Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

On the Construction of Error Estimators for Implicit Runge--Kutta Methods

J.J.B. de Swart, G. Söderlind

Modelling, Analysis and Simulation (MAS)

MAS-R9704 February 28, 1997

Report MAS-R9704
ISSN 1386-3703

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

On the Construction of Error Estimators for Implicit Runge–Kutta Methods

Jacques de Swart [†] and Gustaf Söderlind [‡]

[†] *CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (Jacques.de.Swart@cwi.nl).*

[‡] *Lund University, Department of Computer Science,*
P.O. Box 118, S-221 00 Lund, Sweden (Gustaf.Soderlind@dna.lth.se).

ABSTRACT

For implicit Runge–Kutta methods intended for stiff ODEs or DAEs, it is often difficult to embed a local error estimating method which gives realistic error estimates for stiff/algebraic components. If the embedded method's stability function is unbounded at $z = \infty$, stiff error components are grossly overestimated. In practice some codes "improve" such inadequate error estimates by premultiplying the estimate by a "filter" matrix which damps or removes the large, stiff error components. Although improving computational performance, this technique is somewhat arbitrary and lacks a sound theoretical backing. In this scientific note we resolve this problem by introducing an *implicit* error estimator. It has the desired properties for stiff/algebraic components without invoking artificial improvements. The error estimator contains a free parameter which determines the magnitude of the error, and we show how this parameter is to be selected on the basis of method properties. The construction principles for the error estimator can be adapted to all implicit Runge–Kutta methods, and a better agreement between actual and estimated errors is achieved, resulting in better performance.

1991 Mathematics Subject Classification: 65L06, 65Gxx, 65L70.
1991 Computing Reviews Classification System: G.1.7
Keywords and Phrases: Ordinary differential equations, Runge–Kutta methods, error analysis, error bounds.
Note: Work carried out under project 'Circuit simulation' (no. MAS2.2).

1. Introduction

We shall consider the problem of estimating the local error in a single step when an implicit Runge–Kutta method (IRK) is applied to a stiff system of ordinary differential equations

$$y' = f(y); \qquad y(0) = y_0, \quad t \geq 0\,, \tag{1.1}$$

where $f : \mathbf{R}^d \to \mathbf{R}^d$. Using standard notation, [HW96b], we write an s–stage IRK (A, b) in the form

$$Y = \mathbb{1} \otimes y_n + h(A \otimes I)F(Y)\,, \tag{1.2}$$

$$y_{n+1} = y_n + h(b^T \otimes I)F(Y)\,, \tag{1.3}$$

where y_n approximates $y(t_n)$. Further, h is the stepsize, Y is the sd–dimensional stage vector whose s component stage vectors Y_i approximate $y(t_n + c_i h)$. The abscissae are defined by $c = A\mathbb{1}$, with $\mathbb{1} = (1, 1, \ldots, 1)^T$. Finally, F stands for the component–wise evaluation of f, i.e.

$$F(Y) = (f(Y_1)^T, f(Y_2)^T, \ldots, f(Y_s)^T)^T.$$

By solving the nonlinear system (1.2) we obtain Y and compute y_{n+1} from (1.3). (In this note we leave the option of solving for the stage derivatives $F(Y)$ aside.)

The primary means to control the accuracy of the computational process is to vary the stepsize. In order to do this we need estimates of the error committed in each individual step, the *local error*. Let $\hat{y}(t; \tau, \eta)$ denote a solution to the differential equation with initial value $y(\tau) = \eta$. Then the local error

in y_{n+1} is $\epsilon = y_{n+1} - \hat{y}(t_{n+1}; t_n, y_n)$. It is estimated by computing a second approximation, $\hat{y}_{n+1}$, to $\hat{y}(t_{n+1}; t_n, y_n)$. In embedded IRK methods, this is obtained by taking another linear combination $\hat{b}$ of the stage derivatives.

