NUMERICAL METHODS FOR THE SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

by

ARTHUR CHRISTOPHER ROLLS NEWBERY

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in the Department of

MATHEMATICS

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September, 1958.

ABSTRACT

Families of three- and four-point corrector formulae are derived, which differ from standard formulae in that they express y_n in terms of more than one previously computed ordinate. It is shown that the standard formulae are special cases of the more general formulae derived here. By theoretical argument and by numerical experiments it is shown that the standard formulae are often inferior to others which are developed in this thesis.

The three-point family, with its associated truncation error, is given in (7) and (9) of Chapter 2 on page 12. The four-point family is given in (41) on page 25.

With the help of Rutishauser's method each family is examined for stability. In the four-point case a procedure is described, whereby the magnitude of the coefficient in the error term can be minimized subject to the restriction that the formula shall remain stable. Also a theorem is proved, which states that no stable four-point formula can have a truncation error of degree higher than fifth in the step-size h.

I hereby certify that this abstract_is satisfactory.

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representative. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics

The University of British Columbia,
Vancouver 8, Canada.

Date Sept 17th 1958

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The author wishes to acknowledge his indebtedness to Dr. T. E. Hull for encouragement and advice received throughout the investigation which forms the topic of this thesis.

Thanks are also due to the Computing Centre of the University of British Columbia for the many hours of machine time, without which the practical verification of results would necessarily have been far more limited in scope.

CHAPTER ONE. History of Numerical Methods for Solving Differential Equations

The earliest numerical methods for solving ordinary differential equations date back almost as far as the calculus itself. The fields of astronomy and physics were among the first in which the calculus found practical application, and both these fields gave rise to ordinary differential equations for which no analytic solutions could be found. Examples of such differential equations are those arising from the computation of eccentric anomalies of planetary orbits, and those encountered in connection with simple and compound pendulums. The latter problems gained particular importance in Newton's time, since maritime explorers and traders required accurate chronometers for astronavigation.

The simplest and crudest method for solving first-order initial-value problems is Euler's. Using the standard notation as used, for example, in [1], the problem is this:

Given a differential equation, $y' = f(x,y)$ and an initial point (x_0,y_0), calculate y_n, i.e. find the value of y when $x = x_0 + nh$, where h is a small quantity.

Euler's solution is
$$\begin{aligned} \bar{y}_1 &= y_0 + hf(x_0,y_0) \\ \bar{y}_2 &= \bar{y}_1 + hf(x_1,\bar{y}_1) \\ &\text{etc.,} \end{aligned}$$

where the bars denote approximations. Now suppose the differential equation possessed an analytic solution, $y = F(x)$, then it would be possible to express y_1 in the form of a Taylor expansion thus:

$$y_1 = F(x_0 + h) = F(x_0) + hF'(x_0) + \frac{h^2}{2!}F''(z)$$

where $x_0 \leqslant z \leqslant x_1$, and since $F'(x_0)$ is the derivative of the function at (x_0,y_0), we may write

$$y_1 = y_0 + hf(x_0,y_0) + \frac{h^2}{2!}F''(z).$$

The value of $\bar{y}_1$ obtained by Euler's method is what one would obtain by truncating the Taylor series after the second term. It would be exact if F(x) were linear, because the second and higher derivatives would then be zero. The error, therefore, of Euler's predictor formula is $O(h^2)$. Since h is taken to be a small quantity, attempts were made to produce formulae with truncation errors of higher degree in h, and it was found that various existing numerical quadrature formulae could easily be adapted for the purpose. For example the trapezoidal quadrture formula,

$$\int_{x_0}^{x_1} f(x)dx = F(x_1) - F(x_0) \simeq \frac{x_1-x_0}{2}\left[f(x_0) + f(x_1)\right],$$

where $f(x)$ is the derivative of $F(x)$, became the trapezoidal corrector formula,

$$\bar{y}_1 - y_0 = \frac{h}{2}(y_0' + \bar{y}_1') .$$

Thus, nothing more than a simple change of notation was necessary in order to convert this quadrature formula into a formula for solving differential equations numerically. The above formula could be used iteratively to correct the values predicted by Euler's formula. This is discussed in [1, p.26]. The truncation error for this formula is $O(h^3)$. In order to obtain corrector formulae with truncation errors of higher degree than third, some of the more elaborate Newton-Cotes quadrature formulae were adapted. The best known case is that of Simpson's rule, viz.

$$\int_{x_0}^{x_2} f(x)dx = F(x_2) - F(x_0) \simeq \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right] .$$

After changing the notation it reads:

$$\bar{y}_2 - y_0 = \frac{h}{3}(y_0' + 4y_1' + \bar{y}_2').$$

The truncation error is now $O(h^5)$, but this improvement was bought at a high price, because, before attempting to satisfy the above equation by iteration, it is necessary to know y_0' and y_1' . Without this information Simpson's rule cannot be used. Moreover, as will be explained later, this formula is conditionally unstable.

After this stage was reached in the middle of the eighteenth century, no further progress was made until 1883 when the Adams method was published. Adams' formulae are of interest historically for two reasons: firstly they are the earliest of the more elaborate formulae devised specifically for the solution of differential equations; earlier mathematicians had been content to adapt quadrature formulae for the purpose. The second point of interest about the Adams formulae is that they are stable. In Adams' day it is doubtful whether even the concept of stability of formulae existed; yet in 1952 when Rutishauser published the first comprehensive theory on the subject, he showed that the Adams formulae were not only stable but they possessed optimum stability properties. Although Adams clearly recognized the fact that a good quadrature formula does not necessarily make a good corrector formula, he did not completely sever the long-established bond between the two types of formula. His corrector formulae can easily be transformed into quadrature formulae by applying in reverse the notational changes mentioned above. However, the quadrature formulae so derived might have little application, since they would involve evaluation of the integrand at points outside the range of integration.

