A-Posteriori Error Estimates

by

Seymour V. Parter

Computer Sciences Technical Report #214

May 1974
A-Posteriori Error Estimates

by

Seymour V. Parter*

I. INTRODUCTION

Most of the literature in numerical analysis is concerned with a-priori error estimates, i.e. a statement of form: Let $Y(\xi)$ be the solution of the problem. Let $\bar{Y}(\xi;h)$ be an "approximation" to $Y(\xi)$. Then

$$\|Y-\bar{Y}(\cdot,h)\|_\alpha \leq E(h) \|Y\|_\beta$$

where

(1) $E(h)$ is some error function, typically

$$E(h) = Ch^N$$

(2) $\|Y\|_\beta$ is some norm of the true solution $Y(\xi)$ and usually involves derivatives of $Y(\xi)$ of order $N$

(3) $\|Y-\bar{Y}(\cdot,h)\|_\alpha$ is an appropriate norm of the error function.

For many problems these a-priori estimates are sufficient. On the other hand, there are a large number of problems where either

a) They have not been established, or

b) they exist, but are not adequate because $\|Y\|_\beta$ is unknown.

* Supported by the Office of Naval Research under Contract N00014-67-A-0128-0004
The latter situation is a familiar one and (incredibly) does not seem to upset people.

On the other hand, particularly when dealing with nonlinear Boundary Value problems where non-uniqueness and bifurcation phenomena occur, it is not always possible to establish a-priori estimates. Indeed, it is precisely these difficulties that lead one to employ the "shooting" methods and rely on the usually excellent a-priori estimates for Initial Value problem.

Recently [6], [7] we came upon a class of problems where the need for better estimates was obviously imperative.

Nonlinear boundary value problems of the form

\[
\begin{align*}
  y'' + \frac{1}{x}y' + \beta f(x,y,\tau) &= 0, \quad 0 < x < 1 \\
  y'(0) &= 0, \quad y(1) = 0
\end{align*}
\]

(1.1)

arises in the study of chemically reacting systems with cylindrical symmetry; \( y \) is essentially the temperature, and \( \beta \) is a dimensionless combination of other physical parameters (see Gavalas [4], and Frank-Kamenetskii [3]).

In these problems one is not only concerned with computing accurate approximations to the solutions (there are frequently several solutions!) but one is also concerned with the bifurcation behavior, i.e. determine values of \( \beta, \tau \) at which bifurcation occurs. Indeed knowledge of the bifurcation behavior is of great importance even when one merely seeks one solution for fixed values of \( \beta, \tau \). If we know that we have determined a neighborhood which contains one and only one solution we may proceed with a variety boundary value techniques to solve this problem.
Unfortunately, while the Chemical Engineering literature is replete with computational results for problems of this form, there is essentially no discussion of error estimates. Moreover, the problem contains an apparent singularity which seems to eliminate most of the mathematical a-priori error estimates which require a uniform Lipschitz condition.

Since problems of this type occur in a variety of applications we felt it was desirable to develop the theory of a-posteriori error estimates in detail. Hence this lecture is concerned with a discussion of the many technical problems which arise in the computation of a-posteriori error estimates between a computed function $\tilde{Y}(x, y_0; h)$ and the function $Y(x, y_0)$ which solves the related initial value problem

\[ \begin{cases} Y'' + \frac{1}{x} Y' + \beta f(x, Y(x), \tau) = 0, & 0 \leq x \leq 1 \\ Y'(0, y_0) = 0, & Y(0, y_0) = y_0. \end{cases} \] (1.2)

Much of this discussion is elementary and quite apparent once the basic facts have been re-asserted. Nevertheless, we feel that this question is of sufficient importance that a somewhat detailed discussion is worthwhile. We hope that this is just the beginning of a serious study of the general problem of a-posteriori error estimates.
2. A BASIC ESTIMATE

