AN APPLICATION OF THE METHOD OF STEEPEST DESCENTS TO THE SOLUTION OF SYSTEMS OF NON-LINEAR SIMULTANEOUS EQUATIONS

By A. D. BOOTH

(*Electronic Computer Project, Birkbeck College, London*)

[Received 2 November 1948]

SUMMARY

The solution of non-linear simultaneous equations using the 'method of steepest descents' is discussed. A comparison is made with the Southwell and Synge procedures for the solution of a set of equations, and it is concluded that the proposed method requires only about $1/n$ times the number of 'steps' to reach a solution. It is emphasized, however, that in view of the rather complicated expressions which must be calculated, it is best suited for use in conjunction with a high-speed electronic computer.

Finally a method is given for discriminating between real and false solutions in the case when the problem is one of minimization rather than of obtaining an exact solution.

If
$$\phi_j(Z_1,\ldots,Z_N) = 0 \quad (j = 1,\ldots,M) \tag{1}$$
is a set of M simultaneous equations, in the N unknowns (Z_r), a 'solution' will be defined to be any set of (Z_r) for which
$$\Phi = \sum_{j=1}^{M} \phi_j \bar{\phi}_j \tag{2}$$
is a minimum, $\bar{\phi}_j$ being the complex conjugate of ϕ_j. This definition is convenient in that it covers, not only the case $M = N$ to which the classical definition of solution applies, but also the cases $M < N$, $M > N$ where there may be either an infinity of solutions or no solution at all in the classical sense.

Suppose now that
$$\begin{aligned} \phi_j &= F_j + if_j, \\ \phi_j \bar{\phi}_j &= F_j^2 + f_j^2 \end{aligned} \tag{3}$$
so that Φ is everywhere positive. The problem is now to minimize Φ with respect to the $2N$ variables (X_r, Y_r), where
$$Z_r = X_r + iY_r. \tag{4}$$
For convenience of notation Φ will be defined by the relations
$$\Phi = \sum_{j=1}^{2M} \psi_j^2, \tag{5}$$
where:
$$\begin{aligned} \psi_{2j} &= F_j(x_1,\ldots,x_{2N}), \\ \psi_{2j-1} &= f_j(x_1,\ldots,x_{2N}), \end{aligned} \tag{6}$$
$$\begin{aligned} X_r &= x_{2r-1} \\ Y_r &= x_{2r}. \end{aligned} \tag{7}$$

In order to effect the minimization all that is needed is some systematic procedure, such that if any point $P_0(x_1,\ldots,x_{2N})$ is given in the $2N$ dimensional space defined by the (x_r), and if Φ is not a minimum at this point, then the procedure makes possible the passage to another point P_1 of the space for which Φ is less. Repeated application of the process will then lead to the minimization of Φ.

It is helpful, at this point, to consider a two-dimensional case of the above analysis, since here a graphical interpretation is possible which renders the argument clearer and illustrates various advantages and defects of the different techniques.

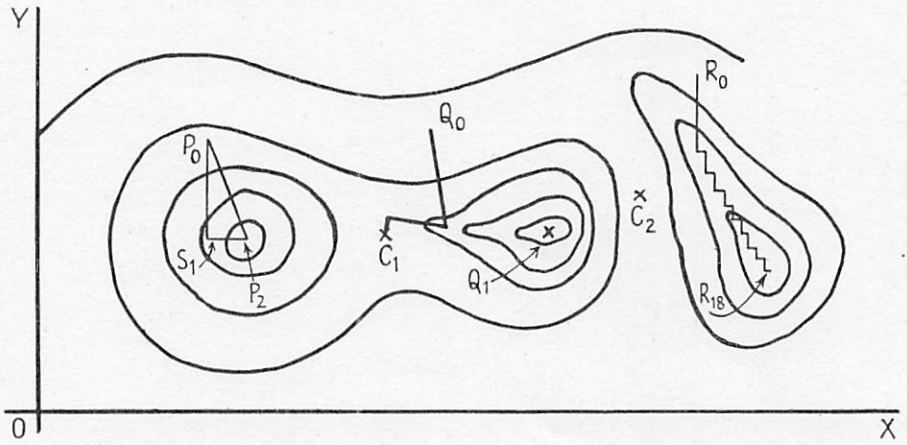

FIG. 1.

