An efficient method for finding the minimum of a function of several variables without calculating derivatives

By M. J. D. Powell*

A simple variation of the well-known method of minimizing a function of several variables by changing one parameter at a time is described. This variation is such that when the procedure is applied to a quadratic form, it causes conjugate directions to be chosen, so the ultimate rate of convergence is fast when the method is used to minimize a general function. A further variation completes the method, and it ensures that the convergence rate from a bad approximation to a minimum is always efficient. Practical applications of the procedure have proved to be very satisfactory, and numerical examples are given in which functions of up to twenty variables are minimized.

1. Introduction

There is no need to stress the importance of the problem of finding values of n parameters $x_1, x_2, \ldots, x_n$, so that the value of a function of these parameters, $f(x_1, x_2 \ldots, x_n)$, is a minimum. Fletcher and Powell (1963) have published an account of Davidon's (1959) conjugate gradient procedure, which has been found very satisfactory when first derivatives of the function are available. However, it is frequently the case that it is laborious or practically impossible to calculate first derivatives, so there is a definite need for minimization procedures which do not require them. Such procedures include Rosenbrock's (1960), the simplex method of Himsworth, Spendley and Hext (1962), Smith's method based on conjugate directions (1962) and, of course, the well-known one of changing one variable at a time. This last procedure and some variations of it have been described by Spang (1962).

A new method has been developed because, of the above procedures, only Smith's will find the minimum of a general quadratic form in a finite number of steps, and it is essential to be able to find such a minimum to ensure ultimate fast convergence for a general function having continuous second derivatives. Smith's method has the disadvantage that the first variable, x_1, is changed n times more frequently than the nth, so it is a little slow in starting from a bad approximation to the minimum. The method to be described has neither of these deficiencies, and it has the further advantage that it is practically invariant under linear transformations of the co-ordinate space. Applications of the new procedure have been entirely satisfactory, and numerical comparisons suggest that it is more efficient than Rosenbrock's method, which is possibly the most used at the present time for minimization without derivatives. However, unlike Rosenbrock's procedure, the new method is not designed to recognize constraints on the variables.

The procedure is defined and the convergence properties are proved in the next four sections. Section 2 describes how the method of minimization which changes one variable at a time is modified to find the minimum of a quadratic form in a finite number of steps; the proof of the efficacy of the modification is given in Section 3. In Section 4 the description of an iteration of the procedure is completed by the more practical modification which ensures that the rate of convergence is reasonable from a bad approximation to the minimum. The theory of this is discussed in Section 5, and, from this discussion, a result is derived which may be used to force efficient convergence in many iterative procedures which search down n independent directions on each iteration. In Section 6 the criterion for ultimate convergence is considered, while in Section 7 a simple application of the procedure is presented. The method chosen for finding a minimum along a line is described in Section 8, in Section 9 some program details are given, and in Section 10 more numerical examples are set out. The paper is concluded with a discussion in which an opinion is given on the effectiveness of the method, the situations in which other known methods may be preferable and the fields in which it is thought that existing methods may be improved.

2. How conjugate directions are chosen

Each iteration of the procedure commences with a search down n linearly independent directions $\xi_1, \xi_2, \ldots, \xi_n$, starting from the best known approximation to the minimum, p_0. These directions are chosen to be the co-ordinate directions initially, so the start of the first iteration is identical to an iteration of the method which changes one parameter at a time. This latter method is modified to generate conjugate directions by making each iteration define a new direction, ξ, and choosing the linearly independent directions for the next iteration to be $\xi_2, \xi_3, \ldots, \xi_n, \xi$. The way in which ξ is defined ensures that, if a quadratic is being minimized, after k iterations the last k of the n directions chosen for the $(k+1)$th iteration are mutually conjugate. After n iterations all the directions are mutually conjugate, and it will be proved that in consequence the exact minimum of the quadratic is found.

* *Applied Mathematics Group, Theoretical Physics Division, A.E.R.E. Harwell, Berks.*

Thus an iteration of the basic procedure is as follows.

(i) For $r = 1, 2, \ldots, n$ calculate λ_r so that $f(\boldsymbol{p}_{r-1}+\lambda_r\xi_r)$ is a minimum and define $\boldsymbol{p}_r = \boldsymbol{p}_{r-1} + \lambda_r\xi_r$.
(ii) For $r = 1, 2, \ldots, n-1$ replace ξ_r by ξ_{r+1}.
(iii) Replace ξ_n by $(\boldsymbol{p}_n - \boldsymbol{p}_0)$.
(iv) Choose λ so that $f(\boldsymbol{p}_n + \lambda\{\boldsymbol{p}_n - \boldsymbol{p}_0\})$ is a minimum and replace $\boldsymbol{p}_0$ by $\boldsymbol{p}_0 + \lambda\{\boldsymbol{p}_n - \boldsymbol{p}_0\}$.

