General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

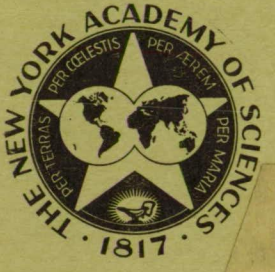
- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

SOLUTION OF NONLINEAR SYSTEMS

L. R. Turner



Reprinted from
ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Volume 86, Article 3, Pages 817-827
May 25, 1960

SOLUTION OF NONLINEAR SYSTEMS

L. R. Turner
Lewis Research Center, National Aeronautics and Space Administration, Cleveland, Ohio

Introduction

The problem of solving systems of nonlinear equations has been relatively neglected in the mathematical literature, especially in the textbooks, in comparison to the corresponding linear problem. Moreover, treatments that have an appearance of generality[1,2] fail to discuss the nature of the solutions and the possible pitfalls of the methods suggested.

Probably it is unrealistic to expect that a unified and comprehensive treatment of the subject will evolve, owing to the great variety of situations possible, especially in the applied field where some requirement of human or mechanical efficiency is always present. Therefore we attempt here simply to pose the problem and to describe and partially appraise the methods of solution currently in favor.

Statement of the Problem

A problem in nonlinear algebra in several variables may arise in several ways. Fairly commonly, a physically oriented problem will contain as many well-established, physically independent conditions as there are unknown quantities in the solution. That this parity between the number of independent statements and unknowns is not required, however, may be demonstrated by a simple exercise.

For any linear system of the form

$$F_i = b_i - \sum_j a_{ij}x_j = 0$$

there is a corresponding nonlinear problem of locating the zeros, if any, or the minima of the test function

$$\phi_1 = \sum_{l,m} F_l c_{l,m} F_m$$

where $c_{l,m}$ is an element of any positive definite matrix. The single function ϕ_1 has the zeros of the system $F_i = 0$, and its minima otherwise represent a best solution of the system in some least-squares sense. The properties of the original system are closely linked with the properties of the Jacobian (or Hessian) matrix

$$J = \left(\frac{\partial^2 \phi_1}{\partial x_i \partial x_j} \right)$$

If $| J | \neq 0$, there will be a unique minimum point of ϕ_1 that will be a solution if the initial system is consistent. If $| J | = 0$, there will be a con-

tinuous locus of minima (or perhaps of zeros) that, depending on definition, may be treated as infinitely many solutions or as no solution.

Problems that consist of a single statement are common in physics; one example is the location of the minimum or other level points of a potential function. With several statements, it is frequently expedient to construct non-negative test functions such as ϕ_1 and to replace the original problem by that of finding the level points, usually minima, of ϕ_1.

In the nonlinear as in the linear case, the Jacobian matrix J of a potential function or of quadratic test function ϕ_1 generally characterizes the nature of the solution; in the nonlinear case, however, there are additional problems. While a nonlinear system may have a unique minimum, it is more likely that it will have several minima or level points. Furthermore, these level points may be distinct or finitely confluent or they may occur in continuous sets, and any combination of these situations may occur in a single problem. If the level points are all distinct, then $| J | \neq 0$ in some open region containing each level point. At any of the points of confluence, $| J | = 0$. If there are two or more nearly coincident level points, $| J |$ will be very small and may be computationally zero.

Milne[3] has described a process by which the difficulty of coincident or nearly coincident roots may be resolved. An attempt is made to solve jointly the initially given system and the added equation $| J | = 0$. If the augmented system has a subsystem with a nonvanishing Jacobian determinant at the space point in question, the point is a multiple root or an approximation to a set of nearly coincident roots of the original system. It is easily demonstrated, however, by consideration of the well-determined problem $\phi = x^2 + y^4 = 0$, that a single use of this process may be inadequate to resolve the difficulty.

A very practical difficulty arises in the course of locating the minima of a potential or test function: if the process of solution is one designed to seek a level point of ϕ_1, the computation may tend toward a maximum or a saddle point, as well as toward the true minimum. Booth[1] points out that these various types of level points may be distinguished by examining the quadratic form

$$\sum_{l,m} \frac{\partial^2 \phi_1}{\partial x_l \partial x_m} \cdot x_l \cdot x_m$$

which is positive definite only at a minimum. This test will fail if the level points being tested are not simple; moreover, no formal test evaluated at a single point will distinguish between a relative and an absolute minimum.

