ON THE NUMERICAL SOLUTION OF SPECIAL SECOND-ORDER INITIAL VALUE PROBLEMS.

Thesis presented to

THE SWISS FEDERAL INSTITUTE OF TECHNOLOGY, ZURICH

for the degree of
Doctor of Mathematics

By

WILLIAM PERRY TIMLAKE

M.A., University of North Carolina
Citizen of the UNITED STATES

Accepted on the Recommendation of

Prof. DR. P. HENRICI and Prof. DR. H. RUTISHAUSER

MURRAY HILL
BELL TELEPHONE LABORATORIES, INCORPORATED
1964
ON THE NUMERICAL SOLUTION OF SPECIAL SECOND-ORDER INITIAL VALUE PROBLEMS

William Perry Timlake

INTRODUCTION

Consider the special second-order initial value problem

\[ X''(t) = f(t, X(t), \ldots, X^{(m)}(t)) \quad \left( n = \frac{d}{dt^2} \right) \]  

\[ X(i)(a) = a_i, \quad X'(i)(a) = b(i), \quad i = 1, \ldots, m. \]  

Since problems of this form frequently occur in mechanical problems (e.g., in orbit calculations), we limit ourselves to this particular form.

Although it is possible to write (0.1) as an equivalent first-order differential problem with the immediate applicability of the results given in [4],* it seems unnatural to introduce the computation of \( X(t) \) unless it is desired for some special use. For these reasons many methods have been devised, particularly by astronomers [6] and [8] to compute an approximation to the solution (if it exists) of (0.1).

*Numbers in brackets refer to the bibliography. A more complete list of literature pertinent to this general area is given in [3].
We shall study those finite difference schemes which, recursively from the equation

\[ A_{n+k}X_{n+k} + \ldots + A_nX_n - h^2(B_{n+k}f(t_{n+k},X_{n+k}) + \ldots + B_nf(t_n,X_n)) = 0, \]

where \( A_1 \) and \( B_1 \) are diagonal matrices, generate a sequence of vectors \( \{X_n\} \) (\( n \geq k \)) whose \( i^{th} \) component approximates \( X_i(t_n) \), the solution of (0.1). A more precise formulation of this class of methods will be given shortly. That diagonal matrices are permitted is a slight generalization over [4] in which only scalar multiples of the identity matrix are used. The basic intention is that if it is known a priori that certain of the components of (0.1) have a large derivative of a particular order but smaller values for subsequent orders, then a more refined technique may be used on this component. But at the same time a cruder (and hence less computation required) scheme may be used for the better behaved components.

Before giving a formal presentation, we offer a brief resume of the results. (For the moment it is assumed the reader will understand the terminology intuitively from its literal connotation.) In Theorem IF a necessary and sufficient condition on \( A_1, B_1 \) (\( i = 1, \ldots, k \)) is given which guarantees that the sequence \( \{X_n\} \) converges to the solution of (0.1). This condition is the natural generalization of the analogous result of Dahlquist [9]. The method of proof
follows quite closely that of [4]. In Theorem IID we obtain an asymptotic formula which depicts the "rate of convergence" in terms of the truncation error and the starting error. In particular

\[ X(i)_n = X(i)(t_n) + h^p e(i)_n(t_n) + h^q \sum_{j=1}^{d} d'_{ij} z_j X_{ij}(t_n) + O(h^r) \]

where \( X(i)_n \) is the \( i \)-th component of the vector generated by the recursion formula; \( X(i)(t_n) \) is the \( i \)-th solution of (0.1); \( e_i(t_n) \) and \( X_{ij}(t_n) \) are smooth functions; \( d'_{ij} \) is either zero or one; \( z_j \) is a complex number with absolute value 1; \( O(h^{p+2}) \) is the order of the truncation error; and \( O(h^{q+1}) \) is the order of the starting error. The method of proof follows in broad outlines that of a similar theorem in [4]. However, since we permit diagonal matrices, \( A_i \), which of course do not commute with arbitrary matrices, certain nontrivial complications arise.

In terms of this asymptotic expansion, statistical estimates of the round-off error are found. As might be expected, the mean of the accumulated round-off error is \( O(\mu/h^2) \) where \( \mu \) is a bound on the mean of the local round-off error. This dangerous growth property has been long recognized at least intuitively, and a summing technique was devised to combat this difficulty.
An extension of this technique is given in Section IV. The terminology necessary for a description of this technique would not be clear from the literal meaning, and we postpone a discussion until the introductory remarks of Section IV. In the final section two illustrative examples are given.

We now begin a formal presentation. Let I denote the interval \( \{ t \mid -\infty < a < t < b < \infty \} \) and let \( E_m \) denote the \( m \)-dimensional linear space over the complex numbers. If \( y \in E_m \) where the symbol \( \in \) is to be read "belongs to," then let

\[
|y| = \sum_{i=1}^{m} |y_i|
\]

where \( y_i \) is the \( i \)th coordinate of \( y \) and where \( |y_i| \) is the absolute value of \( y_i \). If \( A \) is an \( m \times m \) matrix with elements \( a_{ij} \) over the complex numbers, let

\[
|A| = \max_{1 \leq i, j \leq m} \sum_{j=1}^{m} |a_{ij}|
\]

We note that if \( y \in E_m \), \(|Ay| \leq |A| \cdot |y|\).

Matrices will be denoted by capital Roman letters.

Let \( f \) be a function from \( I \times E_m \) to \( E_m \). For \( t \in I \) and \( X \in E_m \), we write \( f(t, X) \) for the value of \( f \) at \( (t, X) \). \( f \) is said to be in class \( L \) (Lipschitz) provided: it is continuous and there exists a constant \( L \) such that for \( t \in I \), \( X \in E_m \) and \( y \in E_m \),

\[
|f(t, X) - f(t, y)| \leq L|X - Y|.
\]

The set of functions from \( I \times E_m \) to \( E_m \), for which all \( k \)th order
partial derivatives \( \frac{\partial^k f}{\partial t^p \partial x_1^{p_1} \cdots \partial x_m^{p_m}} \) exist and are continuous on \( I \times \mathbb{E}_m \), is denoted by \( C^k \). Let

\[
f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}
\]

where \( f_i \) is as in (0.1). It will be assumed in the following that \( f \in L \). Now for the initial value problem (0.1) which can be written as

\[
X''(t) = f(t, X(t)) \quad A, B, X(t) \in \mathbb{E}_m,
\]

\[
X(a) = AX'(a) = B
\]

it is known [1] that there exists a unique solution.

The multistep difference equation problem for special second-order initial value problems (or more briefly, the multistep problem) is as follows: let \( f \) be in class \( L \); given is a set of \((2k+2)m\) real constants \( \alpha_j^1 \), \( \beta_j^1 \), \( j = 0, \ldots, k \), \( i = 1, \ldots, m \) and \( k \) elements of \( \mathbb{E}_m \), say \( S_0, \ldots, S_{k-1} \). Let \( A_t \) and \( B_t \) be diagonal matrices with diagonal elements \( \alpha_t^i \) and \( \beta_t^i \), respectively. Also

\[
\prod_{j=1}^{m} \alpha_k^i \neq 0 \text{ and } |\alpha_0^i| + |\beta_0^i| > 0 \text{ for at least one } i.
\]

Let \( h \) be a given positive constant. Define the sequence \( \{t_n\} \) by

\[
t_n = a + nh \text{ and the set of natural number } J \text{ to be those for which } t_n \in I.
\]
PROBLEM: To find a sequence \( \{X_n\} \) such that

(a) \( X_1 = S_1 \) \( i = 0, \ldots, k-1 \)

(b) for the function \( F_n \) defined from \( J \) and the \((k+1)\)-fold Cartesian product of \( E \) with itself to \( E \) whose value at \( t_n, X_n, \ldots, X_{n+k} \) is given by

\[
F_m(t_n, X_n, \ldots, X_{n+1}) = \sum_{j=0}^{k} A_j X_{n+j} - h^2 \sum_{j=0}^{k} B_j f(t_n+jX_{n+j}), \quad (0.2)
\]

the sequence \( \{X_n\} \) satisfies

\[
F_n(t_n, X_n, \ldots, X_{n+k}) = 0 \quad (n \in J).
\]

We shall speak of \( A_i, B_i \) \( (i = 0, \ldots, k) \) and the resulting function \( F_n \) as a special second-order method \( M \) or, more briefly, as method \( M \).

It is clear that if there is a solution of the multistep difference equation problem, say \( \{X_n\} \), then \( X_n \) depends on \( h \). Such a sequence will be said to approximate the solution \( X(t) \) of the special second-order initial value problem provided that as \( h \to 0 \) and \( n \to \infty \) simultaneously in such a manner that \( t_n = a + nh = t \) remains fixed in \( I \), that

\[
\lim_{n \to \infty} X_n = X(t).
\]

\[
\lim_{h \to 0} X_n = X(t).
\]
The following well-known theorem guarantees that (0.2) does indeed generate a sequence \( \{X_n\} \).

**THEOREM**: If \( f \) is in class \( L \) and \( 0 \leq h \leq h_0 \) where \( h_0 \) is such that \( |A_k^{-1}B_k|h_0^2L < 1 \), then (0.2) has a unique solution which takes on the initial values \( X_0, \ldots, X_{k-1} \).

The proof follows almost exactly that given by [4].

Under the assumption that \( \{X_k\} \ k = 0, \ldots, n+k-1 \) has been found, then Eq. (0.2) at \( n+k \) can be written as

\[
X_{n+k} = h^2A_k^{-1}B_kf(t_{n+k}, X_{n+k}) + C
\]

where \( C \) depends only on \( X_0, \ldots, X_{n+k-1} \).

The theorem follows immediately from [2], provided the function \( h^2A_k^{-1}B_kf(t_{n+k}, X) \) has a Lipschitz constant which is smaller than 1. But by hypothesis

\[
h^2|A_k^{-1}B_k||f(t_{n+k}, X) - f(t_{n+k}, Y)| \leq h^2|A_k^{-1}B_k|L < 1.
\]

**SECTION I: STABILITY, CONSISTENCY, AND CONVERGENCE**

We begin by proving three lemmas, two of which give estimates for the growth of the solution of (0.2).

**LEMMA IA**: If \( R(\xi) = A_k\xi^k + \ldots + A_0 \) where \( A_1 (i = 0, \ldots, k) \) are diagonal matrices such that \( \det A_k \neq 0 \) and \( A_k + \ldots + A_0 = 0 \), then the \( \{\Gamma_n\} \) which are defined by \( (\xi = 1/z) \)
are such that the sum

\[ \Sigma_n = \left( \Gamma_{n-k-1} A_{k-1} + \Gamma_{n-k-2} A_{k-2} + \ldots + \Gamma_{n-2k+1} A_0 \right) \]

\[ + \left( \Gamma_{n-k-2} A_{k-2} + \ldots + \Gamma_{n-2k} A_0 \right) \]

\[ + \ldots + \left( \Gamma_{n-k} A_1 + \Gamma_{n-k-1} A_0 \right) + \left( \Gamma_{n-k} A_0 \right) \]  

(IA.1)

has bounded norm.

PROOF: One can rearrange (IA.1) according to columns and use \( A_0 + \ldots + A_k = 0 \). This gives

\[ \Sigma_n = \Gamma_{n-k} (A_k) + \Gamma_{n-k-1} (A_k + A_{k-1}) + \ldots + \Gamma_{n-2k+1} (A_k + \ldots + A_1). \]  

(IA.2)

Upon further rearrangement

\[ \Sigma_n = \left( \Gamma_{n-k} A_k + \Gamma_{n-k-1} A_{k-1} + \ldots + \Gamma_{n-3k+1} A_1 \right) \]

\[ + \left( \Gamma_{n-k-1} A_k + \ldots + \Gamma_{n-2k} A_2 \right) + \ldots + \left( \Gamma_{n-2k+1} A_k \right). \]  

(IA.3)

From the power series definition of \( \Gamma_n \), one has

\[ \Gamma_n A_k + \ldots + \Gamma_{m-k} A_0 = 0 \quad m \geq k. \]
With this equation, (IA.3) can be reduced to

\[ \Sigma_n = (T_{n-2k}A_0) + (T_{n-2k-1}A_0 + T_{n-2k}A_1) + \cdots + (T_{n-2k}A_{k-3} + \cdots + T_{n-3k+1}A_0). \]

Thus we have arrived at (IA.1) with \(-k\) replaced by \(-2k\). This process can be repeated \(t\) times where \(t\) is such that \(m - tk \leq k\). Thus the sum is independent of \(n\) and hence bounded.

Now consider the difference equation

\[ \sum_{j=0}^{k} A_j z_{n+j} - h^p \sum_{j=0}^{k} E_{n+j}^j(z_{n+j}) = \Lambda_n \quad n = 0, 1, \ldots \quad (IA.4) \]

where \(\{E_{n+j}^j\}\) are functions from \(E_m\) to \(E_m\).

We assume there is a constant \(B\) such that

\[ |E_{n+j}^j(x)| \leq B|x| \quad j = 0, \ldots, k; \quad n = 0, 1, \ldots \]

and furthermore \(\det A_k \neq 0\).

Notice that (0.2) can be brought to the form of (IA.4) by adding and subtracting \(\sum_{j=0}^{k} B_j f(t_{n+j}, 0)\) and setting \(p = 2\).

Our next lemma gives a growth estimate of the solution of (IA.4) under certain hypotheses on the starting values \(z_0, \ldots, z_{k-1}\), viz., that
This particular requirement is in anticipation of the proof of theorem IF in which the equivalence of the convergence of a method to the method's stability and consistency is proved. That \( |z_0 - z_v| \leq H_1 \) is nothing more than the requirement of Theorem IF that the starting values "approximate" the given first derivative.

The second application of our next lemma will be in Section IV in which a requirement different from that of Theorem IF is placed on \( A_0, \ldots, A_k \). For this reason we content ourselves with saying that \( \det A_k \neq 0 \), which is sufficient to guarantee the existence of a solution of (IA.4).

