PREDICTOR-CORRECTOR PROCEDURES FOR SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS

by

RAMSAY VINCENT MICHEL ZAHAR
B.A.Sc. (Engineering Physics),
The University of British Columbia, 1962

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

April, 1964

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 representatives. 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 April 10, 1964

ABSTRACT

Some of the most accurate and economical of the known numerical methods for solving the initial-value problem

$$\frac{dx}{dt} = f(t,x) \quad , \quad x(t_0) = x_0$$

are of the predictor-corrector type.

For systems of equations, the predictor-corrector procedures are defined in the same manner as they are for single equations.

For a given problem and domain of t, a plot of the maximum error in the numerical approximation to $x(t)$ obtained by a predictor-corrector procedure, versus the step-size, can be divided into three general regions - round-off, truncation, and instability. The most practical procedures are stable and have a small truncation error.

The stability of a method depends on the magnitudes of the eigenvalues of a certain matrix that is associated with the matrix

$$G = (g_{ij}) = \left(\frac{\partial f_i(t,x)}{\partial x_j}\right) \quad .$$

When the functions f_i are complicated, predictor-corrector procedures involving two evaluations per step seem to be the most efficient for general-purpose applications.

ACKNOWLEDGEMENTS

I thank Professor T. E. Hull of the Department of Mathematics at the University of Toronto for his valuable theoretical ideas and for his assistance in interpreting the experimental results. I also thank him for the advice and encouragement he has given me throughout the course of my work.

I wish to express my appreciation to Professor R. D. James, Head of the Department of Mathematics at the University of British Columbia, and Professor C. C. Gotlieb, Chairman of the Department of Computer Science at the University of Toronto for the financial aid they presented to me during my graduate studies.

TABLE OF CONTENTS

INTRODUCTION

In this thesis, we shall consider the initial-value problem

$$x' = f(t,x) \quad , \quad x(t_0) = x_0 \tag{1}$$

where in general, x and x_0 are N-dimensional vectors. In fact, (1) represents the system of first order ordinary differential equations

$$\begin{aligned} x_1' &= f_1(t,x_1,x_2,\ldots,x_N) \\ x_2' &= f_2(t,x_1,x_2,\ldots,x_N) \\ &\cdots\cdots\cdots\cdots \\ x_N' &= f_N(t,x_1,x_2,\ldots,x_N) \end{aligned}$$

where the values of the x_i , $i=1,2,\ldots,N$, are specified for an initial time t_0 .

For convenience in expressing the following results, we shall adopt the standard notation of functional analysis. Let S be a complete normed linear space (a Banach space). Let the elements of S be $u_1,u_2,\ldots$. The norm on S is a real-valued function, whose value at u is denoted by $\|u\|$, and which satisfies

(a) $\|u_1+u_2\| \leqslant \|u_1\| + \|u_2\|$

(b) $\|u\| \geqslant 0$ and $\|u\| = 0$ if and only if $u = 0$, the zero element of S

(c) $\|\alpha u\| = |\alpha| \, \|u\|$ for any complex number α .

S becomes a metric space if we define a distance function between any two points u_1 and u_2 of S as

$$d(u_1,u_2) = \|u_1-u_2\| \quad .$$

A sequence of functions $u_1,u_2,\ldots$, satisfies the Cauchy Criterion if

$$\|u_n-u_m\| \longrightarrow 0 \quad \text{as} \quad n,m \longrightarrow \infty \quad .$$

Since S is complete, each such Cauchy sequence possesses a limit u, where u is an element of S. We say that the sequence $\{u_i\}$ converges to u.

We now let S be the real Euclidean N-dimensional space R_N. If x is an element of R_N, x is the vector with real components x_i, $i=1,2,\ldots,N$. R_N is complete (see e.g. [11], p. 96).

Later, we will let

$$\|x\| = |x_1| + |x_2| + \ldots + |x_N| \quad .$$

However, any valid norm may be used in the following discussion.

We introduce the usual notation regarding the signs $<$, $\leqslant$, $>$, $\geqslant$. For example, $x \leqslant y$ if $x_i \leqslant y_i$ for each $i=1,2,\ldots,N$.

Certain existence theorems for (1) exist. The Cauchy-Lipschitz Theorem states that if $f(t,x)$ is continuous and satisfies the Lipschitz-condition

$$\|f(t,x) - f(t,\tilde{x})\| \leqslant L \|x-\tilde{x}\|$$

in a region containing (t_0,x_0) for some positive real number L, then there exists a unique solution $x(t)$ to problem (1) over a suitable interval containing t_0. To prove this

theorem and other similar theorems (see [9]), one first introduces a method such as the Euler polygonal method to generate approximate solutions. Then one proves that these approximate solutions converge to the solution of (1).

Although a given function f may satisfy the conditions of the above theorem, thus assuring the existence and uniqueness of a solution to (1), it may be difficult or even impossible to represent this solution in closed form (as finite combinations of elementary functions such as polynomials, exponential functions and logarithms, and of indefinite integrals of such functions). For example, it has been proven that the equation

$$x' = t+t^2+x^2 \quad , \quad x(0) = 0$$

cannot be solved in terms of elementary functions although its solution does exist and can be tabulated.

However, the proof of the existence theorem provides us with a method of constructing approximations to the solution of (1) when it exists. In practice, only some of these methods, particularly the Runge-Kutta and multi-step methods, are useful in numerical calculations. Consequently, if we wish to generate a numerical approximation to the solution of (1), we must first have a criterion for choosing the numerical method.

In general, the required accuracy is prescribed. Therefore, we choose the numerical method that will give this accuracy at a minimum cost.

It is the purpose of this thesis to show in general,

that predictor-corrector methods can be used to generate numerical solutions to (1) of the required accuracy, and in particular, that certain such methods will yield the prescribed accuracy at a minimum cost. The special case $N=1$, when (1) simplifies to a single equation, has been investigated thoroughly in [8].

In the next three sections are represented the theoretical results upon which our investigation is based. The reasoning for the general case is quite similar to that for the case $N=1$. However, as we shall see, certain generalizations are quite impractical because of their complicated nature.

To specify a predictor-corrector method, we must choose a starting procedure, the predictor and corrector formulas, a rule for iterating with the corrector formula, and a step-size.

In section 5, we present experimental results for a number of problems. We use a variety of different procedures for each problem. Any restrictions on the chosen ranges of parameters and on the considered formulas are based on the results for the case $N=1$.

Throughout this thesis, we shall assume that the function $f(t,x)$ is reasonably complicated. Thus, the cost of any procedure will be directly proportional to the number of times this function is evaluated.

THEORY

In this section, we first define the predictor and corrector formulas and the rules for iteration. Next, we show that the numerical sequence determined by the iteration procedure converges to a limit. We then investigate how closely this limit approximates the actual solution of the system of differential equations. Finally, we obtain specific conditions for the stability of the Adams method.

The predictor formula can be expressed as

$$y_n = \sum_{i=1}^{k} a_i^* y_{n-i} + h \sum_{i=1}^{k+1} b_i^* y'_{n-i} \qquad (2)$$

and the corrector formula as

$$y_n = \sum_{i=1}^{k} a_i y_{n-i} + h \sum_{i=0}^{k} b_i y'_{n-i} \qquad (3)$$

where h is the step-size, $t_n = t_0+nh$, $y_{n-i} = y(t_n-ih)$ and $y'_{n-i} = f(t_n-ih, y_{n-i})$. (The terminology is the same as in [8].) We emphasize that the y's and the y''s are N-dimensional vectors. We have chosen the unknowns a_i^*, b_i^*, a_i and b_i to be one-dimensional constants and have thus restricted ourselves to using the same predictor and corrector formulas for each of the N components. Our justification for doing this lies only in the relative simplicity of the resulting analysis.

Assuming that the values of y_{n-i} and y'_{n-i} needed in (2) are known, a predictor-corrector procedure at time t_n is defined by first using (2) to calculate an approxima-

tion y_n to x_n. Then the y_n so determined is substituted into $f(t,x)$ to evaluate an approximation y_n' to x_n'. (3) is then used to correct, giving a new approximation to x_n. We may then evaluate obtaining a new y_n', and then correct and so on. We shall denote the resulting sequence of values as $y_{n,0}, y_{n,1}, \ldots, y_{n,m}$ where $y_{n,0}$ is the approximation to x_n obtained from (2), $y_{n,j}$ is the approximation to x_n obtained from the j^{th} application of the corrector, and m is the number of iterations. We distinguish between two cases, those that end on a correction and those that end on an evaluation. We denote the former by $P(EC)^m$ and the latter by $PE(CE)^m$. We understand that whichever case is used, m will be constant throughout the domain of t considered.

It is worth noting that under the assumption that $f(t,x)$ is relatively complicated, the cost of a correction will be negligible with respect to the cost of an evaluation. Thus, the method ending on a correction is the most logical to use.

To use any predictor-corrector method on an initial-value problem, the starting values $y_1, \ldots, y_k$ must be secured by another method known as the starting procedure. One-step methods such as the Runge-Kutta method and others (see [1], p. 81) are frequently used.

We now ask the following question. For arbitrarily chosen values $y_{n-k}, \ldots, y_{n-1}$, does the infinite sequence $y_{n,j}$ converge to a unique y_n which satisfies (3) exactly? The answer is given by the following theorem (cf. [6], p. 216)

which is a special case of a classical theorem of functional equations (see e.g. [1], p. 38).

Before we state and prove the theorem, we note that in (2), y_n does not appear in the left hand side. Therefore, (2) determines $y_{n,0}$ explicitly as a function of the $y_{n-1},\dots,y_{n-k-1}$. Also, using (3), $y_{n,j}$ can be expressed in terms of $y_{n,j-1}$ by

$$y_{n,j} = F(y_{n,j-1}) \qquad (4)$$

where $F(y_{n,j-1}) = \sum_{i=1}^{k} a_i y_{n-i} + hb_0 f(t_n, y_{n,j-1}) + \sum_{i=1}^{k} b_i y'_{n-i}$.