It was not until 1895 that a complete break with quadrature formulae was made. In that year Runge published his celebrated method [4], and since then numerous variants on his method have also been published. The relative merits of the so-called Runge-Kutta procedures and the predictor-corrector methods have been widely discussed e.g. [2, pp. 247,248]. With the advent of digital computers the former methods have gained in popularity at the expense of the latter. Nevertheless this thesis is devoted to a study of predictor-corrector methods and to an investigation of means by which they may be made more competitive. The view taken is this: A highly accurate procedure must inevitably be complicated; complexity increases more rapidly with accuracy in the case of Runge-Kutta methods than in the case of predictor-corrector methods; the former class of methods will reach 'saturation point' before the latter; therefore the latter should not be neglected in favour of the former.

The concept of stability is of fundamental importance to the study of numerical methods for solving differential equations. It may be defined verbally thus: A formula is stable if it is insensitive to small errors in the data to which it is applied. Since frequent reference will be made to Rutishauser's paper on the subject [3], the relevant parts of his argument are reproduced here for convenience.

<u>Rutishauser's stability analysis</u>. Let the differential equation be $y' = f(x,y)$, and let $\bar{y}, \bar{y}'$ be approximations to the true values of y, y', so that $\bar{y} = y + s$ and $\bar{y}' = y' + s'$.

Now $\bar{y}' = f(x,\bar{y}) = f(x,y+s) \simeq f(x,y) + sf_y(x,y) = y' + sf_y(x,y)$.

(1) Hence $s' \simeq sf_y(x,y)$.

If Simpson's rule is used, then

$$\bar{y}_{k+1} = \bar{y}_{k-1} + \frac{h}{3}[\bar{y}'_{k+1} + 4\bar{y}'_k + \bar{y}'_{k-1}], \quad \text{or}$$

$$y_{k+1} + s_{k+1} = y_{k-1} + s_{k-1} + \frac{h}{3}[y'_{k+1} + s'_{k+1} + 4(y'_k + s'_k) +$$

$$y'_{k-1} + s'_{k-1}].$$

This yields the difference equation for s:

$$s_{k+1}[1 - \frac{h}{3}f_{y,k+1}] - \frac{4h}{3}f_{y,k}s_k - [1 + \frac{h}{3}f_{y,k-1}]s_{k-1} = 0.$$

If we make the simplifying assumption that f_y is constant, we obtain the result $s_k = p^k$, where p is the root of a quadratic equation whose roots are

$$p_1 = 1 + hf_y + \frac{h^2}{2}f_y^2 + \ldots \simeq e^{hf_y}$$

$$p_2 = -1 + \frac{h}{3}f_y - \frac{h^2}{18}f_y^2 \quad \ldots \simeq -e^{-\frac{1}{3}(hf_y)}.$$

It may be seen that p_1 approximates to a solution of (1), for by writing $kh = x$ and $s_k = p_1^k$ we have

$$s_k = p_1^k \simeq e^{hf_y k} = e^{xf_y} \text{ , therefore}$$

$$s_k' \simeq f_y e^{xf_y} = f_y s_k \, .$$

The other root, p_2 is parasitic, and when f_y is negative the magnitude of p_2 is greater than one; hence in this case it causes an exponentially increasing oscillation. Moreover this situation cannot be remedied by reducing the size of h.

Next the Adams four-point corrector is examined,

$$y_{k+1} = y_k = \frac{h}{24}[9y'_{k+1} + 19y'_k - 5y'_{k-1} + y'_{k-2}] + 0(h^5).$$

Proceeding as before, the difference equation for s is obtained:

$$[1 - \frac{3h}{8} f_{y,k+1}] s_{k+1} - [1 + \frac{19h}{24} f_{y,k}] s_k + \frac{5h}{24} f_{y,k-1} s_{k-1} -$$

$$\frac{h}{24} f_{y,k-2} s_{k-2} = 0.$$

On writing $s_k = p^k$ as before, and taking f_y as constant, we obtain a cubic equation in p with two parasitic roots. This case differs from the three-point case just discussed, in that the parasitic roots tend to zero with h. When $h = 0$ the cubic equation in p reduces to $p^3 - p^2 = 0$, and therefore the parasitic roots vanish with h. Hence Rutishauser concludes that the Adams method is stable for sufficiently small h.

In the following chapter a study of three- and four-point corrector formulae is made. In each case a new family of formulae has been developed, and experimental results are given to aid assessment of the relative merits of the new formulae and the established ones. The restriction to corrector formulae is justified by the fact that in practice it is these formulae rather than the predictors which determine the value of a method. In the theoretical part, attention is focussed upon single first-order initial value problems. In practice the extension to simultaneous and higher-order differential equations is quite simple, though the stability analysis is harder. No assumptions concerning linearity of the equations are made, but it is assumed that the solution function has continuous derivatives of fourth and fifth order in the three- and four-point cases respectively.

CHAPTER TWO. Three- and Four-Point Corrector Formulae

A study of predictor-corrector procedures for solving ordinary differential equations reveals the fact that all the formulae in common use give an expression for the value of a new ordinate in terms of one previously computed ordinate and a linear combination of previously computed derivatives. A collection of such formulae may be found in [1, pp. 48,49]. This fact suggests the following questions:

(i) Is it possible to derive a generalized n-point corrector formula of which all known n-point corrector formulae are special cases, and if so what would be the procedure for derivation?

(ii) Given such a generalized formula could we carry out a stability- and error analysis on it in such a way that this analysis would also apply to the special cases and enable us to assess their various merits?