We now formally state

**LEMMA IB**: If \( \{Z_n\} \) is the solution of (IA.5) with starting values as described and furthermore if \( R(l) = 0 \), then

\[
|z_n| \leq \left( E_n + D_n k (H_1 + H_2) \right) e^{\lambda_n} \text{ as long as } 1 - h^p \gamma_n B > 0
\]

where

\[
\lambda_n = \sum_{j=0}^{n-k} |A_j|; \quad \gamma_n = \max_{0 \leq j \leq n-k} |\Gamma_j|; \quad A = \max_{0 \leq j \leq k-1} |A_k|
\]

\[
\gamma_n^* = \gamma_n (1 - h^p \gamma_n B)^{-1}
\]
We premultiply the equation (IA.4) corresponding to \( j \) by \( \Gamma_{n-j} \) and sum from \( j=k \) to \( j=n \). Recalling the definition of \( \Gamma_{n-j} \), we reduce the left side of the sum to

\[
z_n + \sum_{j=0}^{k-1} \sum_{i=0}^{k-l-j} \Gamma_{n-k-i} A_{k-1-i} z_{k-l-j}
\] (IB.1)

To this expression we add and subtract the sum of the coefficients of \( z_0, \ldots, z_{k-1} \), times \( z_0 \), viz.,

\[
\sum_{j=0}^{k-1} \sum_{i=0}^{j} \Gamma_{n-k-j} A_{1-j} z_0.
\]

Thus, (IB.1) can be written as

\[
z_n + (\Gamma_{n-k-1} A_{k-1} + \ldots + \Gamma_{n-2k+1} A_0)(z_{k-1} - z_0) + \ldots
\]

\[
+ (\Gamma_{n-k} A_{k} + \Gamma_{n-k-1} A_0)(z_1 - z_0) + \sum_{j=0}^{k-1} \sum_{i=0}^{j} \Gamma_{n-k-j} A_{1-j} z_0.
\]

Taking norms and using Lemma IA in which it was proved that \( \sum_{i=0}^{j} \sum_{j=0}^{k-1} \Gamma_{n-k-j} A_{1-j} \) is bounded, one obtains:
\[ |z_n| \leq 2A\gamma_n kH_1 + MH_2 + h^{n-B} \sum_{j=1}^{n-k} |r_j| |z_{n-k-j+1}| + \sum_{j=0}^{n-k} |r_j| |A_{n-k-j}|. \]

For \( h \) such that \( 1 - h^p \gamma_n B > 0 \), one has

\[ |z_n| \leq \frac{2A\gamma_n kH_1 + MH_2 + \gamma_n \lambda_n}{1-h^p \gamma_n B} + \frac{h^p \gamma_n E_k}{1-h^p \gamma_n B} \sum_{j=0}^{n-1} |z_j| \]

or

\[ |z_n| \leq E_n + D_n \sum_{j=0}^{n-1} |z_j| \quad n \geq k \]

where \( E_n \) and \( D_n \) are as previously defined.

For \( n = k \),

\[ |z_k| \leq E_k + D_k \sum_{j=0}^{k-1} |z_j| \leq E_k + D_k k(H_1 + H_2) \]

\[ \leq [E_k + D_k k(H_1 + H_2)][1 + D_k]^k. \]

Suppose the induction assumption,

\[ |z_m| \leq [E_m + D_m k(H_1 + H_2)][1 + D_m]^m \quad m = k, \ldots, n-1 \]

holds. But
Furthermore, since $E_n$ and $D_n$ are monotone increasing,

$$|z_n| \leq E_n + D_n \sum_{j=0}^{n-1} |w_j| \leq E_n + D_n k(H_1 + H_2) + D_n \sum_{j=k}^{n-1} |z_j|$$

$$\leq E_n + D_n k(H_1 + H_2) + D_n \{E_k + D_k k(H_1 + H_2)(1+D_k)^k + \ldots + D_{n-1} k(H_1 + H_2)\}[1+D_{n-1}]^{n-1}$$

The following lemma is a slight modification of the previous one, and its proof will not be given in full. Unless otherwise stated, the notation is that of Lemma 1B.

If $r_i$ is such that the coefficients of the polynomial $p_i(z) = \sum_{t=0}^{k} \alpha_t^i z^t$ satisfy $\alpha_0^i = \ldots = \alpha_{r_1-1}^i = 0$ and $\alpha_{r_1}^i \neq 0$, then $r_i$ will be called the critical index of $p_i$.

In this lemma it is desired to obtain an estimate for the growth of the solution of (IA.4) which will show that this growth is independent (in the sense of the last hypothesis) of the $i^{th}$ component of the starting vectors $S_v$, provided $v < r_i$. 
LEMMA IC: IF, as in Lemma IB, \( \det A_k \neq 0 \) and \( |B_1^{n+j}(x)| \leq B|x| \)
and if furthermore \( |(z_v)_1| \leq H \) \( v = r_1, \ldots, k-1 \) and
\[
\sum_{v=0}^{k-1} |z_v| \leq s, \quad \text{then } |z_n| \leq E_n^* e^{nD_n} \quad n \geq k \text{ as long as } 1 - h^p \gamma_n B > 0
\]
where
\[
E_n^* = [x^2HA\gamma_n + \gamma_n \lambda_n + s D_n][1-h^p \gamma_n B]^{-1}
\]
\[D_n = \gamma_n h^p [1-h^p \gamma_n B].\]

Proof: Observe that in (IB.1) \( z_t \), \( 0 \leq t \leq k-1 \) has
among the \( A_j \) \( j = 0, \ldots, k \) only \( A_0, \ldots, A_t \) as possible coefficients.
Since the \( A_j \)'s are diagonal, the vector \( A_j z_j \) \( j = 0, \ldots, t \) is
independent of \( (W_j)_j \) for \( 0 \leq j \leq r_1 \) where \( r_1 \) is the critical
index of \( \rho_1 \). Thus, \( |A_j W_j| \leq AH, A = \max_{0 \leq j < k} |A_j| \). One obtains
\[
\left| \sum_{j=0}^{k-1} \sum_{i=0}^{k-1-j} \Gamma_{n-k-i} z_{i-j} \right| < \gamma_n k^2 AH + \ldots + \gamma_n k^2 AH \leq \gamma_n k^2 AH.
\]

Now, as in Lemma IB and by the previous discussion, one has for \( n \geq k \)
\[
|z_n| \leq \gamma_n k^2 AH + h^p \gamma_n k B \sum_{j=0}^{n-1} |z_j| + \gamma_n \lambda_n.
\]
We introduce the briefer notation, $D_n$ and $E^{**}$ so

$$|z_n| \leq E_n + h^PD_n \sum_{j=0}^{n-1} |z_j|.$$  

The proof from this point is completely like the previous one, and the details are omitted.

The vector values functions $S_0(h),...,S_{k-1}(h)$ defined on $(0,h_0)$ are said to have property $S$, provided that there is a constant $M$ such that for $h \in (0,h_0)$

$$|S_0(h)| < M$$

$$|S_0(h) - S_v(h)| < Mh \quad v = 0,...,k-1.$$  

Following [7] a method $M$ is said to be stable, if and only if, for every function $f(t,X)$ which is in class $L$ and for initial values $S_0(h)...S_{k-1}(h)$ with property $S$, the solution $\{z_n\}$ of (0.2) with starting values $S_v(h)$ is bounded for all $h \in (0,h_0)$. 

A polynomial $p(z) = \alpha_k z^k + ... + \alpha_0$ with $\alpha_k \neq 0$ is said to satisfy the root condition [5] provided that its roots $z_1$ have the properties

1. $|z_1| \leq 1$

2. $|z_1| = 1$ implies $z_1$ is a root of multiplicity at most 2.
THEOREM ID: A method $M$ is stable if and only if the polynomials $S_1(z) = a_1^1 z^k + \ldots + a_0^1$, where $a_j^1$ is the $j$th main diagonal element of $A_1$, satisfy the root condition.

Proof: The necessity follows from the examples given by [3]. The proof of the sufficiency is obtained via Lemma IB. This requires the computation of $\{\Gamma_n\}$. From Lemma IA

$$\sum_{t=0}^{\infty} \Gamma_t z^t = [A_0 z^k + \ldots + A_k]^{-1} = \text{diag} \ z^{-k} \rho_1(z)$$

where $\text{diag} a_1$ is used to denote a diagonal matrix with main diagonal given by $a_1$. From Lemma 6.2 of [3], one has that for each diagonal element there exist constants $\overline{\Gamma}_1$ and $\overline{\gamma}_1$ such that $|\Gamma_n|_1 \leq \overline{\Gamma}_1 n + \overline{\gamma}_1$. The proof of that lemma follows in broad outline upon an expansion of $z^{-k} \rho_1(z)$ in partial fractions and then each term thereof in Taylor's series. The factor $n$ arises from the roots of multiplicity 2. Define

$$\alpha = \max_{0 \leq k \leq m} \overline{\Gamma}_1, \quad \beta = \max_{0 < i \leq m} \overline{\gamma}_1.$$

Thus $\gamma_n = \alpha n + \beta$. 
Let \( \{X_n\} \) denote a family of solutions of (0.2) with starting values having property S. One can write (0.2) in the form

\[
A_k x_{n+k} + \ldots + A_0 x_n = h^2 \{ B_k [f(t_{n+k}, x_{n+k}) - f(t_{n+k}, 0)] + \ldots + B_0 [f(t_n, x_n) - f(t_n, 0)] \} + \Lambda_n
\]

where

\[
\Lambda_n = h^2 \{ B_k f(t_{n+k}, 0) + \ldots + B_0 f(t_n, 0) \}.
\]

Since \( f \) is continuous, the definition of \( \Lambda \) yields \( \lambda_n \leq hC \) where \( C \) is a constant. As \( f \) is in class L, one can take \( B = bL \) where \( b = \max_{0 \leq j \leq k} |B_j| \).

Let \( h \) be chosen so that \( S = 1 - h^2 B \left( \frac{a(b-a)}{h} + \beta \right) > 0 \). But now Lemma IB which guarantees the boundedness of \( x_n \), provided the "remainder" terms \( \Lambda_n \) are \( O(h^2) \) and the starting values are \( O(h) \), can be applied to finish the proof. More precisely, Lemma IB yields

\[
|X_n| \leq \left[ \frac{2Ak(a+b)h}{6} + \frac{Mh}{6} + \frac{a+b}{6} hC \right] + \left[ \frac{h^2 \beta}{6} (a+b)k(h+M) \right] \exp \left[ \frac{nh^2 (a+b)}{6} \right].
\]
Since $h = \frac{t_n - a}{n}$, and $t_n \leq b$,

$$|X_n| \leq C,$$

where $C$ is a constant.

A method $M$ is said to be consistent if and only if, for every function $f$ which is in class $L$ and for every solution $X(t)$ of (0.1), there holds

$$\sup_{n \in J} |F_n X(t_n), \ldots, X(t_{n+k})| = O(h^2),$$

as $h \to 0$

**THEOREM IE:** A method $M$ is consistent if and only if $\rho_1$ and $\sigma_1(z) = \sum_{j=0}^{k} \beta_j^1 z^j$ where $\beta_j^1$ is the $i^{th}$ diagonal element of $B_1$

satisfy $\rho_1(1) = \rho_1'(1) = 0$, $\rho_1''(1) = 2\sigma_1(1)$; $i = 1, \ldots, m$.

**Proof:** Suppose $\rho_1(1) = \rho_1'(1) = 0$ and $\rho_1''(1) = 2\sigma_1(1)$.

Thus, $\alpha_k^r + \ldots + \alpha_o^r = 0$, $k\alpha_k^1 + \ldots + \alpha_1^r = 0$, and $k^2 \alpha_k^1 + \ldots + \alpha_1^r = 2(\beta_k^r + \ldots + \beta_o^r)$; $i = 1, \ldots, m$. Let $f$ be in class $L$ and $X$ be a solution of (0.1). $X''(t)$ is continuous on its interval of definition. Define

$$\omega(S) = \max_{|s-t| \leq \delta} |X''(x) - X''(t)|$$

for $s, t \in [a, b]$.  


By the mean value theorem applied componentwise (in this theorem we use \( X_i \) to denote the \( i \)th component of \( X \)) one has

\[
X_i(t_n + p) = \frac{X_i(t_n) + pX'_i(t_n) + \frac{p^2}{2} X''_i(t_n)}{X_i(t_n) + pX'_i(t_n) + \frac{p^2}{2} \left( X''_i(t_n) + \theta_{n1} \omega(p) \right)}
\]

where

\[
|\theta_{n1}| < 1.
\]

Now

\[
X(t_n + p) = X(t_n) + pX'(t_n) + \frac{p^2}{2} X''(t_n) + \frac{p^2}{2} \theta_n \omega(p)
\]

where

\[
\theta_n = \text{diag } \theta_{n1}.
\]

Thus,

\[
\left( F_n(x(t_n), \ldots, x(t_{n+k})) \right)_i
\]

\[
= \left( \alpha_k^1 + \ldots + \alpha_0^1 \right) x_i(t_n) + \left( \alpha_k^1 + \ldots + \alpha_1^1 \right) x'_i(t_n)
\]

\[
+ h^2 \left( \alpha_k^1 \frac{k^2}{2} \ldots + \alpha_1^1 \right) x''_i(t_n) + h^2 \left( \alpha_k^1 \frac{k^2}{2} \theta_n \omega(kh) + \ldots + \frac{\alpha_1 \theta_{n1}}{2} \omega(h) \right)
\]

\[
- h^2 \left( \beta_k^1 + \ldots + \beta_0^1 \right) x''_i(t_n) - h^2 \left( \beta_k^1 \theta_n \omega(kh) + \ldots + \beta_1 \theta_{n1} \omega(h) \right)
\]
which reduces to

\[ = h^2 \left( \frac{k^2}{2} \theta_k \omega(kh) + \ldots + \frac{\alpha_1 \theta_1}{2} \omega(h) + \beta_k \theta_k \omega(kh) + \ldots + \theta_1 \beta_1 \omega(h) \right). \]

Since \( \omega(h) \) is monotone increasing,

\[ |(F_nX(t_n), \ldots, X(t_{n+k}))| \leq h^2 k^2 \alpha^* \omega(kh). \]

\[ \alpha^* = \max_{0 \leq j \leq k} |\alpha_j^2| . \]

Thus, \(|F_n(X(t_n), \ldots, X(t_{n+k}))| \leq h^2 k^3 \alpha^* \omega(kh), \) and the consistency follows.

The converse follows by the examples given on p. 301 of [3].

**THEOREM IF:** A method \( M \) is convergent, if and only if \( M \) is stable and consistent.

**Proof:** Suppose \( M \) is a stable, consistent method.

Let \( \{X_n\} \) and \( X(t) \) be solutions of (0.2) and (0.1) respectively. Let \( S_i(h) = 0, \ldots, k-1 \) be starting values with property \( S \) and

\[ \lim_{h \to 0} S_i(h) = A. \]

As at the end of the previous theorem we write

\[ F_n(X_n, \ldots, X_{n+k}) = \Lambda_n \]
where $|\Lambda_n| \leq 6(h)h^2$ with $6(h) \to 0$ as $h \to 0$. Thus

$$\lambda_n = \sum_{j=0}^{n-k} |\Lambda_j| \leq h6(h).$$

Setting $E_n = X_n - X(t_n)$, one has

$$F_n(x_n, \ldots, x_{n+k}) - F_n(x(t_n), \ldots, x(t_{n+k}))$$

$$= -\Lambda_n$$

$$= \sum_{j=0}^{n} A_j E_{n+j} - h^2 \sum_{j=0}^{k} (B_j f(t_{n+j} X_{n+j}) + E_{n+j}) + h^2 \sum_{j=0}^{k} B_j f(t_{n+j} X_{n+j}).$$

Regard this as a difference equation for $E_n$. It is desired to apply Lemma IB. Now

$$B_m^\mu(E_m) = B_\mu \{f(t_m, X(t_m) + E_m) - f(t_m, X(t_m))\}.$$ 

Thus, $|B_m^\mu(E_m)| \leq L |B_m| |E_m|$ and we can take $B = L \max_{0 \leq j \leq k} |B_j|$. Also

$$|E_\mu| \leq |S_\mu - A| + |X(t_\mu) - A| \quad \mu = 0, \ldots, k-1$$

$$\leq H$$

where $H$ is small for $h$ small.
Also

\[ |E_v - E_0| \leq |S_v - X(t_v) - X_0 + X(a)| \]
\[ \leq |S_v - S_0 - B_v h + o(h^2)| \]

where \( H \) is small when \( h \) is small. Since \( M \) is stable,

\[ \gamma^v \leq \frac{\alpha n + 2}{5} \text{ for } h_0 \text{ chosen that} \]

\[ h = 1 - h_0^2 \left( \frac{a(b-a)}{h_0} + 2 \right) > 0. \]

Thus, an application of Lemma 1B yields

\[ |E_n| \leq \left[ 2A \frac{\alpha n + 2}{5} hH + \frac{MH}{2} + \frac{\alpha n + 2}{5} h5(h) + \frac{h^2 B(\alpha n + 2)}{5} \right] C \]
\[ \leq C(H + 5(h)), \]

and since \( H \to 0 \) as \( h \to 0 \), \(|E_m|\) is uniformly bounded and this part of the proof is finished.