Theorem 1: Let the function F be defined for all y in R_N and suppose F satisfies a Lipschitz-condition with $K < 1$. Then the iteration (4) defines a sequence $y_{n,1}, y_{n,2}, \dots$ which converges to a unique solution y_n of (3). Also,

$$\|y_n - y_{n,m}\| \leqslant \frac{K^m}{1-K} \|y_{n,1} - y_{n,0}\| .$$

If L is the Lipschitz constant for $f(t,x)$, then a Lipschitz constant K for F is given by $K = hb_0L$, and $K < 1$ for all $h < h_0 = 1/b_0L$. Now

$$\|y_{n,j} - y_{n,j-1}\| = \|F(y_{n,j-1}) - F(y_{n,j-2})\| \leqslant K \|y_{n,j-1} - y_{n,j-2}\|$$

Therefore $\|y_{n,j} - y_{n,j-1}\| \leqslant K^{j-1} \|y_{n,1} - y_{n,0}\|$.

and since $\|y_{n,\nu+\mu} - y_{n,\nu}\| \leqslant \|y_{n,\nu+\mu} - y_{n,\nu+\mu-1}\| + \dots$

$$\dots + \|y_{n,\nu+1} - y_{n,\nu}\| ,$$

then $\| y_{n,\nu+\mu} - y_{n,\nu} \| \leq (K^{\nu+\mu-1} + \ldots + K) \| y_{n,1} - y_{n,0} \|$

$$\leq \frac{K^{\nu}}{1-K} \| y_{n,1} - y_{n,0} \| \quad .$$

Therefore $\| y_{n,\nu+\mu} - y_{n,\nu} \| \longrightarrow 0$ as $\nu \longrightarrow \infty$.

Thus, the sequence $y_{n,0}, y_{n,1}, \ldots$ is a Cauchy sequence and converges to a limit y_n .

Now, since F satisfies a Lipschitz-condition, F is continuous. Therefore,

$$y_n = \lim_{\nu \to \infty} (y_{n,\nu}) = \lim_{\nu \to \infty} F(y_{n,\nu-1})$$

$$= F(\lim_{\nu \to \infty} y_{n,\nu-1}) = F(y_n) \quad .$$

That is, y_n satisfies (3) exactly.

Suppose the sequence converges to $\tilde{y}_n$ as well. Then $y_n = F(y_n)$ and $\tilde{y}_n = F(\tilde{y}_n)$, and

$$\| y_n - \tilde{y}_n \| \leq K \| y_n - \tilde{y}_n \| \quad .$$

But since $K < 1$, this is a contradiction unless $\| y_n - \tilde{y}_n \| = 0$, which implies that $y_n = \tilde{y}_n$. Therefore, y_n is unique.

Finally, letting $\mu \longrightarrow \infty$ in the inequality for $\| y_{n,\nu+\mu} - y_{n,\nu} \|$, we get

$$\| y_n - y_{n,m} \| \leq \frac{K^m}{1-K} \| y_{n,1} - y_{n,0} \| \quad .$$

This completes the proof of the Theorem.

The inequality in the statement of the theorem is significant because it indicates the existence of an estimate for $\| y_n - y_{n,m} \|$.

Now that the convergence of our iteration process is established, it remains for us to investigate how closely the vector $y_{n,m}$ approximates x_n, the solution of (1) at t_n.

To indicate the relation between y_n and x_n, we write

$$x_n = \sum_{i=1}^{k} a_i^* x_{n-i} + h \sum_{i=1}^{k+1} b_i^* x'_{n-i} + T_n^* \qquad (5)$$

where we define the vector T_n^* to be the truncation error of the predictor formula. Similarly, we let

$$x_n = \sum_{i=1}^{k} a_i x_{n-i} + h \sum_{i=0}^{k} b_i x'_{n-i} + T_n \qquad (6)$$

define the truncation error of the corrector formula T_n.

The $y_{n,m}$ are the values we expect to obtain when we perform the numerical calculations. However, the numerical values we achieve are not exact, for they are truncated or rounded to a finite number of significant figures. We let the vectors $Z_{n,j}$ denote the rounded results. Using the previous notation, $Z_{n,0}$ is the rounded result obtained from the predictor formula at t_n, and $Z_{n,j}$, $j=1,\ldots,m$ is the rounded result after the j^{th} application of the corrector. In the following, we shall consider the procedure $P(EC)^m$, which turns out to be the more complicated. The case $PE(CE)^m$ can be treated similarly and the results will be given later.

We define the vector r_n^*, the round-off error at the predictor by

$$Z_{n,0} = \sum_{i=1}^{k} a_i^* Z_{n-i,m} + h \sum_{i=1}^{k+1} b_i^* Z_{n-i,m-1} - r_n^* \qquad (7)$$

and the vector $r_{n,j}$, the round-off error at the j^{th} application of the corrector by

$$Z_{n,j} = \sum_{i=1}^{k} a_i Z_{n-i,m} + hb_0 Z'_{n,j-1} + h \sum_{i=1}^{k} b_i Z'_{n-i,m-1} - r_{n,j} \qquad (8)$$

We define the propagated error to be the vector $e_n = x_n - Z_{n,m}$. We define the matrix G by the equation $f(t,u) - f(t,v) = G \cdot (u-v)$. Applying the Theorem of the Mean (see e.g. [10], p. 224) to each component f_i of f, we see that

$$G = (g_{i,j}) = \left(\frac{\partial f_i(t,\tilde{u}_i)}{\partial x_j}\right)$$

where the $\tilde{u}_i$ are certain vectors such that $u < \tilde{u}_i < v$.

We now subtract equation (7) from (5) and equations (8) from (6) obtaining

$$\begin{aligned}
e_{n,0} &= \sum_{1}^{k} a_i^* e_{n-i,m} + hG \sum_{1}^{k+1} b_i^* e_{n-i,m} + (T_n^* + r_n^*) \\
e_{n,1} &= \sum_{1}^{k} a_i e_{n-i,m} + hb_0 G e_{n,0} + hG \sum_{1}^{k} b_i e_{n-i,m-1} + (T_n + r_{n,1}) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
e_{n,m} &= \sum_{1}^{k} a_i e_{n-i,m} + hb_0 G e_{n,m-1} + hG \sum_{1}^{k} b_i e_{n-i,m-1} + (T_n + r_{n,m}) .
\end{aligned} \qquad (9)$$

This is a set of $(m+1)$ difference equations in the $(m+1)$ unknowns $e_{n,j} = x_n - Z_{n,j}$, $j=0,1,\ldots,m$. Before solving

these equations, we replace the T_n^* , r_n^* , T_n , $r_{n,j}$ by respective constant values T^* , r^* , T , r . We also assume that G is the same constant matrix for each equation. It is clear that the $e_{n,j}$ satisfying these new equations will approximate the actual $e_{n,j}$ only if T_n^* , r_n^* , T_n , $r_{n,j}$, and G change very little over the intervals under consideration (those intervals used when the Theorem of the Mean is applied). For now, we assume that the intervals are small enough so that this is true. It will turn out that the resulting expressions are only slightly dependent on the changes of these variables. Thus, we would expect our resulting expressions to indicate reasonably the way in which the error propagation depends on the choice of procedure.

Actually, a more rigorous analysis can be outlined as follows. If we replace the a_i and b_i in equations (9) by their corresponding absolute values, and assume that each component of the T_n^* , T_n , r_n^* , r_n is bounded in absolute value by the corresponding component of T^* , T , r^* and r , respectively, we obtain 'dominating' difference equations, the last of which is

$$E_{n,m} = \sum_{i=1}^{k} |a_i| E_{n-i,m} + h|b_0| G E_{n,m-1} + hG \sum_{i=1}^{k} |b_i| E_{n-i,m-1} + T+r$$

where G is defined in the Lipschitz sense by

$$|f_i(t,x_1,\ldots,x_j,\ldots x_N) - f_i(t,x_1,\ldots,\tilde{x}_j,\ldots x_N)| < g_{ij}|x_j-\tilde{x}_j| \quad .$$

Upon solving the dominating difference equations corresponding to (9), we obtain an expression for the $E_{n,m}$. It is easy to see that the $E_{n,m}$ are bounds for the $e_{n,m}$. However, these bounds turn out to be ultra-conservative in general (because the b_i are not all positive, for example) so we shall not consider them further.

Now, we are interested in the propagated error $e_n \equiv e_{n,m}$. Thus, eliminating $e_{n,m-1}$ from the last equation of (9), and then $e_{n,m-2}$ and so on, gives

$$e_{n,m} = (I-\Theta^m)(I-\Theta)^{-1}\Big[\sum_1^k a_i e_{n-i,m} + hG\sum_1^k b_i e_{n-i,m-1} + (T+r)\Big]$$
$$+ \Theta^m\Big[\sum_1^k a_i^* e_{n-i,m} + hG \sum_1^{k+1} b_i^* e_{n-i,m-1} + (T^*+r^*)\Big] \tag{10}$$

where $\Theta = hb_0G$ and I is the identity matrix. For this expression to be correct, it is necessary and sufficient that all eigenvalues of Θ are less than one in absolute value. ([3], p. 60). This condition is quite similar to the sufficient condition for the convergence of our iterations in Theorem 1. However, this condition on Θ is not sufficient to ensure that e_n is bounded.

If $m \longrightarrow \infty$, $e_{n-i,m-1} = e_{n-i,m}$ and $\Theta^m \longrightarrow$ the zero matrix, so (10) becomes the difference equation of order k for the vectors $e_n,\ldots,e_{n-k}$,

$$e_n = (I-\Theta)^{-1}\Big[\sum_1^k a_i e_{n-i} + hG\sum_1^k b_i e_{n-i} + (T+r)\Big]. \tag{11}$$

(11) represents a system of N difference equations for

each of the N components of each of the vectors involved. Because G is a matrix, each of the N equations contains, in general, all of these N(k+1) components. The resulting equations are quite complicated and difficult to solve in this form. Even the case N=1, when (1) and thus (11) become single equations, is not easy. We shall now consider this less general case in detail.

