A rapidly convergent descent method for minimization

By R. Fletcher and M. J. D. Powell

A powerful iterative descent method for finding a local minimum of a function of several variables is described. A number of theorems are proved to show that it always converges and that it converges rapidly. Numerical tests on a variety of functions confirm these theorems. The method has been used to solve a system of one hundred non-linear simultaneous equations.

1. Introduction

We are concerned in this paper with the general problem of finding an unrestricted local minimum of a function $f(x_1, x_2 \ldots, x_n)$ of several variables $x_1, x_2, \ldots, x_n$. We suppose that the function of interest can be calculated at all points. It is convenient to group functions into two main classes according to whether the gradient vector $g_i = \partial f/\partial x_i$ is defined analytically at each point or must be estimated from the differences of values of f. The method described in this paper is applicable to the case of a defined gradient. For the other case a useful method and general discussion are given by Rosenbrock (1960).

Methods using the gradient include the classical method of steepest descents (Courant, 1943; Curry, 1944; and Householder, 1953), Levenberg's modification of damped steepest descents (1944), a somewhat similar variation due to Booth (1957), the conjugate gradient method of Hestenes and Stiefel (1952), similar methods of Martin and Tee (1961), the "Partan" method of Shah, Buehler and Kempthorne (1961), and a method due to Powell (1962). In this paper we describe a powerful method with rapid convergence which is based upon a procedure described by Davidon (1959). Davidon's work has been little publicized, but in our opinion constitutes a considerable advance over current alternatives. We have made both a simplification by which certain orthogonality conditions which are important to the rate of attaining the solution are preserved, and also an improvement in the criterion of convergence.

Because, near the minimum, the second-order terms in the Taylor series expansion dominate, the only methods which will converge quickly for a general function are those which will guarantee to find the minimum of a general quadratic speedily. Only the latter four methods of the last paragraph do this, and the procedures of Hestenes and Stiefel and of Martin and Tee are not applicable to a general function. Of course the generalized Newton–Raphson method (Householder, 1953) has fast convergence eventually, but it requires second derivatives of the function to be evaluated, and frequently fails to converge from a poor approximation to the minimum. The method described has quadratic convergence and is superior to "Partan" and to Powell's method, both in that it makes use of information determined by previous iterations and also in that each iteration is quick and simple to carry out. Furthermore, it yields the curvature of the function at the minimum, so excellent tests for convergence and estimates of variance can be made.

The method is given an elegant theoretical basis, and proofs of stability and of the rate of convergence are included. The results of numerical tests with a variety of functions are also given. These confirm that the method is probably the most powerful general procedure for finding a local minimum which is known at the present time.

2. Notation

It is convenient to describe the method in terms of the Dirac bra-ket notation (Dirac, 1958) applied to real vectors. In this notation the column vector $(x_1, x_2 \ldots, x_n)$ is written as $|x\rangle$. The row vector with these same elements is denoted by $\langle x|$. The scalar product of $\langle x|$ and $|y\rangle$ is written $\langle x|y\rangle$ and we may note that

$$\langle x|y\rangle \equiv \sum_i x_i y_i \equiv \sum_i y_i x_i \equiv \langle y|x\rangle.$$

The construction $|x\rangle\langle y|$, however, denotes a linear operator with matrix elements $D_{ij} = x_i y_j$ so that $|x\rangle\langle y| \not\equiv |y\rangle\langle x|$. A general linear operator or matrix will be denoted by a capital letter in bold type. It then follows that say $\boldsymbol{H}|x\rangle$ is a column vector, $\langle x|\boldsymbol{H}$ is a row vector and $\langle x|\boldsymbol{H}|y\rangle$ is a scalar.

We reserve f to denote the function of interest, $|x\rangle$ to denote its arguments and $|g\rangle$ to denote its gradient.

