SOME REMARKS ON THE NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

by

M.V. Wilkes
Director, Mathematical Laboratory
University of Cambridge

One method of solving the differential equation $\frac{dy}{dx} = f(x,y)$ makes use of the formula :

$$y_1 - y_0 = h\left(f_0 + \frac{1}{2} \nabla f_o + \frac{5}{12} \nabla^2 f_o + \frac{3}{8} \nabla^3 f_o + \ldots\right) \quad (1)$$

where ∇ denotes a backward difference. A more refined method uses (1) as a predictor and the following formula as a corrector:

$$y_1 - y_0 = h\left(f_1 - \frac{1}{2} \nabla f_1 - \frac{1}{12} \nabla^2 f_1 - \frac{1}{24} \nabla^3 f_1 - \ldots\right) \quad (2)$$

In practice the formulae are terminated after a finite number of terms, and I shall assume that the term after the last one taken into account is negligible; in other words, I shall assume that the interval h is always such that truncation error is unimportant. If ∇^4 is neglected, (1) and (2) may be written in Lagrangian form

$$y_1 - y_0 = \frac{h}{24}\left(55 f_o - 59 f_{-1} + 37 f_{-2} - 9 f_{-3}\right) \quad (3)$$

$$y_1 - y_0 = \frac{h}{24}\left(9 f_1 + 19 f_0 - 5 f_{-1} + f_{-2}\right) \quad (4)$$

Books on numerical analysis discuss the way in which errors present in the solution are propagated. I propose to investigate the way in which the maximum error in y at the end of one step depends on the errors in the values of y from which it is obtained.

Suppose it is known that the errors in $y_{-3}, y_{-2}, y_{-1}, y_o$ expressed in units of the last decimal place retained, are all less in magnitude than ϵ. I shall assume that x is given exactly at each step, that the computation of f is performed to a sufficient number of decimal places to ensure that the error in f depends only on the error in y:

i.e. the error in f $= \frac{\partial f}{\partial y}$ x error in y.

Denote by $\frac{1}{M}$ the upper bound of $\frac{\partial f}{\partial y}$ in the total interval of x considered.

Then error in f $\leqslant \epsilon / M$

Consider first the case in which (3) is used alone, and let η_o denote the upper bound of the error in the computed value of y; then

$$\eta_o = \epsilon + \frac{h}{24}(55 + 59 + 37 + 9)\frac{\epsilon}{M} + \frac{1}{2}$$

Note that $\eta_o > \epsilon + \frac{1}{2}$, however small h/M.

If it is required that $\eta_o \leqslant S\epsilon$, where S is some number greater than unity, then we must have

$$\frac{h}{M} \leqslant \frac{3}{20} \; \frac{(S - 1)\epsilon - \frac{1}{2}}{\epsilon}$$

If $\epsilon >> \frac{1}{2}$, i.e. if the rounding-off error in y is small compared with the total error, this reduces to

$$\frac{h}{M} \leqslant \frac{3}{20}(S - 1)$$

If (3) is used as a predictor, followed by one application of (4) as a correction, the maximum error, which can occur in y_1 is given by

$$\eta_1 = \epsilon + \frac{h}{24} \left[\frac{9\eta_o}{M} + (19 + 5 + 1)\frac{\epsilon}{M} \right] + \frac{1}{2}$$

$$= \epsilon + \frac{1}{2} + \frac{h}{M} \; \frac{1}{24} \; (9\eta_o + 25\epsilon)$$

It may be verified that the combination of errors in y_{-3}, y_{-2}, y_{-1}, y_o, which are most unfavourable as regards error in the value of y_1 given by (3) are also most unfavourable as regards the error in the value given by (4).

If the corrector is applied more than once and the maximum error after n applications is denoted by η_n, it is easily seen that

$$\eta_2 = \epsilon + \frac{1}{2} + \frac{h}{M} \; \frac{1}{24} \; (9\eta_1 + 25\epsilon)$$

$\eta_3 = \epsilon + \frac{1}{2} + \frac{h}{M} \; \frac{1}{24} \; (9\eta_2 + 55\epsilon)$, etc., and it may be shown that

$$\eta_n = \left(\frac{3}{8}\,\frac{h}{M}\right)^n \eta_o + \left(\epsilon + \frac{1}{2} + \frac{25}{24}\;\frac{h}{m}\,\epsilon\right) \frac{1 - \left(\frac{3}{8}\,\frac{h}{m}\right)^n}{1 - \frac{3}{8}\,\frac{h}{M}}$$

Note that the iterative process converges if $\frac{h}{M} < \frac{8}{3}$, but that the error when convergence has taken place can be large if $\frac{h}{M}$ is nearly equal to 8/3.

The following table gives the largest permissable values of h/M for various values of n and for two values of S, where $\eta_n < S\epsilon$; it is assumed that $\epsilon >> \frac{1}{2}$. Since η_n is an upper bound to the error, if S = 5/4 the error introduced in one step is likely the very small; if S = 2, the chance of a serious error occurring is higher. The table also gives the number of times a value of f must be calculated in order and cover a total interval in x equal to M.

Figures for the 4th order Runge Kutta process are also given for comparison.