2. Error estimation in RADAU5

Because of difficulties in finding $\hat{b}$ such that the order of the error estimate is suitable, one may have to introduce extra parameters. Let us consider the widely used Radau IIa methods, [HW96b, p.123], where the following formula for $\hat{y}_{n+1}$ is used:

$$\hat{y}_{n+1} = y_n + h\left(\hat{b}_0 f(y_n) + (\hat{b}^T \otimes I)F(Y)\right). \tag{2.1}$$

Here $\hat{b}_0$ is a free parameter and $\hat{b}$ is an s–dimensional vector, which is determined such that $\hat{y}_{n+1}$ is of local order $s+1$, i.e., $\hat{b}$ must satisfy the order conditions

$$C\,\hat{b} = (1 - \hat{b}_0, 1/2, 1/3, \ldots, 1/s)^T.$$

The $s \times s$ matrix C has entries $c_{ij} = c_j^{i-1}$. Note that putting $\hat{b}_0 = 0$ in (2.1) would by the order condition lead to the same formula as (1.3); hence $\hat{b}_0 \neq 0$. Therefore at least one extra parameter is necessary to obtain a nonzero error estimate.

The estimate ϵ is now computed as

$$\epsilon = y_{n+1} - \hat{y}_{n+1}\,, \tag{2.2}$$

and y_{n+1} is accepted as an approximation to $y(t_{n+1})$ if $\|\epsilon\|$ is less than the specified tolerance. As ϵ is dependent on the stepsize, its ratio to the tolerance is also used to compute the next stepsize.

Most IRKs are constructed in such a way that they are at least A–stable. However, the reference formula (2.1) is normally not A–stable. Consequently, $\|\epsilon\|$ can be very large due to large stiff error components. In practice this is typically the case, since IRK methods are indeed intended to solve stiff problems or DAEs.

In RADAU5, [HW96a], which is an implementation of the 3–stage Radau IIa method, Hairer and Wanner use the following remedy, [HW96b, p. 123], which is attributed to Shampine [SB84]. A modified error estimate $\hat{\epsilon}$ is constructed from

$$\hat{\epsilon} = (I - \gamma h J)^{-1}\epsilon\,, \tag{2.3}$$

in which $\hat{y}_{n+1}$ is computed from (2.1) with $\hat{b}_0 = \gamma$, the single real eigenvalue of A. The matrix $(I - \gamma h J)^{-1}$ is then available and factorized from the Newton iteration used to solve (1.2). To see the effect of this transformation, consider the test equation $y' = \lambda y$; we now have $\hat{\epsilon} \to -1$ as $h\lambda \to \infty$, as opposed to $\epsilon \to \infty$. The purpose of the premultiplication by $(I - \gamma h J)^{-1}$ is thus to keep the error estimate bounded also for large values of h by filtering out stiff error components.

3. Case study: The implicit Euler method

The filtering technique has also been used in other contexts where it has a theoretical foundation in terms of the map from a residual to the corresponding error. In the context above, however, it is a trick—albeit a necessary one—in order to restore the full potential of the Radau IIa method.

In order to see where and how the filtering is justified, we consider the simplest Radau IIa method, i.e. the implicit Euler method

$$y_{n+1} = y_n + h f(y_{n+1})\,. \tag{3.1}$$

If we insert the local solution $\hat{y}(t; t_n, y_n)$ into this discretization, there results a defect, or *local residual* δ:

$$\hat{y}(t_{n+1}) = y_n + hf(\hat{y}(t_{n+1})) - \delta\,. \tag{3.2}$$

We find the *local error* $\epsilon = y_{n+1} - \hat{y}(t_{n+1}; t_n, y_n)$ by subtracting (3.2) from (3.1) and obtain an algebraic relation between the residual and the error:

$$\epsilon = hf(\hat{y}(t_{n+1}) + \epsilon) - hf(\hat{y}(t_{n+1})) + \delta\,. \tag{3.3}$$

Linearizing and solving for ϵ we obtain the error/residual relation,

$$\epsilon = (I - hJ)^{-1}\delta\,. \tag{3.4}$$

This equation is the mathematical justification of "filtering". As is well–known, there is an important conceptual as well as numerical difference between a residual and its corresponding error—the defect and error are elements of different spaces. Although this equation is well established, [HNW93, p. 369], it is frequently overlooked. The reason seems to be an overemphasis on asymptotics; as $hJ \to 0$ we have $\epsilon \approx \delta$, i.e. in the nonstiff case it does not matter if one estimates ϵ or δ, but in the stiff case the difference is known to be very significant. This observation has led to the view that a "poor" error estimate can be improved by the premultiplication of a filtering matrix. Even if this works in practice, such arbitrariness in error estimation ought to be replaced by a search for qualitatively correct error estimates. Note that in embedded IRK methods, filtering is in principle *never* justified since one normally estimates a local error, never a local residual. The situation may, however, be different for defect estimation.