The converse follows readily from examples given by [3].

SECTION II: ASYMPTOTIC REPRESENTATION

In the following section, in an attempt to obtain more detailed information about the behavior of the numerical approximation, we increase our assumptions on \( f \) and the starting values. It will hereafter be assumed that the method \( M \) is both stable and consistent.
It is now of some importance to show explicitly the dependence of $F_n$ on $h$, and we denote it at $(t_n, X_k, \ldots, X_{n+k}, h)$ for $0 < h < h_0$ by $F_n(X_n, \ldots, X_{n+k}, h)$.

We define the order of a method to be the largest integer $p$ such that

$$\max_{a \leq t_n \leq b} |F_n(X(t_n), \ldots, X(t_{n+k}), h)| = O(h^{p+2})$$

for all solutions $X$ of all initial value problems (0.1) where $f$ is continuously differentiable of sufficiently high order.

In the one-dimensional case our definition of order is equivalent to that given in [3] where it is proved that the term $O(h^{p+2})$ can be replaced by $h^{p+2}C_{p+2}X(t)^{(p+2)} + O(h^{p+3})$.

We also recall that $F_n(X(t_n), \ldots, X(t_{n+k}), h) = \sum_{j=1}^{k} A_j X(t_{n+j}) - h^2 \sum_{j=1}^{k} B_j X''(t_{n+j})$. But in this form the $i$th component of $F_n$ depends only on the $i$th components of $X$ and $X''$; thus, the order of a method in our sense is the smallest order of the methods defined by $\rho_1, \sigma_1$ in the sense of [3]. Thus, there exists a diagonal matrix $C_{p+2}$ with at least one nonzero element such that

$$F_n(X(t_n), \ldots, X(t_{n+k}), h) = C_{p+2}X(t_n)^{(p+2)}h^{p+2} + O(h^{p+3}).$$
In particular

\[
(C_{p+2})_{11} = \begin{cases} 
0 & \text{if the order of the method defined by } p_1 \delta_1 \text{ is greater than } p. \\
\sigma_1(1)c_1 & \text{where } c_1 \text{ is the error constant (in the sense of [3]) obtained for the method defined by } p_1 \delta_1.
\end{cases}
\]

The next lemma will be used in the derivation of the asymptotic formula. Although this estimate could have practical application, numerical experiments have shown that it is quite gross. For this reason we will be imprecise with the constants and only try for an easily applicable form.

**LEMMA IIA:** If the method \( M \) is of order \( p \), and if the solution \( X \) of (0.1) is such that \( X \in C^{p+2}[a,b] \), and if \( X_n \) results from starting values \( S_i(h) \) with property \( S \) and

\[
h \delta(h) = \max_{0 \leq i \leq k-1} |S_i(h) - X(t_i)|, \text{ then } |S_n - X(t_n)| \leq C(\delta(h) + h^p).
\]

**Proof:** The proof is that of Theorem IF with \( H \) and \( \omega(h) \) replaced by \( \delta(h) \) and \( CX^{(p+2)}h^p \) respectively.

**Lemma IIB,** which is a standard representation theorem for the theory of difference equations, will be given in some detail because of the importance it plays in the asymptotic formula (Theorem IID). It is here that the critical index first plays a role. Together with the explicit form of
the solution of the difference equation
\[ a_k y_{n+k} + \ldots + a_0 y_n = 0, \]
Lemma IIB shows that the solution does not depend on starting values whose index is smaller than the critical index.

More precisely,

**LEMMA IIB:** If \( p(z) = a_k z^k + \ldots + a_0 \) is a stable polynomial, then the difference equation
\[ a_k y_{n+k} + \ldots + a_0 y_n = 0 \]
with starting values \( s_k, \ldots, s_{k-1} \), where \( y_n, s_0, \ldots, s_{k-1} \) are scalars, is satisfied by

\[
y_n = \sum_{k=1}^{d} (a_n + nb_n)z_n^k + \sum_{n=2d+1}^{l} C_n z_n^n + \sum_{n=t-1}^{m} P_n(n)z_n^n
\]

where \( p(z) = 0 \) and in particular \( z_1, \ldots, z_d \) are the distinct roots of modulus 1 and multiplicity 2 (called "essential roots" in [3]).

\[
b_n = \frac{z}{z_n n''(z_n)} \sum_{t=0}^{k-1} \alpha_n t^s t
\]

\( \alpha_n t \) is defined by
\[
\frac{p(z)}{z-z_n} = a_{kn} + \ldots + a_{nk-1} z^{k-1} \]
and \( a_n^r C_n^r P_n^r(k) \)
are defined in the proof. Furthermore, \( y_n \) takes on the starting values \( s_r, \ldots, s_{k-1} \) where \( r \) is the critical index of \( p \).

**Proof:** If one considers the equations

\[
y_n = s_r, \ldots, y_{k-1} = s_{k-1}, \quad (IIB.1)
\]
then there arises a system of $k-1-r$ equations in the $k-1-r$ unknowns $a_n, b_n, c_n, d_n$. This linear system has a unique solution; indeed, the coefficient matrix has a nonvanishing determinant [3]. Thus, it is possible to determine the unknowns so that the theorem is satisfied.

It is of some interest to know an explicit formula for the computation of $b_n$, and one is obtained in the following.

Define $\alpha_{nj} \; j = 0, \ldots, k-1$ by

$$f(z) = \frac{\rho(z)}{z-z_n} = \alpha_{n0} + \ldots + \alpha_{nk-1} z^{k-1} \quad 1 \leq n \leq d.$$ 

Notice at $z_n$

$$f(z_n) = \rho''(z_n) = 2\left\{\alpha_{n1} + \ldots + (k-1)\alpha_{nk-1} z_n^{k-1}\right\}.$$ 

Also

$$f'(z_n) = \rho'(z_n) = p! \alpha_{np} + \ldots + \alpha_{nk-1}(k-1)\ldots(k-p+1)z^{k-p-1}.$$
Multiply the equation \( y_j = s_j \) by \( \alpha_{nj} \) and sum. This gives

\[
\sum_{\lambda=1}^{d} a_{\lambda} \sum_{j=0}^{k-1} \alpha_{nj} z_{nj}^j + \sum_{n=1}^{d} b_{n} \sum_{j=1}^{k-1} \alpha_{nj} z_{nj}^j + \sum_{n=2d+1}^{l} c_{n} \sum_{j=0}^{k-1} \alpha_{nj} z_{nj}^j + \sum_{n=t+1}^{m} d_{no} \sum_{j=0}^{k-1} \alpha_{nj} z_{nj}^j + \ldots + \sum_{n=t+j}^{m} d_{np} \sum_{j=l}^{(j-l)\ldots(j-p)} \alpha_{nj} z_{nj}^j = \sum_{j=0}^{k-1} \alpha_{nj} s_j^n
\]

\((p = \text{largest multiplicity})\) which reduces to

\[
\frac{b_n z_n \rho''(z_n)}{z} = \sum_{j=0}^{k-1} \alpha_{nj} s_j^n
\]

whence

\[
b_n = \frac{2}{z_n \rho''(z_n)} \sum_{j=0}^{k-1} \alpha_{1j} s_j^n.
\]

The \( a_n \) can be determined by using

\[
\frac{\rho(z)}{(z-z_n)^2} = \beta_{no} + \ldots + \beta_{nk-2} z^{k-2}.
\]

The \( b_n \)'s are subtracted from both sides. Again the \( d \) terms produce no contribution.
To determine $P_n(n)$, one can use

$$\frac{\rho(z)}{(z-z_n)^2}, \ldots, \frac{\rho(z)}{(z-z_n)^p}$$

where $p$ is the multiplicity of $z_n$. The precise form of these coefficients is not needed and will not be computed.

One can describe the vector solution $\{y_n\}$ of (IIB.2)

$$A_k y_{n+k} + \ldots + A_0 y_n = 0$$

by an immediate corollary to the previous lemma. Since the $A_j$'s are diagonal, the system is separable; and we can apply the previous lemma $m$ times to

$$\sum_{j=0}^k a_j^i y_{n+j} = 0$$

Thus, we can write

$$(y_n)_1 = \sum_{j=1}^{d_1} \left( a_{ij}^* + nb_{ij}^* \right) z_{ij}^n + \sum_{j=2d+j}^{l_1} c_{ij}^* z_j^n + \sum_{j=l+j}^{n_1} P_j^*(n) z_j^n$$

where $a_{ij}^*, b_{ij}^*, c_{ij}^*, P_j^*(n)$ are respectively $a_n, b_n, c_n, P_n$ of the foregoing lemma with $\rho$ replaced by $\rho_1(z) = \sum_{j=0}^k a_j z^j$. However, this notation is not convenient for the next theorem.
Let the set of roots of the polynomials $p_1(z)$ $i = 1, \ldots, m$ be renumbered so that $z_1(= 1), z_2, \ldots, z_d$ are essential roots of some $p_1$; $z_{2d+1}, \ldots, z_t$ are modulus 1, multiplicity 1 roots of some $p_1$; $z_{t+1}, \ldots, z_m$ are roots of modulus less than 1 of all $p_1$. These three classes are not disjoint. Indeed $z_\lambda$ could be an essential root of, say, $p_1$ but a simple root of $p_2$.

We define

$$b_{ij} = \begin{cases} \ast_{ij} & \text{if } z_{ij} \text{ is an essential root of } p_1 \\ 0 & \text{otherwise} \end{cases}$$

$$a_{ij} = \begin{cases} a_{ij} & \text{as above} \\ 0 & \text{otherwise} \end{cases}$$

$$c_{ij} = \begin{cases} c_{ij} & \text{if } z_{ij} \text{ is a simple, modulus 1 root of } p_1 \\ 0 & \text{otherwise} \end{cases}$$

$$p_y(n) = \begin{cases} p_{ij}^*(n) & \text{if } z_{ij} \text{ is a root of modulus less than 1 of } p_1 \\ 0 & \text{otherwise} \end{cases}$$

Now we can write

$$\left(y_n\right)_1 = \sum_{j=1}^{d} (a_{ij} + nb_{ij})z_j^n + \sum_{j=2d+1}^{t} c_{ij}z_j^n + \sum_{j=t+1}^{m} p_{ij}(n)z_j^n$$

(IIB.3)
and state

**COROLLARY IIC:** Under the just described notation and the assumption that each $p_j(z)$ is a stable polynomial, there follows that the $i^{th}$ component of a solution of (IIB.1), which takes the initial values $(S_v)_i$, $v = r_1, \ldots, k-1$ where $r_1$ is the critical index of $p_1$, is given by (IIB.3).

In the next theorem an asymptotic formula for the numerical solution is given. Basically, the next theorem states that asymptotically the solution of the approximating difference equation consists of three parts which are: the solution of the given differential equation, the "truncation error," and the "starting error." If the method is of order $p$, then the "truncation error" contribution will be $O(h^p)$; if the starting values approximate the given initial conditions to $O(h^{q+1})$, then the "starting error" is $O(h^q)$. The later contribution is unfortunately notationally intricate; however, were the $A_i$'s scalar multiples of the identity matrix or were there no essential roots other than $z = 1$, then a much simpler theorem would result.

We introduce the following notation. Suppose $z_j$ is a member of the set of essential roots. Define


\[ s_{1j} = \begin{cases} 1 & \text{if } z_j \text{ is an essential root of } \rho_1 \\ 0 & \text{otherwise} \end{cases} \]

\[ s_{2j} = \begin{cases} 1 & \text{if } \rho_1'(z_j) \neq 0 \\ 0 & \text{otherwise} \end{cases} \]

\[ s_{3j} = \begin{cases} 1 & \text{if } \rho_1(z_j) = 0 \text{ and } \rho_1'(z_j) \neq 0. \\ 0 & \text{otherwise} \end{cases} \]

We also define

\[ X_{1j}(t) = \begin{cases} 0 & \text{if } s_{1j}' = 0 \\ \text{otherwise the solution of} \end{cases} \]

\[ X_{1j}'' = \frac{2\rho_1(z_j)}{z_j^2} \sum_{t=1}^{m} s_{1t}(t)s_{1j}'X_{tj} \]

\[ X_{1j}(a) = 0, \quad X_{1j}'(a) = b_{1j} \]

\[ Y_{1j} = \begin{cases} 0 & \text{if } s_{2j}' = 0 \\ \text{otherwise the solution of} \end{cases} \]

\[ Y_{1j}'' = \frac{\rho_1(z_j)}{\rho_1'(z_j)} \sum_{t=1}^{m} s_{1t}(t)s_{1j}'X_{tj} \]

\[ Y_{1j}(a) = Y_{1j}'(a) = 0 \]
\[ z_{1j} = \begin{cases} 0 & \text{if } \text{if } b_{3j}^3 = 0 \\ \text{otherwise the solution of} \\ Z_{1j}'' = \frac{\sigma_1(z_j)}{z_j \rho_1(z_j)} \sum_{t=1}^{m} g_{1t}(t) b_{tj} X_{1j} \end{cases} \]

\[ Z_{1j}(a) = Z_{1j}'(a) = 0. \]

The expression \( 2\sigma_1(z_j)/z_j^2 \rho_1''(z_j) \) will be called a growth parameter and will be denoted by \( \mu_1^2 \) or merely by \( \mu^2 \), if \( \sigma_1 = \sigma_t \) and \( \rho_1 = \rho_t \) for all \( 1 \leq i, t \leq m \).

Define \( U_n \) such that

\[ (U_n)_i = \sum_{j=1}^{d} 5_{1j}^3 z_j^3 x_{1j}(t_n), \]

\( V_n \) such that

\[ (V_n)_i = \sum_{j=1}^{d} 5_{1j}^3 z_j^3 x_{1j}(t_n), \]

\( W_n \) such that

\[ (W_n)_i = \sum_{j=1}^{d} 5_{1j}^2 z_j^2 z_{1j}(t_n). \]
THEOREM IID: If \( f \) is a twice-continuously differentiable function; \( A, B \) are given vectors; \( X \), the exact solution of (0.1), is of class \( C^{p+2}[a,b] \); the method \( M \) is of order \( p \); \( \{X_n\} \) arises from the starting values \( S_v(h) \) where \( S_v(h) - X(t_v) = S_v h^{q+1} + O(h^{q+2}) \) with the \( S_v \) as constant vectors; \( b_1, A, C, P(n) \) are as in Corollary IIC; \( G(t) \) is a matrix with elements

\[
\mathcal{G}_{ij} = \frac{\partial f_i(t, X(t))}{\partial x_j}, \quad e(t) \text{ is the solution of } e'' = Ge - CX^{(p+2)} \]

where \( C \) is the truncation matrix of \( M \), then

\[
X_n = X(t_n) + hP e(t_n) + h^q U_n + O(h^{r+1})
\]

where \( r = \min(p,q) \).

Under slightly more restrictive hypothesis a notationally simpler version is given after the proof.