3. Proof that a quadratic is minimized

If the quadratic function to be minimized is

$$f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\boldsymbol{x} + c,$$

then the directions $\boldsymbol{p}$ and $\boldsymbol{q}$ are defined to be conjugate if

$$\boldsymbol{p}\boldsymbol{A}\boldsymbol{q} = 0.$$

For a minimum to be defined it is necessary for the matrix $\boldsymbol{A}$ to be positive definite, but the proof does not make use of this fact. Consequently the procedure of Section 2 will find the stationary point of any quadratic function, if each linear search finds a maximum or minimum. The proof requires two theorems.

Theorem 1. If $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m$, $m \leqslant n$, are mutually conjugate directions, then the minimum of the quadratic $f(\boldsymbol{x})$, where $\boldsymbol{x}$ is a general point in the m-dimensional space containing $\boldsymbol{x}_0$ and the directions $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m$, may be found by searching along each of the directions once only.

The required minimum is the point

$$\boldsymbol{x}_0 + \sum_{i=1}^{m} \alpha_i\boldsymbol{q}_i,$$

where the parameters α_i, $i = 1, 2, \ldots, m$, are such as to minimize

$$f\left\{\boldsymbol{x}_0 + \sum_{i=1}^{m} \alpha_i\boldsymbol{q}_i\right\}$$

$$= \sum_{i=1}^{m} \{\alpha_i^2\boldsymbol{q}_i\boldsymbol{A}\boldsymbol{q}_i + \alpha_i\boldsymbol{q}_i.(2\boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{b})\} + f(\boldsymbol{x}_0).$$

There are no terms in $\alpha_i\alpha_j$, $i \neq j$, because of the mutual conjugacy of the directions. Consequently the effect of searching in the direction $\boldsymbol{q}_i$ is to find α_i to minimize

$$\{\alpha_i^2\boldsymbol{q}_i\boldsymbol{A}\boldsymbol{q}_i + \alpha_i\boldsymbol{q}_i.(2\boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{b})\},$$

and the resultant value of α_i is independent of the other terms of the function. Hence searching in each of the directions once only will find the absolute minimum in the space.

Theorem 2. If $\boldsymbol{x}_0$ is the minimum in a space containing the direction $\boldsymbol{q}$, and $\boldsymbol{x}_1$ is also the minimum in such a space, then the direction $(\boldsymbol{x}_1 - \boldsymbol{x}_0)$ is conjugate to $\boldsymbol{q}$.

By definition $\dfrac{\partial}{\partial\lambda}\{f(\boldsymbol{x}_0 + \lambda\boldsymbol{q})\} = 0$ at $\lambda = 0$.

Therefore $2\lambda\boldsymbol{q}\boldsymbol{A}\boldsymbol{q} + \boldsymbol{q}.(2\boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{b}) = 0$, $\lambda = 0$.

Also $2\lambda\boldsymbol{q}\boldsymbol{A}\boldsymbol{q} + \boldsymbol{q}.(2\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}) = 0$, $\lambda = 0$.

Hence $\boldsymbol{q}\boldsymbol{A}(\boldsymbol{x}_1 - \boldsymbol{x}_0) = 0$,

which is the condition for conjugacy.

The convergence to the minimum of a quadratic function in n iterations will be proved by induction, so it will be assumed that k iterations have been completed and that the directions $\xi_{n-k+1}, \xi_{n-k+2}, \ldots, \xi_n$, defined for the $(k+1)$th iteration, are mutually conjugate. As these were the last k directions of search, applying Theorem 1, the starting approximation for the $(k+1)$th iteration, $\boldsymbol{p}_0$, is the minimum in a space containing the directions. By Theorem 1 again, the point $\boldsymbol{p}_n$, defined in the $(k+1)$th iteration, is also the minimum in such a space. Hence, applying Theorem 2, the new direction defined by the iteration is conjugate to $\xi_{n-k+1}, \xi_{n-k+2}, \ldots, \xi_n$, so the general step of the induction is proved.

The point $\boldsymbol{p}_0$, defined for the second iteration, and the consequent $\boldsymbol{p}_n$ are both minima in the direction ξ_n, so the second iteration yields a pair of conjugate directions, thus commencing the induction. After n iterations all the directions of search are mutually conjugate, so, by Theorem 1, the required minimum will have been found.

4. Ensuring reasonable convergence

The basic procedure described in Section 2 is modified to ensure that the rate of convergence to the minimum is satisfactory, even when the initial approximation is very poor. The reason why a change has to be made is that on occasions the basic procedure may choose nearly dependent directions, and this possibility has been found to be serious if the function to be minimized depends on more than five variables. In particular, in the notation of Section 2, if λ_1 is zero the resultant directions will not span the full parameter space. Therefore the modification described in this section allows a direction other than ξ_1 to be discarded, so that the new direction will always contain an appreciable component of that which is lost.

