Step size adjustment at discontinuities for fourth order Runge-Kutta methods

P. G. O'Regan

Department of Electrical Engineering, University College, Cork, Ireland

With discontinuities in the differential equations of an initial value problem it is in general necessary to alter the effective step size to evaluate the variables at the point of discontinuity. Since this latter point is often determined by the values of some or all of the dependent variables it would normally be observed by a checking procedure just prior to updating these dependent variables. A method of calculating the fraction of the interval to the point of discontinuity and of updating the values of the dependent variables to this point is given. The method which will be subsequently referred to as the Alpha Method is third order, and no new evaluation of the function is required. It compares favourably both for computational time and programming with the normal method of continuing the tabulation beyond the discontinuity and employing inverse interpolation.

(Received November 1969)

1. Mathematical discussion

Discontinuities can occur in the differential equations of initial value problems for a number of reasons. The closing or opening of a switch can lead to a completely new set of differential equations and if the position of this switch is a function of one or more of the dependent variables these discontinuities cannot be anticipated by any simple procedure. Saturation and backlash can also lead to discontinuities. When using Runge-Kutta formulae to integrate the differential equations a constant step size (h) would be employed until a checking procedure indicates that the discontinuity occurs in the present interval of integration. This checking procedure would normally occur at the end of the functional calculations but prior to the updating of the dependent variables. When a point of discontinuity occurs within an interval it is necessary to:

1. Determine the fraction α of the step h along the independent variable to the point of the discontinuity, and
2. Update the dependent variables to this point.

If the set of m simultaneous first order differential equations are

$$\frac{dy_i}{dx} = f_i(x, y_1, y_2 \ldots y_m) \qquad i = 1, 2, \ldots, m \tag{1}$$

and the condition for discontinuity is

$$\phi(x, y_1, y_2 \ldots y_m) = 0 \tag{2}$$

the problem is to find the intersection of equation (2) with the solution to equation (1). A single first order differential equation will be considered and, apart from error analysis, the extension to the system in equation (1) is then straightforward. Equations (1) and (2) with $m = 1$ and writing y and f in place of y_1 and f_1, respectively become

$$\frac{dy}{dx} = f(x, y) \tag{3}$$

and

$$\phi(x, y) = 0 \tag{4}$$

The Taylor series for $y(x_n + \alpha h)$ can be written in the form

$$y(x_n + \alpha h) = y(x) + \alpha hf + \frac{(\alpha h)^2}{2!} Df + \frac{(\alpha h)^3}{3!}(D^2 f + f_y Df) + \frac{(\alpha h)^4}{4!}(D^3 f + f_y D^2 f + f_y^2 Df + 3Df Df_y)|_n + 0(\alpha h)^5 \tag{5}$$

where

$$\left.\begin{aligned} f &= f(x, y) \\ D &= \frac{\partial}{\partial x} + f_n \frac{\partial}{\partial y} \\ f_n &= f(x_n, y_n) \\ y_n &= y(x_n) \end{aligned}\right\} \tag{6}$$

and $|_n$ indicates evaluation at the point (x_n, y_n).

When the Generalised Fourth Order Runge-Kutta method is applied to (3) the relevant equations are

$$y_{n+1} = y_n + \sum_{i=1}^{4} W_i K_i + 0(h^5) \tag{7}$$

where the W_i's are constants and

$$K_i = hf(x_n + \alpha_i h, y_n + \sum_{j=1}^{i-1} \beta_{ij} K_j) \tag{8}$$

The parameter $\alpha_1 = 0$ and the other parameters are selected so that:

1. Equations (7) and (5) with $\alpha = 1$ match up to and including terms in h^4.
2. The parameters are simple with many of them zero as in the Classical Runge Method or a bound of the error term is minimised as in Ralstons Method or some other desirable feature is introduced.

