On increment modification in finite difference methods for solving ordinary differential equations

William Michael Wagner
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Mathematics Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/5343

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
This dissertation has been microfilmed exactly as received 67-2101

WAGNER, William Michael, 1936-
ON INCREMENT MODIFICATION IN FINITE DIFFERENCE METHODS FOR SOLVING ORDINARY DIFFERENTIAL EQUATIONS.

Iowa State University of Science and Technology, Ph.D., 1966 Mathematics

University Microfilms, Inc., Ann Arbor, Michigan
ON INCREMENT MODIFICATION IN FINITE DIFFERENCE METHODS FOR SOLVING ORDINARY DIFFERENTIAL EQUATIONS

by

William Michael Wagner

A Dissertation Submitted to the Graduate Faculty in Partial Fulfillment of The Requirement for the Degree of DOCTOR OF PHILOSOPHY

Major Subject: Applied Mathematics

Approved:

Signature was redacted for privacy. In Charge of Major Work

Signature was redacted for privacy. Head of Major Department

Signature was redacted for privacy. Dean of Graduate College

Iowa State University Of Science and Technology Ames, Iowa 1966
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter I: Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Chapter II: The Transitional Equations</td>
<td>12</td>
</tr>
<tr>
<td>Chapter III: The Two Back-Point Case</td>
<td>43</td>
</tr>
<tr>
<td>Chapter IV: The Three Back-Point Case</td>
<td>51</td>
</tr>
<tr>
<td>Bibliography</td>
<td>68</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>69</td>
</tr>
<tr>
<td>Appendix</td>
<td>70</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

From the theory of ordinary differential equations it is known that for the initial value problem,

\[(1.1) \quad y' = f(x, y), \quad (x_0, y_0),\]

where it is assumed that \(f\) is continuous and satisfies a Lipschitz condition throughout some open region, \(R\), of the \((x, y)\) plane containing the point \((x_0, y_0)\), there exists an open interval, \((a, b)\), on which a unique solution, \(y\), exists. In this paper is developed a method of modifying the increment or, as it is frequently called, the step-size associated with multi-step methods used in obtaining a numerical solution of such initial value problems.

A numerical solution of (1.1) consists of a set of ordered pairs, \(\{ (x_i, y_i) \} \), where \(y_i\) is taken as an approximation of \(y(x_i)\). It is usually assumed that the values, \(x_i\), of the independent variable are exact. The spacing of successive values of this variable is the increment referred to above and is normally constant for a number of consecutive values of the variable. Depending upon the behavior of the solution, \(y\), it is sometimes advantageous to increase or decrease the increment.

There are two basic types of methods used in obtaining a numerical solution of (1.1), known respectively as
single-step and multi-step methods. As the term single-step is intended to imply, such methods generate points of the numerical solution utilizing only the previously determined point of the numerical solution, whereas, the multi-step methods utilize several, so called back-points. The primary advantage of a single-step over a multi-step method is that the former is self-starting, the point \((x_0, y_0)\) serving as the initial back-point; a multi-step method obviously is not. On the other hand, a multi-step method requires only one or two evaluations of a derivative at each point of the numerical solution while a single-step usually requires several, thus, in general, a multi-step method is more efficient. Among the former, the most widely used are those of the Runge-Kutta type. Hybrid methods, such as the one proposed by Butcher (1), are also available. In view of the above statements, a single-step method frequently is used to provide the necessary number of back-points required by a multi-step method, the latter then assuming the task of generating additional points of the numerical solution. Dahlquist (3), (4), and Henrici (7), are primarily responsible for presenting a general theory of multi-step methods. Both single-step and multi-step methods are discussed in detail by the latter in his book.

Multi-step methods usually involve one or two recursive
formulas of the type

\[ y_{n+k} = a_{n-1} y_{n-1+k} + \cdots + a_0 y_k + h \left[ b_n f(x_{n+k}, y_{n+k}) + \cdots + b_0 f(x_k, y_k) \right] \]

with \( h \) as the increment (Note, although \( h \) can be negative, without loss of generality it will be assumed that \( h \) is positive, thus giving meaning to such terms as "rightmost back-points". Also, a non-zero coefficient, \( a_n \), of \( y_{n+k} \) is frequently included; here \( a_n = -1 \).) If \( b_n = 0 \), the right-hand side of (1.2) does not depend on \( y_{n+k} \); such an equation is said to be of the predictor type or simply, a predictor. A corrector is characterized by the condition \( b_n \neq 0 \). Either a predictor or corrector, the latter iterated if \( f \) involves the dependent variable, can be used alone in obtaining a numerical solution to (1.1) or they can be used in conjunction. Widely used is a PECE algorithm, i.e., a predictor is used to generate an approximation \( \tilde{y}_{n+k} \) of \( y(x_{n+k}) \); thinking of (1.2) as representing a corrector, \( \tilde{y}_{n+k} \) is then used in lieu of \( y_{n+k} \) to evaluate the derivative at \( x_{n+k} \) in the right-hand side of (1.2). The corrector yields the accepted approximation \( y_{n+k} \) of \( y(x_{n+k}) \); finally an evaluation of the derivative at \( x_{n+k} \) is made using \( y_{n+k} \). For the method of increment modification presented in this paper PECE algorithm will be used. It is a trivial point, but worth noting, that
if \( f \) does not involve the dependent variable explicitly, the predictor of a PECE algorithm serves no useful purpose. Under such conditions, the first half of the PECE algorithm can be skipped without affecting the result.

Two concepts are fundamental to multi-step methods, that of order and stability. Order is a local property. If (1.2) is of order \( q \) and \( y \) is a polynomial of degree \( q \), then \( y \) satisfies (1.2) exactly for any value of \( k \). To say that (1.2) is stable implies that, as \( k \) increases, any errors already present in the dependent variable are not magnified beyond bound. As an example of instability, consider \( y_{2+k} = 2y_{1+k} - y_k \), applied to \( y' = 0 \), \((x_0, y_0) = (0, C)\). The method, being of order zero, furnishes an exact numerical solution to the initial value problem. However, if an error of \( \epsilon \) is present in \( y_1 \), a few iterations on \( k \) show that \( y_k = C + k\epsilon \) for \( k = 0, 1, \ldots \) and thus, the initial error is magnified with increasing \( k \). It is shown in Henrici (7, pp. 218 ff.) that stability of (1.2) is characterized by the conditions that no zero of the polynomial \( P \), associated with (1.2) and defined by

\[
(1.3) \quad P(z) = -z^n + a_{n-1}z^{n-1} + \ldots + a_0
\]

exceeds one in modulus and that any zeroes of modulus one are simple.

The fundamental theorem underlying multi-step methods is that a stable multi-step method of order at least one is
convergent under the following definition given by Henrici (7, pp. 218-219).

**Definition 1.1.** The linear multi-step method defined by (1.2) is said to be convergent if and only if the following is true for all functions, $f$, satisfying the conditions stated for the unique solution of the initial value problem (1.1) and all values of $y_0$ such that $(x_0, y_0) \in \mathbb{R}$:

\[
(1.4a) \quad \lim_{{h \to 0}} y_n = y(x) \quad \text{as} \quad x_n = x
\]

holds for all values of $x \in (a, b)$ and all solutions \{\{x_m, y_m\}\} of (1.2) having starting values $y_\mu^{(0)} = y^{(0)}(x_0 + h)$ satisfying

\[
(1.4b) \quad \lim_{{h \to 0}} y^{(\mu)}_0 = y_0, \quad \mu = 0, \ldots, n-1.
\]

It should be noted that Definition 1.1 takes into account the increased number of steps necessary to reach the point $x$ from $x_0$ for a decreased increment and it allows for error in the initial set of back-points. Thus, a prescribed accuracy in the dependent variable at the point $x$ can be achieved by choosing $h$ sufficiently small. This condition has certain practical limitations however. For instance, in a digital computer where only a fixed number of digits are carried with each arithmetic operation, the number of significant digits in a difference such as $f(x+h, y) - f(x, y)$ can be
reduced to zero by choosing \( h \) sufficiently small, the "difference" consisting strictly of computational round-off error. It is easy to see that if \((1.2)\) is of order zero then one is a zero of \( P \). With zeroes of modulus one playing the critical role in stability, Crane and Lambert (2) show that, under the assumption \( f_y \) exists and is constant with value \( K \), upon letting \( \bar{h} = Kh \), the stability condition imposed on the polynomial \( P \), where

\[
P(z) = -z^n + a_{n-1}z^{n-1} + \ldots + h(b_nz^n + b_{n-1}z^{n-1} + \ldots + b_0)
\]

lead to practical intervals of stability for \( h \), i.e., intervals throughout which the zeroes of \( P(z) \) lie on or inside the unit circle, any on being simple. This latter type of stability is known as numerical stability, the former then being referred to as zero stability.

A basic property of convergent linear multi-step methods is that, except under special conditions, their order cannot exceed one more than the number of back-points utilized. The exceptions possess undesirable features which are attributable to the fact that all zeroes of the associated polynomial \( P \) lie on the unit circle. For this reason it is assumed, in conjunction with the method of increment modification presented in this paper, that multi-step methods of order less than or equal to \( n+1 \), \( n \) being the number of back-points, are in use in obtaining a
Modification of the increment, \( h \), should be undertaken for several reasons. In practice, it is more likely that \( f_y \) will vary slowly with the solution than remain constant. Thus, if left unchecked, \( h \) may leave an interval of numerical stability. To guard against this, a rough approximation of \( f_y \) can be made periodically via standard differencing techniques.

A second reason for reducing the increment is that the solution, \( y \), to the initial value problem (1.1) cannot be expected to satisfy (1.2) exactly; error is usually introduced at each step of the numerical solution. This error, discussed later, is referred to as discretization or truncation error. An estimation of the truncation error is, in general, difficult to obtain. However, in the event a PECE algorithm is being used to obtain the numerical solution and the predictor and corrector are of the same order, a simple technique for estimating the truncation error, attributed to W. E. Milne, is available. The method is described in Henrici (7, pp. 256-257).

Depending upon the behavior of the solution, \( y \), of (1.1), it is at times possible to increase the size of the increment. This can be done if the accumulated error is well within desired tolerances and if it is possible to increase the increment, \( h \), without also sacrificing
numerical stability.