Proof: It is desired to show that

\[
\bar{z}_n = X(t_n) - X_n + hP e(t_n) + h^q U_n \quad n = 0, \ldots, N
\]

where \( e, U \) are as given before the statement of Theorem IID, have two properties:

(1) \( \mathcal{L}(\bar{z}_n) = O(h^{r+3}) \)

where \( \mathcal{L} \) is the operator defined by

\[
\mathcal{L}(z_n) = \sum_{j=0}^{k} A_k \bar{z}_{n+j} - h^2 \sum_{j=0}^{k} B_j G(t_{n+j}) \bar{z}_{n+j}.
\]

(2) \( (\bar{z}_j)_1 = O(h^{r+2}) \quad r_1 \leq j < k \)

where \( r_1 \) is the critical index of \( \rho_1(z) \).
Once these are established, the theorem follows by an application of Lemma IC. Indeed, we recall that Lemma IC gives a bound for $|Z_n|$ in terms of the starting values and $\Lambda_n$. Here $\Lambda_n$ is the $O(h^{r+2})$ of property 1, and the starting values are the $O(h^{r-2})$ terms of property 2. Roughly speaking, the bound of Lemma IC reduces the $O(h^{r+3})$ to $O(h^{r+1})$ and the $O(h^{r+2})$ to $O(h^{r+1})$, but this will give the desired result that $|Z_n|$ is bounded by a term which is $O(h^{r+1})$.

To facilitate this task one can study

$$Z_n = Z_n + h^{q+1} \left\{ \sum_{x=1}^{d} D_x Z^n_x + \sum_{x=2d+1}^{t} C_x Z^n_x + \sum_{x=t+1}^{m} P_x(n) Z^n_x + W_n + hV_n \right\}$$

(IID.1)

where $(D_x)_1 = a_{1x}$, $(C_x)_1 = c_{1x}$, $(P_x(n))_1 = p_{1x}(n)$ with $a_{1x}$, $c_{1x}$, $p_{1x}(n)$ as in Corollary IIC. Since $|Z^n_x| < 1$, $t + 1 \leq x \leq m$, $|Z^n_x| = 1$, $2d \leq x \leq t$, the term in parentheses is uniformly (independent of $n, t_n$) bounded. Thus, if $Z_n = O(h^{r+1})$, then $Z_n = O(h^{r+1})$.

We first show that $f(Z_n) = O(h^{r+3})$. Since $f$ is linear, then $f$ can be applied separately to each term on the right of Eq. (IID.1)
\[
\mathcal{L}(X(t_n)) = \sum_{j=0}^{k} A_j X(t_{n+j}) - h^2 \sum_{j=0}^{k} B_j C(t_{n+j}) X_{n+j} - f(t_{n+j}, X(t_{n+j}))
\]

\[-h^2 \sum_{j=0}^{k} B_j C(t_{n+j}, X(t_{n+j}))\]

\[= h^{p+2} C_{p+2} X(t_n^p) + h^2 \sum_{j=0}^{k} B_j \left\{ f(t_{n+j}, X(t_{n+j})) - G(t_{n+j}) X(t_{n+j}) \right\} + o(h^{p+3})\]

\[
\mathcal{L} X_n = \sum_{j=0}^{k} A_j X_{n+j} - h^2 \sum_{j=0}^{k} B_j \left\{ G(t_{n+j}) X_{n+j} + f(t_{n+j}, X_{n+j}) - f(t_{n+j}, X_{n+j}) \right\}
\]

\[= h^2 \sum_{j=0}^{k} B_j \left\{ f(t_{n+j}, X_{n+j}) - G(t_{n+j}) X_{n+j} \right\}\]

\[
\mathcal{L} h^p e(t_n) = h^p \sum_{j=0}^{k} A_j e(t_{n+j}) - h^{p+2} \sum_{j=0}^{k} B_j \left\{ C(t_{n+j}) e(t_{n+j}) - C X^{p+2}(t_{n+j}) \right\}
\]

\[= o(h^{2p+2}) - h^{p+2} \sum_{j=0}^{k} B_j C X(t_n^{p+2})\]

\[= -h^{p+2}(B_k + \ldots + B_0) C X(t_n^{p+2}) + o(h^{p+3}).\]
Thus
\[ x_{hP_e}(t_n) = -h^{p+2}C_{p+2}x(t_{n}) + O(h^{p+3}). \]

Now
\[
\mathcal{L}(h^q u_n + h^{q+1}w_n + h^{q+2}v_n)
\]
\[ = h^q \sum_{j=0}^{k} A_j(u_{n+j} + h w_{n+j} + h^2 v_{n+j}) - h^{q+2} \sum_{j=0}^{k} B_j G(t_{n+j}) u_{n+j} + O(h^{q+3}). \]

Consider the \(i^\text{th}\) component of the previous equation.
\[
\left( \mathcal{L}(h^q u_n + h^{q+1}w_n + h^{q+2}v_n) \right)_i
\]
\[ = h^q \sum_{m=0}^{k} \sum_{j=1}^{d} \sum_{i=0}^{n+m} \sum_{l=1}^{m} \alpha_{i} \beta_{m} \gamma_{i,j}(t_{n+m}) \]
\[ + \delta_{i,j}^2 (t_{n+m}) + \delta_{i,j}^3 (t_{n+m}) \]
\[ - h^{q+2} \sum_{n=0}^{k} \sum_{j=1}^{d} \sum_{l=1}^{m} \varepsilon_{i} \zeta_{l,t} (t_{n+m}) \delta_{i,j} \gamma_{i,j}(t_{n+m}) + O(h^{q+3}). \]
We can apply Taylor's theorem which yields

\[ h^q \sum_{j} z_j^m \left( \rho(z_j) \left( \delta_{1j} x_{1j}(t_n) + h^2 \delta_{2j} x_{1j}(t_{n+m}) + h^3 \delta_{3j} x_{1j}(t_{n+m}) \right) \right) \]

\[ + h^2 z_j^2 \rho''(z_j) \left( \delta_{1j} x_{1j}'(t_n) + h^3 \delta_{3j} x_{1j}'(t_n) \right) \]

\[ + h^2 z_j x_{1j}''(z_j) \left( \delta_{1j} x_{1j}'(t_n) \right) \]

\[ - h^2 \sigma_j(z_j) \sum_{t=1}^{m} s_{1t}(t_n) \delta_{1j} x_{1j}(t_n) + O(h^{q+3}). \]

Using the various definitions given before the statement of the theorem, one can show that this expression is \( O(h^{q+3}) \). Indeed, if \( z_j \) is an essential root, then \( \delta_{1j} = 1 \), but also \( x_{1j} \) can be replaced by the right side of its defining equation (given just before the statement of Theorem IID) in which case the expression with coefficient \( \sigma_j(z_j) \) is cancelled; thus, \( O(h^{q+3}) \). If \( z_j \) is not an essential root, then either \( \rho(z_j) \neq 0 \) or \( \rho(z_j) = 0 \) and \( \rho(z_j) \neq 0 \). In the first instance \( \delta_{1j}^2 \neq 0 \) and \( \delta_{1j}^3 = \delta_{1j}^3 = 0 \). Here we may replace \( x_{1j}' \) by its defining equation and again obtain \( O(h^{q+3}) \). Finally, in the second instance \( \delta_{1j}^3 \neq 0 \) and \( \delta_{1j}^3 = \delta_{1j}^2 = 0 \). Now \( Z_{1j} \) may be replaced by its defining equation and we have \( O(h^{q+3}) \). Thus, in each instance the expression is \( O(h^{q+3}) \).
Finally,

\[
\mathcal{L} \left[ h^{q+1} \sum_{x=1}^{d} A_x z_x^n + \sum_{x=2d+1}^{\ell} C_x z_x^n + \sum_{x=\ell+1}^{m} P_k(n) z_k^n \right]
\]

\[
= h^{q+1} \sum_{j=0}^{k} \left( \sum_{x=1}^{d} A_x z_x^{n+j} + \sum_{x=2d+1}^{\ell} C_x z_x^{n+j} + \sum_{x=\ell+1}^{m} P_k(n) z_k^{n+j} \right) + O(h^{p+3})
\]

\[
= O(h^{q+3}).
\]

Collecting these results we have

\[
\mathcal{L}(z_n) = h^{p+2} c_{p+2} x \left( \theta \left( t_n \right) \right) + h^2 \sum_{j=0}^{k} B_j f(t_{n+j}, x(t_{n+j}))
\]

\[
- h^2 \sum_{j=0}^{k} B_j g(t_{n+j}) x(t_{n+j}) + h^2 \sum_{j=0}^{k} B_j f(t_{n+j}, x_{n+j})
\]

\[
- h^2 \sum_{j=0}^{k} B_j g(t_{n+j}) x_{n+j} - h^{p+2} c_{p+2} x(t_n) + O(h^{p+3}).
\]

But since \(|x(t_n) - x_n| = O(h^n)| by Lemma IIA, we have

\[
\mathcal{L}(z_n) = O(h^{n+3}).
\]
We now study \( z_v, \quad v = 0, \ldots, k-1 \)

\[
z_v = - h^{q+1} s_v + o(h^{q+2}) + h^q u_n + o(h^{q+2})
+ h^{q+2} \sum_{k=1}^{d} A_k z_k^v + \sum_{k=2d+1}^{l} C_k z_k^v
+ \sum_{k=t+1}^{m} p_k(v) z_k^v + o(h^{p+2}).
\]

The \( i \) th component of the previous expression is

\[
(z_v)_i = - h^{p+1} (s_v)_i + h^p \sum_{j=1}^{d} b_{ij} v z_j^v x_{1j} (t_v)
+ h^{p+1} \sum_{k=1}^{d} (A_k)_i z_k^v + \sum_{k=2d+1}^{l} (C_k)_i z_k^v
+ \sum_{k=t+1}^{m} (p_k(v))_i z_k^v + o(h^{p+2})
= h^{p+1} \left[ - (s_v)_i + \sum_{j=1}^{d} b_{ij} v z_j^v + \sum_{k=1}^{d} (A_v)_i z_k^v
+ \sum_{k=2d+1}^{l} (C_k)_i z_k^v + \sum_{k=v+1}^{m} (p_k(v))_i z_k^v \right] + o(h^{p+2}).
\]
But if \( i \geq r_1 \) where \( r_1 \) is the critical index of \( \rho_1 \), we know by Corollary IIC that \( b_{ij}, (A_v)_1, (C_\kappa)_1, (P_\kappa(v))_1 \) are such that the expression in brackets vanishes and the proof is finished.

As an immediate corollary we give

**COROLLARY:** If the hypotheses of Theorem IID hold with the added restriction that \( \rho_1 = \cdots = \rho_m, \sigma_1 = \cdots = \sigma_m \), then

\[
X_n = X(t_n) + h^P e(t_n) + h^q U_n + O(h^{r+1})
\]

where

\[
U_n = \sum_{j=1}^{d} z^n_j x_j(t_n) \quad (x_j(t) \in E_m)
\]

for \( z_1, \ldots, z_d \) the essential roots of \( \rho(z) \)

\[
x''_j(t) = \frac{2\sigma(z_j)}{z^2_2 \rho_j'(z_j)} G(t) x_j(t) \quad x_j(a) = 0 \quad x'_j(a) = b_j
\]

and \( b_j \) has elements \( b_{ij} \) as before.

In a later section we require Theorem IID even for \( q = -1 \), but under the added assumption that \( f \) is linear. Actually the theorem is valid under more general starting conditions than will be given; but since our later application only demands starting conditions \( X_v = 0, v = 0, \ldots, k-2, X_{k-1} = 1 \),
we restrict ourselves to this case. With this latter application in mind we also want an asymptotic expansion of a solution of (0.2) with $X_n$ now a matrix. In view of more wieldy results and since in the study of the summed formula the first two terms of the asymptotic formula are needed, we require that in the following the $A_i$'s be scalar multiples of the identity matrix.

THEOREM IIE: If $\rho \delta$ define a stable, consistent method $M$; $G$ is a known matrix whose elements are in $C'[a,b]$; $\sigma(z_n) \neq 0$,

$n = 1, \ldots, d$ where $z_n$ are the essential roots of $\rho$, then the solution of

$$
\sum_{j=0}^{k} \alpha_j X_{n+j} - h^2 \sum_{j=0}^{k} \beta_j G(t_{n+j}) X_{n+j} = 0
$$

with starting values

$$
X_v = \begin{cases} 
0 & v = 0, \ldots, k-2 \\
1 & v = k-1 
\end{cases}
$$

has an asymptotic representation of the form

$$
X_n = \frac{1}{h} \sum_{j=1}^{d} \frac{2\alpha_k z_j^{n-1}}{\rho''(z_j)} D_j(t_n) + \sum_{j=1}^{d} \frac{2\alpha_k z_j^{n-1}}{\rho''(z_j)} E_j(t_n) \\
+ \sum_{j=1}^{d} A_j z_j^{n-1} + \sum_{j=2d+1}^{m} p_j(n) z_j^{n} + \sum_{j=1}^{m} z_j^{n-1} p_j(t_n) + O(h)
$$

(IIE.1)
where the notation is that of Lemma IIC with the exception that the last sum is over the roots of modulus 1 and for 

\[ 1 \leq j \leq d \]

\[ E_j''(t) = \mu_j^2 G(t)E_j(t) - t_jD_j''(t) \quad E_j(a) = E_j'(a) = 0 \]

\[ t_j = \frac{\mu_j^2}{\sigma(z_j)} \left[ \frac{\rho''(z_j)}{z_j^2} - \frac{1}{\mu_j^2} \frac{z_j^2 \rho''(z_j)}{z_j} \right] \]

\[ \frac{\mu_j^2}{\sigma(z_j)} = \frac{\sigma(z_j)}{z_j^2 \rho''(z_j)} \]

\[ D''(t) = \mu_j^2 G(t)D(t) \quad D(a) = 0 \quad D'(a) = 1 \]

\[ F_j''(t) = \mu_j^2 G(t)F_j(t) - \mu_j^2 G(t)A_j' \quad F_j(a) = F_j'(a) = 0. \]

Proof: The proof follows the lines of Theorem IID.

We note that the difference equation 

\[ \sum_{j=1}^{k} \alpha_j X_{n+j} \]

with starting values \( X_0 = \ldots = X_{k-2} = 0; X_{k-1} = 1 \) has a solution of the form

\[ X_n = \sum_{j=1}^{d} \left( A_j + \frac{2\alpha_j n}{\rho''(z_j)} \right) z_j^{n-1} + \sum_{j=2d+1}^{m} P_j(n) z_j^n \]
which takes on the initial values $X_r, \ldots, X_{k-1}$ where $r$ is the critical index of $\rho$. The notation is that of Corollary IIC. The contribution of the modulus 1 simple roots has been included in the $P_j(n)$ term.