Consider the linear initial-value problem

\[
\begin{cases}
\frac{1}{\beta(x)} (\alpha(x)E')' + F(x)E = R(x), & 0 < x \leq 1 \\
E(0) = E'(0) = 0
\end{cases}
\]  

(2.1)

where \( \alpha(x) \), \( \beta(x) \) are positive and "smooth" for \( x > 0 \), \( F(x) \) is bounded and

\[
B(x) = \frac{1}{\alpha(x)} \int_0^x \beta(t)dt \leq M, \ 0 < x \leq 1
\]

(2.2)

Let

\[
K(x,t) = \beta(t) \int_t^x \frac{d(s)}{\alpha(s)}
\]

(2.3)

An application of Fubini's theory shows that \( K(x,t) \in L^1(0,1) \). Moreover, after two integrations of (2.1) and the necessary application of Fubini's theorem, we see that \( E(x) \) satisfies the Volterra integral equation of the second kind

\[
E(x) = \int_0^x K(x,t)R(t)dt - \int_0^x K(x,t)F(t)E(t)dt.
\]

(2.4)

That is, if

\[
Q(x) = \int_0^x K(x,t)f(t)dt
\]

then \( Q(x) \) is the solution of the initial value problem

\[
\begin{cases}
(\alpha(x)Q')' = \beta(x)f(x), & 0 \leq x \leq 1 \\
Q(0) = Q'(0) = 0
\end{cases}
\]

(2.5)
Remark: In the case of the Bessel Operator of (1.2) we have

\[ \alpha(x) = \beta(x) = x \]
\[ B(x) = \frac{x}{2} \]

\[ K(x,t) = x \ln \left( \frac{x}{t} \right). \]

In order to simplify the treatment without losing most of the cases of physical interest we assume \( K(x,t) \) is an \( L^2 \) kernel (see [8]).

The solution of this Volterra equation is easily obtained (see [8]). Let

\[ H(x) = \int_0^x K(x,t)R(t)\,dt \]  \hspace{1cm} (2.6)
\[ G(x,t) = G_1(x,t) = K(x,t)F(t) \]  \hspace{1cm} (2.7)
\[ G_j(x,t) = \int_t^x G(x,s)G_{j-1}(s,t)\,ds, \ j \geq 2. \]  \hspace{1cm} (2.8)

Then,

\[ E(x) = H(x) + \sum_{j=1}^{\infty} (-1)^j \int_0^x G_j(x,t)H(t)\,dt \]  \hspace{1cm} (2.9)

Moreover, this infinite series is absolutely convergent. Thus, if we define

\[ G_j(x) = \int_0^x |G_j(x,t)|\,dt \]

then

\[ |E(x)| \leq \left[ 1 + \sum_{j=1}^{\infty} G_j(x) \right]. \sup \{H(t)|; \ 0 \leq t \leq x\}. \]  \hspace{1cm} (2.10)

and the infinite series is absolutely convergent.
In many applications the function $F(t)$ is not known. However bounds for $F(t)$ are known. The discussion given above is then easily modified. Let

$$|F(x)| \leq F$$

Let

$$K_j(x) = \int_0^x K_j(x,t)dt, \quad K_j(x,t) = \int_t^x K(x,s)K_{j-1}(s,t)ds.$$ 

Then

$$K_j(x) > 0$$

and

$$|E(x)| \leq \left[1 + \sum_{j=1}^\infty K_j(x)F^j\right]\sup\{|H(t)|; 0 \leq t \leq x\}.$$ 

Using the definition of $H(t)$ we may rewrite (2.14a) as

$$|E(x)| \leq \left\{\frac{1}{F} \sum_{j=1}^\infty K_j(x)F^j\right\}\sup\{|R(t)|; 0 \leq t \leq x\}.$$ 

The following Lemma gives a useful alternative formula for the quantities $K_j(x)$.

