A new method of constrained optimization and a comparison with other methods

By M. J. Box*

A new method for finding the maximum of a general non-linear function of several variables within a constrained region is described, and shown to be efficient compared with existing methods when the required optimum lies on one or more constraints. The efficacy of using effective constraints to eliminate variables is demonstrated, and a program to achieve this easily and automatically is described. Finally, the performance of the new method (the "Complex" method) with unconstrained problems, is compared with those of the Simplex method, from which it was evolved, and Rosenbrock's method.

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

1. Introduction

The first part of this paper shows, by recording attempts to solve a practical problem, that constrained optimization is difficult. (By a constrained optimum is meant one for which the solution corresponds to certain variables lying at the edges of their permissible ranges, since if this is not the case a method with no provision for bounding the variables will produce the same result.) The constraints under consideration are inequality constraints, which are assumed to apply to functions of the independent variables as well as to the independent variables themselves, i.e. the problem is to maximize (or minimize) a function $f(x_1, \ldots x_n)$ of n independent variables $x_1, \ldots x_n$ subject to m constraints of the form $g_k \leqslant x_k \leqslant h_k$, $k = 1, \ldots m$, where $x_{n+1}, \ldots x_m$ are functions of $x_1, \ldots x_n$, and the lower and upper constraints g_k and h_k are either constants or functions of $x_1, \ldots x_n$.

As the author had access at the time to only one program for constrained optimization, namely that due to Rosenbrock (Rosenbrock, 1960), this method was applied first, but met with only limited success. For unconstrained optimization, a program incorporating the Simplex method of Spendley *et al.* (Spendley, Hext and Himsworth, 1962) was available, so an obvious line of research was to develop a constrained version of the Simplex method. This has been done, and of the original method only the basic principle remains, all the details having been changed.

At the end of this paper, an objective comparison of the Simplex and Rosenbrock's methods for unconstrained problems is given.

All comparisons are made on the basis of the number of function evaluations, since for many real problems this is vastly in excess of the time to organize the search. An example is the problem of determining parameters in highly non-linear differential equations from experimental data. In this problem, the sum of squared residuals between experimental data and numerically integrated solutions of the differential equations is accumulated for imputed values of the parameters. The parameters are then systematically varied by some hill-climbing procedure so as to locate the minimum of the error surface. It has been known for the evaluation of the error function in this problem to take up to 1,000 times as long as the organization of the search.

For constrained gradient (as opposed to direct search) methods, the reader is referred to Rosen's "Gradient Projection Method" (Rosen, 1960, 1961) and to Carroll's "Created Response Surface Technique" (Carroll, 1961). The latter method is known to work quite well when used in conjunction with the method described originally by Davidon (Davidon, 1959), and in a refined form by Fletcher and Powell (Fletcher and Powell, 1963).

2. Unsatisfactory attempts to solve a practical problem

The author's interest in optimization stems from attempts to solve the following problem:

Problem A—A simple model.

To maximize the function f, of 5 variables, subject to 8 constraints given below;

$$
\begin{aligned}
b &= x_2 + 0{\cdot}01x_3 \\
x_6 &= (k_1 + k_2x_2 + k_3x_3 + k_4x_4 + k_5x_5)x_1 \\
y_1 &= k_6 + k_7x_2 + k_8x_3 + k_9x_4 + k_{10}x_5 \\
y_2 &= k_{11} + k_{12}x_2 + k_{13}x_3 + k_{14}x_4 + k_{15}x_5 \\
y_3 &= k_{16} + k_{17}x_2 + k_{18}x_3 + k_{19}x_4 + k_{20}x_5 \\
y_4 &= k_{21} + k_{22}x_2 + k_{23}x_3 + k_{24}x_4 + k_{25}x_5 \\
x_7 &= (y_1 + y_2 + y_3)x_1 \\
x_8 &= (k_{26} + k_{27}x_2 + k_{28}x_3 + k_{29}x_4 + k_{30}x_5)x_1 + x_6 \\
&\quad + x_7 \\
f &= (a_2y_1 + a_3y_2 + a_4y_3 + a_5y_4 + 7840a_6 - 100000a_0 \\
&\quad - 50800ba_7 + k_{31} + k_{32}x_2 + k_{33}x_3 + k_{34}x_4 \\
&\quad + k_{35}x_5)x_1 - 24345 + a_1x_6
\end{aligned}
$$

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

where x_1, x_2, x_3, x_4 and x_5 are the independent variables, and the required optimum must satisfy the constraints

$$\begin{aligned} 0 &\leqslant x_1 \\ 1{\cdot}2 &\leqslant x_2 \leqslant 2{\cdot}4 \\ 20 &\leqslant x_3 \leqslant 60 \\ 9 &\leqslant x_4 \leqslant 9{\cdot}3 \\ 6{\cdot}5 &\leqslant x_5 \leqslant 7{\cdot}0 \\ 0 &\leqslant x_6 \leqslant 294000 \\ 0 &\leqslant x_7 \leqslant 294000 \\ 0 &\leqslant x_8 \leqslant 277200. \end{aligned}$$

The values of the a_i and the k_i will be found in the Appendix.

It will be seen that f, x_6, x_7 and x_8 are linear in any one variable but are quadratic due to cross-product terms, i.e. the problem is frustratingly just beyond the scope of linear programming.

The given initial point 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 $f = 2{,}351{,}244$.

