JOURNAL OF COMPUTATIONAL PHYSICS 69, 252–257 (1987)

Note

Error Estimation for a Class of Differential Eigenproblems

INTRODUCTION

This note first describes a simple derivation of error estimates for a class of 2-point boundary value problems, in terms of two variational integrals. Second, it explores why, for Riccati and Prüfer transformations of a Sturm–Liouville equation, these integrals are needed anyway, so that the error estimates come almost for free.

The idea is simple and may be applicable to higher order systems, but the author has not explored this possibility. The NAG library Sturm–Liouville code DO2KDF uses a version of this method for its error control and has proved robust over some six years of use.

1. THE ESTIMATES

The kind of BVP we consider is a scalar first-order eigenproblem: determine λ such that a solution exists for

$$\phi' = F(x, \phi; \lambda) \qquad \text{on} \quad a < x < b \tag{1a}$$

$$\phi(a) = \alpha, \qquad \phi(b) = \beta. \tag{1b}$$

For local uniqueness of λ it is desirable that $\partial F/\partial\lambda$ be of one sign. For the author's applcations, existence and uniqueness are guaranteed so they will not be discussed further.

We suppose (1a) and (1b) are solved by shooting. That is, for a trial value of λ the DE is integrated from the given values at a, b by an initial value code, towards a matching point $c \in [a, b]$. The difference between the left and right "legs" $\phi_a(x; \lambda)$ and $\phi_b(x; \lambda)$ at $x = c$ defines a miss-distance function $f(\lambda)$, and a rootfinder is used to determine λ as the solution of $f(\lambda) = 0$.

In practice the initial value code commits an error at each integration step so that we obtain a computed miss-distance $\tilde{f}(\lambda)$; also the final $\tilde{\lambda}$ at which the root-

0021-9991/87 $3.00

finding terminates will not generally have $\bar{f}(\bar{\lambda})$ exactly zero. We use the model (e.g. Gear [1, p. 21]) of step-by-step integration, namely, that it can be regarded as following the exact "local solution" of (1a) through the previous point $(x_{i-1}, \bar{\phi}_{i-1})$ up to the current x_i, and then making a jump of size e_i, the local error at x_i, to the computed value $\bar{\phi}_i$. Further we assume that the integrator estimates e_i as part of its step control.

The solution at the last rootfinding iteration can then be regarded as the exact solution of (1b) together with

$$\bar{\phi}'(x) = F(x, \bar{\phi}(x); \bar{\lambda}) + e(x), \qquad a < x < b, \tag{2}$$

where $e(x) = \sum_i e_i \delta(x - x_i)$, and $\delta(x - \xi)$ denotes the Dirac δ-function at $x = \xi$, and the x_i are an enumeration of the integration meshpoints including $x = c$. Each e_i is either a local error (on the left shooting leg) or minus a local error (on the right leg) or formed of two local errors and the final $\bar{f}(\bar{\lambda})$ value (at $x = c$). Errors in the BCs can be modelled by a contribution from $x = a$ and $x = b$.

The exact eigenfunction $\phi(x)$ and eigenvalue λ satisfy (1b) and

$$\phi'(x) = F(x, \phi(x); \lambda), \qquad a < x < b. \tag{3}$$

Subtracting (2) and (3) and applying the mean value theorem gives

$$\begin{aligned}(\bar{\phi} - \phi)'(x) = &\frac{\partial F}{\partial \phi}(x, \psi(x); \mu(x)) \cdot (\bar{\phi} - \phi)(x) \\ &+ \frac{\partial F}{\partial \lambda}(x, \psi(x); \mu(x)) \cdot (\bar{\lambda} - \lambda) + e(x)\end{aligned} \tag{4}$$

where

$$\psi(x) = t_x \bar{\phi}(x) + (1 - t_x)\, \phi(x), \qquad \mu(x) = t_x \bar{\lambda} + (1 - t_x)\, \lambda,$$

and $0 < t_x < 1$. Using the integrating factor

$$\bar{M}(x) = \exp - \left[\int \frac{\partial F}{\partial \phi}(x, \psi(x); \mu(x))\, dx \right] \tag{5}$$

and the boundary conditions we integrate (4) to obtain

$$0 = [(\bar{\phi} - \phi)(x) \cdot \bar{M}(x)]_a^b = \int_a^b \left\{ \frac{\partial F}{\partial \lambda}(x, \psi(x); \mu(x)) \cdot (\bar{\lambda} - \lambda) + e(x) \right\} \bar{M}(x)\, dx$$

so from the δ-function definition of $e(x)$,

$$\bar{\lambda} - \lambda = - \sum_i \bar{M}(x_i)\, e_i \tag{6}$$

where $\bar{M}(x)$, which is arbitrary up to a scalar factor, has been normalized so that

$$\int_a^b \frac{\partial F}{\partial \lambda}(x, \psi(x); \mu(x)) \cdot \bar{M}(x)\, dx = 1. \tag{7}$$