In the next section we suggest an error estimate which has an inherent damping of stiff error components as a design criterion. No extra filtering is required or permitted (as it cannot be justified). As a starting point we note that the poor asymptotic behavior of ϵ as defined by (2.1) is caused by (2.1) being essentially an explicit formula. Thus, $\hat{y}_{n+1}$ is computed from old data, the stage derivatives *and the explicitly calculated* $hf(y_n)$. This turns the error estimator formula effectively into an explicit method, and consequently all hopes for a proper behavior for large values of h are in vain.

4. An implicit error estimate

Instead of (2.1) we propose to use an implicit reference formula of the structure

$$\hat{y}_{n+1} = y_n + h\left(\hat{b}_0 f(y_n) + (\hat{b}^T \otimes I)F(Y) + \gamma f(\hat{y}_{n+1})\right), \tag{4.1}$$

where γ is such that $(I - \gamma hJ)^{-1}$ is available from the (transformed) Newton process used to solve for Y from (1.2). Solving $\hat{y}_{n+1}$ from (4.1) by a modified Newton process leads to the recursion,

$$\begin{aligned} r^{(j)} &= \hat{y}^{(j)}_{n+1} - y_n - h\left(\hat{b}_0 f(y_n) + (\hat{b}^T \otimes I)F(Y) + \gamma f(\hat{y}^{(j)}_{n+1})\right), \\ \hat{y}^{(j+1)}_{n+1} &= \hat{y}^{(j)}_{n+1} - (I - \gamma hJ)^{-1} r^{(j)}\,. \end{aligned}$$

The natural starting value is $\hat{y}^{(0)}_{n+1} = y_{n+1}$. Since we are computing an error estimate we do not need high accuracy and may consider the first Newton iterate $\hat{y}^{(1)}_{n+1}$ as *the* reference formula itself. This yields

$$\hat{y}^{(1)}_{n+1} = y_{n+1} + h(I - \gamma hJ)^{-1}\left(((\hat{b}^T - b^T) \otimes I)F(Y) + \hat{b}_0 f(y_n) + \gamma f(y_{n+1})\right)\,.$$

In this formula, we determine $\hat{b}$ such that $\hat{y}^{(1)}_{n+1}$ is of local order $s+1$, which means that we require

$$C\,\hat{b} = (1 - \hat{b}_0, 1/2, 1/3, \ldots, 1/s)^T - \gamma \mathbb{1}\,. \tag{4.2}$$

The parameter $\hat{b}_0$ is free but required to be nonzero as taking $\hat{b}_0 = 0$ yields $\hat{y}_{n+1} \equiv y_{n+1}$. For methods with $c_s = 1$ we have $C^{-1}\mathbb{1} = e_s$ and

$$\hat{y}^{(1)}_{n+1} = y_{n+1} + h(I - \gamma h J)^{-1} \left((-\hat{b}_0 e_1^T C^{-T} \otimes I) F(Y) + \hat{b}_0 f(y_n) \right) . \tag{4.3}$$

Consequently, the *error estimator formula*

$$\begin{aligned} \epsilon &= y_{n+1} - \hat{y}^{(1)}_{n+1} \\ &= \hat{b}_0 h (I - \gamma h J)^{-1} \left((e_1^T C^{-T} \otimes I) F(Y) - f(y_n) \right) \end{aligned} \tag{4.4}$$

becomes a homogeneous function of $\hat{b}_0$. In other words, the choice of $\hat{b}_0$ determines the magnitude of the error estimate.

In general, we define the error estimator formula by

$$\begin{aligned} \epsilon &= y_{n+1} - \hat{y}^{(1)}_{n+1} \\ &= h(I - \gamma h J)^{-1} \left(((b^T - \hat{b}^T) \otimes I) F(Y) - \hat{b}_0 f(y_n) - \gamma f(y_{n+1}) \right) . \end{aligned} \tag{4.5}$$