We can write this as

$$X_n = \sum_{j=1}^{d} \frac{2\alpha_k}{\rho''(z_j)} z_j^{n-1} + \sum_{j=1}^{d'} D_j z_j^{n-1} + \sum_{j=2d+1}^{m} P_j'(n) z_j^n$$

where the second sum contains the contribution of the simple roots of modulus 1 and the $O(1)$ contribution of the essential roots. Let

$$z_n = -X_n + \frac{1}{n} \sum_{j=1}^{d} \frac{2\alpha_k z_j^{n-1}}{\rho''(z_j)} D_j(t_n) + \sum_{j=1}^{d^2} A_j z_j^{n-1} + \sum_{j=2d+1}^{m} P_j'(n) z_j^n + \sum_{j=1}^{d} \frac{2\alpha_k z_j^{n-1}}{\rho''(z_j)} E_j(t_n) + \sum_{j=1}^{d} z_j^{n-1} F_j(t_n).$$

As before for $v = r, \ldots, k-1$ ($r$, the critical index $\rho$)

$$z_v = -S_v + \sum_{k=1}^{d} \frac{\nu_k z_v^{n-1}}{\rho''(z_j)} + \sum_{j=1}^{d} A_j z_j^{n-1} + \sum_{j=2d+1}^{m} P_j'(v) z_j^n + O(h^2)$$

$$= O(h^2).$$
We now study $\mathcal{L}(z^n)$ in preparation to the application of Lemma IC. Since the operator $\mathcal{L}$ is linear, each term can be considered separately. Certainly

$$\mathcal{L}(-x_n) = 0.$$ 

After a short computation using the definition of $D_j(t)$ and Taylor's theorem, one finds

$$\mathcal{L}\left(\frac{1}{n} \sum_{j=1}^{d} \frac{2\alpha_j z^{n-1}}{\rho''(z_j)} D_j(t_n)\right) = n^2 \sum_{j=1}^{d} \frac{2\alpha_j r_j z^{n-1}}{\rho''(z_j)} D_j''(t_n) + O(h^3)$$

where

$$r_j = \left[ \frac{\rho''(z_j)}{z_j^3} - \frac{1}{\mu_j} z_j^2 \sigma'(z_j) \right].$$

Similarly,

$$\mathcal{L} \sum_{j=1}^{d} \frac{2\alpha_j z^{n-1}}{\rho''(z_j)} E_j(t_n) = O(h^3) - n^2 \sum_{j=1}^{d} \frac{2\alpha_j z^{n-1}}{\rho''(z_j)} \frac{t_j}{\mu_j} \sigma(z_j) D_j''(t_n).$$

Also

$$\mathcal{L} = \sum_{j=1}^{d'} \frac{A_j z^n}{z_j} = - n^2 \sum_{j=1}^{d'} z_j^{n-1} \sigma(z_j) G(t_n) A_j'.$$
Finally
\[ L \sum_{j=1}^{d'} z_j^{n-1} F_j(t_n) = O(h^3) + h^2 \sum_{j=1}^{k} z_j^{n-1} \sigma(z_j) g(t_n) A_j'. \]

But collecting these results gives
\[ L(z_n) = O(h^3). \]

Now Lemma IC applies to show
\[ z_n = O(h) \]

and the proof is completed.

SECTION III: ROUND-OFF ERROR

Until the present section we have ignored certain practical considerations involved in solving the difference equation (0.2). For example, we have assumed that if \( B_k \neq 0 \), the exact value of \( X_{n+k} \) is obtained before preceding to the next step and that no round-off error is committed during the various other computations involved in finding \( X_{n+k} \). In computational practice this, of course, is not the case. Thus, we are led to study the accumulated round-off error defined by
\[ r_n = \bar{X}_n - X_n \]
where \( \{ \overline{x}_n \} \) is the solution of

\[
\sum_{j=0}^{k} A_j \overline{x}_{n+j} - h^2 \sum_{j=0}^{k} B_j f(t_{n+j}, \overline{x}_{n+j}) = \varepsilon_m.
\]

\( \{ x_n \} \) is the solution of the (0.2) and

\[
\overline{x}_v = x_v \quad v = 0, \ldots, k-1.
\]

Since in practice the \( \varepsilon \)-perturbations arise from rounding, we will call \( \varepsilon_n \) the local rounding.

In the first theorem an order of magnitude estimate for the accumulated round-off error is given.

**THEOREM IIIA:** If \( M \) is a convergent method, \( h \) is such that \( 0 < h < h_0 \) and \( 1 - h^2 \left( \frac{g(h-a)}{h_0} \right) = \delta > 0 \), \( |\varepsilon_n| < \varepsilon \), and \( f \) is in class L, then

\[
|r_n| < \frac{C\varepsilon}{h^2}
\]

where \( C \) is a constant.

Proof: We observe that \( r_n \) satisfies

\[
\sum_{j=0}^{k} A_j r_{n+j} - h^2 \sum_{j=0}^{k} B_j g_{n+j} = \varepsilon_m
\]
where

\[ g_m = f(t_m, \bar{x}_m) - f(t_m, x_m). \]

Since \( f \) is in class \( L \), \( |g_m| \leq L|r_m| \). From the stability of the method, one has

\[ \gamma_n \leq \alpha_n + \beta. \]

Thus, \( h \) can be chosen so that

\[ h > 0. \]

Also

\[ \lambda_n < n\epsilon. \]

Now Lemma IB yields

\[ |r_n| \leq \left( 2A \left( \frac{\alpha n + \beta}{h} \right)^2 + \frac{\alpha_0}{h} + \left( \frac{\alpha n + \beta}{h} \right) \right) \exp \left( h^2 \frac{L(\alpha n + \beta)n}{\delta} \right) \]

or

\[ |r_n| \leq \frac{C\epsilon}{h^2}. \]
Let the sequence \( r'_n \), called the primary component of the accumulated round-off error, be defined as the solution of

\[
A_k r'_{n+k} + \ldots + A_o r'_n - h^2 \left\{ B_k G(t_{n+k}) r'_{n+k} + \ldots + B_o G(t_n) r'_n \right\} = \varepsilon_n
\]

which will be abbreviated to \( \sum r'_n = \varepsilon_n \)

\[
r'_0 = \ldots = r'_{k-1} = 0.
\]

The following theorem implies that the behavior of the accumulated round-off error is essentially that of the primary component.

**THEOREM IIIB:** If \( f \in C^2[a,b] \), \( |\varepsilon_n| \leq Ch^2 \); \( M \) is stable and consistent method; then

\[
|r_n - r'_n| = O(h^2).
\]

**Proof:** We note that

\[
\begin{align*}
\Delta(r_n - r'_n) &= \sum_{j=0}^{k} A_j (r_{n+j} - r'_{n+j}) - h^2 \sum_{j=0}^{k} \left\{ B_j f(t_{n+j}, r_{n+j}) + G(t_{n+j}) r'_{n+j} \right\} - \varepsilon_n \\
&= \sum_{j=0}^{k} A_j (r_{n+j} - r'_{n+j}) - h^2 \sum_{j=0}^{k} B_j G(t_{n+j}) (r_{n+j} - r'_{n+j}) \\
&\quad - \varepsilon_n - h^2 \sum_{j=1}^{k} B_j G_{m+j} = 0.
\end{align*}
\]
But

$$|g_m'| < Ch^2.$$  

Thus,

$$f(r_n - r'_n) = O(h^4).$$

Lemma IC now yields the desired results.

In the remaining part of this section we will concentrate on the primary round-off error. It will be necessary to have a representation for the solution of Eq. (IIIB.1). Such a representation is given by the next two lemmas.

The following lemma was essentially proved in [4] and we do not reproduce the proof.

**LEMMA IIIC:** The solution of $f(y_n) = z_n$, $y_0 = \ldots = y_{k-1} = 0$ is given by

$$y_n = \sum_{p=0}^{n-k} D^p_n [A_k - h^2 B_k G(t_{n+k})]^{-1} z_p$$

where the matrix $D^p_n$ satisfies

$$D^p_n = 0 \quad n < p + k$$

$$D^{p+k}_{p+k} = I \quad \text{(IIIC.1)}$$

$$f(D^p_n) = 0.$$
We can now apply Lemma II in which an asymptotic representation for the solution of the problem defined by (IIIC.1) was derived. This gives

\[ D_k^p = \frac{1}{n} \sum_{j=1}^{d} \frac{2 \alpha k z_{j}^{n-p-1}}{\rho(z_j)} D_j^p(t_n) + O(1) \]

where

\[ \frac{d^2}{dt^2} D_j^p(t) = \mu_j^2 G(t) D_j^p(t); \quad D_j^p(t_{p+1}) = 0 \quad D_j^p(t_{p+1}) = I. \]

A more convenient expression is desired for \( D_j^p \) and we digress to investigate the solution of the differential equation defining \( D_j^p \).

After dropping \( j \) subscript and \( p \) superscript, we study the solution of

\[ D''(t) = \mu^2 G(t) D(t) \]

\[ D(t_{p+1}) = 0 \quad D'(t_{p+1}) = I. \]

Consider the systems

\[ \chi' = A \chi \quad A = \begin{bmatrix} 0 & I \\ \mu^2 G & 0 \end{bmatrix} \quad \chi(a) = I \]
where

\[
\mathbf{X} = \begin{bmatrix}
X_1 & X_2 \\
X_3 & X_4
\end{bmatrix}
\]

\[\theta' = -\theta \mathbf{A} \quad \theta(a) = I\]

where

\[
\theta = \begin{bmatrix}
\theta_1 & \theta_2 \\
\theta_3 & \theta_4
\end{bmatrix}
\]

We observe that

\[\theta(t)\mathbf{X}(t) = I.\]

We also notice that \( \mathbf{X} \) is a fundamental matrix [1] for the system representation of \( \mathbf{D} \). Since

\[
\mathbf{X}(t)\theta(t) = \begin{bmatrix}
X_1 \theta_1 + X_2 \theta_3 & X_1 \theta_2 + X_2 \theta_4 \\
X_3 \theta_1 + X_4 \theta_3 & X_3 \theta_3 + X_4 \theta_4
\end{bmatrix},
\]

we can now write

\[
D(t) = X_1(t)\theta_2(t_{p+1}) + X_2(t)\theta_4(t_{p+n}) = F(t, t_{p+1})
\]
where

\[ x''_1 = \mu^2 g(t) x \quad x_1(a) = I \quad x'_1(a) = 0 \]

\[ \theta''_2 = \mu^2 \theta_2 g \quad \theta_2(a) = 0 \quad \theta'_2(a) = I \]

\[ x''_2 = \mu^2 g x_2 \quad x_2(a) = 0 \quad x'_2(a) = I \]

\[ \theta''_4 = \mu^2 \theta_4 g \quad \theta_4(a) = I \quad \theta'_4(a) = 0 \]

The solution of

\[ D''(t) = \mu^2 GD + g(t) \quad D'(t_{p+1}) = D(t_{p+1}) = 0 \]

may be represented as

\[ D(t) = x_1(t) \int_{t_{p+1}}^{t} \theta_2(\tau) g(\tau) d\tau + x_2(t) \int_{t_{p+1}}^{\tau} \theta_4(\tau) g(\tau) d\tau. \]

This may be verified directly.
The terms \(E_j\) and \(F_j\) of Theorem IIE may be written in this case

\[
E_j(t_n) = x_1(t_n) \int_{t_{p+l}}^{t_n} \theta_2(t)(-t_j)D'_j dt + x_2(t_n) \int_{t_{p+l}}^{t_n} \theta_4(t)(-t_j)D''_j(t) dt
\]

\[
= \varepsilon_j(t_n, t_{p+l})
\]

\[
F_j(t_n) = x_1(t_n) \int_{t_{p+l}}^{t_j} \theta_2(t)(-\phi_j G(t)A'_j) dt + x_2(t_n) \int_{t_{p+l}}^{t_n} \theta_4(-\phi_j G(t)A'_j) dt
\]

\[
= F_j(t_n, t_{p+l}).
\]

Thus

\[
D^p_n = \frac{1}{n} \sum_{j=1}^{d} \frac{2\alpha_k z^{n-p}_j}{\rho''(z_j)} F_j(t_n, t_{p+l}) + \sum_{j=1}^{d} \frac{z\alpha_k z^{n-p}_j}{\rho''(z_j)} \varepsilon_j(t_n, t_{p+l})
\]

\[
+ \sum_{j=1}^{d} A_j z^{n-p}_j + \sum_{j=1}^{d'} z^{n-p}_{j} f_j(t_n, t_{p+l}) + \theta_\zeta(n)
\]
where \( \zeta(h) = \frac{Ch}{(h + \log t^{-1})} \). The later estimate arises as follows. The expression \( \sum_{j=2d+1}^{m} P_j(n)z_j^n \) given in (IIIE.1) results from the roots whose modulus is smaller than 1: i.e.,

\[
t^2 = \max_{2d+1 \leq j \leq m} |z_j| < 1.
\]

We note

\[
|n^kz_j^k| \leq n^k t^{2k} \leq C t^n = C e^{-n \log t^{-1}}
\]

\[
\leq \frac{C}{1 + n \log t^{-1}}
\]

\[
= \frac{C h}{h + (X_n - a) \log t^{-1}}, \quad 2d + 1 \leq j \leq m.
\]

Thus

\[
\left| \sum_{j=2d+1}^{m} P_j(n)z_j^n \right| \leq \frac{Ch}{h + (X_n - a) \log t^{-1}}
\]

But now we have

**THEOREM IIID:** The primary round-off error satisfies

\[
r'_n = \frac{1}{h} \sum_{p=0}^{n-k} \left\{ \sum_{j=1}^{d} \frac{2z_j^n - p}{\rho''(z_j)} P_j(t_n, t_{p+1}) + O(h) \right\} \varepsilon_p.
\]

The \( O(h) \) term can be replaced by \( \theta_p h \) where \( \theta_p \) is a matrix whose norm is uniformly bounded by \( C \).
It has long been known that upper bounds for the "error," which are based merely on \( f \) being in class \( L \), are unrealistic and work has been done in several directions \([9],[3]\) to overcome this. We follow the direction of \([3]\) and assume \( \varepsilon_n \) are random variables with a known distribution. This assumption is justified by the realistic results presented in \([3]\), although \( \{\varepsilon_n\} \) are not strictly random variables. Without repeating the details given there \([3]\), we assume \( \{\varepsilon_n\} \) are random variables whose expected value and covariance are given by

\[
E(\varepsilon_p) = \mu q(t_p)
\]

\[
E(\varepsilon_p - \mu_p)(\varepsilon_q^* - \mu_q^*) = \begin{cases} 0 & p \neq q \\ r^2c(t_p) & p = q \end{cases}
\]

where \( q, C(t), \varepsilon \in C'[a,b] \) and \( * \) denotes conjugate transpose.