In Section 5 it will be shown that sometimes it is unwise to replace any one of $\xi_1, \xi_2, \ldots, \xi_n$ by ξ, so the modification allows the old set of linearly independent directions to be used again. An iteration of the recommended procedure is as follows.

(i) For $r = 1, 2, \ldots, n$ calculate λ_r so that $f(\boldsymbol{p}_{r-1} + \lambda_r\xi_r)$ is a minimum and define $\boldsymbol{p}_r = \boldsymbol{p}_{r-1}+\lambda_r\xi_r$.
(ii) Find the integer m, $1 \leqslant m \leqslant n$, so that $\{f(\boldsymbol{p}_{m-1}) - f(\boldsymbol{p}_m)\}$ is a maximum, and define $\Delta = f(\boldsymbol{p}_{m-1}) - f(\boldsymbol{p}_m)$.
(iii) Calculate $f_3 = f(2\boldsymbol{p}_n - \boldsymbol{p}_0)$, and define $f_1 = f(\boldsymbol{p}_0)$ and $f_2 = f(\boldsymbol{p}_n)$.
(iv) If either $f_3 \geqslant f_1$ and/or

$$(f_1 - 2f_2 + f_3).(f_1 - f_2 - \Delta)^2 \geqslant \tfrac{1}{2}\Delta(f_1 - f_3)^2$$

use the old directions $\xi_1, \xi_2, \ldots, \xi_n$ for the next iteration and use $\boldsymbol{p}_n$ for the next $\boldsymbol{p}_0$, otherwise

(v) defining $\xi = (p_n - p_0)$, calculate λ so that $f(p_n + \lambda\xi)$ is a minimum, use $\xi_1, \xi_2, \ldots, \xi_{m-1}, \xi_{m+1}, \xi_{m+2}, \ldots, \xi_n, \xi$ as the directions and $p_n + \lambda\xi$ as the starting point for the next iteration.

A consequence of the above modification is that one of the mutually conjugate directions may be thrown away, so that more than n iterations are required to find the exact minimum of a quadratic. This is unfortunate, but the next section should convince the reader that the modification is of value; in fact it was found to be essential, to minimize a function of twenty variables.

5. Proof that the convergence is always efficient

The modification given in the last section is based on a very useful result. It is that, if the directions $\xi_1, \xi_2, \ldots, \xi_n$ are scaled so that, in the notation of Section 3,

$$\xi_i A \xi_i = 1, \quad i = 1, 2, \ldots, n,$$

the determinant of the matrix whose columns are the vectors ξ_i takes its maximum value if and only if the directions are mutually conjugate. This is proved in the following way.

Any set of similarly scaled mutually conjugate directions, $\eta_1, \eta_2, \ldots, \eta_n$, is chosen. By definition it will have the property that

$$\eta_i A \eta_j = \delta_{ij} \quad i = 1, 2, \ldots, n; j = 1, 2, \ldots, n.$$

There is a transformation linking ξ and η of the form

$$\xi_i = \sum_{j=1}^{n} U_{ij}\eta_j,$$

and the determinant of the matrix having columns ξ_i is equal to the determinant of the matrix having columns η_i multiplied by the determinant of the transformation matrix $\boldsymbol{U}$. Now

$$\xi_i A \xi_j = \sum_{k=1}^{n} \sum_{l=1}^{n} U_{ik} U_{jl} \eta_k A \eta_l$$

$$= \sum_{k=1}^{n} U_{ik} U_{jk},$$

and, in particular,

$$\sum_{k=1}^{n} U_{ik} U_{ik} = 1, \quad i = 1, 2, \ldots, n.$$

Hence the determinant of $\boldsymbol{U}$ does not exceed unity, and it equals unity only if $\boldsymbol{U}$ is an orthogonal matrix. In this case

$$\xi_i A \xi_j = \delta_{ij},$$

so the directions ξ_i are mutually conjugate.

The consequence of the above theorem is that $\xi_1, \xi_2, \ldots, \xi_n$ should be chosen to make the corresponding determinant as large as possible, and this is a more powerful criterion than one of orthogonality because it is independent of linear transformations of the parameter space. The criterion is applied by using the new direction, ξ, defined by an iteration, if it can cause the determinant to increase, and by rejecting the direction the replacement of which causes the new determinant to be largest. It will now be proved that the direction which should be discarded, if any, is ξ_m, $1 \leqslant m \leqslant n$, where m is such that $\{f(p_{m-1}) - f(p_m)\}$ is a maximum.