(iii) Why is it that the established formulae only make use of one known ordinate? Would it not be possible, by using all the available ordinates, to produce formulae which are superior to those in common use?

The above questions will be investigated for the cases $n = 3$ and $n = 4$. Attention is concentrated on corrector formulae only, because it is upon the corrector formula that the stability and accuracy of a given procedure depend.

We will start with the derivation of a generalized three-point corrector formula. Let the equation to be solved be $y' = f(x,y)$, and let its solution be $y = F(x)$. The first two rows,

$$\begin{array}{ccc} x_0 & y_0 & y'_0 \\ x_1 & y_1 & y'_1, \end{array}$$

are given. The entries for the third row have been estimated by means of a predictor formula in conjunction with the given differential equation. Denote these values by $\bar{y}_2$, $\bar{y}'_2$. The generalized three-point corrector formula is required to give a corrected value for y_2 (denoted by y_{2c}) in terms of y_0, y_1, y'_0, y'_1 and $\bar{y}'_2$.

Since the final formula may involve three ordinates, it is no longer adequate to adapt a quadrature formula. Instead we shall use a method of undetermined coefficients, and the formulae so derived will include as special cases the results obtained by standard procedures. Let the required formula be of the form

$$(1) \qquad y_{2c} = a_0y_0 + a_1y_1 + h[b_0y'_0 + b_1y'_1 + b_2\bar{y}'_2] ,$$

where the a's and b's are coefficients to be determined in such a way that when y_2, y'_2 are substituted for y_{2c} and $\bar{y}'_2$ in (1), the resulting formula shall have the highest order of accuracy that is consistent with stability. The 'order' of accuracy is defined in the usual way thus: If (1) is exact when $y = F(x)$ is any polynomial of degree $\leqslant n$, but not when $F(x)$ is some polynomial of degree $n+1$, then the accuracy of (1) is of n^{th} order.

In order to determine the a's and b's we express (1) in terms of the shift operator E and the differential operator D, so that

$EF(x_0) = F(x_0 + h) = F(x_1) = y_1$ etc. (See [2], Chap. V). We obtain

$$(2) \qquad E^2 y_0 = [a_0 + a_1 E + hD(b_0 + b_1 E + b_2 E^2)]\, y_0 \, .$$

Now $E = e^{hD}$ [2, p.134]. Write $hD = u$, $E = e^u$, and substitute into (2). On cancelling the y_0 and proceeding formally, we obtain

$$(3) \qquad e^{2u} = a_0 + a_1 e^u + u(b_0 + b_1 e^u + b_2 e^{2u}) \, .$$

Now expand each side of (3) as a power series in u, and equate corresponding powers of u. We obtain an infinite set of simultaneous linear equations in the a's and b's.

$$(4) \qquad \begin{array}{llllll} a_0 + a_1 & & & & = & 1 \\ a_1 & + b_0 & + b_1 & + b_2 & = & 2 \\ \frac{a_1}{2!} & & + b_1 & + 2b_2 & = & \frac{2^2}{2!} \\ \frac{a_1}{3!} & & + \frac{b_1}{2!} & + \frac{2^2 b_2}{2!} & = & \frac{2^3}{3!} \\ \frac{a_1}{4!} & & + \frac{b_1}{3!} & + \frac{2^3 b_2}{3!} & = & \frac{2^4}{4!} \end{array}$$

etc.

Since we only have five degrees of freedom, the five coefficients a_0, a_1, b_0, b_1, b_2, we cannot hope to satisfy more than five equations of (4). The n^{th} equation of (4) is derived by equating coefficients of u^{n-1} in (3). If this equation is not satisfied, therefore, we shall have an error term involving u^{n-1} i.e. $h^{n-1}D^{n-1}$. In general we want the error term to have a high degree in h, so if we decide to satisfy p of the equations (4) we shall always choose the first p of the equations. It will be shown that if the coefficients are chosen so as to satisfy the first five equations of (4), then the corrector formula so derived turns out to be Simpson's rule. However Rutishauser [3] pointsout that Simpson's rule is unstable under certain conditions, and therefore it may be asked whether it is possible, by sacrificing the fifth equation, to derive a stable corrector formula from the first four equations of (4). Since we now have only four constraints, but still have five degrees of freedom, an infinite number of solutions is possible, and these can be expressed parametrically.

Using the second, third and fourth of the equations (4), we can obtain all the b's in terms of a_1 . The first equation of (4) then determines a_0 in terms of a_1 . Hence a_1 may be regarded as a parameter which, once a value is assigned to it, determines the values of a_0, b_0, b_1 and b_2 . The solutions of these equations are

$$\begin{aligned} a_0 &= 1 - a_1 , \\ b_0 &= \frac{1}{12}(4 - 5a_1) , \\ b_1 &= \frac{8}{12}(2 - a_1) , \\ b_2 &= \frac{1}{12}(4 + a_1) . \end{aligned} \tag{5}$$

On substituting these values into (1) we obtain

$$y_2 = (1 - a_1)y_0 + a_1 y_1 + \frac{h}{12}[y_0'(4 - 5a_1) + 8y_1'(2 - a_1) + y_2'(4+a_1)], \tag{6}$$

or

$$(7) \qquad y_2 = (1-a_1)y_0 + a_1y_1 + \frac{h}{12}[4y'_0 + 16y'_1 + 4y'_2 - a_1(5y'_0+8y'_1-y'_2)] .$$

Equation (7) represents, in terms of a single parameter a_1, all possible three-point corrector formulae with truncation error of fourth or higher degree in h. In particular, when $a_1 = 0$, (7) reduces to Simpson's rule, and when $a_1 = 1$ we have Adams' three-point formula. For other values of a_1 we have a family of new formulae whose properties are to be investigated. Before proceding to an empirical investigation of these properties, two questions must be raised. First: What is the truncation error associated with (7) in terms of a_1? Second: For what range of values of a_1 is formula (7) stable?