Hence the standard quadratic form in n dimensions

$$f = f_0 + \sum_{i=1}^{n} a_i x_i + \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} G_{ij} x_i x_j$$

becomes in this notation

$$f = f_0 + \langle a|x\rangle + \tfrac{1}{2}\langle x|\boldsymbol{G}|x\rangle \qquad (1)$$

and also

$$|g\rangle = |a\rangle + \boldsymbol{G}|x\rangle. \qquad (2)$$

3. The method

If we consider the quadratic form (1) then, given the matrix $G_{ij} = \partial^2 f/\partial x_i \partial x_j$, we can calculate the displacement between the point $|x\rangle$ and the minimum $|x_0\rangle$ as

$$|x_0\rangle - |x\rangle = -\boldsymbol{G}^{-1}|g\rangle. \qquad (3)$$

In this method the matrix $\boldsymbol{G}^{-1}$ is not evaluated directly;

instead a matrix $\boldsymbol{H}$ is used which may initially be chosen to be any positive definite symmetric matrix. This matrix is modified after the i^{th} iteration using the information gained by moving down the direction

$$|s^i\rangle = -\boldsymbol{H}^i|g^i\rangle \tag{4}$$

in accordance with (3). The modification is such that $|\sigma^i\rangle$, the step to the minimum down the line

$$|x\rangle = |x^i\rangle + \lambda|s^i\rangle$$

is effectively an eigenvector of the matrix $\boldsymbol{H}^{i+1}\boldsymbol{G}$. This ensures that as the procedure converges $\boldsymbol{H}$ tends to $\boldsymbol{G}^{-1}$ evaluated at the minimum.

It is convenient to take the unit matrix initially for $\boldsymbol{H}$ so that the first direction is down the line of steepest descent.

Let the current point be $|x^i\rangle$ with gradient $|g^i\rangle$ and matrix $\boldsymbol{H}^i$. The iteration can then be stated as follows.

Set $|s^i\rangle = -\boldsymbol{H}^i|g^i\rangle$.

Obtain α^i such that $f(|x^i\rangle + \alpha^i|s^i\rangle)$ is a minimum with respect to λ along $|x^i\rangle + \lambda|s^i\rangle$ and $\alpha^i > 0$. We will prove that α^i can always be chosen to be positive.

Set $$|\sigma^i\rangle = \alpha^i|s^i\rangle. \tag{5}$$

Set $$|x^{i+1}\rangle = |x^i\rangle + |\sigma^i\rangle.$$

Evaluate $f(|x^{i+1}\rangle)$ and $|g^{i+1}\rangle$ noting that $|g^{i+1}\rangle$ is orthogonal to $|\sigma^i\rangle$, that is

$$\langle\sigma^i|g^{i+1}\rangle = 0. \tag{6}$$

Set $$|y^i\rangle = |g^{i+1}\rangle - |g^i\rangle.$$

Set $$\boldsymbol{H}^{i+1} = \boldsymbol{H}^i + \boldsymbol{A}^i + \boldsymbol{B}^i \tag{7}$$

where $$\boldsymbol{A}^i = \frac{|\sigma^i\rangle\langle\sigma^i|}{\langle\sigma^i|y^i\rangle}$$

and $$\boldsymbol{B}^i = \frac{-\boldsymbol{H}^i|y^i\rangle\langle y^i|\boldsymbol{H}^i}{\langle y^i|\boldsymbol{H}^i|y^i\rangle}.$$

Set $i = i + 1$ and repeat.

There are two obvious and very useful ways of terminating the procedure, and they arise because $|s^i\rangle$ tends to the correction to $|x^i\rangle$. One is to stop when the predicted absolute distance from the minimum $\langle s^i|s^i\rangle^{\frac{1}{2}}$ is less than a prescribed amount, and the other is to finish when every component of $|s^i\rangle$ is less than a prescribed accuracy. Two additional safeguards have been found necessary in automatic computer programs. The first is to work through at least n (the number of variables) iterations, and the second is to apply the tests to $|\sigma^i\rangle$ as well as to $|s^i\rangle$.

The method of obtaining the minimum along a line is not central to the theory. The suggested procedure given in the Appendix, which uses cubic interpolation, is based on that given by Davidon, and has been found satisfactory.

