Numer. Math. 19, 218—229 (1972)

A Basic Theorem in the Computation of Ellipsoidal Error Bounds

K. G. Guderley and C. L. Keller

Received June 13, 1971

Summary. In this paper we formulate and prove the basic principles of a procedure for computing a bound on the error in the numerical solution of a system of linear differential equations. The bound on the error at each integration step is expressed in terms of an ellipsoid whose size and orientation is determined by the computations. To illustrate the procedure, Bessel's equation (of order zero) is integrated over the interval $2 \leqq x \leqq 3$ at steps of length 0.1 and bounds on the error are given for each step.

In an internal report [1], we have shown a method for computing error bounds to the solution of a system of ordinary linear differential equations. In this report we developed a geometrical visualization of the process. In problems of this kind the motivation of the individual steps is quite helpful; however, in order to get an overview of the procedure, it may be worth-while to extract the mathematical essence and to describe the procedure in terms of a number of lemmas and theorems. This is done in this paper.

Generally speaking, the procedure can be described as the computation of a Liapunov function for the problem. The emphasis in the present approach is different, however, from that usually encountered in Liapunov theory. If one is interested in the stability of a system and has a problem which is rather stable, then the set described by the Liapunov function may be much larger than the actual solution. However, in a procedure to give error bounds, such a characterization may be far too pessimistic. Thus we are interested in determining Liapunov functions which stay close to the actual solutions.

We use ellipsoids for the description of the error bounds. Other characterizations, for instance, parallelpipeds with axes not restricted to being parallel to the coordinate axes can serve as well. Ellipsoids were chosen for their analytic convenience. In restricting error bounding figures to a certain type, one must find a compromise between flexibility and narrowness of the bound. There are examples for which the bounds given by a bounding procedure carried out in terms of one of the usual norms or even pseudo norms are unrealistically wide. The present procedure if carried out in its best possible manner will give bounds which exceed the actual error by a finite factor, which depends upon the number of equations of the system. Thus the present approach is not only better than one which uses a fixed norm, but is good in that no major improvements can be made.

After working on the method for some time it was called to our attention that Professor William Kahan of the University of Toronto had made an announcement[1] concerning such a method. We feel that the presentation given here is sufficiently different to be of interest.

We begin with five lemmas. These are basic steps in the derivation of Theorem 1. Then we state Theorem 1 and sketch a proof. Lastly, we indicate how Theorem 1 is used in the problem of computing error bounds and give two more lemmas which are useful in determining the matrix $R(x)$ in Theorem 1. A computer program which carries out the procedure described here is given in [2] and [3].

All quantities are real. If u and v are n-dimensional vectors with components u_i and v_i relative to some orthogonal basis, then $\langle u, v\rangle$ denotes the inner product of the vetors u and v and

$$\langle u, v\rangle = \sum_i u_i v_i.$$

Lemma 1. Let $\hat{V}$ denote a symmetric, positive definite matrix and let r be a given vector. For any real number $a>0$ and vector y different from the null vector the inequality

$$2|\langle r, \hat{V}y\rangle| \leqq a^{-2}\langle y, \hat{V}y\rangle + \frac{a^2}{\langle y, Vy\rangle}\langle r, \hat{V}y\rangle^2 \tag{1}$$

holds.

Proof. Set

$$k = |\langle r, \hat{V}y\rangle| / \langle y, \hat{V}y\rangle.$$

The desired conclusion follows from

$$\langle aky - (1/a)y, \hat{V}(aky - [1/a]y)\rangle \geqq 0$$

after a little algebraic manipulation and replacing k by its value.

Lemma 2. Let R denote a symmetric nonnegative matrix. Set

$$\hat{E} = \{u | \langle u, Ru\rangle \leqq 1\} \tag{2}$$

and

$$E = \{r | r = Ru \text{ for some } u \text{ in } \hat{E}\}. \tag{3}$$

Then for an arbitrary but fixed vector y

$$\max_{r\in E}\langle r, y\rangle = \langle y, Ry\rangle^{\frac{1}{2}}. \tag{4}$$

This result follows from the Cauchy-Schwarz inequality.