Truncation Error. In deriving formula (7) we ensured that the first four equations of (4) were satisfied identically. We must therefore examine the fifth equation in order to obtain the truncation error. Let the truncation error be R, and add R to the R.H.S. of (7). The fifth equation of (4), when written out in full becomes

$$(8) \qquad \frac{a_1u^4}{4!} + \frac{b_1u^4}{3!} + \frac{2^3b_2u^4}{3!} + R = \frac{2^4u^4}{4!} ,$$

$$\therefore \; R = \frac{u^4}{4!}[2^4 - a_1 - 4b_1 - 4 \cdot 2^3b_2] .$$

On substituting for b_1, b_2 from (5) we obtain

$$(9) \qquad R = \frac{-a_1u^4}{4!} = \frac{-a_1h^4D^4F(z)}{4!} \quad \text{where } x_0 \leqslant z \leqslant x_2 .$$

If we choose $a_1 = 0$, then the fourth order error is zero. This is to be expected, since (7) has now become Simpson's rule, and the truncation error for this formula is known to be of fifth degree in h.

Stability. Rutishauser's criterion for stability, when applied to (7) requires that the parasitic root of the equation $p^2 - a_1p - (1 - a_1) = 0$ should be less than one in absolute value. The parasitic root is $1 - a_1$. We therefore require

(10) $0 < a_1 < 2.$

When $a_1 = 1$, as is the case with Adams' method, the parasitic root is zero, thus we have optimum stability conditions. From (9) and (10) it will be seen that by taking a_1 such that $0 < a_1 < 1$, it is possible to produce stable formulae with less truncation error than that associated with Adams' method.

Several tests were carried out in order to establish the properties of various members of the three-point family. As might be expected, the results confirm that Simpson's rule, when it is stable, is the best formula of the family. However, when Simpson's rule is not stable, it may be far inferior to the other formulae. This is well exemplified by the equation $y' = -2xy^2$, $y(1) = \frac{1}{3}$, whose solution is $y = \frac{1}{x^2+2}$. This equation was solved on the Alwac III-E computer at the University of British Columbia by each of the six three-point formulae corresponding to six equally spaced values of a_1 between 0 and 1 inclusive. At regularly spaced points on the abscissa the errors were computed and output. A step-size $h = \frac{3}{32}$ was taken; 31 binary digits after the point were carried, and x_0 was taken at $\frac{13}{16}$. From the results shown in Table 1 the instability of Simpson's rule is at once apparent. The errors arising from computation by this method were in fact alternating in sign, but this does not show in the table, because the spacing of the outputs shown is an even multiple of the step-size.

x	a_1	10^9E
3.8125	1.	7
	.8	-12
	.6	-20
	.4	-24
	.2	-24
	0.	-66
15.4375	1.	-13
	.8	-7
	.6	-7
	.4	-5
	.2	-3
	0.	-238
27.0625	1.	-4
	.8	-4
	.6	-2
	.4	-4
	.2	-3
	0.	-500
38.6875	1.	-1
	.8	0
	.6	0
	.4	0
	.2	-3
	0.	-815
50.3125	1.	6
	.8	6
	.6	5
	.4	5
	.2	5
	0.	-1131

Table 1 showing the errors corresponding to each of six values of a_1, at five equally spaced points on the abscissa. The equation is $y' = -2xy^2$.

The results shown in Table 1 strongly suggest that most values of the parameter a_1 yield better results than those obtained by Simpson's rule ($a_1 = 0$). The question now arises: Which value of a_1 should be chosen in a particular case? Both theoretical investigations and some preliminary experimental results suggest that values of a_1 close to zero are best, i.e. one should choose the smallest value of a_1 which yields a stable formula. A joint paper by T. E. Hull and the author is planned, in which this subject is treated in more detail.

The four-point case. The procedure adopted for finding the general expression (7) for all possible three-point corrector formulae, in terms of a single parameter a_1, may be extended to the four-point case. However, for reasons to be explained, it is now desirable to use two parameters.

Let the four-point corrector be of the form

$$(11) \qquad y_{3c} = a_0y_0 + a_1y_1 + a_2y_2 + h[b_0y'_0+b_1y'_1+ b_2y'_2+b_3\bar{y}'_3] .$$

On rewriting (11) in terms of the operators E and D we obtain

$$(12) \qquad E^2y_0 = (a_0+a_1E+a_2E^2)y_0 + hD[b_0+b_1E+b_2E^2+b_3E^3]y \quad .$$

Now write $hD = u$, $E=e^{hD} = e^u$ in (12), and drop the y_0 . We obtain

$$(13) \qquad e^{3u} = a_0 + a_1e^u + a_2e^{2u} + u[b_0+b_1e^u+b_2e^{2u}+b_3e^{3u}] .$$

After expressing each side of (13) as an infinite power series in u, and equating the coefficients of successive powers of u on the L.H.S. and R.H.S., we again have an infinite set of simultaneous equations in the a's and b's:

$$(14)\quad \left\{\begin{array}{l} a_0 + a_1 + a_2 = 1 \\[1ex] a_1 + 2a_2 + b_0 + b_1 + b_2 + b_3 = 3 \\[1ex] \dfrac{a_1}{2!} + \dfrac{2^2 a_2}{2!} + b_1 + 2b_2 + 3b_3 = \dfrac{3^2}{2!} \\[2ex] \dfrac{a_1}{3!} + \dfrac{2^3 a_2}{3!} + \dfrac{b_1}{2!} + \dfrac{2^2 b_2}{2!} + \dfrac{3^3 b_3}{2!} = \dfrac{3^3}{3!} \\[2ex] \qquad\qquad \text{etc.} \\[1ex] \text{The Nth row, for } N \geqslant 3 \text{ is} \\[1ex] \dfrac{a_1}{(N-1)!} + \dfrac{2^{N-1} a_2}{(N-1)!} + \dfrac{b_1}{(N-2)!} + \dfrac{2^{N-2} b_2}{(N-2)!} + \dfrac{3^{N-2} b_3}{(N-2)!} = \dfrac{3^{N-1}}{(N-1)!} \end{array}\right.$$