Because $f(p_i)$ is a minimum in the direction ξ_i, if ξ_i is scaled so that

$$\xi_i A \xi_i = 1,$$

the displacement from p_{i-1} to p_i is

$$\sqrt{[f(p_{i-1}) - f(p_i)]} \,.\, \xi_i = \alpha_i \,.\, \xi_i, \text{ say.}$$

The direction defined by the iteration is

$$p_n - p_0 = \alpha_1\xi_1 + \alpha_2\xi_2 + \ldots + \alpha_n\xi_n,$$

so, if $p_n - p_0 = \mu\xi_p$, where

$$\xi_p A \xi_p = 1,$$

the effect of replacing the column vector ξ_i by the vector ξ_p is to multiply the determinant of directions by α_i/μ. Consequently the direction to be discarded, if any, is that for which α_i is largest, and this is the direction ξ_m.

Obviously this replacement should be made only if $\alpha_m \geqslant \mu$, and to calculate μ the function values f_1, f_2 and f_3, defined in Section 4, are used. Because these three function values are equally spaced along ξ_p, the predicted stationary value of the function along the new direction is

$$f_s = f_2 - \frac{1}{8}\frac{(f_1 - f_3)^2}{(f_1 - 2f_2 + f_3)},$$

and this value is predicted to be a minimum if

$$(f_1 - 2f_2 + f_3) > 0.$$

If the above second difference term is negative, a new direction should certainly be defined, otherwise μ is predicted as

$$\sqrt{(f_1 - f_s)} \pm \sqrt{(f_2 - f_s)},$$

the plus or minus sign depending on whether f_3 is greater or less than f_1. In the former case it is straightforward to show that the old directions should be used again; in the latter case new directions should be defined only if, in the notation of Section 4,

$$\sqrt{\Delta} \geqslant \sqrt{(f_1 - f_s)} - \sqrt{(f_2 - f_s)}.$$

The above results have been condensed in the criterion that $p_n - p_0$ should *not* be used for the next iteration if and only if either $f_3 \geqslant f_1$ and/or

$$(f_1 - 2f_2 + f_3)(f_1 - f_2 - \Delta)^2 \geqslant \tfrac{1}{2}\Delta(f_1 - f_3)^2.$$

Because the recommended procedure cannot cause the determinant of directions to decrease, the efficiency of the directions of search $\xi_1, \xi_2, \ldots, \xi_n$ is never less than

that of the original co-ordinate directions. If the co-ordinate directions are poor, improved directions will be found without difficulty. Therefore it is asserted that the rate of convergence is always reasonable.

6. The criterion for ultimate convergence

Ideally the ultimate convergence criterion would cause the iterative procedure to end as soon as the differences between the predicted values of the variables, x_i, and their actual values at the true minimum were less than given amounts, ε_i, $i = 1, 2, \ldots, n$. However, it is believed that it is impossible to choose such a convergence criterion which is effective for the most general function having continuous second derivatives, so a compromise has to be made between stopping the iterative procedure too soon and calculating $f(x_1, x_2, \ldots, x_n)$ an unnecessarily large number of times.

The obvious criterion is to assume convergence when an iteration causes each variable to be changed by less than the required accuracy. It is usually efficacious because the ultimate rate of convergence is better than linear, but it sometimes terminates the procedure before the required accuracy has been achieved. In these cases there are directions along which the function varies very slowly, and, because the method may not allow the function to increase, it must make just small changes in the variables to detect these directions.

As the procedure has been coded as a general subroutine a very safe convergence criterion was chosen. Consequently it has caused many unnecessary function evaluations to be requested, but it has not failed to yield the required accuracy. It is as follows:

(i) Apply the normal procedure until an iteration causes the change in every variable to be less than one-tenth of the required accuracy, say the resultant point is $\boldsymbol{a}$.

(ii) Increase every variable by ten times the required accuracy.

(iii) Apply the normal procedure until an iteration causes the change in every variable to be less than one tenth of the required accuracy again. Say the resultant point is $\boldsymbol{b}$.

(iv) Find the minimum on the line through $\boldsymbol{a}$ and $\boldsymbol{b}$, say it is $\boldsymbol{c}$.

(v) Assume ultimate convergence if the components of $(\boldsymbol{a} - \boldsymbol{c})$ and $(\boldsymbol{b} - \boldsymbol{c})$ are all less than one-tenth of the required accuracy in the corresponding variables, otherwise

(vi) include the direction $(\boldsymbol{a} - \boldsymbol{c})$ in place of ξ_1, and restart the procedure from (i).

It is based on the assumption that, if the function varies slowly along some direction, the procedure will converge to a point on this direction. Therefore two different starting values are likely to yield the required minimum or two different points on such a direction, and, in the latter case, the direction is incorporated in the directions of search.

Stage (vi) may be questioned because a check has not been made that the determinant of directions has increased. To apply this check requires one to solve a set of linear equations as well as obtaining estimates of the relative magnitudes of the directions. Because the new direction is almost certainly a good one, it is considered that extra sophistication is unnecessary.

