A globally convergent procedure for solving a system of nonlinear algebraic equations

1985 J. Phys. A: Math. Gen. 18 2691

(http://iopscience.iop.org/0305-4470/18/14/020)

J. Phys. A: Math. Gen. 18 (1985) 2691-2699. Printed in Great Britain

A globally convergent procedure for solving a system of nonlinear algebraic equations

S M Tang and W C Kok

Department of Physics, National University of Singapore, Kent Ridge, Singapore 0511

Received 13 December 1984, in final form 18 March 1985

Abstract. A reliable computational procedure for solving a system of algebraic equations of the general form $\boldsymbol{F}(\boldsymbol{x})=0$ is presented. Global convergence is achieved through the requirement that the norm $\|\boldsymbol{F}\|$ decreases at each iteration as far as is feasible. A discussion of convergence and illustrative examples are given for two variable systems and these considerations are extended to systems of n equations in n variables with particular reference to three variable systems.

1. Introduction

The numerical solution of a set of nonlinear simultaneous algebraic equations is of importance in a large variety of physical problems. For example in recent years, an explicit expression for the pair correlation function of a hard-sphere fluid has been obtained (in the Percus-Yevick approximation) which involves all poles (k_0+ik_1) of the structure factor $S(z)$ in the complex z plane, k_0 and k_1 being determined numerically by solving two simultaneous equations (Kok 1980). The calculation of phase diagrams and bulk densities of the components in two co-existing liquid-vapour phases (Telo da Gama and Evans 1984) also involves the numerical solution of a set of nonlinear equations. Iterative methods for solving simultaneous nonlinear equations have also been used in band structure calculations in which the density-functional total energy is minimised (Bendt and Zunger 1982, Srivastava 1984). In addition, some physical problems may be formulated in terms of integral equations (Lonseth 1977, Guy *et al* 1984); it is known that one method of solving integral equations numerically is by reduction of an integral equation to a set of simultaneous algebraic equations (see e.g. Keller 1968).

In some cases, one has a fair idea of the approximate solutions when solving a set of equations. However in others, it is not often possible even to venture a crude guess as to where the solutions might be. In the latter situation, different methods have been relied on to provide an initial approximation which, once available, may be used as input to generate a solution to the required accuracy using methods which are usually fast but do require a good initial approximation. Generally, some of the problem areas encountered with the standard methods are as follows.

(a) If the starting point is distant from a root, the iterates might diverge away from the root for iterative methods which are not globally convergent. This problem is inherent in methods like that of Newton-Raphson which is not globally convergent except when solving for real functions (of one variable) whose first derivatives are monotonic. A recent development has been to devise safe starting regions (Moore and

0305-4470/85/142691+09$02.25

Jones 1977) from which successive iterates in the Newton–Raphson scheme converge to a root.

(b) With some methods based on minimisation, there is a possibility that the methods may fail when the iteration converges to a non-zero local minimum of the norm of the function.

In this paper, we report a practical procedure for finding roots of a set of algebraic equations of the general form

$$\boldsymbol{F}(\boldsymbol{x}) = 0. \tag{1.1}$$

The method is globally convergent irrespective of starting point and provides an automatic search for a real root. For a small number of equations, it is likely to be fast and efficient.

2. The computational procedure

We begin by illustrating the procedure for a system of two equations in two variables

$$F_1(x_1, x_2) = 0, \qquad F_2(x_1, x_2) = 0$$

and propose an iteration scheme of the form

$$x_1^{i+1} = x_1^i - F_1(x_1^i, x_2^i)/F_{1,1}(x_1^i, x_2^i) \tag{2.1a}$$

$$x_2^{i+1} = x_2^i - \tfrac{1}{2}F_2(x_1^{i+1}, x_2^i)/F_{2,2}(x_1^{i+1}, x_2^i) \tag{2.1b}$$