The first five equations of (14) enable us to express a_1 and the b's in terms of two parameters a_0 and a_2 thus:

$$(15)\quad \left\{\begin{array}{ll} \multicolumn{2}{c}{a_1 = 1 - a_0 - a_2\ ,} \\[1ex] b_0 = \frac{1}{24}[9a_0 + a_2]\ , & b_1 = \frac{1}{24}[8 + 19a_0 - 13a_2]\ , \\[1ex] b_2 = \frac{1}{24}[32 - 5a_0 - 13a_2]\ , & b_3 = \frac{1}{24}[8 + a_0 + a_2]\ . \end{array}\right.$$

When a_1 and the b's are so chosen, the first five of equations (14) are identically satisfied, regardless of how a_0 and a_2 are chosen. The question may now be raised: Is it possible to choose a_0, a_2 in such a way that

(a) The first six equations of (14) are satisfied, and

(b) The resulting corrector formula is stable?

The sixth equation of (14) is

$$(16) \qquad \frac{a_1}{5!} + \frac{2^5 a_2}{5!} + \frac{b_1}{4!} + \frac{2^4 b_2}{4!} + \frac{3^4 b_3}{4!} = \frac{3^5}{5!}$$

On simplifying and writing a_1 and the b's in terms of a_0, a_2 with the help of (15), this reduces to

$$(17) \qquad 19a_0 + 11a_2 + 8 = 0$$

To test for stability we have to determine the upper and lower bounds of the parasitic roots of the equation

$$(18) \qquad p^3 - a_2p^2 - a_1p - a_0 = 0 \, .$$

If we divide (18) by p-1, the remainder is $1-a_1-a_2-a_0$, which by (15) is zero. On removing the factor (p-1), from (18) we are left with a quadratic equation whose roots are the parasitic roots of (18). This quadratic is

$$(19) \qquad p^2 + (1 - a_2)p + a_0 = 0.$$

Eliminating a_2 between (17) and (19) we have

$$(20) \qquad p^2 + \frac{19}{11}(1 + a_0)p + a_0 = 0 \, ,$$

$$(21) \qquad \therefore p = \frac{1}{22}\left[-19(1 + a_0) \pm \sqrt{19^2(1 + a_0)^2 - 4 \times 11^2 a_0}\right] .$$

The expression under the root sign is positive, therefore we have real roots. Denote these by p_1, p_2 .

From (20) $\quad p_1 + p_2 = \frac{-19}{11}(1 + a_0)$ and $p_1p_2 = a_0$,

$$\therefore p_1^2 + p_2^2 = \frac{19^2}{11^2}(1 + a_0)^2 - 2a_0 \, .$$

The equation whose roots are p_1^2, p_2^2 is

$$x^2 - x\left[\frac{19^2}{11^2}(1 + a_0)^2 - 2a_0\right] + a_0^2 = 0.$$

The equation whose roots are $p_1^2 - 1$, $p_2^2 - 1$ is

$$(x + 1) - (x + 1)\left[\frac{19^2}{11^2}(1 + a_0)^2 - 2a_0\right] + a_0^2 = 0,$$

$$\text{(22)} \qquad \text{i.e.} \quad x^2 + x\left[2 + 2a_0 - \frac{19^2}{11^2}(1 + a_0)\right] + 1 + a_0^2 + 2a_0 - \frac{19^2}{11^2}(1 + a_0)^2 = 0.$$

Now Rutishauser showed that a sufficient condition for stability was $p_1, p_2 < 1$, and a necessary condition was $p_1, p_2 \leqslant 1$. He also showed that in some cases the condition $p_1, p_2 < 1$ was necessary as well as sufficient. Since we do not know the nature of the equations on which our formulae may be used, we shall regard the condition $p_1, p_2 < 1$ as being necessary and sufficient for stability. It is therefore necessary that $p_1^2 - 1$ and $p_2^2 - 1$ shall both be negative, hence the constant term of (22) must be positive. Now this constant term is $(1 - a_0)^2(1 - \frac{19^2}{11^2})$.

Clearly this expression cannot be positive, therefore our question is answered: It is not possible to derive a stable four-point corrector formula satisfying the first six of equations (14). Now the sixth equation of (14) was obtained by equating the coefficients of u^5 in (13). Since we have shown that this equation cannot be satisfied by a stable formula, and since $u = hD$, it follows that a stable four-point formula must have a truncation error of degree five or less in h. This result may be expressed as a

THEOREM. A stable four-point corrector formula cannot have a truncation error of degree higher than fifth in h.

It can now be seen why we chose to satisfy only five equations of (14) and to use two parameters, for if we had satisfied six equations and used only one parameter, then this single parameter could only have generated unstable formulae.

Next we wish to determine whether it is possible to derive a stable four-point formula with less truncation error than the Adams four-point formula. (In view of the above theorem we cannot hope to obtain a truncation error of higher degree in h, but we may be able to reduce the numerical coefficient of h^5). For this purpose we require an expression for the error term in terms of a_0, a_2. Let the error term R be added to the L.H.S. of (16), and rewrite the whole equation in terms of a_0, a_2, using (15). Then

$$(23) \qquad R = -\frac{19a_0 + 11a_2 + 8}{6!} u^5 .$$

In order to investigate stability we must return to (19) and consider separately the cases of real and complex roots. It will be helpful to plot a graph on which each point represents a parameter-pair. We shall then be able to determine which areas on the graph contain points which generate stable formulae. The graph is given on page 27.

