A comparison of several current optimization methods, and the use of transformations in constrained problems

By M. J. Box*

The performances of eight current methods for unconstrained optimization are evaluated using a set of test problems with up to twenty variables. The use of optimization techniques in the solution of simultaneous non-linear equations is also discussed. Finally transformations whereby inequality constraints of certain forms can be eliminated from the formulation of an optimization problem are described, and examples of their use compared with other methods for handling such constraints.

1. Introduction

The type of optimization problem with which this paper is concerned is that in which the objective function is highly non-linear and the number of independent variables is small, with a maximum of about twenty. The conclusions therefore do not apply to either linear or mildly non-linear problems. For linear problems, several thousand independent variables can be handled, whilst for mildly non-linear problems the possible number of independent variables is several hundred.

The first part of this paper describes experimental comparisons of the performances of eight current methods for unconstrained optimization based on a set of test functions with up to twenty independent variables. Little comparative information on these methods has been published, understandably since several of the methods are very recent; in particular, little has been reported on the use of these methods with as many as twenty variables.

There follows a brief demonstration that the problem of solving sets of simultaneous non-linear equations should be attempted by using methods based on the Jacobian, rather than by merely minimizing the sum of squared residuals.

The paper then goes on to consider transformations by which problems of constrained optimization (only inequality constraints of certain forms will be considered) can be reduced to a form in which no constraints explicitly appear, so that they are then suitable for solution by methods incapable of handling constraints. These latter methods include several which are more powerful than the limited number available for constrained optimization of a general non-linear function. Some numerical experiments in the use of this approach in conjunction with one of the more efficient algorithms for unconstrained optimization are described, and the results compared with certain other techniques for constrained optimization.

2. The optimization methods studied

The optimization methods which have been applied to test problems without constraints are listed below. For the sake of compactness, the methods will be identified in the tables of results by the letters indicated.

(i) DSC, a method developed from that of Rosenbrock (1960) by Davies, Swann and Campey (Swann, 1964), and which has been discussed by Fletcher (1965).
(ii) R, Rosenbrock's (1960) method.
(iii) N, the simplex method developed by Nelder and Mead (1965) from that of Spendley, Hext and Himsworth (1962).
(iv) P, Powell's (1964) method for minimizing a general function without calculating derivatives.
(v) F, a method, described by Fletcher and Reeves (1964), which uses the properties of conjugate directions.
(vi) D, the method described originally by Davidon (1959), and in a refined form by Fletcher and Powell (1963).
(vii) PSS, Powell's (1965) method for minimizing a sum of squares.
(viii) B, Barnes' (1965) method for solving sets of simultaneous non-linear equations.

Of these eight methods, DSC, R, N and P are "direct search" methods which require a subroutine to compute function values only, i.e. the first derivatives of the function are *not* required, whereas methods F and D are "gradient methods" which *do* require the first derivatives of the function to be computed as well as the function itself.

In a number of papers it has been stated that a typical use of optimization techniques is to solve sets of simultaneous non-linear equations by minimizing the sum of the squared residuals. The PSS and B methods for

* *Imperial Chemical Industries Limited, Central Instrument Research Laboratory, Bozedown House, Whitchurch Hill, Reading, Berkshire.*

solving such sets of equations work with the residuals for each equation instead of merely the sum of their squares. The test problems which are considered in this paper all arise from "least squares" curve fitting, and each can be reformulated as the solution of a set of non-linear equations, as an alternative to a straight minimization problem. Thus the efficacy of using methods such as PSS and B can be studied. It may not be possible to reformulate every optimization problem as the solution of a set of simultaneous equations, however. Certainly it is often not feasible to attempt this.

When a "gradient method" is being used to optimize a function with n independent variables, each entry to the subroutine to compute the function and its n first derivatives requires $(n + 1)$ items of information to be provided. Thus in this comparison, each entry by a "gradient method" to its subroutine will be regarded as equivalent to $(n + 1)$ function evaluations in a "direct search" method, i.e. one that uses function values only, and the term "equivalent function evaluations" will be introduced. This "$(n + 1)$ factor" is also equivalent to estimating the first derivatives by forward (or backward) differences, and this is considered to be a further justification for its use. (The use of central differences would be equivalent to the use of a factor $(2n + 1)$.)

In the case of methods PSS and B no multiplying factor will be used, however, as the amount of computation in the subroutine is no more than with "direct search" methods. In all cases, all the residuals are computed, the only difference being that the PSS and B methods have access to these residuals individually.

All comparisons of methods will be carried out on the basis of the number of "equivalent function evaluations," as for many real problems the time involved in the computation of the function is vastly in excess of that required to organize the search. An example is the problem of determining "least squares" estimates of parameters in non-linear differential equations from experimental data. In this problem, the residual differences, R_i, between experimental data and numerically integrated solutions of the differential equations are computed for imputed values of the parameters. The parameters are then varied either to minimize the sum of squared residuals, or to find the "least squares" solution of the simultaneous non-linear equations, $R_i = 0$, according to which approach is being used. It has been known for the computation of all the residuals for a single set of parameter values to take up to 1,000 times as long as the organization of the search.

3. The comparison of the methods

The performances of the eight optimization methods have been studied with problems with 2, 3, 5, 10 and 20 independent variables.

3.1 *Two dimensions*

The two-dimensional test problem used was as follows:

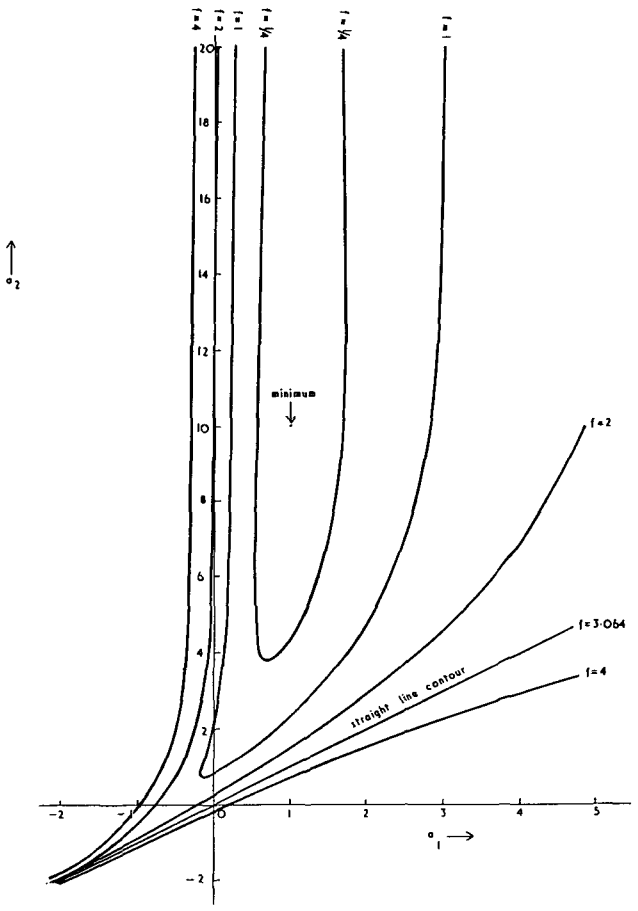

Fig. 1.—Contours of 2-dimensional exponential function

to minimize

$$f(a_1, a_2) = \sum_x [(e^{-a_1 x} - e^{-a_2 x}) - (e^{-x} - e^{-10x})]^2,$$

where the summation is over the values $x = 0{\cdot}1(0{\cdot}1)1$. Thus the minimum is $f = 0$, corresponding to $a_1 = 1$, $a_2 = 10$.

The selection of this test function is motivated by its equivalence to the problem of estimating two parameters in a pair of simultaneous linear differential equations, which is a simple example of the problem of estimating chemical reaction rate constants, a problem on which the author has worked. A plot of the contours of this function, given in **Fig. 1,** indicates the existence of a highly asymmetric curved valley, a feature known to occur in many minimization problems, and noted for the difficulty which it presents.

Five starting points were used:

I.	$a_1 = 0$,	$a_2 = 0$,	$f = 3{\cdot}064$
II.	$a_1 = 0$,	$a_2 = 20$,	$f = 2{\cdot}087$
III.	$a_1 = 5$,	$a_2 = 0$,	$f = 19{\cdot}588$
IV.	$a_1 = 5$,	$a_2 = 20$,	$f = 1{\cdot}808$
V.	$a_1 = 2{\cdot}5$,	$a_2 = 10$,	$f = 0{\cdot}808$

For each starting point the number of equivalent function evaluations necessary to reduce the function to

the values 1, 0·1, 0·01 and 0·00001 are recorded in **Table 1**.

Table 1

Comparison in 2 dimensions

METHOD	CONTOUR	STARTING POINT				
		I	II	III	IV	V
DSC	1·0	23	3	15	7	0
	0·1	45	8	119	23	6
	0·01	55	22	203	28	6
	0·00001	78	57	231	53	6
R	1·0	15	4	13	12	0
	0·1	39	7	26	25	20
	0·01	52	45	73	77	89
	0·00001	96	68	103	103	109
N	1·0	6	2	5	5	0
	0·1	6	8	5	17	12
	0·01	19	20	19	26	13
	0·00001	41	43	39	49	40
P	1·0	33	2	26	16	0
	0·1	52	8	57	16	6
	0·01	56	30	69	37	7
	0·00001	64	51	84	77	23
F	1·0	12	6	21	12	0
	0·1	45	12	75	12	12
	0·01	69	126	87	60	12
	0·00001	93	144	102	84	21
D	1·0	12	6	12	15	0
	0·1	33	9	18	18	6
	0·01	45	42	30	54	9
	0·00001	51	48	45	63	27
PSS	1·0	8	4	28	4	0
	0·1	29	16	36	16	4
	0·01	33	17	40	22	9
	0·00001	38	22	46	29	12
B	1·0	19	4		25	0
	0·1	24	6	N.O	33	6
	0·01	24	12		34	20
	0·00001	27	15		36	24

For method B, the notation "N.O" indicates that results corresponding to starting point III were not obtainable. The program found a direction such that $a_1 \to \infty$, $a_2 \to \infty$ with $a_1 > a_2$, and the minimum along such a direction lies at infinity.

3.2 *Three dimensions*

A three-dimensional test problem was derived from the two-dimensional problem as follows:

to minimize

$$f(a_1, a_2, a_3) = \sum_x [(e^{-a_1 x} - e^{-a_2 x}) - a_3(e^{-x} - e^{-10x})]^2$$

where the summation is over the values $x = 0 \cdot 1(0 \cdot 1)1$.

The desired optimum is $f = 0, a_1 = 1, a_2 = 10, a_3 = 1$. There is, however, the continuum of optima $f = 0$ corresponding to $a_3 = 0$, $a_1 = a_2$, on which various of the methods found solutions with some starting points. This was rarely the case with the nine starting points quoted below, and in the majority of these cases the desired optimum was found by making an alternative selection of initial step-lengths.

In point of fact, the series of computational results presented here had almost been completed before any difficulty in obtaining the desired solution $a_1 = 1$, $a_2 = 10$, $a_3 = 1$ was encountered. It certainly appears now that a superior definition of the three-dimensional test function would have been

$$f(a_1, a_2, a_3) = \sum_x [a_3(e^{-a_1 x} - e^{-a_2 x}) - (e^{-x} - e^{-10x})]^2$$

with the same summation as before.

The optimum $a_1 = 10$, $a_2 = 1$, $a_3 = -1$, has never been obtained with any combination of method and starting point tried to date.

Nine starting points were used:

I.	$a_1 = 0$,	$a_2 = 20$,	$a_3 = 1$,	$f = 2 \cdot 087$
II.	$a_1 = 2 \cdot 5$,	$a_2 = 10$,	$a_3 = 10$,	$f = 275 \cdot 881$
III.	$a_1 = 0$,	$a_2 = 0$,	$a_3 = 10$,	$f = 306 \cdot 401$
IV.	$a_1 = 0$,	$a_2 = 10$,	$a_3 = 1$,	$f = 1 \cdot 885$
V.	$a_1 = 0$,	$a_2 = 10$,	$a_3 = 10$,	$f = 213 \cdot 673$
VI.	$a_1 = 0$,	$a_2 = 10$,	$a_3 = 20$,	$f = 1{,}031 \cdot 154$
VII.	$a_1 = 0$,	$a_2 = 20$,	$a_3 = 0$,	$f = 9 \cdot 706$
VIII.	$a_1 = 0$,	$a_2 = 20$,	$a_3 = 10$,	$f = 209 \cdot 280$
IX.	$a_1 = 0$,	$a_2 = 20$,	$a_3 = 20$,	$f = 1{,}021 \cdot 655$

For each starting point the number of equivalent function evaluations necessary to reduce the function to the values 1, 0·1, 0·01 and 0·00001 are recorded in **Table 2**.