We shall now show that the process is stable, and demonstrate that if $f(|x\rangle)$ is the quadratic form (1) then the procedure terminates in n iterations. We shall also explain the theoretical justification for the manner in which the matrix $\boldsymbol{H}$ is modified.

4. Stability

It is usual for descent methods to be stable because one ensures that the function to be minimized is decreased by each step. It will be shown in this Section that the direction of search $|s^i\rangle$, defined by equation (4), is downhill, so α^i can always be chosen to be positive. Because $|g^i\rangle$ is the direction of steepest ascent, the direction $|s^i\rangle$ will be downhill if and only if

$$-\langle s^i|g^i\rangle = \langle g^i|\boldsymbol{H}|g^i\rangle$$

is positive. We wish the direction of search to be downhill for all possible $|g^i\rangle$ so we must prove that $\boldsymbol{H}^i$ is positive definite. Because $\boldsymbol{H}^0$ has been chosen to be positive definite an inductive argument will be used.

In the proof it is assumed that $\boldsymbol{H}^i$ is positive definite and consequently that α^i is positive. It is proved that, for any $|x\rangle$, $\langle x|\boldsymbol{H}^{i+1}|x\rangle \geqslant 0$. We may define $|p\rangle = (\boldsymbol{H}^i)^{\frac{1}{2}}|x\rangle$ and $|q\rangle = (\boldsymbol{H}^i)^{\frac{1}{2}}|y^i\rangle$ as the square root of a positive definite matrix exists.

From (7)

$$\begin{aligned}\langle x|\boldsymbol{H}^{i+1}|x\rangle &= \langle x|\boldsymbol{H}^i|x\rangle + \frac{\langle x|\sigma^i\rangle\langle\sigma^i|x\rangle}{\langle\sigma^i|y^i\rangle} \\ &\quad - \frac{\langle x|\boldsymbol{H}^i|y^i\rangle\langle y^i|\boldsymbol{H}^i|x\rangle}{\langle y^i|\boldsymbol{H}^i|y^i\rangle} \\ &= \frac{\langle p|p\rangle\langle q|q\rangle - \langle p|q\rangle^2}{\langle q|q\rangle} + \frac{\langle x|\sigma^i\rangle^2}{\langle\sigma^i|y^i\rangle} \\ &\geqslant \langle x|\sigma^i\rangle^2/\langle\sigma^i|y^i\rangle\end{aligned}$$

on account of Schwartz's inequality.

But $$\begin{aligned}\langle\sigma^i|y^i\rangle &= \langle\sigma^i|g^{i+1}\rangle - \langle\sigma^i|g^i\rangle \\ &= -\langle\sigma^i|g^i\rangle \text{ from (6)} \\ &= \alpha^i\langle g^i|\boldsymbol{H}^i|g^i\rangle \text{ from (4) and (5)} \\ &> 0.\end{aligned}$$

Hence $\langle x|\boldsymbol{H}^{i+1}|x\rangle > 0$ for all non-trivial $|x\rangle$. Therefore $\boldsymbol{H}^{i+1}$ is positive definite and the procedure is stable.

5. Quadratic convergence

In this Section it is assumed that f is the quadratic form (1) and that f has a well defined minimum. It is proved that in this case the method finds the minimum in n iterations. The method of proof is to show that $|\sigma^0\rangle$, $|\sigma^1\rangle, \ldots, |\sigma^k\rangle$ are linearly independent eigenvectors of $\boldsymbol{H}^{k+1}\boldsymbol{G}$ with eigenvalue unity. Therefore it will follow that $\boldsymbol{H}^n\boldsymbol{G}$ is the unit matrix.