**Lemma 2.1:** Let

$$\tilde{K}_0(x) = 1$$

and, for all $j \geq 1$ define $\tilde{K}_j(x)$ by the recursion

$$\tilde{K}_j(x) = \int_0^x K(x,t)\tilde{K}_{j-1}(t)dt.$$ 

Then from the remark following (2.4) we see that $\tilde{K}_j(x)$ satisfies

$$\begin{cases} (\alpha(\tilde{x})\tilde{K}_j(x))' = \beta(\tilde{x})\tilde{K}_{j-1}(x) \\ \tilde{K}_j(0) = \tilde{K}'_j(0) = 0 \end{cases}$$
Moreover,

\[(2.18) \quad K_j(x) = \tilde{K}_j(x), \quad j \geq 1\]

**Proof:** It is only necessary to establish (2.18). We proceed by induction. The identity is clearly true for \( j = 1 \). Assume that (2.18) holds for \( j = 1, 2, \ldots (J-1) \). Then

\[
\tilde{K}_J(x) = \int_0^x K(x,t) \tilde{K}_{J-1}(t) \, dt = \int_0^x K(x,t) K_{J-1}(t) \, dt
\]

\[
= \int_0^x K(x,t) \int_0^t K_{J-1}(t,s) \, ds \, dt
\]

\[
= \int_0^x ds \int_s^t K(x,t) K_{J-1}(t,s) \, dt
\]

\[
= \int_0^x K_j(x,s) \, ds = K_j(x)
\]

Thus the lemma is proven.

**Remark:** In the special case of the Bessel Operator we can easily verify [7] that

\[(2.19a) \quad K_j(x) = \left(\frac{x}{2}\right)^{2j} \frac{1}{[j!]}\]

and

\[(2.19b) \quad \frac{1}{F} \sum_{j=1}^{\infty} K_j(x) F^j = \frac{I_0(\sqrt{Fx}) - 1}{F}\]

where \( I_0(x) \) is the modified Bessel function of zeroth order (see Abramowitz and Stegun [1]).
3. ESTIMATES: A-PRIORI AND A-POSTERIORI

We are concerned with the functions $Y(x, y_0)$, $(x, y_0)$ which satisfy

$$\begin{align*}
\frac{1}{\beta(x)}(\alpha(x)Y'(x))' + f(x, Y(x)) &= 0, \quad 0 \leq x \leq 1 \\
\frac{1}{\beta(x)}(\alpha(x)\phi'(x))' + f_y(x, Y(x))\phi(x) &= 0, \quad 0 \leq x \leq 1 \\
Y'(0, y_0) &= \phi'(0, y_0) = 0 \\
Y(0, y_0) &= y_0, \quad \phi(0, y_0) = 1
\end{align*}$$

(3.1)

where $f(x, y)$ is a smooth function and

$$|f_y(x, y)| \leq F$$

(3.2)

In those applications where one is really interested in the initial-value problem, one is usually not interested in the function $\phi(x, y_0)$. However, in those applications where one solves the initial value problem as a device to solve a related boundary value problem (i.e. "shooting"), the function $\phi(x, y_0)$ assumes a great deal of importance, for both the differential equation and the numerical computation (see [5],[7]).

In order to obtain an a-priori bound on $Y(x, y_0)$ we let

$$E(x) = Y(x, y_0) - y_0.$$  Then $E(x)$ satisfies the equation

$$\begin{align*}
\frac{1}{\beta(x)}(\alpha(x)E')' + f_y(x, \zeta(x))E(x) &= -f(x, y_0) \\
E(0) &= E'(0) = 0
\end{align*}$$

(3.3)

Applying the results of section 2, we see that

$$|Y(x, y_0) - y_0| \leq \frac{1}{F} \left[ \sum_{j=1}^{\infty} K_j(x) F_j \right] \sup \{|f(t, y_0)|, 0 \leq t \leq x\}.$$  (3.4)

For the sake of definiteness, let us say
(3.5) \[ |Y(x, y_0)| \leq Q, \quad 0 \leq x \leq 1, \]
where \( Q \) is a computable constant.

