TI-191

REPORTS HOLDINGS LIST INFORMATION

1 REPORT NO. <i>000--1469-0051</i>		2 CLASSIFICATION								
		☒ U	OUO	CDI	CFRD	CRD	SDI	SFRD	SRD	AWD-
3 ANNOUNCEMENT		4 CATEGORY & DISTRIBUTION (Indicate UC or C category)								
☒ NSA ☐ ACR ☐ RC		STD-		MN-32		MN only		No MN.		
5 ORIGIN (Check one. If foreign, give name of country)										
☒ AEC		☐ DOD		☐ NASA		☐ OGA		☐ OTHER US		☐ FOREIGN ()
6 AVAILABILITY (Check. If block 1, give sales agency and price. If block 2, give CAP price)										

- ☒ 1 Unlimited (*CFSTI*)
☐ 2 Government Agencies and Their Contractors + CAP ()
☐ 3 Government Agencies and Their Contractors
☐ 4 Export Control. Domestic Only minus Sales Outlets
☐ 5 AEC Offices and AEC Contractors
☐ 6 AEC Offices Only
☐ 7 Participants in the _____ Program. Others request from _____
☐ 8 No Distribution. Refer requests to _____

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Report No. 221

MAR 16 1967

MASTERNUMERICAL INTEGRATION OF STIFF ORDINARY
DIFFERENTIAL EQUATIONS

by

C. W. Gear

January 20, 1967

DEPARTMENT OF COMPUTER SCIENCE • UNIVERSITY OF ILLINOIS • URBANA, ILLINOIS

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

CFSTI PRICES

Report No. 221

H.C. \$ 3.00; MN. .65

NUMERICAL INTEGRATION OF STIFF ORDINARY DIFFERENTIAL EQUATIONS*

by

C. W. Gear

RELEASED FOR ANNOUNCEMENT
IN NUCLEAR SCIENCE ABSTRACTS

January 20, 1967

Department of Computer Science
University of Illinois
Urbana, Illinois 61801

*This work was supported in part by AEC-(11-1)-1469 and in part by the Argonne National Laboratory, Applied Mathematics Division.

PATENT REVIEW APPROVES RELEASE. PROCEEDURES
GOVERNING PATENT REVIEW AND RELEASE ARE ON
FILE IN RECEIVING SECTION.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

CONTENTS

ABSTRACT	Page
1. The Physical Problem	1
2. The Numerical Problem	2
3. A Simple Error Analysis	3
4. A Theorem of Dahlquist	5
5. The Region of Stability	5
6. Stiffly Stable Methods	7
7. Another Difficulty - Corrector Iteration	10
8. Practical Considerations - Starting and Step Changing	14
9. Finding Stiffly Stable Methods	16
Bibliography	18
Appendix	19

LIST OF FIGURES

Figure 1	1
Figure 2	2
Figure 3	6
Figure 4	8
Figure 5	9

ABSTRACT

Stiff equations are ones in which there are rapidly decaying components which are of no interest because of their small magnitude. Equations of this form frequently arise in physical equilibrium problems such as chemical or nuclear reactions where there are many components with a large range of time constants. Although the rapidly decaying components are of small magnitude, they cause severe numerical problems in most general purpose integration methods. This note will examine the cause of these problems and propose some ways around them. Some specific methods are developed, but the investigation of classes of methods remains to be done.

1. The Physical Problem

Consider the one parameter family of functions

$$y(t) = ce^{-\lambda t} + F(t) \quad (1.1)$$

where λ is fixed and c is the parameter. The family is the solution of

$$y'(t) = -\lambda(y - F(t)) + \frac{dF}{dt}(t) \quad (1.2)$$

If λ (assumed real for the moment) is $\gg 0$, the family (1.1) all tend to $F(t)$ very rapidly as t increases.

Figure 1

The initial value problem for (1.2) for increasing t is therefore very stable. Large initial errors are damped out very quickly.