Lemma 3. Let $\hat{V}$ denote a symmetric positive definite matrix and let V denote the inverse of $\hat{V}$. If a vector y satisfies the condition $\langle y, \hat{V}y\rangle \leqq q_0^2$, $q_0>0$, then

1 See the abstracts for contributed papers to the Symposium on Numerical Solution of Nonlinear Differential Equations, May 11–14, 1966, sponsored by the Mathematics Research Center, US Army and SIAM, Iowa City, Iowa.

the k-th component y_k of the vector y satisfies the condition

$$|y_k| \leqq q_0 \sqrt{V_{kk}}. \tag{5}$$

Here V_{kk} denotes the diagonal element in the k-th row and column of V.

Proof. Let e_k be the unit vector in the direction of the k-th coordinate axis. We have

$$\begin{aligned} y_k &= \langle y, e_k \rangle = \langle y, \hat{V} V e_k \rangle \\ &= \langle \hat{V} y, V e_k \rangle \end{aligned}$$

and

$$|\langle \hat{V} y, V e_k \rangle| \leqq \langle \hat{V} y, V \hat{V} y \rangle^{\frac{1}{2}} \langle e_k, V e_k \rangle^{\frac{1}{2}}.$$

Lemma 4. Let the matrix $A(x)$ be continuous and the matrix $R(x)$ be symmetric, nonnegative and continuous for $x_i \leqq x \leqq x_f$. Let the real function $a^2(x)$ be continuous and different from zero. Let V_0 denote a symmetric and positive definite matrix. Then the matrix $V(x)$ which satisfies the conditions

$$\begin{aligned} &-V'(x) + A(x) V(x) + V(x) A^\dagger(x) + a^{-2}(x) V(x) + \left(a^2(x)/q_0\right) R(x) = 0, \\ &V(x_i) = V_0 \end{aligned} \tag{6}$$

is symmetric and positive definite throughout the interval $[x_i, x_f]$.

Proof. Since $V(x)$ is symmetric there is an orthogonal transformation $S(x)$ such that

$$D(x) = S(x) V(x) S^\dagger(x),$$

where $D(x)$ is a diagonal matrix. Using the expression for $V'(x)$ as given by Eq. (6) and the fact that $S S^\dagger = S^\dagger S = I$ one obtains

$$D' = S' S^\dagger D + D S (S^\dagger)' + D S A^\dagger S^\dagger + S A S^\dagger D + a^{-2} D + (a^2/q_0^2) S R S^\dagger.$$

It follows that the i-th diagonal element d_{ii} of D satisfies an equation of the form

$$d'_{ii} = p(x) d_{ii} + (S R S^\dagger)_{ii}.$$

Since R is positive, $S R S^\dagger$ is positive and therefore $(S R S^\dagger)_{ii} \geqq 0$. It follows that $d_{ii} > 0$ and $V(x)$ is positive definite.

Corollary. If $V(x)$ is a symmetric matrix such that $V(x_i)$ is positive definite and the matrix

$$-V'(x) + A(x) V(x) + V(x) A^\dagger(x) + a^{-2}(x) V(x) + \left(a^2(x)/q_0^2\right) R(x) \tag{7}$$

is nonpositive for $x_i \leqq x \leqq x_f$ then $V(x)$ is positive definite.

Lemma 5. Let $y(x)$ denote a vector and $\hat{V}(x)$ a symmetric, positive definite matrix. Suppose both $y(x)$ and $\hat{V}(x)$ are continuously differentiable for $x_i \leqq x \leqq x_f$. If

$$\frac{d}{dx} \langle y(x), \hat{V}(x) y(x) \rangle \leqq 0, \qquad x_i \leqq x \leqq x_f \tag{8}$$

then

$$\langle y(x), \hat{V}(x) y(x) \rangle \leqq \langle y(x_i), \hat{V}(x_i) y(x_i) \rangle \tag{9}$$

for $x_i \leqq x \leqq x_f$.

Proof.

$$\langle y(x), \hat{V}(x)y(x)\rangle = \langle y(x_i), \hat{V}(x_i)y(x_i)\rangle + \int_{x_i}^{x} \frac{d}{dt}\langle y(t), V(t)y(t)\rangle dt.$$

Theorem 1. (i) Let V_0 be a symmetric, positive definite matrix.