For the informal Sturm–Liouville discussion below we replace (5), (7) by their limits in the case of small errors ($\|e\|_1 = \sum_i |e_i| \to 0$) defining $M(x)$ by

$$M(x) = \exp\left[-\int \frac{\partial F}{\partial \phi}(x, \phi(x); \lambda)\, dx \right] \tag{8}$$

where

$$\int_a^b \frac{\partial F}{\partial \lambda}(x, \phi(x); \lambda) \cdot M(x)\, dx = 1. \tag{9}$$

Computationally, on the other hand, we replace (6) by the estimate

$$\bar{\lambda} - \lambda \simeq \text{errest} \underset{\text{def}}{=} -\sum_i \tilde{M}_i \tilde{e}_i \tag{10}$$

where $\tilde{M}_i = \tilde{M}(x_i)$, $\tilde{M}(x)$ is a numerical version of $\bar{M}(x)$ and $\tilde{e}_i$ is the initial value code's estimate of e_i. In the same way we form an approximate error bound

$$|\bar{\lambda} - \lambda| \lesssim \text{errbnd} \underset{\text{def}}{=} \sum_i \tilde{M}_i \tilde{E}_i \tag{11}$$

where $\tilde{E}_i$ is an estimated bound for $|e_i|$, for instance the local error tolerance used by the code. Suppose that the $\tilde{M}_i$ approximate $\bar{M}(x_i)$ with small relative error. (This amounts to assuming that the eigenfunction is well conditioned. For further discussion of this in the Sturm–Liouville case see the author's paper Pryce [3].) Then if $|e_i| < \tilde{E}_i$ with high probability and the number of terms in the sum (11) is reasonably large, it follows on statistical grounds that the chance of $|\bar{\lambda} - \lambda|$ exceeding errbnd is negligible. Similarly, if $|\tilde{e}_i - e_i|$ is, with high probability, small compared with $\tilde{E}_i$ then almost certainly $|\bar{\lambda} - \lambda - \text{errest}|$ is small compared with errbnd. These assumptions are valid for modern initial-value codes, so (10) and (11) are the basis for a robust error control and estimation process, provided the eigenfunction is well conditioned with respect to the shooting method. This is confirmed by numerical experiments for Sturm–Liouville problems in Pryce [3].

2. Computing the Normalization

The computed $\bar{M}(x)$ can only be normalized after the integration is finished. In fact, we compute scalar multiples $\theta_a \bar{M}(x)$, $\theta_b \bar{M}(x)$ of the true function, on the left

and right shooting legs respectively, and hence "left" and "right" contributions to the normalizing integral (7) and to the sums (10) and (11), which are θ_a and θ_b times their true values. The continuity of $\bar{M}(x)$ at $x=c$ determines the ratio $\theta_a:\theta_b$ and (7) gives a second equation from which θ_a, θ_b are determined. The "left" and "right" sums in (10), (11) are then scaled and added to give errest and errbnd.

Computationally we replace $\psi(x)$, $\mu(x)$ by $\bar{\phi}(x)$, $\bar{\lambda}$ in (5), (7). One method is to adjoin the equations

$$\frac{d}{dx}\log M(x) = -\frac{\partial F}{\partial \phi}(x, \bar{\phi}(x); \bar{\lambda}) \qquad \text{(to compute (5))} \tag{12}$$

$$\frac{dI}{dx} = \frac{\partial F}{\partial \lambda}(x, \bar{\phi}(x); \bar{\lambda})\, M(x) \quad \text{(to compute (7))} \tag{13}$$

to the differential system and solve them with (say) initial values $\log M = I = 0$ at $x=a$ for the left leg ($x=b$ for the right leg). Errors in numerically integrating (12), (13) do not seriously affect the estimation process in the author's experience; more serious are large deviations of $\bar{\phi}$ from ϕ in ill-conditioned cases. Some care is needed to avoid over- or underflow in these computations, which is why it can be important to take $\log M$ rather than M as the dependent variable in (12).

3. The Sturm–Liouville Case

Consider a Sturm–Liouville problem

$$(p(x)\,y')' + (\lambda w(x) - q(x))\,y = 0, \qquad a < x < b \tag{14a}$$

with (for simplicity) regular BCs

$$a_1\,y(a) = a_2\,p(a)\,y'(a), \qquad b_1\,y(b) = b_2\,p(b)\,y'(b). \tag{14b}$$

Various essentially equivalent transformations exist which reduce this to a first order eigenproblem (1a), (1b). Examples are:

(a) Riccati. This changes from the dependent variables y, y' to ϕ, y where $\phi = py'/y$ to obtain the equation pair

$$\phi' = -\frac{\phi^2}{p} - Q = F_1(x, \phi; \lambda) \qquad \text{say;}$$

$$y' = \frac{\phi y}{p};$$

where Q is short for $\lambda w(x) - q(x)$.

