NUMERICAL SOLUTIONS OF INHERENTLY
UNSTABLE ORDINARY DIFFERENTIAL EQUATIONS

By Robert M. Myers

June 1970

Prepared by:

Lockheed Electronics Company
Houston Aerospace Systems Division
Houston, Texas

Under Contract NAS 9-5384

For

COMPUTATION AND ANALYSIS DIVISION

National Aeronautics and Space Administration
Manned Spacecraft Center
Houston, Texas

00019

NASA CR 108537

ABSTRACT

The problem of determining numerical solutions of a differential equation which is inherently unstable is considered. In particular, the asymptotic continuation problem, which is formulated in this report, is analyzed. The development of a definition of error and an error analysis proceed naturally for this problem.

Three algorithms are presented and each is applied to a particular example. In each case the algorithms are shown to be vastly superior to a standard numerical integration. In addition, the algorithms are shown to provide reasonable estimates of accumulated errors. The most surprising theoretical result is that a single initial relative error remains uniformly bounded on an unbounded interval even though different solutions may exhibit different exponential rates of growth.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

In a recent paper by Conte (ref. 1) an algorithm is developed for the numerical solution of inherently unstable linear ordinary differential equations. The algorithm is specifically intended for linear boundary value problems in which the underlying ordinary differential equation has solutions which grow at vastly different rates. For such problems it is expected that the shooting or initial value method (ref. 2, pp. 58-63; or ref. 3, pp. 39-71) will yield inaccurate results. However, Conte showed by example that rather simple procedures can effectively control the error buildup.

A method similar to Conte's is presented in this paper. Also, a theoretical analysis of the error is conducted and two other algorithms are included. The algorithms presented in this paper are intended for the numerical solution of the asymptotic continuation problem, formulated in section 2. This problem has certain mathematical niceties which allow the analysis to develop naturally.

The main topics covered in this paper are:

- Formulation of the asymptotic continuation problem (section 2)
- Definition of error, and analysis of the propagation of an initial error (section 3)
- Development of three algorithms to solve the asymptotic continuation problem, including estimates of accumulated errors (section 3)

The most surprising result is that, for a *reasonable* definition of error, an initial error remains uniformly bounded on an unbounded interval, despite the fact that the underlying differential equation is classified as inherently unstable. The only result concerning accumulated errors is presented as the final step of the algorithms. For several test problems this procedure leads to a fairly good estimate of the accumulated relative error.

The author has used the procedures outlined in this paper to analyze a mathematical model of small oscillations of a viscous atmosphere. It was possible to perform the calculations in single precision and thereby reduce computational time.

2. FORMULATION OF THE ASYMPTOTIC CONTINUATION PROBLEM

In this section a special problem is considered. This problem is the basis of a better understanding of error propagation of numerical solutions of inherently unstable ordinary differential equations.

Consider

$$\frac{d\vec{y}(x)}{dx} = A(x)\vec{y}(x) \tag{1}$$

where $\vec{y}(x)$ is an n-dimensional column vector and $A(x)$ is an $n \times n$ matrix. Suppose $A(x)$ has elements which are bounded analytic functions for $|x| > R$ in the complex x-plane.

Then

$$A(x) = \sum_{k=0}^{\infty} A_k x^{-k} \tag{2}$$

where the infinite sum converges for $|x| > R$. If A_0 has distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then there exist formal asymptotic solutions of equation (1) of the form

$$\hat{\vec{y}}_j(x) = \left(\sum_{k=0}^{\infty} \vec{a}_{kj} x^{-k}\right) x^{\alpha_j} e^{\lambda_j x} \tag{3}$$

for $j = 1,2,\ldots,n$, where $\vec{a}_{0j} \neq \vec{0}$.

For the typical case where the infinite vector sum in equation (3) diverges, it is not possible to compute actual solutions of (1) to arbitrary precision by the use of (3) alone. However, actual solutions of equation (1) do exist which can be approximated by formal truncated solutions (ref. 4, ch. 5; or ref. 5). More precisely, there exist actual solutions $\vec{y}_j(x)$ of (1) such that

$$\left\| \vec{y}_j(x) - \left(\sum_{k=0}^{N} \vec{a}_{kj} x^{-k}\right) x^{\alpha_j} e^{\lambda_j x} \right\| = O\left(\frac{1}{x^{N+1}}\right) \left| x^{\alpha_j} e^{\lambda_j x} \right| \tag{4}$$

as $x \to \infty$ in S, for $N = 0,1,2,\ldots$, $j = 1,2,\ldots,n$, and S is some sector of the complex plane. In general, the sectors are determined by ordering $\mathrm{Re}(\lambda_1 x)$, $\mathrm{Re}(\lambda_2 x)$, $\ldots$, $\mathrm{Re}(\lambda_n x)$. The boundaries of the sectors or Stokes lines are determined by solving for the rays along which

$\mathrm{Re}(\lambda_j x) = \mathrm{Re}(\lambda_k x)$, for $j \neq k$. In moving from one sector to another it is likely that an actual solution of the differential equation will suddenly have a different asymptotic representation, the so-called Stokes phenomenon.

Since the formal solutions are capable of only limited accuracy at prescribed values of x , it is necessary to supplement in some way the calculation of $\vec{y}_j(x)$. Basically, the asymptotic continuation problem involves the accurate continuation of the asymptotic solutions to prescribed values of x .

<u>Asymptotic Continuation Problem</u>: Interior to the sector S , in the complex x-plane, determine the vectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$, such that the vectors $\vec{y}_j(x)$ in equation (4), for $j = 1,2,\ldots,n$, are prescribed by

$$\vec{y}_j(x) = \vec{v}_j \tag{5}$$

at $x = a$, and for all other x , $\vec{y}_j(x)$ is a solution of (1).

To guarantee the existence of a solution to the asymptotic continuation problem, it is assumed that $A(x)$ in equation (1) can be analytically continued to the sector S , bounded by $|a| \leq |x| < \infty$ and two Stokes lines. No Stokes lines are interior to S . In addition, it is assumed that $A(x)$ has no singularities in S for $|x| \geq a$, and $x = a$ is an interior point of the region of analyticity of $A(x)$ and an interior point of S. Without further loss in generality, assume that the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are numbered such that

$$\operatorname{Re}(\lambda_1 x) < \operatorname{Re}(\lambda_2 x) < \ldots < \operatorname{Re}(\lambda_n x) \quad , \quad x \varepsilon S \tag{6}$$

The asymptotic continuation problem will be considered solved if the vectors $\vec{v}_j$ in equation (5) can be determined to arbitrary precision. As a practical requirement, an algorithm is sought which is capable of overcoming the inherent instability of numerically integrating equation (1). In addition, the algorithm must not demand excessive precision or computational time. This additional requirement is difficult if not impossible to quantify; nevertheless, it will be considered an integral part of the formulation of the asymptotic continuation problem.

3. NUMERICAL ANALYSIS OF THE ASYMPTOTIC CONTINUATION PROBLEM

In this section it is shown that if it is possible to compute solutions of equation (1) with an initial error, then the propagated error can be uniformly bounded if the theoretical integration of (1) proceeds in the proper direction (decreasing x). This is shown despite the fact that solutions of (1) exhibit differential exponential rates of growth.

For a simple model of a computational device, the inherent instability of equation (1) is shown to result in a delayed exponential error growth. Since the error growth is delayed, rather simple procedures can effectively control it; for example, Conte's orthogonalization procedure (ref. 1). In addition, two other algorithms are considered.

The asymptotic continuation problem has infinitely many solutions.

THEOREM 1: If the vectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ prescribe the solutions of equation (1) at $x = a$ which are respectively asymptotic to $\hat{\vec{y}}_1(x), \hat{\vec{y}}_2(x), \ldots, \hat{\vec{y}}_n(x)$ in sector S, then $\vec{V}_1, \vec{V}_2, \ldots, \vec{V}_n$ also prescribe solutions of (1) at $x = a$ which are respectively asymptotic to $\hat{\vec{y}}_1(x)$, $\hat{\vec{y}}_2(x), \ldots, \hat{\vec{y}}_n(x)$, if and only if

$$\vec{V}_1 = \vec{v}_1 \tag{7a}$$

$$\vec{V}_j = \vec{v}_j + \sum_{k=1}^{j-1} C_{kj}\vec{v}_k \tag{7b}$$

for $j = 2,3,\ldots,n$.