(ii) Let $R(x)$ denote a symmetric nonnegative matrix which depends continuously on x for $x_i \leqq x \leqq x_f$. Let

$$\hat{E} = \{u(x) | u(x) \text{ is continuous and } \langle u(x), R(x)u(x)\rangle \leqq 1\}.$$

Let $u(x)$ be a member of $\hat{E}$ and set

$$r(x) = R(x)u(x). \tag{10}$$

(iii) Let the vector $y(x)$ satisfy the conditions

$$\begin{aligned} y'(x) &= A(x)y(x) + r(x), \qquad x_i \leqq x \leqq x_f \\ \langle y(x_i), V_0^{-1} y(x_i)\rangle &\leqq q_0^2. \end{aligned} \tag{11}$$

(iv) Let $a^2(x)$ be continuous and also $a^2(x) > 0$.

(v) Let the matrix $V(x)$ satisfy the conditions

$$\begin{aligned} &-V'(x) + A(x)V(x) + V(x)A^{\dagger}(x) + a^{-2}(x)V(x) + \frac{a^2(x)}{q_0^2} R(x) \leqq 0, \\ &V(x_i) = V_0. \end{aligned} \tag{12}$$

Then $y(x)$ satisfies the inequality

$$\langle y(x), V^{-1}(x)y(x)\rangle \leqq q_0^2, \qquad x_i \leqq x \leqq x_f, \tag{13}$$

and the components $y_k(x)$ of $y(x)$ satisfy

$$|y_k(x)| \leqq q_0 \sqrt{V_{kk}(x)}, \qquad x_i \leqq x \leqq x_f. \tag{14}$$

In this theorem we specify a family of x dependent vectors $y(x)$ on an interval $[x_i, x_f]$. Then we assert that any $y(x)$ of this family satisfies the inequality (13) if the matrix $V(x)$ satisfies the conditions (12). Geometrically, the matrices $V(x)$ determine ellipsoids and those vector functions $y(x)$ which initially lie within or on the ellipsoid determined by V_0 and which satisfy the differential equation (11) lie within or at most on the ellipsoids determined by the matrices $V(x)$.

Proof. Let $V(x)$ satisfy conditions (12). Then $V(x)$ is symmetric and positive definite by Lemma 4 and the corollary. Set $\hat{V}(x) = V^{-1}(x)$ and suppose there is a $y(x)$ which satisfies conditions (11) but that

$$\langle y(x), V(x)y(x)\rangle > q_0^2, \qquad x_i \leqq x \leqq x_f. \tag{15}$$

(There is no loss of generality by assuming the inequality (15) holds throughout the interval $[x_i, x_f]$, for the argument which follows is valid for any subinterval

over which the inequality (15) holds.) We have

$$\begin{aligned}\frac{d}{dx}\langle y(x), \hat{V}(x)y(x)\rangle &= \langle y(x), [\hat{V}'(x)+\hat{V}(x)A(x)+A^{\dagger}(x)\hat{V}(x)]y(x)\rangle \\ &\quad +2\langle r(x), \hat{V}(x)y(x)\rangle \\ &\leqq \langle y(x), [\hat{V}'(x)+\hat{V}(x)A(x)+A^{\dagger}(x)\hat{V}(x)]y(x)\rangle \\ &\quad +a^{-2}(x)\langle y(x), \hat{V}(x)y(x)\rangle \\ &\quad +\frac{a^2(x)}{\langle y(x), \hat{V}(x)y(x)\rangle}\langle r(x), \hat{V}(x)y(x)\rangle^2\end{aligned} \tag{16}$$

by Lemma 1. By Lemma 2

$$\max_{r(x)}\langle r(x), \hat{V}(x)y(x)\rangle \leqq \langle y(x), \hat{V}(x)R(x)\hat{V}(x)y(x)\rangle^{\frac{1}{2}}. \tag{17}$$

Using our assumption (15) and the result (17) to modify the last term of (16) we obtain

$$\begin{aligned}&\frac{d}{dx}\langle y(x), \hat{V}(x)y(x)\rangle \\ &\quad \leqq \Big\langle y(x), \Big[\hat{V}'(x)+\hat{V}(x)A(x)+A^{\dagger}(x)\hat{V}(x)+a^{-2}(x)\hat{V}(x) \\ &\qquad +\frac{a^2(x)}{q_0^2}V(x)R(x)\hat{V}(x)\Big]y(x)\Big\rangle.\end{aligned} \tag{18}$$