Fig. 1 is a contour map of a part of the function

$$\Phi = \Phi(x, y).$$

It has minima at P_2, Q_1, R_{18} and cols at C_1 and C_2. Suppose first that P_0 is chosen as the initial point; then to reach the nearby minimum P_2 two procedures are available:

1. Determine the direction from P_0 in which Φ, considered as a function of that axial variable (i.e. y), decreases most rapidly. Proceed in this direction until Φ is a minimum. In the figure this end-point is represented by S_1. Having arrived at S_1, repeat the process to obtain $S_2,\ldots, S_n$. The process evidently converges, in the case drawn, to the minimum P_2.

2. Determine the direction of steepest descent of Φ, considered as a function of x and y. Proceed along this direction until the function reaches a minimum. Iterate until the minimum is reached. Again the process converges to P_2. These two methods are typical of (1) the Southwell relaxation technique (**1, 2, 3**) and (2) the steepest descent technique (**4**).

The relative merits of the two methods in various typical situations will now be examined. In the case of a region, such as that about P_2 in Fig. 1, it is seen that both methods converge fairly rapidly but that, on the average, (1) will require about twice as many steps as (2). For a long valley, as shown about R_{18}, the steepest descent from R_0 would reach R_{18} in about three steps, whereas the relaxation technique would converge very slowly, the approximate drawing suggesting the need for about eighteen steps. When a col is present, as at C_1 or C_2, the situation is more complex. Either (1) or (2), as described above, converge to a *turning-point* of Φ rather than a minimum; so that starting from a point Q_0 in the neighbourhood of C_1 either (1) or (2) would converge to C_1 rather than to the true minima P_2, Q_1, or R_{18}. Similarly, if true maxima of Φ are present in the region, the processes may converge to these. The above intuitive arguments suggest that the steepest descent technique is the best for the purpose in hand, but that some modification is required to guard against ascent to maxima or stabilization in cols. In essence the method, to be derived analytically later, consists in making a parabolic approximation to the curve of Φ considered as (1) a function of (x_r) or (2) a function of (n), the distance along a normal to Φ = const., the normal to Φ = const. being, as is well known from vector analysis, the direction of steepest descent. In the first case,

$$\Phi = \Phi(0)+\epsilon_r\Phi_r+\frac{\epsilon_r^2}{2!}\Phi_{r,r} \tag{8}$$

and in the second,

$$\Phi = \Phi(0)+\epsilon_n\Phi_n+\frac{\epsilon_n^2}{2!}\Phi_{n,n}, \tag{9}$$

where $\Phi(0)$ is the value of Φ at the initial point,

$$\Phi_r = (\partial\Phi/\partial x_r)_0, \tag{10}$$

$$\Phi_{r,s} = (\partial^2\Phi/\partial x_r\,\partial x_s)_0, \tag{11}$$

ϵ_r is an increment in x_r, and ϵ_n denotes a change of the coordinates in the direction of the normal. For the relaxation version Φ has to tend to a minimum, whence

$$d\Phi/d\epsilon_r = 0,$$

that is

$$\epsilon_r = -\Phi_r/\Phi_{r,r}, \tag{12}$$

and similarly for the steepest descent

$$\epsilon_n = -\Phi_n/\Phi_{n,n}. \tag{13}$$

It is easy to see how ascent to a maximum can occur. Suppose that the initial point is chosen between the point of inflexion P_2 and the maximum G (Fig. 2). In this case Φ_n is negative as also is Φ_{nn} (this is equally true for the relaxation version); thus (13) gives a negative value for ϵ_n which means that P_G moves up the curve towards G. In the reverse situation, when P

is at P_L between P_I and the minimum L, Φ_n is again negative, but this time Φ_{nn} is positive making ϵ_n positive so that Φ descends towards L. Another fact which becomes obvious from the geometry of Fig. 2 is that if P is unfortunately chosen at P_I, ϵ will be infinite. These considerations suggest that, however excellent the relaxation or descent to a minimum techniques may be for homogeneous quadratic forms, where it is known

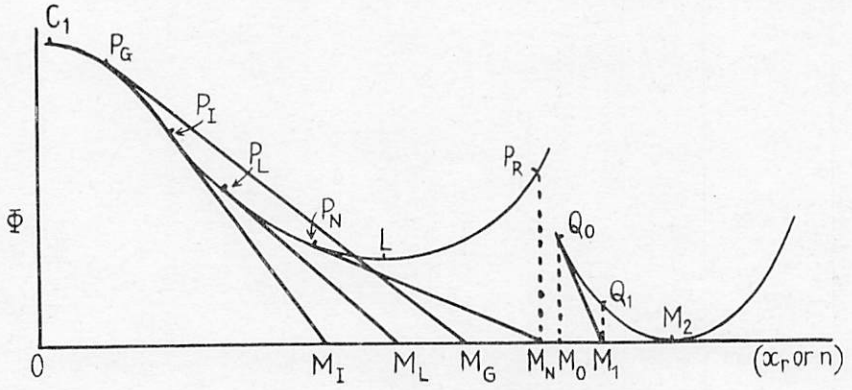

FIG. 2.

that only minima exist, they are not satisfactory in the initial stages of a determination of solutions for non-linear equations of the type given in (1).

