ON THE SOLUTION OF SIMULTANEOUS IMPLICIT EQUATIONS

ON THE SOLUTION OF SIMULTANEOUS IMPLICIT EQUATIONS

by Smbat Abian and Arthur B. Brown, Flushing, N.Y.

(Reçu le 2 septembre 1958.)

In this self-contained paper, generalizing the results obtained in an earlier paper [1] on the case of a single implicit equation, the authors give an explicit method for solving a system of p simultaneous implicit equations $f_i(x_1, \ldots, x_n, y_1, \ldots, y_p) = 0$ for the p unknown functions $y_i = Y_i(x)$. The method consists of successive substitutions.

The hypotheses of the classical implicit function theorem are replaced by weaker hypotheses. In particular, the functions f_i are not required to be differentiable, and there is no requirement that a known point satisfy the given equations.

Two appraisals of the remainder error at the mth stage of approximation are given, one of which is valid regardless of errors made at earlier stages of the computation. It is also proved that if the given functions f_i satisfy Lipschitz conditions in a certain subset of the x's, then the $Y_i(x)$ will also satisfy Lipschitz conditions in the same subset.

Throughout the paper, unless otherwise specified, the indices i, j, k run from 1 to p, the index r runs from 1 to n, $(x) \equiv (x_1, \ldots, x_n)$ and $(y) \equiv (y_1, \ldots, y_p)$. All functions and variables are understood to be real, and the functions singlevalued.

Theorem 1. Given a set of p functions $f_i(x_1, \ldots, x_n, y_1, \ldots, y_p) \equiv f_i(x, y)$ continuous on the closed region $N_1 \subset E^{n+p}$ determined by the relations $| x_r - a_r | \leqq \alpha_{r1}$, $| y_i - b_i | \leqq \beta_{i1}$, where α_{r1}, β_{i1} are positive constants, let there exist a non-singular matrix of constants (C_{ij}) and a matrix of constants (D_{ij}) with

$$\sum_j D_{ij} < 1 , \qquad (1)$$

such that, for $(x, y) \in N_1$,

$$\left| \delta_{ij} \Delta y_j + \sum_k C_{ik} \Delta_j f_k \right| \leqq D_{ij} \left| \Delta y_j \right| , \tag{2}$$

where δ_{ij} is the Kronecker δ, Δy_j is an increment of the variable y_j and $\Delta_j f_k$ is the increment of the function f_k corresponding to the increment Δy_j of y_j.

Then there exist p positive constants $\beta_i \leqq \beta_{i1}$ such that

$$\beta_i - \sum_j D_{ij} \beta_j > 0 . \tag{3}$$

If furthermore $f_i(a, b) = f_i(a_1, ..., a_n, b_1, ..., b_p)$ satisfy

$$\left| \sum_k C_{ik} f_k(a, b) \right| < \beta_i - \sum_j D_{ij} \beta_j , \tag{4}$$

then there exist n positive constants $\alpha_r \leqq \alpha_{r1}$ and a set of p continuous functions $Y_i(x)$ such that if T is the closed region of x-space determined by $|x_r - a_r| \leqq \alpha_r$, the locus of the system of equations $y_i = Y_i(x)$ for $x \in T$ is the same as that of the system $f_i(x, y) = 0$ for $(x, y) \in N$, where $N \subset N_1$ is the closed region determined by

$$|x_r - a_r| \leqq \alpha_r , \quad |y_i - b_i| \leqq \beta_i .$$

We shall prove Theorem 1 simultaneously with Theorem 2.

Theorem 2. The constants α_r of Theorem 1 can be chosen subject only to the conditions

$$\left| \sum_k C_{ik} f_k(x, b) \right| \leqq \beta_i - \sum_j D_{ij} \beta_j , \quad |x_r - a_r| \leqq \alpha_r . \tag{5}$$

Furthermore if we introduce

$$F_i(x, y) \equiv y_i + \sum_k C_{ik} f_k(x, y) , \quad (x, y) \in N_1 , \tag{6}$$

and take $Y_i(x; 0)$ as a function, not necessarily continuous, satisfying

$$|Y_i(x; 0) - b_i| \leqq \beta_i , \quad x \in T , \tag{7}$$

then for $m \geqq 0$ the function

$$\mathrm{Y}_i(x; m+1) = \mathrm{F}_i[x, \mathrm{Y}(x; m)], \tag{8}$$

is well defined for $x \in \mathrm{T}$ and

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

Proof of Theorems 1 and 2. Before beginning the actual proof, we observe that a natural choice for $\mathrm{Y}_i(x; 0)$ is $\mathrm{Y}_i(x; 0) = b_i$. (Cf. Theorem 4.) We observe also that condition (2) is readily satisfied if $f_i(x, y)$ is of class C^1 and the Jacobian of the partial derivatives of the $f_i(x, y)$ with respect to the y_j is not zero at (a, b). For in that case the matrix equation