Once it is decided that the increment is to be modified, the modification can be done in various ways. Certainly the easiest way to increase an increment is to double $h$ and use every other back-point in the multi-step procedure. However, as will be seen later, when a linear multi-step method requiring $n$ back-points is programmed for a digital computer, it frequently is the case that only the $n$ current right-most points of the numerical solution can be referenced. A second modification procedure, used to halve an increment, is interpolation. If, as is quite likely, a single-step method has been used to generate the initial set of back-points required by a multi-step method, the former could be reintroduced to affect the modification. If increment modification is fairly infrequent, the relative inefficiency of the single-step method would not be noticeable. However, if the opposite is true, the method presented in this paper, more in the line of multi-step methods, would be an advantage. Also, if used in conjunction with a digital computer of limited storage, the memory space occupied by the single-step method, once it performed its function of producing an initial set of back-points, could be made available for general storage. Thus, in the case of two or three back-points, a net gain in storage could be realized if the method herein is utilized.
The method of increment modification presented in this paper can be described briefly as follows. A pair of predictor and corrector equations, each utilizing \( n \) uniformly spaced back-points, is used to step the numerical solution ahead by an amount \( \lambda h \), \( \lambda \) in general, being a positive constant. Taking into account the fact that the two rightmost back-points of the numerical solution are now separated by an interval of size \( \lambda h \) while the rest are separated by an interval of size \( h \), a second predictor and corrector pair is used to step the solution ahead by an amount \( \lambda h \). A total of \( n-1 \) such pairs, each taking into account the unique spacing of the relative back-points, are used yielding \( n-1 \) additional points of the numerical solution. Thus there result \( n \) consecutive points being uniformly spaced by an amount \( \lambda h \). At this time the original multi-step can be re-introduced.

A word concerning the programming of multi-step methods on a digital computer is in order since this latter increment modification procedure involves an abundance of equations of the type (1.2). An efficient scheme for programming a PECE algorithm follows. Three vectors, \( Y, P, C \), of consecutive words of storage, the first containing the values of the dependent variables and derivatives and the other two containing the products of \( h \) and the coefficients of the predictor and corrector formulas respectively, are set up
To compute a predicted value of the dependent variable an inner product of the first $2n+1$ elements of the vectors $Y$ and $P$ is formed, the result being stored in the word labeled $Y_n'$. Using this value and the current value of the independent variable, the derivative is evaluated with the result being placed in the word labeled $Y_n'$. Next, an inner product of the first $2n+1$ elements of the vectors $Y$ and $C$ is computed, the result being placed in the word $Y_n'$ and possibly printed on an output device. Using the latter value for the dependent variable, the derivative is again evaluated and placed in $Y_n'$. At this point the contents of the vector $Y$ are shifted left one word, the original contents of the first word being lost. Finally, the cycle is completed by incrementing the independent variable by the amount $h$. Inner products of this nature are readily performed on digital computers; if an indirect addressing feature is available, such as on the IBM 7090 series of
computers, the referencing of the various predictor and corrector coefficient vectors is accomplished with ease.
CHAPTER II: THE TRANSITIONAL EQUATIONS

In deriving the transitional equations, a linear operator \( L \) will be defined for functions, \( y \), of class \( C^1 \) over some real interval \((a, b)\); \( L[y](x) \) to be indicated by \( L[y(x)] \). The parameters \( n \) and \( h \) appearing in the definition of \( L \) will correspond respectively to the number of back-points and the increment of multi-step formulas similar to (1.2) which will be associated with \( L \).

**Definition 2.1.** Let \( y \) be a function of class \( C^1 \) over the real interval \((a, b)\). Let \( n \geq 2, \ h > 0, \) and \( \mu_i, \ i = 0, \ldots, n \) satisfy \( \max \mu_i - \min \mu_i < (b-a)/h \). The linear operator, \( L \), is then defined by the equation

\[
L[y(x)] = \sum_{j=0}^{n} a_j y(x + \mu_j h) + h \sum_{j=0}^{n} b_j y'(x + \mu_j h).
\]

If \( y \) is also of class \( C^{r+1} \), \( r > 0 \), over the interval \((a, b)\), \( L[y(x)] \) can be expanded by Taylor's formula about some point \( \bar{x} \in (a, b) \). Using the definitions

\[
\begin{align*}
(2.2a) \quad & t = (x-\bar{x})/h, \\
(2.2b) \quad & 0(t+\mu_j)^{-1} = 0, \\
(2.2c) \quad & (t+\mu_j)^0 = 1, \quad j = 0, \ldots, n.
\end{align*}
\]
(2.1) can be written

\[ (2.2d) \quad L[y(x)] = \sum_{j=0}^{n} a_j \sum_{i=0}^{r} \frac{y^{(i)}(\bar{x})(t+\mu_j)^{i+1}}{i!} h^i \]

\[ + \sum_{j=0}^{n} b_j \sum_{i=0}^{r} \frac{y^{(i)}(\bar{x})_1(t+\mu_j)^{i-1}h^i}{i!} \]

\[ + \frac{h^{r+1}}{(r+1)!} \sum_{j=0}^{n} (t+\mu_j)^r a_j (t+\mu_j) y^{(r+1)}(\xi_j) + b_j (r+1) y^{(r+1)}(\zeta_j), \]

\[ \xi_j, \zeta_j \in (\bar{x}, x+\mu_j h), \quad j = 0, \ldots, n. \]

Interchanging the order of summation in (2.2d) and expanding the powers of \( t+\mu_j \) by the binomial theorem yields

\[ (2.3a) \quad L[y(x)] = \sum_{i=0}^{1} \frac{y^{(i)}(\bar{x})}{i!} h^i \sum_{j=0}^{n} \left\{ a_j \sum_{i'=0}^{i} \left( \frac{i-1}{i'} \right) \mu_j^{i'} \right\}^{i-1} \]

\[ + b_j \sum_{i'=0}^{1} \left( \frac{i-1}{i'} \right) \mu_j^{i'} \right\}^{i-1} + R_{r+1}(x, \bar{x}; y) \]
where

\[(2.3b) \quad R_{r+1}(x, \xi; y) = \frac{n^{r+1}}{(r+1)!} \sum_{j=0}^{n} \frac{(t+\mu_j)^r}{(r+1)!} \left[ \alpha_j (t+\mu_j) y^{(r+1)}(\xi_j) \right] \]

\[+ (r+1) b_j y^{(r+1)}(\xi_j), \]

\[\xi_j, \xi'_j \in (\bar{x}, x+\mu_j h), \quad j = 0, \ldots, n,\]

and where

\[(2.3c) \quad \begin{pmatrix} -1 \\ i' \end{pmatrix} = 0, \quad i' = 0, -1,\]

\[(2.3d) \quad t^0 = 1,\]

and

\[(2.3e) \quad \mu_j = 1, \quad j = 0, \ldots, n.\]

In view of (2.3c), the identity

\[(2.3f) \quad \begin{pmatrix} i-1 \\ i' \end{pmatrix} = \begin{pmatrix} i \\ i-1 \end{pmatrix}(i-1', \quad i = 0, 1, \ldots; \quad i' = 0, \ldots, i-1,\]

holds. Making use of this identity in the fourth summation of (2.3a), upon substituting \(i''\) for \(i-i'\) in the third and fourth summations, there results
(2.3g) \[ L[y(x)] = \sum_{i=0}^{r} \sum_{j=0}^{n} \sum_{i''=1}^{1} \left\{ a_j \sum_{i''=1}^{0} \binom{1}{i''-1} t^{i''-1} \mu_j^{i''} \right\} + R_{r+1}(x,\bar{x};y). \]

The latter summation of (2.3g) can be extended to include \( i''=0 \); this permits the latter two summations to be combined. Thus,

(2.3h) \[ L[y(x)] = \sum_{i=0}^{r} \sum_{j=0}^{n} \sum_{i''=0}^{1} \binom{i''}{i''-1} t^{i''-1} (a_j \mu_j^{i''} + b_j \mu_j^{i''-1}) + R_{r+1}(x,\bar{x};y). \]

Finally, interchanging the order of summation of the latter two yields
It is convenient at this point to introduce the following. Referring to (2.31), let

\[(2.4a) \quad C_i = \sum_{j=0}^{n} (a_j \mu_j^i + b_j \nu_j^{i-1}), \quad i = 0, 1, \ldots,\]

and

\[(2.4b) \quad T_i(t) = \sum_{i' = 0}^{i} \binom{i}{i'} t^{i-i'} C_i, \quad i = 0, 1, \ldots.\]

These definitions permit (2.31) to be written as

\[(2.5) \quad L[y(x)] = \sum_{i=0}^{r} \frac{y^{(i)}(\bar{x})}{i!} h^{i} T_i(t) + R_{r+1}(x, \bar{x}; y).\]
Definition 2.2. The linear operator $L$, as defined by Definition 2.1, is said to be of order $q$ if and only if $C_i = 0$, $i = 0, \ldots, q$ and $C_{q+1} \neq 0$.

An immediate, but important, consequence of (2.2a), (2.3d), and (2.4b) concerning the first non-vanishing term of (2.5) when $L$ is of order $q$ is stated in that if the linear operator $L$, as defined by Definition 2.1, is of order $q$, and, if $y$ is any function of class $C^{q+2}$, $q \geq 0$, over the interval $(a, b)$, then $T_{q+1}(t)$, as given by (2.4b), is independent of $x$ and $\bar{x}$. In particular,

\[(2.6) \quad T_i(t) = \begin{cases} 0, & i = 0, \ldots, q \\ C_{q+1}, & i = q+1. \end{cases} \]

With $r$ replaced by $q+1$ in (2.5), this implies $L[y(x)]$ can be represented as

\[(2.7a) \quad L[y(x)] = C_{q+1} \frac{y^{(q+1)}(\bar{x})}{(q+1)!} n^{q+1} + R_{q+2}(x, \bar{x}; y) \]

where $x$ and $\bar{x}$ are any points of $(a, b)$. In particular, with the choice $x = \bar{x}$, $t = 0$, hence,
\[ (2.7b) \quad L[y(x)] = c_{q+1} \frac{y^{(q+1)}(x)}{(q+1)!} h^{q+1} \]

\[ + \frac{h^{q+2}}{(q+2)!} \sum_{j=0}^{n} \mu_j^{q+1} \left[ a_j y_j^{(q+2)}(\xi_j) + (q+2) b_j y_j^{(q+2)}(\zeta_j) \right], \]

\[ \xi_j, \zeta_j \in (x, x + \mu_j h), \quad j = 0, \ldots, n. \]

**Definition 2.3.** The constant \( c_{q+1} \), as given by (2.4a) when \( q = q+1 \), will be referred to as the \( q \)th order error constant of the linear operator \( L \) as defined by (2.1).