where the first derivatives of F_1 and F_2 are assumed to exist. There are three other possible iteration schemes similar to this one: one in which the variables x_1 and x_2 are interchanged and another two schemes obtained from the first two by interchanging the functions F_1 and F_2. It will be shown later that at least two of the four iteration schemes exemplified by (2.1) are convergent with the same convergence factor. The other two schemes, if convergent, also have a common convergence factor. The task is hence to find one of these convergent schemes, preferably the one with the smallest convergence factor. In the linear region in the neighbourhood of a root, the norm $\|\boldsymbol{F}\|$ evaluated at successive iterates should decrease if the iteration converges to a root. Our procedure for choosing the scheme is thus based on this trend of decreasing norm, i.e. we select the scheme corresponding to the smallest of the four norms $\|\boldsymbol{F}\|$ evaluated with the iterates generated from all the four schemes with a given starting vector. It is shown in § 3, that in the linear region, the scheme with the smallest convergence factor yields the smallest norm $\|\boldsymbol{F}\|$. If the starting vector is in a nonlinear region, the scheme chosen according to this criterion may not be convergent near the root although the resulting set of iterates is closer to the solution. Therefore, it is necessary to check that the norm $\|\boldsymbol{F}\|$ decreases at successive iteration points. If this is the case, then the same scheme is adopted for subsequent iterations. However, should $\|\boldsymbol{F}\|$ be found to increase at any stage, the search for a scheme that gives the smallest norm of $\boldsymbol{F}$ is repeated using the iterates of the previous stage as the starting vector. The algorithm for the iterative process is as follows.

Step 1. With a selected starting vector, evaluate the next set of iterates using each of the four schemes.

Step 2. For each set of iterates obtained in step 1, calculate the norm $\|\boldsymbol{F}\|$ and identify the scheme that gives the smallest $\|\boldsymbol{F}\|$.

Step 3. With the scheme identified in step 2, generate next set of iterates.

Step 4. Calculate the norm $\|F\|$ with the iterates obtained in step 3 and compare with that calculated with the previous iterates. If the former is smaller than the latter, then go to step 5 else take the previous iterates as starting vector and go to step 1.

Step 5. Apply stopping criterion; if not satisfied go to step 3.

This algorithm has been tested and found to work well in most cases. However, for certain systems of equations, it is possible that at some stages of iteration all the four norms obtained in step 2 exceed the norm evaluated with the previous iterates. When this occurs the convergence can be slow or the algorithm may even fail. Such a situation arises in two ways.

(a) If there exists a region where the separation between the two functions becomes very small but there is no real intersection, the iteration will proceed towards this region. At the stage when the iteration point is in the region of minimum separation, the above mentioned situation will occur. If the iterative process continues, subsequent iteration points will oscillate about this narrow region.

This, however, does not pose a real problem and can be overcome by a slight modification of the algorithm. When the situation of increasing norm is encountered for all the four schemes, the scheme that gives the smallest norm at the next iteration point is selected as prescribed in the above algorithm. But in the subsequent iterations, the choice of scheme is restricted to only two out of the four if the successive norms are still increasing. These two schemes comprise the one currently in use and the one having the same convergence factor. This restriction is lifted when the norm starts to decrease. Hence, the modification only involves some minor changes in the algorithm. Limiting the choice of these two schemes has the effect of constraining the iteration to proceed away from the neighbourhood of minimum separation.

(b) If at any stage, all the first partial derivatives of F_1 and F_2 are small compared with the functions themselves, the iteration equations (i.e. equations (2.1) and similar ones associated with the other schemes) will generate a point which is much further away from the root than the current point. This is usually manifested by an enormous increase of the norm $\|F\|$. However, because of the global convergence property of the present procedure, the iteration will eventually converge back to the root, although this may mean many more iteration steps. To reduce the number of iterations, one may continue the iteration with a point nearby which does not give too small a derivative. This point can be arbitrarily chosen. A choice that is likely to lead to fewer iteration steps for convergence to a root is the inversion of the previous point with respect to the current point. This would certainly be a logical choice if the previous successive iteration points are proceeding towards a root.