Case I. p_1 and p_2 are real. From (19) the condition for real roots is

$$(24) \qquad (1 - a_2)^2 \geqslant 4a_0 .$$

The equation whose roots are $p_1^2 - 1$, $p_2^2 - 1$ is

$$x^2 + x\,[2 + 2a_0 - (1 - a_2)^2] + (1 + a_0)^2 - (1 - a_2)^2 = 0.$$

For stability we require both roots of this equation to be negative, and for this it is necessary and sufficient that

(25) $(1 + a_0)^2 > (1 - a_2)^2$ and

(26) $2(1 + a_0) > (1 - a_2)^2$.

Any parameter-pair satisfying the inequalities (24) to (26) will generate a stable formula. For the purpose of graph-plotting it is convenient to make the transformation:

(27) $x = 1 + a_0$, $y = 1 - a_2$.

The above three inequalities then become

(28) $y^2 \geqslant 4(x - 1)$,

(29) $x^2 > y^2$,

(30) $2x > y^2$.

By substituting equality for inequality signs we can find the boundary of the area which contains points satisfying the three inequalities. From the graph it is seen that the required area is enclosed by the line-pair (29) and the parabola (28).

Case II . p_1 and p_2 are complex conjugate. For stability we require that the modulus shall be less than one. From (19) this condition reduces to $0 < a_0 < 1$, or after the transformation (27),

(31) $1 < x < 2.$

From the graph it is seen that the required area is enclosed by the line $x = 2$ and the parabola (28).

By applying the transformation (27) to equation (23) we can find the locus of points which generate formulae with zero fifth degree error, thus

$$19a_0 + 11a_2 + 8 = 0,$$ or, after transformation,

(32) $19x - 11y = 0$.

Since it is seen that this line does not pass through any area which contains stable points, we have a graphical verification of the theorem.

We can also plot the locus of points which generate formulae with the same error term as the Adams four-point formula. Adams chose $a_2 = 1$, $a_0 = 0$, therefore from (23) the error term for his formula is $\frac{-19}{6!} u^5$. The equation of this locus will be

$$19a_0 + 11a_2 + 8 = 19,$$ or, after transformation,

(33) $19x - 11y - 19 = 0$.

Adams' choice of parameters is represented on the graph by the point A(1,0). It can be seen from (19) that this choice makes both parasitic roots zero, therefore his formula has optimum stability properties.

One more question relating to four-point corrector formulae may be raised: Given that the modulus of the larger parasitic root equals c, where $0 \leqslant c < 1$, what is the smallest possible error term subject to this restriction?

Applying transformation (27) to (19) and (23), we have

$$(34) \qquad p^2 + yp + x - 1 = 0 , \qquad\qquad p = \frac{-y \pm \sqrt{y^2 - 4(x - 1)}}{2} .$$

$$(35) \qquad R = - \frac{19x - 11y}{6!} u^5 .$$

Consider the case of complex roots first. The modulus of the roots is $\sqrt{x - 1}$, $\therefore x - 1 = |c|^2$. This gives a straight line parallel to the y axis, and the condition for complex roots requires that we consider only that segment of the straight line which lies within the concave portion of the parabola (28). The question now reduces to the following: What point on the line segment lies closest to the line (32)? It can be seen from the graph that the required point would have to be indefinitely close to the point where the line segment intersects the parabola (28) and where y is non-negative. Having thus located the point, at least for practical purposes, the minimum error can be determined from (35).

Now consider the case of real roots. From (34) we have

$$\max \left| \frac{1}{2}\left[-y \pm \sqrt{y^2 - 4(x - 1)}\right] \right| = c, \text{ or equivalently}$$

$$(36) \qquad |y| + \sqrt{y^2 - 4(x - 1)} = 2c .$$

On rationalizing (36) we obtain

$$y^2 - 4(x - 1) = 4c^2 - 4c|y| + y^2 ,$$

$$\therefore |y| = \frac{(x - 1)}{c} + c ,$$

$$(37) \quad \therefore y = \pm\left[\frac{x - 1}{c} + c\right].$$

Equation (36) imposes the restriction $y \leqslant 2c$, so (37), subject to the restriction (36), represents a pair of line segments intersecting at $(1 - c^2, 0)$ and with gradients $\pm\frac{1}{c}$. In order to minimize the error corresponding to the given c, we have to determine the point on the line segments which is closest to the line (32). In order to determine this point we need to know what are the extremities of the line segments.

First take the negative sign in (37), $\therefore -(x - 1) = c^2 + yc$.
Substituting for $-(x - 1)$ in (36) we obtain

$$|y| + \sqrt{y^2 + 4c^2 + 4yc} = 2c .$$

On rationalizing, this reduces to

$$y^2 + 4c^2 + 4yc = 4c^2 - 4c|y| + y^2 ,$$

i.e. $y = -|y|$, so y must be non-positive, and one extremity of this line segment must be the point $(1 - c^2, 0)$. Since the vectorial angle of this line is negative acute, it can be seen graphically that the line is directed away from (32). It therefore contains no points closer to (32) than the point $(1 - c^2, 0)$.