Once a root or extremal at the vector location Y has been determined, a further solution may be found by consideration of the successor problem:

$$G_i = \frac{F_i}{|| X - Y ||^p} = 0$$

where $\| X - Y \|$ is a suitable norm. The solution Y is discriminated against weakly if $p = 1$ and more strongly if $p > 1$. A desirable choice of p would seem to be the square distance $\sum_i (x_i - y_i)^2$ if the test function has quadratic minima. If the problem is one of locating minima rather than zeros, the original problem must be returned to once an approximation to a second level point is found.

It has been noted that this method of discrimination involves the theoretical difficulty that, when the locus of a zero is obtained numerically and therefore approximately, the successor problem has a pole and a nearly coincident zero (Morton A. Hyman, personal communication). This will become a real difficulty if the initial problem has nearly coincident zeros. In situations of this kind, Milne recommends the use of extended precision arithmetic.

Finally, we must remember that since it is impossible to locate an infinite number of roots, limits on the number of roots or on their range of locus are always required.

Methods of Solution

With so great a variety of situations it seems unlikely that a general method could be used to solve all problems, even if the number of variables were limited. Experience gained through observing the solution of many problems indicates that use of a mixture of techniques greatly increases the probability of success.

The methods usually recommended are of two types. The first involves use of functional iteration in either an extension of Newton's method to several variables or, in some special cases, a method of successive substitution. The second is a method of "descents" by which the problem of solving a system of nonlinear equations is reduced to that of solving an infinite sequence of nonlinear equations in one variable.

Functional Iteration

Methods of functional iteration include any method in which the initial system of equations $F_i = 0$ is written in the vector form $\mathbf{X} = \mathbf{G}(\mathbf{X})$ and solved by the iteration $\mathbf{X}_{k+1} = \mathbf{G}(\mathbf{X}_k)$, the subscripts indicating successive iterations. In many applied problems the separate equations will represent physically independent statements about the solution, and in such cases it is usually possible to obtain a solution by writing $\mathbf{X}_{k+1} = \mathbf{X}_k - \mathbf{g}F$ where

$$\mathbf{g} = J^{-1} = \left(\frac{\partial F_i}{\partial x_j}\right)^{-1}$$

This is the extension of Newton's method to systems of equations; usually it will be written in the form

$$\left(\frac{\partial F_i}{\partial x_j}\right)(\mathbf{X}_{k+1} - \mathbf{X}_k) = -F_i(\mathbf{X}_k)$$

A problem of minimizing a potential function or a test function can be written in a form appropriate for Newton's method as follows:

$$\left(\frac{\partial^2 \phi_1}{\partial x_i \partial x_j}\right)(\mathbf{X}_{k+1} - \mathbf{X}_k) = \left(\frac{\partial \phi_1}{\partial x_i}\right).$$

Householder[2] has shown that this process converges if the Jacobian matrix is not singular and if the initial estimate X_0 is sufficiently close to X.

The condition that the Jacobian determinant not vanish will be violated at multiple roots. To resolve this difficulty, Milne[3] discards temporarily one of the original equations, $F_{i_1} = 0$, and in place of the original system solves the system

$$F_i = 0 \qquad (i \neq i_1)$$
$$| J | = 0$$

If the revised system leads to a root Z and if $F_{i_1}(Z) = 0$, then within computational accuracy the original system has a pair of coincident roots at Z. More significantly, perhaps, if the initial system is undetermined at or near a point Z_0, the revised system will also be undetermined at this space point. Unfortunately, the converse is not true, as may be shown by applying Milne's technique to the solution of the equation $\phi = x^2 + y^4 = 0$. The component equations are

$$2x = 0 \tag{1}$$

$$4y^3 = 0 \tag{2}$$

From this,

$$| J_1 | = 24y^2 \tag{3}$$

and, if EQUATION 3 replaces EQUATION 2, we have $| J_2 | = 96y$ and, finally, $| J_3 | = 192$. The completion of the chain demonstrates that the original system was well determined. If EQUATION 1 rather than EQUATION 2 had been replaced by $| J_1 | = 24y^2 = 0$, the false conclusion might have been reached that the initial system was indeterminate.