The notation "F" is used to indicate that the DSC and P methods failed on this problem for the six starting points with the highest f-values, in that they produced the following solution:

$$a_1 \simeq 0 \cdot 61, \quad a_2 \to \infty, \quad a_3 \simeq 1 \cdot 32, \quad f \simeq 0 \cdot 076,$$

i.e. regarding the problem as one of curve-fitting, these methods have effectively eliminated a_2 from the problem and then endeavoured to fit one exponential and a multiplicative factor to the data. This failure stems from the fact that both these methods (as implemented in the programs released to the author) set out to locate the minimum along a line too precisely. (The author has made no attempt to restrict the number of steps that are taken along a line.) Any method which does not find the minimum along a line, for instance the methods of Rosenbrock (1960), Nelder and Mead (1965), Spendley, Hext and Himsworth (1962) or some variants of steepest descent, could not fail in this way.

Table 2

Comparison in 3 dimensions

METHOD	CONTOUR	STARTING POINT								
		I	II	III	IV	V	VI	VII	VIII	IX
DSC	1·0	3	F	F	3	F	F	1	F	F
	0·1	8			4			6		
	0·01	20			6			8		
	0·00001	448			313			25		
R	1·0	5	30	31	5	30	43	10	24	41
	0·1	7	57	31	18	30	50	25	55	61
	0·01	150	197	139	38	174	165	224	190	179
	0·00001	347	281	200	198	292	350	273	460	246
N	1·0	2	11	6	2	18	27	11	18	27
	0·1	13	11	29	20	26	92	17	28	109
	0·01	48	42	46	28	70	198	47	127	210
	0·00001	119	128	112	73	110	307	79	164	315
P	1·0	2	F	F	2	F	F	3	F	F
	0·1	8			7			9		
	0·01	33			10			13		
	0·00001	78			56			19		
F	1·0	8	28	44	8	36	40	8	36	40
	0·1	8	28	44	16	44	48	16	56	48
	0·01	116	40	48	28	68	48	116	208	124
	0·00001	564	112	104	56	92	92	344	608	188
D	1·0	8	8	44	8	52	52	8	52	52
	0·1	16	12	52	12	60	60	12	60	60
	0·01	72	16	128	20	128	112	76	120	116
	0·00001	92	68	144	24	148	140	96	148	140
PSS	1·0	5	N.O	N.O	5	6	6	6	6	6
	0·1	7			6	6	6	7	7	7
	0·01	15			15	9	9	22	30	21
	0·00001	34			29	28	28	43	46	33
B	1·0	4	N.O	N.O	5	11	N.O	5	11	21
	0·1	5			6	12		5	11	21
	0·01	21			6	17		21	42	22
	0·00001	42			15	48		37	51	59

The notation "N.O" indicates that with method PSS two, and with method B three, starting points gave solutions on the line $a_1 = a_2$. Although several different combinations of initial step-lengths were tried with the PSS method, the desired solution was not obtainable.

3.3 *Five, ten and twenty dimensions*

The test function used in 5, 10 and 20 dimensions was that originally introduced by Fletcher and Powell (1963): to minimize

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

with respect to the α_j,
i.e. to solve the set of simultaneous non-linear equations

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

by minimizing the sum of squared residuals.

The A_{ij} and B_{ij} were generated as pseudo-random integers rectangularly distributed over the interval $[-100, 100]$, and target values of the variables α_i^*, $i = 1, 2, \ldots n$, were generated as pseudo-random real numbers over the interval $[-\pi, \pi]$. For these values, the right-hand sides of the equations, E_i, were calculated. The initial point used was $\alpha_i^* + \delta_i$, where the δ_i were generated as pseudo-random real numbers over the interval $[-\pi/10, \pi/10]$. In each run the criterion for convergence was that every α_i should have been found to accuracy 0·0001, a comparison of each variable α_i with its correct value α_i^* being carried out every time f was computed, with the number of function evaluations needed to achieve such accuracy being printed out the first time it was obtained. Advantages of this test function are that the number of variables, n, can be readily varied, and, moreover, the results can be replicated by the use of different pseudo-random number sequence initiators.

Replications were performed, and the number of equivalent function evaluations necessary for convergence (using the criterion described above) are recorded in **Table 3** for each combination of method and number of dimensions. The values given for method PSS in 20 dimensions (denoted with an asterisk) may be a little too large as these results have been provided by M. J. D. Powell from printouts not given immediately 10^{-4} accuracy was obtained. "N.A" indicates that the solution of this problem in 10 and 20 dimensions was not attempted with method B. The results in 5, 10 and 20 dimensions for methods D, P and PSS are reproduced from the papers of Fletcher and Powell (1963), Powell (1964) and Powell (1965) respectively.

The results given in Table 3 have been obtained using three different computers, all with different word-lengths, and four different compilers and pseudo-random number sequence generators. Thus detailed comparison of particular runs for different optimization methods is not meaningful as, for the most part, these runs do not correspond to the same member of the family of test functions. As with the two- and three-dimensional results it is felt that to present all the results is preferable to giving average values, particularly so as there is a considerable spread in the results for any combination of method and problem. The variation in the results between different members of this family of test functions is in fact less drastic than the variation obtained with the two- and three-dimensional test functions arising from the use of different starting points.

The qualitative evidence which the results of Table 3 provide will be given considerable weight when conclusions are drawn, as the larger the number of variables, the more searching the test of the optimization method turns out to be. (Many methods which successfully optimize functions of two or three variables are found to be highly inefficient when there are more independent variables.) Further evidence on the relative performances of current optimization methods with as many as twenty or more variables would be welcome, but until such evidence becomes available, Table 3 summarizes much of what is currently known.