We have shown that the negative sign in (37) is associated with non-positive y-values. In the same way it can be shown that the positive sign in (37) is associated with non-negative y-values. In this case the vectorial angle is positive acute and the gradient is $\frac{1}{c}$. The gradient of (32) is $\frac{19}{11}$. We must now distinguish between the two cases

$$\text{(a)} \quad \frac{1}{c} < \frac{19}{11} \qquad \text{and (b)} \quad \frac{1}{c} > \frac{19}{11} .$$

In case (a), if we start from the point $(1 - c^2, 0)$ and proceed along the line segment, then we are moving away from the line (32); therefore in this case the point on the line segment closest to (32) is the point $(1 - c^2, 0)$.

In case (b), as we move along the line segment we are approaching the line (32), therefore in this case we should proceed as far as restriction (36) permits. The largest permissible y-value from (36) is $y = 2c$, and for this it is necessary that $y^2 - 4(x - 1) = 0$. Hence the coordinates of the point on the line segment nearest to (32) are given by

$$(38) \qquad y = 2c, \qquad x = c^2 + 1 .$$

It is now possible with the help of (35) to determine the minimum error coefficient in terms of c where $1 > c \geqslant 0$ thus:

$$(39) \qquad R = -\frac{19}{6!}(1 - c^2)u^5 , \qquad c \geqslant \frac{11}{19} \qquad \text{(Case a)}$$

$$R = -\frac{1}{6!}[19(c^2 + 1) - 22c]u^5 , \qquad c \leqslant \frac{11}{19} \qquad \text{(Case b).}$$

The values of a_0 and a_2 which give minimum error for a given c are:

$$(40) \quad \begin{cases} \text{Case (a)}, \ c \geqslant \frac{11}{19}, & a_0 = -c^2 , \quad a_2 = 1 , \\ \text{Case (b)}, \ c \leqslant \frac{11}{19}, & a_0 = c^2 , \quad a_2 = 1 - 2c . \end{cases}$$

If $c = \frac{11}{19}$, then a_0 and a_2 may be determined by either of the equations (40), since for this value of c the two expressions for R in (39) are equal.

The final form of the four-point corrector formula in terms of two parameters is:

(41) $$y_{3c} = a_0(y_0 - y_1) + y_1 + a_2(y_2 - y_1) +$$

$$\frac{h}{24}[8y_1' + 32y_2' + 8y_3' + a_0(9y_0'+19y_1'-5y_2'+y_3') + a_2(y_0'-13y_1'-13y_2'+y_3')] -$$

$$\frac{1}{6!}[19a_0 + 11a_2 + 8]\, h^5 F^{(V)}(Z) \;.$$

The quadrilateral OBCD contains all the points from which one may derive stable corrector formulae with less error than the Adams formula.

(28) Parabola. Points on the convex side generate real values of parasitic roots.

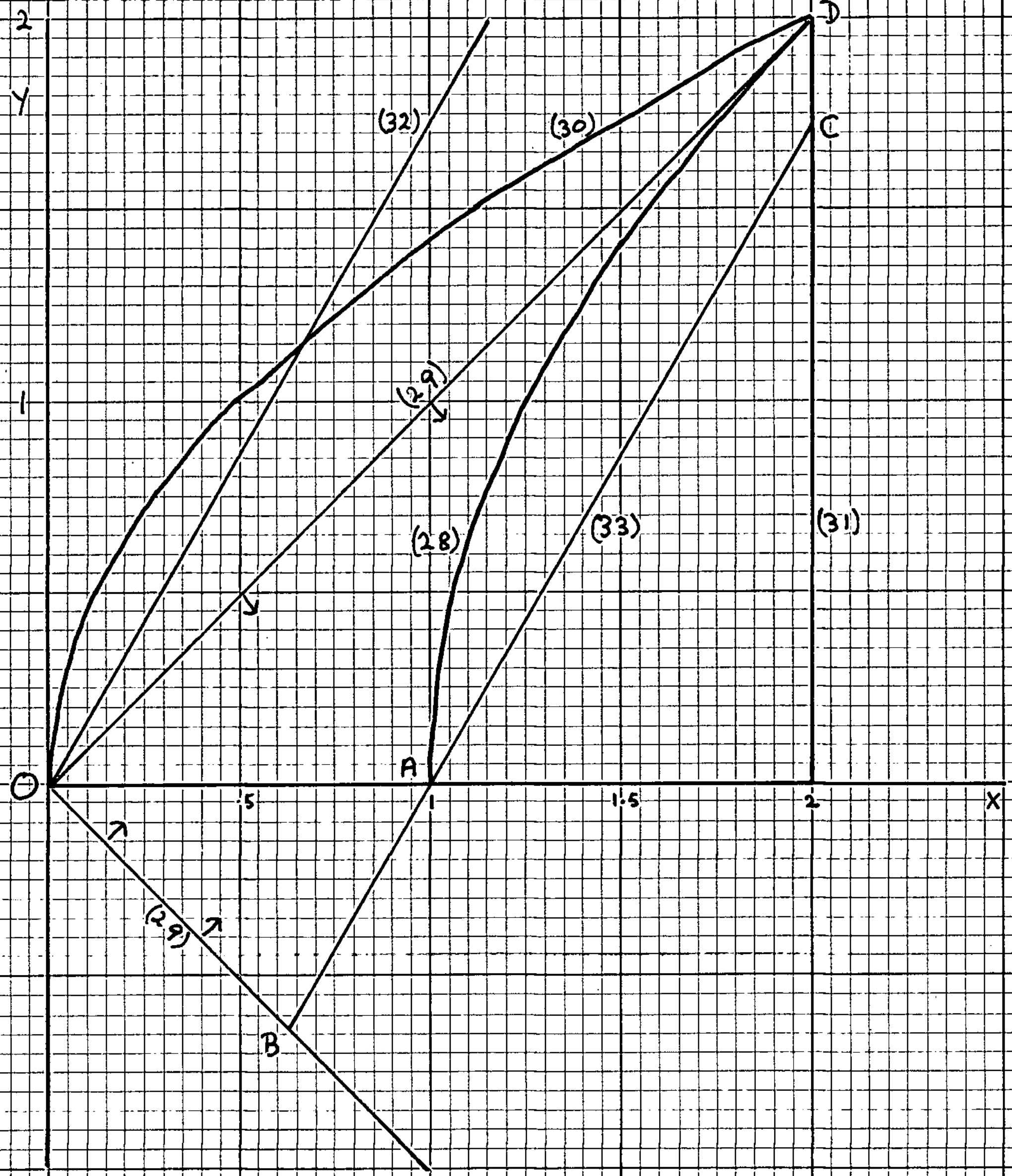
(29) Line pair. Points must be on the arrowed side in order that the formula shall be stable.