This simple example will be treated in the sequel to clarify the presentation, but the methods to be discussed are equally applicable to systems of non-linear equations as to single linear equations. The term $ce^{-\lambda t}$ can be thought of as representing the local behavior of one of the error terms. The number $-\lambda$ is, of course, an eigen value of the matrix $\frac{\partial f}{\partial y}$ if f is a vector valued function of the n component vector y which represents the differential system by

$$y' = f(t, y) \quad (1.3)$$

Therefore λ must be allowed to assume complex values in the discussion. The time constants are $+\frac{1}{\text{Re}(\lambda)}$. They are only defined for decaying solutions

such that $\Re(\lambda) > 0$. We will assume that $F(t)$ is a well behaved function with a moderately sized derivative in relation to $e^{-\lambda t}$.

2. The Numerical Problem

If $F(t)$ is a smooth enough function such that $\frac{dy}{dt} = \frac{dF(t)}{dt}$ could normally be integrated with a step size h_0 , one would hope that equation (1.2) would have the same property. Unfortunately not! Most methods require that h be chosen so that $|h\lambda|$ is bounded by a small number. The actual bound depends on the method, 2 is not untypical. This means that h has to be chosen to see the finest structure of the solution due to terms like $ce^{-\lambda t}$, although these components may be completely insignificant. (They will almost never be zero because of rounding and truncation errors.)

These restrictions occur both in Runge-Kutta and multistep methods. We will restrict ourselves to multistep methods in this discussion after an example of the common method, Euler's method. It illustrates the problem very simply and pictorially.

Figure 2

Suppose the numerical solution has reached (t_n, y_n) with an error ϵ_n from the true solution $F(t)$ shown in Figure 2. Euler's rule projects the numerical solution along the tangent, arriving at (t_{n+1}, y_{n+1}) and so on to (t_{n+3}, y_{n+3}) , each error growing in magnitude. Mathematically, we have

$$y_{n+1} = y_n + h y'(t_n, y_n)$$

Suppose $F(t) = 0$ so that

$$y' = -\lambda y$$

$$\text{Then } y_{n+1} = y_n - \lambda h y_n = (1 - \lambda h) y_n$$

Thus if $1 - \lambda h < -1$ or $\lambda h > 2$

$$|y_{n+1}| > |y_n|,$$

and the process is unstable.

Of course, nobody uses Euler's method in ordinary differential equations--but the problem will arise in most methods. The reason that it arises can be seen very easily in the above example. The true relation of y_{n+1} to y_n is

$$y_{n+1} = (1 - \lambda h) y_n$$

The accuracy with which $1 - \lambda h$ approximates $e^{-\lambda h}$ determines the accuracy of the method. For $\lambda h < 0$, the growth rate of the differential equation is larger than that of the difference equation, for $0 < \lambda h < 2$, the differential and the difference equations are stable, although the difference representation is "relatively unstable" (Hamming's term) when $\lambda h > 1 + e^{-\lambda h}$, since then $|1 - \lambda h| > |e^{-\lambda h}|$. If $\lambda h > 2$, then $|1 - \lambda h| > 1$ and the difference method is unstable.

The stability of the difference equation depends on λh . The usual asymptotic analysis considers only the case $h \rightarrow 0$, but the problem is to perform the integration with a reasonably sized h .

3. A Simple Error Analysis

Multi-step methods are usually written in the form

$$\alpha_0 y_{n+1} = - \left[\sum_{i=1}^k \alpha_i y_{n+1-i} + h \sum_{i=0}^k \beta_i y'_{n+1-i} \right] \quad (3.1)$$

where $\alpha_0 \neq 0$ and $|\alpha_k| + |\beta_k| \neq 0$. If $\beta_0 \neq 0$, the method is called implicit because, for the non-linear equation $y' = f(x, y)$, y_{n+1} appears on the right hand side in y'_{n+1} , otherwise it is called an explicit method. The coefficients α_i and β_i are chosen to make the method accurate in the sense the value of y_{n+1} calculated from (3.1) does not differ from the exact value of $y(t_{n+1})$ by very much when the $\{y_{n+1-i} \mid i = 1, 2, \dots, k\}$ are the exact values for a class of functions $y(t)$. This difference is called the truncation error, defined as

$$T(y, t_n, h) = \sum_{i=0}^k (\alpha_i y(t_{n+1-i}) + h\beta_i y'(t_{n+1-i})) \quad (3.2)$$

with $\alpha_0 = 1$.