When N=1, (11) becomes

$$e_n - \sum_0^k (a_i + hgb_i)\, e_{n-i} = (T+r) \qquad (12)$$

where $g = \dfrac{\partial f(t,\tilde{x})}{\partial x}$ for some $\tilde{x}$. This difference equation can be solved by a standard method ([6], p. 209). The solution e_n can be written in the form

$$e_n = A_1 s_1^n + \ldots + A_k s_k^n + e \qquad (13)$$

where the A_i are constants. The first k terms are solutions of the homogeneous equation obtained from (12) by putting the right hand side equal to 0, and e is the particular solution of the non-homogeneous equation. e is formed by assuming that $e_n = e_{n-1} = \ldots = e_{n-k} = e$, a constant. Thus

$$e = \frac{(T+r)}{1 - \sum_0^k (a_i + hgb_i)} \quad .$$

Substituting $e_{n-i} = As^{n-i}$, $i=0,\ldots,k$ into the homogeneous equation obtained from (12), the results show that the s_i are roots of the polynomial

$$C(s) = s^k - \sum_0^k (a_i + hgb_i)\, s^{k-i} \quad .$$

We shall assume that the s_i are distinct. If they are not, the resulting discussion, and in particular, equation (13) must be modified. However, these modifications are not complicated and are described in [6].

We have written $e_n = x_n - Z_{n,m}$. Thus, using the Theorem of the Mean,

$$e_n' = f(t_n, x_n) - f(t_n, Z_{n,m}) = \tilde{g} e_n$$

where $\tilde{g} = \dfrac{\partial f(t_n, \tilde{x})}{\partial x}$ for some $\tilde{x}$, $x_n < \tilde{x} < Z_{n,m}$.

On the basis of this equation, we would expect the error to be proportional to e^{tg}, which is a solution of this equation for constant $\tilde{g} = g$. In fact, as we shall see later, if $T = O(h^p)$, one root of $C(s)$ is then $s_1 = e^{hg} + O(h^p)$. $s_2, \ldots, s_k$ are extraneous roots and have arisen because we have approximated a first order differential equation with a difference equation of order k.

When m is finite, equation (10) must be modified. (10) now contains the other unknowns $e_{n,m-1}, \ldots, e_{n-k,m-1}$. But (10) and the last equation of (9) can be considered as two simultaneous difference equations in the variables $e_{n-i,m}$ and $e_{n-i,m-1}$ for all $i=0,\ldots,k$. To solve these equations, we put

$$e_{n,m} = As^n \quad , \quad e_{n,m-1} = Bs^n$$

and, as before, the solution e_n will be a linear combination of the n^{th} powers of the resulting roots s, plus a 'particular integral'. After cancellation, the equations

become two homogeneous linear equations in A and B. In order that solutions for A and B exist, the determinant of these equations must be 0. This condition yields the required polynomial in s. The s_i are then the $(2k+1)$ roots of the polynomial

$$s^{k+1}C(s) + \theta^{m-1}\left\{\theta s^k\left[\sum_1^{k+1}(a_i-a_i^*+\theta a_i^*)\, s^{k+1-i}\right]\right.$$

$$+ hg\left[s^{k+1}\sum_1^k b_i s^{k-i}+(1-\theta)\left(\sum_1^k a_i s^{k-i}\sum_1^{k+1} b_i^* s^{k+1-i}\right.\right.$$

$$\left.\left.\left. - \sum_1^{k+1} a_i^* s^{k+1-i}\sum_1^k b_i s^{k-i} - s^k\sum_1^{k+1} b_i^* s^{k+1-i}\right)\right]\right\} \quad . \qquad (14)$$

As $m \to \infty$, $(k+1)$ roots of this polynomial approach 0. The other k roots become the roots of $C(s)$.

Before writing down the complete expression for e_n, we shall discuss how the a_i, a_i^*, b_i and b_i^* are chosen. Closely related with this problem is the concept of stability.

Roughly speaking, a method is defined to be stable if the error e_n is insensitive to small changes in the local errors - the errors at each step of the calculation. Now, one term in the expansion of e_n is proportional to the n^{th} power of the root s_1 that approximates e^{hg} and is to be expected from the differential equation. We want the contributions from the other s_i to be negligible with respect to the contribution from s_1. If g is positive, $e^{hg} > 1$. For the term of e_n containing s_1 to be the dominating term, the other s_i must be at least less than s_1 in absolute value. If g is negative, we want the

terms of e_n to be bounded. Thus, we define a method to be stable if $|s_i| \leq s_1$ for $g > 0$ and if $|s_i| < 1$ for $g < 0$. Stability is defined in this way so that the relative error of a stable method is small in absolute value.

The degree of a predictor or corrector formula is defined to be the largest integer p such that $T = O(h^p)$. Obviously, we would like to choose our constants so that the method has as high a degree as possible.

We now return to equation (5), and using the Taylor expansion

$$x_{n-i} = x(t_n - ih) = x(t_n) + (ih)x'(t_n) + \frac{(ih)^2}{2!} x''(t_n) + \cdots$$

and a similar one for x'_{n-i}, we expand the right-hand side in powers of h. By equating the coefficients of the same powers of h on each side of the resulting equation, we obtain the relations

$$\begin{aligned}
&a_1^* + a_2^* + \cdots + a_k^* = 1 \\
&a_1^* + 2a_2^* + \cdots + ka_k^* = (b_1^* + b_2^* + \cdots + b_{k+1}^*) \\
&a_1^* + 2^2 a_2^* + \cdots + k^2 a_k^* = 2(b_1^* + 2b_2^* + \cdots + (k+1)b_{k+1}^*) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&a_1^* + 2^{k+1} a_2^* + \cdots + k^{k+1} a_k^* = (k+1)(b_1^* + 2^k b_2^* + \cdots + (k+1)^k b_{k+1}^*) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned} \tag{15}$$

We first note that at least the first two of these equations must be satisfied if the true solution $x(t)$ is to satisfy the predictor equation except for terms of $o(h)$ as $h \to 0$.

(These two equations are the necessary and sufficient conditions for the consistency of a method [9].)

Since we have (2k+1) unknowns, it would seem reasonable to solve the first (2k+1) equations of (15) for our unknowns and thus obtain a method with a truncation error of degree 2k+1. However, Dahlquist [2] proved the rather remarkable theorem that the degree of a stable operator of order k cannot exceed (k+2) (or k+3 in an exceptional case which we will not consider). If we use the first (k+2) equations to determine the constants, T_n^* will be $O(h^{k+2})$. Now, the determinant of the coefficients of the b_i^* is a Vandermonde determinant and therefore is not 0. Thus, there is a large number of stable methods with $p = k+2$. In practice, the a_i^* are chosen first, and then the b_i^* are determined in this manner.

The corrector constants are chosen in the same way as the predictor constants. The resulting equations are similar to (15) with b_0 appearing on the right hand side of the second equation. Also, b_{k+1} does not exist.

By substituting $s_1 = e^{hg} + O(h^{k+2})$ into $C(s)$ and using the corrector equations corresponding to equations (15), we see that s_1 is indeed a root of $C(s)$ and thus an approximate root of (14).

We now return to writing down the complete expression for e_n for stable operators. To determine the particular solution of the non-homogeneous equation, we put $e_{n-i,m} = e$ and $e_{n-i,m-1} = \tilde{e}$, a constant, for all i and eliminate $\tilde{e}$ from equation (10) and the last equation of (9), finally

obtaining an expression for e_n. Since we are assuming that the operator is stable, we shall neglect the effects of the roots other than s_1. Also, we neglect the effect of starting errors by assuming that $e_0 = 0$. The resulting approximate expression for e_n is

$$e_n \simeq \left[\left(\frac{\theta^{m-1}(1-\theta)hg \sum_{1}^{k+1} b_i^*}{(1-\theta^m)} \right) \frac{T+r}{hg \sum_{i=0}^{k} b_i} + \frac{\theta^{m-1}(1-\theta)}{(1-\theta^m)} (T^*+r^*) \right] (s_1^n - 1) \quad . \qquad (16)$$

Since T^* and T are $O(h^{k+2})$, they are negligible with respect to r and r^* for small h. This region of h is called the round-off region. We shall see later that in this region, the error can be minimized by using proper computational techniques. As h becomes large, T^* and T become the more important terms. This region of h for stable operators is called the truncation region. For large m, e_n is approximately $O(h^{k+1})$ in the truncation region.

It is important to determine precisely the conditions for a method to be stable. Dahlquist's theorem provides us with a condition on the degree of a method if it is to be stable. In general, we cannot increase h indefinitely and still maintain stability.

For $m = \infty$, the stability requirement places a condition on the root of $C(s)$ of maximum absolute value, say s^*. For given constants a_i, b_i and k, $C(s)$ determines $|s^*|$ as a function of hg. An example of a plot of $|s^*|$

versus hg is given in Appendix I . For $hg > 0$, $s^* = s_1$, the expected root, in the region of small hg . However, as hg decreases in the negative direction, s^* becomes one of the extraneous roots, $|s^*|$ becomes greater than 1, and instability occurs. We say that the hg-region of stability is (d,∞) .

A graph such as that in Appendix I is a practical aid in determining stability. During the course of solving a differential equation, g is calculated - after each step if necessary - and h is chosen so that the value hg lies in the region (d,∞) .

For $m \neq \infty$, and the method $P(EC)^m$, the determination of the hg-region of stability is as above, except that the polynomial (14) is used instead of $C(s)$. For the case $PE(CE)^m$ and finite m , the polynomial is even simpler. It results from equation (10) with $e_{n-i,m-1}$ replaced by $e_{n-i,m}$. It turns out to be

$$sC(s) + \theta^m \sum_{i=1}^{k+1} (a_i - a_i^* + \theta a_i^* + hg(b_i - b_i^* + \theta b_i^*))s^{k+1-i} . \qquad (17)$$

The hg-regions of stability for values of $k=1,\ldots,8$ and certain values of m for procedures based on the formulas of Adams type are shown in Appendix II. The Adams formulas are defined by taking $a_1^* = a_1 = 1$, but otherwise $a_i^* = a_i = 0$. The predictor then becomes the Adams-Bashforth formula and the corrector, the Adams-Moulton.

We note three characteristics of the table in Appendix II. For given m , d increases as k goes from $k=2$ to $k=8$.

For given k and m, the hg-region for $PE(CE)^m$ contains the hg-region for $P(EC)^m$ and $P(EC)^{m+1}$. Thirdly, the hg-region for $m = \infty$ is the largest for each given k.

To conclude this section, we shall investigate the extent to which the preceding results for $N=1$ can be generalized.