Proof: Let $\vec{y}_j(x)$ and $\vec{Y}_j(x)$ be vector solutions of equation (1) which are defined in S and are asymptotic to $\hat{\vec{y}}_j(x)$ for $j = 1,2,\ldots,n$. Due to the different asymptotic rates of growth of $\vec{y}_1(x), \vec{y}_2(x), \ldots, \vec{y}_n(x)$, it follows that these n vectors form a fundamental set of solutions. Similarly, $\vec{Y}_1(x), \vec{Y}_2(x), \ldots, \vec{Y}_n(x)$ form a fundamental set of solutions. Hence, there exist constants C_{kj} such that

$$\vec{Y}_j(x) = \sum_{k=1}^{n} C_{kj}\vec{y}_k(x) \tag{8}$$

for all x in S and $|a| \leq |x| < \infty$. If $\vec{y}_j(x)$ is asymptotic to $\hat{\vec{y}}_j(x)$, then $\vec{Y}_j(x)$ is also asymptotic to $\hat{\vec{y}}_j(x)$, if and only if

$$\left|\left|\vec{Y}_j(x) - \vec{y}_j(x)\right|\right| = O\left(\frac{1}{x^N}\right)\left|x^{\alpha_j}e^{\lambda_j x}\right| \tag{9}$$

in sector S for $N = 0,1,2,\ldots,$ [1]. This follows from relation (4).

Recall that $Re(\lambda_k x) < Re(\lambda_j x)$ in S if and only if $k < j$. Hence, relation (9) can be satisfied if and only if the constants C_{kj} in (8) are such that

$$C_{kj} = \begin{Bmatrix} 0 & \text{if } k > j \\ 1 & \text{if } k = j \\ \text{arbitrary} & \text{if } k < j \end{Bmatrix} \qquad \text{Q.E.D.}$$

Definition 1: Denote by $\{\vec{y}_j(x)\}$ the family of all solutions of equation (1) which exist in S for $|a| \leq x < \infty$ and such that each member of the family is asymptotic to $\hat{\vec{y}}_j(x)$.

If the vectors in $\{\vec{y}_j(x)\}$ are considered position vectors in an n-dimensional vector space, then for a

---

[1]Unless otherwise stated, the Euclidean norm will be used.

prescribed value of x the family $\{\vec{y}_j(x)\}$ determines a j-1 dimensional hyperplane.

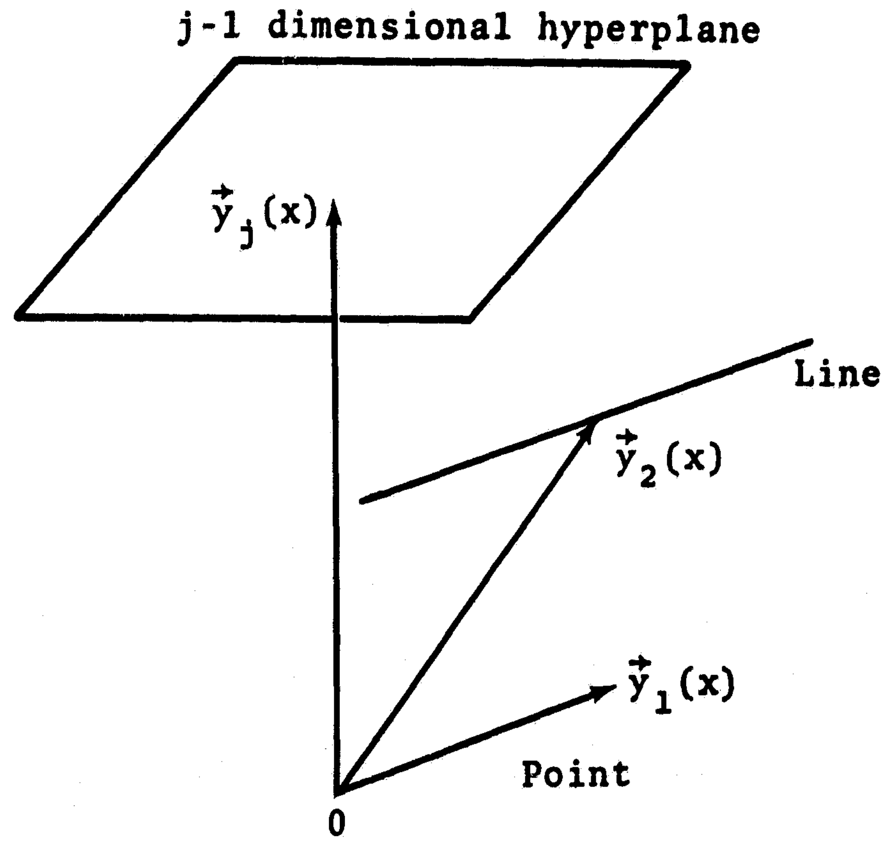

Figure 1. – Representation of the asymptotic families of solutions.

The asymptotic continuation problem is solved if it is possible to determine some member of each of the families to arbitrary precision; that is, with arbitrarily small error. However, the families for $j > 1$ consist of infinitely many vector solutions of (1), and, hence, it is not immediately obvious how errors should be defined. Theorem 1 will be the basis of a reasonable definition of error.

If $\vec{y}_j(x)$ is a member of $\{\vec{y}_j(x)\}$, and $\vec{\varepsilon}_{0j}$ is some vector, then consider

$$\vec{y}_{je}(x_0;x_0) = \vec{y}_j(x_0) + \vec{\varepsilon}_{0j} \tag{10}$$

where x_0 is interior to sector S, $\vec{y}_{je}(x;x_0)$ is a solution of equation (1) for all other x in S, and $|a| \leq |x| \leq |x_0| < \infty$. Clearly, if $j > 1$ and $\vec{\varepsilon}_{0j}$ is a linear combination of members of $\{\vec{y}_{j-1}(x)\}$, ..., $\{\vec{y}_1(x)\}$ at $x = x_0$, then $\vec{y}_{je}(x;x_0)$ would be a member of $\{\vec{y}_j(x)\}$. For such a case it would be reasonable to say that $\vec{y}_{je}(x;x_0)$ defined in equation (10) has no error, even though $\vec{\varepsilon}_{0j}$ may represent in some way an inability to calculate precisely a particular member of $\{\vec{y}_j(x)\}$ at $x = x_0$. If $\vec{\varepsilon}_{0j}$ in (10) can not be represented as some linear combination of members of $\{\vec{y}_{j-1}(x)\}$, ..., $\{\vec{y}_1(x)\}$, then $\vec{y}_{je}(x;x_0)$ will have a nonzero error.

For the asymptotic continuation problem, any member of $\{\vec{y}_j(x)\}$ which can be determined at $x = a$ will be an acceptable solution. A vector $\vec{y}_{je}(x;x_0)$ is a useful approximation of some member of this j-1 dimensional hyperplane if the distance of $\vec{y}_{je}(x;x_0)$ to the hyperplane is small.

Definition 2: The vector solution of equation (1), defined in relation (10) and denoted by $\vec{y}_{je}(x;x_0)$, is said to have the absolute error $e_j(x;x_0)$ with respect to the family $\{\vec{y}_j(x)\}$, where $e_j(x;x_0)$ is the Euclidean distance of $\vec{y}_{je}(x;x_0)$ from the hyperplane defined by $\{\vec{y}_j(x)\}$ for each value of x.

Due to the different rates of exponential growth of solutions of equation (1), it will be more useful to deal with relative errors rather than with absolute errors.

Definition 3: The vector solution of equation (1), defined in relation (10) and denoted by $\vec{y}_{je}(x;x_0)$, is said to have the relative error $r_j(x;x_0)$ with respect to the family $\{\vec{y}_j(x)\}$, where

$$r_j(x;x_0) = \frac{e_j(x;x_0)}{||\vec{y}_{jmin}(x)||}$$

The vector $\vec{y}_{jmin}(x)$ is the Euclidean normal to the hyperplane $\{\vec{y}_j(x)\}$, from the origin.

