An iterative method for finding stationary values of a function of several variables

By M. J. D. Powell

An iterative method for finding stationary values of a function of several variables is described. In many ways it is similar to the method of steepest descents; however this new method has second order convergence.

1. Introduction

Eighteen months ago Rosenbrock (1960) published a paper in this journal on finding the greatest or least value of a function of several variables. A number of methods were listed and they all have first-order convergence. Six months ago Martin and Tee (1961) published a paper in which they mentioned gradient methods which have second-order convergence for finding the minimum of a quadratic positive definite function. In this paper will be described an iterative method which is not unlike the conjugate gradient method of Hestenes and Stiefel (1952), and which finds stationary values of a general function. It has second-order convergence, so near a stationary value it converges more quickly than Rosenbrock's variation of the steepest descents method and, although each iteration is rather longer because the method is applicable to a general function, the rate of convergence is comparable to that of the more powerful of the gradient methods described by Martin and Tee.

The efficiency of this new procedure is discussed, and two numerical examples are given as a comparison with the method of steepest descents and a variation of it. In addition there is a section on the problems which arise in programming the method.

2. Some Definitions and Assumptions

The number of independent variables of the function is defined to be n, the independent variables are $x_1, x_2, \ldots, x_n$ and the function is called $f(x)$. The independent variables define an n-dimensional space and the equations $f(x) = C$ define contours within the space. The stationary value we are trying to find is defined to be at ξ. It is assumed that the first and second derivatives of $f(x)$ exist and are continuous in the neighbourhood of ξ; the following notation is used for derivatives

$$f_i(x) = \frac{\partial}{\partial x_i} f(x)$$

and

$$f_{ij}(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x).$$

Then, as a consequence of the definitions and the assumptions,

$$f_i(\xi) = 0 \quad i = 1, 2, \ldots, n$$

and

$$f_{ij}(\xi) = f_{ji}(\xi).$$

Of course the assumptions have been made so that $f(x)$ can be approximated in the neighbourhood of the stationary value by the first few terms of the Taylor series expansion. We define the approximation to $f(x)$ to be $g(x)$, and therefore

$$g(x) = f(\xi) + \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \xi_i)(x_j - \xi_j) f_{ij}(\xi).$$

The iterative procedure will have second-order convergence if and only if when $f(x) \equiv g(x)$ then the iteration leads one from an arbitrary estimate of the stationary value, say η, to ξ in a single cycle. In the next section the method is described from a geometrical point of view and it is clear that the method does in fact have second-order convergence.

3. The Method

Throughout this section it is assumed that $f(x) = g(x)$ because if the method has second-order convergence in this simple case it will have second-order convergence in the most general case. The method is inductive, and we describe how the stationary value of the function can be found in the n-dimensional space if it can be found in any $(n-1)$-dimensional subspace. In such a subspace there will be a constraint on the variables $x_1 x_2, \ldots, x_n$ so the function will in effect be a function of $(n-1)$ independent variables. Such a description is sufficient because there are efficient methods of finding the stationary values of a function of a single variable. The method hinges on a corollary of the theorem that because $f(x)$ is quadratic in the independent variables any line which passes through ξ intersects the members of the family of contours $f(x) = C$ at equal angles. The corollary is that if the normal at t to the contour $f(x) = f(t)$ is parallel to the normal at t' to $f(x) = f(t')$ then the line joining t to t' passes through ξ.

To find the stationary value of $f(x)$ in n dimensions, given an initial estimate η, first find the line which passes through η and which is normal to the contour $f(x) = f(\eta)$. Proceed along this line to the point ϵ where the derivative of $f(x)$ with respect to the distance along the line is zero. In fact the point ϵ may be any point on the line which is a finite distance from η; however, if it is chosen in the way specified the convergence of the process is assured. Find the stationary value of $f(x)$ in the $(n-1)$-dimensional hyperplane which contains ϵ and which is such

that if t is a general point in the hyperplane then the line joining t to ϵ is perpendicular to the line joining ϵ to η. Say this point is δ. Then since the normal to $f(x) = f(\delta)$ at δ is parallel to the normal at η the required stationary value in the n-dimensional space will be that point on the line joining η to δ where the derivative of $f(x)$ with respect to the distance along the line is zero.