3. Discussion on convergence

The convergence factor α is defined as

$$\alpha = \frac{e_j^{i+1}}{e_j^i} = \frac{x_j^{i+1} - r_j}{x_j^i - r_j} \qquad \text{for } j = 1, 2$$

where e_j^i is the error in the ith iterate x_j^i, and the root $\boldsymbol{r} = (r_1, r_2)$.

For the iteration scheme (2.1) hereafter referred to as the scheme $x_1F_1x_2F_2$, it can be shown that the convergence factor is

$$\alpha_1 = \tfrac{1}{2}(m_2/m_1 + 1) \tag{3.1a}$$

where m_1 and m_2 are the gradients of the curves $F_1 = 0$, $F_2 = 0$ at the root respectively. The convergence factors corresponding to the other three alternative schemes $x_2F_2x_1F_1$ (in which the variable x_2 is iterated first using the iteration function F_2), $x_2F_1x_1F_2$ and $x_1F_2x_2F_1$ are respectively

$$\alpha_2 = \tfrac{1}{2}(m_2/m_1 + 1), \qquad \alpha_3 = \tfrac{1}{2}(m_1/m_2 + 1), \qquad \alpha_4 = \tfrac{1}{2}(m_1/m_2 + 1). \tag{3.1b, c, d}$$

It is readily seen from these expressions for the α's that there are only two distinct convergence factors, the smaller of which is less than one except in the special case where $m_1 = m_2$.

To achieve convergence, one needs to choose one of two schemes that have a convergence factor less than one when the iteration has reached the linear region near a root. We now show that starting from a given point specified say by the ith iterates in such a region, the scheme that yields the smallest norm evaluated at the $(i+1)$th iterates is that associated with the smaller convergence factor. This fact is used in our proposed algorithm (step 2) to search for a suitable scheme. To prove this, we give below expressions for the square of the norm, $N = F_1^2 + F_2^2$, evaluated with the $(i+1)$th iterates in the linear region near a root for each of the four iteration schemes mentioned earlier in the section (taken in the same order):

$$N_1 = \tfrac{1}{4}(a_2^2 + b_2^2)[(1 - m_2/m_1)x_2^i]^2 \tag{3.2a}$$

$$N_2 = \tfrac{1}{4}(a_1^2 + b_1^2)[(1 - m_2/m_1)x_1^i]^2 \tag{3.2b}$$

$$N_3 = \tfrac{1}{4}(a_1^2 + b_1^2)[(1 - m_1/m_2)x_1^i]^2 \tag{3.2c}$$

$$N_4 = \tfrac{1}{4}(a_2^2 + b_2^2)[(1 - m_1/m_2)x_2^i]^2 \tag{3.2d}$$

where

$$a_j = F_{1,j}(r_1, r_2), \qquad b_j = F_{2,j}(r_1, r_2).$$

From these expressions for the N's, one obtains

$$N_4/N_1 = N_3/N_2 = m_1^2/m_2^2. \tag{3.3}$$

If, for example, $m_2^2/m_1^2 < 1$, then the first two schemes have the smaller convergence factor according to equations (3.1). In this case, $N_4 > N_1$ and $N_3 > N_2$, and hence the iteration scheme associated with the smallest norm has a convergence factor less than unity. Convergence to a root is thus guaranteed once in the linear region near a root.

4. Systems of n equations in n variables

The iterative procedure described in the preceding sections can be extended to a system of n nonlinear equations in n variables of the general form:

$$\boldsymbol{F}(\boldsymbol{x}) = 0 \tag{4.1a}$$

or

$$F_j(x_1, x_2, \ldots, x_n) = 0, \qquad j = 1, 2, \ldots, n \tag{4.1b}$$

where F is an n-component vector-valued function. The set of iteration equations we propose, analogous to (2.1), is as follows.