The optimum value is now known to be $f = 5{,}280{,}334$ corresponding to

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

As Rosenbrock's program (Rosenbrock, 1960) was the only program available for constrained optimization of a general function, this method was tried first. The program ceased to make any progress whatsoever after 689 trials, although the run was continued up to 1,388 trials. The following values had then been produced:

$$\left.\begin{aligned} x_1 &= 4{\cdot}58353 \\ x_2 &= 2{\cdot}21545 \\ x_3 &= 31{\cdot}77327 \\ x_4 &= 9{\cdot}29997 \\ x_5 &= 6{\cdot}99863 \end{aligned}\right\} \text{ corresponding to } \begin{aligned} f &= 5{,}222{,}459 \\ x_6 &= 78{,}989{\cdot}85 \\ x_7 &= 194{,}722{\cdot}60 \\ x_8 &= 277{,}177{\cdot}96. \end{aligned}$$

On starting the optimization from a slightly different point, a similar function value was obtained, but rather different values of the independent variables were produced, notably $x_3 \simeq 37$. This raised doubt as to just how near to the true optimum these solutions were. A number of modifications were made in turn to Rosenbrock's program in an attempt to obtain a larger final function value. These included:

(i) varying the widths of the boundary regions,

(ii) projecting the vector of perturbations existing at the end of a stage on to the new axes of search,

(iii) resetting all the perturbations at the beginning of a stage to a constant value, or to a fraction of the total progress achieved during the last stage,

(iv) relating the perturbation to be used at the beginning of a stage for the axis in the direction of total progress of the last stage to the magnitude of that progress, all other perturbations to be only one-tenth as large,

(v) trying alternative forms for the attenuating function in the boundary region,

(vi) using the logarithms of all variables.

All these modifications led to values of f between 5,170,923 and 5,233,320 which values, it transpired, are respectively 2·1% and 0·9% below the true optimum. The indeterminacy in the values of the x_i is of more interest, however. In all these runs, progress had ceased completely by, or was very slow after over 1,000 trials.

A constrained "one variable at a time" method (constructed by removing those parts of Rosenbrock's program which rotate the axes at the conclusion of a stage and those which modify the function when the boundary regions are entered) was then considered. The value $f = 5{,}214{,}820$ was obtained in 150 trials, whereupon no further progress was made (although the run was continued up to 5,350 trials).

A random search was then carried out. Random points within a cell about the best point to date were selected, and if all the constraints were satisfied and the function value was larger than the previous best, then this point became the centre of the cell. Of 3 runs with different pseudo-random number initiators, the best results obtained were $f = 4{,}659{,}433$ after 954 trials, with successful points being selected rarely and giving rise to only a small improvement in the function.

3. The new method (Complex)

The constrained Simplex (Complex) method searches for the maximum value of a function $f(x_1, \ldots x_n)$ subject to m constraints of the form $g_k \leqslant x_k \leqslant h_k$, $k = 1, \ldots m$, where $x_{n+1}, \ldots x_m$ are functions of $x_1, \ldots x_n$, and the lower and upper constraints g_k and h_k are either constants or functions of $x_1, \ldots x_n$. (To find a minimum, $-f$ is maximized.) It has been developed from the Simplex method of Spendley *et al.* (Spendley, Hext and Himsworth, 1962). It is assumed that an initial point $x_1^\circ, \ldots x_n^\circ$, which satisfies all the m constraints is available.

In this method, $k \geqslant n + 1$ points are used, of which one is the given point. The further $(k - 1)$ points required to set up the initial configuration are obtained one at a time by the use of pseudo-random numbers and ranges for each of the independent variables, viz. $x_i = g_i + r_i(h_i - g_i)$ where r_i is a pseudo-random deviate rectangularly distributed over the interval (0, 1).

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

A point so selected must satisfy the explicit constraints, but need not satisfy all the implicit constraints. If an implicit constraint is violated, the trial point is moved halfway towards the centroid of those points already selected (where the given initial point is included). Ultimately a satisfactory point will be found. (It is assumed that the feasible region is convex.) Proceeding in this way, $(k - 1)$ points are found which satisfy all the constraints.

The function is evaluated at each vertex, and the vertex of least function value is replaced by a point $\alpha \geqslant 1$ times as far from the centroid of the remaining points as the reflection of the worst point in the centroid, the new point being collinear with the rejected point and the centroid of the retained vertices. If this trial point is also the worst, it is moved halfway towards the centroid of the remaining points to give a new trial point. The above procedure is repeated until some constraint is violated.

If a trial vertex does not satisfy some constraint on some independent variable x_i, $i = 1, \ldots n$, that variable is re-set to a value 0·000001 inside the appropriate limit; if some implicit constraint x_j, $n + 1 \leqslant j \leqslant m$ is violated, the trial point is moved halfway towards the centroid of the remaining points. Ultimately a permissible point is found. Thus as long as the complex has not collapsed into the centroid, progress will continue.

The use of over-reflection by a factor $\alpha > 1$ tends to cause a continual enlargement of the complex and thus to compensate for the moves halfway towards the centroid. Furthermore, it enables rapid progress to be made when the initial point is remote from the optimum. It is also an aid towards maintaining the full dimensionality of the complex. So too is the use of $k > n + 1$ points, since with $k = n + 1$ points only, the complex is liable to collapse into a subspace. In particular, it tends to flatten itself against the first located constraint and thus be unable to move along an additional constraint when a corner is reached.

The ability of the complex to turn a corner can be explained in the following way. Consider an optimization in which the $k > n + 1$ points of the complex lie roughly in a subspace parallel to a constraint. Then if the contours of the function change considerably (or the intersection of two constraints is reached), progress may be maintained by one of the following features:

(i) The over-reflection factor $\alpha > 1$ may immediately enlarge the complex and move it in the desired direction (if the points do not strictly lie in the subspace).

(ii) The complex may shrink from being long and narrow to a very small size, and then behave as in (i).

(iii) When a corner is reached, the device of setting an explicit variable 0·000001 inside its limit usually means that a point not lying in the subspace of the other points is introduced.

The method of setting up the initial complex avoids the difficulty of constructing a regular simplex which satisfies all the constraints and is of reasonable size; furthermore, the initial array is roughly scaled to the orders of the various variables, i.e. the programmer does not need to scale his problem.

The only stopping criterion built into the program is a conservative one, namely that the program shall stop itself when five consecutive equal function evaluations have occurred, which give values of f which are "equal" to the accuracy of the computer word-length being used. This means the program will not terminate when there is any chance of further improvement in the function, but avoids fruitless machine time when the complex has shrunk to such a size that changes in the function are smaller than one digit in the least significant place. The usual method for checking that the global rather than a local maximum has been found is to restart the program from different points, and infer that if they all converge to the same solution then a global optimum has been found. In several dimensions, for a problem for which the feasible region of parameter space is small, the discovery of an alternative permissible initial point can present considerable difficulty. With the Complex method, there is no difficulty in using the same initial point with different pseudo-random number initiators to perform such a rough check as to whether the optimum is indeed global.