Now consider the test equation $y' = \lambda y$, for which

$$\hat{y}^{(1)}_{n+1} = \hat{R}^{(1)}(z) y_n , \qquad z := h\lambda .$$

The value of $\hat{R}^{(1)}(\infty)$ of the reference formula is known to be of relevance to the size of the estimated error in the stiff components. Although this is not a matter of stability, it is desirable that $\hat{R}^{(1)}(\infty)$ is fairly small. A straightforward derivation yields

$$\hat{R}^{(1)}(z) = R(z) + \frac{z}{1 - \gamma z} \left((\hat{b}^T - b^T)(I - zA)^{-1} \mathbb{1} + \hat{b}_0 + \gamma R(z) \right) , \tag{4.6}$$

where $R(z)$ is the stability function of the implicit Runge–Kutta method. If A is nonsingular, we thus obtain

$$\lim_{z \to \infty} \hat{R}^{(1)}(z) = -\frac{\hat{b}_0}{\gamma} . \tag{4.7}$$

Thus the stiff error components are damped if $|\hat{b}_0/\gamma| < 1$. This damping is desirable as the error estimator will "see" a stiff error component from the previous step's iteration error multiplied by $|\hat{b}_0/\gamma|$.

For an s–stage Radau IIa method one can easily give an explicit formula for our new error estimator. If we write $R(z) = P_R(z)/Q_R(z)$ and normalize P_R and Q_R such that $Q_R(z)$ is monic (e.g. for $s = 3$ we have $Q_R(z) = z^3 - 9z^2 + 36z - 60$), then by (4.6), $R(z) - \hat{R}^{(1)}(z)$ is a rational function with denominator $(1 - \gamma z) Q_R(z)$. Thus the degree of the denominator is $s + 1$. By (4.7) the numerator then has degree at most $s+1$. As the local order of the error estimator is $s+1$, however, the numerator only contains a single power of z, viz. z^{s+1}. It follows that

$$R(z) - \hat{R}^{(1)}(z) = - \frac{\hat{b}_0 z^{s+1}}{(1 - \gamma z) Q_R(z)} . \tag{4.8}$$

For the 3–stage Radau IIa, $R(z) - \hat{R}^{(1)}(z)$ thus has a four–fold zero at $z = 0$ and the same poles as $R(z)$ with the exception that $z = 1/\gamma$ is a double pole.

Remarks. Note that if $b^T = e_s^T A$, where e_s is the s^{th} canonical basis vector of $\mathbf{R}^s$, then (4.5) and (2.3) differ only by a factor $\hat{b}_0/\gamma$. The condition $b^T = e_s^T A$ ("stiff accuracy"), holds for all Radau IIa methods as well as for the Lobatto IIIa and IIIc methods. For these methods our implicit error estimator justifies filtering by providing an estimate with the same effect. For other methods, however, one must be more careful. Thus e.g., it is incorrect to use filtering for the implicit midpoint method, which is a Gauss method, but harmless to use it for the trapezoidal rule, which falls into the Lobatto IIIa category. In order to avoid mistakes, we suggest that the construction of implicit estimators is

considered to be the normal route instead of filtering. Finally we remark that even in cases when the error estimator formula is not a homogeneous function of $\hat{b}_0$, we may select the magnitude of the error estimator with a multiplicative factor; we may consider $E(z) = \theta(R(z) - R^{(1)}(z))$ as the error estimator, where the parameter θ is to be carefully determined so that the estimator gives a proper approximation to the actual error. This technique may be of particular importance for DAEs.

5. Choosing $\hat{b}_0$

We shall finally discuss the choice of the free parameter $\hat{b}_0$, and limit ourselves to methods with $c_s = 1$ such as Radau IIa methods. Specifically, we will motivate a suitable choice of $\hat{b}_0$ for the 3–stage, 5th order Radau IIa method. For such methods the new error estimator is a homogeneous function of $\hat{b}_0$, i.e. $\hat{b}_0$ determines the magnitude of the error estimate.

We argue that the most important design goal is that the error estimator does not significantly underestimate the error. On the other hand, a too large value of $\hat{b}_0$ will degrade performance. A small value is also desirable to reduce $\hat{R}^{(1)}(\infty)$. To find a suitable value, we model the error of the Radau IIa method by first considering the linear test equation with $z = h\lambda$. The actual relative local error, $|R(z) - \mathrm{e}^z|$, is investigated on two domains: $\mathcal{A}$, where the method operates in its asymptotic regime and relative accuracy is high, and $\mathcal{B}$, where the method is able to yield accurate results. $\mathcal{B}$ is considerably larger than $\mathcal{A}$.