$$(\delta_{ij}) + (\mathrm{C}_{ik})\left(\frac{\partial f_k}{\partial y_j}\right) = 0, \quad (x, y) = (a, b), \tag{10}$$

is solvable for (C_{ik}), and it follows that if every D_{ij} is a positive constant, (2) will hold if N_1 is taken as a sufficiently small neighborhood of (a, b). From (10) we infer that (C_{ik}), so obtained, is non-singular.

Returning now to the actual proof, we first observe that, in view of (1), relations (3) are easily satisfied, for example by taking $\beta_i = \min\limits_j (\beta_{j1})$. We now assume that the β_i have been so chosen and that (4) is satisfied.

Since the f_i are continuous, we see from (4) that constants $\alpha_r \leqq \alpha_{r1}$ can be chosen so that (5) is satisfied. We assume that such constants α_r have been chosen.

Let $\mathrm{N} \subset \mathrm{N}_1$ be defined as in the statement of Theorem 1. If (x, y) and $(x, z) \in \mathrm{N}$, from (6) we obtain

$$\mathrm{F}_i(x, z) - \mathrm{F}_i(x, y) = z_i - y_i + \sum_k \mathrm{C}_{ik}[f_k(x, z) - f_k(x, y)] =$$
$$= \sum_j \delta_{ij}\,\Delta y_j + \sum_j \sum_k \mathrm{C}_{ik}\,\Delta_j f_k.$$

Hence, in view of (2), we infer that

$$\left| \mathrm{F}_i(x, z) - \mathrm{F}_i(x, y) \right| \leqq \sum_j \mathrm{D}_{ij} \left| z_j - y_j \right|, \tag{11}$$

for (x, y) and (x, z) belonging to N.

We now introduce (8) and prove inductively that, for $m \geqq 0$, $Y_i(x; m)$ is well defined, and

$$|Y_i(x; m) - b_i| \leqq \beta_i, \quad x \in T. \tag{12}$$

From (7) we see that (12) is true for $m = 0$. Now let us assume that (12) is true for $m = s$, so that for $x \in T$ the point $[x, Y(x; s)] \in N$. This, in view of (8), implies that $Y_i(x; s + 1)$ is well defined for $x \in T$. From (6) and (5) we see that

$$|F_i(x, b) - b_i| \leqq \beta_i - \sum_j D_{ij} \beta_j, \quad x \in T. \tag{13}$$

From (8) we obtain

$$|Y_i(x; s + 1) - b_i| \leqq |F_i[x, Y(x; s)] - F_i(x, b)| + |F_i(x, b) - b_i| ,$$

a relation which, in view of (11), (12) with $m = s$ and (13), implies (12) with $m = s + 1$. Hence we infer that for $x \in T$ and $m \geqq 0$, $Y_i(x; m)$ is well defined, and (12) holds, so that the point $[x, Y(x; m)] \in N$.

From (8) and (11), if $m \geqq 1$, we have for $x \in T$

$$\left| Y_i(x; m + 1) - Y_i(x; m) \right| \leqq \sum_j D_{ij} \left| Y_j(x; m) - Y_j(x; m - 1) \right|. \tag{14}$$

Let

$$D = \max_i \left(\sum_j D_{ij} \right). \tag{15}$$

From (1) and (2) we see that

$$0 \leqq D < 1. \tag{16}$$

From (14) and (15) we infer that, for $m \geqq 1$ and $x \in T$,

$$\left[\max_i |Y_i(x; m + 1) - Y_i(x; m)|\right] \leqq D \left[\max_j |Y_j(x; m) - Y_j(x; m - 1)|\right]. \tag{17}$$

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

$$\left| Y_i(x; m + 1) - Y_i(x; m) \right| \leqq D^m \left[\max_j \left| Y_j(x; 1) - Y_j(x; 0) \right| \right]. \tag{18}$$

For $x \in T$, the bracket on the right is bounded by $2 \max_j (\beta_j)$. Thus, in view of (16) and (18), the sequence $\{ Y_i (x; m) \}$, $x \in T$, is uniformly convergent for each i. Hence $Y_i (x)$, as defined in (9), exists. Moreover, from (9) and (12) we conclude that, for $x \in T$, $| Y_i (x) - b_i | \leqq \beta_i$, and therefore the locus $y_i = Y_i (x)$ is contained in N.