Intuitively the Complex method is likely to find a higher optimum than Rosenbrock's method if the permissible region contains several local maxima, as Rosenbrock's method will converge to that local maximum which is "nearest" to the initial point in some sense. In the Complex method, some of the randomly generated points will be relatively remote from the initial point and may be in the vicinity of a higher peak; moreover, the initial few over-reflections may well throw the program from end to end of the permitted region, i.e. the first few trials scan the whole permitted region. No systematic search for alternative optima is made, however.

The problem of locating a feasible point is usually tackled as follows; suppose all the constraints are of the form $g_i \geqslant 0$. Then we maximize $\sum_{i=1}^{m} g_i$ where the summation is only taken over those constraints which are violated. When a maximum of zero is obtained we have a feasible point. This approach is plausible for use with any optimizing technique which does not make use, even implicit use, of the continuity of the first derivatives of the function it is optimizing. Of course it does not prove the non-existence of a feasible point when it is unable to find one.

4. The selection of suitable values of α and k

As constraint-bound optimization problems occur sparsely in the literature, the following problem was introduced to assist in the selection of suitable values of α and k:

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

Problem B—This problem has a different type of constraint to problem A, namely one in which the current limit for one of the variables (x_2) depends on the current value of another variable (x_1) instead of the constraints all being constants.

To maximize the function f, of 2 variables, subject to 3 constraints given below;

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

subject to $0 \leqslant x_1$

$$0 \leqslant x_2 \leqslant \frac{x_1}{\sqrt{3}}$$

$$0 \leqslant x_3 = x_1 + \sqrt{3}(x_2) \leqslant 6.$$

The initial point used in each case was

$$x_1 = 1$$
$$x_2 = 0{\cdot}5$$

corresponding to $f = 0{\cdot}01336$.

The optimum value is 1 at $x_1 = 3$, $x_2 = \sqrt{3}$.

In **Tables 1** and **2** are tabulated the best function values obtained for problems A and B after 200 function evaluations for a range of values of α and k. Those entries marked with an asterisk indicate that the stopping criterion ended the run before 200 trials had been performed.

After examination of these tables, it was decided to take as reflection factor $\alpha = 1{\cdot}3$ and $k = 2n$ as the number of vertices for constrained problems, the choice apparently not being critical.

5. The solution of problem A by the Complex method

With the above choice of α and k, and using $x_1 \leqslant 5$ as the upper limit on x_1 for setting up the initial complex, the program stopped after 1,440 trials with a function value $f = 5{,}279{,}932$. Of these trials only 881 satisfied all the constraints, i.e. were permissible, so that it would be advantageous to test all the explicit and implicit constraints before evaluating the function. The point is that the program could always be made to test whether

Table 1

Problem A—Function value after 200 trials

NUMBER OF VERTICES k	REFLECTION FACTOR α						AVERAGE
	1·0	1·1	1·2	1·3	1·4	1·5	
6	5,175,501	5,177,024	5,156,428	5,174,663	5,171,262	5,168,845	5,170,620
7	5,127,218	5,167,767	5,167,704	5,170,132	5,183,046	5,193,077	5,168,157
8	5,198,689	5,162,963	5,216,529	5,153,434	5,169,406	5,141,928	5,173,824
9	5,145,139	5,218,889	5,223,309	5,190,207	5,160,370	5,195,258	5,188,862
10	5,196,802	5,184,167	5,194,993	5,213,805	5,188,108	5,158,949	5,189,470
11	5,167,050	5,207,356	5,193,535	5,204,301	5,213,832	5,218,021	5,200,682
12	5,187,422	5,148,227	5,194,280	5,171,596	5,184,185	5,192,269	5,179,663
Average	5,171,117	5,180,913	5,192,396	5,182,591	5,181,458	5,181,192	5,181,611

Table 2

Problem B—Function value after 200 trials

NUMBER OF VERTICES k	REFLECTION FACTOR α						AVERAGE
	1·0	1·1	1·2	1·3	1·4	1·5	
3	0·96808	0·98309	0·99916	1·00000	1·00000*	1·00000*	0·99172
4	0·99436	0·99996	0·99987	0·99999*	0·99999*	0·99998	0·99902
5	0·98541	0·99944	0·99934	1·00000*	1·00000	0·99999	0·99736
6	0·99396	0·99970	0·99908	0·99999	0·99959	0·99997	0·99871
7	0·99708	0·99974	0·99990	0·99601	0·99652*	0·99990	0·99819
8	0·99951	0·99965	0·99727	0·99856	0·99980	0·99831	0·99885
Average	0·98973	0·99693	0·99910	0·99909	0·99931	0·99969	0·99731

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

the explicit constraints are satisfied if these are constant, without entering any auxiliary. If, however, the limits of some of the independent variables are not constant, i.e. depend on the values of other independent variables, or if the problem has implicit constraints, an auxiliary sequence must be entered before the permissibility of any point is known. Rosenbrock considers that the program is easier to use if only one auxiliary has to be provided, even if it has to serve three purposes: evaluating constraints which are not constant, evaluating implicit variables and of course computing the function itself. Runs with several constraint-bound test problems have shown that for these problems, separating the function evaluation from the evaluation of the constraints and implicit variables would lead on average to Rosenbrock's program requiring 11% less function evaluations, and to the Complex program requiring 36% less function evaluations than they do at present.

The function value $f = 5{,}236{,}850$, which is better than that obtained by any modification of Rosenbrock's method, was obtained after 517 trials (300 permissible). The program was restarted from the above best point, stopping after a further 679 trials (320 permissible) with $f = 5{,}280{,}327$, which compares well with the true solution 5,280,334. (Note that Rosenbrock's program makes no provision to re-start at the best point if this lies within one or more boundary regions.)