Obviously $\mathcal{A}$ and $\mathcal{B}$ must contain a neighbourhood of the origin. We take $\mathcal{A}$ to be a disk of radius ρ,

$$\mathcal{A}(\rho) = \{z \in \mathbf{C} \,:\, |z| \leq \rho\}.$$

The selection of the radius is based on several criteria. First, $\mathcal{A}(\rho)$ must exclude the poles of $R(z)$ which for the 3–stage Radau IIa are located at $1/\gamma \approx 3.6378$ and $2.6811 \pm 3.0504\,\mathrm{i}$, respectively. By (4.8), the poles of the reference formula $\hat{R}^{(1)}(z)$ are then also excluded. Furthermore, $\mathcal{A}(\rho)$ should cover the central portion of the order star of the method, [IN91, p. 7], as this corresponds to the domain of high relative accuracy. Last, the intersection with the imaginary axis is an important criterion of relevance for oscillatory systems. To resolve an angular frequency of ω, the stepsize must satisfy $h\omega < \pi$ by the sampling theorem. In practice, however, the numerical method is unable to accurately resolve this frequency with stepsizes exceeding $h\omega = \pi/2$. Based on these considerations, we have taken $\rho = \pi/2$. The selected asymptotic domain $\mathcal{A}(\pi/2)$ meets all the criteria above.

$\mathcal{B}(\rho)$ should contain $\mathcal{A}(\rho)$ as well as a large portion of the negative halfplane. Again, high frequencies cannot be resolved, but $\mathcal{B}(\rho)$ should cover the negative real axis if the method—like the Radau IIa—is able to produce accurate solutions there. We have chosen to consider the parabolic domain

$$\mathcal{B}(\rho) = \{z = x + \mathrm{i}\omega \,:\, x \leq (\rho - \omega)(\rho + \omega)/\rho\},$$

and $\mathcal{A}(\pi/2)$, $\mathcal{B}(\pi/2)$ and the order star of $R(z)$ are plotted in Figure 1. The Radau IIa method is able to provide reasonable accuracy inside $\mathcal{B}(\pi/2)$. The method is still of use in large portions of the complex plane outside $\mathcal{B}(\pi/2)$, e.g. in all of $\mathbf{C}^-$; A–stability implies that $|R(z)| \leq 1$ on $\mathbf{C}^-$ just like $|\mathrm{e}^z| \leq 1$, even if the relative local error $|R(z) - \mathrm{e}^z|$ cannot be considered to be "small" on all of $\mathbf{C}^-$.

As $R(z) - \mathrm{e}^z$ is an analytic function in the domain of accuracy $\mathcal{B}(\pi/2)$, $\max|R(z) - \mathrm{e}^z|$ is attained on $\partial\mathcal{B}(\pi/2)$ by virtue of the maximum modulus theorem. Thus we find that $\max|R(z) - \mathrm{e}^z| = 0.067$ in $\mathcal{B}(\pi/2)$, and we may choose $\hat{b}_0$ (i.e. the magnitude of the error estimator) so that $\max|R(z) - \hat{R}^{(1)}(z)|$ comes close to the maximum of the actual error. This suggests choosing $\hat{b}_0/\gamma = 0.067$, or $\hat{b}_0 \approx 0.018$, and in Figure 2 (left), we plot $|R(z) - \mathrm{e}^z|$ and the error estimator $|R(z) - \hat{R}^{(1)}(z)|$ on $\partial\mathcal{B}(\pi/2)$ for $\hat{b}_0 = 0.02$. Beacuse the maxima may not occur at the same points, we verify in Figure 2 (right) that the error estimator with its chosen magnitude does not exhibit any significant underestimation of the error on the negative real axis.

To investigate the new estimator in the asymptotic regime, we have plotted $|R(z) - \mathrm{e}^z|$ and $|R(z) - \hat{R}^{(1)}(z)|$ on $\partial\mathcal{A}(\pi/2)$ in Figure 3 (left), showing that their magnitudes are similar there.