Consider equation (1.2) with solution (1.1). Define the difference between the calculated values y_n and the true values $y(t_n) = F(t_n) = F_n$ as $\epsilon_n = y_n - F_n$.

$$\begin{aligned} \text{By (3.1)} \quad \alpha_0 y_{n+1} &= - \sum_{i=1}^k \alpha_i y_{n+1-i} - h \sum_{i=1}^k y'_{n+1-i} \\ \alpha_0 (F_{n+1} + \epsilon_{n+1}) &= - \sum_{i=1}^k \alpha_i (F_{n+1-i} + \epsilon_{n+1-i}) - h \sum_{i=1}^k \beta_i (-\lambda (F_{n+1-i} + \epsilon_{n+1-i} - F_{n+1-i}) \\ &\quad + F'_{n+1-i}) \end{aligned}$$

$$\sum_{i=0}^k (\alpha_i - h\lambda\beta_i) \epsilon_{n+1-i} + T(F, t_n, h) = 0 \quad (3.3)$$

Therefore the errors ϵ_n obey an inhomogeneous k -step difference equation whose inhomogeneous term is the truncation error for the exact solution. Since F is assumed to be well behaved, this can be assumed to be small for reasonable methods, it is not influenced by the presence of rapidly decaying terms.

It is necessary to bound the solution of the inhomogeneous difference equation (3.3). This is done in all of the standard texts. (e.g. Henrici^[2] Discrete Variable Methods in Differential Equations.) It depends on the stability of the corresponding homogeneous equation which arises when T is set to zero. The difference equation is stable if and only if all roots of the polynomial equation

$$\sum_{i=0}^k (\alpha_i - h\lambda\beta_i) \xi^{k-i} = 0 \quad (3.4)$$

are inside the unit circle or on the unit circle and simple.

Asymptotic theory considers the case $h\lambda = 0$ only. If stiff equations are to be integrated with large h , then large values of $h\lambda$ must not make (3.4) unstable. Since there may also be components with large time constants (small λ), small values of $h\lambda$ must not make (3.4) unstable either. One root of (3.4) must approximate $e^{-h\lambda}$ for values of $h\lambda$ for which accuracy is required, the other roots must be small.

4. A Theorem of Dahlquist

The order p of a multistep method is defined as the highest degree of polynomial functions y for which $T(y, t, h)$ vanishes identically in t and h . Since the truncation error is then of the form

$$h^{p+1} k(y, t, h)$$

where $k = O(1)$ as $h \rightarrow 0$, it is usually desirable to make p large. Dahlquist^[1] has shown that if (3.4) is to be stable for all λ such that $\text{Re}(h\lambda) \geq 0$ (all stable equations) then $p \leq 2$. He also showed that the "best" method is the trapezoidal rule (with $p = 2$). This is a negative result that appears discouraging. However, let us examine the requirement $\text{Re}(h\lambda) \geq 0$.

5. The Region of Stability

The roots of (3.4) are to be small, except that one root must be close to $e^{-h\lambda}$ for those values of $h\lambda$ for which accuracy is necessary. However, they cannot be bounded everywhere except in trivial cases, in which case the method cannot be accurate. Therefore we must ask "for what regions of the $h\lambda$ plane is stability necessary and for what regions is accuracy necessary?"

Consider first the region where $\text{Re}(h\lambda) \geq D > 0$. In this region the term $ce^{-\lambda t}$ is reduced by $|e^{-\lambda h}|$ in magnitude in one step of size h . Now $|e^{-\lambda h}| \leq e^{-D}$, so if D is chosen so that e^{-D} is insignificant at the accuracy we are interested in, then we are not interested in accurate representation of this term, only in stability of the difference equation. For $\text{Re}(h\lambda) < D$ we may also be interested in accuracy. Suppose that $h\lambda = x + iy$. In one step the term $ce^{-\lambda t}$ changes by $e^{-x} e^{-iy}$. Thus, in addition to a magnitude change of e^{-x} , it oscillates $\frac{y}{2\pi}$ complete cycles. For $x < D$ we are interested in representing the function accurately, and we will do this by calculating the

values at points with a spacing of h . It is impossible to represent a band limited function with less than two samples per cycle of the maximum frequency present, and in practice, 6 to 10 samples per cycle are needed for any sort of accuracy. Therefore, if $x < D$, we must arrange that $\frac{y}{2\pi} < \theta$ where θ is smaller if higher accuracy is required. Typically θ will be around $1/5$. If the solution is growing ($x < 0$), then the step size must be chosen small enough to provide sufficient points on the curve, so that x must be limited below by some number $-\alpha$, where e^α is the maximum amount of growth to be allowed in one step. Inside the region $D \geq x \geq -\alpha$, $\theta \geq y \geq -\theta$, accurate representation of $e^{-h\lambda}$ by one root of (3.4) is required, while other roots of (3.4) must not interfere with the solution. (This could be satisfied by relative stability, stability or a combination, depending on the problem.)