In order to obtain an a-priori bound on \( \phi(x, y_0) \)
we let

\[ E(x) = \phi(x, y_0) - 1. \]

Then \( E(x) \) satisfies

\[
\begin{cases}
\frac{1}{\alpha(x)}(\alpha(x) E')' + f_y(x, Y(x)) E = -f_y(x, Y(x)) \\
E(0) = E'(0) = 0
\end{cases}
\]

(3.6)

In this case the results of section 2 lead to the estimate

\[
\phi(x, y_0) - 1 \leq \frac{\sum_{j=1}^{\infty} K_j(x) P^j}{1}
\]

(3.7)

For the sake of definiteness, let us say

\[
\phi(x, y_0) \leq L, \quad 0 \leq x \leq 1
\]

(3.8)

where \( L \) is a computable constant.

It often happens that these estimates are rather
large "over-estimates", nevertheless they are useful starting
points for further estimates. In particular cases one can
use special properties of the special function to obtain
better bounds, see [7].

Suppose \( \bar{Y}(x, y_0), \psi(x, y_0) \) are functions which satisfy
the appropriate initial conditions and "approximate"
\( Y(x, y_0), \phi(x, y_0) \). Let

\[
\begin{cases}
E_1(x) = Y(x, y_0) - \bar{Y}(x, y_0) \\
E_2(x) = \phi(x, y_0) - \psi(x, y_0)
\end{cases}
\]

(3.9)
Then these functions satisfy
\[
\begin{align*}
\frac{1}{\beta(x)}(\alpha(x)E_1')' + f_y(x, \xi(x))E_1(x) &= -R_1(\bar{Y}(x, y_0)) \\
\frac{1}{\beta(x)}(\alpha(x)E_2')' + f_y(x, y(x))E_2(x) &= -R_2(\psi(x, y_0), \bar{Y}(x, y_0)) \\
&- f_{yy}(x, \zeta(x))E_1(x)\psi(x, y_0) \\
E_1(0) &= E_2(1) = E_1'(0) = E_2'(0) = 0
\end{align*}
\]

where
\[
\begin{align*}
R_1(a(x)) &= \frac{1}{\beta(x)}(\alpha(x)a')' + f(x, a(x)) \\
R_2(b(x), a(x)) &= \frac{1}{\beta(x)}(\alpha(x)b')' + f_y(x, a(x))b
\end{align*}
\]

Therefore, if we can estimate or bound \( R_1(\bar{Y}(x, y_0)) \), \( R_2(\psi(x, y_0), \bar{Y}(x, y_0)) \) we may use the basic estimates of section 2 to establish a-posteriori error estimates. Of course, if we are not interested in the function \( \Phi \) we may ignore the second equation.
4. CONSTRUCTION OF APPROXIMATING FUNCTIONS

One of the difficulties in implementing the procedure described in the preceding sections is that fact \( \bar{Y}(x,y_0), \psi(x,y_0) \) must be functions which are piecewise in \( C^2[0,1] \). Most computational procedures compute values \( \bar{Y}(x_k,y_0,h), \bar{Y}'(x_k,y_0,h), \psi(x_0,y_0,h), \psi'(x_k,y_0,h) \) which are "approximations" to the functions \( Y(x,y_0), \phi(x,y_0) \) and their derivatives at a discrete set of points

\[
0 = x_0 < x_1 < \cdots < x_N = 1.
\]

However, it is an easy matter to use these values and the differential equations to obtain approximate values for derivatives of any desired order. These values may now be used to construct Hermite approximates, i.e. functions which are piecewise polynomials and satisfy the differential equation at these "knots".

In our work [7] with M. L. Stein and P. R. Stein we used a standard Runge-Kutta procedure and constructed functions which are quintic polynomials in each interval.