It will be assumed that this selection has been carried out. Now suppose that the intersection between (3) and (4) occurs between x_n and x_{n+1} and furthermore that a trial y_{n+1} must be computed before the test for detection of the intersection can be applied. (Normally it would be sufficient to compare the signs of $\phi(x_n, y_n)$ and $\phi(x_{n+1}, y_{n+1})$.) Hence the K_i's have been evaluated and (7) must be replaced by

$$y(x_n + \alpha h) = y_n + \sum_{i=1}^{4} A_i K_i + 0(h^4) \tag{9}$$

where the A_i's are new parameters which are functions of α(Lim $A_i = W_i$) so that (9) and (5) match as accurately
(Lim Ai = Wi)
$\alpha > 1$
as possible for all values of α. Since there are four free A_i parameters the match can be accurate up to and including terms in h^3 (hence error term of $0(h^4)$) and the equations for the A_i's then are:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & \alpha_2 & \alpha_3 & \alpha_4 \\ 0 & \alpha_2^2 & \alpha_3^2 & \alpha_4^2 \\ 0 & 0 & \alpha_2\beta_{32} & \alpha_2\beta_{42}+\alpha_3\beta_{43} \end{bmatrix} \begin{vmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{vmatrix} = \begin{vmatrix} \alpha \\ \alpha^2/2 \\ \alpha^3/3 \\ \alpha^3/6 \end{vmatrix} \tag{10}$$

These equations when inverted can be written in the form

$$\begin{vmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{vmatrix} = \begin{bmatrix} 1 & C_{12} & C_{13} \\ 0 & C_{22} & C_{23} \\ 0 & C_{32} & C_{33} \\ 0 & C_{42} & C_{43} \end{bmatrix} \begin{vmatrix} \alpha \\ \alpha^2 \\ \alpha^3 \end{vmatrix} \tag{11}$$

where with

$$D = \alpha_3(\alpha_3 - \alpha_2)(\alpha_2\beta_{42} + \alpha_3\beta_{43}) - \alpha_2\beta_{32}\alpha_4(\alpha_4 - \alpha_2) \tag{12}$$

the C_{ij} coefficients can be evaluated in the following sequence

$$\begin{aligned} C_{42} &= (\alpha_2^2\beta_{32})/2D \\ C_{43} &= \{\alpha_3(\alpha_3 - \alpha_2) - 2\alpha_2\beta_{32}\}/6D \\ C_{32} &= -\frac{(\alpha_2\beta_{42} + \alpha_3\beta_{43})}{2D}\alpha_2 \\ C_{33} &= \frac{2(\alpha_2\beta_{42} + \alpha_3\beta_{43}) - \alpha_4(\alpha_4 - \alpha_2)}{6D} \\ C_{22} &= +1/\alpha_2(\tfrac{1}{2} - \alpha_3 C_{32} - \alpha_4 C_{42}) \\ C_{23} &= -1/\alpha_2(\alpha_3 C_{33} + \alpha_4 C_{43}) \\ C_{12} &= -(C_{22} + C_{32} + C_{42}) \\ C_{13} &= -(C_{23} + C_{33} + C_{43}) \end{aligned} \tag{13}$$

As an example in the case of the Classical Runge Method ($\alpha_2 = \alpha_3 = \beta_{21} = \beta_{32} = 1/2$, $\alpha_4 = \beta_{43} = 1$, all other $\beta_{ij} = 0$) the Coefficient Matrix in (11) becomes:

$$\begin{bmatrix} 1 & -3/2 & 2/3 \\ 0 & 1 & -2/3 \\ 0 & 1 & -2/3 \\ 0 & -1/2 & 2/3 \end{bmatrix} \tag{14}$$

while for Ralstons Method (see Ralston, 1962; Ralston, 1965) it is:

$$\begin{bmatrix} 1 & -1{\cdot}42715746 & 0{\cdot}60191774 \\ 0 & -2{\cdot}31621878 & 1{\cdot}76473812 \\ 0 & 4{\cdot}25693059 & -3{\cdot}05139499 \\ 0 & -0{\cdot}51355434 & 0{\cdot}68473912 \end{bmatrix} \tag{15}$$

Equation (9) can now be written in the form

$$y(x_n + \alpha h) = y_n + \sum_{i=1}^{4} K_i \sum_{j=1}^{3} C_{ij}\alpha^j + 0(\alpha h^4) \tag{16}$$

if $C_{11} = 1$ and $C_{21} = C_{31} = C_{41} = 0$. Equation (16) can in turn be written in the form:

$$\begin{aligned} y(x_n + \alpha h) = y_n &+ \alpha K_1 \\ &+ \alpha^2 \sum_{i=1}^{4} K_i C_{i2} + \alpha^3 \sum_{i=1}^{4} K_i C_{i3} + \epsilon_4 \end{aligned} \tag{17}$$

where ϵ_4 is the error term containing $(\alpha h)^4$ and higher terms. Clearly since (17) and (5) match for terms up to and including terms in α^3 the coefficients of the powers of α must be identical and so