If a multi-step method requiring \( n \) back-points spaced at intervals of size \( h \) is used to furnish a numerical solution of (1.1), in modifying the increment by a factor of \( \lambda \), \( n-1 \) additional points of the numerical solution, spaced at intervals of size \( \lambda h \), must be generated before there are again available \( n \) uniformly spaced back-points. These \( n-1 \) points will be generated by \( n-1 \) pairs of predictor and corrector formulas, each stepping the numerical solution ahead by an amount \( \lambda h \), the \( k\text{th} \), \( k = 1, \ldots, n-1 \), pair utilizing \( k-1 \) back-points spaced at intervals of size \( \lambda h \) and \( n+1-k \) back-points spaced by an amount \( h \). Thus, taking into account the new point generated by the \( k\text{th} \) pair, there are \( k \) intervals of size \( \lambda h \) and \( n-k \) of size \( h \) associated with the
Definition 2.4. For each \( k, \ k = 1, \ldots, n-1 \), the predictor operator \( L^p_k \lambda \), and the corrector operator \( L^c_k \lambda \), are operators as defined by Definition 2.1 characterized by the following conditions:

i) the coefficient corresponding to \( a^n \) in (2.1) is -1 in both \( L^p_k \lambda \) and \( L^c_k \lambda \);

ii) the coefficient corresponding to \( b^n \) in (2.1) is zero in \( L^p_k \lambda \) and non-zero in \( L^c_k \lambda \);

iii) for both \( L^p_k \lambda \) and \( L^c_k \lambda \),

\[
\mu_i = \begin{cases} 
1, & i = 0, \ldots, n-k, \\
(n-k)+(i-n+k), & i = n-k+1, \ldots, n.
\end{cases}
\]

The notation \( L^p[y(x);k,\lambda] \) will be used to designate \( L^p_k \lambda [y](x) \) and \( L^c[y(x);k,\lambda] \) to designate \( L^c_k \lambda [y](x) \). Hence, with these definitions, for \( y \), a function of class \( C^1 \) over \((a,b)\),

\[
(2.8a) \quad L^p[y(x);k,\lambda] = \sum_{j=0}^{n-k} \alpha_{kj} y_{x+jh} 
\]

\[
+ \sum_{j=n-k+1}^{n} \alpha_{kj} y_{x+(n-k)h+(j-n+k)\lambda h}
\]
\[
\begin{align*}
&\text{for } k = 1, \ldots, n-1,
\end{align*}
\]
(2.9a) \[ \Gamma_{k1} = \sum_{j=0}^{n-k} \alpha_{kj} j + \sum_{j=n-k+1}^{n} \alpha_{kj} [(n-k)+(j-n+k)\lambda]^j \]

\[ + \sum_{j=0}^{n-k} \beta_{kj} j^i + \sum_{j=n-k+1}^{n} \beta_{kj} [(n-k)+(j-n+k)\lambda]^{j-1} \],

\[ k = 1, \ldots, n-1; \quad i = 0, 1, \ldots, \]

and,

(2.9b) \[ C_{k1} = \sum_{j=0}^{n-k} a_{kj} j^i + \sum_{j=n-k+1}^{n} a_{kj} [(n-k)+(j-n+k)\lambda]^j \]

\[ + \sum_{j=0}^{n-k} b_{kj} j^{i-1} + \sum_{j=n-k+1}^{n} b_{kj} [(n-k)+(j-n+k)\lambda]^{-1} \],

With each of the 2(n-1) operators as given by Definition 2.4, a multi-step formula will be associated. It is assumed that a multi-step method requiring n back-points is being used to generate a numerical solution of (1.1), y being the true solution, and, at the onset of increment modification, \((x_N, y_N)\) is the rightmost point of this solution.

For each k, the set
(2.10) \( B_k = \left\{ (x_{N-n+j+k}, y_{N-n+j+k}) \right\}_{j=0}^{n-1} \)

will constitute the set of back-points for the formulas associated with \( I_{k \lambda}^P \) and \( I_{k \lambda}^C \). The formula associated with \( L_{k \lambda}^P \) will yield \( \hat{y}_{n+k} \), a predicted value of \( y(x_{n+k}) \). The formula associated with the corrector operator \( L_{k \lambda}^C \), will utilize the set \( B_k \) together with the point \( (x_{n+k}, \hat{y}_{n+k}) \). This formula will serve to generate \( y_{n+k} \), the corrected, i.e., accepted, approximation of \( y(x_{n+k}) \). Thus, the formulas associated with (2.8a) and (2.8b) are respectively,

\[
(2.10a) \quad \hat{y}_{n+k} = \sum_{j=0}^{n-1} \alpha_{kj} y_{N-n+j+k+h} + \sum_{j=0}^{n-1} \beta_{kj} y'_{N-n+j+k},
\]

and

\[
(2.10b) \quad y_{n+k} = \sum_{j=0}^{n-1} a_{kj} y_{N-n+j+k+h} + \sum_{j=0}^{n-1} b_{kj} y'_{N-n+j+k},
\]

where

\[
(2.10c) \quad y'_{N-n+j+k} = f(x_{N-n+j+k}, y_{N-n+j+k}),
\]

and
Note, the formulas imply a PECE type algorithm.

**Definition 2.5.** The order of a predictor or corrector formula will be that of the associated operator.

In general it cannot be expected that the points of the solution, \( y \), of (1.1) satisfy the equations of (2.10) exactly or, conversely that points generated by these equations coincide with points of the true solution, \( y \), even if it were the case that the points of the initial set of back-points, \( B_1 \), did. Thus, error is introduced at each application of a formula of (2.10) and, since in the general case, it must be assumed that error is also in the points of \( B_k \), each application will propagate a modification of inherited error. In order to examine these errors, let

\[
(2.11a) \quad \varepsilon_i = y_i - y(x_i), \quad i = 0, \ldots, N+n-1,
\]

\[
(2.11b) \quad \tilde{\varepsilon}_i = \widetilde{y}_i - y(x_i), \quad i = N+1, \ldots, N+n-1.
\]

Under these definitions and the equations of (2.8) and (2.10), upon forming the differences \( \widetilde{y}_{N+k} - L^p[y(x_{N+k});k,\lambda] \) and \( y_{N+k} - L^c[y(x_{N+k});k,\lambda] \), there results the error equations

\[
(2.10d) \quad \widetilde{y}_{N-n+j+k} = f(x_{N-n+j+k}, \widetilde{y}_{N-n+j+k}),
\]

\[ k = 1, \ldots, n-1; \quad j = 0, \ldots, n-1. \]
(2.12a) \[ \varepsilon_{N+k} = \sum_{j=0}^{n-1} \alpha_{kj} \varepsilon_{N-n+k+j} \]
\[ + h \sum_{j=0}^{n-1} \beta_{kj} \left[ y'_{N-n+k+j} - y'_{(x_{N-n+k+j})} \right] \]
\[ + L^p \left[ y(x_{N+k}); k, \lambda \right] \]

and

(2.12b) \[ \varepsilon_{N+k} = \sum_{j=0}^{n-1} a_{kj} \varepsilon_{N-n+k+j} + h b_{kn} \left[ y'_{N+k} - y'_{(x_{N+k})} \right] \]
\[ + h \sum_{j=0}^{n-1} b_{kj} \left[ y'_{N-n+k+j} - y'_{(x_{N-n+k+j})} \right] \]
\[ + L^c \left[ y(x_{N+k}); k, \lambda \right], \quad k = 1, \ldots, n-1. \]

The assumption will now be made that \( f_y \) exists and is constant of value \( K \) over the relatively small transition interval. By the mean value theorem it follows that
(2.13a) \( y_1' - y'(x_1) = K\epsilon_i \), \( i = N-n+1, \ldots, N+n-2 \)

and

(2.13b) \( \tilde{y}_1' - y'(x_1) = K\tilde{\epsilon}_i \), \( i = N+1, \ldots, N+n-1 \).

With \( h \) as the increment separating at least the \( n \) points to the left of \( x_N \), define

(2.14) \( \tilde{h} = Kh \).

The error equations of (2.12) can be now written respectively

(2.15a) \( \tilde{\epsilon}_{N+k} = \sum_{j=0}^{n-1} \alpha_{kj} \epsilon_{N-n+k+j} \)

and

\[ +\tilde{h} \sum_{j=0}^{n-1} \beta_{kj} \epsilon_{N-n+k+j} + L \left[ y(x_{N+k}); \kappa, \lambda \right] \]
(2.15b) \( \epsilon_{N+k} = \sum_{j=0}^{n-1} a_{kj} \epsilon_{N-n+k+j} + \tilde{h} b_{kn} \epsilon_{N+k} \)

\[ + \tilde{h} \sum_{j=0}^{n-1} b_{kj} \epsilon_{N-n+k+j} + L^c[\gamma(x_{N+k});k,\lambda], \]

\( k = 1, \ldots, n-1. \)

Definition 2.6. For the solution, \( y, \) of (1.1), \( L^p[y(x);h,\lambda] \) and \( L^c[y(x);k,\lambda] \) are known as the discretization or truncation errors of, respectively, the predictor and corrector formulas as given by the equations of (2.10) at \( x, k=1, \ldots, n-1. \)

Thus, the discretization error is the error introduced upon applying one of the methods of (2.10). Substituting \( \tilde{\epsilon}_{n+k} \) as given by (2.15a), into (2.15b), yields the total error after one cycle of the PECE algorithm has been completed (Note, the error is established at the correction phase of the PECE cycle, the subsequent evaluation of the derivative does not affect the error of the current cycle but does enter into the next \( n \) cycles.). Thus,
\[ (2.16) \quad \epsilon_{N+k} = \sum_{j=0}^{n-1} \left[ a_{kj}^l \alpha_{kj}^l k_j + b_{kj}^l \beta_{kj}^l h + b_{kj}^l \beta_{kj}^l h^2 \right] \epsilon_{N-n+k+j} \]

\[ + h b_{kn}^l \left[ y(x_{N+k}); k, \lambda \right] + h b_{kn}^l \left[ y(x_{N+k}); k, \lambda \right], \]

\[ k = 1, \ldots, n-1. \]

With \( \lambda = 1 \) and the coefficients of the equations of (2.10) independent of \( k \), (2.16) describes the error in the dependent variable for the usual PECE algorithm of uniform step-size. Because of the factor of \( h \) present in the last term of (2.16), it is not uncommon to find a predictor in use that is of order one less than that of the associated corrector. However, it is important to distinguish between the order of a method and the magnitude of its truncation error. Depending upon the magnitude of \( h \), e.g. \( |K| \) roughly of magnitude \( 1/h \), the term of the error (2.16) associated with the predictor may dominate excessively. Hamming (6, p. 198) adopts the attitude that \( |h| \) be kept less than 0.4 and it is not uncommon to find multi-step methods that are numerically stable for values of \( |h| \) larger than this. Increment modification, and especially reduction, can be expected to occur under relatively critical conditions such as \( h \) leaving a region of stability and/or
the truncation error exceeding desired bounds; the choice will be made to select predictors and correctors of the same order, this order being n+1, n being the number of back-points. As has been mentioned, the commonly used multi-step methods utilizing n back-points spaced uniformly are of order n+1. Thus, the truncation error of the transition formulas can be expected to be compatible with the multi-step formulas of uniform increment with respect to order. With \( y \) a function of class \( C^{n+3} \), (2.16) can be written

\[
\epsilon_{n+k} = \sum_{j=0}^{n-1} \left[ a_{kj} + (b_{km} \alpha_{kj} + b_{kj}) \tilde{h}^j + b_{km} \beta_{kj} \tilde{h}^2 \right] \epsilon_{n-k+j}
\]

\[ + \frac{1}{(n+2)!} (c_{k,n+2} + h b_{km} \tilde{h} \Gamma_{k,n+2}) y^{(n+2)} (x_{n+k} h^{n+2} + o(h^{n+3})) \]

\( k = 1, \ldots, n-1. \)

**Definition 2.7.** The truncation coefficient of the error equation (2.17) will be defined to be the quantity

\[
T_k(\tilde{h}) = c_{k,n+2} + h b_{km} \tilde{h} \Gamma_{k,n+2}, \quad k = 1, \ldots, n-1.
\]

In order that the predictor and corrector formulas as given by the equations of (2.10) be of order n+1 for a particular value of k, it is necessary that

\[
\Gamma_k = 0,
\]
To achieve this order there are available $2n$ parameters, namely, $a_{ki}, b_{ki}, i = 0, \ldots, n-1$. Imposing the $n+2$ conditions of (2.19a) on these parameters implies at least $n-2$ free parameters. For the corrector (2.10b), there is available one more parameter and, since the same number of conditions are imposed, this implies one extra free parameter associated with the corrector. Also, since the equations of (2.10) are linear in the parameters, the solutions to those of (2.19) will be linear in the respective free parameters.