On the other hand, if we approach the problem with these estimates in mind, we are led to the construction of certain implicit Runge-Kutta schemes, i.e. collocation schemes (see [2]). For example, let

\[
0 = u_1 < u_2 < \cdots < u_M = 1
\]

be \( M \) distinct points in the closed interval \([0,1]\). Then in each interval \([x_k,x_{k+1}]\) of the partition of \([0,1]\) described by (4.1) we can try to construct polynomials \( \rho_k(x), \psi_k(x) \) of order \( M + 2 \) which satisfy the collocation equations.
\[
\begin{align*}
\frac{\alpha(\xi_{kj})}{\beta(\xi_{kj})} \rho_k''(\xi_{kj}) + \frac{\alpha'(\xi_{kj})}{\beta(\xi_{kj})} \rho_k'(\xi_{kj}) + f(\xi_{kj}, \rho(\xi_{kj})) = 0 \\
\frac{\alpha(\xi_{kj})}{\beta(\xi_{kj})} \psi_k''(\xi_{kj}) + \frac{\alpha'(\xi_{kj})}{\beta(\xi_{kj})} \psi_k'(\xi_{kj}) + f_y(\xi_{kj}, \rho(\xi_{kj}))\psi_k(\xi_{kj}) = 0
\end{align*}
\]

with

\(4.3a\) \( \xi_{kj} = x_k + [x_{k+1} - x_k]u_j \)

As a check on the calculations in [7] John H. Cerutti used this method with \( M = 4 \) and

\[
u_2 = \frac{1}{2} - \frac{\sqrt{5}}{10}, \quad u_3 = \frac{1}{2} + \frac{\sqrt{5}}{10},
\]

i.e., the Lobatto points. Cerutti did four calculations with \( y_0 = 1, 2, 3, 4 \). The results agreed with the Runge-Kutta runs to six significant figures and the residuals were noticeably smaller.
5. **BOUNDING THE RESIDUAL.**

At first glance, the problem of finding realistic bounds on the residual appears trivial. After all, finding the maximum of a real valued function \( R(t) \) on the interval \([0,1]\) is a problem in elementary calculus. After a moments thought, this problem appears impossible. Fortunately, after reconsideration it is indeed possible, and a bit complicated.

For the moment, let \( R(t) \) be a general smooth function defined on the interval \([0,1]\). We will assume we have crude \((O(1))\) bounds on \( R(t) \) and its derivates, say

\[
|R^{(j)}(t)| \leq \rho_j \quad , \quad j = 1, 2, \ldots, M
\]

Let us consider several methods for our problem

**Method 1:** Choose points \( x_k \), \( 0 \leq x_1 < x_2 < \ldots < x_N = 1 \).

Let

\[
\text{MAX} [x_{k+1} - x_k] = h
\]

Then

\[
\text{MAX}|R(t)| \leq \text{MAX}|R(x_k)| + h^2 \rho_2
\]

**Method 2:** Isolate the zeros of \( R'(t) \), i.e., determine disjoint intervals \([\alpha_j, \beta_j]\) which contain all the zeros of \( R'(t) \) in \([0,1]\). Then

\[
\text{MAX}|R(t)| \leq \text{MAX}\{|R(\alpha_j)| + \frac{h_j^2}{8} \rho_2, \quad |R(\beta_j)| + \frac{h_j^2}{8} \rho_2\}
\]

where

\[
h_j = \beta_j - \alpha_j.
\]
Method 3: Sample $R(t)$ to determine (via sign change) that in the subinterval $[x_j, x_{j+1}]$ $R(t)$ has at least \( k \) roots, say $\xi_1, \xi_2, \ldots, \xi_k$, which in general are not known precisely. Then, since

$$R(t) = \left[ \prod_{j=1}^{k} (t-\xi_j) \right] R[\xi_1, \xi_2, \ldots, \xi_k, t]$$

we see that

$$|R(t)| \leq \frac{[x_{j+1}-x_j]^k}{k!} \rho_k , \quad x_j \leq t \leq x_{j+1}$$

This method is particularly applicable in the case where $\bar{Y}(t, y_0)$ is obtained by Collocation. In that case we are assured of at least \( M \) zeros of $R(t)$ in each interval $[x_j, x_{j+1}]$.