In the remainder of the half plane $\text{Re}(h\lambda) < D$, we do not care what happens.

This discussion can be summarized graphically.

Figure 3

A method which satisfies these properties will be called "stiffly stable with respect to D , θ , and α ".

Roughly speaking, if more accuracy is required, then larger D and the smaller θ and α should be used. Note that, for a given method, if $h\lambda$ is in the left half plane, then h must be chosen so that $h\lambda$ lies inside the region of accuracy. This is a reasonable thing to do; it represents a reduction in h to achieve the desired accuracy. If $h\lambda$ is in the right half plane (not including the imaginary axis) then either h can be made small enough to bring $h\lambda$ inside the region of accuracy or else it can be made large enough to put $h\lambda$ in the "don't-care" region of stability, depending on whether the solution for that component is of interest or not.

6. Stiffly Stable Methods

The existence of stiffly stable methods depends on the parameters D , θ , and α and on the definition of accuracy. The usual definition of accuracy is that of order. Dahlquist has shown that if D is zero then the order p cannot exceed 2. The only positive result so far is that methods of higher order than 2 do exist if $D > 0$. To be precise, methods of order as high as 6 have been obtained and shown to be stiffly stable for suitable parameters D , θ , and α .

These methods were found by observing that as $h\lambda$ changes from 0 to plus infinity the roots of (3.4) move from the roots of the polynomial $\rho(\xi) = \sum_{i=0}^k \alpha_i \xi^{k-i}$ to those of the polynomial $\sigma(\xi) = \sum_{i=0}^k \beta_i \xi^{k-i}$. From this it is obvious that $\sigma(\xi)$ must itself have stable roots, and must have degree k . The most stable possibility is $\sigma(\xi) = \xi^k$ which has zero roots. If the unique method of degree k is calculated for this $\sigma(\xi)$, it can be shown to be stable for $k \leq 6$. Greater k have not been investigated. The unique method of degree k can be found by simple arithmetic. This is discussed in Appendix 1. The stability of the method can then be investigated in the following way.

We are interested in values of $h\lambda$ for which

$$\rho(\xi) - h\lambda\sigma(\xi) = 0 \quad (6.1)$$

has roots inside the unit circle. This region will meet the region in which this is not true at a line where one or more roots of (6.1) are on the unit

Figure 4

Figure 5

circle.

This boundary can be found in the $h\lambda$ plane as the locus of $\frac{\rho(\xi)}{\sigma(\xi)}$ for $\xi = e^{i\theta}$, $\theta \in [0, 2\pi]$. This locus can be plotted in the $h\lambda$ plane. It is known that at $h\lambda = +\infty$, the method is stable so that any region connected to that point will be stable by a continuity argument. These plots are shown for $k = 2$ (1) 6 in Figure 4 and 5. It can be seen that up to $k = 6$ methods are stiffly stable for $D > 6.1$, $\theta < .5$ and for some α that has not been determined. These methods have been used with success on a set of three equations arising in a chemistry reaction where the time constants are $1/.01$ ($\pm$ a small amount) and $1/3500$ respectively.