Table 3

Comparison in 5, 10 and 20 dimensions

METHOD	NUMBER OF DIMENSIONS		
	5-DIM.	10-DIM.	20-DIM.
DSC	303 281 307	2,269 938 1,378	5,183 5,924 8,254
R	465 465 388 384	1,210 1,258 1,298	10,208 4,681 8,411
N	229 195 298	752 962 970	6,970 12,100 10,246
P	104 103	329 369	1,519 2,206
F	354 288 216	1,639 2,860 1,276	4,200 7,854 12,348
D	114 138	396 319	1,764 1,428 2,541
PSS	21 22	35 34	46* 65*
B	42 41 37	N.A	N.A

3.4 *Conclusions*

Of the methods tried for unconstrained optimization, that due to Fletcher and Powell (1963) (Davidon's method), was the most consistently successful.

Powell's (1964) method performed remarkably similarly to Davidon's method for the five-, ten- and twenty-dimensional test functions, and in consequence was concluded to be virtually as efficient as Davidon's method. Both these methods, Powell (1964) and Fletcher and Powell (1963), possess quadratic convergence, i.e. the property that they will converge to the minimum of a quadratic function in a finite number of steps, and although such functions rarely occur in practice, it is nevertheless found that methods with this feature converge more rapidly, particularly of course

in the vicinity of the optimum. The method of Fletcher and Reeves (1964) also possesses quadratic convergence, but was found to be less successful, however. A possible explanation is that this method, having searched in a manner that would locate the exact minimum of a quadratic surface, periodically discards all the information which it has collected about the actual surface, and commences a completely new and independent search from the best point yet obtained. (The motivation for this drastic but necessary procedure will be found in Fletcher and Reeves' paper.) The similarity of the performances of the methods of Powell (1964) and Fletcher and Powell is considered to be a justification for the manner whereby gradient and non-gradient methods were compared by means of the "$(n + 1)$-factor."

That two methods possessing quadratic convergence should perform substantially better with the five-, ten- and twenty-dimensional test functions than methods without this property, prompts the question as to whether the members of this family of test functions are essentially quadratic. The contours of a two-dimensional member of this family are therefore given in **Fig. 2,** where for convenience the minimum has been taken to lie at the origin of the co-ordinates. As with the function plotted in Fig. 1, the existence of a curved valley and of asymmetric contours, rising more rapidly on one side of the valley than the other, are again noted. However, these features are much less marked than with the function plotted in Fig. 1. The six crosses indicate typical possible starting points. Whilst it seems likely that the minimization of this surface would be easier than finding the optimum of the function given in Fig. 1, it is felt that the assumption of a quadratic nature for the surface would not aid the optimization very much. If the departure of surfaces of this family from quadratic were only slight, then the method of Fletcher and Reeves could be expected to perform very much better than was in fact observed. In higher dimensions, cross-sections of the surface resemble Fig. 2, presumably.

The method of Nelder and Mead (1965) performed well in two dimensions, and also to a lesser extent in three dimensions, but for more dimensions it was progressively less successful. In their paper, Nelder and Mead compare the performances of their method and the method of Powell (1964) with functions of two, three and four variables. Their method is shown to perform better than that of Powell with these problems, a result which is surprising, bearing in mind just what the two algorithms do. This "anomaly" is put into perspective by the observation of this paper that the superior performance of Nelder and Mead's method does not continue as the number of independent variables is increased.

The conclusion was also reached that, when solving sets of simultaneous non-linear equations, methods such as those due to Powell (1965) or Barnes (1965) should be used in preference to merely minimizing the sum of squared residuals, which reaffirms Powell's conclusion.

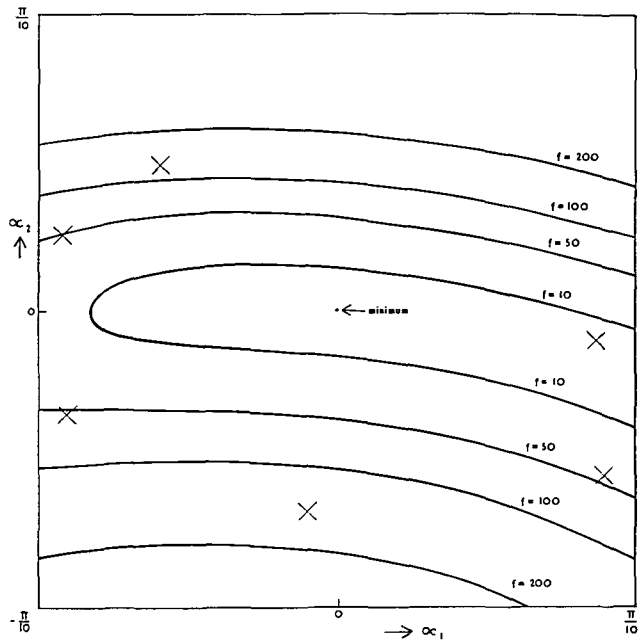

Fig. 2.—Contours of 2-dimensional trigonometric function

It is important to note that in this paper, the case where all the residuals cannot be simultaneously reduced to zero, i.e. no exact solution exists, has not been considered.

For the two- and three-dimensional test functions, directions exist along which these functions are monotonic decreasing, and some methods which seek the minimum along a line precisely, and especially those which require to straddle the minimum along this line, will break down when using such a direction. However, it should be possible to recognize this situation, which can occur with any of the methods considered in this paper except those of Rosenbrock and of Nelder and Mead, and to limit the number of steps taken along such a direction, but the author has not attempted to do this.

4. Transformations

In Section 3.4 it was concluded that the property of quadratic convergence is definitely desirable for a method designed to find unconstrained optima. As the author had access to no quadratically convergent method for optimizing a function of arbitrary form subject to the most general non-linear inequality constraints, a study was made of some cases when it is possible to eliminate the explicit appearance of inequality constraints from the formulation of the problem. It is known that Davidon's method, D, has been successfully used with Carroll's (1961) "Created Response Surface Technique," i.e. an optimizer with quadratic convergence is used after the objective function has been modified by the introduction of a penalty function. The approach that will be adopted here is to *transform the independent variables and leave the objective function unaltered.* Powell's (1964) method will be used rather

than Davidon's, as for many problems the computation of the first derivatives is a major task.