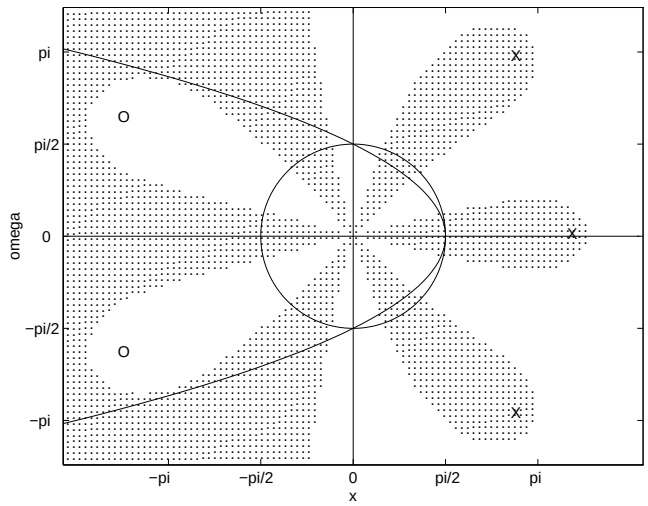

Figure 1: Plots of $\partial\mathcal{A}(\pi/2)$ and $\partial\mathcal{B}(\pi/2)$, together with the order star of the 3–stage Radau IIa method. Poles and zeros of $R(z)$ are denoted by $\times$ and $\circ$, respectively.

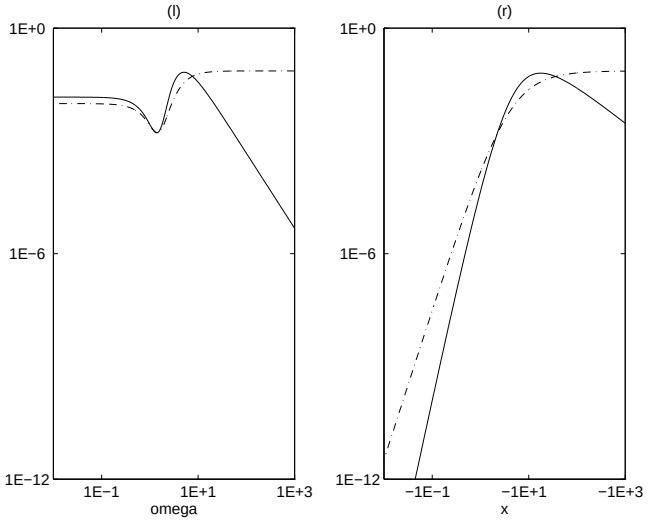

Figure 2: Plots of relative error $|R(z) - \mathrm{e}^z|$ (solid) and error estimator $|R(z) - \hat{R}^{(1)}(z)|$ (dash–dotted) on $\partial\mathcal{B}(\pi/2)$ for $10^{-2} \le \omega \le 10^3$ (left) and on the negative real axis for $-10^3 \le x \le -10^{-2}$ (right). The plot on the right clearly shows slopes of 6 and 4 when $|x| < 1$, and a slight underestimation of the error near $x = -10$. The plots were obtained using $\hat{b}_0 = 0.02$ in order to match the levels of the error and its estimate.

Note that because the error estimator has lower order than the method, it is still likely to significantly overestimate the error at sharp tolerances. This is seen in Figure 3 (right), where we study the ratio

$$K(z) = \frac{R(z) - \mathrm{e}^z}{R(z) - \hat{R}^{(1)}(z)} \tag{5.1}$$

and have plotted

$$k(\rho) = \max_{|z| \le \rho} |K(z)| \tag{5.2}$$

for $0 < \rho \le \pi/2$. The plot suggests that the error estimator underestimates the error outside $\mathcal{A}(1.3)$. This underestimation is benign, however, as verified by Figure 4, which shows the level curves $|K(z)| = \kappa$ for $\kappa = 0.2(0.2)1.2$. Thus, the underestimation occurs only in the right half–plane for $|z| > 1.3$, where, in the absence of dissipativity, the method is less likely to proceed with large steps.