It should be noted that

$$||\vec{y}_{jmin}(x)|| \leq ||\vec{y}_j(x)||$$

for all $\vec{y}_j(x)\varepsilon\{\vec{y}_j(x)\}$. In addition, $\vec{y}_{jmin}(x)$ is not generally a member of $\{\vec{y}_j(x)\}$ for $j \geq 2$; that is, a single solution of equation (1) is not expected to be normal to the hyperplane for all x . However, for each value of x , there will correspond a unique member of $\{\vec{y}_j(x)\}$ such that this member will equal $y_{jmin}(x)$. As x varies, infinitely many different members of $\{\vec{y}_j(x)\}$ will generally correspond. Only for $j = 1$ is $\vec{y}_{jmin}(x)$ always a member of $\{\vec{y}_j(x)\}$, since this family has only one member.

$$\vec{y}_{1min}(x) = \vec{y}_1(x) \tag{11a}$$

$$\vec{y}_{2min}(x) = \vec{y}_2(x) - \left(\vec{y}_2(x), \vec{y}_1(x)\right) \frac{\vec{y}_1(x)}{||\vec{y}_1(x)||^2} \tag{11b}$$

$$\vec{y}_{jmin}(x) = \vec{y}_j(x) - \sum_{k=1}^{j-1} \left(\vec{y}_j(x), \vec{y}_{kmin}(x)\right) \frac{\vec{y}_{kmin}(x)}{||\vec{y}_{kmin}(x)||^2} \tag{11c}$$

where

$$(\vec{\mu}, \vec{v}) = \sum_{i=1}^{n} \mu_i \bar{v}_i$$

μ_i and v_i are the *ith* components of $\vec{\mu}$ and $\vec{v}$, respectively, and $\bar{v}_i$ is the conjugate of v_i.

Suppose $\vec{\varepsilon}_{0j}$ in relation (10) is expanded in terms of $\vec{y}_1(x_0)$, $\vec{y}_2(x_0)$, ..., $\vec{y}_n(x_0)$, where $\vec{y}_j(x)$ is a particular member of $\{\vec{y}_j(x)\}$; that is,

$$\vec{\varepsilon}_{0j} = \sum_{k=1}^{n} \varepsilon_{kj} \vec{y}_k(x_0) \tag{12a}$$

or

$$\vec{\varepsilon}_{0j} = \sum_{k=1}^{j-1} \varepsilon_{kj} \vec{y}_k(x_0) + \varepsilon_{jj} \vec{y}_j(x_0) + \sum_{k=j+1}^{n} \varepsilon_{kj} \vec{y}_k(x_0) \tag{12b}$$

Due to theorem 1 and the subsequent definition of $e_j(x;x_0)$, it follows that

$$e_j(x;x_0) \leq |\varepsilon_{jj}| \; ||\vec{y}_j(x)|| + \sum_{k=j+1}^{n} |\varepsilon_{kj}| \; ||\vec{y}_k(x)||$$

and

$$r_j(x;x_0) \leq |\varepsilon_{jj}| \frac{||\vec{y}_j(x)||}{||\vec{y}_{jmin}(x)||} + \sum_{k=j+1}^{n} \varepsilon_{kj} \frac{||\vec{y}_k(x)||}{||\vec{y}_{jmin}(x)||} \tag{13}$$

THEOREM 2: If $\hat{\vec{y}}_{j,N}(x)$ is the truncated formal asymptotic solution of equation (1) defined by

$$\hat{\vec{y}}_{j,N}(x) = \left(\sum_{k=0}^{N} \vec{a}_{kj} x^{-k}\right) x^{\alpha_j} e^{\lambda_j x}$$

and $x = x_0$ and $x = a$ are on the same ray $\arg x = \theta$ which is interior to sector S, then

$$r_j(a;x_0) = O\left(\frac{1}{x_0^{N+1}}\right)$$

if

$$\vec{y}_{je}(x_0;x_0) = \hat{\vec{y}}_{j,N}(x_0)$$

Note: Theorem 2 is established by showing that an initial error can only grow by a bounded factor even though solutions of equation (1) grow at different exponential rates and x_0 can be chosen arbitrarily large. Ordinarily it would be expected that an initial error would experience an exponential growth. Hence, it might even be anticipated that the truncated asymptotic solutions would be inadequate to determine members of $\{\vec{y}_j(x)\}$ for all j, since the initial relative error decreases only algebraically fast as $x_0 \to \infty$ and may be magnified exponentially fast on the ray connecting $x = x_0$ to $x = a$. However, theorem 2 precludes this possibility.

Proof (Sketch Only): Relation (4) implies that

$$r_j(x_0;x_0) = O\left(\frac{1}{x_0^{N+1}}\right)$$

if

$$\vec{y}_{je}(x_0;x_0) = \hat{\vec{y}}_{j,N}(x_0)$$

since, for any particular member of $\{\vec{y}_j(x)\}$, it is easily shown that

$$\frac{||\vec{y}_j(x)||}{||\vec{y}_{j\min}(x)||} \leq M \tag{14}$$

for $|a| \leq |x| < \infty$. For different members of $\{\vec{y}_j(x)\}$, the bound M in (14) will vary, and no uniform bound exists for $j \geq 2$.

For the ε_{kj} defined in relations (12) and (13) it can be shown that

$$|\varepsilon_{kj}| \frac{||\vec{y}_k(x_0)||}{||\vec{y}_{jmin}(x_0)||} = O\left(\frac{1}{x_0^{N+1}}\right) \tag{15}$$

In order to complete the proof of theorem 2 it is necessary only to show that an initial relative error can increase by, at most, a bounded factor which is independent of x_0 . Every solution of (1) can grow by only a finite factor if the independent variable x is restricted to a compact subset of S and $|a| \leq |x|$. This follows since the assumptions on (1) in section 2 eliminate the possibility of a singularity for finite x such that $|x| \geq |a|$ and $x \varepsilon S$.

For a particular member $\vec{y}_j(x)$ of $\{\vec{y}_j(x)\}$, the truncated formal solution $\hat{\vec{y}}_{j,0}(x)$ will approximate $\vec{y}_j(x)$ for sufficiently large x and $j = 1,2,\ldots,n$. Consequently, for a prescribed $\varepsilon > 0$ there will correspond a sufficiently large $x_1(\varepsilon)$ such that

$$\frac{||\vec{y}_k(x)||}{||\vec{y}_j(x)||} \leq (1 + \varepsilon) \times \frac{||\vec{y}_k(x_0)||}{||\vec{y}_j(x_0)||}$$

for $|x_1(\varepsilon)| \leq |x| \leq |x_0| < \infty$ and $j < k$. This seems reasonable since $\mathrm{Re}(\lambda_j x) < \mathrm{Re}(\lambda_k x)$ and, hence, $\vec{y}_k(x)$ grows exponentially faster than $\vec{y}_j(x)$. For $|a| \leq |x| \leq |x_1(\varepsilon)|$ it is possible to determine a constant L such that

$$\frac{||\vec{y}_k(x)||}{||\vec{y}_j(x)||} \leq L \times \frac{||\vec{y}_k\left(x_1(\varepsilon)\right)||}{||\vec{y}_j\left(x_1(\varepsilon)\right)||}$$

Hence,

$$\frac{||\vec{y}_k(x)||}{||\vec{y}_j(x)||} \leq (1 + \varepsilon)L \times \frac{||\vec{y}_k(x_0)||}{||\vec{y}_j(x_0)||} \tag{16}$$

for $|a| \leq |x| \leq |x_0| < \infty$ and $j < k$. Relations (13), (14), (15), and (16) establish theorem 2. Q.E.D.

Theorem 2 implies that the asymptotic continuation problem is not inherently unstable, since initial errors can grow by only a bounded factor. However, theorem 2 requires an exact determination of solutions of (1). Since it is not generally possible to determine an exact solution of a differential equation, it is necessary to assess error buildup on a reasonable model of a computational device.

Consider a hypothetical computing device which is capable of determining solutions of equation (1) to a finite number of significant figures. Instead of determining $\vec{y}_{je}(x;x_0)$ defined in (10), it is possible to determine $\vec{Y}_{je}(x;x_0)$, where

$$\frac{||\vec{Y}_{je}(x;x_0) - \vec{y}_{je}(x;x_0)||}{||\vec{y}_{je}(x;x_0)||} < 10^{-N} \tag{17}$$

The vector $\vec{Y}_{je}(x;x_0)$ may be a useless approximation of $\vec{y}_{je}(x;x_0)$, even though the leading N significant figures of each component of $\vec{Y}_{je}(x;x_0)$ may agree with the leading N digits of the corresponding component of $\vec{y}_{je}(x;x_0)$. The reason is that all N digits of each component of $\vec{Y}_{je}(x;x_0)$ may merely approximate the useless multiples $\varepsilon_{1j}\vec{y}_1(x), \ldots, \varepsilon_{(j-1)j}\vec{y}_{j-1}(x)$ which were initiated at x_0 (see relations (12) and (13)).