Two techniques will be considered for selecting the free parameters. The first is statistical in nature.

For each $k, k = 1, \ldots, n-1$, let the parameter $b_{kn}$ of the respective corrector equations as given by (2.10b) be among the free parameters. Referring to (2.9b), it can be seen that $c_{k,n+2}$ is linear in $b_{kn}$, which permits the implicit definition of $m_k$ and $\tilde{c}_k$ by

$$c_{k,n+2} = m_k b_{kn} + \tilde{c}_k$$

where $\tilde{c}_k$ does not depend upon $b_{kn}$. Thus, (2.18) can be written

$$T_k (\tilde{h}) = \tilde{c}_k + (m_k + h \Gamma_{k,n+2}) b_{kn}$$
The assumption will be made that for a particular multi-step method, possibly in conjunction with a restricted set, F, of initial value problems, an educated guess can be made of the mean value, $\bar{h}$, of $h$ at which increment modification will occur. For instance, the property $f_y(x,y) = 0$ might apply to each element of F and stability of the multi-step method might extend to $\bar{h} = -1/2$. Taking into account that the modification is also triggered by the truncation error, if some prior knowledge of this were available, such a guess might be feasible. If this were the case, the parameter $b_{kn}$, for each $k = 1, \ldots, n-1$, can be selected to minimize the expected value of the magnitude of the truncation coefficient $T_k$. Furthermore, if a variance, $\sigma^2$, of $\bar{h}$ can also be specified, $b_{kn}$ can be selected to minimize $T_k^2$. In either case, the selection can be done under fairly weak conditions imposed on g, the frequency distribution of the values of $\bar{h}$ over which modification will occur. This is shown by the following theorem.

**Theorem 2.1.** Let the truncation coefficient, $T_k$, be given by (2.21) with $m_k$ and $\tilde{c}_k$ as defined implicitly by (2.20). Let g be any density function such that the first moment, $\bar{h}$, of g exists. If

\[(2.22a) \quad m_k + \sum_{k=n+2} \bar{h} = 0,\]

the value, \( \hat{b}_{kn} \), of \( b_{kn} \) (in terms of the other free parameters), which minimizes the magnitude of the expected value of the truncation coefficient is zero. If (2.22a) does not hold, the value, \( \hat{b}_{kn} \), of \( b_{kn} \) which minimizes the magnitude of the expected value of the truncation coefficient is

\[
(2.22b) \quad \hat{b}_{kn} = -\frac{\tilde{c}_k}{m_k + \int_{k,n+2}^2 \bar{h}} , \quad k = 1, \ldots, n-1.
\]

Also, if the variance, \( \sigma^2 \), of \( g \) exists, the value \( \tilde{b}_{kn} \), of \( b_{kn} \) which minimizes the expected value of \( T_k^2 \) is zero if (2.22a) holds, otherwise, it is

\[
(2.22c) \quad \tilde{b}_{kn} = -\frac{\tilde{c}_k}{(m_k + \int_{k,n+2}^2 \bar{h}) + \frac{\int_{k,n-2}^2 0^2}{(m_k + \int_{k,n+2}^2 \bar{h})}}.
\]

Proof: The second assertion only will be proved, the first being done in a similar manner. The expected value of the square of the truncation error is given by

\[
(2.23) \quad E\{T_k^2\} = \int_{-\infty}^{\infty} T_k^2(\bar{h})g(\bar{h})d\bar{h}.
\]

Upon adding and subtracting \( \bar{h}\int_{k,n+2}^2 b_{kn} \) to the right-hand side of (2.21) there follows
(2.24) \( T_k(\bar{h})^2 = \left\{ \left[ (m_k + \sum_{k,n+2}^{\bar{h}} b_{kn} + \tilde{c}_k \right] + (\bar{h} - \bar{h}) \right\}^2 \)

thus,

(2.25) \[
E\left\{ T_k^2 \right\} = \left[ (m_k + \sum_{k,n+2}^{\bar{h}} b_{kn} + \tilde{c}_k \right] \int_{-\infty}^{\infty} g(h) \, dh \\
+ 2\left[ (m_k + \sum_{k,n+2}^{\bar{h}} b_{kn} + \tilde{c}_k \right] \int_{-\infty}^{\infty} (\bar{h} - h) g(h) \, dh \\
+ \int_{k,n+2}^{2} b_{kn} \int_{-\infty}^{\infty} (\bar{h} - h)^2 g(h) \, dh.
\]

Since \( g \) is a density function with mean \( \bar{h} \) and variance \( \sigma^2 \), the three integrals of (2.25) are one, zero, and \( \sigma^2 \) respectively, i.e.,

(2.26) \[
E\left\{ T_k^2 \right\} = \left[ (m_k + \sum_{k,n+2}^{\bar{h}} b_{kn} + \tilde{c}_k \right] \int_{-\infty}^{\infty} g(h) \, dh \\
+ \int_{k,n+2}^{2} b_{kn} \int_{-\infty}^{\infty} (\bar{h} - h)^2 g(h) \, dh.
\]

which is non-negative. Under (2.22a) the conclusion is immediate. If (2.22a) does not hold, consider

(2.27a) \[
\frac{1}{2} \frac{dE\left\{ T_k^2 \right\}}{\frac{d}{b_{kn}}} = \left[ (m_k + \sum_{k,n+2}^{\bar{h}} b_{kn} + \sigma^2 \int_{k,n+2}^{2} b_{kn} \right] \\
+ (m_k + \sum_{k,n+2}^{\bar{h}} b_{kn}) \tilde{c}_k
\]

and
which exist for all values of $b_{kn}$. The theorem follows by equating $\frac{dE_k^T}{db_{kn}}$ to zero.

Note, when $\bar{\bar{n}}_{n+2} = 0$, the values $\hat{b}_{kn}$ and $\tilde{b}_{kn}$ as given by (2.22b) and (2.22c) respectively imply $C_{kn,n+2} = 0$ which in turn implies the order of the associated corrector is increased. Likewise, when $\bar{n} = 0$, the corresponding value of $\hat{b}_{kn}$ implies increased order of the corrector.

Theorem 2.1 describes a method of minimizing the truncation error at each step of the increment modification procedure. Consideration will now be given to the propagation of error when $f$ is a polynomial of degree $n+1$, $n$ being the number of back-points and the procedure is iterated to continually modify the increment by a factor of $\lambda$.

That $f$ is a polynomial of degree $n+1$ implies $K$ and hence $\bar{n}$, are zero; the solution, $y$, of (1.1) is a polynomial of degree $n+2$; and the interval $(a,b)$ over which the solution to (1.1) exists, is the real line. Also, because $f$ does not involve the dependent variable, predictors are unnecessary.

It is convenient to associate an error vector,
reflecting back-point error, with each step of the process. In the following, a prime will be used to indicate the transpose of a matrix.

**Definition 2.8.** Let $x_N$ be the value of the independent variable at which increment modification commences and let $h$ be the increment separating the points $x_{N-n+1}$ through $x_N$. The $n$ dimensional vector, $\vec{e}_{Nk}$, defined by

$$
(2.28) \quad \vec{e}_{Nk}(h) = (e_{N+k}, e_{N+k-1}, \ldots, e_{N+k-n+1})', \quad k = 0, \ldots, n-1,
$$

whose components are given by (2.11a), will be referred to as the $k$th error vector associated with $x_N$.

Thus, the errors in the back-points the $k$th formula of the modification encounters are the components of $\vec{e}_{N,k-1}$. This formula generates the first component of the vector $\vec{e}_{Nk}$ according to (2.17) with $h=0$.

**Definition 2.9.** The $n \times n$ propagation matrix $A_k = \left\{ u_{ij}^{(k)} \right\}$, is defined by

$$
(2.29) \quad u_{ij}^{(k)} = \begin{cases} a_{k,n-j} & i=1 \\ \delta_{i-1,j} & i=2, \ldots, n; \ k=1, \ldots, n-1
\end{cases}
$$

where $\delta_{ij}$ is Kronecker's delta.

With $y$, a polynomial of degree $n+2$ having leading coefficient $c$, and the modification formulas of order $n+1$, it follows from (2.7a) that the discretization error,
Definition 2.10. The n dimensional truncation vector $\vec{v}_k$ is defined by

$$(2.30) \quad \vec{v}_k = (c_k, n+2, 0, \ldots, 0)' , \quad k = 1, \ldots, n-1 .$$

With these definitions, the recurrence relation

$$(2.31) \quad \vec{\epsilon}_{Nk}(h) = A_k \vec{\epsilon}_N, k-1(h) + \vec{v}_k c_h^{n+2} , \quad k = 1, \ldots, n-1 ,$$

holds among the error vectors of the modification procedure. The initial error vector inherited by the procedure is $\vec{\epsilon}_{NO}$ and the last one generated by the n-1 modification formulas is $\vec{\epsilon}_{N, n-1}$. Upon letting

$$(2.32a) \quad A = \left( \begin{array}{c|c} n-1 \\ \hline & A_k \end{array} \right)' ,$$

and

$$(2.32b) \quad \vec{v} = \sum_{k=1}^{n-1} \left( \begin{array}{c|c} n-1 \\ \hline & A_j \end{array} \right)' \vec{v}_k ,$$

where

$$(2.32c) \quad \left( \begin{array}{c|c} n-1 \\ \hline & A_j \end{array} \right)' = I , \quad j = n .$$
\( \hat{e}_{N,n-1} \) can be expressed in terms of \( \hat{e}_{N0} \) and the \( n-1 \) truncation vectors as

\[
(2.33) \quad \hat{e}_{N,n-1}(h) = A \hat{e}_{N0}(h) + \hat{v} \chi h^{n+2}.
\]

After the modification of the increment by a factor of \( \lambda \) has been completed, the value of the independent variable of the rightmost point of the numerical solution is \( x_{N+n-1} = x_N + (n-1)\lambda h \). If the modification procedure is again applied, Definition 2.8 implies \( \hat{e}_{N+(n-1),0}(\lambda h) = \hat{e}_{N,n-1}(h) \).

In general, after \( j \) modifications,

\[
(2.34a) \quad x_{N+j(n-1)} = x_N + (n-1) \frac{1 - \lambda}{1 - \lambda} \lambda^j h, \quad \lambda \neq 1,
\]

and

\[
(2.34b) \quad \hat{e}_{N+j(n-1),0}(\lambda^j h) = \hat{e}_{N+(j-1)(n-1),n-1}(\lambda^{j-1} h),
\]

\( j = 1, 2, \ldots \).