Just how the iteration proceeds should now be clear; however, we will illustrate the two-dimensional case, since a diagram may assist the reader to understand the procedure. In Fig. 1, A is the point η, B is the point ϵ, D is the point δ, C is the point ξ, AB is the tangent to the conic through B and is normal to the conic through A, and BD is a tangent to the conic through D and is perpendicular to AB. Clearly C lies on AD and B could have been chosen anywhere on AB.

If, as is usual, $f(x)$ is not equal to $g(x)$, the same recipe is used for each cycle, and, in general, ξ will not be found by a single iteration. The method will converge if there is no ambiguity in identifying the correct point on a line where the derivative of the function with respect to the distance along the line is zero. This will only prove troublesome if, within the region where the point is being sought, third and higher order terms in the Taylor series expansion of $f(x)$ about ξ swamp $g(x)$. In these cases it is probable that the value of η is not sufficiently close to the required stationary value of $f(x)$ to define it uniquely, and it is advisable to cause the program to report.

In the special case when a maximum or minimum is being sought, however, there is less likelihood of ambiguity because it will be known whether the required zero value of the derivative corresponds to a maximum or a minimum. Indeed, if the largest or smallest function value which is found within some finite region of the n-dimensional space is always chosen, the process will converge.

4. The Efficiency of the Method

The efficiency of the method described will depend upon the function $f(x)$, and it is encouraging to discuss the efficiency when $f(x) = g(x)$. In order to determine $g(x)$, $\frac{1}{2}(n+1)(n+2)$ pieces of information are required as this is the number of coefficients in the general form

$$g(x) = C + \sum_{i=1}^{n} b_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{i} a_{ij} x_i x_j.$$

In principle $g(x)$ could be found by calculating $\frac{1}{2}(n+1)(n+2)$ distinct function values of $f(x)$ and then solving the resultant linear simultaneous equations to determine the constant coefficients. This is likely to be a bad method because, if the points where the function values are evaluated all happen to be near ξ, the equations will probably be ill-conditioned and, if they are not near the stationary value, the process may fail to converge when $f(x) \neq g(x)$.

In theory it is only necessary to evaluate derivatives when applying the method of section (3) and, as it is not necessary to know the value of C to determine ξ,

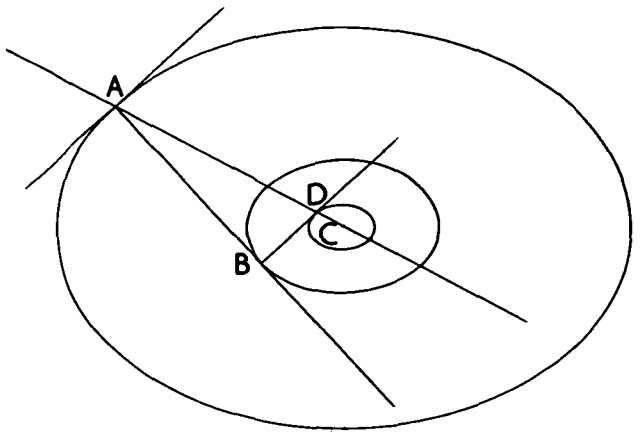

Fig. 1.—A cycle of the iteration in two dimensions

at least $\frac{1}{2}n(n+3)$ derivatives must be calculated. In the process described the steepest descent at a point has to be evaluated firstly in n dimensions, then in $(n-1)$ dimensions, . . ., and finally in one dimension. This requires at least $\frac{1}{2}n(n+1)$ derivatives. Also, on $(2n-1)$ occasions, knowing the derivative at a point on a line and knowing that if we assume $f(x) = g(x)$ we are assuming that the derivative varies linearly with the distance along the line, we have to calculate the point where the derivative is zero. Therefore on each occasion one more derivative on the line is sufficient to ensure second-order convergence. We have recognized that the first $(n-1)$ times in a cycle that this is done it is not essential, while it is essential to find the stationary value the last n times. Therefore we have to evaluate $\frac{1}{2}n(n+3)$ essential derivatives and $(n-1)$ unessential ones. The number of essential derivatives corresponds with the minimum number so, in a sense, we can consider the iterative method to be thoroughly efficient.