Next, using the facts that $V(x)\hat{V}(x)=I$ and $V(x)\hat{V}'(x)=-V'(x)\hat{V}(x)$, we obtain

$$\begin{aligned}&\frac{d}{dx}\langle y(x), \hat{V}(x)y(x)\rangle \\ &\quad \leqq \Big\langle \hat{V}(x)y(x), \Big[-\hat{V}(x)+A(x)V(x)+V(x)A^{\dagger}(x)+a^{-2}(x)V(x) \\ &\qquad +\frac{a^2(x)}{q_0^2}R(x)\Big]\hat{V}(x)y(x)\Big\rangle \\ &\quad \leqq 0\end{aligned} \tag{19}$$

by hypothesis, condition (12), of this theorem. It follows from Lemma 5 that $\langle y(x), \hat{V}(x)y(x)\rangle \leqq q_0^2$. That is, the assumption (15) is untenable.

Next we show how Theorem 1 is used in determining error bounds. There are many quantities, in particular $r(x)$, $R(x)$, V_0, q_0, and $a^2(x)$ appearing in the statement of Theorem 1. We will indicate the role of these quantities and how they are determined.

We consider two cases. In the first case we suppose that the matrix operator A may be treated without error. In particular this is the case if A is a constant matrix or has entries which are polynomials in the independent variable x. Suppose that $w(x)$ is determined by the conditions.

$$\begin{aligned}&w'(x)=A(x)w(x)+f(x) \qquad \alpha \leqq x \leqq \beta, \\ &w(\alpha)=w_0.\end{aligned} \tag{20}$$

In general, a numerical solution of (20) is a known function $z(x)$ which satisfies

$$\begin{aligned} z'(x) &= A(x)z(x) + g(x), \qquad x_i \leqq x \leqq x_f, \\ z(x_i) &= z_i \end{aligned} \tag{21}$$

on a set of subintervals $[x_i, x_f]$ of $[\alpha, \beta]$. These subintervals $[x_i, x_f]$ collectively cover $[\alpha, \beta]$. The Eq. (21) determines $g(x)$.

The error $y(x)$ is the difference,

$$y(x) = w(x) - z(x), \tag{22}$$

between the vector functions $w(x)$ and $z(x)$. In this paper we assume, for convenience, that the eror due to round-off is negligible relative to the truncation error. Then we prescribe a method for computing a bound on the error resulting from truncation and propagation of initial error.

The equation for the error on a subinterval $[x_i, x_f]$ is

$$\begin{aligned} y'(x) &= A(x)y(x) + r(x), \qquad x_i \leqq x \leqq x_f, \\ y(x_i) &= w(x_i) - z(x_i). \end{aligned} \tag{23}$$

We call $r(x)$ the residual but it is called the deviation also. Here

$$\begin{aligned} r(x) &= f(x) - g(x) \\ &= f(x) + A(x)z(x) - z'(x) \end{aligned} \tag{24}$$

for $x_i \leqq x \leqq x_f$.

Lemma 6. Let v be a given vector and set

$$R = vv^{\dagger}. \tag{25}$$

Set

$$\hat{E} = \{u \mid \langle u, Ru \rangle \leqq 1\}$$

and

$$E = \{r \mid r = Ru \text{ for some } u \text{ in } \hat{E}\}.$$

Then v is in E.

The (degenerate) ellipsoid E is determined, as indicated in the statement of Lemma 6, by the matrix R. It is clear that the ellipsoid E is contained in every ellipsoid which contains the vector v. Set $R(x) = r(x)r^{\dagger}(x)$ and take $u(x) = r(x)/\langle r(x), r(x) \rangle$. Then $\langle u(x), R(x)u(x) \rangle = 1$ and $r(x) = R(x)u(x)$.

For any integration step other than the first, we may take for V_0, $V(x)$, as determined from the previous integration step, evaluated at the final point of the preceding subinterval. For the first integration step, lacking a better choice, one may take for V_0 the identity matrix.