In terms of \( \hat{e}_{N0} \), it can be shown that

\[
(2.35) \quad \hat{e}_{N+j(n-1),0}(\lambda^j h) = A^j \hat{e}_{N0}(h) + \sum_{i=0}^{j-1} A^i \lambda^{(j-i-1)(n+2)} h^{n+2} \chi.
\]

Now, if \( S \) is a matrix which transforms \( A \) to its Jordan normal
form $J$, i.e., if

$$(2.36) \quad J = SAS^{-1},$$

then (2.35) can be written

$$(2.37) \quad \hat{\epsilon}_{N+j(n-1),0}^j = S^{-1}J^jS\hat{\epsilon}_{N0}^j(h)$$

$$= S^{-1}\left\{ \sum_{i=0}^{j-1} j^i \lambda^{j-i}(n+2) \right\} e^{Scvh n+2}.$$ 

Thus, the error at the $j^{th}$ iteration is composed of two parts, one due to the initially inherited error, $\hat{\epsilon}_{N0}^j$, and the other due to generated truncation error generated at each step in the modification process. Now it is shown in Varga (8, p.14) that if $J_e = (d_{rs}(e))_1 \leq r, s \leq n_e$ is a Jordan block of $J$ of dimension $n_e$ corresponding to the eigenvalue $\mu_e$, then its $j^{th}$ power, $J_e^j = (d_{rs}(j)(e))$ is given by

$$$(2.38) \quad d_{rs}(j)(e) = \begin{cases} 
0, & s < r, \\
\frac{J^j}{(s-r)\mu_e^{j-s+r}}, & r \leq s \leq \min(n_e, j+r) \\
0, & j+r < s \leq n_e.
\end{cases}$$$


Recall that the powers, $J^j$, of $J$ retain the quasi-diagonal form of $J$ with respect to the Jordan blocks of $J$.

At this point, the following matrix and vector forms are introduced.

**Definition 2.11.** If $B = (b_{ij})$ is an $n \times n$ complex matrix and $\mathbf{z} = (z_1, \ldots, z_n)'$ is an $n$ dimensional complex vector, then the norm of $B$, $\|B\|$, and the norm of $\mathbf{z}$, $\|\mathbf{z}\|$, are defined as follows:

\begin{align}
(2.39a) \quad \|B\| &= \max_{1 \leq i \leq n} \sum_{j=1}^{n} |b_{ij}|, \\
(2.39b) \quad \|\mathbf{z}\| &= \max_{1 \leq i \leq n} |z_i|.
\end{align}

That these definitions satisfy the properties of a norm and that for two $n \times n$ matrices $B$ and $C$,

\begin{align}
(2.39c) \quad \|BC\| &\leq \|B\|\|C\|,
\end{align}

is shown in Faddeeva (5, p.54 ff.). Also, it is shown that the definition of the vector and matrix norms are compatible in the sense that

\begin{align}
(2.39d) \quad \|B\mathbf{z}\| &\leq \|B\|\|\mathbf{z}\|.
\end{align}
With this matrix norm applied to \( J_e \), it follows from (2.38) that \( \lim_{j \to \infty} \| J_e^j \| = 0 \) if and only if \( |\mu_e| < 1 \). Also, for any eigenvalue \( \mu_e \) of modulus one that is simple, \( \| J_e \| = 1 \). This, in turn, implies that for sufficiently large \( j \), say \( j_0 \), and all \( j > j_0 \), \( \| J^j \| \leq 1 \) if the eigenvalues of \( J \) are, in modulus, less than or equal to one and if those of modulus one are simple. Referring to (2.31) and (2.37), then

\[
(2.40) \quad \| A^j_{N,0}(h) \| \leq \| s^{-1} \| \| S \| \| \xi_{N,0}(h) \|, \quad j > j_0,
\]

and thus, under the above conditions, this component of (2.35) remains bounded as \( j \to \infty \), independent of \( \lambda \). Furthermore, for \( |\lambda| < 1 \), and \( j > j_0 \),

\[
(2.41) \quad \left\| \sum_{i=0}^{j-1} J^i \lambda^{(j-1-i)(n+2)} \right\| \leq \sum_{i=0}^{j} \| J^i \| |\lambda|^{(j-1-i)(n+2)}
\]

\[
+ \sum_{i=j_0+1}^{j} \| J^i \| |\lambda|^{(j-1-i)(n+2)}, \quad j > j_0,
\]

or

\[
(2.42) \quad \left\| \sum_{i=0}^{j-1} J^i \lambda^{(j-1-i)(n+2)} \right\| \leq \sum_{i=0}^{j} \| J^i \| \sum_{i=j_0+1}^{j} |\lambda|^{(j-1-i)(n+2)}.
\]
Hence the total error remains bounded as \( j \to \infty \) when \( |\lambda| < 1 \).

From the foregoing, it is seen that a desirable feature of the procedure is that the spectral radius of the matrix \( A \), i.e., the maximum of the moduli of the eigenvalues of \( A \), be no greater than one, and that eigenvalues of modulus one be simple. That the spectral radius of \( A \) is not less than one is shown by the following theorem.

**Theorem 2.2.** The matrix \( A \), as defined by (2.32), has an eigenvalue of one and hence, the spectral radius of \( A \) is at least one.

**Proof:** From (2.32) it is evident that \( A \) is a finite product of matrices \( A_k^\ell \), \( k = 1, \ldots, n-1 \), whose elements are given by (2.29). Referring to (2.29), it is seen that the sum of the elements of each of the latter rows of such a matrix is one. However, this is also true for the first row because under the condition \( C_{k0^\ell} = 0 \), as given by (2.9b),

\[
(2.43\text{a}) \quad \sum_{j=0}^{n} a_{kj} = 0, \quad k = 1, \ldots, n-1,
\]

and, since \( a_{kn} = -1 \),

\[
(2.43\text{b}) \quad \sum_{j=0}^{n-1} a_{kj} = 1.
\]
But this latter sum is precisely that of the elements of the first row of the matrix $A$. Thus, every matrix $A_{k}$, $k=1,\ldots,n-1$, has the property that the sum of each of its rows is one. It is easily shown that the product of such matrices enjoys this property also and that if the row (or column) sums of a matrix are the same, the matrix has an eigenvalue of that common sum.

With sufficient free parameters available, the following theorem is useful in guaranteeing that an eigenvalue of $A$ is zero, thus reducing the task of assuring that the spectral radius of $A$ is no greater than one.

**Theorem 2.3.** The matrix $A$ of (2.32) has an eigenvalue of zero if and only if $a_{k0}=0$ for at least one of its factors, $A_{k}$, as given by (2.29).

**Proof:** From (2.32),

$$
\begin{align*}
(2.44) \quad \det A = \prod_{k=1}^{n-1} \det A_{k}
\end{align*}
$$

and hence $\det A = 0$ if and only if $\det A_{k} = 0$ for some $k$.

Referring to the definition of $A_{k}$ as given by (2.29), all the elements in the last column of $A_{k}$ other than $a_{k0}$ are zero. Deleting the row and column in which $a_{k0}$ appears from the matrix $A_{k}$ leaves an identity matrix. Hence, $\det A_{k} = \pm a_{k0}$.

**Definition 2.10.** A set of modification formulas,
will be said to be compatible if and only if the product, $A$, of the associated propagation matrices $A_k$, $k = 1, \ldots, n-1$ as given by (2.29), is such that its spectral radius is one, and all eigenvalues of $A$ of modulus one are simple.

In concluding this chapter, note that a generalization of the matrix $A_k$ as given in Definition 2.7 is $A_k(h) = \{u_{ij}(h)\}$ where, referring to (2.16),

\[(2.45) \quad u_{i,n-j}(h) = \begin{cases} a_{kj} + (b_{kj} \alpha_{kj} + b_{kj})h + b_{kn} h_{kj} h^2, & i = 1 \\ \delta_{i-1,j} & i = 2, \ldots, n; \quad j = 1, \ldots, n, \end{cases}\]

which reduces to $A_k$ when $h = 0$. Corresponding to the matrix $A$ of (2.32a) is a matrix $A(h)$ defined in the same manner.

If the eigenvalues of $A(h)$ are treated as functions of $h$, they can be expanded as Taylor's series about $h = 0$. Restricting attention to those eigenvalues which lie on the unit circle when $h = 0$, the parameters can be selected such that the linear terms of the respective series are zero and the quadratic terms negative at least in the three back-point case for both doubling and halving an increment. Thus, the eigenvalues of modulus one of $A(h)$ when $h = 0$, move inside
the unit circle for \( \hat{n} \neq 0 \). However, analogous to \( A^j \)
in (2.35), is the product \( \prod_{i=0}^{j-1} A(\lambda^j \hat{n}) \) in the more general case which did not encourage further pursuit.
CHAPTER III: THE TWO BACK-POINT CASE

In the case of two back-points, only one pair of predictor and corrector formulas is involved. For notational convenience then, the subscript $k$, which in chapter two is used to designate one of several pairs of formulas, will be dropped. If the formulas are to be of order three, referring to the equations of (2.9) with $n=2$ and $k$ suppressed, it is necessary that $\gamma_i = 0, \delta_i = 0, i=0,...,3$. Selecting $b_2$ as the only free parameter, these conditions can be written respectively in matrix form as

\[(3.1a)\]
\[
\begin{bmatrix}
1 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 \\
0 & 1 & 0 & 2 \\
0 & 1 & 0 & 3
\end{bmatrix}
\begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\beta_0 \\
\beta_1
\end{bmatrix}
= \begin{bmatrix}
1 \\
1+\lambda \\
(1+\lambda)^2 \\
(1+\lambda)^3
\end{bmatrix}
\]

and

\[(3.1b)\]
\[
\begin{bmatrix}
1 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 \\
0 & 1 & 0 & 2 \\
0 & 1 & 0 & 3
\end{bmatrix}
\begin{bmatrix}
a_0 \\
a_1 \\
b_0 \\
b_1
\end{bmatrix}
= \begin{bmatrix}
1 \\
1+\lambda \\
(1+\lambda)^2 \\
(1+\lambda)^3
\end{bmatrix} - \begin{bmatrix}
0 \\
1+\lambda \\
(1+\lambda)^2 \\
(1+\lambda)^3
\end{bmatrix}b_2
\]

The inverse of the coefficient matrix of (3.1) is readily
verified to be

\[
(3.2) \quad \begin{bmatrix}
1 & 0 & -3 & 2 \\
0 & 0 & 3 & -2 \\
0 & 1 & -2 & 1 \\
0 & 0 & -1 & 1 \\
\end{bmatrix} = \begin{bmatrix}
1 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 \\
0 & 1 & 0 & 2 \\
0 & 1 & 0 & 3 \\
\end{bmatrix}^{-1}
\]

which, after a suitable amount of algebraic manipulation, yields the following solutions to (3.1a) and (3.1b) respectively.