In cases when $f(x) \neq g(x)$ this method can be compared with the method of steepest descents. It is obvious that the method described in this paper will converge much faster near ξ, so this method will probably be superior if high accuracy is required. However, away from the stationary value, it is more difficult to judge between the two, and the comparison must be based on the choice of directions along which a stationary value is sought. Because the first n directions of a cycle in this method are chosen to be mutually orthogonal, this method will not ignore long shallow contours in the way the steepest-descents method is inclined to do. However, if there are no long shallow contours, the steps along the last few of the n orthogonal directions will almost certainly be less profitable than steps down the steepest descent. This disadvantage is unimportant if derivatives are calculated from function values, because r derivatives have to be evaluated to find the steepest descent from a point in an r-dimensional space, and while in the steepest-descents method r is always equal to n or $(n-1)$, in our method r takes integral values from n down to unity. Therefore, if derivatives are calculated from function values, the less profitable searches

will only take a fraction of the time of full steepest-descent searches. Unfortunately there will be no corresponding saving of time if derivatives are evaluated from analytic formulae and, in these cases, there may be situations in which the steepest-descents method finds the approximate position of the stationary value more quickly. However, the numerical examples given in section 6 suggest that the method of this paper is so much more powerful than that of steepest descents that it is always more suitable for a general program.

5. Some Practical Points of the Method

In this section the numerical problems which would be encountered in programming the method are pointed out. They are not peculiar to this method and they exist in all procedures to find stationary values in which derivatives have to be evaluated. As is usual the solutions of the problems are more satisfactory if derivatives of the function can be calculated from analytic formulae.

The first hurdle is to evaluate the steepest descent of $f(\mathbf{x})$ at $\boldsymbol{\eta}$, say, in an r-dimensional space which is contained in the n-dimensional space. It is always convenient to define the r-dimensional space by the directions of r mutually orthogonal directions through $\boldsymbol{\eta}$ which lie in the space, and we will define them to be $\boldsymbol{l}_i$, $i = 1, 2, \ldots, r$. Therefore, if $\boldsymbol{t}$ is a general point in the r-dimensional space, there will exist r unique coefficients C_i which are such that

$$\mathbf{t} = \boldsymbol{\eta} + \sum_{i=1}^{r} C_i \boldsymbol{l}_i.$$

The direction of the steepest descent is defined to be s so a general point on the steepest descent will be at $\boldsymbol{\eta} + \sigma s$. Furthermore there will exist r coefficients θ_i such that

$$s = \sum_{i=1}^{r} \theta_i \boldsymbol{l}_i.$$

If first derivatives of $f(\mathbf{x})$ could be evaluated from analytic formulae, $f_i(\boldsymbol{\eta})$ for $i = 1, 2, \ldots, n$ would be calculated and then

$$\theta_i = N \cdot \sum_{j=1}^{n} l_{ji} f_j(\boldsymbol{\eta}) \quad i = 1, 2, \ldots, r$$

would be worked out; N is a normalization factor. So in this case finding the steepest descent is straightforward but tedious. However, if derivatives have to be calculated from function values there are numerical difficulties.