7. The procedure applied to a function of three variables

To illustrate the elements of the method a maximum of the function

$$f = \frac{1}{1 + (x - y)^2} + \sin\{\tfrac{1}{2}\pi yz\} + \exp\left\{-\left(\frac{x + z}{y} - 2\right)^2\right\}$$

will be found. The function is bounded, and its maxima occur at

$$x = y = z = \pm\sqrt{(4n + 1)},\ n \text{ integral.}$$

The starting point for the procedure is chosen to be (0, 1, 2), and the progress of the first six iterations is set out in Table 1.

The most striking feature of this table is that the function attains the value 2·0000 very quickly, it remains at this value for six linear searches, and then, on the last search of the third iteration, the progress is remarkable. The reason is that the first iteration defines values of the parameters such that the exponential term of the function is practically negligible. Therefore searching along x on the second iteration maximizes the first term of the function, searching along y maximizes the second term, and consequently the function value is 2·0000. Because the function depends primarily upon $(x - y)$ and yz at this point, there is a direction along which it varies very slowly, so, as has been asserted, the changes in the parameters are small although the best approximation is far from the required minimum. To determine this direction a set of mutually conjugate directions is built up, the last of the set being $\boldsymbol{n}_2$ which is forced to be conjugate to both z and $\boldsymbol{n}_1$. Therefore $\boldsymbol{n}_2$ is the direction along which the function varies slowly, so a search along it results in a spectacular change in the variables. After the third iteration moderately efficient directions have been defined, and the resultant convergence to the minimum is very satisfactory. Note that an inadequate convergence criterion might have accepted the values of the variables at the end of iteration No. 2.

8. Finding the minimum along a line

To find the minimum on a line, the following data must be provided:

(i) a point on the line, $\boldsymbol{p}$,

(ii) the direction of the line, ξ,

(iii) an upper bound to the length of step along the line, m,

(iv) an order of magnitude of the length of step along the line, q, assumed to be less than m, and

(v) the accuracy to which the minimum is required, e.

The method of minimization must be such as to find the minimum of a quadratic form, so it is primarily based on the quadratic defined by three function values.

Initially $f(\boldsymbol{p})$ and $f(\boldsymbol{p}+q\boldsymbol{\xi})$ are calculated, and then either $f(\boldsymbol{p}-q\boldsymbol{\xi})$ or $f(\boldsymbol{p}+2q\boldsymbol{\xi})$ is worked out depending on whether $f(\boldsymbol{p})$ is less than or greater than $f(\boldsymbol{p}+q\boldsymbol{\xi})$. These three function values are now used in the general formula which predicts the turning value of the quadratic defined by $a, f(\boldsymbol{p}+a\boldsymbol{\xi})$, $b, f(\boldsymbol{p}+b\boldsymbol{\xi})$, c and $f(\boldsymbol{p}+c\boldsymbol{\xi})$ to be at $(\boldsymbol{p}+d\boldsymbol{\xi})$, where

$$d=\frac{1}{2}\cdot\frac{(b^2-c^2)f_a+(c^2-a^2)f_b+(a^2-b^2)f_c}{(b-c)f_a+(c-a)f_b+(a-b)f_c}.$$

It is a minimum if

$$\frac{(b-c)f_a+(c-a)f_b+(a-b)f_c}{(a-b)(b-c)(c-a)}<0.$$

If the turning value is predicted to be a maximum, or if the value of d is such that to calculate $f(\boldsymbol{p}+d\boldsymbol{\xi})$ a step greater than m must be taken, the maximum allowed step is taken in the direction of decreasing f, and the function value at the point which is furthest from $(\boldsymbol{p}+d\boldsymbol{\xi})$ is discarded, so the prediction may be repeated.

Otherwise d is compared with a, b and c, and, if it is within the required accuracy of one of them, that point is chosen as the minimum. If it is not, $f(\boldsymbol{p}+d\boldsymbol{\xi})$ is calculated so that the quadratic prediction may be repeated; the function value which is thrown away out of $f(\boldsymbol{p}+a\boldsymbol{\xi})$, $f(\boldsymbol{p}+b\boldsymbol{\xi})$ and $f(\boldsymbol{p}+c\boldsymbol{\xi})$ is normally the greatest, but it is not if rejecting a smaller one can yield a definite bracket on a minimum, which would not be obtained otherwise.