Instead of considering $r_j(x;x_0)$, it is necessary to consider $R_j(x;x_0)$, where

$$R_j(x;x_0) = \frac{\min\limits_{\vec{y}\varepsilon\{\vec{y}_j\}} ||\vec{Y}_{je}(x;x_0) - \vec{y}(x)||}{||\vec{y}_{jmin}(x)||} \tag{18a}$$

or

$$R_j(x;x_0) \leq 10^{-N} \frac{||\vec{y}_{je}(x;x_0)||}{||\vec{y}_{jmin}(x)||} + r_j(x;x_0) \tag{18b}$$

Notice that if infinitely many figures can be maintained in the computation of $\vec{y}_{je}(x;x_0)$, then $R_j(x;x_0)$ reduces to $r_j(x;x_0)$. The multiples of $\vec{y}_1(x)$, $\vec{y}_2(x)$, ...,

$\vec{y}_{j-1}(x)$ initiated at $x = x_0$ are neglected in $r_j(x;x_0)$ (see relation (13)); however, $\varepsilon_{kj}\vec{y}_k(x)$ defined in (12) can greatly influence $||\vec{y}_{je}(x;x_0)||/||\vec{y}_{jmin}(x)||$. If $|x| << |x_0|$ and $k < j$, then it is quite possible that $|\varepsilon_{kj}|\ ||\vec{y}_k(x)||$ may greatly exceed $||\vec{y}_{jmin}(x)||$. Roughly speaking, inherent exponential error growth begins when $||\varepsilon_{kj}\vec{y}_k(x)||$ approximately equals $||\vec{y}_{jmin}(x)||$ for $k < j$. A further decrease in $|x|$ causes $||\vec{y}_{je}(x;x_0)||/||\vec{y}_{jmin}(x)||$ to experience an exponential growth (see figure 2, page 29). In order to combat inherent error growth, it is desirable to bound the multiples of $\vec{y}_k(x)$ present in $\vec{y}_{je}(x;x_0)$; that is, to append some process to the numerical integration of (1) which bounds $||\vec{y}_{je}(x;x_0)||/||\vec{y}_{jmin}(x)||$. For example, Conte's orthogonalization procedure can maintain $||\vec{y}_{je}(x;x_0)||/||\vec{y}_{jmin}(x)||$ within reasonable bounds.

So far, theorem 1 has been exploited to yield a reasonable definition of error which subsequently resulted in theorem 2 (uniformly bounded relative errors despite different exponential rates of growth). For a simple model of a computational device, it was shown that relation (13) had to be modified, and the net result was a phenomenon which will be called delayed inherent error growth. Now theorem 1 will be exploited again to eliminate delayed inherent error growth on our hypothetical computational device.

Suppose, to some desired precision, it has been possible to determine $\vec{y}_1(x)$, $\vec{y}_{2min}(x)$, ..., $\vec{y}_{(j-1)min}(x)$. Is it possible to modify the calculation of $\vec{Y}_{je}(x;x_0)$ so that whenever $\vec{Y}_{je}(x;x_0)$ is an N-digit approximation of

$\vec{y}_{je}(x;x_0)$ it is likely to be a useful approximation of some member of $\{\vec{y}_j(x)\}$?

THEOREM 3: If $\vec{y}_{jemin}(x;x_0)$ is defined by

$$\vec{y}_{jemin}(x;x_0) = \vec{y}_{je}(x;x_0) - \sum_{k=1}^{j-1} \left(\vec{y}_{je}(x;x_0), \vec{y}_{kmin}(x)\right) \times \frac{\vec{y}_{kmin}(x)}{||\vec{y}_{kmin}(x)||^2}$$

and $\vec{Y}_{je}(x;x_0)$ satisfies

$$||\vec{Y}_{je}(x;x_0) - \vec{y}_{jemin}(x;x_0)|| \leq 10^{-N}||\vec{y}_{jemin}(x;x_0)||$$

then the relative error $R_j(x;x_0)$ associated with $\vec{Y}_{je}(x;x_0)$ satisfies

$$R_j(x;x_0) \leq 10^{-N} + (1 + 10^{-N})r_j(x;x_0)$$

where

$$R_j(x;x_0) = \frac{\min\limits_{\vec{y}_j \varepsilon \{\vec{y}_j\}} ||\vec{Y}_{je}(x;x_0) - \vec{y}_j(x)||}{||\vec{y}_{jmin}(x)||}$$

and

$$r_j(x;x_0) = \frac{\min\limits_{\vec{y}_j \varepsilon \{\vec{y}_j\}} ||\vec{y}_{je}(x;x_0) - \vec{y}_j(x)||}{||\vec{y}_{jmin}(x)||}$$

Proof:

$$\min_{\vec{y}_j \varepsilon \{\vec{y}_j\}} ||\vec{Y}_{je}(x;x_0) - \vec{y}_j(x)|| \leq ||\vec{Y}_{je}(x;x_0) - \vec{y}_{jemin}(x;x_0)||$$

$$+ ||\vec{y}_{jemin}(x;x_0) - \vec{y}_{jmin}(x)||$$

Thus,

$$R_j(x;x_0) \leq \frac{||\vec{Y}_{je}(x;x_0) - \vec{y}_{jemin}(x;x_0)||}{||\vec{y}_{jemin}(x;x_0)||} \frac{||\vec{y}_{jemin}(x;x_0)||}{||\vec{y}_{jmin}(x)||}$$

$$+ \frac{||\vec{y}_{jemin}(x;x_0) - \vec{y}_{jmin}(x)||}{||\vec{y}_{jmin}(x)||} \tag{19}$$

Since $\vec{y}_{jemin}(x;x_0)$ and $\vec{y}_{jmin}(x)$ are both orthogonal to the hyperplane defined by $\{\vec{y}_j(x)\}$, it follows that the second term on the right side of (19) is $r_j(x;x_0)$. In addition, $||\vec{y}_{jemin}(x;x_0)||/||\vec{y}_{jmin}(x)||$ can be bounded by $1+r_j(x;x_0)$. The hypothesis on $\vec{Y}_{je}(x;x_0)$,

$$||\vec{Y}_{je}(x;x_0) - \vec{y}_{jemin}(x;x_0)||/||\vec{y}_{jemin}(x;x_0)|| \leq 10^{-N}$$

implies that

$$R_j(x;x_0) \leq 10^{-N}\left(1 + r_j(x;x_0)\right) + r_j(x;x_0) \qquad \text{Q.E.D.}$$

Notice in the statement and proof of theorem 3 that the errors in $\vec{y}_{1e}(x;x_0)$, $\vec{y}_{2emin}(x;x_0)$, ..., $\vec{y}_{(j-1)emin}(x;x_0)$ were neglected; that is $\vec{y}_1(x)$, $\vec{y}_{2min}(x)$, ..., $y_{(j-1)min}(x)$ were considered to be known. If the errors in the first j-1 solutions are truly negligible, then it seems reasonable to attempt to determine $\vec{y}_{jemin}(x;x_0)$ instead of $\vec{y}_{je}(x;x_0)$ if only finitely many digits are avialable for the computation of solutions of (1). Of course, the goal is to bound the multiples of $\vec{y}_k(x)$ present in $\vec{y}_{je}(x;x_0)$, and in order to accomplish this, it is not necessary to continuously reduce $\vec{y}_{je}(x;x_0)$ approximately to $\vec{y}_{jemin}(x;x_0)$ since $\vec{y}_{jemin}(x;x_0)$ varies continuously with x . It is necessary only to reduce $\vec{y}_{je}(x;x_0)$ approximately to $\vec{y}_{jemin}(x;x_0)$ several times over the portion of the ray $\arg x = \theta$ which passes through $x = a$ and $x = x_0$. This recursive procedure of determining $\vec{y}_{jemin}(x;x_0)$ when $\vec{y}_{1e}(x;x_0)$, ..., $\vec{y}_{(j-1)emin}(x;x_0)$ are known must obviously begin with an accurate determination of $\vec{y}_{1e}(x;x_0)$. Since $\vec{y}_1(x)$ has dominant exponential growth for x large and $|x|$ decreasing, there should be little difficulty in determining an accurate approximation of $\vec{y}_{1e}(x;x_0)$; that is, the solution with dominant growth should be easily computed.

Algorithm 1 (Similar to Conte's, Reference 1)

(a) Divide the interval $|a| \leq |x| \leq |x_0|$ into L equal subintervals. Choose L sufficiently large so that the asymptotic estimates predict that $||\vec{y}_1(x)||/||\vec{y}_n(x)||$ will increase by less than a factor of 10 in each subinterval as x is decreased.

(b) Numerically integrate equation (1) from x_0 to $x_0 - \frac{(x_0 - a)}{L}$, starting with the initial condition $\hat{\vec{y}}_{j,N}(x_0)$ for $j = 1,2,\ldots,N$.