By definition

$$\begin{aligned}|y^i\rangle &= |g^{i+1}\rangle - |g^i\rangle \\ &= \boldsymbol{G}|x^{i+1}\rangle - \boldsymbol{G}|x^i\rangle \quad \text{from (2)} \\ &= \boldsymbol{G}|\sigma^i\rangle. \end{aligned} \tag{8}$$

Also from (8)

$$\begin{aligned} \boldsymbol{H}^{i+1}\boldsymbol{G}|\sigma^i\rangle &= \boldsymbol{H}^{i+1}|y^i\rangle \\ &= \boldsymbol{H}^i|y^i\rangle + |\sigma^i\rangle - \boldsymbol{H}^i|y^i\rangle \text{ by using (7)} \\ &= |\sigma^i\rangle. \end{aligned} \tag{9}$$

The equations

$$\langle\sigma^i|\boldsymbol{G}|\sigma^j\rangle = 0 \qquad 0 \leqslant i < j < k \tag{10}$$

and

$$\boldsymbol{H}^k\boldsymbol{G}|\sigma^i\rangle = |\sigma^i\rangle \quad 0 \leqslant i < k \tag{11}$$

will now be considered. It is clear from (9) that they are true if $k = 1$. It will be proved that if they are true for k they are true for $k + 1$.

From (2)

$$\begin{aligned} |g^k\rangle &= |a\rangle + \boldsymbol{G}|x^k\rangle \\ &= |a\rangle + \boldsymbol{G}(|x^{i+1}\rangle + |\sigma^{i+1}\rangle + |\sigma^{i+2}\rangle + \ldots \\ &\qquad + |\sigma^{k-1}\rangle) \\ &= |g^{i+1}\rangle + \boldsymbol{G}(|\sigma^{i+1}\rangle + \ldots + |\sigma^{k-1}\rangle). \end{aligned}$$

Therefore from (10) and (6)

$$\begin{aligned} \langle\sigma^i|g^k\rangle &= \langle\sigma^i|g^{i+1}\rangle \\ &= 0 \qquad 0 \leqslant i < k. \end{aligned} \tag{12}$$

Hence from (11)

$$\langle\sigma^i|\boldsymbol{G}\boldsymbol{H}^k|g^k\rangle = 0$$

so from (4) and (5)

$$-\alpha^i\langle\sigma^i|\boldsymbol{G}|\sigma^k\rangle = 0.$$

Therefore

$$\langle\sigma^i|\boldsymbol{G}|\sigma^k\rangle = 0 \qquad 0 \leqslant i < k. \tag{13}$$

Also from (8), (11) and (13)

$$\begin{aligned} \langle y^k|\boldsymbol{H}^k\boldsymbol{G}|\sigma^i\rangle &= \langle y^k|\sigma^i\rangle \\ &= \langle\sigma^k|\boldsymbol{G}|\sigma^i\rangle \\ &= 0 \qquad 0 \leqslant i < k. \end{aligned}$$

Therefore using the above result and equations (7), (11) and (13)

$$\begin{aligned} \boldsymbol{H}^{k+1}\boldsymbol{G}|\sigma^i\rangle &= \boldsymbol{H}^k\boldsymbol{G}|\sigma^i\rangle \\ &= |\sigma^i\rangle \quad 0 \leqslant i < k. \end{aligned} \tag{14}$$

Equations (9), (13) and (14) prove the induction. Equation (10) proves that the vectors $|\sigma^0\rangle, |\sigma^1\rangle, \ldots, |\sigma^{n-1}\rangle$ are linearly independent and therefore $\boldsymbol{H}^n = \boldsymbol{G}^{-1}$.

That the minimum is found by n iterations is proved by equation (12). $|g^n\rangle$ must be orthogonal to $|\sigma^0\rangle, |\sigma^1\rangle, \ldots, |\sigma^{n-1}\rangle$ which is only possible if $|g^n\rangle$ is identically zero.

6. Improving the matrix H

The matrix $\boldsymbol{H}^i$ is modified by adding to it two terms $\boldsymbol{A}^i$ and $\boldsymbol{B}^i$. $\boldsymbol{A}^i$ is the factor which makes $\boldsymbol{H}$ tend to $\boldsymbol{G}^{-1}$ in the sense that for a quadratic

$$\boldsymbol{G}^{-1} = \sum_{i=0}^{n-1} \boldsymbol{A}^i. \tag{15}$$

This result can be proved from the orthogonality conditions (10) because these imply that $\boldsymbol{S}'\boldsymbol{G}\boldsymbol{S} = \boldsymbol{\Lambda}$, where $\boldsymbol{S}$ is the matrix of vectors $|\sigma^i\rangle$, and $\boldsymbol{\Lambda}$ is a diagonal matrix with elements $\langle\sigma^i|\boldsymbol{G}|\sigma^i\rangle$.