From \((\text{IIIID.2})\) we have

\[
E[r_n^t] = \frac{h}{n} \left[ \sum_{p=0}^{n-k} \sum_{j=0}^d \frac{2z_j^{n-p}}{p^n(z_j)} F_j(t_n, t_{p+1})q(t_p) + h\theta_p q(t_p) \right]
\]

\[
= \frac{h}{n} \sum_{j=1}^d \frac{2}{p^n(z_j)} \sum_{p=0}^{n-k} z_j^{n-P} F_j(t_n, t_{p+1})q(t_p) + \frac{h}{n} \sum_{j=1}^d \sum_{p=0}^{n-k} h\theta_p q(t_p).
\]
By Lemmas 3.6 and 3.7 of [4], which are stated at the end of the proof, this reduces to

\[
E[r'_n] = \frac{n}{h^2} \int_0^{t_n} F_2(t_n,t)q(t)dt + O\left(\frac{1}{h}\right).
\]

The integral can be replaced by \(H\), the solution of

\[
H'' = GH + q \quad H(a) = H'(a) = 0.
\]

Lemma 3.6. Let \(f\) be of class \(C'[a,b]\). If \(t_n = t \in [a,b]\) is fixed and \(h \to 0\), then

\[
h \sum_{p=0}^{n-k} f(t_p) = \int_a^t f(t)dt + O(h).
\]

Lemma 3.7. Let \(z\) be a complex number of modulus 1 and let the function \(f\) of class \(C'[a,b]\). Then as \(h \to 0\), \(z \neq 1\)

\[
h \sum_{p=0}^{n-k} z^p f(t_p) = O(h)
\]

uniformly for

\(t_n \in [a,b]\).
We now study the covariance. We write (by Theorem IIID)

\[ r'_n = \sum_{p=0}^{n-k} \Delta^p_n \varepsilon_p \]

where

\[ \Delta^p_n = \frac{1}{h} \sum_{j=1}^{d} \frac{2z_j^{n-p}}{\rho''(z_j)} F_j(t_n, t_{p+1}). \]

Now

\[
\text{covar}(r'_n) = E(r'_n - E(r'_n))(r'_n - E(r'_n)) \\
= \sum_{p=0}^{n-k} \sum_{q=0}^{n-k} \Delta^p_n (\varepsilon_p - E_p)(\varepsilon_q - E_q)^* \Delta^q_n \\
= \sigma^2 \sum_{p=0}^{n-k} \Delta^p_n c(t_p) \Delta^p_n. \\

\]

Let

\[ \Delta^p_{nj} = F_j(t_n, t_{p+1}). \]
Now

\[ \Delta^p_n = \sum_{j=1}^{d} \frac{2z_j^{n-p}}{h \rho''(z_j)} \Delta^p_{n-j} + o(1). \]

Therefore

\[ \text{Covar}(r'_n) = \frac{a^2}{h^3} \sum_{j=1}^{d} \sum_{\mu=1}^{d} \Gamma_{j\mu} \Delta_{n \mu j} \]

where

\[ \Gamma_{j\mu} = \frac{1}{\rho''(z_j) \rho''(z_\mu)} \]

\[ \Delta_{n \mu j} = h \sum_{p=0}^{n-k} (z_\mu z_j)^{n-p} \Delta^p_{n \mu} \Delta^p_{n j}. \]

But by Lemma 3.6 and 3.7 of [4]

\[ \Delta_{n \mu} = \begin{cases} \int_{0}^{t_n} F_\mu(t_n, t) C(t) F_\mu^*(t_n, t) \, dt & \mu = \mu \\ 0(h) & \mu \neq \mu \end{cases} \]
Finally

\[
\text{Covar}(r'_n) = \sigma^2 \frac{2}{h^3} \sum_{j=1}^{d} \frac{1}{|\rho''(z_j)|^2} \int_0^{t_n} F_j(t_n,t) c(t) F_j^*(t_n,t) \, dt + o\left(\frac{1}{h^2}\right)
\]

where

\[
\frac{d^2}{d\tau^2} F_j(\tau,t) = \mu_j^2 G(\tau) F_j(\tau,t)
\]

\[
F_j(\tau,\tau) = v \left. \frac{d}{d\tau} F_j(\tau,t) \right|_{\tau=t} = I
\]

\[
\mu_j^2 = \frac{2\sigma(z_j)}{z_j^2 |\rho''(z_j)|}
\]

These last theorems point out that it is in general questionable to use the approach of (0.2). If one replaces (0.1) by

\[
X' = u, \quad U' = f(t,u), \quad U(a) = B X(a) = A
\]

and then approximates this by standard first order techniques [3],[4], the round-off error behaves like $1/h$ instead of $1/h^2$. Attempts have been made in other directions to control the strong growth of the round-off error, and we shall study these in the next section.

Since $h(z)$ has roots only on $|Z| = 1$, one can write

\[
\frac{1}{z^j h(1/z)} = \sum_{k=0}^{\infty} \gamma_k z^k.
\]
SECTION IV: SUMMED FORMULA

In [3] it is pointed out that the "second summed method" or $\Sigma^2$ procedure which has long been used by astronomers is "equivalent" to Cowell's method and that its merit lies in the reduction of round-off error. In this study the summed formula was extended to more general techniques than Cowell's method, viz. to methods whose characteristic equation is of the form $R(z) = (z-1)^2 f(z)$ where $f$ has no zeros of modulus 1 and multiplicity 2. We propose to give an extension to yet more general characteristic polynomials.

In general outline we follow the derivation of [3] but rather than merely summing, we form a weighted sum. By using certain properties of the weights, we are able to find a new set of difference equations* which would yield the same sequence $\{x_n\}$ were there no round-off error and

\[ x(t) = a_1 x_1 + \ldots + a_n x_n. \]

*The idea may also be viewed as a difference equation analog to Lobatto's method of finding a particular solution of an inhomogeneous linear differential equation with constant coefficients. Roughly, $\alpha_n y^{(n)} + \ldots + \alpha_0 y = \alpha_k (D-m_1)(D-m_2)\ldots(D-m_n)y = x(t), (D = d/dt)$. Then $y = \frac{1}{(D-m_1)(D-m_2)\ldots(D-m_n)} \frac{x(t)}{\alpha_k}$

\[ = \frac{a_1}{D-m_1} x + \ldots + \frac{a_n}{D-m_n} x = a_1 e^{m_1 t} \int e^{-m_1 t} x dt + \ldots + a_n e^{m_n t} \int e^{-m_n t} x dt. \]

In our instance the factors $D-m_1$ appear as squares and $D$ means difference. We only divide by one of the square's factors, and the final integration becomes a summation.
were the starting values the same. But we shall prove that with respect to the propagation of round-off error the performance of the new algorithm is quite different.

It is clear that every stable characteristic polynomial can be written in the form \( p(z) = h^2(z)f(z) \) where all essential roots of \( p \) are included in the \( h^2 \) factor. We require that if \( z_1 \) is a root of \( h(z) = 0 \), then \( z \) is a \( n \)th root of unity. (This requirement is not completely artificial. Indeed, we know from [3] p. 342 that all optimal \( p \) has only essential roots], monic, stable polynomials with integral coefficients can have among their roots only zero or roots of unity.) For definiteness we suppose \( h \) is of degree \( t \) and \( p \) of degree \( J \). First we observe \( \{\gamma_k\} \) is periodic with period \( L(\geq t) \). Indeed, \( h(z) \) has only roots of unity as zeros. It is clear that \( L = \text{least common multiple of } \{n_1, \ldots, n_t\} \). We write

\[
h^2(z) = \sum_{j=0}^{t} a_j z^j \quad h(z) = \sum_{j=0}^{t} \alpha_j z^j.
\]

Thus, \( \sum_{j=0}^{t} \alpha_{t-j} z^j = \sum_{j=0}^{t} a_{t-j} z^j \sum_{k=0}^{\infty} \gamma_k z^k \). From this one concludes

\[
\alpha_t = a_{2t} \gamma_0
\]
\[
\alpha_{t-1} = a_{2t} \gamma_1 + a_{2t-1} \gamma_0
\]
\[ a_0 = a_2 t'_{t+1} + \ldots + a_{t-1} t' \]

\[ 0 = a_2 t'_{k} + \ldots + a_{2l-k} t' \quad t + 1 \leq k \leq 2l \]

\[ 0 = a_2 t'_{k} + \ldots + a_{0} t'_{k-2l} \quad k \geq 2l. \]

Define the operator \( L \) by \( L x_n = x_{n+1} \).

The difference equation (0.2) for \( p \) as hypothesized can now be written as

\[
f(L)\{a_2 t' x_{m+2l} + \ldots + a_0 x_m\} = h^2 \sum_{j=0}^{k} \beta_j f(t_{m+j}, x_{m+j}).
\]

For \( t = 0, \ldots, n-J \) we multiply the equation corresponding to \( m = n-J-t \) by \( \gamma_t \) and sum the resulting equations. The right side of the sum has the form

\[
h^2[\gamma_0 [\beta_k f(t_{n+j}, x_{n+j}) + \ldots + \beta_0 f(t_{n}, x_{n})] + \gamma_1 [\beta_k f(t_{n+j}, x_{n+j}) + \ldots + \beta_0 f(t_{j}, x_{j})] + \ldots + \gamma_n [\beta_k f(t_{n}, x_{n}) + \ldots + \beta_0 f(t_{0}, x_{0})].
\]

Making use of the relations between \( \gamma_k, a_k, c_k \) one can reduce the left side to

\[
f(L)\{a_2 t'_{n+2l} + \ldots + a_0 t'_{n+l}\} + f(L)\{\gamma_{n} a_{n} t'_{n} + \ldots + x_{2l-1} (a_{2l-1} \gamma_{n} + \ldots + a_{0} \gamma_{n-2l})\}.
\]

Since \( \gamma_n \) is periodic, \( \gamma_n = \gamma_j \) where \( 0 \leq j < L \) and the second part of the sum can be written as

\[
f(L)\{\gamma_{j} a_{n} t'_{0} + \ldots + x_{2l-1} (a_{2l-1} \gamma_{j} + \ldots + a_{0} \gamma_{j-2l})\}.
\]
No such reduction was found for the right side. Thus, our weighted sum appears as

\[ f(L)\{\alpha \gamma^a x_{n+2l} + \cdots + \alpha \gamma^b x_{n+l}\} + f(L)\{\gamma^a x_{n} + \cdots + x_{2l-1}(a_{2l-1}\gamma^a + \cdots + a_{2l-1} \gamma_{2l+1})\} \]

\[ = h^2 \left\{ \sum_{j=0}^{J} \sum_{j=0}^{n} \gamma_{n-j} f(t_{n+j}, x_{n+j}) \right\}. \]

(IVA.1)

Now suppose we define \( H^j \) by

\[ h^2 (L) H^j = f(L)\{\gamma^a x_{n} + \cdots + x_{2l-1}(a_{2l-1}\gamma^a + \cdots + a_{2l-1} \gamma_{2l+1})\}. \]

For the same \( m \) and hence same \( j \) we define

\[ P^j_{n+j} = h \sum_{j=0}^{n} \gamma_{n-j} f(t_{j+n}, x_{j+n}) - H^j \]

\[ \vdots \]

\[ P^0_{n+j} = h \sum_{j=0}^{n} \gamma_{n-j} f(t_j, x_j) - H^j. \]
Let us study $h(L)F_{n+k}^{kJ}_{n+j-k}$. From the relations between $\{a_j\}$ and $\{\gamma_j\}$ and since $h(1) = 0$, we have

\[
\begin{align*}
(\alpha^L + \cdots + \alpha_o)[\gamma^o f_{j+n-l} + \cdots + \gamma^f_J] \\
= \alpha^L[\gamma^o f_{j+n} + \cdots + \gamma^o f_{k+n-l} + \cdots + \gamma^f_J] + \cdots + \alpha_o[\gamma^o f_{j+n-l} + \cdots + \gamma^f_J] \\
= \alpha^L[\gamma^o f_{j+n} + \cdots + \alpha_o \gamma^o f_{j+n-l} + \cdots + \gamma^f_J] \\
+ \cdots + \alpha_o[\gamma^o f_{j+n-l} + \cdots + \gamma^f_J]
\end{align*}
\]

where

\[f_m = f(t_m, X_m).\]

Finally

\[h(L)F_{n+k-l}^{kJ} = f(t_{n+j}, X_{n+j}).\]

Thus, we may write (IVA.1) and (IVA.2) (noticing the definition of $H^J$) as

\[f(L)\{a^L X_{n+2l} + \cdots + a_o X_{n+l}\} = h_f^j f_{n+j} + \cdots + \beta_o F_{n+j}^o J\}

\[h(L)F_{n+j}^J = h_f(t_{n+j}, X_{n+j}) \quad \text{(IVA.3)}\]

\[h(L)F_{n+j-l}^J = h_f(t_n, X_n).\]
For the particular \( n \) and more generally for \( n \equiv j \mod L \) and if the second set of equations in (IVA.3) have starting values as discussed, this new system produces the same \( X_{n+j} \) as (0.2) (ignoring round-off error). That (0.2) implies (IVA.3) follows from our derivation. The converse can be obtained by replacing \( n \) by \( n-t \) and applying the operator \( h(L) \) to the first equation given under (IVA.3).

But the new system still is incomplete in that we can compute \( X_{n+j} \) only for \( n \equiv j \mod L \). However, for each \( n \) one can form the analogous sum. There result \( L \) different cases. But it is clear that in each case the equations of (IVA.3) will have the same form; but since the \( H^j \) are different, the second set of equations will have different starting values. We now can write

\[
f(L)h(L)X_n = h \sum_{k=0}^{J} \beta_k p^kj \n_{n+j}
\]

\[
h(L)P^j_{n+j-l} = hf(t_{n+j}, X_{n+j})
\]

(IVA.4)

\[
\vdots
\]

\[
h(L)P^j_{n+j-l} = hf(t_n, X_n).
\]
For $\beta_J \neq 0$ the computation of $X_{n+J}$ proceeds as follows: first, one decides which second formula is applied, i.e., what is $J$, $X_{n+J}$ is guessed and then $F_{n+J}^J$ is computed. Since $F_{n+i}^J 0 \leq k < J$ involve previously computed values, can we use the first part of (IVA.3) to compute a new $X_{n+J}$. With this value for $X_{n+J}$, we can begin anew and so on recursively.

That the process converges follows from the choice of $h$ and that $f$ is class $L$ by the fixed point theorem given in the introduction.

Unfortunately, our previous theorems, e.g., Lemma IB, do not permit estimates of the growth of the round-off error for such a system as (IVA.4); therefore, we must write (IVA.4) in an equivalent form which will be tractable by our methods. We write

$$f(L)[\alpha_L X_{n+2L} + \ldots + \alpha_0 X_{n+L}]$$

$$= h\left[\beta_0 \left[ \delta_{0,n \mod L} F_{n+J}^0 + \delta_{1,n \mod L} F_{n+J}^1 + \ldots + \delta_{L-1,n \mod L} F_{n+J}^{L-1} \right] \right.$$

$$+ \ldots + \beta_0 \left[ \delta_{0,n \mod L} F_{n+J}^0 + \ldots + \delta_{L-1,n \mod L} F_{n+J}^{L-1} \right]\left] \right.$$  

(IVA.5)

$$h(L)F_{n+J}^J = h F_{n+J}$$

$$\vdots$$

$$J = 0, \ldots, L-1$$

$$h(L)F_{n+J}^0 = h F_n$$
where $\delta_{j} n \mod L$ is the Kronecker delta. It is clear that the solution of (IVA.5) is identical to that of (IVA.4), provided both have the same starting values.

**THEOREM IVA:** If $f(t,X)$ is in class $L$, the local rounding is bounded by $\epsilon$; $h(z), f(z)$ are as previously stated, then the accumulated round-off error of the numerical solution of (0.1) by (IVA.5) satisfies

$$|r_n| \leq \frac{(X_n-a)}{h} \frac{\Gamma_0}{(1h^2B)} e^c$$

where $\Gamma, c$ and $B$ are constants.