There are a variety of other methods which come to mind. The reader is encouraged to think about methods for solving this problem in an efficient manner.

Suppose now that $\bar{Y}(t, y_0), \psi(t, y_0)$ are in fact piecewise polynomials functions, being a polynomial on the subinterval $[x_j, x_{j+1}]$. Then

$$R(t) = R_1(t) + R_2(t)$$

where

$$\begin{cases} R_1(t) = \frac{1}{\beta(t)} [\alpha(t) \bar{Y}(t, y_0)'] \\ R_2(t) = f(t, \bar{Y}(t, y_0)) \end{cases}$$

If we have bounds on the derivatives of

$$\left( \frac{3}{\beta^2} \right)^k f(t, y)$$

we have no difficulty in obtaining crude bounds on the derivatives of $R_2(t)$. Similarly, with some effort we can obtain crude bounds on $R_1(t)$. 
In the special case of the Bessel operator \( (\alpha(x) = \beta(x) = x) \) \( R_1(t) \) is a rational function and it is not too difficult to obtain the necessary crude bounds. For example

\[
\left( \frac{d}{dt} \right)^2 R_1(t) = \frac{t^3 \bar{\gamma}(iv) + t^2 \bar{\gamma}('') - 2t \bar{\gamma}''' + 2\bar{\gamma}'}{t^3}
\]

\[
\left( \frac{d}{dt} \right)^3 R_1(t) = \frac{t^4 \bar{\gamma}(v) + t^3 \bar{\gamma}(iv) - 3t^2 \bar{\gamma}('') + 6t \bar{\gamma}'''}{t^4}
\]

Thus we can determine the zeros of \( R_1'''(t) \). That is, we find disjoint intervals \([\alpha_j, \beta_j] \) which contain the zeros of \( R_1'''(t) \). Then

\[
(5.9a) \quad R_1(iv)(t) = \frac{P(t)}{t^5}, \quad \alpha_j < t < \beta_j
\]

where \( P(t) \) is a polynomial, say

\[
(5.9b) \quad P(t) = \sum_{s=0}^{M} A_s t^5.
\]

Employing the techniques described above, let \( S_j \) be a good bound for \( P(t) \) on \([\alpha_j, \beta_j] \) (obtained on the basis of estimates on \( P'(t), P''(t) \)), then

\[
(5.10) \quad |R_1^{iv}(t)| \leq \frac{S_j}{(\alpha_j)^5}, \quad \alpha_j \leq t \leq \beta_j.
\]
Or, if one is willing to work a bit more, we can proceed as follows. Observe that

\[ R_1^{(v)}(t) = \frac{P_1(t)}{t^6}, \quad \alpha_j < t < \beta_j \]

where \( P_1(t) \) is another polynomial. Thus, by careful root finding one can frequently find (smaller) disjoint intervals \((\alpha_j', \beta_j')\) on which \(|R_1''(t)|\) assumes its maxima and on which \(R_1^{iv}(t)\) is monotone. Hence, crude (but reasonable) bounds on \(|R_1^{iv}(t)|\) can be obtained. Using the methods described above one can obtain crude \([O(1)]\), but reasonable, bounds on \(|R_1''(t)|\). Finally, combining these bounds with \(O(1)\) bounds on \(|R_2''(t)|\), we have bounds on \(|R'(t)|\). Thus, using the methods described above we get good bounds on \(|R(t)|\).
REFERENCES


A-posteriori error estimates are described and methods for obtaining such estimates are discussed for the case of "singular" initial-value problems.
<table>
<thead>
<tr>
<th>KEY WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-posteriori Error Estimates</td>
</tr>
<tr>
<td>Ordinary Differential Equations</td>
</tr>
<tr>
<td>Volterra Equations</td>
</tr>
<tr>
<td>Bounds</td>
</tr>
<tr>
<td>Residual</td>
</tr>
</tbody>
</table>