Hence by definition $\boldsymbol{G} = (\boldsymbol{S}')^{-1}\boldsymbol{\Lambda}(\boldsymbol{S}^{-1})$

$$= (\boldsymbol{S}\boldsymbol{\Lambda}^{-1}\boldsymbol{S}')^{-1}.$$

Therefore $\boldsymbol{G}^{-1} = \boldsymbol{S}\boldsymbol{\Lambda}^{-1}\boldsymbol{S}'$

and as $\boldsymbol{\Lambda}$ is a diagonal matrix this reduces to

$$\begin{aligned} \boldsymbol{G}^{-1} &= \sum_i (\boldsymbol{\Lambda}^{-1})_{ii}|\sigma^i\rangle\langle\sigma^i| \\ &= \sum_i \langle\sigma^i|\boldsymbol{G}|\sigma^i\rangle^{-1}|\sigma^i\rangle\langle\sigma^i|. \end{aligned}$$

Therefore from the definition of $\boldsymbol{A}^i$ and equation (8), equation (15) is proved.

The form of the term $\boldsymbol{B}^i$ can be deduced because equation (9) must be valid. For a quadratic we must have

$$\begin{aligned} \boldsymbol{H}^{i+1}\boldsymbol{G}|\sigma^i\rangle &= |\sigma^i\rangle \\ &= \boldsymbol{H}^i\boldsymbol{G}|\sigma^i\rangle + \boldsymbol{A}^i\boldsymbol{G}|\sigma^i\rangle + \boldsymbol{B}^i\boldsymbol{G}|\sigma^i\rangle. \end{aligned}$$

Therefore as $\boldsymbol{A}^i\boldsymbol{G}|\sigma^i\rangle = |\sigma^i\rangle$ the equation

$$\boldsymbol{B}^i\boldsymbol{G}|\sigma^i\rangle = -\boldsymbol{H}^i\boldsymbol{G}|\sigma^i\rangle = -\boldsymbol{H}^i|y^i\rangle$$

must be satisfied.

This implies that the simplest form for $\boldsymbol{B}^i$ is

$$\boldsymbol{B}^i = -\frac{\boldsymbol{H}^i|y^i\rangle\langle z|}{\langle z|y^i\rangle}$$

and as $\boldsymbol{B}^i$ is to be symmetric this gives

$$\boldsymbol{B}^i = -\frac{\boldsymbol{H}^i|y^i\rangle\langle y^i|\boldsymbol{H}^i}{\langle y^i|\boldsymbol{H}^i|y^i\rangle}.$$

Although Davidon's method involves these relations, some of the other ideas used by him can cause $\boldsymbol{H}$ not to tend to $\boldsymbol{G}^{-1}$ even in the quadratic case. The effect in the non-quadratic case would depend upon the function in question but might well lead to slower convergence.

7. Numerical results—comparison with other procedures

As a comparison with other methods we use the function given by Rosenbrock

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

starting at $(-1{\cdot}2,\ 1{\cdot}0)$. This function is difficult to minimize on account of its having a steep sided valley following the curve $x_1^2 = x_2$. Eighteen iterations were required to reach the minimum, each one requiring the minimum to be calculated in only one direction. Table 1 shows how this procedure compares with the classical steepest descent method and Powell's method, one of the procedures with quadratic convergence. The table takes into account that the latter method requires minima to be found in three directions for each iteration. It will be seen that this method is considerably more efficient than that of Powell, both of these being far more efficient than steepest descents.