In order to reduce the number of times $f(x_1, x_2, \ldots, x_n)$ has to be calculated, advantage may be taken of the fact that three function values are sufficient to predict

$$\frac{\partial^2}{\partial\lambda^2}\{f(\boldsymbol{p}+\lambda\boldsymbol{\xi})\}.$$

The prediction of the second derivative is

$$D=-2\cdot\frac{(b-c)f_a+(c-a)f_b+(a-b)f_c}{(a-b)(b-c)(c-a)}$$

so, if after finding the minimum in the direction $\boldsymbol{\xi}$ the components of $\boldsymbol{\xi}$ are scaled by $1/\sqrt{D}$, the next time a minimum is sought in the same direction the unit second deviative may be used. In this case just $f(\boldsymbol{p})$ and $f(\boldsymbol{p}+q\boldsymbol{\xi})$ are sufficient to predict the minimum to be at $(\boldsymbol{p}+d\boldsymbol{\xi})$,

$$d=\tfrac{1}{2}q-\frac{f(\boldsymbol{p}+q\boldsymbol{\xi})-f(\boldsymbol{p})}{q}.$$

The criterion given in Section 4 is such that, if a new direction is defined, three function values on this direction will have been calculated, so a prediction of the minimum may be made immediately. Therefore, it is only during the first n linear searches of the first iteration of the procedure that it is necessary to calculate two function values, in addition to $f(\boldsymbol{p})$, for the initial prediction of the minimum in the direction $\boldsymbol{\xi}$.

Table 1

A function of three variables

ITERATION	DIRECTION	x	y	z	f
0	—	0·0000	1·0000	2·0000	1·5000
1	x	0·3674	1·0000	2·0000	1·5879
1	y	0·3674	0·4799	2·0000	1·9857
1	z	0·3674	0·4799	2·0827	1·9876
2	x	0·4799	0·4799	2·0827	2·0000
2	y	0·4799	0·4802	2·0827	2·0000
2	z	0·4799	0·4802	2·0821	2·0000
2	n_1	0·4801	0·4802	2·0821	2·0000
3	y	0·4801	0·4803	2·0821	2·0000
3	z	0·4801	0·4803	2·0815	2·0000
3	n_1	0·4802	0·4803	2·0815	2·0000
3	n_2	0·7449	0·7511	0·8648	2·8320
4	z	0·7449	0·7511	0·9217	2·8387
4	n_1	0·6401	0·7509	0·9222	2·8670
4	n_2	0·6357	0·7463	0·9426	2·8683
4	n_3	0·5505	0·7426	1·0033	2·8768
5	z	0·5505	0·7426	1·0472	2·8813
5	n_2	0·5581	0·7503	1·0125	2·8853
5	n_3	0·5627	0·7505	1·0092	2·8853
5	n_4	0·7995	0·9050	1·1236	2·9731
6	n_2	0·8159	0·9218	1·0482	2·9870
6	n_3	0·8656	0·9239	1·0127	2·9904
6	n_4	0·9272	0·9641	1·0424	2·9968
6	n_5	0·9969	0·9964	0·9982	3·0000

9. Other programming details

The procedure of this paper has been coded, in FORTRAN, for the IBM 7030 computer. It is programmed as a FORTRAN subroutine so a number of parameters are communicated through the calling sequence. These are

(i) the number of variables, n,

(ii) initial values of the variables $x_1, x_2, \ldots, x_n$,

(iii) the requested accuracy in the components of the calculated position of the minimum, $\varepsilon_1, \varepsilon_1, \ldots, \varepsilon_n$,

(iv) a scalar, E, which defines m for each linear search as the largest number such that the components of $m\xi$ do not exceed $E\varepsilon_1, E\varepsilon_2, \ldots, E\varepsilon_n$,

(v) a parameter controlling the amount of information printed while the procedure is being carried out,

(vi) a parameter to select one of the two ultimate convergence criteria given in Section 6 of this paper, and

(vii) an integer to limit the total number of iterations. In addition the user of the subroutine must provide a sequence of orders to calculate $f(x_1, x_2, \ldots, x_n)$ for any values of the variables.

As recommended in Section 2, the directions ξ_i are chosen initially to be unit vectors in the co-ordinate directions, so the value of q for the search in the direction ξ_i is set to $\frac{1}{10}.E.\varepsilon_i$. After each linear search the directions are scaled to have unit second derivative; therefore, after the first iteration, the value of q is chosen to be independent of i. The value of e, the accuracy to which the minimum is required along a line, is set so that each parameter is determined to at least 0·05 of the required final accuracy, except that a relative error of 3% in a linear search is tolerated.

When an iteration has been completed the ultimate convergence criterion is tried, and a check is made on the number of iterations. If these do not cause the subroutine to be left, the total change in the function due to the last iteration, Δf, is noted. If it is zero an error return results, otherwise q is set to $0{\cdot}4\sqrt{(\Delta f)}$. If it is necessary to reduce q so that the step size, which is limited by E, is not too large, this is done before a new iteration is initiated.

After the execution of the subroutine has been completed, the components of the calculated position of the minimum are set in $x_1, x_2, \ldots, x_n$.