$$K_1 = 0(h)$$

$$\sum_{i=1}^{4} K_i C_{i2} = 0(h^2)$$

and

$$\sum_{i=1}^{4} K_i C_{i3} = 0(h^3) \tag{18}$$

Assuming $K_1 \neq 0$ it is possible to rewrite (17) in the form

$$\alpha = A - B\alpha^2 - C\alpha^3 - \epsilon_4/K_1 \tag{19}$$

where

$$\begin{aligned} A &= \frac{y(x_n + \alpha h) - y_n}{K_1} = 0(1) \\ B &= \frac{1}{K_1}\sum_{i=1}^{4} K_i C_{i2} \quad = 0(h) \\ C &= \frac{1}{K_1}\sum_{i=1}^{1} K_i C_{i3} \quad = 0(h^2) \end{aligned} \tag{20}$$

and *neglecting the* $-\epsilon_4/K_1$ *term* a solution to (19) can be written in the form

$$\begin{aligned} \alpha = A - BA^2 + (2B^2 - C)A^3 - 5A^4B(B^2 - C) \\ + A^5(14B^4 - 21B^2C - 3C^2) + 0(h^5) \end{aligned} \tag{21}$$

where the terms in this solution are sequentially 0(1), $0(h)$, $0(h^2)$, . . . This solution is explicit only if $y(x_n + \alpha h)$ is independent of α, i.e. if the discontinuity occurs at

$$y = y_d \tag{22}$$

where y_d is a constant whereupon

$$A = \frac{y_d - y_n}{K_1}$$

Also equation (21) cannot be applied if $K_1 = 0$ and as will be illustrated in the example in Section 3 it suffers from the serious disadvantage that it is not always rapidly convergent. Hence it is recommended that (17) be solved by an iterative procedure and if Newton's

method is employed the iteration equation is

$$\alpha_{j+1} = \alpha_j - \frac{y_n - y(x_n + \alpha_j h) + K_1\alpha_j + P\alpha_j^2 + Q\alpha_j^3}{-hm(\alpha_j) + K_1 + 2P\alpha_j + 3Q\alpha_j^2} \tag{23}$$

where

$$P = \sum_{i=1}^{4} K_i C_{i2}$$

$$Q = \sum_{i=1}^{4} K_i C_{i3} \tag{24}$$

$$m(\alpha_j) = \frac{dy}{dx}\bigg|_{\substack{\phi(x,y)=0 \\ X=X_n+\alpha_j h}} = -\frac{\partial\phi/\partial x}{\partial\phi/\partial y}\bigg|_{\substack{X=X_n+\alpha_j h \\ y=y(X_n+\alpha_j h)}} \tag{25}$$

and $y(x_n + \alpha_j h)$ is obtained by solving

$$\phi\{x_n + \alpha h, y(x_n + \alpha h)\} = 0 \tag{26}$$

with $\alpha = \alpha_j$. The initial value α_0 for α can be obtained from the relation

$$\alpha_0 = \phi(x_n, y_n)/\{\phi(x_n, y_n) - \phi(x_{n+1}, y_{n+1})\} \tag{27}$$

When the iterative procedure in (23) converges to α_e (i.e. $\lim_{j \Rightarrow \infty} \alpha_j = \alpha_e$) then because the error ϵ_4 in (17) is neglected there is an error ϵ_α in the estimate of α. If the correct value of α is α_t then

$$\epsilon_\alpha = \alpha_t - \alpha_e \tag{28}$$

Since α_e satisfies (17) with ϵ_4 neglected

$$y(x_n + \alpha_e h) = y_n + \alpha_e K_1 + \alpha_e^2 P + \alpha_e^3 Q \tag{29}$$

and since α_t satisfies (17)

$$y(x_n + \alpha_t h) = y_n + \alpha_t K_1 + \alpha_t^2 P + \alpha_t^3 Q + \epsilon_4 \tag{30}$$