Table 1

A comparison in two dimensions

EQUIVALENT n	STEEPEST DESCENTS $f(x_1, x_2)$	POWELL'S METHOD $f(x_1, x_2)$	OUR METHOD $f(x_1, x_2)$
0	24·200	24·200	24·200
3	3·704	3·643	3·687
6	3·339	2·898	1·605
9	3·077	2·195	0·745
12	2·869	1·412	0·196
15	2·689	0·831	0·012
18	2·529	0·432	1×10^{-8}
21	2·383	0·182	—
24	2·247	0·052	—
27	2·118	0·004	—
30	1·994	5×10^{-5}	—
33	1·873	8×10^{-9}	—

Table 2

A quadratic function

ITERATION	0	1	2
x	−4 +2	−1·69 −1·08	0 0
f	+40	+1·54	10^{-15}
$\boldsymbol{H}$	+1 0 0 +1	+0·781 +0·361 +0·361 +0·411	1 ½ ½ ½
$\lvert\sigma\rangle$	+2·31 −3·08	+1·69 +1·08	—
$\boldsymbol{A}$	+0·069 −0·092 −0·092 +0·123	+0·931 +0·592 +0·592 +0·377	— —

A similar comparison was made with the function given by Powell:

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

starting at (3, −1, 0, 1). In six iterations the method reduced f from 215 to $2{\cdot}5 \times 10^{-8}$. Powell's method took the equivalent of seven iterations to reach 9×10^{-3}, whereas steepest descents only reached 6·36 in seven iterations. The method also brought out the singularity of $\boldsymbol{G}$ at the minimum of f, the elements of $\boldsymbol{H}$ becoming increasingly large.

To compare this variation of Davidon's method with his original method the simple quadratic

$$f(x_1, x_2) = x_1^2 - 2x_1x_2 + 2x_2^2$$

was used. The complete progress of the method described is given in Table 2, showing that it does terminate in two iterations and that $\boldsymbol{H}$ does converge to $\boldsymbol{G}^{-1}$ which for this function is

$$\begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

It will be noticed also, as proved, that $\boldsymbol{G}^{-1} = \sum_i \boldsymbol{A}^i$.

In Davidon's method, although a value of f of similar order of magnitude had been reached in two iterations, $\boldsymbol{H}$ had only reached

$$\begin{pmatrix} 0{\cdot}95 & 0{\cdot}47 \\ 0{\cdot}47 & 0{\cdot}48 \end{pmatrix}.$$

This was due to one of the alternatives allowed by Davidon. Also his procedure for terminating the process was unsatisfactory, and the computation had to be stopped manually.

A non-quadratic test in three dimensions was also made, by using a function with a steep sided helical valley. This function

$$f(x_1, x_2, x_3) = 100\{[x_3 - 10\theta(x_1, x_2)]^2 + [r(x_1, x_2) - 1]^2\} + x_3^2$$

where
$$2\pi\theta(x_1, x_2) = \arctan(x_2/x_1), \quad x_1 > 0$$
$$= \pi + \arctan(x_2/x_1), \quad x_1 < 0$$

and
$$r(x_1, x_2) = (x_1^2 + x_2^2)^{\frac{1}{2}}$$

has a helical valley in the x_3 direction with pitch 10 and radius 1. It is only considered for

$$-\pi/2 < 2\pi\theta < 3\pi/2$$

that is
$$-2{\cdot}5 < x_3 < 7{\cdot}5.$$

It has a minimum at the point (1, 0, 0).

Both methods were started from (−1, 0, 0) and $\boldsymbol{H}^0$ set to the unit matrix. The method given in this paper converged in eighteen iterations, whereas Davidon's method required only ten. However, on account of the more complicated nature of Davidon's iterations, the minimum often being sought along more than one direction in a single iteration, the time taken by the two procedures was almost identical. The progress of this method on the function is given in Table 3.

8. Numerical results—functions of a large number of variables

Tests were also made to find out whether the method is suitable for finding the minimum of a function of a large number of variables. In these tests the Stretch computer was used to solve non-linear simultaneous equations in up to a hundred variables.