For the first iterate,

$$x_1^{i+1} = x_{1'}^{i+1} \tag{4.2a}$$

and for the other $n-1$ iterates,

$$x_j^{i+1} = \tfrac{1}{2}(x_j^i + x_{j'}^{i+1}) \tag{4.2b}$$

where

$$x_{j'}^{i+1} = x_j^i - F_{j'}^i / F_{j',j}^i$$

with

$$F_{j'}^i = F_j(x_{1'}^{i+1}, x_{2'}^{i+1}, \ldots, x_{(j-1)'}^{i+1}, x_j^i, \ldots, x_n^i).$$

It can be shown that near a root the relation between the error vectors, $e = x - r$, of successive iterations is

$$e^{i+1} = Me^i \tag{4.3}$$

where $M = \frac{1}{2}[I - (L+D)^{-1}U]$ and D, L, U are the respective diagonal matrix, lower and upper triangular matrices with zeros in each position of the leading diagonals such that

$$D + L + U = F'(r).$$

Explicitly, for the jth component of the error vector e,

$$e_j^{i+1} = \frac{e_j^i}{2} + \frac{1}{2}\left(F_{j,1} + \sum_{l=1}^{n-1} B_l \frac{F_{j-l,1}}{F_{j-l,j-l}}\right) \sum_{k=2}^{n} \frac{F_{1,k} e_k^i}{F_{j,j} F_{1,1}}$$

$$- \frac{1}{2} \sum_{k=j+1}^{n} \frac{F_{j,k}}{F_{j,j}} e_k^i - \frac{1}{2} \sum_{l=1}^{n-1} \sum_{k=j}^{n} \frac{B_l F_{j-l,k}}{F_{j-l,j-l} F_{j,j}} e_k^i, \qquad j = 2, 3, \ldots, n \tag{4.4}$$

where $B_l = (B_{l-1} A_{j-l} - F_{j,j-l})(1 - \delta_{j,l+1})$, $B_0 = 0$, and $A_{j-l} = -F_{j-l+1,j-l}$, all derivatives being evaluated at the root.

The linear convergence factor α satisfies the secular equation

$$\det(S - \alpha I) = 0 \tag{4.5}$$

where $S = (M_{jk})_{j,k=2}^n$ is the $(n-1)\times(n-1)$ matrix formed from the coefficients of $e_2^i, e_3^i, \ldots, e_n^i$ on the RHS of (4.4). In general α may be complex. The necessary condition for convergence is $|\alpha| < 1$. This condition will be guaranteed if the spectral radius $\rho(S) < 1$, i.e. all the roots of the secular equation (4.5) lie within the unit hypercircle with centre at the origin.

For a system of two equations in two variables, there are four different iteration schemes, i.e., the schemes $x_1F_1x_2F_2$, $x_2F_2x_1F_1$, $x_2F_1x_1F_2$ and $x_1F_2x_2F_1$ described in § 2.

For $n = 3$, there are 36 different iteration schemes. It can easily be proved that the number of convergence factors is only 12, each associated with three schemes. For example, the scheme $x_1F_1x_3F_2x_2F_3$ and the other two schemes $x_3F_2x_2F_3x_1F_1$ and $x_2F_3x_1F_1x_3F_2$, which are obtained through a cyclic permutation of the xF groups, have the same convergence factor.

Of these twelve generally distinct convergence factors, it is of interest to examine how many of them have absolute values less than one. An analytical approach for

doing this could be quite involved. We adopt a simple way that makes use of the Monte Carlo technique.

The secular equation (4.5) reduces, for $n = 3$, to a quadratic equation whose larger root is given by

$$\alpha_l = (p+s)\{1+[1-4(ps-qr)/(p+s)^2]^{1/2}\} \tag{4.6}$$

where p, q, r and s are functions of the various first partial derivatives of F_j at the root. For the scheme $x_1F_1x_2F_2x_3F_3$,

$$p = \tfrac{1}{2}(1+a_2b_1/a_1b_2), \qquad q = \tfrac{1}{2}(a_3b_1/a_1b_2 - b_3/b_2)$$

$$r = \tfrac{1}{2}(a_2c_1/a_1c_3 - a_2b_1c_2/a_1b_2c_3), \qquad s = \tfrac{1}{2}(1 - a_3b_1c_2/a_1b_2c_3 + b_3c_2/b_2c_3 + a_3c_1/a_1c_3)$$