7. Another Difficulty - Corrector Iteration

An implicit corrector formula can not, in general, be solved explicitly for y_{n+1} . Usually a first guess (the predicted value) is used to start an iterative solution, usually of the successive substitution form. Specifically, if the corrector formula is

$$\alpha_0 y_{n+1} = -h\beta_0 y'(x_{n+1}, y_{n+1}) - \sum_{i=1}^k (\alpha_i y_{n+1-i} + h\beta_i y'_{n-i+1})$$

and the predicted value of y_{n+1} is $y_{n+1}^{(0)}$, the corrector iteration is usually

$$\alpha_0 y_{n+1}^{(m+1)} = -h\beta_0 y'(x_{n+1}, y_{n+1}^{(m)}) - \sum \dots$$

This will converge if $\left| \frac{h\beta_0}{\alpha_0} \frac{\partial y'}{\partial y} \right|$ is less than 1 in a region encompassing a large enough neighborhood of the solution. Since this depends on h , it is usual to say that for sufficiently small h the bound is satisfied. Unfortunately we are interested in large values of $h \frac{\partial y'}{\partial y}$. Instead, let us consider the corrector iteration

$$\begin{aligned} \alpha_0 (y_{n+1}^{(m+1)} + \frac{h\beta_0}{\alpha_0} D^{(m)} y_{n+1}^{(m+1)}) = \\ - h\beta_0 (y'(x_{n+1}, y_{n+1}^{(m)}) - D^{(m)} y_{n+1}^{(m)}) - \sum \dots \end{aligned} \quad (7.1)$$

where $D^{(m)}$ is a matrix in case we are talking of a system of equations. $D^{(m)}$

is not specified at this time except that $I + \frac{h\beta_0}{\alpha_0} D^{(m)}$ is to be non-singular so that (7.1) can be solved for $y_{n+1}^{(m+1)}$. Note that if $D^{(m)}$ is $\frac{\partial y'}{\partial y}(y_{n+1}^{(m)})$ then (7.1) is the Newton iteration to solve the corrector formula.

If a sequence $D^{(m)}$ can be chosen to make (7.1) converge, then it converges to the solution of the corrector formula. Therefore, if an approximation to $\frac{\partial y'}{\partial y}$ can be calculated, it can be used for $D^{(m)}$. This need not be calculated for every iterate, nor yet for every step if $\frac{\partial y'}{\partial y}$ changes slowly.

It is convenient to re-express (7.1) to simplify the arithmetic. Subtract (7.1) with $m-1$ replacing m from (7.1) to get

$$\left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right] y^{(m+1)} - \left[I + \frac{h\beta_0}{\alpha_0} D^{(m-1)} \right] y^{(m)} = - \frac{h\beta_0}{\alpha_0} \left[y'(y^{(m)}) - D^{(m)} y^{(m)} - y'(y^{(m-1)}) + D^{(m-1)} y^{(m-1)} \right] \quad (7.2)$$

(Note that the subscript 'n+1' has been dropped and that $y'(y_{n+1}^{(m)}, y_{n+1}^{(m)})$ has been abbreviated to $y'(y^{(m)})$.)

Rearranging (7.2), we get

$$\left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right] (y^{(m+1)} - y^{(m)}) = - \frac{h\beta_0}{\alpha_0} \left[y'(y^{(m)}) - \left\{ y'(y^{(m-1)}) + D^{(m-1)} (y^{(m)} - y^{(m-1)}) \right\} \right] \quad (7.3)$$

Define the variable $d_{n+1}^{(m)} = hy'(y_{n+1}^{(m-1)}) + hD^{(m-1)}(y_{n+1}^{(m)} - y_{n+1}^{(m-1)})$

Hence (7.3) can be written as

$$y^{(m+1)} = y^{(m)} - \frac{\beta_0}{\alpha_0} \left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right]^{-1} \left[hy'(y^{(m)}) - d^{(m)} \right] \quad (7.4)$$

We also need a recurrence relation for $d^{(m)}$

$$\begin{aligned} d^{(m+1)} &= hy'(y^{(m)}) + hD^{(m)}(y^{(m+1)} - y^{(m)}) \\ &= hy'(y^{(m)}) + hD^{(m)} \left\{ - \frac{\beta_0}{\alpha_0} \left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right]^{-1} \left[hy'(y^{(m)}) - d^{(m)} \right] \right\} \end{aligned}$$

$$\begin{aligned}
&= hy' (y^{(m)}) - \left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right] \left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right]^{-1} \left[hy' (y^{(m)}) - d^{(m)} \right] \\
&\quad + \left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right]^{-1} \left[hy' (y^{(m)}) - d^{(m)} \right]
\end{aligned}$$

Therefore

$$d^{(m+1)} = d^{(m)} + \left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right]^{-1} \left[hy' (y^{(m)}) - d^{(m)} \right] \quad (7.5)$$