The particularly interesting feature of this solution was the values of the variables x_i:

$x_1 = 4{\cdot}537459$ $\quad x_2 = 2{\cdot}399963$ $\quad x_3 = 59{\cdot}999500$
$x_4 = 9{\cdot}300000$ $\quad x_5 = 6{\cdot}999994$ $\quad x_8 = 277{,}199{\cdot}91$

x_6 and x_7 being nowhere near their upper constraints. Thus x_2, x_3, x_4, x_5 and the additional constraint x_8 are at their respective upper limits, the upper constraint on x_8 being effectively an upper constraint on x_1. It had not been found possible to predict this by an examination of the function f. The nature of the true solution had not even vaguely been suggested by Rosenbrock's method, it now being apparent that the optimum obtained by this method was 1·1% below the true value.

In obtaining the solution, it appeared that the complex had flattened itself against the constraint on x_8 and had then rolled along it with relative ease before flattening itself against further constraints until it ended up in the appropriate corner of the permissible region.

At the optimum point, it was found that

$$\frac{\partial f}{\partial x_1} = 1{,}169{,}093$$

$$\frac{\partial f}{\partial x_2} = 682{,}940$$

$$\frac{\partial f}{\partial x_3} = -711$$

$$\frac{\partial f}{\partial x_4} = 2{,}161{,}951$$

$$\frac{\partial f}{\partial x_5} = 3{,}309{,}977$$

Understandably many methods tried on this problem have had a lot of difficulty driving x_3 to its limit (and, to a lesser degree x_2), because $\frac{\partial f}{\partial x_3}$ is so small. The compatibility of the negative sign for $\frac{\partial f}{\partial x_3}$ and x_3 being set to its *upper* limit rests upon the fact that $\frac{\partial x_8}{\partial x_3}$ is negative at the optimum. At the optimum the angle between the normal to the implicit constraint $x_8 = 277{,}200$ and the gradient of f is $\cos^{-1}(\cdot 99347) \sim 6\frac{1}{2}°$, i.e. as had been suspected, this constraint is not far from being parallel to the contours of f.

6. Solutions of other problems by the Complex method

Others problems have been solved with similar precision and efficiency, the brief details being as follows.

(i) Rosenbrock's method solved his four cases of the Post Office Parcel Problem (Rosenbrock, 1960) in 600 trials, the function value being in error by 1 unit in the fourth significant figure in each case. The Complex program had corresponding errors of 1, 1, 1 and 5 units in the eighth significant figure, the program stopping after 310, 342, 272 and 576 trials, respectively (205, 258, 156 and 354 permissible).

(ii) Problem B was solved by Rosenbrock's method (several variations) with an error of 2 units in the fourth significant figure in about 310 trials; the Complex program stopped after 159 trials (76 permissible) with $f = 0{\cdot}999995$.

7. Further investigations into the problem of constraint-bound optima

The suggestion, originating from Rosenbrock himself, that possibly wider boundary regions should be used in his program has been found to be of no benefit as far as problem A is concerned.

Rosenbrock-type boundary regions of varying width have been used with the Complex program in an attempt to discover whether they effectively introduce a false maximum, bearing in mind the finite word-length of the computer. In this event, the instance of Rosenbrock's method "getting stuck" would be explained. However, in every case, the first run of the Complex program led us to solutions of problem A of the order of 5,279,800. Therefore no alternative boundary region schemes for Rosenbrock's program were studied.

Rosenbrock's program has successfully solved problem A as follows. The solution originally obtained showed x_8 to have closely approached its constraint and therefore the equation $x_8 = h_8$ was used to eliminate x_1, and the modified problem with only four independent variables re-run. By sequentially using effective constraints to eliminate variables, the solution $f = 5{,}280{,}323$ was

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

ultimately obtained, but only at the expense of reprogramming the problem several times.

The conclusion was then reached that the constraints had not received adequate attention from previous workers. In constraint-bound problems, the constraints are of as much importance as the contours and gradient of the function (cf. linear programming). This suggested that optimizers which are essentially for unconstrained problems do not necessarily become efficient constrained optimizers merely by the addition of some penalty function concept.

The following variant of Rosenbrock's method was then tried. Suppose that at the end of a stage one constraint only is effective, i.e. its boundary region has been entered. The gradient of the normal to this constraint can be estimated by perturbations about the current point. Then one axis of search for the next stage is set parallel to this normal and a further $(n - 1)$ mutually orthogonal axes are constructed. These latter must be locally parallel to the constraint, and so we might hope for reasonable progress along the constraint. For problem A, using the normal to the constraint $x_8 = h_8$, no advantage was gained, presumably because the curvature of the constraint is excessive. It was found that the current point moved even nearer to the constraint rather than along it.

Alternatively, again considering problem A, suppose up to $(n - 1)$ constraints are effective. Then it is possible to compute a direction which is locally parallel to all these constraints, and to set up the further $(n - 1)$ mutually orthogonal axes in the usual way. This, however, was no more successful.

8. The RAVE program

The methods of the two previous paragraphs are attempts to set up implicitly a search parallel to the effective constraint(s). It had been thought that implicit elimination of variables would be of more general application than explicit elimination, but as attempts to implement it were unsuccessful, the less general method was then considered. Accordingly explicit variable elimination has been incorporated into a program RAVE (Rosenbrock Automatic Variable Elimination). In this method, the user is able to specify that some variable be eliminated whenever the current point is found to have entered a boundary region, that of the ith variable say, i.e. within $0{\cdot}0001 \times$ (range of variable x_i) of the lower or upper constraint. This check is only performed at the end of each stage (rotation of axes). The current point will thereafter be made to lie on this constraint, as there is no facility provided whereby the program can test whether it should at any subsequent time be free to leave it. In fact, as Rosenbrock's program cannot be restarted at a point within a boundary region, and in particular on a boundary, it would seem that unlatching from the constraints cannot be done with Rosenbrock's program without some considerable modification.

It is not necessary to use these facilities, when of course the program performs in the normal manner. Experience indicates that even if only one or two variables can be eliminated in this way, then a substantial easing of the problem will have been achieved. The resulting reduction in the total number of independent variables also eases the problem of course. Should the program find n constraints effective and be provided with the appropriate n elimination auxiliaries, then it stops, having solved the problem exactly (subject, of course, to the limitation that boundaries once entered cannot be left) with the solution given as a vertex of the n-dimensional feasible region. In this program all the constraints, implicit and explicit, are tested before the function is evaluated, thus effecting a worthwhile saving in the number of function evaluations, as mentioned earlier.