\[
(3.3a) \quad \begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\beta_0 \\
\beta_1 \\
\end{bmatrix} = \begin{bmatrix}
\lambda(2\lambda+3) \\
-(\lambda+1)^2(2\lambda+1) \\
\lambda(\lambda+1) \\
\lambda(\lambda+1)^2 \\
\end{bmatrix}
\]

and

\[
(3.3b) \quad \begin{bmatrix}
a_0 \\
a_1 \\
b_0 \\
b_1 \\
\end{bmatrix} = \begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\beta_0 \\
\beta_1 \\
\end{bmatrix} + \begin{bmatrix}
-6\lambda(\lambda+1) \\
6\lambda(\lambda+1) \\
-\lambda(3\lambda+2) \\
-(\lambda+1)(3\lambda+1) \\
\end{bmatrix}^{b_2}
\]

With
the third order error constants \( \Gamma_4 \) and \( C_4 \) can be determined from (2.9) to be

\[(3.5a) \quad \Gamma_4 = -\lambda^2 (\lambda+1)^2 \]

and

\[(3.5b) \quad C_4 = \Gamma_4 + mb_2.\]

From Theorem 2.1, when \( m + \Gamma_4 \neq 0 \), the value of \( b_2 \) that minimizes the magnitude of

\[(5.6a) \quad E(c^2 + hbg^2) = g + \Gamma_2^2 \]

is

\[(3.6b) \quad \hat{b}_2 = -\frac{\Gamma_4}{m + \Gamma_4} , \]

and that which minimizes

\[(3.7a) \quad E\left\{ (C_4 + \bar{h}b_2 \Gamma_4)^2 \right\} = \left( \bar{\Gamma}_4 + \Gamma_4 + mb_2 \right)^2 + b_2^2 \Gamma_4^2 \sigma^2 \]

is
Note that when $\bar{n} = 0$, the value of $\hat{b}_2$ as given by (3.6b) implies that $C_4 = 0$, i.e., the order of the corrector is at least four.

In the two back-point case there is a single propagation matrix. Thus, corresponding to (2.32a),

\[(3.8) \quad A = \begin{bmatrix} a_1 & a_0 \\ 1 & 0 \end{bmatrix} \]

The characteristic equation of the propagation matrix is readily seen to be

\[(3.9) \quad \mu^2 - a_1 \mu - a_0 = 0.\]

Making use of the fact that

\[(3.10) \quad a_0 + a_1 = 1,\]

the roots of (3.9) are found to be one and $-a_0$. Thus, the spectral radius of the propagation matrix is one and the roots of (3.9) are simple if and only if $-1 < a_0 \leq 1$. It follows from the equations of (3.3), that this condition can be stated, in terms of the parameter $b_2$, as
(3.11) \[
\frac{(\lambda+1)(2\lambda+1)}{6\lambda} \leq b_2 < \frac{\lambda(2\lambda+3)+1}{6\lambda(\lambda+1)}
\]

For halving, \(\lambda = 1/2\) and for doubling, \(\lambda = 2\), in which case (3.3a) and (3.3b) are respectively,

\[
(3.12a) \quad \begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\beta_0 \\
\beta_1
\end{bmatrix} = \begin{bmatrix}
1 \\
0 \\
3/8 \\
9/8
\end{bmatrix} \begin{bmatrix}
28 \\
-27 \\
12 \\
18
\end{bmatrix}
\]

\(\lambda = 1/2, \lambda = 2\)

and

\[
(3.12b) \quad \begin{bmatrix}
a_0 \\
a_1 \\
b_0 \\
b_1
\end{bmatrix} - \begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\beta_0 \\
\beta_1
\end{bmatrix} = \begin{bmatrix}
-9/2 \\
9/2 \\
-7/4 \\
-15/4
\end{bmatrix} \begin{bmatrix}
b_2 \\
b_2 \\
b_2 \\
b_2
\end{bmatrix}
\]

\(\lambda = 1/2, \lambda = 2\).
Corresponding to (3.5a) and (3.5b) are the error constants

\[(3.13a) \quad \hat{\gamma}_4 = -9/16, \lambda = 1/2; \quad \hat{\gamma}_4 = -36, \lambda = 2,\]

and

\[(3.13b) \quad c_4 = -9/16 + 3b_2, \lambda = 1/2; \quad c_4 = -36 + 60b_2, \lambda = 2,\]

The respective values of \(b_2\), as given by (3.6b), are for \(\lambda = 1/2\) and \(\lambda = 2\),

\[(3.14) \quad \hat{b}_2 = \frac{3}{16 - 3h}, \lambda = 1/2; \quad \hat{b}_2 = \frac{3}{5 - 3h}, \lambda = 2.\]

From (3.11), the ranges of \(\hat{b}_2\) for which the spectral radius of the propagation matrix is one is

\[(3.15) \quad 0 \leq \hat{b}_2 < 4/9, \lambda = 1/2; \quad 3/4 < \hat{b}_2 < 29/36, \lambda = 2\]

Note that the values 3/16 and 3/5 respectively of \(\hat{b}_2\) corresponding to \(h = 0\) lies within the interval defined in (3.15) when \(\lambda = 1/2\) and without when \(\lambda = 2\). This is in contrast to the situation when \(\lambda = 1\). In the latter case, the interval given by (3.11) is \([1/3, 1/2]\), the value of \(\hat{b}_2\) is 1/3 and the second root of the characteristic equation of the propagation matrix lies on the unit circle at -1.

The unique third order predictor coefficient for halving and doubling an interval are given in Table 3.1.
In Table 3.2 two sets of corrector coefficients are given each for doubling and for halving. The first set in each case corresponds to the value of $b_2$ given by (3.14) when $h = 0$ which implies fourth order formulas. The second set in each case of halving and doubling corresponds to the value of $b_2$ chosen from the midpoint of the intervals of (3.15) and lead to the second root of the respective characteristic equation (3.10) being zero. The fourth order corrector constants are unique.

Table 3.1. Third order predictor coefficients for halving ($\lambda = 1/2$) and doubling ($\lambda = 2$) an increment.

<table>
<thead>
<tr>
<th></th>
<th>$\lambda = 1/2$</th>
<th>$\lambda = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0</td>
<td>-27</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>3/8</td>
<td>12</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>7/8</td>
<td>18</td>
</tr>
<tr>
<td>$\Gamma_4$</td>
<td>-9/16</td>
<td>-36</td>
</tr>
</tbody>
</table>
Table 3.2. Fourth and third order corrector coefficients for halving ($\lambda = 1/2$) and doubling ($\lambda = 2$) an increment.

<table>
<thead>
<tr>
<th></th>
<th>$\lambda = 1/2$</th>
<th></th>
<th>$\lambda = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>5/32</td>
<td>0</td>
<td>32/5</td>
</tr>
<tr>
<td>$a_1$</td>
<td>27/32</td>
<td>1</td>
<td>-27/5</td>
</tr>
<tr>
<td>$b_0$</td>
<td>3/64</td>
<td>-1/72</td>
<td>12/5</td>
</tr>
<tr>
<td>$b_1$</td>
<td>27/64</td>
<td>1/12</td>
<td>27/5</td>
</tr>
<tr>
<td>$b_2$</td>
<td>3/16</td>
<td>2/9</td>
<td>3/5</td>
</tr>
<tr>
<td>$c_4$</td>
<td>0</td>
<td>5/48</td>
<td>0</td>
</tr>
</tbody>
</table>
Two pairs of predictor and corrector formulas are required to effect increment modification when three back-points are in use. The subscript $k$ will again be used exclusively to designate a particular pair; $k=1$ implies the pair requiring the three back-points to be spaced by two intervals of size $h$ and which steps the numerical solution ahead by an amount $\lambda h$; $k=2$ implies the pair requiring the three back-points to be spaced by one interval of size $h$ and one of size $\lambda h$ and which also steps the numerical solution ahead by an amount $\lambda h$. All formulas will be of order at least four which implies one free parameter for each of the predictor equations and two free parameters for each of the corrector equations.

That the formulas be of order four implies

\begin{align*}
(4.1a) & \quad \sum_{ki} = 0, \\
(4.1b) & \quad c_{ki} = 0, \quad i = 0, \ldots, 4; \quad k = 1, 2.
\end{align*}

For the two predictor formulas, $x_{k0}, \quad k = 1, 2$ will be chosen to be the respective free parameters and for the two correctors, $a_{k0}$ and $b_{k3}, \quad k = 1, 2$, respectively. Referring to the
equations of (2.9), the conditions of (4.1) can be represented in matrix notation as follows. Defining:

\[
\begin{bmatrix}
1 & 1 & 0 & 0 & 0 \\
1 & p_k & 1 & 1 & 1 \\
1 & p_k^2 & 0 & 2 & 2p_k \\
1 & p_k^3 & 0 & 3 & 3p_k^2 \\
1 & p_k^4 & 0 & 4 & 4p_k^3
\end{bmatrix}
\]

where

(4.2b) \( p_k = (k-1)\lambda + (3-k) \);

(4.2c) \( \tilde{a}_k = (a_{k1}, a_{k2}, b_{k0}, b_{k1}, b_{k2})' \),

(4.2d) \( \tilde{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, \beta_{k0}, \beta_{k1}, \beta_{k2})' \),

(4.2e) \( \tilde{u}_k = (1, q_k, q_k^2, q_k^3, q_k^4)' \),

(4.2f) \( \tilde{v}_k = (1, 0, 0, 0, 0)' \),

and

(4.2g) \( \tilde{w}_k = (0, 1, 2q_k, 3q_k^2, 4q_k^3)' \),

where
(4.2h) \( q_k = k\lambda + (3-k), \quad k = 1, 2; \)

the conditions \( \sum_{k=1}^{4} \) are given by the matrix equation

\[
(4.3a) \quad B_k \hat{\alpha}_k = \hat{u}_k - \alpha_{k0} \hat{v}_k, \quad k = 1, 2;
\]

the conditions \( c_{ki} = 0, \quad i = 0, \ldots, 4, \)

by

\[
(4.3b) \quad B_k \hat{a}_k = \hat{u}_k - a_{k0} \hat{v}_k - b_k \hat{w}_k, \quad k = 1, 2.
\]

