0(x)|; \tag{21}$$

and the factor $R^m/(1 - R)$ *is minimized by taking*

$$k = 2/(D + C). \tag{22}$$

Furthermore,

$$|\, Y_m(x) - Y(x) \,| \leqq \frac{R}{1 - R} |\, Y_m(x) - Y_{m-1}(x) \,|; \tag{23}$$

$$|\, Y_m(x) - Y(x) \,| \leqq \frac{|\, g[x; Y_m(x)] \,|}{C}; \tag{24}$$

and, if (22) *holds, then*

$$|\, Y_m(x) - Y(x) \,| \leqq \frac{|\, g[x; Y_{m-1}(x)] \,| (D - C)}{C(D + C)}. \tag{25}$$

Moreover, relations (23), (24), (25) *are valid regardless of errors in calculation through* $Y_{m-1}(x)$ *for* (23) *and for* (25) *and through* $Y_m(x)$ *for* (24), *provided merely that* $|\, Y_m(x) - b \,| \leqq \beta$ *for* (24), *and that, for* (23) *and* (25), $|\, Y_{m-1}(x) - b \,| \leqq \beta$ *and* $Y_m(x)$ *is correctly calculated from* $Y_{m-1}(x)$.

Proof. Since

$$Y(x) - Y_m(x) = [Y_{m+1}(x) - Y_m(x)] + [Y_{m+2}(x) - Y_{m+1}(x)] + \cdots,$$

relation (21) follows from (9), (18), and the formula for the sum of a geometric series.

From (4) we see that the value of k given in (22) satisfies (5). From (10) and (12) we see that R, hence also $R^m/(1 - R)$, is minimized when $1 - kC = kD - 1$, so that (22) holds.

Relation (23) follows easily from (9), (17), and the formula for the sum of a geometric series.

Relation (24) is proved easily from (20) by taking $u = Y_m(x)$ and $v = Y(x)$ in (3).

If $g[x; Y_{m-1}(x)] = 0$, repeated application of (8) shows that each side of (25) is zero, so that (25) is true. If $g[x; Y_{m-1}(x)] \neq 0$, from (8) with m replaced by $m - 1$ we find

$$\frac{Y_m(x) - Y(x)}{g[x; Y_{m-1}(x)]} = \frac{Y_{m-1}(x) - Y(x)}{g[x; Y_{m-1}(x)]} - k.$$

From (3) and (20) we have

$$\frac{1}{D} \leqq \frac{Y_{m-1}(x) - Y(x)}{g[x; Y_{m-1}(x)]} \leqq \frac{1}{C}.$$

Hence

$$\frac{1}{D} - k \leqq \frac{Y_m(x) - Y(x)}{g[x; Y_{m-1}(x)]} \leqq \frac{1}{C} - k. \tag{26}$$

By (22), the left and right members of (26) are respectively

$$\frac{C - D}{D(D + C)} \quad \text{and} \quad \frac{D - C}{C(D + C)}.$$

Since $C \leqq D$, (25) then follows from (26).

To prove that (23) is valid regardless of earlier errors in calculation, we observe that the right member of (23) depends only on $Y_m(x)$ and $Y_{m-1}(x)$ and that, as we see by comparing (7) with the given relation

$$| Y_{m-1}(x) - b | \leqq \beta,$$

$Y_{m-1}(x)$ can be considered to be a new function $Y_0(x)$, in which case (21) with $m = 1$ gives (23). The corresponding proofs for (24) and (25) are similar.

THEOREM 4. *If* $Y_0(x) \equiv b$, $x \in T$, *then*

$$| Y_m(x) - Y(x) | \leqq \beta R^m. \tag{27}$$

The right member of this inequality is minimized by choosing k *as in* (22).

Proof. In view of (8), we have

$$Y_1(x) - Y_0(x) = f(x; b) - b. \tag{28}$$

Relation (27) now follows from (21), (28), and (16).

The final statement of Theorem 4 follows from the property, already proved, that R is minimized by taking k as in (22).

THEOREM 5. *Under the hypotheses of Theorem* 1, *and with the* α_i*'s chosen as in Theorem* 2, *if* $g(x; y)$ *satisfies a Lipschitz condition in a subset of the* x_i*'s, the function* $Y(x)$ *will also satisfy a Lipschitz condition in the same* x_i*'s.*

Proof. With $p \leqq n$ and $x_i = x_i'$ for $i > p$, suppose that if $(x; y)$ and $(x'; y) \in N$,

$$| g(x'; y) - g(x; y) | \leqq \textstyle\sum_{j=1}^{p} H_j | x_j' - x_j |,$$

where the H_j's are nonnegative constants. Since

$$| f[x', Y(x')] - f[x, Y(x)] |$$
$$\leqq | f[x', Y(x')] - f[x, Y(x')] | + | f[x, Y(x')] - f[x, Y(x)] |,$$

we then infer, in view of (6) and (14), that

$$| f[x'; Y(x')] - f[x; Y(x)] | \leqq \textstyle\sum_{j=1}^{p} kH_j | x_j' - x_j | + R | Y(x') - Y(x) |.$$

Using (19) we obtain

$$| Y(x') - Y(x) | \leqq \sum_{j=1}^{p} \frac{kH_j}{1 - R} | x_j' - x_j |,$$

and the proof is complete.