For all steps, except possibly the first, there is some initial error, namely, the error generated over the preceding step and the propagated error from previous steps. If we allow that there is some small error at the start and we take for V_0 the identity matrix then q_0 denotes the radius of the n-sphere which bounds the initial error. In [3] we describe a modification which permits V_0 to be zero.

Lastly, we need the positive real function $a^2(x)$. This quantity $a^2(x)$ influences the size and shape of the bounding ellipsoids determined by $V(x)$. The

quantity a was introduced in Lemma 1. It is clear that for an appropriate value of a the inequality (1) becomes equality. The same holds for the inequality (16). Thus if the real function $a^2(x)$ could be appropriately determined, the boundaries of the error figure and the bounding ellipsoid have points in common. On the other hand forcing the bounding ellipsoid to be tangent to the error figure at some ill chosen point could result in an undesirable elongation of the bounding ellipsoid in some direction.

Let $\mathrm{Tr}\,R$ denote the trace of the matrix R, that is, $\mathrm{Tr}\,R$ denotes the sum of the diagonal elements of R. An approximation to the quantity

$$q_0[\mathrm{Tr}\,V(x)/\mathrm{Tr}\,R(x)]^{\frac{1}{2}} \tag{26}$$

is a convenient and usually a satisfactory choice for $a^2(x)$.

In [3] we give a procedure for determining the quantities discussed above and a computer program for calculating the error for this first case. Before discussing the second case we need to give one more lemma.

Let R_1 and R_2 be two symmetric nonnegative matrices. Set

$$\hat{E}_1 = \{u \mid \langle u, R_1 u\rangle \leqq 1\}, \tag{27}$$

$$E_1 = \{y \mid y = R_1 u, \text{ for } u \text{ in } \hat{E}_1\}, \tag{28}$$

and similarly,

$$\hat{E}_2 = \{v \mid \langle v, R_2 v\rangle \leqq 1\}, \tag{29}$$

$$E_2 = \{z \mid z = R_2 v, \text{ for } v \text{ in } \hat{E}_2\}. \tag{30}$$

E_1 and E_2 are ellipsoids, possibly degenerate. We say that the ellipsoid E_1 is determined by the matrix R_1.

Set

$$S = E_1 + E_2 = \{s \mid s = y + z, \text{ for } y \text{ in } E_1 \text{ and } z \text{ in } E_2\}. \tag{31}$$

S is called the vector sum of E_1 and E_2.

Lemma 7. Let R_1 and R_2 be symmetric nonnegative matrices. For any real number $a > 0$ let the matrix R be given by the equation

$$R = R_1 + a R_1 + R_2 + (1/a) R_2. \tag{32}$$

Then the ellipsoid E determined by the matrix R contains the vector sum of the ellipsoids determined by R_1 and R_2.

The result expressed in Lemma 7 was given to us by Professor William Kahan. A derivation and discussion of Lemma 7 is given in [4] and also [5].

In this second case we suppose that $w(x)$ is determined by the conditions

$$\begin{aligned} w'(x) &= B(x)w(x) + f(x), \qquad \alpha \leqq x \leqq \beta, \\ w(\alpha) &= w_0. \end{aligned} \tag{33}$$

Suppose also that the numerical solution obtained satisfies conditions (21) and $r(x)$ is given by (24) just as in the first case. That is, $A(x)$ is some approximation

to the matrix $B(x)$ and $r(x)$ as we define it is a partial residual in this second case. Then the equation for the error may be written as

$$\begin{aligned} &y'(x) = A(x)y(x) + r(x) + [B(x) - A(x)]z(x) + [B(x) - A(x)]y(x). \\ &y(x_i) = w(x_i) - z(x_i). \end{aligned} \tag{34}$$

As in the first case set $R_1(x) = r(x)r^\dagger(x)$. Let $R_2(x)$ and $R_3(x)$ denote symmetric, nonnegative matrices that determine ellipsoids which contain the vectors $[B(x) - A(x)]z(x)$ and $[B(x) - A(x)]y(x)$ respectively. Let $R(x)$ denote a symmetric nonnegative matrix obtained from $R_1(x)$, $R_2(x)$ and $R_3(x)$ by repeated application of Lemma 7. Then a matrix $V(x)$ which satisfies

$$\begin{aligned} &-V'(x) + A(x)V(x) + V(x)A^\dagger(x) + \left(1/a^2(x)\,V(x) + \left(a^2(x)/q_0^2\right)R(x) \leqq 0 \right. \\ &V(x_i) = V_0, \end{aligned}$$

where V_0 and $a^2(x)$ are obtained in the same fashion as in case 1, determines an ellipsoid which contains the error $y(x)$ given by (34).