A numerical discrimination between indeterminacy and multiplicity of roots is sometimes available. If the degree of nullity of a singular Jacobian matrix at point X_0 is only 1, the characteristic vector $\mathbf{V}$ associated with the zero root is easily calculated. Then to terms in the first order the equation $J = 0$ is satisfied by $\mathbf{X}_1(h) = \mathbf{X}_0 + h\mathbf{V}$. If by a test computation with $\mathbf{X}_1(h)$ as a starting point the root X_0 is again found, X_0 is probably a multiple root. If each trial $\mathbf{X}_1(h)$ leads to a different apparent root, the system is probably underdetermined or at least has many nearly coincident roots. Any computation of this sort is difficult because of rounding errors.

Aside from the difficulties caused by coincident roots and the general difficulties that may occur in any solution of linear simultaneous equations, a very important difficulty is that convergence in the large, that is, convergence to some solution from an arbitrary starting point, is not assured. For example, the system

$$F_1 = x^2 - y^2 = 0$$

$$F_2 = 1 + 2xy = 0$$

will not converge by Newton's method for any starting point $x = y$. This difficulty is analogous to the impossibility of finding a complex zero of an algebraic equation with the use of only real arithmetic. In this and similar cases, an initial estimate nearly equal to a starting point that causes nonconvergence is expected to cause slow convergence.

Even in relatively favorable cases, Newton's method will often overshoot during the early iterations. In some cases of initial overshoot,[4] a significant improvement in the rate of convergence has been achieved merely by limiting the magnitude of the vector change in the variables in each iteration.

These experiences suggest strongly that, when Newton's method is used, alternatives such as step limitation or inverse interpolation might be used when some measure of convergence, such as a norm of the residuals, is not reduced. If a truly nonconvergent sequence is encountered, a restart will be required.

A second type of method of functional iteration involves only successive substitution in the application. Briefly, it is required that the equations be written in the form $\mathbf{X} = \mathbf{G}(\mathbf{X})$ and that the function $\mathbf{G}(\mathbf{X})$ have a formal derivative $\mathbf{G}'$ less than unity. Specifically, if $\mathbf{X}_1$ and $\mathbf{X}_2$ are any vectors in the neighborhood S of a point for which $\mathbf{X} = \mathbf{G}(\mathbf{X})$ and

$$\| \mathbf{G}(\mathbf{X}_2) - \mathbf{G}(\mathbf{X}_1) \| \leqq K \| \mathbf{X}_2 - \mathbf{X}_1 \| \qquad K < 1$$

the solution point X is unique. If, on the contrary, there were two solutions U_1 and U_2 in S, then clearly we would have

$$\| \mathbf{G}(\mathbf{U}_1) - \mathbf{G}(\mathbf{U}_2) \| = \| \mathbf{U}_1 - \mathbf{U}_2 \|$$

in violation of the hypothesis. Furthermore, if the hypothesis is satisfied, convergence is clearly established.

Unfortunately, this method yields a domain of convergence that usually is small unless the function $\mathbf{G}$ is obtained by some elaborate process such as the formal expansion of Newton's method; therefore, more often than not, divergence of an algorithm does not mean that the problem has no solution. On the other hand, convergence of an algorithm likewise does not of itself imply uniqueness, which must be established independently.

This method has been very effective for applications in which it arises

naturally, such as the solution of integral equations and quasi-linear partial differential equations and the numerical approximation of the Picard iteration process as applied to systems of ordinary differential equations. Early efforts to apply this method to the problem of chemical equilibrium were abandoned in favor of the use of Newton's method, because many different algorithms were required to achieve convergence in general situations. This experience suggests that the method should be considered for problems of limited scope only.

Methods of Descent

The common feature of methods of descent is an explicit emphasis on the location of a minimum of a potential function or of some nonnegative (test) function of the residuals of a system of equations. The variants described in the literature range from the very formal "method of steepest descents" to the almost purely experimental "downhill" methods.