Remark. Given a function $g(x; y)$ satisfying a relation like (3) but with C and D both negative, we can obtain one satisfying (3) with positive C and D by introducing $G(x; y) = -g(x; y)$. We observe also that a relation like (3) does exist if $\partial g/\partial y$ is continuous and not zero at the point $(a; b)$.

The following extension of Theorems 1 and 2 can be used to obtain more rapid convergence.

THEOREM 6. *Let* $b(x_1, \cdots, x_n) \equiv b(x)$ *be a continuous function defined on the closed region* $P \subset E^n$ *determined by* $| x_i - a_i | \leqq \alpha_i$, *and let*

$$g(x_1, \cdots, x_n ; y) \equiv g(x; y)$$

be a continuous function defined on the closed region $N \subset E^{n+1}$ *determined by* $x \in P$, $| y - b(x) | \leqq \beta$, *where* α_i *and* β *are positive constants. Let, moreover, there be positive constants* C_1, D_1, *and functions* $C(x)$, $D(x)$, *not necessarily continuous, such that for* $x \in P$

$$| g[x; b(x)] | < \beta C(x),$$

and, if $(x; u)$, $(x; v)$ *are distinct points of* N,

$$C_1 \leqq C(x) \leqq \frac{g(x; u) - g(x; v)}{u - v} \leqq D(x) \leqq D_1.$$

Then there exists a function $Y(x)$ *continuous on* P, *such that the locus of the equation* $g(x; y) = 0$ *for* $(x; y) \in N$ *is the same as that of* $y = Y(x)$ *for* $x \in P$.

Furthermore, if $k(x)$ *is any function, not necessarily continuous, such that for* $x \in P$

$$0 < k(x)\{\beta D(x) + | g[x; b(x)] | \} < 2\beta, \tag{29}$$

and if

$$f(x; y) = y - k(x)g(x; y), \qquad (x; y) \in N,$$

and

$$Y_0(x) = f[x; b(x)],$$

then

$$Y_{m+1}(x) = f[x; Y_m(x)], \qquad m \geqq 0,$$

is well defined for $x \in P$, *and* $Y(x) = \lim_{m \to \infty} Y_m(x)$.

The proof is similar to that of Theorems 1 and 2, but one first requires $k(x)$ to be continuous and

$$k(x)D(x) \leqq E < 2, \tag{30}$$

in order to establish the continuity of the unique function $Y(x)$. However, examination of the proof shows that even if $k(x)$ is allowed to be discontinuous, and satisfy only (29) rather than (29) and (30), the sequence of functions $\{Y_m(x)\}$ is convergent for each fixed x, from which, in view of the uniqueness of the continuous solution $y = Y(x)$, the truth of the theorem follows.

Theorems 3 and 5 remain valid with R, C, D, k replaced by $R(x)$, $C(x)$, $D(x)$, $k(x)$. Their proofs present no special difficulties.

The results above are easily applied to the problem of solving an equation $G(y) = 0$ in one unknown, and can be used when Newton's method is not

applicable, since $G(y)$ is not required to be differentiable. It is clear that with x unrestricted, the hypotheses of Theorem 1 will be satisfied if $G(y)$ satisfies corresponding hypotheses. We assume that $G(b) \neq 0$ and can thus introduce an equality sign in (5). The following theorem holds.

THEOREM 7. *Let $G(y)$ be a continuous function defined on the interval I determined by $| y - b | \leqq \beta$, where β is a positive constant; and let there be positive constants C and D such that*

$$0 < | G(b) | < C\beta,$$

and, if u, v are distinct points of I,

$$C \leqq \frac{G(u) - G(v)}{u - v} \leqq D.$$

Then there exists a unique solution $Y \in I$ of the equation $G(y) = 0$. Furthermore, if k is a constant satisfying

$$0 < k[\beta D + | G(b) |] \leqq 2\beta,$$

and if

$$f(y) = y - kG(y), \qquad Y_0 = f(b),$$

then

$$Y_{m+1} = f(Y_m), \qquad m \geqq 0,$$

is well defined, and $Y = \lim_{m\to\infty} Y_m$.

The appraisals of the remainder error given in Theorems 3 and 4 remain valid. Also, we note that if (22) holds and $D < 3C$, then $\beta R^m < \beta/2^m$, so that the appraisal (27) is more favorable than that obtained by the method of taking successive midpoints of intervals at the endpoints of which $G(y)$ has opposite signs.

References. For related ideas in a more general setting, cf. *Implicit functions and their differentials in general analysis* by T. H. HILDEBRANDT and LAWRENCE M. GRAVES, Trans. Amer. Math. Soc., vol. 29 (1927), pp. 127–153. For the solution of $G(y) = 0$, cf. Chapter II of *Numerical calculus* by W. E. MILNE, Princeton, Princeton University Press, 1949, or *Practical analysis* by F. A. WILLERS, New York, Dover Publications, 1948, p. 209.

QUEENS COLLEGE
FLUSHING, NEW YORK