A discussion of the relation between the bound obtained by this procedure and the actual error is given in [1].

The inequalities occurring in the above equations, for example, in the formulae for the residual and the matrix inequality for V, must be satisfied throughout the interval of integration. With the usual integration methods (Runge-Kutta or predictor-corrector methods) it is difficult to establish that these inequalities are satisfied. We found it practical to express all quantities in terms truncated power series and to bound the remainders. Accordingly, we assume that the matrix A is given in the individual intervals by a polynomial and that an estimate for the matrix $B - A$ is available in a corresponding form. This information should be regarded as input data.

A program for the case where $B \equiv A$ has been shown and explained in detail in Ref. [2]. In order to illustrate the application of the above results for the case $A \neq B$ we consider the integration of Bessel's equation of order zero over the interval $2 \leqq x \leqq 3$. Write the equation as

$$\begin{pmatrix} w_1'(x) \\ w_2'(x) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & -1/x \end{pmatrix} \begin{pmatrix} w_1(x) \\ w_2(x) \end{pmatrix}, \quad \begin{pmatrix} w_1(2) \\ w_2(2) \end{pmatrix} = \begin{pmatrix} J_0(2) \\ -J_1(2) \end{pmatrix}. \tag{35}$$

Let H denote the step size and set $x_i = 2 + iH$, $i = 0, 1, \ldots$. Suppose that the approximate solution $z(x)$ and the matrix $V(x)$ have been determined for $2 \leqq x \leqq x_i$. We discuss the next step in the integration process.

In this example the matrix A is the same as B in Eq. (35), except the entry $-1/x$ is replaced by the first three terms of the development of $-1/x$ about $x = x_i$.

The components of $z(x)$ are taken as polynomials of the third degree in $h = x - x_i$. Equating the coefficients of like powers of h in the two expressions $z'(x)$ and $A(x)z(x)$ fixes the coefficients for the polynomial components of $z(x)$. The remaining terms, that is, $A(x)z(x) - z'(x)$ constitute the residual $r(x)$,

Eq. (24). The components of $r(x)$ are polynomials in h also. Then

$$R_1(x) = r(x)r^\dagger(x). \tag{36}$$

The term $[B(x) - A(x)]z(x)$ in Eq. (34) is a vector whose magnitude does not exceed $h^3|z_2(x)|/x_i^4$, where $z_2(x)$ is the second component of $z(x)$. Hence, the matrix for an ellipsoidal bound for this vector is given by

$$R_2(x) = h^6 \begin{pmatrix} 0 & 0 \\ 0 & [z_2(x)]^2/x_i^8 \end{pmatrix}. \tag{37}$$

Similarly,

$$R_3(x) = h^6 \begin{pmatrix} 0 & 0 \\ 0 & q_0^2 V_{22}(x)/x_i^8 \end{pmatrix} \tag{38}$$

is the matrix for an ellipsoidal bound for the term $[B(x) - A(x)]y(x)$. Here $V_{22}(x)$ is the indicated diagonal element of the matrix $V(x)$. The quantity $V_{22}(x)$ is not known at this point, however. Hence we use a known value $V_{22}(x_i)$ as an initial estimate and revise later in an obvious way.

The various contributions to the residual as characterized by R_1, R_2, and R_3 are now combined using Lemma 7. We found the value $a = [\mathrm{Tr}\, R_3(x_i)/\mathrm{Tr}\, R_2(x_i)]^{\frac{1}{4}}$ a computationally convenient choice for the quantity a occurring in Lemma 7. Reasons for this choice are discussed in [1]. Thus R_2 and R_3 are combined to give a matrix $\hat{R}$. The matrices $\hat{R}$ and R_1 are combined in exactly the same way to give the matrix R of Theorem 1. The matrix R is of the form

$$R(x) = h^6(R_{ij}(x))$$

with the elements R_{ij} polynomials in h and $R_{ij} = R_{ji}$.