From (9) and (8), in view of the continuity of $F_i (x, y)$ on N, we see that

$$Y_i (x) \equiv F_i [x, Y (x)] , \quad x \in T . \tag{19}$$

Since (C_{ik}) is non-singular, we then infer from (6) that

$$f_i [x, Y (x)] \equiv 0 , \quad x \in T . \tag{20}$$

We thus see that the locus of the system of equations $y_i = Y_i (x)$ is contained in the locus of the system of equations $f_i (x, y) = 0$, for $(x, y) \in N$.

Next we prove that, for $x \in T$, $y_i = Y_i (x)$, given by (9), gives the complete locus of the system of equations $f_i (x, y) = 0$ for $(x, y) \in N$. Suppose that $f_i (\xi, \eta) = 0$ with $(\xi, \eta) \in N$. From (6) we infer that

$$\eta_i = F_i (\xi, \eta) . \tag{21}$$

From (19), (21) and (11) we have

$$\left| \eta_i - Y_i (\xi) \right| \leqq \sum_j D_{ij} \left| \eta_j - Y_j (\xi) \right| ,$$

and from (15) we further infer that

$$\left[\max_i | \eta_i - Y_i (\xi) | \right] \leqq D \left[\max_i | \eta_i - Y_i (\xi) | \right] .$$

In view of (16) we now infer that $\eta_i - Y_i (\xi) = 0$, so that $\eta_i = Y_i (\xi)$. We thus conclude that $y_i = Y_i (x)$ for $x \in T$ gives the complete locus of the system of equations $f_i (x, y) = 0$ for $(x, y) \in N$.

It remains only to prove that $Y_i (x)$ is continuous. For this purpose, take $Y_i (x; 0) = b_i$, which satisfies (7) and makes $Y_i (x; 0)$ continuous. Examination of the above proof then shows that $Y_i (x; m)$ is continuous for $m \geqq 0$. Since the sequence $\{Y_i(x; m)\}$ has been proved to be uniformly convergent for each i, we infer that $\left[\lim_{m\to\infty} Y_i(x; m)\right]$ is continuous. But we have already shown that for each $x \in T$ there is a set of uniquely determined values $Y_i(x)$ with $|Y_i(x) - b_i| \leqq \beta_i$, and satisfying (20). Hence the functions $Y_i(x)$ given by (9) are continuous, and the proof is complete.

We now give two appraisals of the remainder error.

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

$$|Y_i(x; m) - Y_i(x)| \leqq \frac{D^m}{1 - D}\left[\max_j |Y_j(x; 1) - Y_j(x; 0)|\right], \tag{22}$$

$$|Y_i(x; m) - Y_i(x)| \leqq \frac{D}{1 - D}\left[\max_j |Y_j(x; m) - Y_j(x; m - 1)|\right]. \tag{23}$$

Moreover, relation (23) is valid regardless of errors in computation through the $Y_i(x; m - 1)$, provided merely that $|Y_i(x; m - 1) - b_i| \leqq \beta_i$ and that $[Y(x; m)]$ is calculated correctly from $[Y(x; m - 1)]$.

Proof. Since $Y_i(x) - Y_i(x; m) = [Y_i(x; m + 1) - Y_i(x; m)] + [Y_i(x; m + 2) - Y_i(x; m + 1)] + \ldots$, relation (22) follows from (9), (16), (18) and the formula for the sum of a geometric series.

By comparing the given relation $|Y_i(x; m - 1) - b_i| \leqq \beta_i$ with (7), we see that $[Y(x; m - 1)]$ can be considered to be a new $[Y(x, 0)]$. If we apply (22) with $m = 1$ and this new $[Y(x; 0)]$, we obtain (23).

The proof given in the preceding paragraph makes clear the truth of the final assertion of Theorem 3.

We observe that this same procedure of considering $[Y(x; m - 1)]$ to be a new $[Y(x; 0)]$ shows that a finite number of errors of calculation will not prevent the sequence $\{Y_i(x, m)\}$ from converging to the function $Y_i(x)$.

Theorem 4. If $Y_i(x; 0) = b_i$, $x \in T$, then, for $x \in T$ and $m \geqq 1$,

$$|Y_i(x; m) - Y_i(x)| \leqq \frac{D^m}{1 - D}\left[\max_k \left(\beta_k - \sum_j D_{kj}\,\beta_j\right)\right]. \tag{24}$$

Proof. With $Y_i(x; 0) = b_i$, we have, by (8), for $x \in T$,

$$Y_i(x; 1) - Y_i(x; 0) = F_i(x; b) - b_i, \tag{25}$$

Relation (24) now follows from (22), (25) and (13). This completes the proof.