Subtracting (29) from (30) using (25), (28) and the Mean Value Theorem and rearranging yields

$$\epsilon_\alpha = \alpha_t - \alpha_e =$$

$$\frac{\epsilon_4}{hm(\eta) - \{K_1 + (\alpha_t + \alpha_e)P + (\alpha_t^2 + \alpha_t\alpha_e + \alpha_e^2)Q\}} \tag{31}$$

where η lies between α_t and α_e. Neglecting terms of $0(h^2)$ in the denominator of (31) then gives

$$\epsilon_\alpha = \frac{\epsilon_4/h}{m(\eta) - f_n} \tag{32}$$

which shows that the error in α is normally $0(h^3)$ but can get large if $\phi(x, y) = 0$ has the same slope as the differential equation in the neighbourhood of the intersection point. The error ϵ_x in the x coordinate of the intersection point is

$$\epsilon_x = h\epsilon_\alpha \tag{33}$$

and if (26) is employed to find y then the resulting error in the y coordinate is

$$\epsilon_y = h\epsilon_\alpha m(\eta) = \frac{m(\eta)}{m(\eta) - f_n}\,\epsilon_4 \tag{34}$$

If (17) is employed to find y then the expression for ϵ_y is again that of (34). It should perhaps be pointed out that in a system of equations such as (1), if some of the variables had no influence on the point of intersection then these variables would have to be evaluated using (17).

With the aid of Ralston (1965) it can be shown that (neglecting terms in h^5 and higher order terms)

$$\begin{aligned}\frac{4!\epsilon_4}{h^4} &= [\alpha^4 - 4(\alpha_2^3 A_2 + \alpha_3^3 A_3 + \alpha_4^3 A_4)]\, D^3 f \\ &+ [\alpha^4 - 12\{\alpha_2^2\beta_{32}A_3 + (\alpha_2^2\beta_{42} + \alpha_3^2\beta_{43})A_4\}] f_y D^2 f \\ &+ [3\alpha^4 - 24\{\alpha_2\alpha_3\beta_{32}A_3 + (\alpha_2\beta_{42} + \alpha_3\beta_{43})\alpha_4 A_4\}] Df Df_y \\ &+ [\alpha^4 - 24\alpha_2\beta_{32}\beta_{43}A_4] f_y^2 Df\end{aligned} \tag{35}$$

Since the A_i quantities in (35) can by means of (11) be replaced by third degree polynomials in α, ϵ_4 is clearly a function of α and the derivatives of f. A bound can be obtained on ϵ_4 as given in (35) when in the usual manner M and L are introduced by the inequalities

$$|f(x, y)| < M$$

and

$$\left|\frac{\partial^{i+j} f}{\partial x^i \partial y^j}\right| < \frac{L^{i+j}}{M^{j-1}}, \qquad i + j \leqslant 4 \tag{36}$$

in a region R about the point (x, y) containing all points in equation (8) thus making $D^3 f$, $f_y D^2 f$, $Df Df_y$, $f_y^2 Df$ respectively less than 8, 4, 4 and 2 times ML^3. This bound does not simplify very much unless the numerical values of the α_i and β_{ij} quantities are inserted in (35). When this is done in the Classical Runge case the result is (again neglecting terms in h^5 and higher order terms):

$$\epsilon_{4\ \text{Runge}} < \frac{1}{4!}(18\alpha^4 - 44\alpha^3 + 26\alpha^2)ML^3h^4, \qquad 0 < \alpha < 1 \tag{37}$$

and thus

$$\operatorname*{Max}_{0<\alpha<1}(\epsilon_{4\ \text{Runge}}) < 0{\cdot}0917\, ML^3h^4 \tag{38}$$

the maximum occurring at $\alpha = 0{\cdot}573$.

The corresponding equation in the Ralston case is

$$\operatorname*{Max}_{0<\alpha<1}(\epsilon_{4\ \text{Ralston}}) < 0{\cdot}0937\, ML^3h^4 \tag{39}$$

which occurs at $\alpha = 0{\cdot}539$.

It should perhaps be stressed that (37) is a *Bound* on the error term for the Classical Runge case and (38) gives the Maximum over the range of α of this bound. These figures are accordingly normally much greater than the true error. Likewise the error bound in equation (39) is pessimistic.

2. Comparison with interpolation method

The error in estimation of the intersection of (3) and (4) (and also for the intersection of (1) and (2)) is $0(h^4)$ and hence in the same order as a four point interpolation scheme. Since the nature of the discontinuities may be such that only one point is available prior to the intersection—and in any case space and time are lost in storing previous points—it is not unreasonable to assume that for a comparable interpolation method the tabulation should be extended three points beyond the intersection point. This requires eight extra evaluations of the function in (3) or (1) and this in itself could involve

considerable calculation. In addition the remaining Runge-Kutta procedure would have to be executed twice and this then followed by the interpolation routine which would have to be applied to the independent and each dependent variable. By contrast the alpha method does not require any evaluation of the function in (3) or (1), only one execution of that part of the Runge-Kutta procedure involved in the updating of the variables is required and the calculation of α would hardly be as extensive or involved as the interpolation routine. The procedure in the case of the alpha method merely consists of a simple algorithm to evaluate α and then using (11) the A_i quantities. Thus the alpha method has a decided advantage when the function evaluation is time consuming and apart from this it is equally if not more simple and straightforward from a programming point of view than the interpolation method.