To determine the coefficients θ_i it is necessary to calculate the derivative at $\boldsymbol{\eta}$ of $f(\mathbf{x})$ with respect to the distance along each of the r mutually orthogonal directions through $\boldsymbol{\eta}$. This is equivalent to calculating

$$\frac{\partial}{\partial \lambda} \cdot f(\boldsymbol{\eta} + \lambda \boldsymbol{l}_i) \text{ for } i = 1, 2, \ldots, r.$$

As the task of finding the derivative of a single variable numerically is discussed in all the textbooks on numerical analysis we will make just one more comment on this topic. If $\boldsymbol{\eta}$ is a poor approximation to $\boldsymbol{\xi}$, as it usually will be at the start of the iteration, it is often not worthwhile to attempt to evaluate the derivatives to high accuracy. Probably an estimate for a derivative based on just one more function value near $\boldsymbol{\eta}$ will suffice. However, if the derivative is evaluated from such an estimate the method of this paper will not have second-order convergence. Therefore, when it is thought that the iteration has nearly converged, the derivatives should be calculated from a formula in which there are no errors of the order of the second derivative. Indeed this is essential very near the stationary value in all methods, because all first derivatives are zero at $\boldsymbol{\xi}$. Because of the difficulty of recognizing rates of convergence it will often be worthwhile to use the more accurate formula for the derivatives on every third or fourth iteration, even if it appears that the best estimate of $\boldsymbol{\xi}$ is some way from the required stationary value.

The next problem which would be encountered is how to find the point on a line where the derivative of $f(\mathbf{x})$ with respect to the distance along the line is zero. Procedures for doing this have been given by both Booth (1957) and Rosenbrock (1960).

Finally, as in all iterative procedures, it is necessary to judge when the process has converged. This is difficult when searching naïvely for stationary values because near any stationary value $f(\mathbf{x})$ will be insensitive to changes in the variables $x_1, x_2, \ldots, x_n$. We will mention two possible and obvious criteria for convergence.

The first is to consider changes in the variables due to successive iterations and to be content when these changes are less than specified amounts.

The second, which is the better if derivatives can be determined from analytic formulae, is to find some point $\boldsymbol{e}$ such that

$$f_i(\boldsymbol{e}) < E_i \text{ for } i = 1, 2, \ldots, n$$

where the numbers E_i are specified before the start of the iteration.

6. Two Examples of the Method

In this section two numerical examples will be given which compare the method of this paper with the method of steepest descents and the modification of the method of steepest descents mentioned by Booth (1957). In each cycle of Booth's method the steepest descent at the starting point A is calculated and then, as in the steepest descents method, the stationary value on this steepest descent is found; say it is B. The variation is that B is only taken as the new approximation to the required stationary value on every fifth cycle. On the other cycles the new approximation is assumed to be at distance 0·9AB from A on AB. In both these examples the required minima and the required steepest descents are calculated exactly. For this reason it is not possible to make a direct comparison with Rosenbrock's method (Rosenbrock, 1960) which does not depend upon a knowledge of derivatives.

For the first example a function which Rosenbrock

Table 1

A Comparison in Two Dimensions

	STEEPEST DESCENTS		BOOTH'S VARIATION		NEW METHOD
n	$f(x_1, x_2)$	n	$f(x_1, x_2)$	n	$f(x_1, x_2)$
0	24·200	0	24·200	0	24·200
3	3·704	3	4·123	1	3·643
6	3·339	6	4·105	2	2·898
9	3·077	9	4·086	3	2·195
12	2·869	12	4·067	4	1·412
15	2·689	15	4·049	5	0·831
18	2·529	18	4·029	6	0·432
21	2·383	21	4·008	7	0·182
24	2·247	24	3·988	8	0·052
27	2·118	27	3·960	9	0·004
30	1·994	30	3·864	10	5×10^{-5}
33	1·873	33	3·746	11	8×10^{-9}

chose has been taken. It is

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

and the starting value for the iteration is at (−1·2, 1·0). The results are given in Table 1. (In Tables 1 and 2, n is the number of iterations necessary to arrive at the given value of f.) Because in two dimensions the method of this paper requires a minimum to be calculated in three distinct directions in order to complete a cycle, the values of $f(x_1, x_2)$ tabulated in the second and fourth columns of Table 1 are those calculated after every third cycle. The values of x_1 and x_2 for the final row of the table are

$$x_1 = -0{\cdot}3634 \qquad x_2 = 0{\cdot}1441$$
$$x_1 = -0{\cdot}9336 \qquad x_2 = 0{\cdot}8632$$

and

$$x_1 = 1{\cdot}0001 \qquad x_2 = 1{\cdot}0001$$

for the steepest descents method, Booth's method, and the new method respectively.