The variable elimination sequences must satisfy the following rules:

For each of the m upper constraints h_i, an indicator u_i must be set to 0, 1, 2, ... or n, where the value 0 indicates that if x_i enters its upper boundary region, no action is to be taken, but if $u_i = k$, $1 \leqslant k \leqslant n$, then an auxiliary to compute x_k from $x_1, \ldots x_{k-1}$ is provided at label $100 + i$. (In particular x_k may be set to a constant.) Similarly, for each lower constraint g_i, there exists an indicator l_i, and possibly a corresponding auxiliary at label i.

As an example, we shall consider Rosenbrock's modified Post Office Parcel Problem, viz.:

To maximize $f = x_1x_2x_3$ subject to the constraints

$$0 \leqslant x_1 \leqslant 20$$
$$0 \leqslant x_2 \leqslant 11$$
$$0 \leqslant x_3 \leqslant 42$$
$$0 \leqslant x_4 \leqslant 72, \text{ where } x_4 = x_1 + 2x_2 + 2x_3.$$

The solution is $x_1 = 20$, $x_2 = 11$, $x_3 = 15$, $x_4 = 72$, $f = 3{,}300$.

Thus we set $l_1 = 0$ $l_2 = 0$ $l_3 = 0$ $l_4 = 0$
$u_1 = 1$ $u_2 = 2$ $u_3 = 0$ $u_4 = 3$

and provide auxiliaries as follows

at label 101 to set $x_1 = 20$
at label 102 to set $x_2 = 11$
at label 104 to set $x_3 = 36 - 0{\cdot}5x_1 - x_2$.

The program found the exact optimum solution in 129 trials and then stopped itself.

The standard Post Office Parcel Problem, i.e. with the constraints $0 \leqslant x_1 \leqslant 42$ and $0 \leqslant x_2 \leqslant 42$, but otherwise as above, has the solution $f = 3{,}456$, $x_1 = 24$, $x_2 = 12$, $x_3 = 12$ and $x_4 = 72$, in which only the constraint on x_4 is effective. An auxiliary was written to determine x_3 from x_1 and x_2 when this upper constraint on x_4 became effective as before, and the solution $f = 3{,}456{\cdot}00$ was obtained in 133, 109 and 136 trials for three starting points. (As there is only one constraint effective, the program does not terminate, but continues searching in two dimensions.)

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

To solve problem A by this method, it is necessary to interchange x_1 and x_5, i.e. the old x_1 becomes x_5^* and the old x_5 becomes x_1^*, but both names will be used to avoid confusion.

The auxiliaries required are as follows:

(i) when $x_1^*(x_5)$ enters its upper boundary region, set it to the upper limit

(ii) when x_2 enters its upper boundary region, set it to the upper limit

(iii) when x_3 enters its upper boundary region, set it to the upper limit

(iv) when x_4 enters its upper boundary region, set it to the upper limit

(v) when x_8 enters its upper boundary region, use the upper limit to determine $x_5^*(x_1)$ from $x_1^*(x_5)$, x_2, x_3 and x_4.

The results were as follows:

(i) after 68 trials $x_5^*(x_1)$ was eliminated

(ii) after 117 trials $x_1^*(x_5)$ was eliminated

(iii) after 124 trials x_4 was eliminated

(iv) after 188 trials x_2 was eliminated

(v) after 220 trials x_3 was eliminated and $f = 5{,}280{,}334$ was obtained.

(Note that Rosenbrock's method is an inefficient 1-dimensional search procedure, requiring 32 trials to find the optimum.)

Table 3

Function value after 200 trials ($n = 2$, $\lambda = 100$, $\theta = 25$)

NUMBER OF VERTICES k	REFLECTION FACTOR α						PRODUCT
	1·0	1·1	1·2	1·3	1·4	1·5	
3	$4{\cdot}5 \times 10^{-12*}$	$2{\cdot}3 \times 10^{-12}$	1·6	$1{\cdot}1 \times 10^{-3}$	$2{\cdot}7 \times 10^{-5*}$	$2{\cdot}8 \times 10^{-10*}$	$\sim 10^{-40}$
4	$1{\cdot}2 \times 10^{-13}$	$1{\cdot}1 \times 10^{-10}$	$3{\cdot}0 \times 10^{-7}$	$1{\cdot}2 \times 10^{-2}$	$6{\cdot}7 \times 10^{-9*}$	$7{\cdot}6 \times 10^{-14}$	$\sim 10^{-53}$
5	$7{\cdot}4 \times 10^{-13}$	$1{\cdot}3 \times 10^{-8}$	$1{\cdot}8 \times 10^{-4}$	$2{\cdot}0 \times 10^{-1}$	$1{\cdot}5 \times 10^{-9}$	$1{\cdot}4 \times 10^{-7}$	$\sim 10^{-40}$
6	$3{\cdot}6 \times 10^{-7}$	$4{\cdot}4 \times 10^{-6}$	$3{\cdot}0 \times 10^{-3}$	$3{\cdot}0 \times 10^{-1}$	31	$7{\cdot}2 \times 10^{-9}$	$\sim 10^{-22}$
7	$3{\cdot}0 \times 10^{-8}$	$7{\cdot}6 \times 10^{-5}$	$3{\cdot}7 \times 10^{-2}$	1·0	$1{\cdot}0 \times 10^{-5}$	$6{\cdot}3 \times 10^{-6}$	$\sim 10^{-24}$
8	$1{\cdot}7 \times 10^{-6}$	$3{\cdot}1 \times 10^{-3}$	$4{\cdot}1 \times 10^{-2}$	1·7	5·6	$2{\cdot}6 \times 10^{-5}$	$\sim 10^{-14}$
Product	$\sim 10^{-57}$	$\sim 10^{-42}$	$\sim 10^{-16}$	$\sim 10^{-6}$	$\sim 10^{-25}$	$\sim 10^{-48}$	