As an example of the type of constraint which can be eliminated, suppose that the three independent variables x_1, x_2, x_3 are to be constrained by $0 \leqslant x_1 \leqslant x_2 \leqslant x_3$. Then consider the transformations

$$\begin{aligned} x_1 &= y_1^2 \\ x_2 &= y_1^2 + y_2^2 \\ x_3 &= y_1^2 + y_2^2 + y_3^2. \end{aligned}$$

It will be seen that an optimizing routine with no provision for incorporating constraints can now be used to seek an *unconstrained* optimum in y-space.

Transformations which have been considered include:

(i) $x_i = y_i^2$
(ii) $x_i = e^{y_i}$
(iii) $x_i = |y_i|$
(iv) $x_i = \sin^2 y_i$
(v) $x_i = \dfrac{e^{y_i}}{e^{y_i} + e^{-y_i}}$.

(i), (ii) and (iii) constrain the x_i to be positive, whilst (iv) and (v) restrict x_i to the range $0 \leqslant x_i \leqslant 1$.

An important practical case of constrained optimization is that when each independent variable is subject to constant lower and upper constraints, $g_i \leqslant x_i \leqslant h_i$, i.e. the permissible region consists of a rectilinear "box" in n dimensions. Then the transformations $x_i = g_i + (h_i - g_i)\sin^2 y_i$ mean that now an *unconstrained* optimum in y-space need be sought. The periodicity of optimal solutions in y-space should not cause any difficulty provided the optimizer in use does not take steps so large that it jumps from peak to peak.

All the transformations which have been considered are such that the neighbourhood of any point Y in y-space maps into the neighbourhood of X in x-space, where Y maps into X. Hence although some of the transformations are not 1 : 1 they *cannot* introduce additional (essentially distinct) local optima.

The advantage of the transformations is that they have been found to result in the correct solutions being obtained easily for problems for which alternative methods made only slow progress, or ceased to make any progress whatsoever once one or two constraints were reached, even when the current point was still a long way from the optimum.

5. Applications of the transformations and comparisons with other methods

The method of Powell (1964) has been used in conjunction with transformations to solve six problems, each of which Box (1965) has solved in three ways. These methods of solution were:

(i) that of Rosenbrock (1960)
(ii) the Complex (constrained simplex) method described by Box (1965)
(iii) the *RAVE* method (*R*osenbrock with *A*utomatic *V*ariable *E*limination) described by Box (1965).

Rosenbrock's method and the Complex method are applicable to the constrained optimization of a general function subject to general inequality constraints, whereas the *RAVE* method and the method of Powell used with transformations are more *approaches* to the problem of the constrained optimization of a general function subject to inequality constraints of somewhat restricted forms.

The six test problems were:

(i) *Box's* (1965) *Problem A*—a simple model

The aim is to maximize the function f, of 5 independent variables $x_1, \ldots x_5$, subject to 8 constraints $g_i \leqslant x_i \leqslant h_i$, $i = 1, \ldots 8$, where x_6, x_7 and x_8 are functions of $x_1, \ldots x_5$ and represent implicit constraints. This problem is set out in Appendix 1, where it will be seen that f, x_6, x_7 and x_8 are all linear in any one independent variable, but quadratic due to cross-product terms. In this example, all the lower bounds g_i, and upper bounds h_i, are constants.

An analytic study of this problem has shown that the lower and upper constraints on x_6 and x_7 and the lower constraint on x_1 can never be approached if the constraints on x_2, x_3, x_4, x_5 and x_8 are to be observed, and consequently these constraints need not be considered further. It is now possible to regard x_2, x_3, x_4, x_5 and x_8 as the five independent variables, each of which is subject to constant lower and upper constraints, and thus the transformations

$$x_i = g_i + (h_i - g_i)\sin^2 y_i$$

may be applied, and Powell's method used to seek an unconstrained optimum in y-space. It is of course to be expected that Rosenbrock's method etc. would perform better with the transformed problem than with the original formulation, but this has not been checked.

(ii) *Box's* (1965) *Problem B*

$$\text{To maximize } f = [9 - (x_1 - 3)^2]\frac{x_2^3}{27\sqrt{3}}$$

subject to
$$\begin{aligned} &0 \leqslant x_1 \\ &0 \leqslant x_2 \leqslant x_1/\sqrt{3} \\ &0 \leqslant x_1 + \sqrt{3}(x_2) \leqslant 6. \end{aligned}$$

The initial point used was

$x_1 = 1$, $x_2 = 0{\cdot}5$, corresponding to $f = 0{\cdot}01336$.

The solution is

$x_1 = 3$, $x_2 = \sqrt{3}$, $f = 1$.

Consider for example the transformations

$$\begin{aligned} X_1 &= x_1 + \sqrt{3}(x_2) \\ X_2 &= x_2 - x_1/\sqrt{3} \end{aligned}$$

and constraints
$$\begin{aligned} 0 &\leqslant X_1 \leqslant 6 \\ -2\sqrt{3} &\leqslant X_2 \leqslant 0. \end{aligned}$$

Thus $x_1 = \frac{1}{2}(X_1 - \sqrt{3}X_2)$

and $x_2 = \frac{1}{2}(X_1/\sqrt{3} + X_2)$.

This formulation includes all the constraints present in the original problem with the exception of the constraint $x_2 \geqslant 0$.

Thus in **Fig. 3,** the original permissible region in x-space is the triangle OAB. The transformation described above gives this triangle plus its reflection in the axis $x_2 = 0$, i.e. the rhombus OABC. Hence it is necessary to consider the region in X-space,

$$0 \leqslant X_1 \leqslant 6$$
$$-2\sqrt{3} \leqslant X_2 \leqslant 0$$

and the transformations

$$x_1 = \tfrac{1}{2}(X_1 - \sqrt{3}X_2)$$
$$x_2 = \tfrac{1}{2}|X_1/\sqrt{3} + X_2|$$

which corresponds to the triangle OAB in x-space as required. The usual transformations

$$X_1 = 6 \sin^2 y_1$$
$$X_2 = -2\sqrt{3} + 2\sqrt{3} \sin^2 y_2$$

can now be applied to give an unconstrained problem in y-space.

(iii) *Rosenbrock's* (1960) *Post Office Parcel Problem*

The P.O.P. problem is to maximize

$$f = x_1 x_2 x_3$$

subject to constraints

$$0 \leqslant x_1 \leqslant 42$$
$$0 \leqslant x_2 \leqslant 42$$
$$0 \leqslant x_3 \leqslant 42$$
$$0 \leqslant x_1 + 2x_2 + 2x_3 \leqslant 72.$$

The solution is $f = 3{,}456$ corresponding to $x_1 = 24$, $x_2 = 12$, $x_3 = 12$.