Three comments should be made on these results. The first is that in this example the method of this paper is definitely superior to the other two methods. The second is that the fact that Booth's method appears to be slower to converge than the steepest-descents method is fortuitous, and in fifty iterations it has reduced $f(x_1, x_2)$ to 0·71. In the same number of iterations the steepest-descents method has found values of x_1 and x_2 such that $f(x_1, x_2) = 1{\cdot}20$. Thirdly, in two dimensions the steepest descents method is equivalent to varying x_1 and then x_2 for some orientation of the axes, so it is not particularly effective or interesting.

For the second example the function

$$f(x_1, x_2, x_3, x_4) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$$

was chosen. In all three methods the iteration to seek the minimum at (0, 0, 0, 0) was started at (3, −1, 0, 1) and this example was supposed to illustrate that, away from the minimum where fourth order terms are significant, and where there are more than two variables, the steepest-descents method can be as good as the method of this paper. It completely failed to do this as Table 2 shows. The reason that the method of this paper does not have second-order convergence in this case is that near the minimum $g(x_1, x_2, x_3, x_4) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2$, so it does not determine the minimum uniquely.

Table 2

A Comparison in Four Dimensions

	STEEPEST DESCENTS		BOOTH'S VARIATION		NEW METHOD
n	$f(x_1, x_2, x_3, x_4)$	n	$f(x_1, x_2, x_3, x_4)$	n	$f(x_1, x_2, x_3, x_4)$
0	215·000	0	215·000	0	215·000
7	6·355	7	5·352	1	0·009
14	3·743	14	0·620	2	9×10^{-5}
21	2·269	21	0·135	3	2×10^{-6}
28	1·420	28	0·051	4	2×10^{-6}
35	0·919	35	0·009	5	1×10^{-6}
42	0·614	42	0·008	6	5×10^{-8}
49	0·423	49	0·008	7	4×10^{-9}

Table 3

The Second Iteration of the New Method

POINT	x_1	x_2	x_3	x_4	f
A	0·1266	−0·0123	0·1455	0·1428	0·00852
B	0·1263	−0·0104	0·1337	0·1441	0·00699
C	0·1260	−0·0124	0·1333	0·1435	0·00659
D	0·1259	−0·0111	0·1331	0·1391	0·00632
E	0·1229	−0·0107	0·1332	0·1392	0·00632
F	0·1229	−0·0107	0·1332	0·1392	0·00632
G	0·1166	−0·0114	0·1323	0·1303	0·00583
H	0·0396	−0·0044	0·0300	0·0332	0·00009

The steps of the second iteration are given in Table 3. This particular iteration has been chosen because it does demonstrate very clearly how powerful the choice of directions can be. The first and second steps from A to B and B to C correspond to the ordinary steepest-descents method; the steps from C to D and from D to E correspond to steepest descents in two and one dimensions, respectively. Because, and this is a common occurrence, E nearly coincides with D, the minimum on CE, F, nearly coincides with E. The finding of G, the minimum on BF, gives a slight improvement in $f(x_1, x_2, x_3, x_4)$, and the final step, finding H on AG, reduces the value of $f(x_1, x_2, x_3, x_4)$ handsomely.

7. Conclusion

Both these examples, and other applications which the author has made, show that this new method is more

powerful than the usual methods of finding a stationary value of a function of several variables. It takes a little while to program, but so will any iterative method which has second-order convergence near a stationary value and which will converge from a poor approximation to the stationary value.

References

BOOTH, A. D. (1957). *Numerical Methods*, London: Butterworths, pp. 95–100, p. 158.