Table 4

Function value after 200 trials ($n = 5$, $\lambda = 100$, $\theta = 25$)

NUMBER OF VERTICES k	REFLECTION FACTOR α						PRODUCT
	1·0	1·1	1·2	1·3	1·4	1·5	
6	1·6	2·0	1·8	4·8	18	11	$\sim 10^{3}$
7	2·1	1·3	2·2	5·2	$3{\cdot}2 \times 10^{-1}$	2·0	$\sim 10^{1}$
8	$3{\cdot}3 \times 10^{-1}$	$3{\cdot}1 \times 10^{-1}$	$7{\cdot}0 \times 10^{-1}$	6·3	$1{\cdot}5 \times 10^{-1}$	$2{\cdot}5 \times 10^{-1}$	$\sim 10^{-2}$
9	$4{\cdot}0 \times 10^{-1}$	$4{\cdot}9 \times 10^{-1}$	1·0	7·4	23	3·3	$\sim 10^{2}$
10	$1{\cdot}6 \times 10^{-1}$	$3{\cdot}0 \times 10^{-1}$	$5{\cdot}4 \times 10^{-1}$	5·0	4·1	$1{\cdot}3 \times 10^{-1}$	$\sim 10^{-2}$
11	$9{\cdot}4 \times 10^{-2}$	$9{\cdot}4 \times 10^{-2}$	$4{\cdot}3 \times 10^{-1}$	6·5	3·3	$2{\cdot}6 \times 10^{-1}$	$\sim 10^{-2}$
12	$3{\cdot}6 \times 10^{-2}$	$7{\cdot}2 \times 10^{-2}$	$8{\cdot}1 \times 10^{-1}$	10	6·9	$6{\cdot}8 \times 10^{-2}$	$\sim 10^{-3}$
Product	$\sim 10^{-4}$	$\sim 10^{-4}$	$\sim 10^{-1}$	$\sim 10^{5}$	$\sim 10^{3}$	$\sim 10^{-2}$	

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

9. A comparison of the performances of Rosenbrock's method and the Simplex and Complex methods with unconstrained optimization problems

As the Complex method had proved to be useful in solving constraint-bound optimization problems, it was decided to examine its effectiveness when applied to unconstrained problems. It is necessary to give ranges for each variable in order to set up the initial configuration as before, but a parameter is set to ensure that no testing of the constraints is carried out. In order to examine the variation of effectiveness of each method with n, the number of dimensions, a family of n-dimensional test functions $f(x_1, \ldots x_n; n, \lambda, \theta)$ was constructed as follows:

Let $\boldsymbol{O}_j$ be a rotation matrix such that

$o_{j,j} = \cos\theta, \quad o_{j+1,j+1} = \cos\theta$
$o_{j,j+1} = -\sin\theta, \quad o_{j+1,j} = \sin\theta$
$o_{i,i} = 1$ for $i \neq j$ or $j+1$
with all other elements zero.

e.g. $\boldsymbol{O}_1 = \begin{bmatrix} \cos\theta & -\sin\theta & & & & \\ \sin\theta & \cos\theta & & & & \\ & & 1 & & (\text{zeros}) & \\ & & & 1 & & \\ & (\text{zeros}) & & & \ddots & \\ & & & & & 1 \end{bmatrix}$

Then form the matrix $\boldsymbol{O} = \boldsymbol{O}_{n-1}\boldsymbol{O}_{n-2} \ldots \boldsymbol{O}_2\boldsymbol{O}_1$.

Let $\boldsymbol{\Lambda}$ be the diagonal matrix $(\lambda_1, \ldots \lambda_n)$

where $\lambda_1 = 1, \dfrac{\lambda_i}{\lambda_{i+1}} = \lambda^{\frac{1}{n-1}}$ for $i = 1, 2, \ldots n-1$

i.e. $\dfrac{\lambda_i}{\lambda_n} = \lambda$

Then the test function is taken to be

$f = \boldsymbol{x}'\boldsymbol{O}'\boldsymbol{\Lambda}\boldsymbol{O}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A} = \boldsymbol{O}'\boldsymbol{\Lambda}\boldsymbol{O}$,
$\boldsymbol{x}' = (x_1, \ldots x_n)$, and $'$ denotes transposition.

Thus the test function is a homogeneous quadratic form in $x_1, x_2 \ldots x_n$ with a minimum $f = 0$ at $x_i = 0$, $i = 1, 2, \ldots, n$, having a ratio of largest to smallest principal axes of $\sqrt{\lambda}$, and in which the directions of the principal axes are obtained by $(n-1)$ rotations through an angle θ.

Note that it would be inappropriate to test optimization methods possessing quadratic convergence on this test function.

All 3 programs were applied to minimizing this test function with $\theta = 25°$, $n = 2, 3, 5, 10$ and $\lambda = 1, 10, 100$ and using two starting points $A = (20, 0, 0, \ldots 0)$
$B = (0, 0, \ldots 20)$.

The steplength used by the Simplex program was chosen so that initially the rate of progress (i.e. distance moved by the centroid per reflection) was the same as that of Rosenbrock's method. For the Complex method, the ranges of each variable for setting up the initial complex were $(-21, 21)$. It was considered that for unconstrained problems, the values of α and k previously selected might be inappropriate. Accordingly, experiments were carried out on 2- and 5-dimensional members of the above family of test functions, with the results given in **Tables 3** and **4**. As before those entries marked with an asterisk indicate runs terminated by the stopping criterion in under 200 trials.

After examination of these tables, it was decided to take $\alpha = 1{\cdot}0$ for reflection factor, and $k = 2n$ for the number of vertices, as standard values for unconstrained problems of the type under consideration. Note that $\alpha = 1{\cdot}0$ means the complex cannot expand at all, which, whilst appropriate to the present case in which the initial configuration straddles the optimum, would clearly be inappropriate for an initial configuration remote from the optimum.

The results of this comparison are given in **Tables 5–8,** where R = Rosenbrock's method, S = Simplex and C = Complex, and A and B are the two starting points. For each value of λ the number of trials necessary to reduce the function value to the indicated level is recorded.