A possible transformation for this problem is to set

$$x_1 = X_1$$
$$x_2 = X_2$$
$$x_3 = \tfrac{1}{2}(X_3 - X_1 - 2X_2)$$

subject to

$$0 \leqslant X_1 \leqslant 42$$
$$0 \leqslant X_2 \leqslant 42$$
$$0 \leqslant X_3 \leqslant 72$$

which is the standard form for the transformation

$$X_i = g_i + (h_i - g_i) \sin^2 y_i \quad \text{as before.}$$

This incorporates all the original constraints except $x_3 \geqslant 0$. However, as the constraints $x_1 \geqslant 0$ and $x_2 \geqslant 0$ are included, and the initial points used correspond to positive values of f, no maximum-seeking program would search for negative x_3 values, as this would result in negative (smaller) values of f, i.e. obviously no alternative solution exists if the constraint $x_3 \geqslant 0$ is not included.

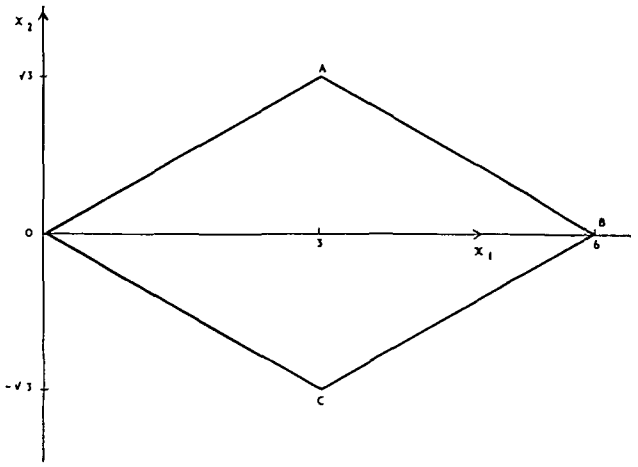

Fig. 3.—Permissible region for Problem B

Cases I, II and III of this problem correspond to the starting points:

I. $x_1 = 10$, $x_2 = 10$, $x_3 = 10$
II. $x_1 = 5$, $x_2 = 10$, $x_3 = 10$
III. $x_1 = 15$, $x_2 = 10$, $x_3 = 10$, respectively.

Case IV also uses the starting point $x_1 = 10$, $x_2 = 10$, $x_3 = 10$ but now the constraints are modified to be

$$0 \leqslant x_1 \leqslant 20$$
$$0 \leqslant x_2 \leqslant 11$$
$$0 \leqslant x_3 \leqslant 42$$
$$0 \leqslant x_1 + 2x_2 + 2x_3 \leqslant 72.$$

The optimum is now $f = 3{,}300$ corresponding to $x_1 = 20$, $x_2 = 11$, $x_3 = 15$. The same transformation may be used:

$$x_1 = X_1$$
$$x_2 = X_2$$
$$x_3 = \tfrac{1}{2}(X_3 - X_1 - 2X_2),$$

but now with constraints

$$0 \leqslant X_1 \leqslant 20$$
$$0 \leqslant X_2 \leqslant 11$$
$$0 \leqslant X_3 \leqslant 72.$$

As before, the constraint $x_3 \geqslant 0$ is not implied by this formulation, but again it can be excluded upon consideration of the form of f, the existence of the constraints $x_1 \geqslant 0$, $x_2 \geqslant 0$ and the fact that the initial point corresponds to a positive f.

The results are given in **Table 4.** The notation p/x is used to denote that the value of f obtained was in error in the pth significant figure by an amount x, and the number of function evaluations necessary to achieve this accuracy is recorded under the heading "trials". The notation "N.A" indicates that the solution of problem B by the *RAVE* method was not attempted, and "EXACT" denotes that the *RAVE* method leads to the exact solution for those cases where the number of constraints effective at the optimum equals the number of independent variables in the problem.

Table 4

Comparison of constrained methods

PROBLEM	ROSENBROCK		COMPLEX		RAVE		TRANSFORMS	
	ERROR	TRIALS	ERROR	TRIALS	ERROR	TRIALS	ERROR	TRIALS
A	3/6	1388	7/7	1201	EXACT	220	8/5	42
B	4/2	310	6/5	76	N.A		10/4	23
P.O.P. I	4/1	600	8/1	205	7/4	133	10/4	55
P.O.P. II	4/1	600	8/1	258	7/4	109	10/4	55
P.O.P. III	4/1	600	8/1	156	7/4	136	10/4	57
P.O.P. IV	4/1	600	8/5	354	EXACT	129	10/4	48

The number of trials entered in this table for the Complex method is for a program which tests all the constraints before entering the function evaluation subroutine, and thus avoids the computation of non-feasible f-values. Box (1965) has found that with these problems Rosenbrock's method would require on average roughly 10% fewer function evaluations if the computation of non-feasible f-values were to be excluded.

Although the run to solve problem A using Rosenbrock's method was continued until 1,388 trials had been performed, in fact no progress whatsoever was made after the 689th trial.

6. Conclusions on the application of transformations

The self-evident conclusion on the use of transformations of the type described in Section 4 is that the solution of constrained problems is eased considerably by eliminating the constraints, so that one of the more powerful methods for unconstrained optimization, e.g. that of Powell (1964), may be employed. For many problems, of course, no transformations can be found, but the scope for ingenuity is considerable, and the advantages of the transformations equally so. If only a few constraints can be eliminated, this will be a worthwhile step forward. The experience of the application of these transformations is limited, but such experience as is available strongly suggests that the possibility of using them in any constrained optimization problem that arises should definitely be borne in mind.

These transformations have successfully been used (again in conjunction with Powell's method) for problems with up to twenty independent variables. These problems arose from a "design of experiments" study (see, for example, Box and Lucas (1959), Behnken (1964)). The aim in each case was to minimize the size (hypervolume) of the confidence region of the estimates of the parameters in the assumed model. Specifically,

(i) In the first model, the transformations $x_i = y_i^2$ were applied to ensure that all free variables remained positive. It had been found that a much smaller confidence region would formally result if certain experiments could be carried out with negative x_i values, a physically absurd situation. In the optimum solution no x_i is zero, however. It can be seen that the constraints are necessary to guide the optimizer to the only local optimum within the permissible region, so that although the optimum does not lie on any constraint, the constraints are necessary for the correct answer to be obtained.