10. More numerical examples

The procedure was developed using a bounded trigonometrical function suggested by Fletcher and Powell. It is

$$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} (A_{ij} \sin x_j + B_{ij} \cos x_j) - E_i \right\}^2.$$

The matrix elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ are generated to be random integers between -100 and $+100$, and values of $x_1, x_2, \ldots, x_n$ are chosen randomly between $-\pi$ and π. For these values the parameters E_i are calculated to be

$$E_i = \sum_{j=1}^{n} (A_{ij} \sin x_j + B_{ij} \cos x_j) \quad i = 1, 2, \ldots, n,$$

and then the starting values of the variables $x_1, x_2, \ldots, x_n$ are displaced from the values defined above by random increments of up to $\pm 0{\cdot}1\pi$. The known minima were found for values of n ranging from 3 to 20, and the resultant number of function evaluations is given in Table 2.

Table 2

Functions of many variables

n	n_f	n_c	e_c
3	61	96	3×10^{-8}
3	84	120	4×10^{-9}
5	104	166	2×10^{-8}
5	103	167	7×10^{-8}
10	329	483	4×10^{-7}
10	369	524	4×10^{-7}
20	1,519	1,954	8×10^{-7}
20	2,206*	2,823*	3×10^{-6}

n = number of variables
n_f = number of functions to achieve 10^{-4} accuracy
n_c = number of functions for convergence
e_c = final accuracy in variables
* See text (Section 10)

The more stringent convergence criterion was used, and the table shows clearly that it can require many extra iterations. The latter twenty-variable case is marked with an asterisk because it was on this case that all other convergence criteria tried were ineffective. The difficulty of this particular minimization was stressed when the procedure had appeared to converge for the first time. At this stage the parameters were changing by less than 10^{-5} per iteration, and the function had been reduced to 0·0043. Increasing the parameters by only 0·001 each caused the function to increase to 3·6. Finding the final minimum forced a variable to change by 0·014. A random displacement of this amount would probably alter the function by about 200, the actual displacement changed the function by only 0·0043.

M. J. Box has minimized the same test function using Rosenbrock's procedure. Three five-dimensional cases required 465, 466 and 388 function evaluations, and three ten-dimensional cases needed 1,210, 1,258 and 1,298 evaluations. The program was stopped immediately an accuracy of 10^{-4} had been obtained, by calculating the difference between the predicted and the known positions of the minima. Therefore these figures should be compared with those in the column of Table 2 headed n_f, which suggest that the method of this paper can be more than three times as fast as Rosenbrock's.

Another comparison was made using Rosenbrock's function

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

The progress of thirteen iterations is recorded in Table 3. As Rosenbrock points out, this function is interesting because it corresponds to a deep parabolic valley, so an efficient minimization method has to choose new directions frequently. The method of this paper has been entirely successful.

For the final example the procedure was applied to the function, used by Powell (1962),

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

This test was made as the function cannot be approximated by a quadratic in the neighbourhood of the minimum because of the two fourth-power terms; the second derivative matrix A is doubly singular, and the required scaled directions ξ tend to have infinite components. It is expected that the fast convergence of the method will break down, and Table 4 shows that the rate of approach to the minimum is approximately linear.

The procedure is started at (3, −1, 0, 1), and it soon finds a fair approximation to the minimum at which each of the four terms which are summed to form $f(x)$ are comparable. From then on it is effectively searching in the two-dimensional space defined by

$$x_1 + 10x_2 = 0 \text{ and } x_3 - x_4 = 0;$$

this is illustrated by the last four columns of the table. Consequently the new directions of search chosen have large components in the space, which explains why even the linear convergence rate is quite fast.

Table 3

A function of two variables

ITERATION	FUNCTION VALUES	x_1	x_2	$f(x_1, x_2)$
0	1	−1·2000	1·0000	24·2000
1	14	−0·9912	0·9927	3·9753
2	25	−0·7674	0·5485	3·2863
3	35	−0·5017	0·2064	2·4608
4	46	−0·2840	0·0307	1·8978
5	57	−0·0123	−0·0408	1·1927
6	71	0·2568	0·0369	0·6366
7	84	0·4379	0·1624	0·4026
8	97	0·6810	0·4478	0·1274
9	109	0·8341	0·6818	0·0469
10	122	0·8894	0·7948	0·0137
11	131	1·0014	0·9997	0·0010
12	142	0·9926	0·9850	6×10^{-5}
13	151	1·0000	1·0000	7×10^{-10}

11. Discussion

The method described in this paper finds an unconstrained minimum of a function of several variables without calculating derivatives. The examples presented and the theory behind the method suggest that it is significantly more efficient than other methods which have been referred to, but it does contain two unsatisfactory features. The first is that as the number of variables increases there is a tendency for new directions to be chosen less often. This could be overcome by always using a new direction and forcing the remaining directions to be conjugate to the new one by a projection technique, but this would require each iteration to demand approximately three times as many function values. The ultimate convergence criterion is also unsatisfactory, but it is not an essential part of the method and any improved criterion could easily be incorporated.