For $n = 5$, the Simplex method with $\lambda = 1$ had failed to obtain a function value as small as 0·01 after over 5,000 trials. This failure appeared to depend only on the starting point chosen, since with starting points $C = (9, 9, 9, 9, 9)$ and $D = (17, 8, 5, 4, \sqrt{6})$, which are approximately as far from the origin, the results given at the bottom of **Table 7** were obtained. The zero entries given for the contour $f = 10$ in **Tables 7** and **8** indicate that starting point B lies inside this contour when $\lambda = 100$.

Table 5

Comparison in 2 dimensions

		FUNCTION VALUE			
		10	1	0·1	0·01
$\lambda = 1$	SA	25	53	70	87
	SB	25	53	70	87
	RA	13	23	33	34
	RB	14	24	44	50
	CA	16	23	37	50
	CB	19	30	40	54
$\lambda = 10$	SA	32	48	60	75
	SB	28	49	63	78
	RA	14	24	38	45
	RB	19	37	43	50
	CA	18	31	39	49
	CB	11	20	30	44
$\lambda = 100$	SA	23	36	46	70
	SB	17	34	42	60
	RA	33	55	61	71
	RB	16	41	50	56
	CA	14	26	101	134
	CB	13	23	35	36

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

Table 6

Comparison in 3 dimensions

		FUNCTION VALUE			
		10	1	0·1	0·01
$\lambda = 1$	S*A*	42	65	84	109
	S*B*	42	65	84	109
	R*A*	19	34	49	50
	R*B*	21	36	74	78
	C*A*	25	36	69	88
	C*B*	31	54	73	86
$\lambda = 10$	S*A*	32	57	74	96
	S*B*	31	47	77	97
	R*A*	21	35	48	74
	R*B*	21	44	69	70
	C*A*	20	49	97	115
	C*B*	20	44	67	82
$\lambda = 100$	S*A*	32	42	71	93
	S*B*	4	54	78	99
	R*A*	20	50	78	118
	R*B*	5	23	83	121
	C*A*	14	39	106	242
	C*B*	17	48	69	77

Table 7

Comparison in 5 dimensions

		FUNCTION VALUE			
		10	1	0·1	0·01
$\lambda = 1$	S*A*	763	796	829	—
	S*B*	770	799	839	875
	R*A*	31	56	81	82
	R*B*	35	60	130	135
	C*A*	56	84	119	153
	C*B*	68	111	146	184
$\lambda = 10$	S*A*	40	117	151	190
	S*B*	37	72	111	153
	R*A*	38	58	84	115
	R*B*	35	59	75	111
	C*A*	48	83	123	183
	C*B*	48	90	177	234
$\lambda = 100$	S*A*	38	66	114	202
	S*B*	0	70	109	145
	R*A*	33	57	85	170
	R*B*	0	35	87	112
	C*A*	39	75	121	193
	C*B*	0	78	168	211
$\lambda = 1$	S*C*	110	151	191	230
	S*D*	111	146	180	231

The averages of all the numbers of function evaluations given in **Tables 5–8** are given in **Table 9** for each method and each value of n, the five dimensional Simplex results S*C* and S*D* being used in place of S*A* and S*B*. The Complex method had performed particularly poorly in ten dimensions, since it had used too many vertices, namely 20. (It will be recalled that $k = 2n$ was selected on the basis of experiments in 2 and 5 dimensions.) The results of **Table 9** can be conveniently summarized as follows, where T is the average number of trials:

Rosenbrock's method: $T = 16{\cdot}1 + 11{\cdot}3n$

Simplex: $T = -7{\cdot}6 + 26{\cdot}5n$

Complex: $T = -14{\cdot}5 + 26{\cdot}2n$

where, in fitting these straight lines (which are quite reasonable fits), the ten-dimensional datum for the Complex method has been omitted.

From this comparison it was concluded that for unconstrained problems, Rosenbrock's method is more efficient than the Simplex and Complex methods, which do not differ significantly in their performance. It was also found that the number of trials needed to locate an optimum to a given precision increased about twice as rapidly with the number of dimensions n for the Simplex and Complex methods as it did with Rosenbrock's method.

10. Conclusions

For Rosenbrock's program, the modifications described have been shown not to change its performance significantly. Also, the selection of the parameters α and k in the Complex method does not appear to be critical. Therefore the author doubts whether any modification of either of these methods or the original Simplex method could be of very great advantage, although Nelder and Mead (Nelder and Mead, 1965) do seem to have developed a successful Simplex-type method. The view of other workers, that random methods and "one variable at a time" methods without rotation of the axes are inefficient, is endorsed.

The Complex and RAVE programs are believed to be of considerable use in the solution of constraint-bound problems, but for unconstrained problems Rosenbrock's method is superior to both the Simplex and Complex methods by a factor of two. The interesting problem of modifying the RAVE program to incorporate criteria to release it from the constraints on to which it has locked would be spurred on by the provision of a problem requiring this facility.

This work has led the author to the view that there are two problems in optimization which must be distinguished:

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

Table 8

Comparison in 10 dimensions

		FUNCTION VALUE			
		10	1	0·1	0·01
$\lambda = 1$	S*A*	181	272	357	438
	S*B*	181	272	357	438
	R*A*	61	111	161	162
	R*B*	70	170	215	275
	C*A*	154	266	373	459
	C*B*	197	312	397	531
$\lambda = 10$	S*A*	130	214	299	378
	S*B*	76	148	226	293
	R*A*	61	112	162	203
	R*B*	70	71	151	172
	C*A*	127	210	354	497
	C*B*	123	286	572	670
$\lambda = 100$	S*A*	117	253	422	507
	S*B*	0	124	224	297
	R*A*	73	113	142	214
	R*B*	0	70	110	150
	C*A*	115	249	572	1264
	C*B*	0	235	833	1207

Table 9

Overall summary of comparison

NUMBER OF DIMENSIONS, *n*	ROSENBROCK	SIMPLEX	COMPLEX
2	37	51	37
3	52	66	65
5	73	124	116
10	129	258	418

(i) efficient unconstrained optimization,

(ii) continuation of the search when one or more constraints become effective,

and at the moment most workers are concerning themselves with the former problem.