Now set

$$V(x) = (V_{ij}(x)) + cIh^3,$$

where the $V_{ij}(x)$ are polynomials in h of the second degree and $V_{ij}(x) = V_{ji}(x)$. The coefficients for these polynomials are determined in the same way as the coefficients for the components of $z(x)$. The constant c is to be determined so that the inequality (12) holds. The details for computing a c are given in [2] and also in [3].

The quantities $a^2(x)$ and q_0^2 are needed to complete the description of (12). Invoking again the trace criterion, we obtain

$$a^2(x) = q_0 h^{-3}[\mathrm{Tr}\, V(x_i)/\mathrm{Tr}\, R(x_i)]^{\frac{1}{4}}.$$

For the numerical example which we are presenting here, we took $z_1(2) = 0.22389$ and $z_2(2) = -0.57622$. Then assuming we know that

$$[(J_0(2) - z_1(2))^2 + (J_1(2) - z_2(2))^2]^{\frac{1}{2}} \leqq 5 \times 10^{-6},$$

we take $q_0 = 5 \times 10^{-6}$.

Proceeding as outlined above with step size $H = 0.1$ we obtained the following numerical results:

x	2.0	00	x	2.0	00
J_0	2.238907791	−01	J_1	5.767248077	−01
z_1	2.2389	−01	$-z_2$	5.7672	−01
Bound	5.	−06	Bound	5.	−06
J_0-z_1	7.791	−07	J_1+z_2	4.8077	−06
Δ_1	4.2209	−06	Δ_2	1.923	−07
	2.1	00		2.1	00
	1.666069803	−01		5.682921357	−01
	1.666070675	−01		5.682861762	−01
	7.343883432	−06		7.184714598	−06
	−8.72	−08		5.9595	−06
	7.256683432	−06		1.225214598	−06
	2.2	00		2.2	00
	1.103622669	−01		5.559630498	−01
	1.103634596	−01		5.559561155	−01
	9.556004353	−06		9.194212488	−06
	−1.1927	−06		6.934300	−06
	8.363304353	−06		2.259912488	−06
	2.3	00		2.3	00
	5.553978445	−02		5.398725326	−01
	5.554229865	−02		5.398648228	−01
	1.167270713	−05		1.107555173	−05
	−2.51420	−06		7.7098	−06
	9.15850713	−06		3.3657 5173	−06
	2.4	00		2.4	00
	2.50768330	−03		5.201852681	−01
	2.511708385	−03		5.201770019	−01
	1.368924663	−05		1.283445827	−05
	−4.025085	−06		8.2662	−06
	9.66416163	−06		4.56825827	−06
	2.5	00		2.5	00
	−4.83837764	−02		4.970941024	−01
	−4.837808051	−02		4.970855160	−01
	1.559077492	−05		1.446628283	−05
	−5.69589	−06		8.5864	−06
	9.89488492	−06		5.87988283	−06
	2.6	00		2.6	00
	−9.68049544	−02		4.708182665	−01
	−9.679745953	−02		4.708096101	−01
	1.758139147	−05		1.620415755	−05
	−7.49487	−06		8.6564	−06
	1.008652147	−05		7.54775755	−06
	2.7	00		2.7	00
	−1.424493700	−01		4.416013791	−01
	−1.424399823	−01		4.415929135	−01
	1.964078509	−05		1.802555795	−05
	−9.3877	−06		8.4656	−06
	1.025308509	−05		9.55995795	−06
	2.8	00		2.8	00
	−1.850360333	−01		4.097092468	−01
	−1.850246946	−01		4.097012398	−01
	2.174650315	−05		1.990692022	−05
	−1.13387	−05		8.0070	−06
	1.040780315	−05		1.189992022	−05

x	2.9	00	x	2.9	00
J_0	−2.243115457	−01	J_1	3.754274818	−01
z_1	−2.242982348	−01	$-z_2$	3.754202042	−01
Bound	2.387404161	−05	Bound	2.182367289	−05
J_0-z_1	−1.33109	−05	J_1+z_2	7.2776	−06
Δ_1	1.056314161	−05	Δ_2	1.454607289	−05
	3.0	00		3.0	00
	−2.600519549	−01		3.390589585	−01
	−2.600366884	−01		3.390526805	−01
	2.599721503	−05		2.375051295	−05
	−1.52665	−05		6.2780	−06
	1.073071503	−05		1.747251295	−05

The values for $J_0(x)$ and $J_1(x)$ were obtained from [7]. The quantity Δ_1 is defined by

$$\Delta_1 = \text{Bound} - |J_0 - z_1|$$

and similarly for Δ_2.