(c) At $x = x_0 - \frac{(x_0 - a)}{L}$, determine approximate values of $\vec{y}_{jemin}(x;x_0)$ from $\vec{Y}_{je}(x;x_0)$, where $\vec{Y}_{je}(x;x_0)$ is the result of numerical integration of equation (1). That is, approximate $\vec{y}_{1e}(x;x_0)$ by $\vec{Y}_{1e}(x;x_0)$ and approximate $\vec{y}_{jemin}(x;x_0)$ by

$$\vec{Y}_{jemin}(x;x_0) = \vec{Y}_{je}(x;x_0) - \sum_{k=1}^{j-1} \left(\vec{Y}_{je}(x;x_0),\vec{Y}_{kemin}(x;x_0)\right) \times \frac{\vec{Y}_{kemin}(x;x_0)}{||\vec{Y}_{kemin}(x;x_0)||^2}$$

(d) Numerically integrate equation (1) from $x_0 - \ell\left(\frac{x_0 - a}{L}\right)$ to $x_0 - (\ell + 1)\left(\frac{x_0 - a}{L}\right)$, starting with the initial vector which approximates $\vec{y}_{jemin}(x;x_0)$ at $x = x_0 - \ell\left(\frac{x_0 - a}{L}\right)$. At the end of the subinterval, once again determine approximate values of $\vec{y}_{jemin}(x;x_0)$, which become initial vectors for the next subinterval. Repeat step d for $\ell = 1,2,\ldots,L-1$.

(e) In order to check the calculations, numerically integrate equation (1) from $x = a$ to $x = x_0$, with the initial vector being the approximate value which has been obtained for $\vec{y}_{nemin}(x;x_0)$ at $x = a$. At $x = x_0$, reduce the result of the numerical integration of (1) to an approximation of $\vec{y}_{nmin}(x_0)$ by adding the proper multiples of the initial vectors $\hat{\vec{y}}_{1,N}(x_0)$, ..., $\hat{\vec{y}}_{(n-1)N}(x_0)$ which approximate actual solutions $\vec{y}_1(x)$, ..., $\vec{y}_{n-1}(x)$ at $x = x_0$. Also reduce the initial approximation of $\vec{y}_n(x_0)$, namely $\hat{\vec{y}}_{n,N}(x_0)$, to an approximation of $\vec{y}_{nmin}(x_0)$. The difference of the two approximations of $\vec{y}_{nmin}(x_0)$ is then an estimate of accumulated errors.

(e') If $|x_0| >> |a|$, then compare the result of numerically integrating equation (1), with the approximation $\vec{Y}_{nemin}(a;x_0)$ as the initial vector, from $x = a$ to $x = x_0$ with $\hat{\vec{y}}_{n,N}(x_0)$. If $|x_0| >> |a|$, then the multiples of $\vec{y}_1(x)$, ..., $\vec{y}_{n-1}(x)$ present in the approximation of $\vec{y}_n(x)$ should be negligible at $x = x_0$.

Note: Step e' provides a check for large errors which might be present in the initial vectors $\vec{y}_{j,N}(x_0)$ even if these errors occur through blunders; that is, if the asymptotic solutions are improperly determined. Surprisingly, step e is not a suitable check for blunders. If $\vec{y}_{je}(x;x_0)$ can be determined exactly, then it can be shown that step e will result in a zero estimate of error regardless of the initial error present at $x = x_0$ as long as the initial vectors $\vec{y}_{1e}(x_0;x_0)$, $\vec{y}_{2e}(x_0;x_0)$, ..., $\vec{y}_{ne}(x_0;x_0)$ are linearly independent. However, if the initial error present in $\vec{y}_{je}(x_0;x_0)$ is less than or comparable with the additional errors introduced via numerical integration of equation (1), then step e provides an estimate of the maximum relative error $\max\limits_{1 \le j \le n} R_j(a;x_0)$ in the numerical solutions of (1) at $x = a$. It should not be inferred that step e' is a foolproof check.

The basis for steps e and e' is that $y_n(x)$ is exponentially dominant for $|x|$ increasing. Hence, there should be little difficulty in integrating equation (1) from $x = a$ to $x = x_0$ since only an approximation of the dominant solution is required. In addition, the approximations of $\vec{y}_1(x), \ldots, \vec{y}_{n-1}(x)$ are used to stabilize the calculations of $\vec{y}_n(x)$ for numerical integration of (1) in the direction of decreasing $|x|$, and, hence, inaccuracies in the first n-1 solutions should influence the numerical approximation of $\vec{y}_n(x)$.

Algorithm 2 (Modification of Conte's)

Proceed as in algorithm 1 but define the inner product of vectors by requiring $\vec{a}_{0j}$ for $j = 1,2,\ldots,n$ (see relation (4)) to be an orthonormal basis of the n-dimensional vector space which contains the solutions of equation (1). Theorems 1, 2, and 3 are valid for this new norm. However, in reducing $\vec{y}_{je}(x;x_0)$ to $\vec{y}_{jemin}(x;x_0)$, the vector $\vec{y}_{jemin}(x;x_0)$ will not differ vastly from the truncated asymptotic expansion $\vec{a}_{0j}x^{\alpha_j}e^{\lambda_j x}$ for large x.

Before developing algorithm 3, some preliminary analysis must be performed. Recall that $\{\vec{y}_1(x)\}$ consisted of a unique member $\vec{y}_1(x)$. Thus, at each value of x there exists a component of $\vec{y}_1(x)$ which is greatest in modulus. This component will be called the maximum component of $\vec{y}_1(x)$. If more than one component, at a specified value of x, achieved the maximum modulus, then the component with lowest index will be called the maximum component. Hence, for each value of x there corresponds a unique maximum component of $\vec{y}_1(x)$.

Consider

$$\vec{y}_{2c}(x|x_1) = \vec{y}_2(x) + c_{21}\vec{y}_1(x) \tag{20}$$

where c_{21} is chosen in equation (20) so that $\vec{y}_{2c}(x_1|x_1)$ has a zero component corresponding to the maximum component of $\vec{y}_1(x)$ at $x = x_1$. It is easily shown that $\vec{y}_{2c}(x_1|x_1)$ is a uniquely defined member of $\{\vec{y}_2(x)\}$ if $\vec{y}_2(x)$ in (20) is any member of this family of solutions of equation (1). At $x = x_1$, $\vec{y}_{2c}(x|x_1)$ has a component which is greatest in modulus. This component will be called the maximum component of $\vec{y}_{2c}(x_1|x_1)$. If more than one component of $\vec{y}_{2c}(x_1|x_1)$ achieves the maximum, then the component with lowest index will be called the maximum component.

Consider

$$\vec{y}_{jc}(x|x_1) = \vec{y}_2(x) + \sum_{k=1}^{j-1} c_{jk}\vec{y}_{kc}(x|x_1) \tag{21}$$

where $c_{j1}, \ldots, c_{j(j-1)}$ in equation (21) are chosen so that $\vec{y}_{jc}(x_1|x_1)$ has j-1 zero components corresponding to the distinct maximum components of $\vec{y}_1(x_1)$, $\vec{y}_{2c}(x_1|x_1)$, ..., $\vec{y}_{(j-1)c}(x_1|x_1)$. It is easily shown that $\vec{y}_{jc}(x|x_1)$ is a unique member of $\{\vec{y}_j(x)\}$ which depends on x_1, but is independent of the particular choice of $\vec{y}_j(x)\varepsilon\{\vec{y}_j(x)\}$ in (21).

If the Euclidean norm in the definitions of $e_j(x;x_0)$ and $r_j(x;x_0)$ is replaced by the maximum norm; that is,

$$||\vec{y}||_{max} = \max_{1 \leq i \leq n} |y_i|$$

where y_i is the *ith* component of $\vec{y}$,

$$e_j(x;x_0) = \min_{\substack{\text{(at each fixed x)} \\ \vec{y}_j(x)\varepsilon\{\vec{y}_j(x)\}}} ||\vec{y}_{je}(x;x_0) - \vec{y}_j(x)||_{max}$$

and

$$r_j(x;x_0) = \frac{e_j(x;x_0)}{f_1(x)}$$

where

$$f_1(x) = \min_{\substack{\text{(at each fixed x)} \\ \vec{y}_j(x)\varepsilon\{\vec{y}_j(x)\}}} ||\vec{y}_j(x)||_{max}$$

then theorems which are analogous to theorems 2 and 3 can be established. Namely, if $\vec{y}_{je}(x;x_0)$ can be determined exactly, then

$$r_j(a;x_0) = O\left(\frac{1}{x_0^{N+L}}\right)$$

In addition, $\vec{y}_{jc}(x|x)$ satisfies

$$\frac{||\vec{y}_{jc}(x|x)||_{max}}{f_1(x)} \leq n! \qquad (22)$$

for $j = 1,2,\ldots,n$. Hence, the canonical form $\vec{y}_{jc}(x|x)$ of the family $\{\vec{y}_j(x)\}$ bounds the multiples of $\vec{y}_{j-1}(x), \ldots, \vec{y}_1(x)$ which are present in $\vec{y}_{jc}(x|x)$.