It is shown in Appendix III, that e_n can be represented as the first N components of the vector $A^n y_0$ where A is an Nk by Nk matrix and y_0 is the 'initial-condition' vector. The growth of e_n is determined by the eigenvalues of A. As shown in Appendix III, $e^{h\lambda} + O(h^{k+2})$ where λ is any eigenvalue of G, is an eigenvalue of A. These are to be expected as before. Therefore, we define stability in a manner analogous to the case $N=1$.

Let $\lambda_1 = e^{h\lambda}$ be the eigenvalue of e^{hG} of maximum absolute value ($|\lambda_1|$ is called the spectral radius of e^{hG}). Let the other eigenvalues of A be denoted by λ_i. We say a method is stable if $|\lambda_i| < |\lambda_1|$ for $|\lambda_1| > 1$ and if $|\lambda_i| < 1$ for $|\lambda_1| \leq 1$.

The definition of stability is rather arbitrary. Stability should be defined to suit the requirements placed on the numerical solution. Unless we note otherwise, we shall use the above definition which is suitable for our purposes. If a method satisfies this definition of stability, the relative error of the solution vector of the system of equations, as measured by our norm for example, will remain small. However, the above definition does not ensure the smallness of the relative error of each component of the

numerical solution. If we replace λ_1 in the above definition by the eigenvalue of e^{hG} of minimum absolute value, we obtain a stronger stability condition. This new condition is sufficient - but not necessary, as we shall see later - to ensure the smallness of the relative error of each component of the solution.

We define absolute stability by saying that a method is absolutely stable if the extraneous eigenvalues λ_i satisfy $|\lambda_i| < 1$. Another quite restrictive definition of stability requires that the λ_i satisfy $|\lambda_1| < |\lambda_2|$, where λ_2 is the eigenvalue of e^{hG} of minimum absolute value, even if $|\lambda_2| < 1$.

It turns out that for a stable method, an approximate expression for e_n is

$$e_n \simeq (S_1^n - I)\left[(I-\theta^{m-1}(I-\theta)(I-\theta^m)^{-1} hG \sum_1^{k+1} b_i^*) \frac{G^{-1}(T+r)}{h\sum_0^k b_i}\right.$$

$$\left. + \theta^{m-1}(I-\theta)(I-\theta^m)^{-1}(T^*+r^*)\right] \qquad (18)$$

where $S_1 = e^{hG} + O(h^{k+2})$.

Noting the similarity between equations (18) and (16), we define the round-off and truncation regions for the general case in the same way as before. The right hand side of (18) provides an estimate for $\|e_n\|$.

For Adams method, the region of stability can be determined if the eigenvalues of A are determined in terms of h. Methods for finding these eigenvalues are given in Appendix III. In general, these eigenvalues are difficult

to determine, particularly if the matrix G varies considerably during the course of a calculation. It is quite impractical to calculate the eigenvalues of A with any regularity while the method is used. However, a preliminary study of the system might yield pertinent information on the λ_i. For instance, if G is a constant matrix over the domain of t under consideration, the λ_i can be predetermined for a given k and h, giving the conditions on h necessary for stability.

ECONOMY CONSIDERATIONS

In deciding which numerical methods to use for solving (1), the cost of such procedures should be taken into account. We shall compare the accuracy of methods that cost the same. Since we are assuming that f is relatively complicated, the cost of a procedure is proportional to the number of times f is evaluated. Thus, m/h is a reasonable measure of cost for the predictor-corrector procedures.

A typical plot of the relative error that results when Adams method is used versus h is given in Appendix IV. In the round-off region, the relative error is quite constant. As h increases through the truncation region, the relative error increases. Finally, instability occurs. Depending on the accuracy required, we would like to choose h so that the method is operating in the extreme left hand side of the truncation region. The problem then becomes one of determining the most practical m. Obviously, we would like to keep m as small as possible.

We assume that T^* and T are the dominant terms in (18). It may happen that h is so small that the first term of (18) dominates. In this case, increasing m has little effect in changing e_n. Otherwise, the second term dominates. In this case it is better to decrease h, thus decreasing the magnitude of the terms of θ^{m-1} and of T^*, rather than increasing m. In any case, it is clear that we need consider only small values of m.

Before making a specific choice of m, we should con-

sider the stability question. For the case N=1 and Adams method, it is clear from the values in Appendix II that the PEC method is rather unstable for general purposes. The stability regions for cases with more evaluations per step than PECEC are not considerably magnified. Bearing in mind that we wish to keep m small, we must therefore choose between the methods PECE and PECEC.

The procedure PECE has a larger region of stability than PECEC for each given k . However, it may be that the extra accuracy obtained by the added correction - which costs practically nothing - warrants the use of PECEC over PECE .

The stability question for the general case is harder to discuss since we do not have tables such as those in Appendix II. However, we expect stability to behave the same qualitatively as in the simple case.

EXPERIMENTAL TECHNIQUES

Because the round-off error r appears where it does in equation (16), it can be of considerable importance unless its magnitude is kept small. The use of double-precision arithmetic decreases the magnitudes of the individual round-off errors. However, round-off effects are minimized more simply and more economically by using 'partial double-precision' ([6], p. 94). The values of y_n in (2) and (3) are stored in double-precision and the additions of the terms involving h are done in double-precision. But the terms in (2) and (3) involving h are calculated in single-precision. For instance, if Z_{n-i} denotes the single-precision rounded value of the number y_{n-i}, then partial double-precision for the Adams corrector is defined by

$$y_n = y_{n-1} + h \sum_{i=0}^{k} b_i Z'_{n-i} \quad ,$$

where the second term is left unrounded and is added in its entirety to y_{n-1}.

This procedure is successful because the method involves the addition of small terms to relatively larger ones. In fact, h can now be decreased to a considerably lower value than in the single-precision case, without fear of accumulation of round-off error.

A measure of the local truncation error - the truncation error from a single step of a calculation - can be found by experimental methods. First, we write

$$T_p = C_p h^{k+2} \quad \text{and} \quad T_c = C_c h^{k+2}$$

where T_p and T_c are the truncation errors due to the particular applications of the predictor and corrector respectively at the point t_n, and

$$C_p = \frac{R^* x^{(k+2)}}{(k+1)!} + O(h) \quad \text{and} \quad C_c = \frac{R x^{(k+2)}}{(k+1)!} + O(h) .$$

Therefore, neglecting round-off errors,

$$(y^p - y^c) = (C_p - C_c) h^{k+2}$$

where y^p is the y_n calculated from (2) and y^c the value of y_n calculated from (3). Thus,

$$\|T_c\| = \frac{\|C_c\|}{\|C_p - C_c\|} \|y^p - y^c\| \simeq \left| \frac{R}{R^* - R} \right| \cdot \|y^p - y^c\| . \quad (19)$$

R and R^* depend on the a_i and b_i, a_i^* and b_i^*. For Adams method and $k=6$, for example, $R/(R^*-R) = -1375/38174$.

It is a simple matter to calculate this approximation to $\|T_c\|$ at any stage of the procedure. In practice, the maximum local error to be allowed would be prescribed. During a calculation, h would be chosen so that the measure of the truncation error given by (19) is less than the prescribed maximum.

We shall compare the results obtained by the application of various predictor-corrector methods to the popular Runge-Kutta method defined by

$$y_{n+1} = y_n + \frac{1}{6}(k_1+2k_2+2k_3+k_4)$$

where

$$k_1 = hf(t_n, x_n)$$

$$k_2 = hf(t_n+h/2,\ y_n+k_1/2)$$

$$k_3 = hf(t_n+h/2,\ y_n+k_2/2)$$

$$k_4 = hf(t_n+h,\ y_n+k_3)$$

The truncation error of this method is $O(h^5)$. We note that this method involves four evaluations of f per step and thus costs the same as the PECEC method with a step-size of $h/2$.

Before concluding this section, we note that in Appendices V, VI, VII and VIII, the step-sizes for the Runge-Kutta procedure are chosen so that the methods in the same row of the tables cost the same. That is, for the PEC table, the Runge-Kutta step-size is equal to four times the predictor-corrector step-size. For the PECE and PECEC tables, the Runge-Kutta step-size is twice the predictor-corrector step-size. In these Appendices, h denotes the predictor-corrector step-size.

EXPERIMENTAL RESULTS

We tested the preceding theory on four separate problems. We considered only systems of equations. The first two problems are simple examples of circular motion [7]. Problem (A) is linear whereas (B) is nonlinear. Each has the same trigonometric solution. For problem (C) we chose a linear system of equations with an exponential solution. The equations in problem (D) are the equations of motion of a vehicle reentering the earth's atmosphere [8]. A theoretical solution of the last problem is not known - as is generally the case in actual practice. Therefore, problem (D) was solved under realistic conditions.

We restricted ourselves to procedures using formulas of Adams type only. Much of our theory is based on these formulas. Actually, procedures using Adams formulas are as reliable as most of the other procedures and are generally considered to be representative 'general-purpose' procedures.

We considered procedures based on the iteration method $P(EC)^m$ for $m=1,2,3$ and on $PE(CE)^m$ for $m=1,2$. In each of these cases, we took $k=4,5,6,7$. Problems (A) and (B) were run over the domain $t=0$ to $t=10\pi$, problems (C) and (D) (after normalization) over $t=0$ to $t=30$. For each m and k we used $h=2^0,2^{-1},\ldots,2^{-7}$.

Some of the results for problem (A) are representative of the results for the other problems so we shall discuss (A) in detail. Then we will note where the results from the other problems differ.

Problem (A) is defined by

$$\begin{aligned} x_1'(t) &= x_2 & x_1(0) &= 1 \\ x_2'(t) &= -x_1 & x_2(0) &= 0 \\ x_3'(t) &= x_4 & x_3(0) &= 0 \\ x_4'(t) &= -x_3 & x_4(0) &= 1 \quad . \end{aligned} \qquad (A)$$

It is easy to see that the solution is given by

$$\begin{aligned} x_1(t) &= \cos t \\ x_2(t) &= -\sin t \\ x_3(t) &= \sin t \\ x_4(t) &= \cos t \quad . \end{aligned}$$

The matrix G is a constant matrix and its eigenvalues are i and $-i$, each repeated. We note that all the eigenvalues of e^{hG} have absolute value equal to one. Thus, if another eigenvalue of A has absolute value greater than one, instability will occur.

We use the norm

$$E = \|e\| = |e_1| + |e_2| + |e_3| + |e_4| \quad ,$$

where e_i is the error of the i^{th} component of the numerical solution, as a measure of the error.