(ii) The transformations $x_i = g_i + (h_i - g_i)\sin^2 y_i$ were applied to all variables in a problem where every x_i had to lie between a pair of constant values. The optimum solution, for the case of twenty independent variables, corresponded to five variables being at their lower limits and five variables being at their upper limits.

Finally the possibility of bringing any optimization problem within the scope of the well-known technique of linear programming (see, for example, Graves and Wolfe (1963)) by means of transformations should not be overlooked. As an example it will be seen that problem A (see Appendix 1) becomes a linear programming problem if new variables are defined as follows:

$$z_1 = x_1$$
$$z_2 = x_1x_2$$
$$z_3 = x_1x_3$$
$$z_4 = x_1x_4$$
$$z_5 = x_1x_5$$

This transformation means that only some half dozen steps are necessary for the solution of the problem to be obtained.

A suitable form for writing the constraints is given in Appendix 2.

7. Acknowledgement

The author is indebted to Mr. M. J. D. Powell for providing results for the 5-, 10- and 20-dimensional test problems, for the cases where the published results for the methods of Fletcher and Powell (1963) and Powell (1964 and 1965) applied to a convergence criterion differing from that used in this paper.

Appendix 1

Problem A—a simple model

The form given below is a simplification of that originally given by Box (1965), but nevertheless the two forms are equivalent. The aim is to maximize $f(x_1, x_2, x_3, x_4, x_5)$, where

$$f = c_0 + c_1x_1 + c_2x_1x_2 + c_3x_1x_3 + c_4x_1x_4 + c_5x_1x_5$$

subject to constraints

$$\begin{aligned} 0 &\leqslant x_1 \\ 1{\cdot}2 &\leqslant x_2 \leqslant 2{\cdot}4 \\ 20 &\leqslant x_3 \leqslant 60 \\ 9 &\leqslant x_4 \leqslant 9{\cdot}3 \\ 6{\cdot}5 &\leqslant x_5 \leqslant 7{\cdot}0 \\ 0 &\leqslant x_6 \leqslant 294{,}000 \\ 0 &\leqslant x_7 \leqslant 294{,}000 \\ 0 &\leqslant x_8 \leqslant 277{,}200 \end{aligned}$$

where

$$\begin{aligned} x_6 &= c_6x_1 + c_7x_1x_2 + c_8x_1x_3 + c_9x_1x_4 + c_{10}x_1x_5 \\ x_7 &= c_{11}x_1 + c_{12}x_1x_2 + c_{13}x_1x_3 + c_{14}x_1x_4 + c_{15}x_1x_5 \\ x_8 &= c_{16}x_1 + c_{17}x_1x_2 + c_{18}x_1x_3 + c_{19}x_1x_4 + c_{20}x_1x_5 \end{aligned}$$

and where

$$\begin{aligned} c_0 &= -24{,}345 & c_1 &= -8{,}720{,}288{\cdot}849 & c_2 &= 150{,}512{\cdot}5253 \\ c_3 &= -156{\cdot}6950325 & c_4 &= 476{,}470{\cdot}3222 & c_5 &= 729{,}482{\cdot}8271 \\ c_6 &= -145{,}421{\cdot}402 & c_7 &= 2{,}931{\cdot}1506 & c_8 &= -40{\cdot}427932 \\ c_9 &= 5{,}106{\cdot}192 & c_{10} &= 15{,}711{\cdot}36 & c_{11} &= -155{,}011{\cdot}1084 \\ c_{12} &= 4{,}360{\cdot}53352 & c_{13} &= 12{\cdot}9492344 & c_{14} &= 10{,}236{\cdot}884 \\ c_{15} &= 13{,}176{\cdot}786 & c_{16} &= -326{,}669{\cdot}5104 & c_{17} &= 7{,}390{\cdot}68412 \\ c_{18} &= -27{\cdot}8986976 & c_{19} &= 16{,}643{\cdot}076 & c_{20} &= 30{,}988{\cdot}146. \end{aligned}$$

The reader should not assume that the c_i are accurate to all the figures given. The problem was presented in the form of a computer program to calculate f, x_6, x_7 and x_8 from x_1, x_2, x_3, x_4 and x_5. Many intermediate quantities were calculated by this original program using regression coefficients to far fewer than ten significant figures. The amount of arithmetic necessary to compute f, x_6, x_7 and x_8 could be much reduced by eliminating the explicit computation of these intermediate quantities, as they were not needed in the optimization problem. It was, however, necessary to retain all the figures given in the c_i in order that the values of f calculated by the two programs should agree to 7 or 8 significant figures.

The starting point used was

$$\begin{aligned} x_1 &= 2{\cdot}52 \\ x_2 &= 2 \\ x_3 &= 37{\cdot}5 \\ x_4 &= 9{\cdot}25 \\ x_5 &= 6{\cdot}8 \end{aligned}$$

corresponding to $f = 2{,}351{,}244$ and $x_8 = 130{,}368$.

The optimum is $f = 5{,}280{,}334$ corresponding to

$$\begin{aligned} x_1 &= 4{\cdot}53743 \\ x_2 &= 2{\cdot}4 \\ x_3 &= 60 \\ x_4 &= 9{\cdot}3 \\ x_5 &= 7{\cdot}0 \\ x_6 &= 75{,}570 \\ x_7 &= 198{,}157 \\ x_8 &= 277{,}200 \end{aligned}$$