We need to define $d^{(0)}$. It should be chosen to make (7.4) correct for $m = 0$.
From (7.1)

$$\begin{aligned}
&\left[I + \frac{h\beta_0}{\alpha_0} D^{(0)} \right] (y^{(1)} - y^{(0)}) = \\
&- \frac{h\beta_0}{\alpha_0} (y' (y^{(0)}) - D^{(0)} y^{(0)}) - \sum_{j=1}^k \frac{1}{\alpha_0} (\alpha_j y_{n-j+1} + \beta_j hy'_{n-j+1}) \\
&- \left[I + \frac{h\beta_0}{\alpha_0} D^{(0)} \right] y^{(0)} \\
&= - \frac{\beta_0}{\alpha_0} \left(hy' (y^{(0)}) + \left[\frac{\alpha_0 y^{(0)}}{\beta_0} + \sum_{j=1}^k \frac{1}{\beta_0} (\alpha_j y_{n-j+1} + \beta_j hy'_{n-j+1}) \right] \right)
\end{aligned}$$

$$\text{Hence } d_{n+1}^{(0)} = \frac{-\alpha_0 y_{n+1}^{(0)}}{\beta_0} - \sum_{j=1}^k \frac{1}{\beta_0} (\alpha_j y_{n-j+1} + h\beta_j y'_{n-j+1}) \quad (7.6)$$

Since $y_{n+1}^{(0)}$, the predicted value is a linear combination of earlier function values and derivatives, so is $d_{n+1}^{(0)}$. Note that if $y_{n+1}^{(0)}$ is the value of y_{n+1} which satisfies the corrector equation, then

$$\begin{aligned}
d_{n+1}^{(0)} &= -\frac{1}{\beta_0} \left[\sum_{j=0}^k \alpha_j y_{n-j+1} + h \sum_{j=1}^k \beta_j y'_{n-j+1} \right] \\
&= \frac{1}{\beta_0} h\beta_0 y'_{n+1} = hy'_{n+1}
\end{aligned}$$

Therefore, we identify $d_{n+1}^{(0)}$ as a "predicted" value of hy'_{n+1} . When (7.5) converges, $d^{(m)} = hy' (y^{(m)})$ so that $d_{n+1}^{(m)}$ converges to hy'_{n+1} .

The predictor and corrector formulae can now be rewritten in the form given in Gear.^[3] Define the vector

$$\underline{z}_n^{(m)} = \left[y_n^{(m)}, d_n^{(m)}, y_{n-1}, y_{n-2}, \dots, hy'_{n-1}, hy'_{n-2}, \dots \right]^T$$

The predictor formula and (7.5) can be expressed as

$$\underline{z}_{n+1}^{(0)} = B \underline{z}_n \quad (7.7)$$

where B is a suitable matrix. The corrector formulae (7.4) and (7.5) can be expressed as

$$\underline{z}_{n+1}^{(m+1)} = \underline{z}_{n+1}^{(m)} + k F(\underline{z}_{n+1}^{(m)}) \quad (7.8)$$

$$\underline{z}_{n+1} = \underline{z}_{n+1}^{(M)}$$

where $k = \left[-\frac{\beta_0}{\alpha_0}, 1, 0, \dots, 0 \right]^T$ and $F(\underline{z}_{n+1}^{(m)}) =$

$$\left[I + \frac{h\beta_0}{\alpha_0} D^{(m)} \right]^{-1} \left[hy' (x_{n+1}, y_{n+1}^{(m)}) - d_{n+1}^{(m)} \right]$$

Notice that (7.7) and (7.8) correspond to the formulae for non-stiff equations given in [3] if $D^{(m)} = 0$ in the definition F. In particular, the transformations used in [3] can be used to express these methods in equivalent forms. The Nordsieck form is very convenient for our purposes so we will carry the saved values in the vector

$$\underline{a}_n = \left[y_n, hy'_n, \frac{h^2}{2} y_n^{(2)}, \dots, \frac{h^k y_n^{(k)}}{(k)!} \right]^T$$

and transform (7.7) and (7.8) into

$$\underline{a}_{n+1}^{(0)} = A \underline{a}_n$$

$$\text{and } \underline{a}_{n+1}^{(m+1)} = \underline{a}_{n+1}^{(m)} + \ell F(\underline{a}_{n+1}^{(m)}),$$