3. Example

The differential equation selected is

$$\frac{dy}{dx} = xy^{1/3} \tag{40}$$

with $x_n = y_n = 1$. The solution is

$$y = \left(\frac{x^2 + 2}{3}\right)^{3/2} \tag{41}$$

The Classical Runge method was employed. The procedure adopted consists of first selecting a value of α which will be called α_t since it is the true value of α. The true value of y (called y_t) is then calculated using (41), i.e.

$$y_t = \left[\frac{(1 + \alpha_t h)^2 + 2}{3}\right]^{3/2} \tag{42}$$

In the example it is assumed that

$$\phi(x, y) = y - y_t = 0$$

so that $m = 0$ and $y(x_n + \alpha h)$ can be replaced in (20) or (23) by y_t as calculated in (42). The estimates of α calculated using (23) and (21) can then be compared with α_t. In **Table 1** the estimate of α obtained using (23) is marked α_n and α_{III}, α_{IV} and α_V are the estimates of α obtained using respectively the first three, four and five terms in (21). Clearly the results indicate that Newton's iteration formula (23) is far more accurate than (21) especially at the higher α values. The maximum errors occurred in the neighbourhood of $\alpha = 0{\cdot}6$ as expected. Since M and L are only slightly greater than unity the error bound on $\epsilon_{4\ \text{Runge}}$ is approximately $0{\cdot}1h^4$ giving a bound of $0{\cdot}1h^3$ in ϵ_α. The actual error is only about 1/50th of this figure which illustrates the usual conservative nature of Runge-Kutta error bounds. Clearly from **Table 2** for fixed α_t the error $\alpha_t - \alpha_n$ is proportional to h^3 except in the neighbourhood of $\alpha_t = 1$ where the error is proportional to h^4. The results were computed using a 31 bit binary mantissa (9 plus decimal digits).

4. Conclusions

The method is an alternative to the normal inverse interpolation scheme with a decided advantage over the latter when the evaluation of the function is time consuming. It gives automatic switching from one set of differential equations to another and a sequence of switches in one basic integration interval can be sequentially handled. From a programming point of view the method presents no difficulties.

Table 1

$h = 0{\cdot}1$

α_t	$10^6(\alpha_t-\alpha_n)$	$10^6(\alpha_t-\alpha_{III})$	$10^6(\alpha_t-\alpha_{IV})$	$10^6(\alpha_t-\alpha_V)$
0	0	0	0	0
0·1	−0·200	−0·298	−0·198	−0·200
0·2	−0·643	−2·24	−0·597	−0·640
0·3	−1·15	−9·32	−0·793	−1·13
0·4	−1·58	−27·7	−0·034	−1·52
0·5	−1·86	−66·4	2·96	−1·72
0·6	−1·91	−137·4	10·3	−1·75
0·7	−1·72	−256·8	25·1	−1·78
0·8	−1·34	−440·0	52·0	−2·26
0·9	−0·792	−711·9	97·0	−3·99
1·0	−0·219	−1097·0	168·3	−8·35

Table 2

α_t	$10^6(\alpha_t - \alpha_n)$	
	$h = 0{\cdot}1$	$h = 0{\cdot}2$
0	0	0
0·1	−0·200	−1·58
0·2	−0·643	−5·10
0·3	−1·15	−9·07
0·4	−1·58	−12·4
0·5	−1·86	−14·5
0·6	−1·91	−14·9
0·7	−1·72	−13·7
0·8	−1·34	−11·1
0·9	−0·792	−7·40
1·0	−0·219	−3·24

References

SCHEID, F. (1968). *Theory and Problems of Numerical Analysis*, Schaum's Outline Series, McGraw Hill Book Company, pp. 202–204.

RALSTON, A. (1962). Runge-Kutta Methods with Minimum Error Bounds, *Math. Comput.*, Vol. 16, pp. 431–437.

RALSTON, A. (1965). *A First Course in Numerical Analysis*, McGraw Hill Book Company, pp. 191–201.