and $a_j = \partial F_1/\partial x_j$, $b_j = \partial F_2/\partial x_j$, $c_j = \partial F_3/\partial x_j$, $j = 1, 2, 3$, all evaluated at the root. As mentioned earlier, if the absolute value of α_l deduced from (4.6) is less than one, the corresponding scheme is convergent. Our approach involves the generation of nine random values of the a's, b's and c's and the calculation of $|\alpha_l|$ for all the 12 schemes. The frequency distribution of the occcurrence of α's having absolute values less than one is shown in table 1. It is obtained from 10 000 sets of randomly generated a's, b's and c's, which are restricted to lie between -1000 and $+1000$.

Table 1.

No of convergent schemes	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	0	0	1782	1548	3334	1808	1168	163	158	39	0	0	0

The table shows that for every set of a's, b's and c's, at least two and as many as nine schemes are convergent. Although this does not amount to a conclusive proof, it is highly plausible that convergence in the linear region is assured for a system of three equations involving functions of three variables when a solution exists.

If the iteration proceeds with a fixed sequence of x_1, x_2 and x_3, then there are altogether only $3! = 6$ possible schemes, e.g. $x_1F_1x_2F_2x_3F_3$, $x_1F_1x_2F_3x_3F_2$, $x_1F_2x_2F_1x_3F_3$, $x_1F_2x_2F_3x_3F_1$, $x_1F_3x_2F_1x_3F_2$, $x_1F_3x_2F_2x_3F_1$. It can be shown similarly that there is at least one convergent scheme among the six.

This same technique may be used to study the convergence of higher order systems of equations.

5. Illustrative examples

In this section, we illustrate the use of the present procedure in solving some systems of two nonlinear equations in two variables and subsequently we apply it to a system involving three variables. The examples are chosen to demonstrate the problems encountered in some of the test runs which have led to the modifications described in the last part of § 2.

Example 1. Solve

$$F_1(x_1, x_2) = x_1^2 + x_2^2 - 1 = 0$$

$$F_2(x_1, x_2) = x_1^2 - x_2 = 0.$$

The Euclidean norm is used throughout this example. Table 2 shows the successive iterates with (200, −100) as starting point. The stopping criterion used is

$$\left(\sum_{j=1}^{2} (x_j^i - x_j^{i-1})^2\right)^{1/2} < 0.0001.$$

The superscript 'a' denotes where step 1 in the algorithm described in § 2 is required and the iterates are obtained from the chosen scheme; 'b' indicates where the second problem mentioned in § 2 occurs and the iteration point is replaced by the inversion of the 8th point about the 9th. If the iteration is allowed to continue at step 10 without such a change, then 23 iteration steps would be required for convergence to the root with the same accuracy. It can be seen from the table that convergence is fast outside the linear region. Because the linear convergence factor is close to one in this particular case, a large proportion of the computing time is actually spent in refining the approximate solution in the vicinity of the root.

Table 2.

Iteration	(x_1, x_2)	Iteration	(x_1, x_2)
1	99.750 000, −50.127 344[a]	11	0.551 697, −0.396 941[a]
2	49.623 735, −25.319 200	12	1.039 345, 0.341 649
3	24.556 756, −13.044 955	13	0.944 592, 0.616 952
4	12.012 770, −7.037 316	14	0.800 147, 0.628 593
5	5.713 475, −4.153 844	15	0.778 048, 0.616 976
6	2.493 225, −2.801 447	16	0.787 033, 0.618 199
7	0.684 800, −2.148 476	17	0.786 023, 0.618 015
8	−1.226 288, −1.552 736	18	0.786 166, 0.618 036
9	0.019 960, −1.325 494	19	0.786 150, 0.618 034
10	1.266 208, −1.098 252[b]		

Example 2. Consider solving

$$F_1(x_1, x_2) = x_1 \sin x_1 - x_2 = 0$$

$$F_2(x_1, x_2) = 1/x_1 + 2x_1 - x_2 - 5 = 0$$

as an example of the problem (a) described in § 2. With (20, 10) as starting point and the same stopping criterion as in example 1, the iteration converges to (2.7954, 0.9485) in 11 steps (see table 3).