Algorithm 3

Proceed as in algorithm 1 but instead of reducing $\vec{y}_{je}(x;x_0)$ approximately to $\vec{y}_{jmin}(x)$ at $x = x_0 - \ell(x_0 - a)/L$ for $\ell = 1,2,\ldots,L$, reduce $\vec{y}_{je}(x;x_0)$ approximately to $\vec{y}_{jc}(x|x)$ at these same values of x . That is, replace the process of orthogonalization by the process of reduction to canonical form.

Example: In order to crystallize the ideas presented, a particular problem will be considered. Suppose the coefficient matrix in equation (1) is specified by

$$A(x) = T^{-1}(x)\left\{D(x)T(x) - \frac{dT(x)}{dx}\right\} \qquad (23a)$$

where

$$T(x) = UV\left(I + \frac{W}{x+3}\right) \tag{23b}$$

$$U = \begin{bmatrix} u_{11} & 0 & 0 \\ u_{21} & u_{22} & 0 \\ u_{31} & u_{32} & u_{33} \end{bmatrix} \tag{23c}$$

$$V = \begin{bmatrix} 1 & v_{12} & v_{13} \\ 0 & 1 & v_{23} \\ 0 & 0 & 1 \end{bmatrix} \tag{23d}$$

$$I + \frac{W}{x+3} = \begin{bmatrix} 1 & \frac{w_{12}}{x+3} & \frac{w_{13}}{x+3} \\ 0 & 1 & \frac{w_{23}}{x+3} \\ 0 & 0 & 1 \end{bmatrix} \tag{23e}$$

and

$$D(x) = \begin{bmatrix} -a + \frac{b}{x+2} & 0 & 0 \\ 0 & \frac{c}{x+1} & 0 \\ 0 & 0 & d + \frac{f}{x+4} \end{bmatrix} \tag{23f}$$

Then a fundamental solution of equation (1) has the form

$$Y(x) = T^{-1}(x) \exp\left\{\int^x D(s)\, ds\right\} \tag{24a}$$

where

$$\exp\left\{\int^x D(s)\, ds\right\} = \begin{bmatrix} (x+2)^b e^{-ax} & 0 & 0 \\ 0 & (x+1)^c & 0 \\ 0 & 0 & (x+4)^f e^{dx} \end{bmatrix} \tag{24b}$$

For this constructed example the formal expansions in equation (3) are convergent. An error will be introduced into the initial vector $\vec{y}_{je}(x_0;x_0)$ in order to simulate the more realistic situation when the formal solutions are divergent.

For the particular problem defined by relations (23) and (24), a numerical integration of equation (1) was performed for the constants in (24b) specified by $a = 1$, $b = 2$, $c = 1$, $d = 1$, and $f = 3$. Hence, the scalar growths of solutions of (1) are approximately $(x+2)^2 e^{-x}$, $(x+1)$, and $(x+4)^3 e^x$. A numerical integration of (1) was performed over the interval, starting at $x = 40$ and terminating at $x = 0$. An error was introduced into the initial conditions for each of these solutions. In all cases a fourth-order Runge-Kutta method was used to integrate (1), and the numerical step size was $h = 1/32$. (See figures 2 through 6.)

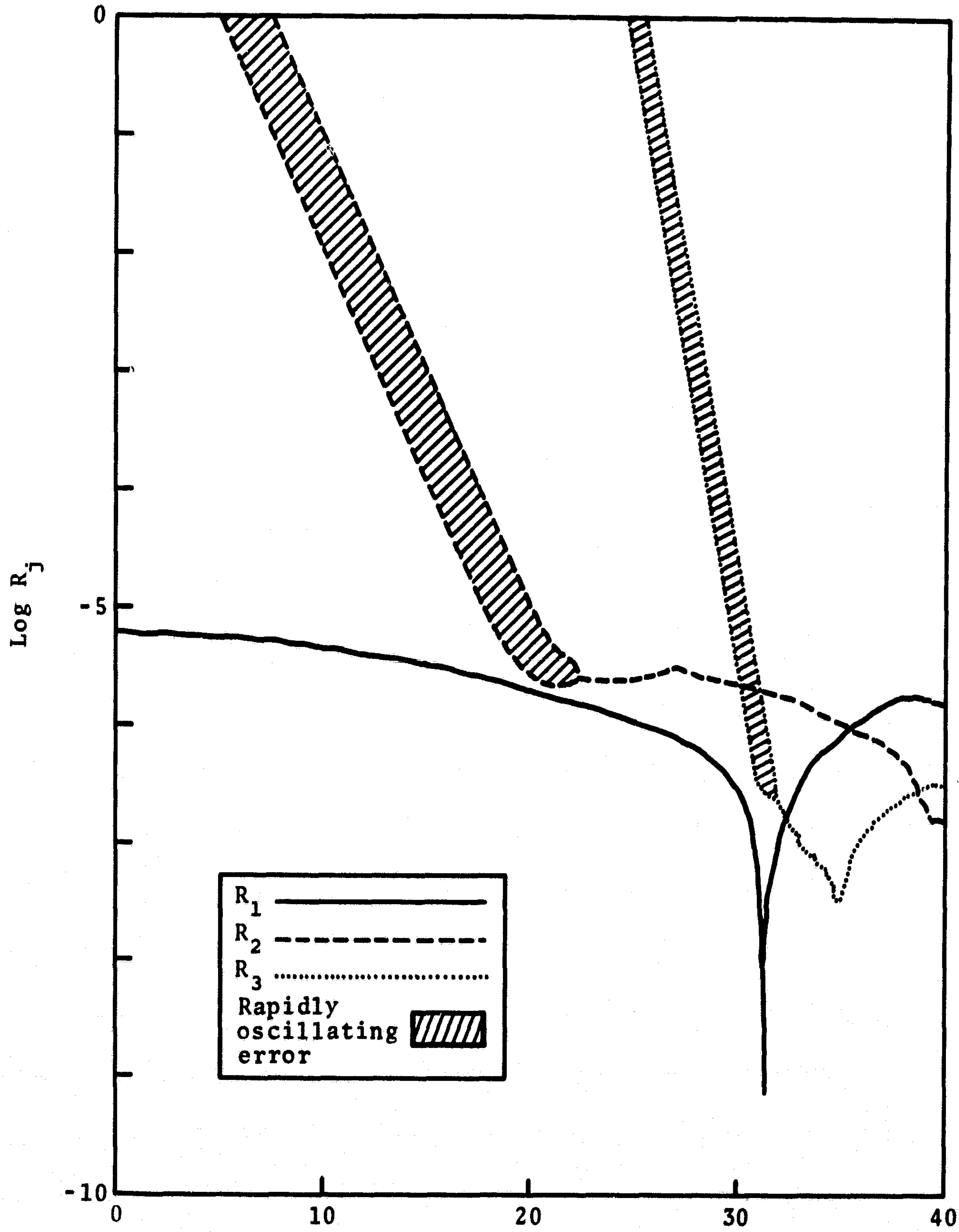

Figure 2. — No error control provided.