It can readily be verified that

\[
(4.4a) \quad \det B_1 = 12,
\]

and less readily that

\[
(4.4b) \quad \det B_2 = 2\lambda^4 (\lambda+1)(\lambda+2),
\]

hence, for positive \( \lambda \), unique solutions to the equations of (4.3) exist. These equations, which involve a considerable amount of algebraic manipulation, have been solved by Cramer's method. Their solutions follow. Defining:
\[ (4.5a) \quad \frac{1}{r_1} = \begin{bmatrix} 12\lambda^2(\lambda+2)^2 \\ -12(\lambda+1)^2(\lambda^2+2\lambda-1) \\ -\lambda(\lambda+1)^2 \\ 4\lambda^2(\lambda+1)(2\lambda+5) \\ \lambda(\lambda+1)^2(5\lambda+12) \end{bmatrix} \]

\[ (4.5b) \quad \frac{1}{r_2} = \begin{bmatrix} 2\lambda(\lambda+1)(17\lambda+10) \\ -16\lambda(\lambda+1)(2\lambda+1) \\ -4\lambda^8 \\ 4\lambda^5(\lambda+1)(\lambda^2+4\lambda+2) \\ 4\lambda^5(4\lambda^2+9\lambda+4) \end{bmatrix} \]

\[ (4.5c) \quad \frac{1}{r_3} = \begin{bmatrix} 0 \\ -12 \\ 4 \\ 16 \\ 4 \end{bmatrix} \]
\[ (4.5d) \]

\[
\begin{bmatrix}
-2\lambda(\lambda-1)(\lambda+1)^4 \\
-2\lambda(\lambda+1)(2\lambda+1) \\
(\lambda+1)^2 \\
\lambda(\lambda+1)^2 \\
\lambda^2(\lambda+1)^2
\end{bmatrix}
\]

\[ (4.5e) \]

\[
\begin{bmatrix}
-48\lambda(\lambda+1)(\lambda+2) \\
48\lambda(\lambda+1)(\lambda+2) \\
2\lambda(\lambda+1)(2\lambda+1) \\
-4\lambda(\lambda+2)(8\lambda+5) \\
-2(\lambda+1)(\lambda+2)(10\lambda+3)
\end{bmatrix}
\]
(4.5f) \[
\begin{bmatrix}
-48\lambda^3(\lambda+1)(2\lambda+1) \\
48\lambda^3(\lambda+1)(2\lambda+1) \\
12\lambda^7 \\
-2\lambda(\lambda+1)(2\lambda+1)(3\lambda+10) \\
-4\lambda(2\lambda+1)(5\lambda+8)
\end{bmatrix};
\]

the solutions to (4.3a) are given by

(4.6a) \( (\det B_k) \alpha_k = \hat{\alpha}_k + \alpha \hat{s}_k \),

those of (4.3b) are given by

(4.6b) \( (\det B_k) \hat{\alpha}_k = \hat{\alpha}_k + \alpha \hat{s}_k + b \hat{t}_k \), \( k = 1, 2 \).

If \( \lambda \neq 0, -1, -2 \), in which case \( \det B \neq 0 \), the constants \( \Gamma_{k5} \) and \( C_{k5} \), \( k = 1, 2 \), are found to be

(4.7a) \( \Gamma_{k5} = \hat{\alpha}_k + \hat{s}_k \alpha_{k0} \),

and

(4.7b) \( C_{k5} = \hat{\alpha}_k + \hat{s}_k \alpha_{k0} + \hat{t}_k \beta_{k3} \), \( k = 1, 2 \),

where
\[
\begin{align*}
(4.7c) \quad \tilde{r}_k &= \begin{cases} 
-\lambda^2(\lambda+1)^2(3\lambda+7)/3, & k = 1, \\
-2\lambda(4\lambda^2+11\lambda+5)/(\lambda+2), & k = 2;
\end{cases} \\
(4.7d) \quad \tilde{s}_k &= \begin{cases} 
4/3, & k = 1, \\
(\lambda+1)^3/(2\lambda+4), & k = 2;
\end{cases} \\
\text{and} \\
(4.7e) \quad \tilde{t}_k &= \begin{cases} 
\lambda(\lambda+1)(\lambda+2)(15\lambda+7)/3, & k = 1, \\
2\lambda^3(2\lambda+1)(7\lambda+15)/(\lambda+2), & k = 2.
\end{cases}
\end{align*}
\]

In reference to the equations of (2.22),

\[
(4.8a) \quad m_k = \tilde{t}_k, \quad k = 1, 2,
\]

and

\[
(4.8b) \quad \tilde{C}_k = \tilde{s}_ka_k\tilde{t}_kb_k, \quad k = 1, 2.
\]

Now, the product of the propagation matrices as given by (2.32) is seen to be
the characteristic polynomial of which is $P$, where

$$(4.10) \quad P(\mu) = -\mu^3 + \mu^2(a_{11} + a_{12}a_{22} + a_{21})$$

$$+ \mu(a_{10}a_{22} - a_{11}a_{21} + a_{12}a_{20} + a_{10}a_{20}).$$

By Theorem 2.3, one is a zero of $P(\mu)$. Making use of the identity

$$(4.11) \quad a_{k0} + a_{k1} + a_{k2} = 1, \quad k = 1, 2,$$

$P$ can be factored as

$$(4.12) \quad P(\mu) = -(\mu - 1)[\mu^2 + (1 - a_{11} - a_{12}a_{22} - a_{21})\mu + a_{10}a_{20}].$$

In the event $a_{10}$ or $a_{20}$ is zero, zero is a zero of $P$ which leaves the negative of the coefficient of $\mu$ in the quadratic factor as the third zero. Thus, with $a_{10}$ or $a_{20}$ zero, the zeroes of $P$ will be within or on the unit circle, those on being simple, if and only if

$$(4.13) \quad 0 \leq a_{11} + a_{12}a_{22} + a_{21} < 2.$$
In halving an increment, \( \lambda = 1/2 \); the solutions, \((4.6)\) to the equations \((4.3)\) are given in matrix form by

\[
(4.14a)
\begin{align*}
\begin{bmatrix}
\alpha_{11} & a_{11} \\
\alpha_{12} & a_{12}
\end{bmatrix}
\begin{bmatrix}
25/16 \\
-9/16
\end{bmatrix}
\begin{bmatrix}
0 \\
-1
\end{bmatrix}
\begin{bmatrix}
-15/2 \\
15/2
\end{bmatrix}
\end{align*}
\]

and

\[
(4.14b)
\begin{align*}
\begin{bmatrix}
\alpha_{21} & a_{21} \\
\alpha_{22} & a_{22}
\end{bmatrix}
\begin{bmatrix}
37/5 \\
-32/5
\end{bmatrix}
\begin{bmatrix}
27/5 \\
-32/5
\end{bmatrix}
\begin{bmatrix}
-192/5 \\
192/5
\end{bmatrix}
\end{align*}
\]
Also,

\[(4.14c)\]
\[
\begin{align*}
  \left( I_{15}^*, c_{15} \right) &= \left( -\frac{51}{32} \right) (1,1) + \left( \frac{4}{3} \right) (a_{10}^*, a_{10}^*) + \left( \frac{145}{16} \right) (0, b_{13}^*), \\
  \left( I_{25}^*, c_{25} \right) &= \left( -\frac{23}{40} \right) (1,1) + \left( \frac{27}{40} \right) (a_{20}^*, a_{20}^*) + \left( \frac{37}{10} \right) (0, b_{23}^*).
\end{align*}
\]

Finally, corresponding to \((2.22c)\) are

\[(4.15a)\]
\[
\hat{b}_{13} = \frac{-153 + 128 \alpha_{10}}{870 + (-153 + 128 \alpha_{10}) \hat{h}} + \frac{(-153 + 128 \alpha_{10})^2 \varphi^2}{870 + (-153 + 128 \alpha_{10}) \hat{h}}
\]

and

\[(4.15b)\]
\[
\hat{b}_{23} = \frac{-23 + 27 \alpha_{20}}{148 + (-23 + 27 \alpha_{20}) \hat{h}} + \frac{(-23 + 27 \alpha_{20})^2 \varphi^2}{148 + (-23 + 27 \alpha_{20}) \hat{h}}.
\]

The values of \(\hat{b}_{b3}^*, k = 1, 2\), can be obtained by setting \(\varphi^2 = 0\) in \((4.14a)\) and \((4.14b)\) respectively.

In Table 4.1 is presented selected predictor-coefficient sets. The relatively large numerical values of the fifth order predictor for \(h = 2\) are a disadvantage of this set if round-off error is not substantially smaller than the discretization error. In Table 4.2 is presented two coefficient sets of compatible correctors. Referring to
(4.12) the compatibility of the first set is established. For the second set, the zeroes of the quadratic factor of (4.11) must be computed.

The asterisks indicate the coefficients of the formulas used in the sample problem described in the appendix.

Table 4.1. Selected predictor coefficient sets for halving an increment; order at least four.

<table>
<thead>
<tr>
<th>$\lambda = 1/2$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{k0}$</td>
<td>153/128</td>
<td>7/16</td>
</tr>
<tr>
<td>$\alpha_{k1}$</td>
<td>25/16</td>
<td>25/16</td>
</tr>
<tr>
<td>$\alpha_{k2}$</td>
<td>-225/128</td>
<td>-1</td>
</tr>
<tr>
<td>$\beta_{k0}$</td>
<td>45/128</td>
<td>19/192</td>
</tr>
<tr>
<td>$\beta_{k1}$</td>
<td>75/32</td>
<td>4/3</td>
</tr>
<tr>
<td>$\beta_{k2}$</td>
<td>225/128</td>
<td>289/192</td>
</tr>
<tr>
<td>$\Gamma_{k5}$</td>
<td>0</td>
<td>-97/96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4.2. Selected compatible corrector coefficient sets for halving an increment; order at least four.

<table>
<thead>
<tr>
<th>$\lambda = 1/2$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{k0}$</td>
<td>$1/16$</td>
<td>0</td>
<td>$1/8$</td>
<td>$-1/8$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_{k1}$</td>
<td>$5/16$</td>
<td>$53/37$</td>
<td>$5/8$</td>
<td>$13/40$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_{k2}$</td>
<td>$5/8$</td>
<td>$-16/37$</td>
<td>$1/4$</td>
<td>$4/5$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$b_{k0}$</td>
<td>$1/64$</td>
<td>$-1/444$</td>
<td>$5/192$</td>
<td>$-3/80$</td>
<td>$1/192$</td>
<td>$1/192$</td>
</tr>
<tr>
<td>$b_{k1}$</td>
<td>$5/24$</td>
<td>$10/37$</td>
<td>$43/96$</td>
<td>$-41/240$</td>
<td>$-1/32$</td>
<td>$-7/96$</td>
</tr>
<tr>
<td>$b_{k2}$</td>
<td>$35/64$</td>
<td>$88/111$</td>
<td>$149/192$</td>
<td>$31/60$</td>
<td>$61/192$</td>
<td>$3/8$</td>
</tr>
<tr>
<td>$b_{k3}$</td>
<td>$1/6$</td>
<td>$23/148$</td>
<td>$1/8$</td>
<td>$1/6$</td>
<td>$5/24$</td>
<td>$37/192$</td>
</tr>
<tr>
<td>$c_{k5}$</td>
<td>0</td>
<td>0</td>
<td>$-113/384$</td>
<td>$-41/960$</td>
<td>$113/384$</td>
<td>$53/384$</td>
</tr>
</tbody>
</table>
Doubling An Increment