The table of maximum errors corresponding to this problem is given in Appendix V. A plot of maximum error versus h is given in Appendix IV for the special case $k=6$ and

the PECEC method. In each instance, the maximum error occurred very near $t = 10\pi$.

We first note that for each k and h, the error in the round-off region is approximately constant. In the case $k=7$ and the PECEC method, the round-off error varies only slightly over the domain $h = 2^{-10}$ to $h = 2^{-4}$. This is a result of the partial double-precision technique.

Instability occurs for each method and each value of k . The transition between the truncation region and the instability region is sharply defined in each case. However, instability arises at different values of h for different methods. For a given k, instability appears at smaller values of h for the PEC method than for other methods. For the PEC method, the h at which instability occurs decreases rather markedly as k increases; however, for methods other than PEC , this phenomenon does not appear. In fact, no indication as to which k is best exists for methods other than PEC .

For the methods PECE to $P(EC)^3$, the round-off regions and the errors in these regions are very nearly the same. Also, the increase of the errors through the truncation regions in these cases behaves approximately the same, although the error rises slightly faster for $k=4$ than for $k=5,6,7$. However, instability appears for smaller h in the cases PECE and PECEC than in the cases $PE(CE)^2$ and $P(EC)^3$ (it occurs for slightly smaller h in the case PECEC than in PECE). But since we are interested in using values of h such that the method operates in the left hand

side of the truncation region, the PECE and PECEC methods are as suitable for this problem as any other method.

For methods other than PEC , and each value of k , the error in the truncation region does not increase as quickly as expected. Assuming that the terms involving θ in equation (18) can be neglected, we expect the error in this region to be approximately $O(h^{k+1})$. However, since m is small, it might be that our assumptions about θ are not strictly true. For this problem and all methods, the truncation region is not particularly wide.

Comparing the predictor-corrector methods PECE and PECEC to the Runge-Kutta method of the same cost (the Runge-Kutta method with twice the step-size), we see first that in both of their round-off regions, the error is approximately the same. However, the truncation region for the Runge-Kutta method starts at much smaller h than for the other methods. From the value of h at which the truncation region of the Runge-Kutta method starts to the points where instability of the predictor-corrector methods occurs, the error of the Runge-Kutta method is considerably higher than that of the predictor-corrector methods. And it is precisely in this region that we wish to operate our predictor-corrector method, thus obtaining the maximum accuracy for the minimum cost.

Problem (B) is

$$x_1'(t) = x_2 \qquad x_1(0) = 1$$

$$x_2'(t) = -x_1 r^{-3} \qquad x_2(0) = 0$$

$$x_3'(t) = x_4 \qquad x_3(0) = 0$$
$$x_4'(t) = -x_3 r^{-3} \qquad x_4(0) = 1 \qquad \text{(B)}$$

where $r = \sqrt{x_1^2 + x_3^2}$. The solution is the same as in (A) .

Even though (A) and (B) are similar problems and have identical solutions, we cannot expect the experimental results to be the same for both problems, for the terms involving r in (B) affect the eigenvalues of G . The characteristic equation for G is not difficult to construct. However, it depends on x_1 and x_3 and is quite complicated. The results for (B) are tabulated in Appendix VI.

The instability regions for methods other than PEC are not well-defined for problem (B). In fact, it appears as if instability does not occur for these methods.

The errors in all the round-off regions are approximately the same. The truncation region for the PEC method starts at lower values of h than for other methods, as is the case with (A). Also, for all given methods other than PEC , the truncation region for $k=4$ starts at smaller h than for $k=5,6,7$. For the latter methods, the truncation regions for (B) start at approximately the same h as those for (A). Again, the truncation error does not rise as fast as expected.

The predictor-corrector methods used on (B) compare with the Runge-Kutta method in essentially the same way as

before.

For problem (C), we chose

$$\begin{aligned} x_1'(t) &= x_2 & x_1(0) &= 1 \\ x_2'(t) &= x_1 & x_2(0) &= 0 \\ x_3'(t) &= x_4 & x_3(0) &= 0 \\ x_4'(t) &= x_3 & x_4(0) &= 1 \quad . \end{aligned} \qquad (C)$$

The solution is

$$\begin{aligned} x_1(t) &= \tfrac{1}{2}(e^t+e^{-t}) = x_4(t) \\ x_2(t) &= \tfrac{1}{2}(e^t-e^{-t}) = x_3(t) \quad . \end{aligned}$$

The eigenvalues of G are 1 and -1, each repeated.

We use the same norm as before and we let $E/2(e^t)$ be a measure of the relative error. The table of maximum relative errors is given in Appendix VII. For each method, the maximum relative error occured at $t=30$.

The eigenvalues of e^{hG} for this problem are e^h and e^{-h}. Since $e^h > 1$, we expect at least one component of the error to behave like a positive exponential. Also, it is reasonable to expect e^h to be the eigenvalue of the matrix A of maximum absolute value for most methods. Thus, these methods would be stable.

The table implies that instability does not occur for methods other than PEC. And since each component of the solution is proportional to e^t, each component has a small relative error. We note, however, that a simple transformation

changes equations (C) into a system in which one of the components has a solution equal to e^{-t}. By its very nature, this component would not have a small relative error.

The truncation region for methods with $k=4$ other than PEC start at smaller h than for methods with $k=5,6,7$. The truncation errors for $k=5,6,7$ are similar. The methods PECE and PECEC are considerably better than the corresponding Runge-Kutta method.

Finally, problem (D) is defined by

$$x_1'(s) = -\frac{C_D S}{m}\rho x_1 - 2(g_r \sin x_3 - g_\lambda \cos x_3 \cos x_2 - g_\mu \cos x_3 \sin x_2)$$

$$- 2(\omega^2 x_6 \cos x_4)(\sin x_3 \cos x_4 + \cos x_3 \sin x_4 \cos x_2)$$

$$x_2'(s) = \frac{C_L S}{2m}\rho \frac{1}{\cos x_3} + \frac{\cos x_3 \sin x_2 \sin x_4}{x_6 \cos x_4}$$

$$- \frac{1}{x_1 \cos x_3}(g_\lambda \sin x_2 - g_\mu \cos x_2)$$

$$+ \frac{\omega^2 x_6 \cos x_4 \sin x_4 \sin x_2}{x_1 \cos x_3} + \frac{2\omega}{\sqrt{x_1}}\left(\frac{\sin x_3 \cos x_2 \cos x_4}{\cos x_3} + \sin x_4\right)$$

$$x_3'(s) = -\frac{C_L S}{2m}\cos\varphi \cdot \rho - \left(\frac{1}{x_6} + \frac{g_r}{x_1}\right)\cos x_3$$

$$- \frac{1}{x_1}(g_\lambda \sin x_3 \cos x_2 + g_\mu \sin x_3 \sin x_2) - \frac{2\omega}{\sqrt{x_1}}\cos x_4 \sin x_2$$

$$- \frac{\omega^2 x_6 \cos x_4}{x_1}(\cos x_3 \sin x_4 - \sin x_3 \sin x_4 \cos x_2)$$

$$x_4'(s) = \frac{\cos x_3 \cos x_2}{x_6}$$

$$x_5'(s) = \frac{\cos x_3 \sin x_2}{x_6 \cos x_4}$$

$$x_6'(s) = -\sin x_3 \quad , \qquad \text{(D)}$$

where $\rho = \rho_0 e^{-\beta y}$

$$y = x_6 - R_0\left(1 - f\sin^2 x_4 - \frac{f^2}{2}\left(\frac{R_0}{6} - \frac{1}{4}\right)\sin^2 2x_4\right)$$

$$g_r = -\frac{GM}{(x_6)^2}\left(1 - \frac{3C_{20}}{2}\left(\frac{R_0}{x_6}\right)^2 (3\sin^2 x_4 - 1)\right)$$

$$g_\lambda = \frac{3GM}{(x_6)^2} C_{20} \left(\frac{R_0}{x_6}\right)^2 \sin x_4 \cos x_4$$

$$g_\mu = 0 \quad .$$

A physical interpretation of the problem and its parameters is given in Appendix IX. The important variables are x_1 , x_3 and x_6 . We normalized the problem by putting $t = 3s \times 10^{-5}$. The domain of t allowed the vehicle to make more than one complete 'skip' through the earth's outer atmosphere.

Each component of the solution to (D) behaved as expected. For all methods, a region of h existed in which each component was constant; these respective constants were the

same for all methods. These regions were the round-off regions. Similarly, truncation regions existed.

As a measure of the error in each x_i, we used the deviation $|x_i-\tilde{x}_i|$ where $\tilde{x}_i$ was the solution in the round-off region. The greatest relative deviations were observed for x_3. The results for the other x_i were very similar. The maximum deviations for x_3 are given in Appendix VIII.

We observe that the PEC method gives results that are similar to those of the previous problems. Instability occurs only for the PEC method with $k=6,7$. The PECE and PECEC methods yield deviations that are as small as the corresponding deviations for the other predictor-corrector methods and considerably smaller than the corresponding deviations for the Runge-Kutta method.

CONCLUSIONS

In general, we have shown that predictor-corrector methods can be used effectively to generate accurate numerical solutions to systems of ordinary differential equations.

We have described the theory and usage of predictor-corrector procedures in detail. We considered methods of the form $P(EC)^m$ and $PE(CE)^m$ for different values of k. We have discussed the concept of stability for single equations and have generalized this concept for systems of equations. We have obtained an approximate expression for the error vector for stable methods and we have discussed the behaviour of the error for varying step-sizes.

In particular, we have investigated theoretically the behaviour of the error for Adams method applied to the case $N=1$. For a given method, the error in the truncation region is lowered as k increases. However, the table in Appendix II indicates that if $g < 0$, instability will occur at smaller h as k becomes larger. On the other hand, for either $P(EC)^m$ or $PE(CE)^m$ and a given k, similar reasoning indicates that instability will occur at larger h as m increases. The experimental results in [8] exhibit this behaviour. Finally, for a given m and k, the method $PE(CE)^m$ is stable over a wider range of h than either $P(EC)^m$ or $P(EC)^{m+1}$. On the basis of these facts and the additional requirement that m be kept small, we decided that the methods PECE and PECEC with the intermediate values of $k=5,6$ are the most efficient for $N=1$.