(30) Parabola. For stability it is necessary that points be chosen on the concave side. However this condition is dominated by (29).

(31) Straight line. For stability it is necessary that points be chosen to the left of this line. This condition together with (29) forms a necessary and sufficient condition for stability, provided h is sufficiently small.

(32) Locus of points giving zero fifth degree error.

(33) Locus of points giving the same error as the Adams formula.


2
Y
1
O
.5
1
1.5
2
X
D
C
(32)
(30)
(29)
(28)
(33)
(31)
A
(29)
B

To illustrate the properties of the various members of the family of four-point corrector formulae, three second-order equations were chosen with respective solutions

(a) $y = \sin x$

(b) $y = \sin \frac{x}{2}$

(c) $y = \sin 2x$.

In all three cases the modulus of the maximum truncation error is bounded. In (a) the upper bound of the fifth derivative is one, in case (b) it is 2^{-5} and in case (c) 2^5. It may therefore be expected that the truncation error is greatest in case (c) and least in case (b).

The examples were computed on the Alwac III-E digital computer at the University of British Columbia. A step-size $h = \frac{1}{16}$ was used in each computation. In each case the values of a_0 and a_2 were automatically computed from the given c by use of formulae (40). An extract of the results is given in Table 2, followed by an explanation of the table and some conclusions suggested by these results.

x (radians)	c	$10^8 E$		
		$y'' = -y$	$y'' = -\frac{y}{4}$	$y'' = -4y$
		$y = \sin x$	$y = \sin \frac{x}{2}$	$y = \sin 2x$
5	0.	-44	419	5636
	.25	-24	463	2725
	.5	-13	486	1607
	.75	-8	502	1204
	1.		-1947	
10	0.	332	-557	-6335
	.25	164	-657	-2842
	.5	86	-708	-1597
	.75	60	-738	-1174
	1.		-8481	
15	0.	421	361	-1074
	.25	214	535	-1056
	.5	113	617	-809
	.75	80	659	-660
	1.		-11628	
20	0.	-349	86	15114
	.25	-169	-148	8065
	.5	-85	-252	5011
	.75	-58	-298	3830
	1.		-13507	
25	0.	-951	-557	-29928
	.25	-478	-327	-15152
	.5	-250	-230	-9166
	.75	-172	-193	-6932
	1.		-9981	
30	0.	-123	792	37599
	.25	-72	667	18419
	.5	-44	618	10937
	.75	-32	606	8217
	1.		-3688	

Table 2.

Explanation of Table 2. The three right-hand columns each indicate true-minus-computed errors which arise when solving the differential equation at the head of the column. These errors are multiplied by 10^8. The column headed c indicates the largest parasitic root permitted in the formula used. The column headed x gives the values of x in intervals of five radians. Since each problem was worked by four or five different formulae corresponding to different values of c there are four or five entries in each error column corresponding to each value of x. For example in the first error column the first figure corresponding to $x = 10$ is 332. The column is headed $y'' = -y$, and the corresponding entry in the c column is 0. This means that when solving the equation $y'' = -y$ by a formula which permitted a maximum parasitic root of zero (i.e. by Adams' formula), the true-minus-computed error was $10^{-8} \times 332$ at the point where $x = 10$ radians.

Conclusions drawn from Table 2. Consider first the second error column of Table 2, giving the errors arising from computation of the equation $y'' = -\frac{1}{4}y$. This was the only equation on which an unstable formula was tried. The errors arising from use of the unstable formula corresponding to $c = 1$ were such that a repetition of the experiment did not seem to be justified. It may further be noted that this equation is the only one of the three in which the Adams method, corresponding to $c = 0$, compares favourably with the other methods. Even here the superiority of the Adams method is not consistent, though it lasts for 20 radians. An explanation of the success of the Adams method applied to this equation is not hard to find. The fifth derivative of the solution is less than 2^{-5}, therefore the truncation error is bound to be small. In these circumstances it is clearly better to use a formula with optimum stability properties, rather than one which is designed, at some sacrifice of stability, to reduce the already small truncation error.

The results of the other two series of experiments display the Adams method as being consistently inferior to all other methods of the family, therefore it is hard to escape the conclusion that the Adams method can generally be bettered.

The foregoing investigations may only be regarded as a first step towards the improvement of corrector formulae. Of the questions which remain unanswered the most important would appear to be: What choice of parameters will give the 'best' four-point formula for application to a given differential equation? This question and analogous questions relating to five- and six-point formulae would form a good field for further investigation.

BIBLIOGRAPHY

[1] Milne, W.E. Numerical Solution of Differential Equations, New York - London 1953.

[2] Hildebrand, F.B. Introduction to Numerical Analysis, New York 1956.

[3] Rutishauser, H. Über die Instabilität von Methoden zur Integration gewöhnlicher Differentialgleichungen, ZAMP vol. 3, 1952, pp. 65-74.

[4] Runge, C. Über die numerische Auflösung von Differentialgleichungen, Math. Ann. vol. 46, 1895, pp. 167-178.