TABLE

No. of iterations	Max. permissible value of h/M		No. of calculations of f per interval M	
	S = 5/4	S = 2	S = 5/4	S = 2
0	.025	.1	40	10
1	.14	.41	14.3	5.0
2	.16	.52	18.8	5.8
∞	.17	.56		
Runge-Kutta	.22	.69	18.2	5.8

When a solution is required to high accuracy, the maximum value of h/M which can be used is more likely to be determined by the truncation error in (3) and (4) than by the considerations discussed here. In low accuracy work, however, particularly in situations in which a digital machine is used instead of an analogue machine such as a simulator, the limits given in the above table may be relevant. In these circumstances it will be seen that the minimum number of computations of a value of f (which is a measure of the amount of work required) per interval M in x is obtained by applying the corrector once only. It is interesting to observe that the Runge-Kutta method is only slightly inferior, as regards the number of computations of f required, to the use of the predictor with one application of the corrector. It has already been pointed out that the fact that convergence takes place after sufficient iterations affords no guarantee that the error is small.

DISCUSSION

Mr. H.L.F. Orde, National Cash Register Co.

When working out the possible error, you assumed everything worked against you. However, would not this mean that the function oscillated violently and thus contradicted one of the other assumptions made, namely that the derivatives never became higher than 1/M and that therefore a function which satisfied all the assumptions made, would have considerably smaller upper bound to the error term than that derived?

Dr. M.V. Wilkes (In Reply)

Yes. If one knew something about the function satisfying the equation it is just possible that one could obtain a better estimate of the upper bound of the error. This is the very worst that could happen.

Mr. H.L.F. Orde, National Cash Register Co.

But could it in fact happen?

Dr. M.V. Wilkes (In Reply)

No, but circumstances might still be very bad and the upper bound is probably the safest estimate.

Professor T.M. Cherry, University of Melbourne.

I recall a method I saw some time ago of the predictor corrector form. The method was somewhat as follows:

First: Use a standard type of predictor formula.

Second: Assuming the error in y is small in the equation $\frac{dy}{dx} = f(x,y)$, for correction we need $f(x,y)$ for a different y, i.e. compute $f(x,y+h)$ and if h is small enough this is given by $f\ (xy) + h\ (\frac{\partial f}{\partial y}) f\ xy$.

That is we use predictor to find y approximately and if we go through the corrector formula with h as an unknown we express the condition that the value of y, shall agree with the value found from the integration and we get an equation which determines the correct value of h.

In the problem which I was investigating, I was integrating along a ridge with a steep drop on each side. By using this method it was possible to take fairly long steps without striking trouble.

Dr. M.V. Wilkes (In Reply)

You are really solving an algebraic equation instead of iterating.

Professor T.M. Cherry, University of Melbourne.

Going to the limit in one step in fact.

However, I wondered if there was any experience as to just what value this method might have.

Dr. M.V. Wilkes (In Reply)

It works better for linear equations than non-linear equations.

Professor T.M. Cherry, University of Melbourne.

In the context where I was using it, it was non-linear and we had the algebraic formula for $\frac{\partial f}{\partial y}$ and calculated it at the same time as $f(x,y)$. Roughly speaking the calculation of the derivative was equivalent to 1.5 calculations of the function but the power obtained from the method was equal to calculating the function 4 or 5 times if one was iterating.

Mr. B.Z. de Ferranti, International Business Machines.

As regards digital simulation. Do you think the limits of engineering accuracy will lead to the design of machines using fewer bits?

Dr. M.V. Wilkes (In Reply)

At least one machine being built now has a basic word length of 40 bits, with arrangements to divide a word into four units each of ten bits so that calculations of ten bit accuracy can be carried out. In fact four additions of such an accuracy can be carried out simultaneously.

Mr. B.Z. de Ferranti, International Business Machines.

Do you think this is a general trend?

Dr. M.V. Wilkes (In Reply)

For machines being used in this field it may well be. Not that such a machine lends itself to general mathematical calculations, but possibly for any kind of simulation. This particular machine is the TX2.

Mr. B.Z. de Ferranti, International Business Machines.

The MIT Whirlwind is also a short word length machine.

Dr. M.V. Wilkes (In Reply)

I think a machine with a very short word length would be very tiresome to programme, since it is very useful to have a few spare bits.

Dr. J.M. Bennett, University of Sydney.

The figures in Dr. Wilkes table indicate that in his opinion there is not much in it, one can use a suitable predictor-corrector with one iteration or the Runge-Kutta in the body of the integration of a d.e. (apart from starting values) and the error will probably be about the same.

Dr. M.V. Wilkes (In Reply)

Yes! And the amount of labour too.

If you were limited by this sort of consideration this would be true. I think one can argue more generally that one application of the corrector is all that is worthwhile. Only in exceptional cases would it be worthwhile programming more than one application of the corrector.

Mr. R.H. Merson, Royal Aircraft Establishment.

The difficulty with the predictor-corrector is that it will not go round corners! To go round a corner one has to cut down the interval of integration and this is difficult with predictor corrector formulae.

With the Runge-Kutta this is easy and I have solved the Bessel equation, which has a singularity at $z = 0$, with a Runge-Kutta process by cutting down the interval from 2^{-1} sec well away from the origin to 2^{-24} sec for 3 or 4 steps near the origin.

This couldn't be done using the predictor-corrector!

Dr. S. Gill, Ferranti Ltd.

While we are on the subject of the pros and cons of being able to alter step length I might mention that if you are in a situation where you wish to find the point at which one of the dependent variables takes a particular value, then being able to alter step length is a great advantage, because the step can be adjusted using a process similar to the Newton-Raphson for finding the zero of a function.