The advantages of methods of descent as compared with methods of functional iteration are as follows: convergence in the large seems almost always to be assured and it is relatively easy, especially with the simplified methods, to prepare stored program computer coding for approximating the solution of an arbitrary problem and finding successive solutions when several are required.

Starting from a geometric interpretation of a function ϕ in n variables as a surface in a space of $n + 1$ dimensions, we designate the vector with components $\partial\phi/\partial x_i$ the gradient of ϕ (grad ϕ). If ϕ is everywhere differentiable, grad ϕ vanishes at the minima, maxima, and saddle points of ϕ. Elsewhere, the gradient is in the direction of the most rapid increase of ϕ. Further, if $\mathbf{V}$ is any unit vector and if we denote differentiation in the direction of $\mathbf{V}$ by primes, then $\phi' = \mathbf{V} \cdot \text{grad } \phi$ and

$$\phi'' = \sum_{i,j} v_i v_j \frac{\partial^2 \phi}{\partial x_i \partial x_j}$$

if the second derivatives exist. Provided that derivatives higher than the second are unimportant, the minimum of ϕ in the direction of $\mathbf{V}$ from point X_k occurs approximately at X_{k+1}, where

$$X_{k+1} = X_k - (\mathbf{V}\phi'/\phi'')$$

If $\mathbf{V}$ is the unit vector in the direction of grad ϕ, the iterative method is a method of steepest descent. The foregoing description is essentially that given by Householder[2] and Booth.[1]

Because Newton's method is an intermediate step, this descent method does not assure convergence and, when achieved, convergence may occur at a maximum or level point. Furthermore, the ultimate rate of convergence is first order.

Although Booth[1] used this detailed method, he noted that the effort in-

volved in computing second derivatives is excessive, and replaced the calculation of directional derivatives by finite difference approximations with the following result:

$$X_{k+1} = X_k + \frac{V_k}{2} \frac{3\phi(X_k) - 4\phi(X_k + V_k) + \phi(X_k + 2V_k)}{\phi(X_k) - 2\phi(X_k + V_k) + \phi(X_k + 2V_k)}$$

where V_k is a vector step size. Any use of this variant requires effective control of the direction and magnitude of the vector step-size. Booth suggests the use of the step

$$V = -\frac{1}{2} \frac{\phi \text{ grad } \phi}{(\text{grad } \phi)^2}$$

which is effective if the process tends toward a zero of ϕ, but may fail if too large a step-size is predicted when the process tends toward a level point that is not a zero of ϕ.

Regardless of the specific variant selected, certain rules of thumb may be used to avoid excursion toward a maximum. For example, negative values of ϕ'' or of

$$\phi(X_k) - 2\phi(X_k - V_k) + \phi(X_k + 2V_k)$$

indicate that the computation is tending toward a maximum or a saddle point, and a step in the opposite direction may be effective. Further, one should expect that $\phi(X_1)$ will be smaller than any other computed value; if not, the location of the smallest value of ϕ may be used for further calculation.

The recent trend in descent methods stresses simplicity of coding for a stored program computer; in particular, the calculation of derivatives tends to be avoided. The extremes of this trend occur in the downhill methods of Ward[5] and of Lance,[6] in which the zeros or minima of test functions are sought by systematic sampling procedures. The methods as published have been applied to the location of zeros of analytic functions of a complex variable.

By Ward's method, values of the test function are computed at a starting point X_0 and at $X_0 \pm he_i$. If the value of the test function is reduced, the locus of the smallest value is used at the next approximation X_1. When no reduction is found, the step size h is halved. The calculation continues until h falls below some assigned limit.

Lance has used similar data to approximate the gradient of the test function ϕ from the relations

$$-h \frac{\partial \phi}{\partial x_i} \approx \phi(X_0) - \phi(X_0 + he_i)$$

or

$$-h \frac{\partial \phi}{\partial x_i} \approx \frac{1}{2} [\phi(X_0 - he_i) - \phi(X_0 + he_i)]$$

which lead, respectively, to these approximations for X_1 :

$$x_{i_1} = x_{i_0} + \phi(X_0) - \phi(X_0 + he_i)$$

or

$$x_{i_1} = x_{i_0} + \frac{1}{2}[\phi(X_0 - he_i) - \phi(X_0 + he_i)]$$

Of these approximations Lance prefers the first because fewer values of the test function need be computed.