The values for the matrix $V(x)$ are

x	v_{111}	v_{112}	v_{122}
2.0	1	0	1
2.1	2.157304954 00	−5.000000000−03	2.064804954 00
2.2	3.652688768 00	−2.353540908−02	3.381341731 00
2.3	5.450083671 00	−6.393127746−02	4.906713849 00
2.4	7.495818942 00	−1.344831661−01	6.588932764 00
2.5	9.722890508 00	−2.426766734−01	8.370933562 00
2.6	1.236421304 01	−3.944798706−01	1.050298888 01
2.7	1.543041756 01	−5.949407713−01	1.299682958 01
2.8	1.891641598 01	−8.488117775−01	1.585141891 01
2.9	2.279879452 01	−1.160242212 00	1.905090794 01
3.0	2.703420758 01	−1.532428184 00	2.256347462 01

We present the pertinent information from the above tables pictorially in Fig. 1 below. The line segments emanating from the origin are the error vectors. The components of these vectors are the rows

$$J_0 - z_1 \quad \text{and} \quad -J_1 - z_2.$$

The error boxes are determined from the values in the row labeled Bound. The two ellipses were obtained from the v_{lij}'s corresponding to $x = 2.5$ and $x = 3.0$, respectively.

Fig. 1 shows that one indeed obtains bounds for the error and that these bounds are not too wide.

The comparison between the bound and the actual error is not quite fair. Ultimately, the residual $r(x)$ will always be characterized by describing a certain neighborhood of the origin to which it is confined. We chose a characterization

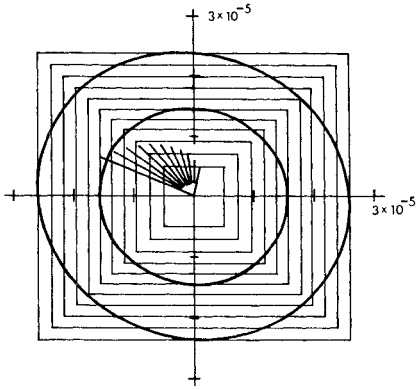

Fig. 1. The smooth closed curves are ellipses. The flatness is an optical illusion

given by ellipsoids $R(x)$. There exists, as a consequence of this characterization, a smallest convex set $K(x)$, the totality of solutions $y(x)$ of (11) of Theorem 1, which is the best characterization of the error once the characterization of the residual is given. A good method of error bounding is one for which the bounding figure, in our case an ellipsoid, is not much larger than the set $K(x)$. The error itself, by fortuitous cancellations, might be quite small.

References

1. Guderley, K. G., Keller, C. L.: Ellipsoidal bounds for the solutions of systems of ordinary linear differential equations. ARL Technical Report, January 1969.
2. Breiter, M. C., Keller, C. L., Reeves, 1st Lt T. E.: A program for computing error bounds for the solution of a system of differential equations. ARL Technical Report, March 1969.
3. Keller, C. L., Reeves, Capt T. E.: Computing error bounds when the truncation error and propagated error are of the same order. ARL Technical Report, April 1970.
4. Guderley, K. G., Keller, C. L.: Ellipsoidal bounds for the vector sum of two ellipsoids. ARL Technical Report, January 1969.
5. Guderley, K. G., Keller, C. L.: Enclosing the vector sum of two ellipsoids with an ellipsoid. Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg, Band 36, Juli 1971.
6. Cargo, G. T., Shisha, O.: The Bernstein form of a polynomial. Journal of Research **70**B, No. 1, 79–81, January-March 1966.
7. Cambi, E.: Bessel functions. New York: Dover Publications, Inc. December 1946.

K. G. Guderley
C. L. Keller
Department of the Air Force
Aerospace Reasearch Laboratories (AFSC)
Wright Patterson Air Force Base
Ohio 45433, USA