(b) Prüfer. This changes from y, y' to r, θ where $py' = r\cos\theta$, $y = r\sin\theta$ to obtain the equation pair

$$\theta' = \frac{1}{2}\left[\frac{1}{p} + Q + \left(\frac{1}{p} - Q\right)\cos 2\theta\right] = F_2(x, \theta; \lambda) \qquad \text{say;}$$

$$r' = \frac{1}{2}\left[\frac{1}{p} - Q\right] r \sin 2\theta.$$

In either case the first equation of the pair, together with BCs derived from (14b), suffices to determine the eigenvalues

The remarkable fact is that in both cases, $M(x)$ as defined by (8) reduces to the square of the second variable, i.e., y^2 for Riccati and r^2 for Prüfer while the normalizing integral in (9) reduces to $\int_a^b y^2 w\, dx$, i.e., to the square of the usual 2-norm for the eigenfunctions. Hence, if we seek to compute the normalized eigenfunction then the error estimation process comes almost for free, the only extra overhead being the accumulation of the sums (10), (11).

We now show that essentially the same is true of any reduction of the Sturm–Liouville problem to the form (1). Since a multiple of a solution of (14a) is a solution, there must be a functional relation between the new variable ϕ, and x, and y'/y. We write this as

$$py'/y = f(x, \phi) \tag{15}$$

for some function f. Direct calculation verifies that ϕ then satisfies the DE

$$\begin{aligned} \phi' &= -f_\phi(x, \phi)^{-1}[Q + f(x, \phi)^2/p + f_x(x, \phi)] \\ &= F(x, \phi; \lambda) \qquad \text{say,} \end{aligned} \tag{16}$$

and that, if $\phi(x)$ is a solution of (16) then

$$-\frac{\partial F}{\partial \lambda} = \frac{d}{dx}\log[y^2 f_\phi(x, \phi)]. \tag{17}$$

Also, the only place where λ enters (16) is in Q, so that

$$\frac{\partial F}{\partial \lambda} = -\frac{w}{f_\phi}. \tag{18}$$

By (17), $M(x)$ is $y^2 f_\phi$ and the normalizing integral (9) reduces to $\int_a^b y^2 w\, dx$ on using this and (18). We deduce that, for any choice (15) of a new variable ϕ, there is a natural choice of a second variable ρ, namely,

$$\rho = cy^2 f_\phi(x, \phi) \qquad \text{for some constant } c. \tag{19}$$

The Sturm–Liouville equation then becomes a first-order pair of the form

$$\begin{aligned} \phi' &= F(x, \phi; \lambda) \\ (\log \rho)' &= -\frac{\partial F}{\partial \phi}(x, \phi; \lambda) \end{aligned} \tag{20}$$

and error estimation is ready to hand if one integrates the two equations (20). As examples:

(a) The *Riccati* method above has $f(x, \phi) = \phi$, hence $f_\phi = 1$ and $\rho = y^2$.

(b) The *Prüfer* method has $f(x, \phi) = \cot \phi$, hence $f_\phi = -\operatorname{cosec}^2 \phi$ and $\rho = -r^2$ (though r^2 would do as well).

(c) The *scaled Prüfer* method used by the author in the NAG library code DO2KDF has the form $f(x, \phi) = S(x) \cot \phi$, where $S(x)$ is a suitable positive scaling function. This also leads to $\rho = -r^2$, where now r is defined by

$$py' = S^{1/2} r \cos \phi, \qquad y = S^{-1/2} r \sin \phi.$$

4. Conclusions and Comments

The method described above has proved very reliable in the NAG routine DO2KDF, which uses the scaled Prüfer transformation, shooting, and an error control based on keeping errbnd below a user-specified tolerance. The a posteriori normalization method described in Section 2 can readily be adapted to multiple shooting, which is appropriate for extremely ill-conditioned Sturm–Liouville problems (e.g., symmetric double potential wells in quantum mechanics). It would be interesting to study the deeper reasons why, for Sturm–Liouville problems, the error estimation comes ready-made in the sense of Section 3, and perhaps thereby to extend the method to linear differential eigenproblems of higher order.

References

1. C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations* (Prentice-Hall, Englewood Cliffs, N.J., 1971).
2. NAG Library routine documents DO2KAF, DO2KDF, DO2KEF (Numerical Algorithms Group, Oxford, U.K., 1978 (Mark 7) onwards).
3. J. D. Pryce, *IMA J. Numer. Anal.* **6**, 103 (1985).

Received January 22, 1986

J. D. Pryce
School of Mathematics,
University of Bristol,
University Walk, Bristol,
BS8 1TW, England