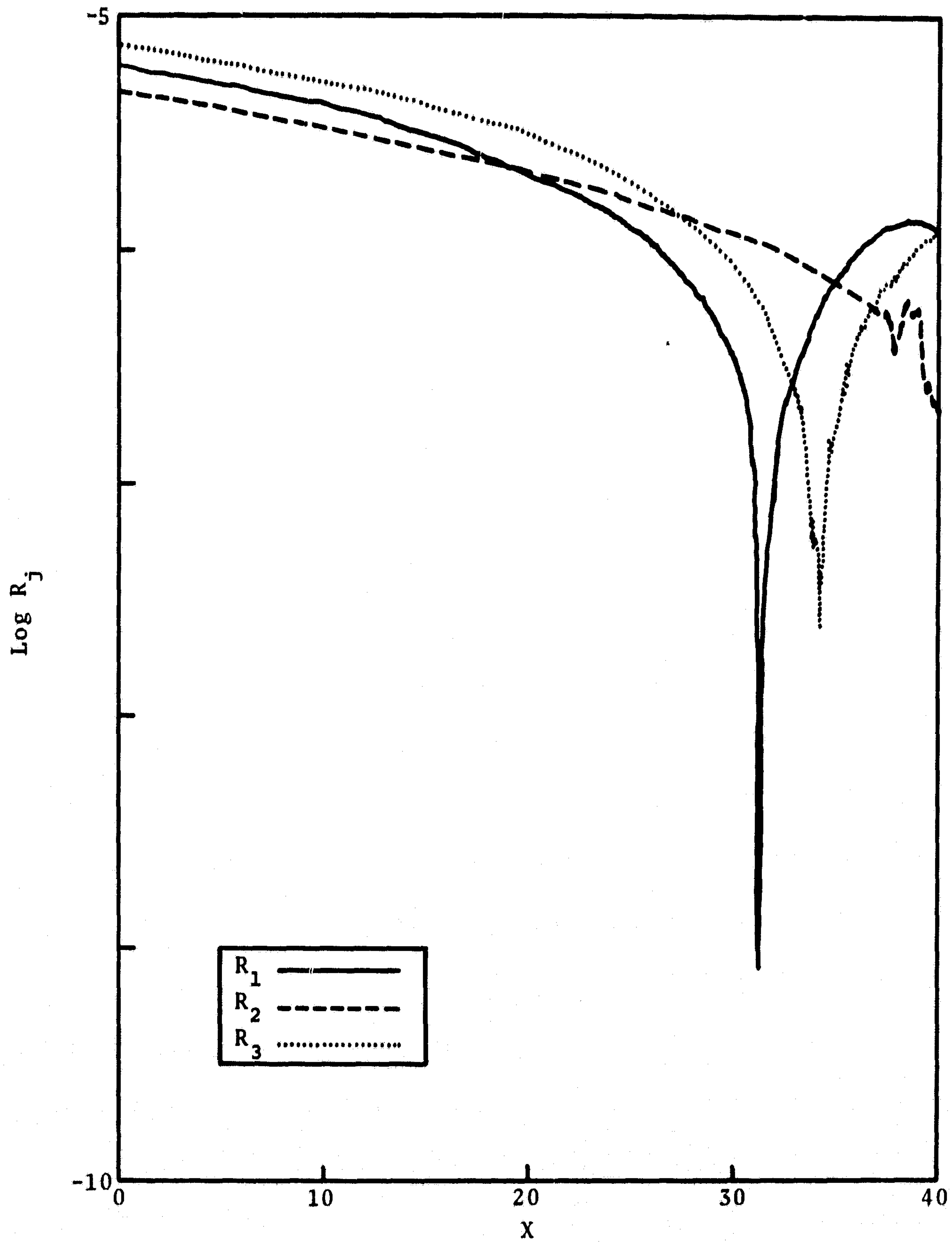

Figure 3. – Inherent error control provided by algorithm 1 (Conte's orthogonalization). Step e of algorithm 1 predicted a relative error of 0.71×10^{-6} and step e' predicted 0.24×10^{-5}.

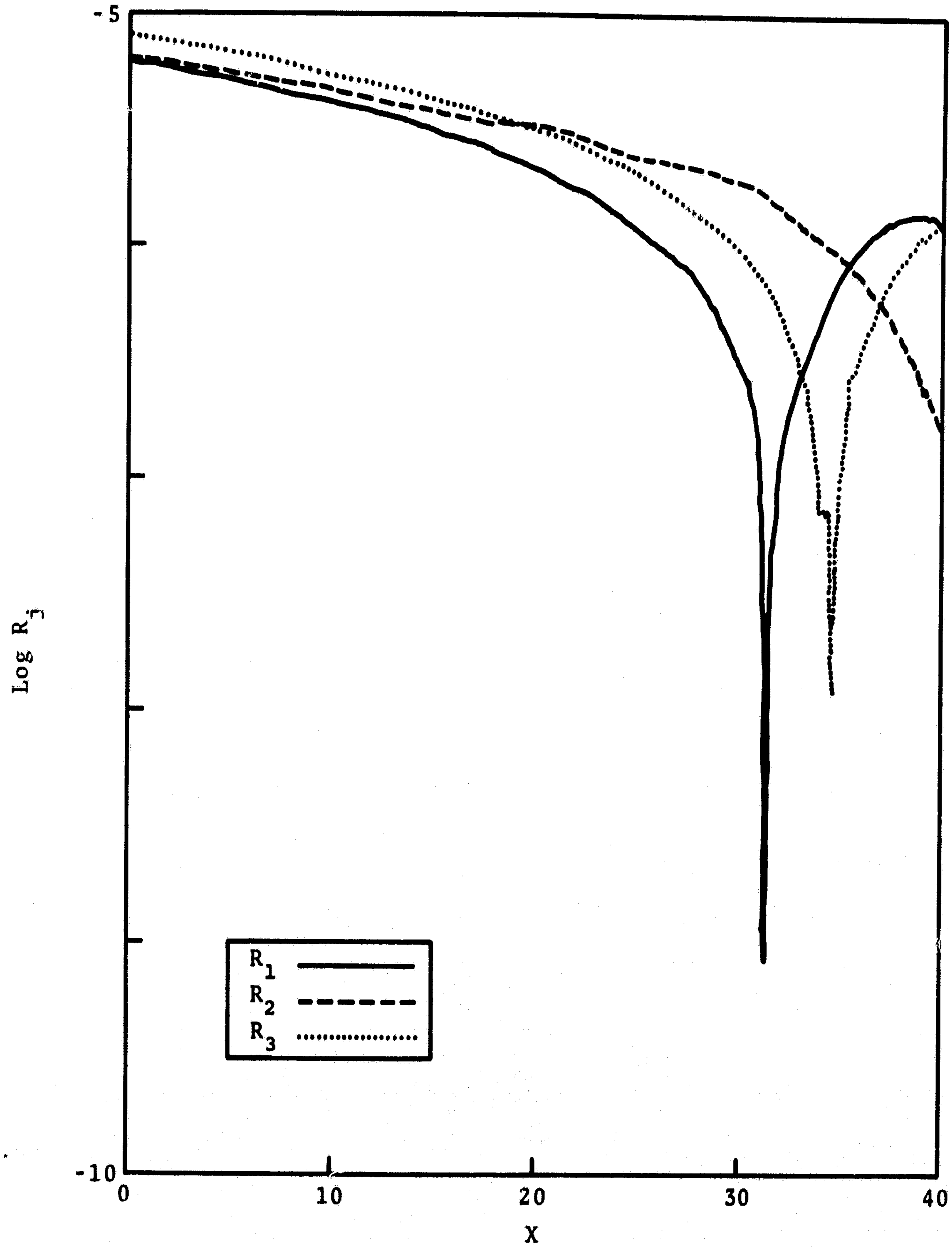

Figure 4. – Inherent error control provided by algorithm 2. Step e of algorithm 2 predicted a relative error of 0.31×10^{-6} and step e' predicted 0.87×10^{-6}.