Fortunately another method is available. Approximating Φ by a linear function of ϵ_n, (9) becomes

$$\Phi \doteq \Phi(0)+\Phi_n \epsilon_n. \tag{14}$$

Φ can be decreased by taking

$$\epsilon_n = -\Phi(0)/\Phi_n. \tag{15}$$

This corresponds to moving down the tangent to the (Φ, n) curve until it intersects the n-axis.

It is seen from Fig. 2 that for P_G, P_I, and P_L this technique leads to abscissae M_G, M_I, and M_L at which Φ is *less* than at the initial points. It is also seen that if the initial point is too near to the minimum, as at P_N, the resulting abscissa M_N may correspond to an increase in Φ. This, however, is no disadvantage since a simple working rule is available: *Use* (15) *until a refinement is reached which increases* Φ; *then change to the more accurate formula* (13). One other point is worthy of mention in connexion with the tangent descent. If the minimum value of Φ is actually zero as at M_2, Fig. 2, and the initial point is sufficiently close to M_2 for the curve to be considered parabolic, then it is easily shown that

$$M_0 M_1 = \tfrac{1}{2} M_0 M_2,$$

so that a more profitable estimate of ϵ_n than (15) is

$$\epsilon_n = -2\Phi(0)/\Phi_n, \tag{16}$$

and this should be used in preference to the latter. The amount $\Phi(0)-\Phi$ by which Φ is changed by the descent processes (12) and (13) is easily found by substitution into (8) and (9). For the relaxation method

$$\Phi(0)-\Phi = \Phi_n^2/2\Phi_{n,n}. \tag{17}$$

Synge (**3**) has suggested that instead of taking $\max(\Phi_r)$ as the direction of descent in the relaxation method, a more rapidly convergent process would be to take that direction for which the actual descent, given by (17), is a maximum. This version of the method involves the determination of the $\Phi_{r,r}$ which may well be a laborious process in complicated systems.

The disadvantage of the full steepest descent method, typified by (13), is that, as will be seen in the next section, it necessitates the calculation of all derivatives of the type $\Phi_{r,s}$ and requires a considerable amount of multiplication. On the other hand, axial descent methods of the type suggested by Southwell and Synge tend to become inefficient with a large number of variables. It was seen that for two dimensions, about two relaxations are, in most favourable cases, equivalent to one steepest descent; in N dimensions this number will be of the order N. With an electronic computer it would be preferable to use either the tangent descent method or the steepest descent to a minimum, since neither of these requires much discrimination on the part of the machine. For manual computation, however, it would seem that the Southwell relaxation process is the least laborious as it involves only the calculation of all Φ_r and one $\Phi_{r,r}$ and the discrimination of the largest of the Φ_r is easy.

A simpler method than any of the above, however, is the following:

1. Determine the values of Φ at the point given by applying the correction of equation (15), $\phi(1)$ say, and at the point given by applying half these corrections, $\phi(\frac{1}{2})$.

2. Using ordinary quadratic interpolation (or extrapolation if appropriate) on the values $\phi(0)$, $\phi(\frac{1}{2})$, $\phi(1)$, find the distance along the normal to the minimum.

3. Repeat process using this point as origin.

It will be seen later (equation (25)) that the above process requires the calculation of only one set of first derivatives and values of Φ at three points per refinement.

The normal derivatives required for the steepest descent process will

now be evaluated. From generalized vector analysis

$$\Phi_n = \partial\Phi/\partial n = |\mathrm{grad}\,\Phi| = \Big\{\sum_{r=1}^{2N}(\Phi_r)^2\Big\}^{\frac{1}{2}}, \tag{18}$$

$$\cos(n, x_r) = \Phi_r/\Phi_n, \tag{19}$$

where (n, x_r) is the angle between the normal, n, and the x_r-axis.