$$\underline{a}_{n+1} = \underline{a}_{n+1}^{(m)}.$$

A is the Pascal triangle matrix

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & \dots & 1 \\ & 1 & 2 & 3 & 4 & & k \\ & & 1 & 3 & 6 & & \\ & & & 1 & 4 & & \\ & & & & 1 & \ddots & 1 & k \\ \bigcirc & & & & & & & 1 \end{bmatrix}$$

and ℓ is determined by the method. The equivalent methods to those discussed in Section 6 are given by the following ℓ .

k \	2	3	4	5	6
ℓ_0	2/3	6/11	24/50	120/274	720/1764
ℓ_1	3/3	11/11	50/50	274/274	1764/1764
ℓ_2	1/3	6/11	35/50	225/274	1624/1764
ℓ_3	0	1/11	10/50	85/274	735/1764
ℓ_4	0	0	1/50	15/274	175/1764
ℓ_5	0	0	0	1/274	21/1764
ℓ_6	0	0	0	0	1/1764

8. Practical Considerations - Starting and Step Changing

To write a useful code, a number of practical problems must be solved. These include:

- (a) How to start
- (b) How to choose the step size
- (c) How to decide when to re-evaluate an approximation to

$$\left[I + \frac{h\beta_0}{\alpha_0} D \right]^{-1}$$

- (d) What order method to use.

Insufficient numerical work has been done so far to substantiate the suggestions below, but preliminary tests suggest that the following provides a practical solution to these problems. (Parenthetically it should be noted that the following remarks also apply to other than stiff equations, which also raises the question of when to use stiffly-stable formulae and evaluate an approximation to

$\left[I + \frac{h\beta_0}{\alpha_0} D \right]^{-1}$ and when not to bother. If stiffly stable methods with reasonable error terms can be found, then it may be worth always using them. The following method is based on this assumption.)

The method consists of the following steps. Initially it is assumed that some initial values exist for the function values, their derivatives and possibly higher derivatives. All unknown derivatives are set to zero. The order p of the method first selected has been set so that a sufficient number of derivatives are known. (If only the function values and first derivatives are known, then p is 1.) The formulae used must be such that the predictor and the corrector have the same order so that the difference between the predictor and the corrector can be used to estimate the first ignored derivative, and hence the truncation error.

- Step 1 Predict the value at the next point by $a_{n+1}^{(0)} = A a_n$.
- Step 2 Iterate the corrector without changing the value of D until it converges or until too many iterations have been used (5?).
- Step 3A If too many iterations were used, and this is the first attempt of step 2, re-evaluate D and repeat step 2. If then is the second attempt, halve the step size and return to step 1.
- Step 3B If the corrector converges, examine the difference between the predicted and corrected values of the functions. This is an estimate of $k h^{p+1} y^{(p+1)}$ where k depends on the method. Since the leading term of the error has the form $C h^{p+1} y^{(p+1)}$ where C is also known, an estimate of the error can be made. This should be compared with Eh where E is a measure of the error to be permitted on the unit interval. (If the equations are known to be very stable or unstable, then E should be changed throughout the interval.) If the estimated error is too large the step must be rejected; if it is too small, it

can be accepted, but in either case, the next step should be an order determining type (Step #4). If the error is within bounds, return to Step 1 for the next integration.

Step 4 If the error is too large or too small the step size must be adjusted (by a factor of 2 down or up on a binary machine). However, do not increase the step until at least p steps have been made since the last change. (From a multistep point of view, this would amount to using points further back in history than the last step change, where presumably, things were less well behaved.)

Having adjusted the step, perform a test for order as follows. An estimate of the scaled (p+1)st derivative is available from the last successful step (except in the case of the first step). Use this and perform an integration with the above technique of order p+1. Do not check the error; instead use the predictor corrector difference to estimate the error, and compare this with the estimated error for methods of order p and p-1 which can be obtained using the available values of the scaled derivatives $\frac{h^{p+1} y^{(p+1)}}{(p+1)!}$ and $\frac{h^p y^{(p)}}{p!}$.

Choose the order method which gives the smallest error (in relation to the components of E if a system is being integrated). Repeat the last step if necessary.