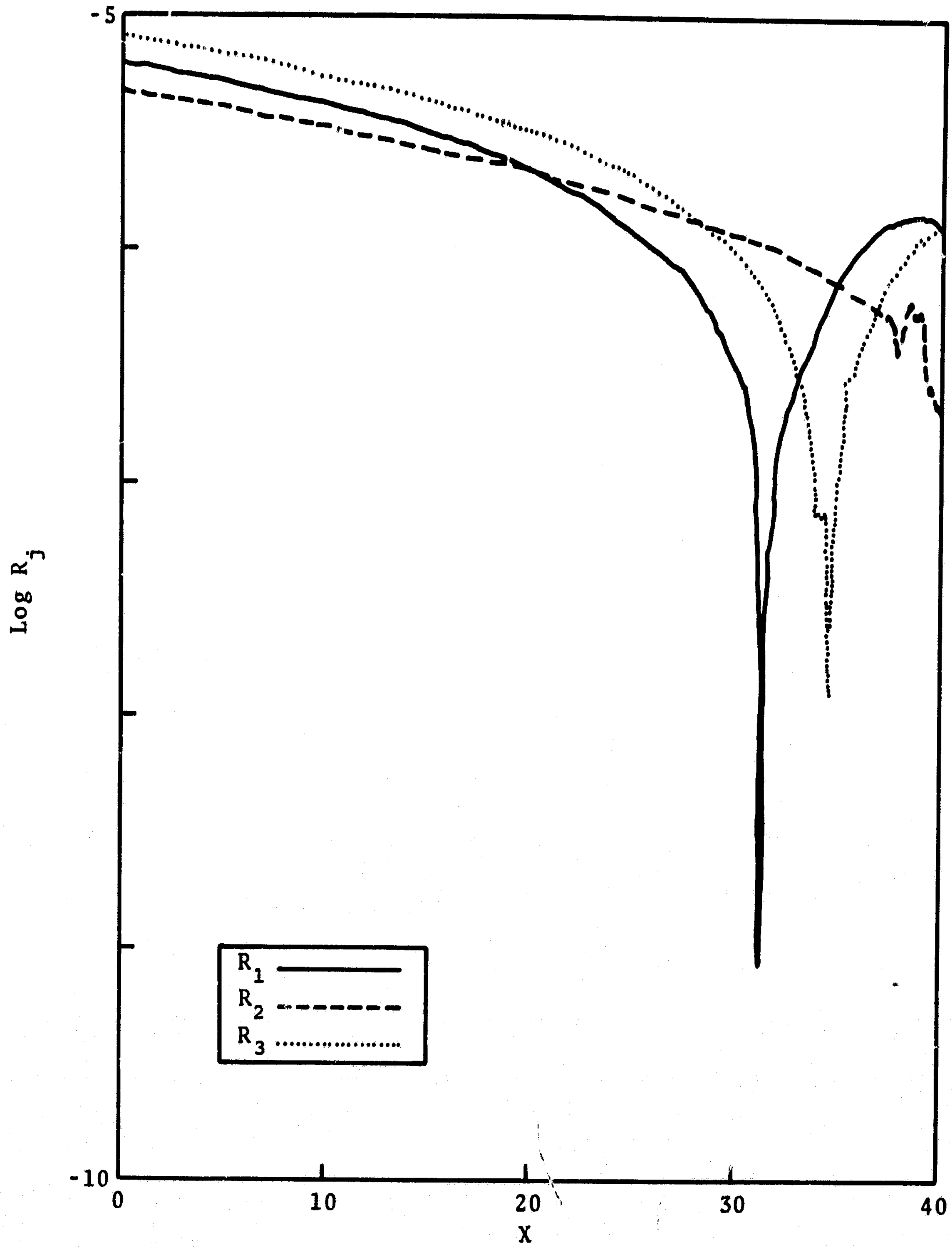

Figure 5. — Inherent error control provided by algorithm 3. Step e of algorithm 3 predicted a relative error of 0.12×10^{-6} and step e' predicted 0.22×10^{-5}.

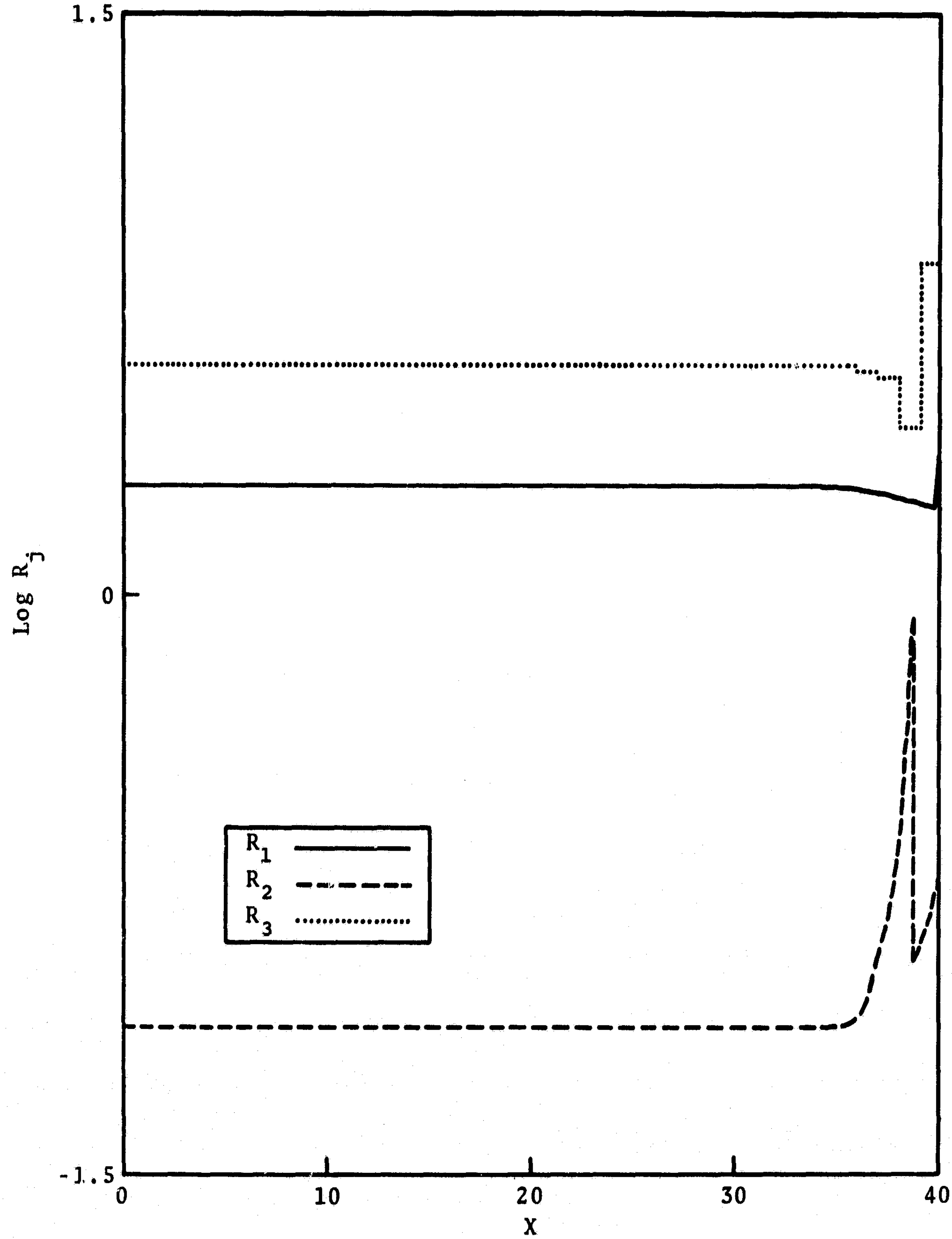

Figure 6. – Blunders in initial vectors. Inherent error control provided by algorithm 1. Step e of algorithm 1 predicted a relative error of 0.9×10^{-6} and step e' predicted 0.73 .

4. CONCLUSIONS

The algorithms 1, 2, and 3 considered in this paper have two main properties:

1. The algorithms make use of a reduction procedure which singles out unique members in each of the families $\{\vec{y}_j(x)\}$ at prescribed values of x .
2. The vector to which $\vec{y}_j(x)$ is reduced limits the multiples of $\vec{y}_{j-1}(x), \ldots, \vec{y}_1(x)$ which are present in the reduced form of $\vec{y}_j(x)$. As a bonus, the problem of inherent error growth is effectively controlled.

It should be noted that whenever the predictions of the maximum relative errors from step e and step e' differ by several orders of magnitude and

$$\left| e^{\lambda_n (x_0 - a)} x_0^{\alpha_n} \right| >> \left| e^{\lambda_j (x_0 - a)} x_0^{\alpha_j} \right|$$

for $j = 1,2,\ldots,n-1$, then it is likely that a blunder has been made (see figure 6).

The author has also considered a problem where the lead coefficient matrix, A_0 in relation (2), has some repeated eigenvalues. For this problem the algorithms are easily modified. However, it is also necessary to modify theorems 2 and 3 since initial errors may experience an algebraic magnification proportional to $|x_0/a|^\alpha$, where α is real and bounded. Generally speaking, algebraic magnification of initial errors is preferable to exponential magnification, and theoretically it is still possible to solve the asymptotic continuation problem.

REFERENCES

1. Conte, S. D.: "The Numerical Solution of Linear Boundary Value Problems." *SIAM Review*, vol. 8, no. 3, July 1966, pp. 309-321.

2. Fox, L.: *Numerical Solution of Ordinary and Partial Differential Equations.* Pergamon Press, 1962.

3. Keller, H. B.: *Numerical Methods for Two-Point Boundary-Value Problems.* Blaisdell Publishing Co., 1968.

4. Coddington, E. A.; and Levinson, N.: *Theory of Ordinary Differential Equations.* McGraw-Hill Book Co., Inc., 1955.

5. Wasow, N.: *Asymptotic Expansions for Ordinary Differential Equations.* John Wiley and Sons, Inc., 1955.

NASA — MSC — Coml., Houston, Texas