Again
$$\Phi_{n,n} = \partial(\Phi_n)/\partial n = \sum_{r=1}^{2N}\frac{\partial}{\partial x_r}(\Phi_n)\frac{\partial x_r}{\partial n}$$
$$= \Big[\sum_{r,s=1}^{2N}\Phi_{r,s}\Phi_r\Phi_s\Big]\Big/\Big[\sum_{r=1}^{2N}(\Phi_r)^2\Big]. \tag{20}$$

From (18), (20), and (13) it follows that

$$\epsilon_n = -\Big[\sum_{r=1}^{2N}(\Phi_r)^2\Big]^{\frac{3}{2}}\Big/\Big[\sum_{r,s=1}^{2N}\Phi_{r,s}\Phi_r\Phi_s\Big], \tag{21}$$

and consequently, the changes, ϵ_r, in *individual coordinates* (x_r), for steepest descent are given by
$$\epsilon_r = \epsilon_n\cos(n, x_r)$$
$$= -\Big[\sum_{r=1}^{2N}(\Phi_r)^2\Big]\Phi_r\Big/\Big[\sum_{r,s=1}^{2N}\Phi_{r,s}\Phi_r\Phi_s\Big]. \tag{22}$$

Similarly the steepest descent along the tangent to the (Φ, n) curve gives

$$\epsilon_n = -\Phi(0)\Big/\Big[\sum_{r=1}^{2N}(\Phi_r)^2\Big]^{\frac{1}{2}}, \tag{23}$$

from which the changes in individual coordinates are seen to be

$$\epsilon_r = -\Phi(0)\Phi_r\Big/\Big[\sum_{r=1}^{2N}(\Phi_r)^2\Big]. \tag{24}$$

In the interpolative version of the method the values of ϵ_r are taken to be

$$\epsilon_r = -\frac{[\Phi(1)-4\Phi(\frac{1}{2})+3\Phi(0)]}{4[\Phi(1)-2\Phi(\frac{1}{2})+\Phi(0)]}\frac{\Phi(0)\Phi_r,}{\Big[\sum_{r=1}^{2N}(\Phi_r)^2\Big]}, \tag{25}$$

where $\Phi(1)$ is the value of Φ at the point given by applying the corrections from equation (24), and $\Phi(\frac{1}{2})$ the value with half these corrections. The Φ_r and $\Phi_{r,s}$ are obtained from (5), namely

$$\Phi_r = 2\sum_{j=1}^{2M}\psi_j\psi_{j,r} \tag{26}$$

and
$$\Phi_{r,s} = 2\sum_{j=1}^{2M}(\psi_{j,s}\psi_{j,r}+\psi_j\psi_{j,r,s}), \tag{27}$$

where
$$\left.\begin{aligned}\psi_{j,r} &= \partial\psi_j/\partial x_r\\ \psi_{j,r,s} &= \partial^2\psi_j/\partial x_r\,\partial x_s\end{aligned}\right\}. \tag{28}$$

The forms of the ψ_j derivatives depend, of course, on those of the particular ϕ_j of the problem.

The method has been used by the author in the solution of sets of equations of the type

$$\phi_j \equiv \Big[\sum_{r=1}^{N} f_r \cos jx_r\Big]^2 + \Big[\sum_{r=1}^{N} f_r \sin jx_r\Big]^2 - A_j = 0,$$

and for three such equations in three unknowns it was found that a solution, accurate to 0·01 radian, could be obtained in four steepest descents which took about 90 minutes.

An interesting comparison of the relative efficiency of the various methods is to be had via the solution of a set of simple simultaneous equations. The equations

$$x+2y = 7$$
$$2x+y = 5$$

were taken, and assuming initially that $x = y = 0$, it was found that eighteen applications of the Southwell or Synge relaxation technique gave the values $x = 0{\cdot}98, \quad y = 3{\cdot}02, \quad \Phi = 0{\cdot}0008.$

Five applications of the tangent descent method gave

$$x = 1{\cdot}04, \quad y = 2{\cdot}97, \quad \Phi = 0{\cdot}003.$$

After which Φ increased slightly, indicating breakdown of the method in the manner of M_N of Fig. 2, and finally, four applications of descent to a minimum or the interpolative method gave the correct solution:

$$x = 1{\cdot}00, \quad y = 3{\cdot}00, \quad \Phi = 0.$$

During the calculation of the relaxation version it became obvious that the convergence was very slow in the region of the minimum.

The five procedures for obtaining a solution of (1) via minimization of (2) may now be summarized.