The equations were

$$\sum_{j=1}^{n} A_{ij} \sin \alpha_j + B_{ij} \cos \alpha_j = E_i \qquad i = 1, 2, \ldots, n$$

Table 3

A function with a steep-sided helical valley

n	x_1	x_2	x_3	f
0	−1·000	0·000	0·000	$2{\cdot}5 \times 10^4$
1	−1·000	2·278	1·431	$5{\cdot}2 \times 10^3$
2	−0·023	2·004	2·649	$1{\cdot}1 \times 10^3$
3	−0·856	1·559	3·429	74·080
4	−0·372	1·127	3·319	24·190
5	−0·499	0·908	3·285	10·942
6	−0·314	0·900	3·075	9·841
7	0·059	1·069	2·408	6·304
8	0·146	1·086	2·261	6·093
9	0·774	0·725	1·218	1·889
10	0·746	0·706	1·242	1·752
11	0·894	0·496	0·772	0·762
12	0·994	0·298	0·441	0·382
13	0·994	0·191	0·317	0·141
14	1·017	0·085	0·133	0·058
15	0·997	0·070	0·110	0·013
16	1·002	0·009	0·014	8×10^{-4}
17	1·000	0·002	0·040	3×10^{-6}
18	1·000	10^{-5}	10^{-5}	7×10^{-8}

so that the function to be minimized was

$$f = \sum_{i=1}^{n} \{E_i - \sum_{j=1}^{n} (A_{ij} \sin \alpha_j + B_{ij} \cos \alpha_j)\}^2.$$

The matrix elements of A and B were generated as random integers between −100 and +100, and the values of the variables α_i, $i = 1, 2, \ldots, n$ were generated randomly between $-\pi$ and π. For these values the right-hand sides of the equations, E_i, were worked out. The method of this paper was applied to find optimum values of α_i starting from $(\alpha_i + 0{\cdot}1\delta_i)$ where the δ_i's were also generated as random numbers between $-\pi$ and π. In each run the criterion for convergence was that every α should be found to accuracy 0·0001.

The method was entirely successful. Table 4 shows that the number of times f and its derivatives had to be calculated was approximately linear in the number of variables. The total time taken for *all* the runs was fifteen minutes, ten minutes of which was spent on the final case. That a different minimum was found on five occasions was not surprising because if $A \equiv B$ it may be shown that there are up to 2^n real solutions to the equations such that $|\alpha_i| \leqslant \pi$. This abundance of minima emphasizes the power of the method because in every case it converged to a reasonable solution.

The progress of these tests is interesting. For the first n iterations the changes in the function were similar to those experienced with the method of steepest descents, that is a substantial change occurred initially due to descending into a nearby valley, after which convergence was slow. However, after n iterations had been completed a good approximation to the final matrix H had been accumulated, after which the function was decreased substantially at each iteration. For example in the hundred-variable trial the function to be minimized was decreased from 293407 to 54165 in the first ten iterations, and to 14686 after one hundred iterations. After 120 iterations it was down to 1342, and after 140 to 147. The function was reduced to 0·44 by 160 iterations, and the minimum was found on the 162nd. The second fifty-variable trial was even more striking. Ten iterations reduced the function from 205380 to 4264, fifty iterations reduced it to 3526, and a further ten iterations reduced it to 0·027.

Table 4

Application to a function of many variables

n	NO. OF TIMES f EVALUATED	WHETHER EXPECTED MINIMUM FOUND
5	19	Yes
5	23	Yes
10	36	Yes
10	29	Yes
20	89	No
20	84	Yes
20	68	Yes
20	121	Yes
30	86	No
30	92	Yes
30	118	No
30	113	Yes
50	169	No
50	119	Yes
100	318	No

The conclusion to be drawn from this behaviour is that for many applications of the method a substantial number of the iterations required will be spent on setting up the inverse of the matrix of second derivatives. Therefore, if a good positive definite approximation to H can be calculated initially, as is the case when the method is being applied to solving simultaneous equations, then this approximation should be chosen for H^0.

9. Conclusion

The numerical examples show clearly that the type of method given by Davidon is considerably superior to other methods previously available. The simplifications we have made enable programs to be written more easily, and they do not seem to impair the speed of convergence. It is obviously practicable to apply this method to find a local minimum of a general function of a large number of variables whose first derivatives can be evaluated quickly, even if only poor initial approximations to a solution are known.