Theorem 5. Under the hypotheses of Theorem 1, and with the α_r's chosen as in Theorem 2, if the $f_i(x, y)$ satisfy Lipschitz conditions in a subset of the x_r's, the functions $Y_i(x)$ will also satisfy Lipschitz conditions in this same subset.

Proof. With $q \leqq n$ and $x_t = \xi_t$ for $t > q$, suppose that, if (x, y) and $(\xi, y) \in N$,

$$|f_i(\xi, y) - f_i(x, y)| \leqq \sum_{t=1}^{q} H_{it} |\xi_t - x_t|, \tag{26}$$

where the H_{it}'s are non-negative constants. Since

$$|F_i[\xi, Y(\xi)] - F_i[x, Y(x)]| \leqq |F_i[\xi, Y(\xi)] - F_i[x, Y(\xi)]| + \\ + |F_i[x, Y(\xi)] - F_i[x, Y(x)]|,$$

we infer from (6), (26) and (11) that

$$|F_i[\xi, Y(\xi)] - F_i[x, Y(x)]| \leqq \sum_k |C_{ik}| \sum_{t=1}^{q} H_{kt} |\xi_t - x_t| + \\ + \sum_j D_{ij} |Y_j(\xi) - Y_j(x)|. \tag{27}$$

From (27), (19) and (15), and letting $\gamma_t = \max_i \left(\sum_k |C_{ik}| H_{kt}\right)$, we obtain

$$|Y_i(\xi) - Y_i(x)| \leqq \sum_{t=1}^{q} \gamma_t |\xi_t - x_t| + D\left[\max_j |Y_j(\xi) - Y_j(x)|\right].$$

Therefore

$$|Y_i(\xi) - Y_i(x)| \leqq \sum_{t=1}^{q} \frac{\gamma_t}{1 - D} |\xi_t - x_t|.$$

Hence the theorem is true.

The results above are easily applied to the problem of solving p equations $g_i(y_1, \ldots, y_p) = 0$ in p unknowns, considered as a special case of the system $f_i(x, y) = 0$ in which the f_i are independent of x. In this case the functions $Y_i(x)$ become constants Y_i. The following theorem corresponds to Theorems 1 and 2.

Theorem 6. Given the functions $g_i(y_1, \ldots, y_p) \equiv g_i(y)$ continuous on the closed region $N_1 \subset E^p$ determined by the relations $|y_i - b_i| \leqq \beta_{i1}$, where the β_{i1} are positive constants, let there exist a non-singular matrix of constants (C_{ij}) and a matrix of constants (D_{ij}) with $\sum_j D_{ij} < 1$, and such that, for $y \in N_1$,

$$\left| \delta_{ij} \Delta y_j + \sum_k C_{ik} \Delta_j f_k \right| \leqq D_{ij} \left| \Delta y_j \right| .$$

Then there exist p positive constants $\beta_i \leqq \beta_{i1}$ such that $\beta_i - \sum_j D_{ij} \beta_j > 0$. If furthermore the quantities $g_k(b) = g_k(b_1, \ldots, b_p)$ satisfy

$$\left| \sum_k C_{ik} g_k(b) \right| < \beta_i - \sum_j D_{ij} \beta_j ,$$

then the system of simultaneous equations $g_i(y) = 0$ has a unique solution $y_i = Y_i$ in the closed region $N \subset N_1$ determined by $|y_i - b_i| \leqq \beta_i$.

Moreover, if for $y \in N_1$ we define $G_i(y) = y_i + \sum_k C_{ik} g_k(y)$, and if $Y_i(0)$ is any constant satisfying $|Y_i(0) - b_i| \leqq \beta_i$, then for $m \geqq 0$ the constants $Y_i(m+1) = G_i[Y(m)]$ are well defined, and $Y = \lim_{m \to \infty} Y_i(m)$.

The appraisals of the remainder error given in Theorems 3 and 4 remain valid.

REFERENCES

1. S. Abian and A. B. Brown, On the Solution of an Implicit Equation. *Illinois Journal of Mathematics.* (Accepted for publication.)
2. T. H. Hilderbrandt and L. M. Graves, Implicit Functions and their Differentials in General Analysis. *Trans. Amer. Math. Soc.*, Vol. 29 (1927), pp. 127-153.