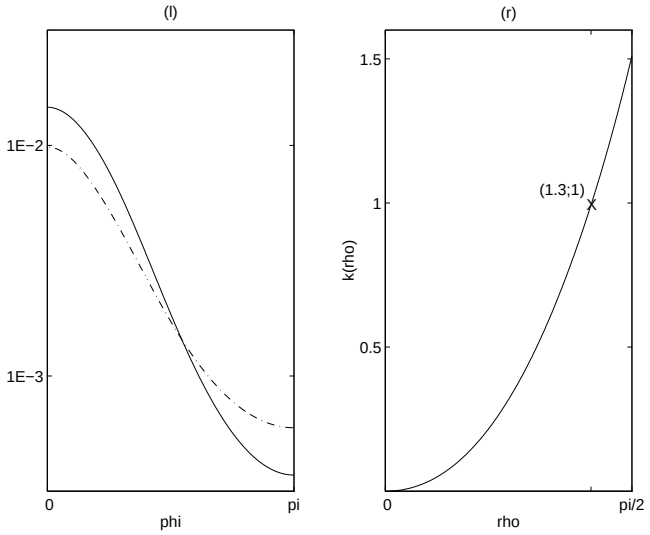

Figure 3: Error $|R(z) - \mathrm{e}^z|$ (solid) and error estimator $|R(z) - \hat{R}^{(1)}(z)|$ (dash-dotted) for $\hat{b}_0 = 0.02$ on $\partial\mathcal{A}(\pi/2) = \{z = \pi \mathrm{e}^{\mathrm{i}\varphi}/2;\, 0 \le \varphi \le \pi\}$ (left), and the maximum ratio $k(\rho) = \max |(R(z) - \mathrm{e}^z)/(R(z) - \hat{R}^{(1)}(z))|$ for $|z| \le \rho$ and $0 \le \rho \le \pi/2$ (right).

Let us now consider linear constant coefficient systems

$$y' = Jy$$

solved with the method pair $(R, \hat{R}^{(1)})$. Because the error estimator is a rational function analytic in $\mathcal{A}(\rho)$, it follows from the spectral theorem, and the maximum modulus theorem, that the estimated relative error in the system is bounded by

$$\|R(hJ) - \hat{R}^{(1)}(hJ)\|_2 \le \max_{\partial\mathcal{A}(\rho)} |R(z) - \hat{R}^{(1)}(z)| = \Delta(\rho)$$

for all matrices J with $\|hJ\|_2 \le \rho$. Since the error estimator has local order 4, we have $\Delta(\rho) = \mathcal{O}(\rho^4)$. ¿From Figure 4 we see that the estimated error exceeds the actual error on $\mathcal{A}(1.3)$, and in the following we may therefore take $0 < \rho < 1.3$. By formally approximating the matrix exponential e^{hJ} by a polynomial $P_{\exp}(hJ)$ such that $\|\mathrm{e}^{hJ} - P_{\exp}(hJ)\|_2 \le \delta$ on $\mathcal{A}(\rho)$, it follows that the actual relative error in the system is bounded by

$$\begin{aligned} \|R(hJ) - \mathrm{e}^{hJ}\|_2 &\le \|R(hJ) - P_{\exp}(hJ)\|_2 + \delta \\ &\le \max_{|z| \le \rho} |R(z) - P_{\exp}(z)| + \delta \end{aligned}$$

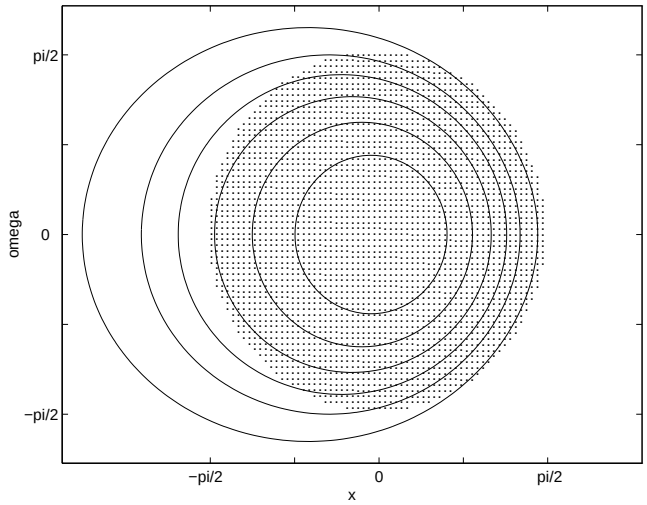

Figure 4: Level contours of $|R(z) - \mathrm{e}^z|/|R(z) - \hat{R}^{(1)}(z)| = \kappa$ for $\kappa = 0.2(0.2)1.2$. The shaded area is $\mathcal{A}(\pi/2)$. The plot was obtained for $\hat{b}_0 = 0.02$.

$$\begin{aligned} &\leq \max_{|z|\leq\rho} |R(z) - \mathrm{e}^z| + 2\delta \\ &\leq \max_{|z|\leq\rho} |R(z) - \hat{R}^{(1)}(z)| + 2\delta \\ &= \Delta(\rho) + 2\delta \end{aligned}$$

for all J with $\|hJ\|_2 \leq \rho$. Note that δ can be made arbitrarily small. Thus we have a bound on the actual error in linear systems, in terms of the error estimator, uniform with respect to the conditioning of J.