In conclusion the author would like to mention that preliminary results indicate that, for unconstrained problems, gradient methods are more efficient than direct-search methods. It is hoped that a comparison of several of the optimization techniques which have recently emerged will form the basis of a future paper.

Appendix

Values of the constants used in problem A

The values of the constants a_i were

$a_0 = 9$ $a_1 = 15$ $a_2 = 50$ $a_3 = 9{\cdot}583$
$a_4 = 20$ $a_5 = 15$ $a_6 = 6$ $a_7 = 0{\cdot}75$.

The values of the coefficients k_i were

$k_1 = -145{,}421{\cdot}402$ $k_2 = 2{,}931{\cdot}1506$ $k_3 = -40{\cdot}427932$ $k_4 = 5{,}106{\cdot}192$
$k_5 = 15{,}711{\cdot}36$ $k_6 = -161{,}622{\cdot}577$ $k_7 = 4{,}176{\cdot}15328$ $k_8 = 2{\cdot}8260078$
$k_9 = 9{,}200{\cdot}476$ $k_{10} = 13{,}160{\cdot}295$ $k_{11} = -21{,}686{\cdot}9194$ $k_{12} = 123{\cdot}56928$
$k_{13} = -21{\cdot}1188894$ $k_{14} = 706{\cdot}834$ $k_{15} = 2{,}898{\cdot}573$ $k_{16} = 28{,}298{\cdot}388$
$k_{17} = 60{\cdot}81096$ $k_{18} = 31{\cdot}242116$ $k_{19} = 329{\cdot}574$ $k_{20} = -2{,}882{\cdot}082$
$k_{21} = 74{,}095{\cdot}3845$ $k_{22} = -306{\cdot}262544$ $k_{23} = 16{\cdot}243649$ $k_{24} = -3{,}094{\cdot}252$
$k_{25} = -5{,}566{\cdot}2628$ $k_{26} = -26{,}237$ $k_{27} = 99$ $k_{28} = -0{\cdot}42$
$k_{29} = 1{,}300$ $k_{30} = 2{,}100$ $k_{31} = 925{,}548{\cdot}252$ $k_{32} = -61{,}968{\cdot}8432$
$k_{33} = 23{\cdot}3088196$ $k_{34} = -27{,}097{\cdot}648$ $k_{35} = -50{,}843{\cdot}766$.

The reader should not assume that the k_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 nine decimal places. The amount of arithmetic necessary to compute f, x_6, x_7, x_8 could be much reduced by eliminating the explicit computation of these intermediate quantities, as they are not needed in the optimization problem. It is, however, necessary to retain all the figures given in the k_i in order that the values of f calculated by the two programs agree to 7 or 8 significant figures.

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024

References

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

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

ROSEN, J. B. (1960). "The Gradient Projection Method for Nonlinear Programming. Part I. Linear Constraints," *J. Soc. Indust. Appl. Math.*, Vol. 8, p. 181.

ROSEN, J. B. (1961). "The Gradient Projection Method for Nonlinear Programming. Part II: Nonlinear Constraints," *J. Soc. Indust. Appl. Math.*, Vol. 9, p. 514.

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.

To the Editor,
The Computer Journal.

Estimation of the truncation error in Runge–Kutta and allied processes

Sir,

I should like to draw attention to a serious limitation of one of the methods suggested in the paper by R. E. Scraton, published in the October 1964 issue of this *Journal*.

In this paper is given the formula,

$$y_0 = \frac{17}{162}k_0 + \frac{81}{170}k_2 + \frac{32}{135}k_3 + \frac{250}{1377}k_4 + \frac{qr}{s} + O(h^6), \quad (1)$$

where q, r, and s are also certain *linear* combinations of the k's. It is held that the use of this formula reduces to five the number of function evaluations necessary to achieve agreement of all terms of up to and including order h^5.

Unfortunately, this is true only in the case of a single equation, say

$$\frac{dy}{dx} = f(x, y), \quad (2)$$

and does not apply to the system

$$\frac{dy_i}{dx} = f_i(y_j), \qquad i, j = 1, 2 \ldots n. \quad (3)$$

For in this general case, the coefficient of h^5 on the right of (1) may be written

$$\frac{1}{4320}\left\{ - f_{i.j}f_{j.klm}f_kf_lf_m - 3f_{i.j}f_{j.kl}f_{k.m}f_lf_m + 9f_{i.j}f_{j.k}f_{k.l}f_{l.m}f_m - f_{i.j}f_{j.k}f_{k.lm}f_lf_m \right.$$
$$\left. + \frac{[f_{i.jk}f_jf_kf_l + 3f_{i.jk}f_{j.l}f_kf_l - 9f_{i.j}f_{j.k}f_{k.l}f_l + f_{i.j}f_{j.kl}f_kf_l][f_{i.j}f_{j.k}f_k]}{f_{i.j}f_j} \right\}$$

in which the summation convention is used and, for example

$$f_{j.kl} \text{ denotes } \frac{\partial^2 f_j}{\partial y_k \partial y_l}.$$

Although in the case of a single equation the summations do not reduce to single terms, they do become factorizable and the whole expression vanishes; but for the system, the nature of the summations is such that this is not so generally.

It is perhaps desirable to call attention to another danger of considering only a single equation when deriving Runge–Kutta formulae. In the general expression for the Taylor expansion of a solution of (3) there are nine terms in the coefficient of h^5, (and to produce a valid formula the coefficients of these nine terms have to be equated to the corresponding terms occurring in the general form of the right-hand side of (1)). But for a single equation, two of these terms coalesce, and two equations to be separately satisfied are replaced by their sum. For the sixth-order terms another five coalescences occur. Thus, even when producing standard linear Runge–Kutta formulae of order higher than the fourth, there is a danger of failing to produce formulae of the intended order of accuracy if the analysis is restricted to the single equation.

Yours faithfully,

A. R. CURTIS.

Mathematics Division,
National Physical Laboratory,
Teddington, Middlesex.

15 January 1965.

Downloaded from https://academic.oup.com/comjnl/article/8/1/42/489862 by guest on 10 April 2024