9. Finding Stiffly Stable Methods

At present, no technique is known for finding stiffly stable methods except that of using common sense and investigating the properties of the guess. A possible approach might be the following (although it has not been successful yet).

$\frac{\rho(\xi)}{\sigma(\xi)}$ should approximate $\log \xi$ for an appropriate region of the ξ -plane corresponding to the accurate region of Figure 3. The part of $|\xi| > 1$ that does not correspond to the accurate region should map into the don't care region while the part of $|\xi| < 1$ that does not correspond to the accurate region may map into anything. $\frac{\rho(\xi)}{\sigma(\xi)}$ is a rational fraction, so we ask the question

"how is a rational fraction chosen to satisfy the aims above?" Such a method

method may not be consistent, but that does not matter since we do not want convergence as $h \rightarrow 0$, the point of the work is to handle large h .

BIBLIOGRAPHY

- [1] Dahlquist, G. G., "A Special Stability Problem for Linear Multistep Methods," BIT 3 (1963) pp. 27-43.
- [2] Henrici, P., Discrete Variable Methods for Ordinary Differential Equations, Wiley, New York.
- [3] Gear, C. W., "Numerical Solution of Ordinary Differential Equations of Various Orders," Argonne National Laboratory Report, ANL 7126, Argonne, Illinois, 1966.

APPENDIX

Determining $\rho(\xi)$ when $\sigma(\xi) = \xi^k$.

It is known (Henrici^[2]) that if the method is of degree p , then

$$\rho(\xi) - \log \xi \sigma(\xi) \sim c(\xi-1)^{p+1} \text{ as } \xi \rightarrow 1 \quad (\text{A1.1})$$

(This is the statement that the root of $\rho(\xi) - \log(\mu) \sigma(\xi) = 0$ is close to μ .)

If we set $p = k$ and $\sigma(\xi) = \xi^k$, (A1.1) determines $\rho(\xi)$ in the following way.

Set $\xi-1 = x$

$$\rho(x+1) = \log(x+1) \cdot (1+x)^k \text{ truncated to terms in } x^k \text{ and lower.}$$

If we represent $\rho(\xi)$ by the column vector

$$\underline{\rho} = [\alpha_k, \alpha_{k-1}, \dots, \alpha_0]^T$$

where $\rho(\xi) = \sum_{j=0}^k \alpha_j \xi^{k-j}$, then $\underline{\rho}$ can be expressed as

$$\underline{\rho} = \begin{bmatrix} 1 & -1 & 1 & -1 & \dots & (-1)^k \\ 0 & 1 & -2 & 3 & & (-1)^{k-1} k \\ 0 & 0 & 1 & -3 & & (-1)^{k-2} \binom{k}{2} \\ & & & 1 & & \\ 0 & 0 & & & & 1 \end{bmatrix} \begin{bmatrix} 1 \\ k & 1 \\ \binom{k}{2} & k & 1 \\ \binom{k}{3} & \binom{k}{2} & k & 1 \\ 1 & k & \binom{k}{2} & \dots & \binom{k}{3} & \binom{k}{2} & k & 1 \end{bmatrix} \begin{bmatrix} 0 \\ +1 \\ -1/2 \\ +1/3 \\ \dots \\ -(1)^k/k \end{bmatrix}$$

The last column consists of the coefficient of the logarithm series for $(1+x)$. The second matrix corresponds to multiplying by $(1+x)^k$ and ignoring terms above x^k . The first matrix arises from the conversion back to $\xi = 1+x$.

If this is done for $k = 2(1)6$, we get the examples:

$$y_{n+1} = \alpha_1 y_n + \alpha_2 y_{n-1} + \dots + \alpha_k y_{n-k+1} + h\beta_0 y_{n+1}$$

where $\{\alpha_i\}$ and β_0 are given by:

k	β_0	α_1	α_2	α_3	α_4	α_5	α_6
2	2/3	4/3	-1/3	0	0	0	0
3	6/11	18/11	-9/11	2/11	0	0	0
4	12/25	48/25	-36/25	16/25	-3/25	0	0
5	60/137	300/137	-300/137	200/137	-75/137	12/137	0
6	60/147	360/147	-450/147	400/147	-225/147	72/147	-10/147

2