HESTENES, M. R., and STIEFEL, E. (1952). "Methods of Conjugate Gradients for Solving Linear Systems," *J. Res. N.B.S.*, Vol. 49, p. 409.

MARTIN, D. W., and TEE, G. J. (1961). "Iterative Methods for Linear Equations with Symmetric Positive Definite Matrix," *The Computer Journal*, Vol. 4, p. 242.

ROSENBROCK, H. H. (1960). "An Automatic Method for finding the Greatest or Least Value of a Function," *The Computer Journal*, Vol. 3, p. 175.

Book reviews

A Fortran Program for Elastic Scattering Analyses with the Nuclear Optical Model, by MICHAEL A. MELKANOFF, DAVID S. SAXON, JOHN S. NODVIK and DAVID G. CANTOR, 1961; 116 pp. (Berkeley, and Los Angeles: *University of California Press*, $4.50; London: *Cambridge University Press*, 34s. 0d.)

To quote from the introduction, "the purpose of the present report is to describe in complete detail a FORTRAN code named Program SCAT 4 written by the UCLA group in order to analyze elastic scattering of various particles against complex nuclei by means of the diffuse surface optical model of the nucleus." The publication is similar to, but more elaborate than, the many technical reports which come from the large scientific institutions such as the Atomic Energy laboratories, aircraft companies and defence research establishments, especially in America where people seem to be particularly good at writing up work. There is a full account of the mathematical basis of the calculation; a description of the program, first in general terms, then in detail, routine by routine; a listing of the program which occupies 40 pages and runs to 21 routines totalling nearly 2,000 FORTRAN statements; the input data and output listing of a simple calculation, the scattering of 9·75 MeV protons by copper; and a bibliography of related calculations. There is also an offer to send the program card deck to anyone who is willing to pay the mailing charges.

The essence of the calculation is the numerical integration of the Schrödinger equation for the system, reduced to a set of ordinary differential equations in a single radial variable by a process of expansion in eigen-functions, followed by a matching of the numerical solution to an asymptotic solution expressed in terms of Coulomb wave functions. The matching process leads to the determination of some important parameters called *phase shifts* which, together with the computed solution, enable one to calculate the cross-section for the reaction in question. A very practical consequence is that one can compute from basic nuclear data the values of certain physical quantities which are needed, for example, in the design of nuclear reactors; in fact, this kind of computational procedure is already beginning to supplement programmes of experimental work, and is quite likely later on to replace a good deal of this.

The attention to detail in the report is very impressive. It is written for the man who wants to use the program and who may wish to extend or modify it, and the impression one gets is that every point has been considered. The mathematical section (33 pages) is concentrated and is certainly not for the uninitiated, but it does give a complete account of the analyses so that one can find out just what has been put into the program; in addition to the mathematical physical arguments, it goes into detail about the numerical processes, including a full account of the special Runge–Kutta process used for the numerical integration of the differential equations. The description of the program itself is equally complete, giving full details of the structure and operation of every routine and of its relations with whatever others it calls or is called by. The writing is terse but very clear, and the publication fulfils admirably its purpose of providing a technical reference manual to a specialized and complicated piece of work.

The corresponding reports issued by industrial and government laboratories are mostly circulated privately, although some, especially those coming from the UKAEA in England and the AEC in America, can be bought by anyone who knows of their existence. Open publication, as in this case, seems to me to be a move to be welcomed; a very large amount of thought, effort and experience has gone into the construction of a computer program of the size and complexity of one such as SCAT 4, and it is all to the good that as many people as possible should be able to take advantage of it. There is quite literally a universal interest in the calculation described here, and because the program language FORTRAN is now used on many of the larger machines, the program itself can be used quite widely just as it stands with enormous saving in scientific effort. Clearly, the value of a publication of this kind is very greatly increased if a more widely accepted language is used; the actual value for money represented by this manual at a price of around £2 is quite remarkable.

J. HOWLETT.