Appendix 2

Linear programming formulation of Problem A

To maximize

$$f = c_0 + c_1z_1 + c_2z_2 + c_3z_3 + c_4z_4 + c_5z_5$$

subject to

$$\begin{aligned} z_1 &\geqslant 0 \\ z_2 &\geqslant 0 \\ z_3 &\geqslant 0 \\ z_4 &\geqslant 0 \\ z_5 &\geqslant 0 \end{aligned}$$

$$\begin{aligned} z_6 &= 2{\cdot}4z_1 - z_2 &&\geqslant 0 \\ z_7 &= -1{\cdot}2z_1 + z_2 &&\geqslant 0 \\ z_8 &= 60z_1 - z_3 &&\geqslant 0 \\ z_9 &= -20z_1 + z_3 &&\geqslant 0 \\ z_{10} &= 9{\cdot}3z_1 - z_4 &&\geqslant 0 \\ z_{11} &= -9{\cdot}0z_1 + z_4 &&\geqslant 0 \\ z_{12} &= 7{\cdot}0z_1 - z_5 &&\geqslant 0 \\ z_{13} &= -6{\cdot}5z_1 + z_5 &&\geqslant 0 \\ z_{14} &= c_6z_1 + c_7z_2 + c_8z_3 + c_9z_4 + c_{10}z_5 &&\geqslant 0 \\ z_{15} &= -c_6z_1 - c_7z_2 - c_8z_3 - c_9z_4 - c_{10}z_5 + 294{,}000 &&\geqslant 0 \\ z_{16} &= c_{11}z_1 + c_{12}z_2 + c_{13}z_3 + c_{14}z_4 + c_{15}z_5 &&\geqslant 0 \\ z_{17} &= -c_{11}z_1 - c_{12}z_2 - c_{13}z_3 - c_{14}z_4 - c_{15}z_5 + 294{,}000 &&\geqslant 0 \\ z_{18} &= c_{16}z_1 + c_{17}z_2 + c_{18}z_3 + c_{19}z_4 + c_{20}z_5 &&\geqslant 0 \\ z_{19} &= -c_{16}z_1 - c_{17}z_2 - c_{18}z_3 - c_{19}z_4 - c_{20}z_5 + 277{,}200 &&\geqslant 0 \end{aligned}$$

References

BARNES, J. G. P. (1965). "An algorithm for solving non-linear equations based on the secant method," *The Computer Journal*, Vol. 8, p. 66.

BEHNKEN, D. W. (1964). "Estimation of Copolymer Reactivity Ratios: An Example of Nonlinear Estimation," *Journal of Polymer Science:* Part A, Vol. 2, p. 645.

BOX, G. E. P., and LUCAS, H. L. (1959). "Design of Experiments in Non-Linear Situations," *Biometrika*, Vol. 46, p. 77.

BOX, M. J. (1965). "A new method of constrained optimization and a comparison with other methods," *The Computer Journal*, Vol. 8, p. 42.

CARROLL, C. W. (1961). "The Created Response Surface Technique for Optimizing Nonlinear Restrained Systems," *Operations Research*, Vol. 9, p. 169.

DAVIDON, W. C. (1959). "Variable Metric Method for Minimization," A.E.C. Research and Development Report, ANL-5990 (Rev).

FLETCHER, R. (1965). "Function minimization without evaluating derivatives—a review," *The Computer Journal*, Vol. 8, p. 33.

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

FLETCHER, R., and REEVES, C. M. (1964). "Function minimization by conjugate gradients," *The Computer Journal*, Vol. 7, p. 149.

GRAVES, R. L., and WOLFE, P. (Eds) (1963). *Recent Advances in Mathematical Programming*, New York: McGraw-Hill Book Co.

NELDER, J. A., and MEAD, R. (1965). "A simplex method for function minimization," *The Computer Journal*, Vol. 7, p. 308.

POWELL, M. J. D. (1964). "An efficient method of finding the minimum of a function of several variables without calculating derivatives," *The Computer Journal*, Vol. 7, p. 155.

POWELL, M. J. D. (1965). "A method for minimizing a sum of squares of non-linear functions without calculating derivatives," *The Computer Journal*, Vol. 7, p. 303.

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

SPENDLEY, W., HEXT, G. R., and HIMSWORTH, F. R. (1962). "Sequential Applications of Simplex Designs in Optimisation and Evolutionary Operation," *Technometrics*, Vol. 4, p. 441.

SWANN, W. H. (1964). "Report on the development of a new direct searching method of Optimisation," I.C.I. Ltd., Central Instrument Laboratory Research Note 64/3.

Correspondence (*continued from p.* 66)

$\delta_1 = 0$ designated a flight or flights Sydney–Melbourne and Melbourne–Sydney prior to Sydney–Adelaide flight (direct or via Melbourne depending on δ_2).

Other variables such as the number of flights between any points were required to be integral.

The resulting non-linear integer programming problem was eventually reduced to a series of linear integer problems and run on a Honeywell H-400 6-tape computer.

The initial results appeared excellent at first sight, with annual usages of 3,701·4 hours and 3,688·4 hours. However, the time of loading of passengers at Adelaide for Sydney was not satisfactory to the airline both from the point of view of poor loading and "customer goodwill." Savings could be made in the degree of daylight servicing provided in the schedule, but the other factors were more important and this reflects back to Elizabeth Barraclough's desire for parameters in the input to relate to the "degree of success" of a timetable.

At this stage, it appeared that to rectify matters and incorporate other commercial judgements, it would be difficult to quantify them all within the memory confines of even very large computers. Matters such as mathematically specifying the nature and interdependence of the frequency of a service and the expected passenger traffic arose. The program was modified to avoid some unacceptable time periods. Memory size quickly limited any further conditions being incorporated.

The results showed that the departure time Sydney–Perth was critical but it was observed that, with several cases, it would be possible to improve this transcontinental service to one flight every day by extracting one daily Sydney–Melbourne —Sydney flight at a low loading time, and using the overall weekly time gained to give almost exactly the time required to achieve seven flights per week to Perth instead of five. This was without disturbing the remainder of the schedule, and was a matter which was apparent by some human observation but not apparent to the computer method. By the time the computer study was completed, the results agreed with the best manual schedule derived independently, and it was verified by the machine that the optimum had been reached for this case.

An additional point which arose seems worthy of mention based on the experience gained regarding formulating commercial judgements: the desirability of parameters in the input related to the degree of success and the advantage of some human intervention. It may be that Dr. Miller's type of linear programming study which indicated very excessive operating costs including "overscheduling", may have suffered due to problem simplification rather than have overscheduling costs introduced in explanations. The only factors which would seem to contribute to operating costs—low load factors, low utilization and uneconomic aircraft—would cover "overscheduling" and were discussed in the study.

B. S. THORNTON.

Honeywell Pty. Limited,
E.D.P. Division,
55 Macquarie Street,
Sydney, Australia.

21 January 1966.