**Proof:** The numerical values actually calculated satisfy

$$f(L)h(L)\tilde{X}_{n+\ell} = h\left\{\beta_j \Delta F_{n+J}^j + \ldots + \beta_0 \Delta F_{n+J}^0 \right\} + \epsilon_{n+J}$$

$$h(L)\hat{F}_{n+J-\ell}^j = hf(t_n, \hat{X}_n) + y_{n+J}^j$$

$$h(L)\hat{F}_{n+J-\ell}^j = hf(t_n, \hat{X}_n) + y_{n+J}^j$$

where $\Delta F_{n+J}^j = \delta_0 n \mod L \hat{F}_{n+J}^0 + \ldots + \delta_{L-1} n \mod L \hat{F}_{n+J}^{kL-1}$

$\epsilon_{n+J}, \eta_{n+k}^{ij}$ are the local round-off errors. Let

$$r_n = \tilde{X}_n - X_n \quad F_n^{kj} = F_n^{kj} - F_n^{kj}.$$
Subtracting from (IVA.6) the corresponding relation in (IVA.5), we find

\[ f(L)h(L)r_{n+1} = h\left[ \beta_j \left( \Delta R^j_{n+J} \right) + \ldots + \beta_0 \left( \Delta R^0_{n+J} \right) \right] + \varepsilon_{n+k} \]

(IVA.7)

\[ h(L)R^i_{n+J-1} = hG(t_n)\Delta^i_{n+J} + \eta^i_{n+J} \quad i = 0, \ldots, J \]

\[ j = 0, \ldots, L-1 \]

where

\[
(G(t_n))_{ij} = \begin{cases} 
\frac{(f(t_n, x_n) - f(t_n, x_n))_i}{(x)_j - (x)^i_j} & (x)_j \neq (x)_j \\
0 & \text{otherwise.}
\end{cases}
\]

We wish to apply Lemma IB to (IVA.7). Since the starting values for (IVA.7) are 0, we may take \( H_1 = H_2 \). Since \( h(z), f(z) \) and \( h(z) \) satisfy the hypothesis of Lemma 5.5 of [3], we have \( |\gamma_n| \leq \Gamma \) where \( \Gamma \) is a constant. We can take

\[
\lambda_n = n \max_{0 \leq t \leq n-J} \left( |\varepsilon_t|, |\eta^k t| \right) = n \varepsilon.
\]
We have

\[ B_{n+j}(x_n) = \left[ B_1 \left( b_n \mod L^{n+j} + \cdots + b_{L-1} \mod L^{n+j} \right) \right] . \]

Therefore

\[ |B_{n+j}(x_n)| \leq |B_1| (b_n \mod L + \cdots + b_{L-1} \mod L) \sum_{j=1}^{L-1} |R_{n+j}^k| + mt|r_{n+k}| \]

\[ \leq (|B_1|L + mt) \left( \sum_{i=1}^{L-1} |R_{n+i}^k| + |r_{n+k}| \right) . \]

Thus, we may take

\[ B = \max_{0 \leq t \leq J} \{|B_1|L + mt\} . \]

But now we have verified the hypothesis of Lemma IB and the desired results follow.

In order to perform a statistical analysis, we must require that the only modulus 1 roots of \( \rho \) are essential roots. We again must restrict our attention to the primary component of the round-off error.
We notice that (IVA.6) can be written in the form

\[ f(L)h^2(L)\overline{x}_{n+\ell} = h\left\{ \beta_j h(L)\Delta^J_{n+J} + \ldots + \beta_0 h(y)\Delta^0_{n+J} \right\} + h(L)e_{n+J} \]

or

\[ = h\left\{ \beta_j \left\{ \delta \text{ on mod } L \left( h(t_{n+J+\ell}, \overline{x}_{n+J+\ell}) + y_{n+\ell+J}^0 \right) \right\} + \ldots + \delta_{L-1} \text{ n mod } L \left( h(t_{n+\ell}, \overline{x}_{n+\ell}) + y_{n+\ell}^0 \right) \right\} + \ldots + \beta_0 \left\{ \delta \text{ on mod } L \left( h(t_{n+\ell}, \overline{x}_{n+\ell}) + y_{n+\ell}^0 \right) \right\} + \ldots + \delta_{L-1} \text{ n mod } L \left( h(t_{n+\ell}, \overline{x}_{n+\ell}) + y_{n+\ell}^{0L-1} \right) \right\} + h(L)e_{n+J} \]

Let \( \Delta^m_n = \delta \text{ on mod } L \overline{t}_n^{m0} + \ldots + \delta_{L-1} \text{ n mod } L \overline{t}_n^{mL-1} \]

\[ m = 0, \ldots, J. \]

Now

\[ f(L)h^2(L)\overline{x}_{n+\ell} = h^2\left\{ \beta_j f(t_{n+J+\ell}, \overline{x}_{n+J+\ell}) + \ldots + \beta_0 f(t_{n+\ell}, \overline{x}_{n+\ell}) \right\} \]

\[ + h\left\{ \beta_j \Delta^J_m + \ldots + \beta_0 \Delta^0_{n-J} \right\} + h(L)e_{n+J}. \]
Let

\[ X_{n+t} = \beta_j \Delta_j^n + \ldots + \beta_0 \Delta_0^{n-J}. \]

By the previous theorem and with \( \varepsilon = O(h^2) \) the primary round-off error \( R_n \) satisfies

\[ f(L)h^2(L)R_{n+t} = h^2 \left\{ \sum_{k=0}^{J} \beta_k g(t_{n+k+t})R_{n+t+k} \right\} + h\chi_{n+t} + h(L)\varepsilon_{n+J}. \]

Thus, from Lemma IIIA in which the solution of an inhomogeneous difference equation is represented in terms of a solution of the homogeneous equation, we can write

\[ R_{n+t} = \sum_{p=0}^{n+t-J} D^p_{n+t} \{ h\chi_p + h(L)\varepsilon_{p+J-t} \}. \]

We will study the \( \varepsilon \) and \( \chi \) contributions separately. First, since \( \varepsilon_o = \ldots = \varepsilon_{J-1} = 0 \)

\[ \sum_{p=0}^{n+t-J} D^p_{n+t} \left\{ \sum_{k=0}^{t} \varepsilon_{p+J-k} \right\} = \sum_{p=0}^{n+t-J} \alpha_p D^p_{n+t} + \alpha_{-1} D^{p+1}_{n+t} + \ldots + \alpha_{-o} D^{p+o}_{n+t} \varepsilon_{p+J}. \]
But from the remarks preceding Theorem (IIIID), we can write

\[ p_{n+t}^{p+m} = \frac{1}{h} \sum_{j=1}^{d} a_j^* \frac{z_n-p-m}{\rho''(z_j)} F_j(t_n t_{p+m+1}) + \sum_{j=1}^{d'} \frac{2a_j^* z_n-p-m}{h''(z_j)} e_j(t_n t_{n+m}) \]

\[ + \sum_{j=1}^{d} A_j z_n-p-m-l + \sum_{j=1}^{d} z_j^{n-p-m} F_j(t_n t_{p+m}) + \theta_2(h) \]

where \( a_j^* \) is the leading coefficient of \( \rho(z) \). Now since \( z_1, \ldots, z_d \) are roots of \( h(z) \), we have

\[ \sum_{k=0}^{d} \frac{d}{n+t} = \sum_{j=1}^{d} \frac{2a_j^* z_n-p-t+1}{\rho''(z_j)} h''(z_j) \varepsilon_{2j}(t_n t_{p+1}) \]

\[ + h \sum_{j=1}^{d} \frac{2a_j^* z_n-p-t+1}{h''(z_j)} \varepsilon_{2j}(t_n t_{p+1}) \]

\[ + h \sum_{j=1}^{d} z_j^{n-p-t-l} h'(z_j) F_j(t_n t_{p+1}) + \theta_2(h), \]

where the subscript 2 denotes the partial derivative with respect to the second variable.
Apparently no simplification can be effected in the \( x_p \). We now can write

\[
R_{n+t} = \sum_{p=0}^{n+t-J} \left\{ \sum_{j=1}^{d} \left( \frac{2}{\rho''(z_j)} \right) z_j^{n-p} F_j(t_n, t_p+1) x_p \right. \\
+ \sum_{j=1}^{d} \frac{2z_j^{n-p-t+1}}{h'(z_j) F_2(t_n, t_{n+1}) + O(h)} \left. \right\} \varepsilon_{p+J}.
\]

Let us assume that the local round-off errors \( \varepsilon_{p+J} \) are vectors whose components are random variables with known expected values which can be expressed by

\[
E(\varepsilon_p) = \mu q(t_p)
\]

where \( q, q^{kt} \in C_p[a,b] \)

\[
E(\eta_{p}^{kt}) = \mu^{kt} q^{kt}(t_p)
\]

where \( \mu \) is independent of \( t_p \) but is \( O(h^3) \). A discussion of \( q \) is given by [3]. Certainly

\[
E(R_{n+t}) = \frac{1}{h} (M_n t + \mu N_n t + Z_{n+t})
\]
where

\[ M_{n+t} = h \sum_{p=0}^{n+t-J} \sum_{j=1}^{d} \frac{2z^{n-p}}{p''(z_j)} F_j(t_n, t_{p+j}) E(x_p) \]

\[ N_{n+t} = h \sum_{p=0}^{n+t-J} \sum_{j=1}^{d} \frac{2z^{n-p-J+1}}{p''(z_j)} h'(z_j) F_{2j}(t_n t_{p+1}) q(t_{p+1}) \]

\[ Z_{n+t} = h \sum_{p=0}^{n+t-J} \zeta(h) q(t_{p+j}). \]

We begin by studying \( N_{n+t} \). By 3.6 and Lemma 3.7 of [4], we can write

\[ N_{n+t} = \int_a^{t_n} \frac{2h'(1)}{\rho''(1)} F_{2j}(t_n, t) q(t) dt. \]

Next notice that

\[ |Z_{n+t}| \leq \max_{a \leq t \leq b} |q| h^2 \sum_{p=0}^{n+t-J} \frac{C}{h + (n-p-1) \log t^{-1}} \]

\[ \leq \max_{a \leq t \leq b} |q(t)| \hat{O}(h \log h^{-1}). \]
Now

\[ E(\chi_p) = \sum_{n=0}^{J} \beta_n \left( \Delta_p^{n-l-k} \right) \]

\[ = \beta_J \left\{ \delta_{q_{p+J}} \sum_{n=0}^{J-L-1} (J-L-1) q_{p+J} \right\} + \ldots + \delta_{n \mod L} \sum_{n=0}^{J-L-1} q_{p+J} \]

\[ + \ldots + \beta_0 \left\{ \delta_{q_{p+J}} \sum_{n=0}^{J-L-1} q_{p+J} \right\} \]

Let us consider a typical term in \( M_{n+l} \), say

\[ \left( \sum_{p=0}^{n+l-J} \sum_{j=1}^{d} \frac{2z^{n-p}_j}{\rho^n(z_j)} \right) F_j(t_n, t_{p+1}) \delta_{n \mod L} \sum_{n=0}^{J-L-1} k_1 q_{p+J} \]

\[ = \sum_{j=1}^{d} \frac{1}{\rho^n(z_j)} \left( \sum_{p=0}^{n+l-J} 2z^{n-p}_j F_j(t_n, t_{p+1}) \delta_{n \mod L} \sum_{n=0}^{J-L-1} k_1 q_{p+J} \right). \]

But again by Lemmas 3.6 and 3.7,

\[ = 2^{n-1} \delta_{n \mod L} \int_{0}^{t_n} F_1(t_n, t) q^{k_1}(t) dt + O(h). \]

Thus, we have

\[ M_{n+l} = \sum_{0}^{J} \sum_{i=0}^{l-1} \mu^{i} m_{n \mod L} \int_{0}^{t_n} F_1(t_n, t) q^{i}(t) dt. \]
Finally

\[ E(R_{n+t}) = \frac{1}{h} \sum_{j=1}^{d} \left[ \sum_{i=0}^{J} \beta_i \sum_{m=0}^{\mu_1 \delta_{m \mod L}} \int_{0}^{t_n} F_j(t_{n}, t) q_{i}^{j \mu m} \, dt \right. \]

\[ + \int_{0}^{t_n} \frac{2b'(z_j)}{\rho''(z_j)} F_{2j}(t_{n}, t) q(t) \, dt \left. \right] + O(1). \]

Now let us study the covariance. We assume

\[ E\{[X_p - E(X_p)][X_q^T - E^T(X_q)]\} = \delta_{pq} Q(t_p) \]

\[ E\{[\epsilon_{p+J} - E(\epsilon_{p+J})][\epsilon_{q+J}^T - E(\epsilon_{q+J})]\} = \delta_{p+J, q+J} C(t_{p+J}) \]

\[ E\{[\epsilon_{p+J} - E(\epsilon_{p+J})][X_q^T - E^T(X_q)]\} = 0 \]

where \(Q\) and \(C\) are in \(C'(a,b]\). Let us write

\[ R_{n+t} = \sum_{p=0}^{n+t-J} D_{np} X_p + H_{np} \epsilon_{p+J}. \]
Now
\[
\text{covar}(R_{n+t}) = E\left( \sum_{p=0}^{n+t-J} D_{np} X_p + H_{np} \varepsilon_{p+J} - E(R_{n+t}) \right) - E(R_{n+t})
\]

\[
= \sum_{q=0}^{n+t-J} X_q^T D_{nq} \varepsilon_q + \varepsilon_{q+J}^T H_{nq} - E(\varepsilon_{q+J})
\]

\[
= E\left( \sum_{p=0}^{n+t-J} D_{np} X_p + H_{np} \varepsilon_{p+J} - \sum_{p=0}^{n+t-J} D_{np} E(X_p) + H_{np} E(\varepsilon_{p+J}) \right)
\]

\[
= E\left( \sum_{q=0}^{n+t-J} X_q^T D_{nq} \varepsilon_q + \varepsilon_{q+J}^T H_{nq} - \sum_{q=0}^{n+t-J} E(\varepsilon_q)^T D_{nq} \varepsilon_q - E(\varepsilon_{q+J})^T H_{nq} \right)
\]

can be written as

\[
\sum_{p=0}^{n+t-J} \sum_{q=0}^{n+t-J} D_{np} E\{[X_p - E(X_p)][X_q^T - E^T(X_q)]\}^T D_{nq}
\]

\[
+ D_{np} E\{[X_p - E(X_p)][\varepsilon_{q+J}^T - E^T(\varepsilon_{q+J})]\}^T H_{nq}
\]

\[
+ H_{np} E\{[\varepsilon_{p+J} - E(\varepsilon_{p+J})][\varepsilon_{q+J}^T - E(\varepsilon_{q+J})]\}^T H_{nq}
\]

\[
+ H_{np} E\{[\varepsilon_{p+J} - E(\varepsilon_{p+J})][X_q^T - E^T(X_q)]\}^T D_{nq}
\]
which reduces to

\[
\sum_{p=0}^{n+t-J} D_{np} Q(t_q) D_{nq}^T + \sum_{p=0}^{n+t-J} H_{np} C(t_{p+j}) H_{nq}^T
\]

But a lengthy calculation, which is very similar to the one made to compute the expected value, gives

\[
\text{Covar}(R_{n+t'}) = \frac{1}{n} \sum_{j=1}^{d} \frac{1}{\rho''(z_j)} \int_{0}^{t_n} F_j(t_n,t)Q(t)F_j^T(t_n,t)dt
\]
\[
+ \frac{1}{h} \sum_{j=1}^{d} \frac{2h'(z_j)}{\rho''(z_j)} \int_{0}^{t_n} F_{2j}(t_n,t)C(t)F_{2j}^T(t_n,t)dt + O(1).
\]
SECTION V: TWO EXAMPLES OF CIRCULAR MOTION

In this section the theorems of Section IV will be applied to two simple initial value problems. The computations will be kept as independent of a particular method as possible.