The superscript 'a' is as in example 1 and 'b' denotes iterates obtained from a scheme chosen from the two having the same convergence factor.

Example 3. Solve

$$F_1 = x_3 - x_1 x_2 x_3 = 0$$

$$F_2 = x_1 + x_2 + x_3 = 0$$

$$F_3 = x_1 x_2 + x_2 x_3 + x_3 x_1 = 0.$$

Table 3.

Iteration	(x_1, x_2)	Iteration	(x_1, x_2)
1	7.459 324, 8.442 957[a]	7	2.731 629, 0.952 051
2	7.870 026, 9.655 036[a]	8	2.797 391, 0.948 009
3	7.259 054, 7.833 452[a]	9	2.795 121, 0.948 589
4	6.339 118, 4.093 914[b]	10	2.795 431, 0.948 525
5	4.444 508, −0.096 038[b]	11	2.795 397, 0.948 532
6	2.284 817, 0.815 339[b]		

This system of equations is solved using a set of twelve iteration schemes having distinct convergence factors. With (1, 2, 3) as starting point, the iteration converges to the solution (0, 0, 0) in 13 steps when

$$\left(\sum_{j=1}^{3}(x_j^i - x_j^{i-1})^2\right)^{1/2} < 0.0001.$$

A rudimentary version of the procedure given here has been used (Kok 1980, Kok and Tang 1982) as the basis for the numerical solution of a set of two equations in which the functions are expressed in integral form and are evaluated using a separate subroutine.

In the examples given above and many others that have been tested, global convergence is a universal feature.

6. Final remarks

The iteration scheme (2.1) is similar to the one used in successive overrelaxation (Lieberstein 1968, Ortega and Rheinboldt 1970) for which the equations of iteration are as follows:

$$x_1^{i+1} = x_1^i - \omega F_1(x_1^i, x_2^i)/F_{1,1}(x_1^i, x_2^i)$$
$$x_2^{i+1} = x_2^i - \omega F_2(x_1^{i+1}, x_2^i)/F_{2,2}(x_1^{i+1}, x_2^i).$$

The present procedure, however, has two important features to cater for convergence on a global scale: (i) that of restricting the successive iterates to a trend of decreasing norm $\|\boldsymbol{F}\|$ and (ii) provision of alternative schemes to facilitate convergence.

Although the iteration equations adopted in the procedure are of Newton-Raphson form, they may be replaced by equations of other standard techniques such as the secant or bisection method.

Acknowledgments

This work is supported by a grant from the National University of Singapore. One of us (WCK) is indebted to the School of Chemical Engineering, Cornell University where part of this work was done, for the kind hospitality extended during the stay.

References

Bendt P and Zunger A 1982 *Phys. Rev.* B **26** 3114
Guy J, Mangeot B and Salès A 1984 *J. Phys. A: Math. Gen.* **17** 1403
Keller H B 1968 *Numerical Methods for Two-Point Boundary-Value Problems* (Waltham, MA: Blaisdell)
Kok W C 1980 *Phys. Lett.* **78A** 273
Kok W C and Tang S M 1982 *J. Chem. Phys.* **77** 4227
Lieberstein H M 1968 *A Course in Numerical Analysis* (New York: Harper and Row) p 113
Lonseth A T 1977 *SIAM Review* **19** 241
Moore R E and Jones S T 1977 *SIAM J. Numer. Anal.* **14** 1051
Ortega J M and Rheinboldt W C 1970 *Iterative solution of Nonlinear Equations in Several Variables* (New York: Academic)
Srivastava G P 1984 *J. Phys. A: Math. Gen.* **17** L317
Telo da Gama M M and Evans R 1984 *Mol. Phys.* **48** 229