It is hoped that the method will prove suitable for the majority of problems which require the position of an unconstrained maximum or minimum to be found. If the derivatives of the function can be calculated as easily as function values, Davidon's method will sometimes be more effective. If the function to be minimized is a sum of squares which tends to zero at the minimum, for example, when non-linear equations are being solved, and if a good approximation to the minimum is known,

Table 4

A function of four variables

ITERATION	FUNCTION VALUES	f	x_1	$x_1 + 10x_2$	x_3	$x_3 - x_4$
0	1	215·0	3·0000	−1·0000	0·0000	1·0000
2	41	3·7759	1·5224	0·5283	0·5054	−0·5418
4	72	0·8857	0·6742	−0·6891	0·3044	−0·0972
6	102	0·0043	0·1969	−0·0282	0·0594	−0·0081
8	126	3×10^{-4}	0·0658	0·0066	−0·0051	−0·0002
10	148	8×10^{-5}	0·0115	-3×10^{-5}	−0·0365	0·0004
12	177	3×10^{-6}	0·0056	-7×10^{-4}	−0·0137	0·0002
14	208	4×10^{-8}	0·0112	1×10^{-4}	0·0037	-3×10^{-6}
16	235	5×10^{-9}	0·0055	3×10^{-5}	0·0038	2×10^{-7}
18	266	8×10^{-11}	0·0016	7×10^{-7}	0·0004	3×10^{-7}
20	295	2×10^{-12}	0·0001	-1×10^{-7}	−0·0004	-5×10^{-8}
30	433	1×10^{-21}	-4×10^{-6}	2×10^{-11}	-3×10^{-6}	4×10^{-13}

the method of least squares can converge much more quickly. This is because the function to be minimized is of form

$$F(x) = \sum_{k=1}^{m} \{f_k(x)\}^2$$

so it can be reasonable to assume that

$$\frac{\partial^2}{\partial x_i . \partial x_j} F(x) = 2 \sum_{k=1}^{m} \frac{\partial f_k}{\partial x_i} \cdot \frac{\partial f_k}{\partial x_j}.$$

Therefore an evaluation of all the derivatives at a single point is sufficient to predict the position of the minimum. It is likely that a method for minimizing a sum of squares without using derivatives will be developed, and one would expect it to be faster than the method of this paper by a factor of order n. Of course, all of these procedures may find a local minimum rather than the absolute minimum, and this is a difficult problem to overcome. If there are many variables it will certainly prove too arduous to apply a searching technique, so it is recommended that different initial approximations are tried to see if they cause more favourable extrema to be found.

12. Acknowledgements

It is a pleasure to acknowledge the interest and encouragement of Mr. A. R. Curtis. Many of the ideas presented in this paper are a direct consequence of his questions and criticisms. The author is also most grateful to Mr. M. J. Box who provided the results of applying Rosenbrock's procedure to a function of several variables.

References

DAVIDON, W. C. (1959). "Variable metric method for minimization," A.E.C. Research and Development Report, ANL-5990 (Rev.).

FLETCHER, R., and POWELL, M. J. D. (1963). "A rapidly convergent descent method for minimization," *The Computer Journal*, Vol. 6, p. 163.

HIMSWORTH, F. R., SPENDLEY, W., and HEXT, G. R. (1962). "The sequential application of simplex designs in optimisation and evolutionary operation," *Technometrics*, Vol. 4, p. 441.

POWELL, M. J. D. (1962). "An iterative method for finding stationary values of a function of several variables," *The Computer Journal*, Vol. 5, p. 147.

SMITH, C. S. (1962). "The automatic computation of maximum likelihood estimates," N.C.B. Scientific Dept., Report No. S.C. 846/MR/40.

SPANG, H. A. (1962). "A review of minimization techniques for nonlinear functions," *S.I.A.M. Review*, Vol. 4, p. 343.

ROSENBROCK, H. H. (1960). "An automatic method for finding the greatest or least value of a function," *The Computer Journal*, Vol. 3, p. 175.

Errata to Errata

"Elementary Divisors of the Liebmann Process," by G. A. Miles, K. L. Stewart and G. J. Tee. *The Computer Journal*, Vol. 6, No. 4, pp. 352–355 (January, 1964).

We regret that misprints occurred in the Errata to this article, published in *The Computer Journal*, Vol. 7, No. 1, p. 39 (April 1964). These previous Errata should be replaced by the following corrections to the original article:

(1) P. 353, line before (2.12), replace "m_i" by "λ^{m_i}".

(2) P. 354, third line after (3.13), replace "$\frac{1}{2}(n - m)$" by "$\eta^{\frac{1}{2}(n-m)}$".

(3) P. 355, in (3.18), replace "$\eta^{1/2(n-m+\nu+a)}$" by "$\eta^{\frac{1}{2}(n-m+\nu+a)}$".