Both authors use the Gerschgorin vector norm, $\phi_2 = \sum_i |F_i|$, as a test function in preference to various possible quadratic forms. In applications to analytic functions of a complex variable, the real and imaginary parts are treated as separate functions. Ward has proved that in this case the test function ϕ_2 cannot have relative minima.

In addition to the obvious fact that an initially chosen step size may be inappropriate and hence may require many adjustments, an important difficulty inherent in these methods is the development of stalemate situations in which the algorithm fails to predict a correct downhill path for any step size.

Lance found that Ward's method sometimes degenerated into a stalemate situation when the test function "contained a 'trough' which was inclined to the axes."[6] This particular difficulty was treated effectively by Lance's variation.

However, Lance's method failed on the simple problem

$$F_1(x,y) = x = 0$$

$$F_2(x,y) = y^2 = 0$$

with the test function $\phi = |x| + y^2$, upon reaching the point $x = 0$, $y = -\frac{1}{4}$. It is easily verified that from this point there is no step size leading to a reduction of ϕ by the first of Lance's methods. In general, then, it can be said that Lance's approximate gradient method may tend toward a stalemate situation when the test function has troughs along the coordinate axes.

The probability of success can be increased in many ways, such as by the use of an alternative method or of several methods simultaneously. One method that may be suggested is the conditional use of a variable test function. In particular, the test function $\phi_2 = \sum |F_i|$ fails to have a gradient at any point where one of its elements F_i changes sign. If the derivatives of F_i are sufficiently large, the minimum value and the subsequent descent path will lie in the hypersurface $F_i = 0$.

This property of the test function suggests an algorithm that is known to be effective in simple cases:

If a trial step of a downhill method fails to reduce the value of the test function and if at least one residual has changed sign, (1) interpolate along the step for the zero of the residual F_j that changed sign, and (2) proceed for one step with the test function $\phi_3 = \sum_{i \neq j} |F_i|$. If the modified test function is reduced, return to the original problem.

In conclusion of this discussion of descents and inverse interpolation methods, a few other properties may be noted:

(1) The descent or downhill methods seem to assure, in general, convergence in the large, but the rate of convergence and accuracy may be poor at multiple roots. The development of stalemate situations may prevent ultimate convergence.

(2) For a wide variety of problems, a general technique can be devised requiring almost no modification from problem to problem.

(3) Simple and effective methods are available for finding successive roots.

(4) The one important fault of these methods is that they will cause convergence to a point of indeterminacy of a system of equations, and generally will do so without giving warning of any kind. This fault, however, may not be important in physically oriented problems.

A Mixed Method

Recently, Davidon[7] has described methods for locating a minimum of a potential function or quadratic test function, involving the calculation of an approximate inverse of the Jacobian matrix $(\partial^2\phi/\partial x_i \partial x_j)$ without requiring that the Jacobian matrix be computed explicitly. The methods may include linear constraints also.

Starting with an estimate X_0 of the solution and H_0 of the inverse J^{-1} of the Jacobian matrix, Davidon computed the gradient ∇_0 of ϕ at X_0 and a new test point $\bar{X}_0$,

$$\bar{X}_0 = X_0 - H_0\nabla_0$$

and the gradient $\bar{\nabla}_0$ at $\bar{X}_0$. If $\phi(\bar{X}_0) < \phi(X_0)$ or if there is a minimum of ϕ between X_0 and $\bar{X}_0$, the approximate inverse is modified so that the smallest value of ϕ would have been attained in one step:

$$X_1 = X_0 - H_1\nabla_0$$

The difference matrix $H_1 - H_0$ is so chosen that step $X_1 - X_0$ is multiplied by a proper scalar and that all perpendicular steps would be unchanged. The process is repeated until the predicted change in position is less than some specified limit.