10. Acknowledgement

One of us, R.F., is indebted to Dr. C. M. Reeves for his constant help and encouragement, and also to the D.S.I.R. for the provision of a research studentship.

Appendix

The minimum on a line

A simple algorithm is given for estimating the parameter α^i. A point $|y^i\rangle$ is chosen on $|x^i\rangle + \lambda|s^i\rangle$ with $\lambda > 0$. Let f_x, $|g_x\rangle$, f_y and $|g_y\rangle$ denote the values of the function and gradient at the points $|x^i\rangle$ and $|y^i\rangle$. Then an estimate of α^i can be formed by interpolating cubically, using the function values f_x and f_y and the components of the gradients along $|s^i\rangle$.

This is given by

$$\frac{\alpha^i}{\lambda} = 1 - \frac{\langle g_y|s^i\rangle + w - z}{\langle g_y|s^i\rangle - \langle g_x|s^i\rangle + 2w}$$

where $w = (z^2 - \langle g_x|s^i\rangle \langle g_y|s^i\rangle)^{\frac{1}{2}}$

and $z = \frac{3}{\lambda}(f_x - f_y) + \langle g_x|s^i\rangle + \langle g_y|s^i\rangle.$

A suitable choice of the point $|y^i\rangle$ is given by $|y^i\rangle = |x^i\rangle + \eta|s^i\rangle$ where

$$\eta = \text{MINIMUM OF} \left\{1, \frac{-2(f_x - f_0)}{\langle g_x|s^i\rangle}\right\}.$$

f_0 is the predicted lower bound of $f(|x\rangle)$, for example zero in least-squares calculations. This value of η ensures that the choice of $|y^i\rangle$ is reasonable.

It is necessary to check that $f(|x^i\rangle + \alpha^i|s^i\rangle)$ is less than both f_x and f_y. If it is not, the interpolation must be repeated over a smaller range. Davidon suggests one should ensure that the minimum is located between $|x^i\rangle$ and $|y^i\rangle$ by testing the sign of $\langle g_y|s^i\rangle$ and comparing f_x and f_y before interpolating. The reader is referred to Davidon's report for more extensive details of this stage.

References

BOOTH, A. D. (1957). *Numerical Methods*, London: Butterworths.

COURANT, R. (1943). "Variational methods for the solution of problems of equilibrium and vibrations," *Bull. Amer. Math. Soc.*, Vol. 49, p. 1.

CURRY, H. D. (1944). "The method of steepest descent for non-linear minimization problems;" *Qu. App. Maths.*, Vol. 2, p. 258.

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

DIRAC, P. A. M. (1958). *The Principles of Quantum Mechanics*, Oxford: O.U.P.

HESTENES, M. R., and STIEFEL, E. (1952). "Methods of conjugate gradients for solving linear systems," *J. Res. N.B.S.*, Vol. 49, p. 409.

HOUSEHOLDER, A. S. (1953). *Principles of Numerical Analysis*, New York: McGraw-Hill.

LEVENBERG, K. (1944). "A method for the solution of certain non-linear problems in least squares," *Qu. App. Maths.*, Vol. 2, p. 164.

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.

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.

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

SHAH, B. V., BUEHLER, R. J., and KEMPTHORNE, O. (1961). "The method of parallel tangents (Partan) for finding an optimum," Office of Naval Research Report, NR-042-207 (No. 2).

Book Review (*continued from p. 143*)

here. It seems entirely wrong that modern source languages should be dominated by the sequential regimes of early machine codes, and any move towards conventional mathematical forms is to be welcomed. Again, the approach here seems rather tentative and some major benefits are lost. I think it is preferable to make sequential coding subordinate to definitions rather than the other way round: here lies the key to the very important problem of integrating the translator with a realistic operating system.

Amongst the other problems tackled are the handling of complex variables, recurrence relations, and direct transfer of control to parts of the program not similarly accessible in ALGOL. It may be regarded as a tribute to ALGOL that an attempt has been made to graft such a system onto the same tree. At the same time it must accept a measure of responsibility for the fact that the above ideas were not more fully developed and in use three years ago.

J. K. ILIFFE.