For the general case, the stability conditions are not easy to analyse. However, from the results for our four different and representative problems, we are able to draw certain conclusions.

The PEC method yielded the poorest accuracy and was quite unstable, especially for large values of k. The $PE(CE)^2$ and $P(EC)^3$ methods were not significantly better than the methods PECE and PECEC for any of the problems. The results from the PECE and PECEC methods were quite similar. However, for $k=4$, the truncation region started at lower values of h than for $k=5,6,7$.

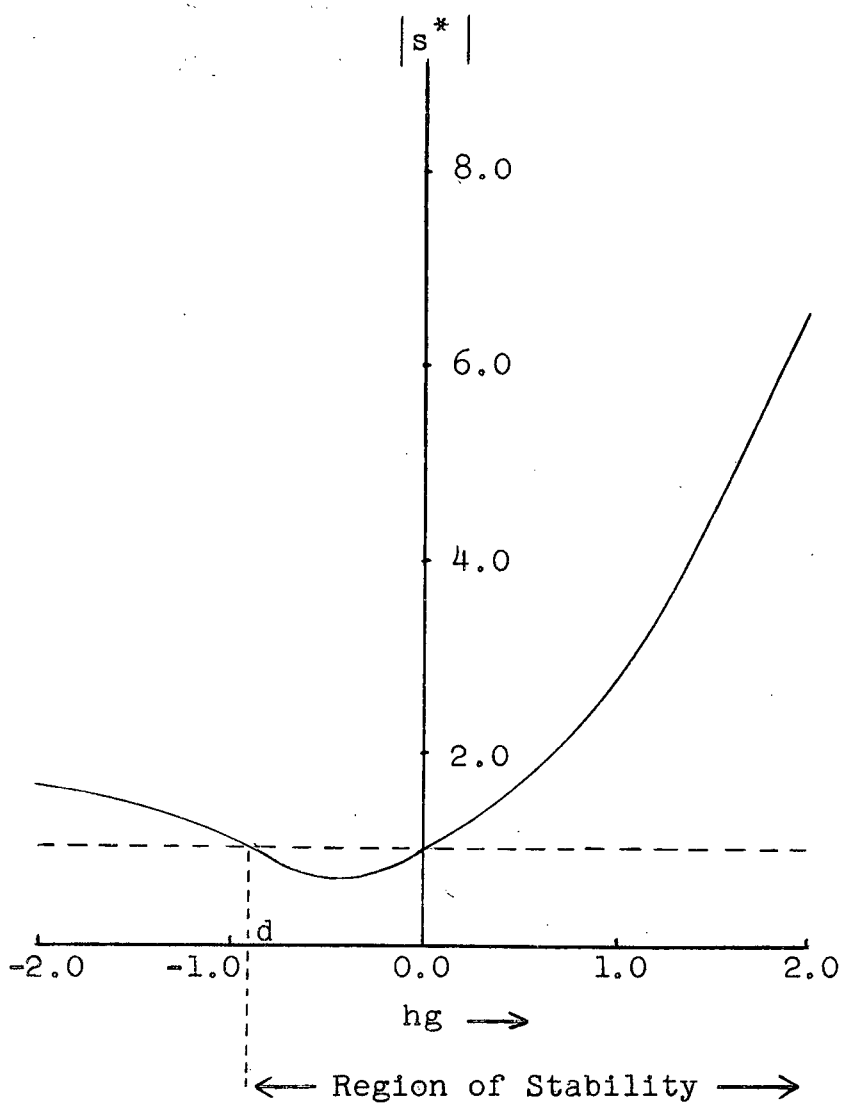
Of the methods considered, the PECE and PECEC methods with $k=5,6,7$ are the best for general purposes. For the problems we investigated, these methods gave considerably better results than the Runge-Kutta method of the same cost.

It is possible that significant results may be obtained by investigating methods using larger values of k. The possibility of determining the eigenvalues of the matrix A and controlling stability during the course of a calculation should be investigated.

APPENDIX I

$|s^*|$ vs. hg

(Adams PECEC, k=3)



APPENDIX II

The hg-Regions of Stability

Method / k	PEC	PECE	$P(EC)^2$	$PE(CE)^2$	$P(EC)^3$	$P(EC)^\infty$
1	-1.00	-1.37	+0.60	+0.38	+0.50	$-\infty$
2	-0.30	-1.70	-1.13	-1.25	-1.00	$(-\infty)$
3	-0.15	-1.25	-0.87	-1.10	-0.87	$(-\infty)$
4	-0.	-1.00	-0.62	-0.87	-0.70	-1.80
5	-0.	-0.70	-0.50	-0.70	-0.55	-1.13
6	-0.	-0.50	-0.38	-0.50	-0.45	-0.75
7	-0.	-0.38	-0.25	-0.38	-0.35	-0.50
8	-0.	-0.30	-0.20	-0.25	-0.25	-0.35

The values of d, where (d,∞) is the region of stability, are plotted. The equations used are

$$P(EC)^m: \quad s^{k+1}C(s)+\theta^{m-1}\left\{\theta s^k\left[\sum_1^k a_i s^{k+1-i}-(1-\theta)\sum_1^{k+1} a_i^* s^{k+1-i}\right]\right.$$

$$+hg\left[s^{k+1}\sum_1^k b_i s^{k-i}+(1-\theta)\left(\sum_1^k a_i s^{k-i}\cdot\sum_1^{k+1} b_i^* s^{k+1-i}\right.\right.$$

$$\left.\left.\left.-\sum_1^{k+1} a_i^* s^{k+1-i}\cdot\sum_1^k b_i s^{k-i}-s^k\sum_1^{k+1} b_i^* s^{k+1-i}\right)\right]\right\} \quad = \quad 0$$

$$PE(CE)^m: \quad sC(s)+\theta^m\sum_1^{k+1}(a_i-a_i^*+\theta a_i^*+hg(b_i-b_i^*+\theta b_i^*))s^{k-i+1} \quad = \quad 0$$

$$m=\infty: \quad C(s) \quad = \quad (1-\theta)s^k-\sum_1^k(a_i+hgb_i)s^{k-i} \quad = \quad 0 \quad ,$$

where for Adams method, $a_1 = a_1^* = 1$ and otherwise $a_i = 0$. $\theta = hb_0 g$.

APPENDIX III

The Generalized Error Equation

For the general case $(N \neq 1)$, the error equation is (11) for $m=\infty$ and a more complicated equation determined by (10) and the last equation of (9) for finite m. We rewrite the homogeneous equation in the form

$$A_1 e_{n+k} + A_2 e_{n+k-1} + \ldots + A_{k+1} e_n = 0 \qquad (20)$$

where, for $m=\infty$,

$$A_1 = I - hb_0G$$

$$A_2 = -I - hb_1G$$

$$A_i = -hb_{i-1}G \quad , \quad i=3,\ldots,k+1 \quad .$$

The A_i are polynomials in G for $m \neq \infty$.

The transformation

$$y_{n1} = e_n$$

$$y_{n2} = e_{n+1}$$

$$\ldots\ldots$$

$$y_{nk} = e_{n+k-1}$$

$$y_n = \begin{pmatrix} y_{n1} \\ y_{n2} \\ \vdots \\ y_{nk} \end{pmatrix} \qquad y_o = \begin{pmatrix} e_0 \\ e_1 \\ \vdots \\ e_{k-1} \end{pmatrix}$$

results in the equation $y_n = A^n y_0$, where

$$A = \begin{pmatrix} \emptyset & I & \emptyset & \emptyset & \ldots. & \emptyset \\ \emptyset & \emptyset & I & \emptyset & \ldots. & \emptyset \\ . & . & . & . & \ldots. & . \\ \emptyset & \emptyset & \emptyset & \emptyset & \ldots. & I \\ -A_1^{-1}A_{k+1} & & . & . & & -A_1^{-1}A_2 \end{pmatrix}$$

(assuming that A_1 is nonsingular) and I and $\emptyset$ are the N by N identity and zero matrices respectively.

The growth of e_n - which is represented as the first N components of y_n - is determined by the eigenvalues of A. For example, if all of the eigenvalues of A are less than one in absolute value, $A^n \longrightarrow$ the zero matrix as $n \longrightarrow \infty$. The eigenvalues λ of A satisfy the characteristic equation

$$\det(a_1\lambda^k + A_2\lambda^{k-1} + \ldots + A_{k+1}) = 0 \quad . \qquad (21)$$

Returning to equation (20) with the A_i for $m = \infty$, we replace e_{n+i} with the $(n+i)^{th}$ power of the matrix

$$S_1 = e^{hG} + O(h^{k+2}) \quad .$$

Using the corrector equations corresponding to (15) and the relation

$$e^Q = I + \sum_{j=1}^{\infty} \frac{Q^j}{j!} \quad ,$$

we see that S_1 satisfies

$$A_1 S_1^{n+i} + A_2 S_1^{n+i-1} + \ldots + A_{k+1} = 0 \quad .$$

Therefore, S_1 satisfies the scalar equation $g(S_1) = 0$ where

$$g(\lambda) = \det(A_1\lambda^k + \ldots + A_{k+1})$$

(see [5], p. 228). Now, if T is a matrix such that $T^{-1}S_1T$ is the Jordan cononical form of S_1, the equation $T^{-1}g(S_1)T = 0$ implies that each eigenvalue of S_1

satisfies $g(\lambda) = 0$. That is, each eigenvalue of S_1 is an eigenvalue of A. However, the eigenvalues of S_1 are of the form $e^{h\lambda_i} + O(h^{k+2})$ where λ_i is an eigenvalue of G.

Therefore, $S_1^n \bar{E}$ where $\bar{E}$ is a constant vector, satisfies (20). The complete solution of (20) can be written as $S_1^n \bar{E}$ plus other terms that depend on A and whose behaviour depends on the eigenvalues of A. For stable methods, the solution of (20) is approximately $S_1^n \bar{E}$.