The method of modification depends on the following results from matrix algebra. If H_0 is a symmetric positive definite matrix, **V** any column vec-

tor, $\mathbf{V}^T$ its transpose, and b a scalar, then the matrix

$$H_1 = H_0 + b\frac{(H_0\mathbf{V})(H_0\mathbf{V})^T}{\mathbf{V}^T(H_0\mathbf{V})}$$

has the property that $H_1\mathbf{V} = (1 + b)H_0\mathbf{V}$ and, if $\mathbf{Y}$ is any vector such that $(H_0\mathbf{V})^T\mathbf{Y} = 0$, then $H_1\mathbf{Y} = H_0\mathbf{Y}$. Further, if Δ_0 is the determinant of H_0 and Δ_1 the determinant of H_1, then $\Delta_1 = (1 + b)\Delta_0$. Finally, if H_0 is singular and nullifies a vector $\mathbf{Z}$, then H_1 also nullifies the vector $\mathbf{Z}$.

To impose a linear constraint of the form $\sum a_i x_i = \alpha$, the initial matrix H_0 is chosen so that $H_0 a_i = 0$, and the point X_0 is chosen to satisfy the constraint condition.

Davidon's paper is devoted chiefly to the details of the flow diagrams for the execution of the method on a stored program computer, and it includes valuable details concerning interpolation for the minimum of the test function. The process outlined is terminated by random perturbation of the estimated locus as a check on significance.

Davidon also describes a second method for the progressive adjustment of a trial matrix H_0, a method equivalent to a conjugate gradient method if ϕ is a quadratic form in the independent variables. Starting from a point X_0 where the gradient is ∇_0 with an estimate H_0 of the inverse matrix, a matrix H_1 is sought such that the steps

$$\bar{X}_0 = X_0 - H_0\nabla_0$$

$$X_1 = \bar{X}_0 - H_1\bar{\nabla}_0$$

would be achieved by the single step

$$X_1 = X_0 - H_1\nabla_0$$

This can always be achieved by the adjustment

$$H_1 = H_0 - b(H_0\bar{\nabla}_0)(H_0\bar{\nabla}_0)^T$$

where

$$b = 1/\bar{\nabla}_0 \cdot H_0(\bar{\nabla}_0 - \nabla_0)$$

if H_0 is symmetrical and if b can be computed.

In the case of general potential and test functions derived from arbitrary nonlinear systems, this change might be excessive. Davidon lists rules by which the change in the determinant of H may be bounded.

Owing to the recent appearance of this method in the literature, no basis exists for a practical appraisal of it other than Davidon's comment that it had worked effectively in all applications up to the time of publication, and that it was believed to be faster and more likely to succeed than computing methods for solution of the same problem.

Conclusions

Three methods for solving systems of nonlinear equations seem to be reasonably effective: Newton's method, methods of descent, and Davidon's method.

Although they require more preparatory effort, Newton's and Davidon's methods are mathematically preferable to the descent methods because of their ultimate quadratic convergence and because the existence of multiple roots or indeterminacy can be detected easily. If computational procedures for the formal development of derivatives and the codes for their numerical evaluation were available, Newton's method would be preferred over Davidon's.

Descent methods, especially the downhill variety, are advantageous in their ease of coding and the relative ease of finding successive solutions, although their ultimate rate of convergence is only first order. As a result, the accuracy of the results may be inferior.

Each method reviewed may fail where an alternative method is effective. All the methods may converge slowly or occasionally may fail to converge, but difficulties with convergence can be removed by the judicious use of mixed methods.

References

1. Booth, A. C. 1955. Numerical Methods. : 154–160. Butterworth's Scientific Publications. London, England.
2. Householder, A. S. 1953. Principles of Numerical Analysis. : 132–137. McGraw-Hill. New York, N. Y.
3. Milne, W. E. 1949. Numerical Calculus. : 44–53. Princeton Univ. Press. Princeton, N. J.
4. Gordon, S., F. J. Zeleznik & V. Huff. 1959. A general method for automatic computation of equilibrium compositions and theoretical rocket performance of propellants. NASA TN D-132.
5. Ward, J. A. 1957. The downhill method of solving $f(Z) = 0$. J. Assoc. Comp. Mach. 4(2): 148–150.
6. Lance, G. N. 1959. Solution of algebraic and transcendental equations on an automatic digital computer. J. Assoc. Comp. Mach. 6(1): 97–101.
7. Davidon, W. C. 1959. Variable metric method for minimization. ANL-5990. Argonne National Laboratory. Lemont, Ill.