The vector $X(t)$ will now have two components denoted by $X$ and $Y$. The problems are:

$$
X'' = -X \quad X(0) = 1 \quad X'(0) = 0 \quad (V.1)
Y'' = -Y \quad Y(0) = 0 \quad Y'(0) = 1
$$

$$
X'' = -\frac{X}{(x^2+y^2)^{3/2}} \quad X(0) = 1 \quad X'(0) = 0 \quad (V.2)
Y'' = -\frac{Y}{(x^2+y^2)^{3/2}} \quad Y(0) = 0 \quad Y'(0) = 1
$$

It may be verified directly that both systems have the identical solution

$$
X(t) = \cos(t) \quad (V.3)
$$

$$
Y(t) = \sin(t)
$$

If these initial value problems are solved by a multistep method $M$ of order $P$ and if the starting values are in error by at least $O(h^{P+2})$, then the asymptotic representation given by Theorem IID is
\[
\begin{pmatrix}
X_n \\
Y_n
\end{pmatrix} = \begin{pmatrix}
X(t_n) \\
Y(t_n)
\end{pmatrix} + h^p \begin{pmatrix}
e_1(t_n) \\
e_2(t_n)
\end{pmatrix} + o(h^{p+1})
\]

where \(e'' = G e - CX(t)^{(P+2)}\), \(e(0) = e'(0) = 0\)

\[
(0)_{ij} = \frac{\partial f_j(t, X(t))}{\partial (X)_i}
\]

\(C = \text{diag } C_1\)

and \(C_1, C_2\) are constants which depend on the method \(M\).

A short calculation gives that for (V.1)

\[
G_1 = \begin{pmatrix}
-1 & 0 \\
0 & 1
\end{pmatrix}
\]

and for (V.2)

\[
G_2 = \begin{pmatrix}
\frac{3}{2} \cos 2t + \frac{1}{2} & \frac{3}{2} \sin t \\
\frac{3}{2} \sin 2t & -\frac{3}{2} \cos 2t + \frac{1}{2}
\end{pmatrix}
\]

For \(G\) replaced by \(G_1\), (V.4) reduces to

\[
e''_1 = -e_1 - C_1 X(t)^{(P+2)}
\]

\[
e_1(0) = e_1'(0) = 0 \quad i = 1, 2.
\]

\[
e''_2 = -e_2 - C_2 Y(t)^{(P+2)}
\]
If $P = 2K$, then $X^{(P+2)}(t) = (-1)^{K} \cos t$, $Y^{(P+2)}(t) = (-1)^{K} \sin t$

and we may verify directly that

$$e_{1}(t) = (-1)^{K} \frac{t}{2} C_{1} \sin t$$

$$e_{2}(t) = (-1)^{K+1} C_{2} \frac{t}{2} \cos t + \frac{(-1)^{K+1}}{2} C_{2} \sin t.$$ 

Whereas, if $P = 2K+1$, then

$$X^{(P+2)}(t) = (-1)^{K} \sin t$$

$$Y^{(P+2)}(t) = (-1)^{K+1} \cos t$$

and now

$$e_{1}(t) = \frac{(-1)^{K} C_{1} t \cos t}{2} + \frac{C_{1} (-1)^{K+1} \sin t}{2}$$

$$e_{2}(t) = \frac{t}{2} (-1)^{K} C_{2} \sin t.$$ 

Thus, in both cases the error behaves (roughly) as a linear function of $t$.

For the nonlinear system (V.2) the situation is more complicated. As in the computation following Lemma IIIC, (V.4) can be written in the equivalent form.
\[
\begin{bmatrix}
  u_1 \\
  u_2 \\
  e_1 \\
  e_2
\end{bmatrix} =
\begin{bmatrix}
  0 & I \\
  G & 0
\end{bmatrix}
\begin{bmatrix}
  u_1 \\
  u_2 \\
  e_1 \\
  e_2
\end{bmatrix} -
\begin{bmatrix}
  0 \\
  0 \\
  C_1 x^{(P+2)} \\
  C_2 y^{(P+2)}
\end{bmatrix} +
\begin{bmatrix}
  u_1(0) \\
  u_2(0) \\
  e_1(0) \\
  e_2(0)
\end{bmatrix}
\]

(V.5)

where

\[
G =
\begin{bmatrix}
  3 \cos 2t + \frac{1}{2} & 3 \sin 2t \\
  3 \sin 2t & -3 \cos 2t + \frac{1}{2}
\end{bmatrix}
\]

If the matrix \( Z(t,\tau) = (Z_{ij}(t,\tau)) \) \( i,j = 1, \ldots, 4 \) satisfies

\[
\frac{d}{dt} Z(t,\tau) = \begin{pmatrix} 0 & I \\ G & 0 \end{pmatrix} Z(t,\tau) \quad Z(t,\tau) = I,
\]

then the solution of (V.5) can be written as

\[
\begin{bmatrix}
  u_1 \\
  u_2 \\
  e_1 \\
  e_2
\end{bmatrix} = \int_{0}^{t} Z(t,\tau) \begin{bmatrix}
  0 \\
  0 \\
  -C_1 x^{(P+2)}(\tau) \\
  -C_2 y^{(P+2)}(\tau)
\end{bmatrix} d\tau
\]
or in particular,

\[
\begin{pmatrix}
e_1(t) \\
e_2(t)
\end{pmatrix} = \begin{bmatrix}
\int_0^t (-c_1z_{33}(t,\tau)x(\tau)^{P+2} - c_2z_{34}(t,\tau)y(\tau)^{P+2})d\tau \\
\int_0^t (-c_1z_{43}(t,\tau)x(\tau)^{P+2} - c_2z_{44}(t,\tau)y(\tau)^{P+2})d\tau
\end{bmatrix}
\]

(V.6)

A permutation of the matrix \(Z(t,\tau)\) was calculated in [3], namely, the matrix \(W(t,\tau)\) was determined so that

\[
\frac{d}{dt} W(t,\tau) = GW(t,\tau) \quad W(\tau,\tau) = I
\]

where

\[
G = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
\frac{3}{2} \cos 2t + \frac{1}{2} & 0 & \frac{3}{2} \sin 2t & 0 \\
0 & 0 & 0 & 1 \\
\frac{3}{4} \sin 2t & 0 & -\frac{3}{2} \cos 2t + \frac{1}{2} & 0
\end{bmatrix}
\]

For

\[
T = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]
one finds that \( T^{-1}GT = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \) and thus

\[
Z(t, \tau) = TW(t, \tau).
\]

But now from the already calculated \( W(t, \tau) \), we can take the elements which we need. Thus,

\[
Z_{33} = \frac{1}{4} \left\{ 6 \cos 2t + 2 \sin 2(t-\tau) + 8 \sin(t-\tau) + 12 \sin \tau \cos t \\
- 12(t-\tau) \sin \tau \sin t \right\}
\]

\[
Z_{43} = \frac{1}{4} \left\{ -6 \sin 2t + 2 \cos 2(t-\tau) + 8 \cos(t-\tau) - 12 \sin \tau \sin t \\
- 12(t-\tau) \sin \tau \cos t \right\}
\]

\[
Z_{34} = \frac{1}{4} \left\{ 6 \cos 2t + 2 \cos 2(t-\tau) + 8 \cos(t-\tau) - 12 \cos \tau \cos t \\
+ 12(t-\tau) \cos \tau \sin t \right\}
\]

\[
Z_{44} = \frac{1}{4} \left\{ -6 \sin 2t - 2 \sin 2(t-\tau) - 8 \sin(t-\tau) + 12 \cos \tau \sin t \\
+ 12(t-\tau) \cos \tau \sin t \right\}.
\]

If \( P = 2K \), then again \( X^{(P+2)}(t) = (-1)^{K+1} \cos t \); \( Y^{(P+2)}(t) = (-1)^{K+1} \sin t \)
and (V.6) yields

\[
c_1(t) = 0(t)
\]

\[
c_2(t) = 0(t).
\]
If \( P = 2K+1 \), then \( X^{(P+2)}(t) = (-1)^K \sin t \); \( Y^{(P+2)}(t) = (-1)^{(K+1)} \cos t \) and (V.6) yields

\[
e_1(t) = (-1)^{K+1} 6t^2 \sin t + o(t)
\]

\[
e_2(t) = (-1)^{K+1} 6t^2 \cos t + o(t).
\]

Thus, if the order is even, the growth is linear; whereas, if the order is odd, the growth is quadratic. This dichotomy also occurred in [3] for multistep approximations to first-order differential equations.

If the starting error were \( O(h^{P+1}) \), the asymptotic representation (Theorem IID) would also include the term \( h^P U_n \) where \((U_n)_1 = \sum_{j=1}^{d} S_j^1 Z_1^n X_j(t_n) \). If the matrices \( A_1 \), \( i = 0, \ldots, K \) are each scalar multiples of the identity matrix, then \( S_i^j = 1 \) and \( X_j \) is the component of the solution of

\[
X^{''} = \frac{2\sigma(Z_j)}{Z_j \rho''(Z_j)} g(t)X_j(t) \quad X_j(0) = 0 \quad X_j'(0) = b_j^*.
\]

Thus, computation of \( X_j \) would involve specialization to a particular method. It should be noted that for the "classical" methods (Nyström, Cowell [3]) one is the only essential root and \( \frac{\sigma(1)}{\rho''(1)} = 1 \).
In order to apply the results of the statistical approach we must hypothesize the behavior of the mean and the covariance of the local round-off error. If the rounding is by chopping, the appropriate hypothesis on the mean [3] is that

$$E(\varepsilon_n) = \mu X(t_n)$$

where $\mu$ is a suitable scalar. But the mean of the accumulated round-off error is given as the solution of

$$m'' = Gm + \mu X(t) \quad m(0) = m'(0) = 0$$

and we have already made this calculation with the results that both in the linear problem

$$m = 0(t)$$

and in the nonlinear problem

$$m = 0(t).$$

The computation of the covariance matrix given by

$$\text{covar}(r'_j) = \frac{\sigma^2}{h^3} \sum_{j=1}^{d} \frac{1}{|\rho''(z_j)|^2} \int_0^{t_n} \mathcal{P}_j(t_n, t) C(t) \mathcal{P}_j^*(t_n, t) dt$$
where

\[
\frac{d^2}{d\tau^2} F_j(\tau, t) = \mu_n^2 (\tau) F_j(\tau, t)
\]

\[
F_j(\tau, t) = \theta \frac{d}{d\tau} F_j(\tau, t) \bigg|_{\tau=t} = I
\]

is considerably more complicated. Since the covariance matrix depends on the growth parameters, one must commit oneself to a particular method. Thus, rather than make further computations based on Eqs. (V.1) and (V.2) we turn our attention to the following theorems.

Theorems VA and VB which will now be given require that \( \mu^{\prime} G(t) \) be skew-symmetric \( \left[ \mu_n^2 G(t) - \mu_n^2 G^*(t) \right] \). This hypothesis is motivated by the matrix \( G \) associated with (V.1) in which case the matrix \( \begin{bmatrix} 0 & J \\ G & 0 \end{bmatrix} \) is similar to a skew-symmetric matrix.

**LEMMA VA:** If \( \mu' G \) is skew-symmetric and \( F(t, \tau) \) satisfies \( F'(t, \tau) = \mu^2 G F(t, \tau) \left( i = \frac{d}{dt} \right) F(t, \tau) = I \) then

\[
F_j(t, \tau) F^*(t, \tau) = I.
\]

**Proof:** Certainly \( F(t, \tau) F^*(t, \tau) = I \). Also

\[
[F(t, \tau) F^*(t, \tau)] = \mu^2 G(t) F(t, \tau) F^*(t, \tau) + F(t, \tau) F^*(t, \tau) G* \mu^2
\]
But, as may be directly verified, the solution of

\[ Z'(t, \tau) = \mu^2 G + \mu^2 G^* + \mu^2 \bar{G}(t)Z(t, \tau) + Z(t, \tau)G^* \mu^2 \]

\[ Z(\tau, \tau) = 0 \]

may be represented by

\[ Z(t, \tau) = \int_{\tau}^{t} F(t, y)(\mu^2 G(y) + \mu^2 G^*(y))F^*(t, y)dy. \]

Thus, since \( \mu G(t) \) is skew-symmetric,

\[ Z(t, \tau) = 0. \]

But

\[ Z(t, \tau) = F(t, \tau)F^*(t, \tau) - I \]

and the proof is finished.

It is frequently [3],[4] assumed that the local covariance matrix \( C \) is proportional to the unit matrix. We follow this practice in the next theorem although with no additional effort the theorem can be extended to any scalar multiple of the unit matrix.

**THEOREM VB:** If \( \mu^2 G(t) \) is skew-symmetric, and if \( C = \alpha I \), then

\[ \int_{0}^{t} F(t, \tau)CF(t, \tau)d\tau = \alpha tI. \]

The proof follows from an immediate application of Lemma VA.
As a final study on the growth of summands of the covariance matrix, we consider

**THEOREM VC:** If \( \mu^2 G(t) \) is skew-symmetric and constant and if \( C \) is a constant matrix, then there exist constants \( k_1 \) and \( k_2 \) such that

\[
\left| \int_0^t F(t, \tau) CF^*(t, \tau) d\tau \right| \leq k_1 t + k_2
\]

Proof: Since \( \mu^2 G \) is constant, then certainly \([1]\)

\[
F = \exp \mu^2 G(t-\tau).
\]

Since \( \mu^2 G \) is skew-symmetric, there exists a matrix \( U \) (\( UU^* = I \)) such that

\[
\exp[\mu^2 G(t-\tau)] = U \exp[(t-\tau)D]U^*
\]

where \( D = \text{diag } \lambda_k, \lambda_k \text{ real} \). Now

\[
\int_0^t F(t, \tau) CF(t, \tau) d\tau = U \int_0^t \text{diag } e^{i(t-\tau)\lambda_k D} \text{diag } e^{-i(t-\tau)\lambda_k D} d\tau
\]

where

\[
B = U^*CU.
\]

But the integrand on the right has as typical element

\[
b_{k\ell} e^{(\lambda_k - \lambda_\ell)(t-\tau)}
\]
whose integral is given by

\[- \frac{b_{kt}}{(\lambda_k - \lambda_l)} \left[ 1 - e^{i(\lambda_k - \lambda_l)t} \right] \quad \lambda_k - \lambda_l \neq 0\]

or

\[b_{kt} t \quad \lambda_k - \lambda_l = 0.\]

But since \( U \) is a constant matrix, \( \int_0^t F(t,\tau)CF(t,\tau)d\tau \) is a linear combination of terms of the form \( 1, t, e^{iat} \) (a real), and the proof is finished.
BIBLIOGRAPHY


VITAE CURRICULUM

I was born on August 22, 1933, in Corinth, Mississippi, U.S.A. After completing my high school education there, I studied at the University of North Carolina from which I obtained my B.S. (1955) and M.A. (1957). I was then employed by the Atomic Energy Division of Babcock & Wilcox. At this time my principal concern was with neutron flux calculations and reactor kinetics.

After spending the academic years 1960 and 1961 at the Georgia Institute of Technology, I transferred to the University of California (Los Angeles) where I began my studies under Professor P. Henrici to whom I am particularly grateful for his guidance of my work. After two years at the University of California, Professor Henrici most graciously arranged for me to continue my studies at the Department of Mathematics and Physics at the Swiss Federal Institute of Technology. Here, I completed the work on my thesis. I also received some particularly incisive comments from Professor H. Rutishauser to whom I am grateful.


William Perry Timlake