In doubling an increment, \( \lambda = 2 \); the solutions, (4.6), to the equations (4.3) are given in matrix form by

\[(4.16a)\]

\[
\begin{bmatrix}
\alpha_{11} & a_{11} \\
\alpha_{12} & a_{12} \\
\beta_{10} & b_{10} \\
\beta_{11} & b_{11} \\
\beta_{12} & b_{12}
\end{bmatrix}
= 
\begin{bmatrix}
64 \\
-63 \\
-3 \\
36 \\
33
\end{bmatrix}
+ 
\begin{bmatrix}
0 \\
-1 \\
1/3 \\
4/3 \\
1/3
\end{bmatrix}
+ 
\begin{bmatrix}
96 \\
96 \\
5 \alpha_{10}^{10} \\
-56 \\
-46
\end{bmatrix}
\]

and

\[(4.16b)\]

\[
\begin{bmatrix}
\alpha_{21} & a_{21} \\
\alpha_{22} & a_{22} \\
\beta_{20} & b_{20} \\
\beta_{21} & b_{21} \\
\beta_{22} & b_{22}
\end{bmatrix}
= 
\begin{bmatrix}
11 \\
-10 \\
-8/3 \\
14 \\
38/3
\end{bmatrix}
+ 
\begin{bmatrix}
-27/32 \\
-5/32 \\
3/8 \alpha_{20}^{20} \\
27/32 \\
3/32
\end{bmatrix}
+ 
\begin{bmatrix}
-15 \\
15 \\
4 \beta_{23} \\
-20 \\
-15
\end{bmatrix}
\]
The quantities $\Gamma_{k5}$ and $c_{k5}$, $k=1,2$, are given by

\begin{equation}
\Gamma_{15}c_{15} = (-156)(1,1) + (4/3)(\alpha_{10},a_{10}) + (296)(0,b_{13})
\end{equation}

and

\begin{equation}
\Gamma_{25}c_{25} = (-344)(1,1) + (27/8)(\alpha_{20},a_{20}) + (580)(0,b_{23})
\end{equation}

Likewise, $\bar{b}_{k3}$, $k=1,2$, is given by

\begin{equation}
\bar{b}_{13} = \frac{-117+a_{10}}{\left[222+(-117+a_{10})h\right] + \frac{(-117+a_{10})^2}{222+(-117+a_{10})h}}
\end{equation}

and

\begin{equation}
\bar{b}_{23} = \frac{-2752+27a_{20}}{\left[4640+(-2752+27a_{20})h\right] + \frac{(-2752+27a_{20})^2}{4640+(-2752+27a_{20})h}}
\end{equation}

Again, the values of $\bar{b}_{k3}$, $k=1,2$, can be recovered from (4.17a) and (4.17b) respectively by setting $\sigma^2 = 0$.

The fifth order predictor coefficients, i.e., those such that $\Gamma_{k5} = 0$, $k=1,2$, are not practical unless round-off error is negligible. Referring to (4.14a), it is seen that $\alpha_{11} = 64$ which forces numerically large values of the
coefficients even in the fourth order case. This is a disadvantage of these formulas. However, doubling occurs when $\bar{h}$ is relatively small, thus, if doubling does not occur for $|\bar{h}| > .1$, from (2.16), it is evident that roughly one digit less accuracy can be tolerated with computations involving the predictor than those involving the corrector. Also, if enough significant digits are present so that the round-off error is a hundredth less than the discretization error in magnitude, round-off error should present no problem with the fourth order coefficients. In Table 4.3 is presented predictor coefficients for doubling an increment.

Table 4.4 contains three pairs of predictor coefficients. It can be shown that, with $\lambda = 2$, no fifth order corrector pairs are compatible. This can be done directly by setting $C_{1} = 0$, $k = 1,2$ in (4.14c), writing the corrector coefficients of (4.14a) and (4.14b) respectively in terms of one parameter each and showing that for all values of these parameters, the zeroes of the quadratic factor of (4.11) exceed one in modulus.
Table 4.3. Selected predictor coefficients for doubling an increment; order at least four.

<table>
<thead>
<tr>
<th>$\lambda = 2$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{k0}$</td>
<td>117</td>
<td>-36</td>
</tr>
<tr>
<td>$\alpha_{k1}$</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>$\alpha_{k2}$</td>
<td>-180</td>
<td>-27</td>
</tr>
<tr>
<td>$\beta_{k0}$</td>
<td>36</td>
<td>-15</td>
</tr>
<tr>
<td>$\beta_{k1}$</td>
<td>192</td>
<td>-12</td>
</tr>
<tr>
<td>$\beta_{k2}$</td>
<td>72</td>
<td>21</td>
</tr>
<tr>
<td>$\Gamma_{k5}$</td>
<td>0</td>
<td>-204</td>
</tr>
</tbody>
</table>
Table 4.4. Selected corrector coefficients for doubling an increment, the latter two pairs compatible; order at least four.

<table>
<thead>
<tr>
<th>$\lambda = 2$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{k0}$</td>
<td>-13/4</td>
<td>-32/27</td>
<td>1/3</td>
<td>1/3</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$a_{k1}$</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>55/32</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_{k2}$</td>
<td>-31/4</td>
<td>-22/27</td>
<td>2/3</td>
<td>-101/96</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$b_{k0}$</td>
<td>-11/8</td>
<td>-32/45</td>
<td>4/9</td>
<td>-17/120</td>
<td>1/3</td>
<td>4/15</td>
</tr>
<tr>
<td>$b_{k1}$</td>
<td>4/3</td>
<td>1</td>
<td>-8/9</td>
<td>73/32</td>
<td>-4/3</td>
<td>-2/3</td>
</tr>
<tr>
<td>$b_{k2}$</td>
<td>7</td>
<td>32/9</td>
<td>22/9</td>
<td>355/96</td>
<td>7/3</td>
<td>5/3</td>
</tr>
<tr>
<td>$b_{k3}$</td>
<td>13/24</td>
<td>3/5</td>
<td>2/3</td>
<td>3/5</td>
<td>2/3</td>
<td>11/15</td>
</tr>
<tr>
<td>$c_{k5}$</td>
<td>0</td>
<td>0</td>
<td>376/9</td>
<td>41/8</td>
<td>124/3</td>
<td>244/3</td>
</tr>
</tbody>
</table>


ACKNOWLEDGEMENT

The author wishes to express his sincere appreciation to Dr. Robert J. Lambert for his guidance and encouragement during the preparation of this thesis and to his gracious wife, Kathryn, the vast majority of whose contributions have never been realized, much less acknowledged.
The fourth order formulas indicated by asterisks in chapter four were used to obtain numerical solutions, over the interval $[0,1.0]$ of the initial value problem

(A.1) \[ y' = -2xy^2, \quad (x_0, y_0) = (0,1), \]

the solution of which is

(A.2) \[ y = (1+x^2)^{-1}. \]

This problem is mentioned in Hamming (6, p. 209) as one whose solution is often troublesome to approximate by polynomials.

In generating the numerical solution, the known solution was used to furnish the necessary initial back-points. In the two examples, the initial increment was 1/32 and 1/64 respectively. The numerical solutions were generated by alternately applying the two formulas which serve to double an increment and then the two which halve. Thus, in the first example, an increment of 1/64 exists relative to the doubling formulas and an increment of 1/32 relative to the halving formulas. Similarly in the second example, the increments are 1/32 and 1/16 respectively.

Obviously,

(A.3) \[ f_y = -4x(1+x^2)^{-1} \]
which attains a minimum of -2 at $x = 1$. Also,

\[(A.4)\]  
\[y^{(5)}(x) = xy^4(-3 - 16x^2y - 16x^4y^4)\]

which has the value 60 at $x = 1$. Referring to (2.17) and the coefficient tables of chapter four, the respective error terms of the four predictor and corrector pairs are

\[(A.5)\]  
\[
\frac{1}{5!}(c_k + h b_{k3} k_5) y^{(5)}(x) h^5
\]

\[
= \begin{cases} 
\frac{1}{5!}\left[(-113/384) + (1/8)(-97/96)h\right] y^{(5)}(x) h^5, & \lambda = 1/2, k = 1; \\
\frac{1}{5!}\left[(-41/960) + (1/6)(-109/80)h\right] y^{(5)}(x) h^5, & \lambda = 1/2, k = 2; \\
\frac{1}{5!}\left[(124/3) + (2/3)(-204)h\right] y^{(5)}(x) h^5, & \lambda = 2, k = 1; \\
\frac{1}{5!}\left[(244/3) + (11/15)(-317)h\right] y^{(5)}(x) h^5, & \lambda = 2, k = 2; 
\end{cases}
\]

The latter truncation term is approximately .0000015 when $x = 1$ and $h = 1/32$; this is to be compared with the solution of the second example.

The computations were performed in single precision floating point on an IBM 360 which implies about seven digits are carried in each computation. Partial results are tabulated in Table A.1.
Table A.1. Selected values of numerical solution to
sample problem (A.1).

<table>
<thead>
<tr>
<th>X</th>
<th>Y True</th>
<th>Predicted</th>
<th>Corrected</th>
<th>Predicted</th>
<th>Corrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/64</td>
<td>.999756</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/64</td>
<td>.999024</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/64</td>
<td>.996109</td>
<td>.996124</td>
<td>.996109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6/64</td>
<td>.991287</td>
<td>.991288</td>
<td>.991287</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/64</td>
<td>.988178</td>
<td>.988177</td>
<td>.988178</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/64</td>
<td>.984615</td>
<td>.984615</td>
<td>.984615</td>
<td>.984619</td>
<td>.984615</td>
</tr>
<tr>
<td>10/64</td>
<td>.976168</td>
<td>.976181</td>
<td>.976167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/64</td>
<td>.966038</td>
<td>.966039</td>
<td>.966037</td>
<td>.966046</td>
<td>.966036</td>
</tr>
<tr>
<td>13/64</td>
<td>.960375</td>
<td>.960373</td>
<td>.960375</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14/64</td>
<td>.954334</td>
<td>.954333</td>
<td>.954333</td>
<td>.954335</td>
<td>.954333</td>
</tr>
<tr>
<td>16/64</td>
<td>.941176</td>
<td>.941193</td>
<td>.941176</td>
<td>.941175</td>
<td>.941175</td>
</tr>
<tr>
<td>1/2</td>
<td>.800000</td>
<td>.799999</td>
<td>.799999</td>
<td>.800140</td>
<td>.799991</td>
</tr>
<tr>
<td>1</td>
<td>.500000</td>
<td>.500001</td>
<td>.499998</td>
<td>.499982</td>
<td>.499986</td>
</tr>
<tr>
<td>2</td>
<td>.200000</td>
<td>.199999</td>
<td>.199999</td>
<td>.200148</td>
<td>.199974</td>
</tr>
<tr>
<td>4</td>
<td>.0588235</td>
<td>.0588226</td>
<td>.0588221</td>
<td>.0588182</td>
<td>.0588183</td>
</tr>
<tr>
<td>6</td>
<td>.0270270</td>
<td>.0270265</td>
<td>.0270266</td>
<td>.0270259</td>
<td>.0270258</td>
</tr>
<tr>
<td>8</td>
<td>.0153846</td>
<td>.0153843</td>
<td>.0153843</td>
<td>.0153843</td>
<td>.0153842</td>
</tr>
</tbody>
</table>
| 10  | .00990099| .00990045| .00990070| .00990073| .00990073