1. *The Southwell relaxation method.* Find that axis (x_r) in the $2N$-dimensional space of the variables (x_r) for which Φ_r is greatest. Change x_r to $(x_r+\epsilon_r)$ where

$$\epsilon_r = -\Phi_r/\Phi_{r,r}$$

and repeat the process until Φ is minimized.

2. *The Southwell–Synge relaxation method.* Find that axis (x_r) for which $\Phi_r^2/2\Phi_{r,r}$ is greatest. Change x_r to $(x_r+\epsilon_r)$ where ϵ_r is as given in 1. Repeat until Φ is minimized.

3. *The steepest descent along a tangent.* Change all coordinates (x_r) to $(x_r+\epsilon_r)$ where the ϵ_r are given by

$$\epsilon_r = -\Phi(0)\Phi_r \Big/ \Big[\sum_{r=1}^{2N} (\Phi_r)^2\Big].$$

Repeat as in 1 and 2.

4. *The steepest descent to a minimum.* As in 3, but with the ϵ_r given by

$$\epsilon_r = -\Big[\sum_{r=1}^{2N} (\Phi_r)^2\Big]\Phi_r \Big/ \Big[\sum_{r,s=1}^{2N} \Phi_{r,s}\Phi_r\Phi_s\Big].$$

5. *The interpolative descent.* Calculate corrections as in 3. Evaluate Φ at the point given by applying these corrections ($\Phi(1)$, say) and also at the point given by applying one-half of these ($\Phi(\frac{1}{2})$). The corrections are then taken as

$$\epsilon_r = -\frac{[\Phi(1)-4\Phi(\frac{1}{2})+3\Phi(0)]}{4[\Phi(1)-2\Phi(\frac{1}{2})+\Phi(0)]} \frac{\Phi(0)\Phi_r}{\Big[\sum_{r=1}^{2N} (\Phi_r)^2\Big]}.$$

One point remains to be mentioned, the desirability of knowing whether the method has converged to a col, as at C_1 and C_2 in Fig. 1. In the case of a system having solutions in the classical sense, convergence to a col, or to a relative minimum, can be seen directly from the fact that Φ is not zero. When, however, the problem is one of minimization only, the distinction between cols and minima becomes important. The simplest test is to find $\Phi(\min)$ and then to consider the quadratic form

$$Q = \sum_{r,s=1}^{2N} \epsilon_r \epsilon_s \Phi_{r,s}. \tag{29}$$

If $\Phi(\min)$ is a true minimum, then Q is a positive definite form. This fact can be ascertained by the following procedure. In matrix notation

$$Q = \epsilon' A \epsilon, \tag{30}$$

where ϵ' is the transpose of ϵ and A is the matrix of the $\Phi_{r,s}$.

Let C be an orthogonal matrix, then a transformation

$$\epsilon = Ce \tag{31}$$

can be found such that

$$Q = \epsilon' A \epsilon = e' C' A C e = e' K e, \tag{32}$$

where K is a diagonal matrix, the quadratic form being thus reduced to

$$Q = \sum_{r=1}^{2N} k_{r,r} e_r^2. \tag{33}$$

Now let k be any element of K; it is evident that

$$|K - kI| = 0, \tag{34}$$

whence, from (32),

$$|C'AC - kI| = 0.$$

But since C is orthogonal $C'C = CC' = 1$, hence

$$\begin{aligned} |C'AC - kI| &= |C'AC - C'kIC| \\ &= |C'(A - kI)C| \\ &= |C'| \cdot |A - kI| \cdot |C| \\ &= |A - kI| \end{aligned}$$

whence

$$|A - kI| = 0. \tag{35}$$

This shows that the elements, k_{rr}, of the diagonal matrix K are the latent roots of A, i.e. the solutions of the secular equation

$$\begin{vmatrix} a_{11}-k & a_{12} & \cdots & a_{1,2N} \\ \vdots & \ddots & & \vdots \\ & & a_{rr}-k & \\ \vdots & & & \vdots \\ a_{2N,1} & & \cdots & a_{2N,2N}-k \end{vmatrix} = 0, \quad (36)$$

and, from (33), it is evident that Q is positive definite if and only if all $k_{rr} \geqslant 0$.

REFERENCES

1. R. V. Southwell, *Relaxation Methods in Theoretical Physics* (Oxford, 1946).
2. R. E. Gaskell, *Quarterly J. App. Math.* **1** (1943), 237.
3. J. L. Synge, ibid. **2** (1944), 87.
4. G. Temple, *Proc. Roy. Soc.* A, **169** (1939), 476.