It is also of interest to bound the actual error directly in terms of the estimated error, i.e. we would like to find a constant $C(\rho) < 1$ such that for all vectors y,

$$\|(R(hJ) - \mathrm{e}^{hJ})y\|_2 \leq C(\rho)\|(R(hJ) - \hat{R}^{(1)}(hJ))y\|_2. \tag{5.3}$$

This can be obtained in a similar manner. By (4.8) and (5.1),

$$K(z) = \frac{R(z) - \mathrm{e}^z}{R(z) - \hat{R}^{(1)}(z)} = \frac{(\gamma z - 1)(P_R(z) - Q_R(z)\mathrm{e}^z)}{\hat{b}_0 z^4}.$$

Note that $P_R(z) - Q_R(z)\mathrm{e}^z = Q_R(z)\mathcal{O}(z^6)$, hence

$$K(z) = \frac{(\gamma z - 1)Q_R(z)}{\hat{b}_0}\,\mathcal{O}(z^2)$$

because of the pole–zero cancellation at the origin. Thus $K(z)$ is regular in $\mathcal{A}(\rho)$ with a double zero at the origin; this is also clearly seen in Figure 3 (right). It follows from (5.3) by the pole–zero cancellation that

$$C(\rho) = \sup_{\|hJ\|_2\leq\rho} \|K(hJ)\|_2.$$

Now, in order to apply the spectral theorem, we again approximate e^{hJ} by $P_{\exp}(hJ)$ and consider instead

$$\tilde{K}(z) = \frac{(\gamma z - 1)(P_R(z) - Q_R(z)P_{\exp}(z))}{\hat{b}_0 z^4}.$$

By taking the degree of $P_{\exp}(z)$ suitably high, we have $\|\tilde{K}(hJ) - K(hJ)\|_2 \leq \delta$, therefore

$$\begin{aligned} \|K(hJ)\|_2 &\leq \|\tilde{K}(hJ)\|_2 + \delta \\ &\leq \max_{|z|\leq\rho} |\tilde{K}(z)| + \delta \\ &\leq \max_{|z|\leq\rho} |K(z)| + 2\delta \\ &= k(\rho) + 2\delta \end{aligned}$$

for all J with $\|hJ\|_2 \leq \rho$. Thus, the actual error is never underestimated on $\mathcal{A}(1.3)$ for linear constant coefficient systems.

We finally remark that the latter result depends on the pole–zero cancellation at the origin. This implies that the result is not valid for more general classes of problems. This comes as no surprise, however, as the error estimator does not contain the same elementary differentials as the actual error; it is therefore not possible to prove that the error estimator is an upper bound for the error in general nonlinear problems.

Concluding remarks. Since the RADAU5 code uses $\hat{b}_0 = \gamma$, [HW96a], our new estimator with $\hat{b}_0 = 0.02$ has approximately 14 times smaller magnitude without significant underestimation of the error. This leads to approximately 70% larger steps, a better agreement between requested and achieved accuracy, and, for a given tolerance, improved performance.

The design process above has also been used in the code PSIDE, [SLV97], which is based on the 4–stage Radau IIa method, and obtained $\hat{b}_0 = 0.01$. Practical experience with these error estimators is affirmative, although extensive testing must be reported elsewhere.

Acknowledgements

JdS is supported via the Dutch Technology Foundation (STW, grant CWI22.2703). GS is supported by the Swedish Research Council for Engineering Sciences (TFR, grant 222–95–546), which also supported JdS's visit to Lund University in November 1996, when this research was initiated. The authors would also like to acknowledge stimulating discussions with Claus Bendtsen, Claus Führer and Hans Olsson.

References

[HW96a] E. Hairer and G. Wanner. *RADAU5*, July 1996. Available via `WWW` at `URL` `ftp://ftp.unige.ch/pub/doc/math/stiff/radau5.f`.

[HNW93] E. Hairer, S.P. Nørsett and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems.* Springer-Verlag, second revised edition, 1993.

[HW96b] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-algebraic Problems.* Springer-Verlag, second revised edition, 1996.

[IN91] A. Iserles and S.P. Nørsett. *Order Stars.* Chapman & Hall, 1991.

[SB84] L.F. Shampine and L. Baca. *Error estimators for stiff differential equations.* J. Comp. and Appl. Math. 11 (1984), 197–207.

[SLV97] J. J. B. de Swart, W. M. Lioen, and W. A van der Veen. Specification of PSIDE. *In preparation*, 1997.