Under special circumstances, the above analysis can be simplified. We assume that there exists a matrix T such that $T^{-1}GT = D$, a diagonal matrix with the N eigenvalues of G (possibly repeated) as the diagonal elements. We let $e_{n+i} = Tv_{n+i}$, $i=0,\ldots,k$ and substitute these expressions into equations (20). Premultiplying (20) by T^{-1} and noting the special form of the A_i, we obtain a system of N uncoupled difference equations, each of the components of v_{n+i} appearing in only one of the equations. Each of these equations may be solved by the usual method. However, $P_1 = e^{hD} + O(h^{k+2})$ is a solution of the resulting system of equations and thus $e^{h\lambda_i} + O(h^{k+2})$ is a characteristic root of the i^{th} difference equation. (A sufficient condition for the smallness of the relative error of each component of v_n can be found by applying the stability condition for $N=1$ to each difference equation.) For stable systems, v_n is approximately $P_1^n \bar{F}$ for some vector $\bar{F}$ and thus e_n is approximately $TP_1^n \bar{F} = S_1^n \bar{E}$ as before.

To write down the complete expression for e_n, we pro-

ceed as in the case $N=1$. We first find the particular integral of the nonhomogeneous equation. We put $e_{n-i,m} = E$ and $e_{n-i,m-1} = \tilde{E}$, both constant vectors, for all i and eliminate $\tilde{E}$ from (10) and the last equation of (9) . We obtain the expression

$$E = (-I+\theta^{m-1}(I-\theta)(I-\theta^m)^{-1}hG \sum_1^{k+1} b_i^*) \frac{G^{-1}(T+r)}{h\sum_0^k b_i}$$

$$- \theta^{m-1}(I-\theta)(I-\theta^m)^{-1}(T^*+r^*) .$$

We then assume that e_n can be written approximately as $S_1^n \bar{E} + E$, and, neglecting starting errors by putting $e_0 = 0$, we find that $\bar{E} = -E$. The final expression for e_n is given by (18).

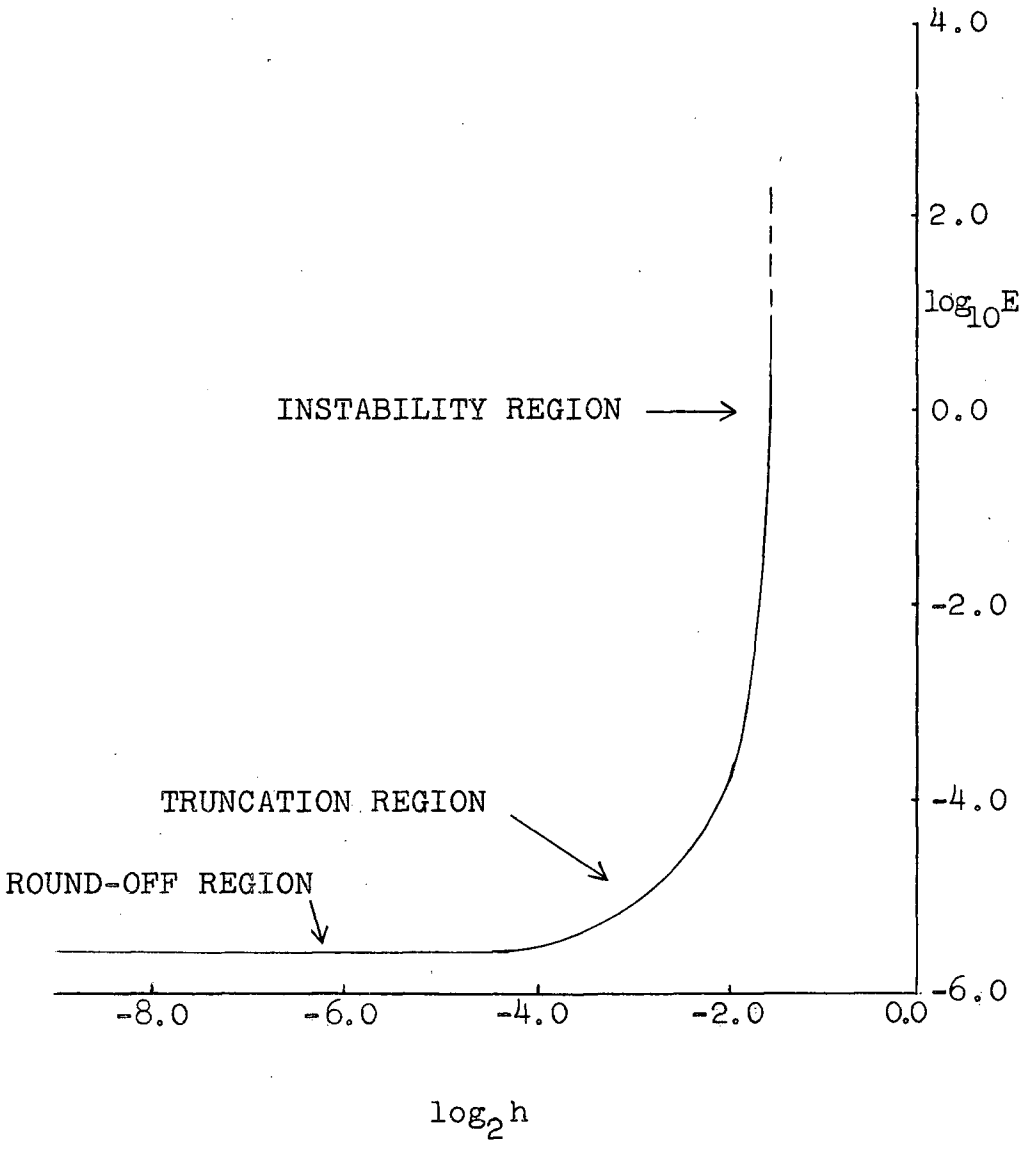
To determine the eigenvalues of A , we use equation (21) or other methods that take advantage of the special form of the A_i .

We note that the above analysis specializes to the analysis for single equations when $N=1$.

APPENDIX IV

$\log_{10}$(Error) vs. $\log_2$(h)

(PECEC, k=6, Problem (A))



APPENDIX V

Maximum Errors - Problem (A)

The units of E are 10^{-6}.

	h \ k	4	5	6	7	Runge-Kutta
PEC	2^{-7}	1.669	2.101	2.488	2.131	1.952
	2^{-6}	1.654	2.116	2.503	—	12.636
	2^{-5}	1.669	2.146	—	—	177.920
	2^{-4}	1.714	8259.982	—	—	2978.951
	2^{-3}	3441.267	—	—	—	47223.382
	2^{-2}	—	—	—	—	610237.444
	2^{-1}	—	—	—	—	—
	2^{0}	—	—	—	—	—
PECE	2^{-7}	1.654	2.086	2.474	2.161	1.326
	2^{-6}	1.654	2.086	2.474	2.146	1.952
	2^{-5}	1.684	2.101	2.488	2.161	12.636
	2^{-4}	2.086	2.235	2.593	2.310	177.920
	2^{-3}	56.505	10.304	7.302	7.197	2978.951
	2^{-2}	2542.719	582.278	266.694	174.493	47223.382
	2^{-1}	137785.740	55772.796	—	—	610237.444
	2^{0}	—	—	—	—	—
$P(EC)^2$	2^{-7}	1.669	2.101	2.474	2.176	1.326
	2^{-6}	1.654	2.116	2.488	2.176	1.952
	2^{-5}	1.699	2.116	2.459	2.176	12.636
	2^{-4}	2.444	2.265	2.563	2.339	177.920
	2^{-3}	52.944	10.595	6.924	7.123	2978.951
	2^{-2}	2204.105	553.556	198.469	223241.962	47223.382
	2^{-1}	111165.665	—	—	—	610237.444
	2^{0}	—	—	—	—	—
$PE(CE)^2$	2^{-7}	1.654	2.101	2.474	2.176	
	2^{-6}	1.654	2.116	2.474	2.161	
	2^{-5}	1.699	2.086	2.503	2.190	
	2^{-4}	2.444	2.280	2.593	2.310	
	2^{-3}	49.874	10.312	6.810	7.078	
	2^{-2}	1830.876	462.681	177.965	161.231	
	2^{-1}	73939.659	31200.409	14849.722	5329.117	
	2^{0}	—	—	—	—	
$P(EC)^3$	2^{-7}	1.654	2.101	2.474	2.176	
	2^{-6}	1.654	2.116	2.488	2.146	
	2^{-5}	1.699	2.086	2.503	2.205	
	2^{-4}	2.421	2.220	2.563	2.310	
	2^{-3}	48.175	10.103	6.869	7.093	
	2^{-2}	1634.471	424.854	168.204	161.350	
	2^{-1}	58997.743	26176.870	11493.862	4910.804	
	2^{0}	—	—	—	—	

A dash denotes an error greater than 1.

APPENDIX VI

Maximum Errors - Problem (B)

The units of E are 10^{-6}.

	h \ k	4	5	6	7	Runge-Kutta
PEC	2^{-7}	.909	5.670	2.787	4.195	10.371
	2^{-6}	.492	10.148	4.880	—	97.185
	2^{-5}	2.913	15.266	—	—	2193.421
	2^{-4}	112.526	—	—	—	68238.030
	2^{-3}	—	—	—	—	—
	2^{-2}	—	—	—	—	—
	2^{-1}	—	—	—	—	—
	2^{0}	—	—	—	—	—
PECE	2^{-7}	1.058	1.438	1.334	.529	5.618
	2^{-6}	.633	3.859	2.198	1.281	10.371
	2^{-5}	3.070	8.285	2.503	1.095	97.185
	2^{-4}	138.827	3.919	9.872	3.047	2193.421
	2^{-3}	3093.675	501.588	98.258	115.767	68238.030
	2^{-2}	6787.375	51689.416	10482.252	4979.767	—
	2^{-1}	—	—	—	—	—
	2^{0}	—	—	—	—	—
$P(EC)^2$	2^{-7}	.671	.656	1.207	2.138	5.618
	2^{-6}	1.192	.477	.596	4.716	10.371
	2^{-5}	5.242	2.876	.671	5.908	97.185
	2^{-4}	153.504	1.535	2.131	8.099	2193.421
	2^{-3}	4193.246	148.892	69.492	103.004	68238.030
	2^{-2}	122627.400	11467.785	2791.904	—	—
	2^{-1}	—	—	—	—	—
	2^{0}	—	—	—	—	—
$PE(CE)^2$	2^{-7}	.671	.656	1.229	2.213	
	2^{-6}	1.036	.559	.507	4.992	
	2^{-5}	4.150	2.496	1.192	5.610	
	2^{-4}	155.605	2.578	1.818	9.507	
	2^{-3}	4334.085	158.772	66.489	99.644	
	2^{-2}	139126.979	14779.426	2462.968	6070.942	
	2^{-1}	—	—	137551.375	503854.632	
	2^{0}	—	—	—	—	
$P(EC)^3$	2^{-7}	.671	.656	1.214	1.825	
	2^{-6}	1.013	.827	.507	4.575	
	2^{-5}	5.640	1.974	1.110	4.120	
	2^{-4}	156.805	1.810	4.299	8.129	
	2^{-3}	4403.017	159.815	67.525	97.781	
	2^{-2}	147231.624	16434.923	2359.815	5988.881	
	2^{-1}	—	—	191579.305	—	
	2^{0}	—	—	—	—	

A dash denotes an error greater than 1.

APPENDIX VII

Maximum Errors - Problem (C)

The units of $E/2(e^t)$ are 10^{-6}.

	h \ k	4	5	6	7	Runge-Kutta
PEC	2^{-7}	.675	.846	.932	.785	.736
	2^{-6}	.675	.871	.932	—	4.108
	2^{-5}	.662	.908	—	—	55.463
	2^{-4}	.724	497905.390	—	—	793.547
	2^{-3}	38.703	—	—	—	10275.044
	2^{-2}	—	—	—	—	104165.800
	2^{-1}	—	—	—	—	555766.100
	2^{0}	—	—	—	—	—
PECE	2^{-7}	.675	.846	.944	.797	.515
	2^{-6}	.675	.846	.932	.797	.736
	2^{-5}	.650	.859	.932	.797	4.108
	2^{-4}	.392	.895	.981	.858	55.463
	2^{-3}	2.661	1.840	2.257	2.441	793.547
	2^{-2}	120.224	63.718	48.374	49.098	10275.044
	2^{-1}	8950.645	4198.864	2344.865	1716.027	104165.800
	2^{0}	203513.730	137079.090	94853.131	69640.762	555766.100
$P(EC)^2$	2^{-7}	.675	.834	.944	.797	.515
	2^{-6}	.675	.846	.944	.797	.736
	2^{-5}	.650	.834	.932	.822	4.108
	2^{-4}	.209	.846	.981	.895	55.463
	2^{-3}	11.701	.834	2.195	.250	793.547
	2^{-2}	257.925	10.904	33.386	46.092	10275.044
	2^{-1}	1581.858	518.892	1034.155	1243.215	104165.800
	2^{0}	86833.257	68949.960	55101.161	46501.988	555766.100
$PE(CE)^2$	2^{-7}	.675	.834	.944	.797	
	2^{-6}	.675	.846	.944	.797	
	2^{-5}	.650	.834	.932	.821	
	2^{-4}	.209	.846	.981	.895	
	2^{-3}	12.425	.711	2.183	2.502	
	2^{-2}	322.808	24.101	30.663	45.578	
	2^{-1}	5424.528	268.096	527.024	1056.293	
	2^{0}	21536.469	6708.022	18865.521	25314.158	
$P(EC)^3$	2^{-7}	.675	.834	.944	.797	
	2^{-6}	.675	.846	.944	.797	
	2^{-5}	.650	.834	.932	.822	
	2^{-4}	.209	.846	.981	.895	
	2^{-3}	12.830	.699	2.183	.250	
	2^{-2}	355.569	30.246	29.473	45.344	
	2^{-1}	7237.439	1464.038	324.158	985.768	
	2^{0}	69446.628	17377.982	6117.562	18375.696	

A dash denotes a relative error greater than 1.

APPENDIX VIII

Maximum Deviations in x_3 - Problem (D)

The units are 10^{-8} radians.

	h \ k	4	5	6	7	Runge-Kutta
PEC	2^{-7}	.1	.0	.1	.0	1.1
	2^{-6}	.8	.7	.8	.7	1.5
	2^{-5}	.3	.6	.8	.1	.6
	2^{-4}	.3	.4	1.5	—	1.0
	2^{-3}	1.0	.5	7.8	—	11.1
	2^{-2}	2.4	.6	—	—	290.8
	2^{-1}	122.8	4128.0	—	—	6686.8
	2^{0}	11912.8	498316.8	—	—	
PECE	2^{-7}	.1	.0	.2	.1	.0
	2^{-6}	.8	.7	.7	.7	1.1
	2^{-5}	.2	.3	.3	.2	1.5
	2^{-4}	.3	.2	.2	.1	.6
	2^{-3}	.1	.1	.3	.3	1.0
	2^{-2}	.9	.9	.5	.5	11.1
	2^{-1}	24.2	10.6	10.8	4.8	290.8
	2^{0}	2191.5	921.0	71.9	431.7	6686.8
$P(EC)^2$	2^{-7}	.1	.0	.2	.0	.0
	2^{-6}	.8	.7	.7	.7	1.1
	2^{-5}	.2	.3	.3	.2	1.5
	2^{-4}	.2	.2	.1	.2	.6
	2^{-3}	.3	.4	.2	.0	1.0
	2^{-2}	.5	.9	.8	.8	11.1
	2^{-1}	36.1	16.6	3.5	.7	290.8
	2^{0}	528.5	575.7	593.4	319.0	6686.8
$PE(CE)^2$	2^{-7}	.1	.0	.2	.0	
	2^{-6}	.8	.7	.7	.7	
	2^{-5}	.2	.3	.3	.2	
	2^{-4}	.3	.2	.1	.1	
	2^{-3}	.3	.4	.0	.1	
	2^{-2}	.5	.8	.9	1.0	
	2^{-1}	31.8	11.8	2.6	2.5	
	2^{0}	210.0	518.9	267.7	36.4	
$P(EC)^3$	2^{-7}	.1	.0	.2	.0	
	2^{-6}	.8	.7	.7	.7	
	2^{-5}	.2	.3	.3	.2	
	2^{-4}	.3	.2	.1	.2	
	2^{-3}	.3	.4	.1	.3	
	2^{-2}	.5	.8	.9	1.0	
	2^{-1}	29.5	10.8	.3	3.4	
	2^{0}	651.7	527.7	129.0	87.4	

A dash denotes an overflow on the I.B.M. 7090.

APPENDIX IX

The Atmospheric Reentry Problem

The important variables of problem (D) are defined by

x_1 = the square of the velocity of the vehicle, relative to the earth;

x_2 = the azimuthal angle (measured from the north);

x_3 = the flight path angle (the angle of depression of the velocity vector from the "horizontal" through the vehicle;

x_4 = the latitude of the vehicle (positive north);

x_5 = the longitude (measured from Greenwich);

x_6 = the distance of the vehicle from the center of the earth;

$s = 3t \times 10^5$ = the length of the flight path .

In addition,

y = the distance of the vehicle from the earth's surface;

ρ = the density of the atmosphere;

g_λ , g_μ , g_r are the components of the earth's gravitational force in the coordinate system determined by x_4 , x_5 and x_6 respectively.

These variables are expressed in terms of the following parameters:

m = the mass of the vehicle;

ω = the angular velocity of the earth relative to inertial space;

C_D = $C_D(\alpha)$, the drag coefficient;

C_L = $C_L(\alpha)$, the lift coefficient;

α = the angle of attack of the vehicle above the zero lift line;

S = the reference area of the vehicle;

φ = the angle of bank of the vehicle .

The constants are defined by

β^{-1} = 23,500 ft.;

ρ_0 = 0.0027 slugs/ft.^{3} ;

R_0 = the equatorial radius of the earth = 20925840 ft.;

f = the earth flattening factor = 1/298.28;

GM = the gravitational constant

= 1.4076536 x 10^{16} ft.^{3}/sec.^{2} ;

C_{20} = -.00108248 .

We used the realistic values C_DS/m = .977 , $C_LS/2m$ = .489 and $\varphi = \pi/3$ radians.

As initial conditions, we used

$x_1(0)$ = (25960 ft./sec.)^2

$x_2(0)$ = $\pi/4$ radians

$x_3(0)$ = 0.06 radians

$x_4(0)$ = $\pi/6$ radians

$x_5(0)$ = 0 radians (arbitrary)

$x_6(0)$ = (R_0+350,000) ft. = 21275840 ft.

Using the above definitions and initial conditions, the flight of a vehicle entering the earth's atmosphere at 350,000 ft. above the surface of the earth can be fully des-

cribed. For the particular values we used, the vehicle descends to approximately 200,000 ft. (at which point x_3 becomes negative), returns to 265,000 ft., and then turns toward the earth again, thus completing one 'skip'. (The above equations (D) must be modified if the vehicle descends below $y = 100{,}000$ ft.)

At the end of our interval, $s = 9 \times 10^6$ ft.,

$x_1 = (20812.365 \text{ ft./sec.})^2$

$x_2 = 1.2943885$ radians

$x_3 = .028790246$ radians

$x_4 = .73785486$ radians

$x_5 = .45546708$ radians

$x_6 = (R_0 + 238795.24)$ ft.

Since $x_3 > 0$, the vehicle's height y is decreasing. These exact values for x_i were the solutions in the round-off regions of all the predictor-corrector methods and of the Runge-Kutta method. This fact indicates that these x_i are good approximations to the solution of (D).

BIBLIOGRAPHY

1. Collatz, L. The Numerical Treatment of Differential Equations, (1960).
2. Dahlquist, G. Convergence and stability in the numerical integration of ordinary differential equations, Math. Scand., 4 (1956), pp. 33-53.
3. Faddeeva, V.N. Computational Methods of Linear Algebra, (1959).
4. Fine, J. Doctoral Thesis, Aerospace Institute, Toronto.
5. Gantmacher, F.R. Matrix Theory, (1959).
6. Henrici, Peter. Discrete Variable Methods in Ordinary Differential Equations, (1962).
7. Henrici, Peter. Error Propagation for Difference Methods, (1963).
8. Hull, T.E. and Creemer, A.L. Efficiency of predictor-corrector procedures, JACM, 10 (1963), pp. 291-301.
9. Hull, T.E. and Luxemburg, W.A.J. Numerical methods and existence theorems for ordinary differential equations, Numerische Math., 2 (1960), pp. 30-41.
10. Taylor, A.E. Advanced Calculus, (1955).
11. Taylor, A.E. Functional Analysis, (1958).