



331602 — SIAM — SINUM 26/5 — Batch 602 — K36/Pierrette — Galley 1

SINUM 602

SIAM J. NUMER. ANAL.

Vol. 26, No. 5, pp. 000-000, October 1989

© 1989 Society for Industrial and Applied Mathematics

000

## ROBUST DEFECT CONTROL WITH RUNGE-KUTTA SCHEMES\*

DESMOND J. HIGHAM†

**Abstract.** Enright [Numerical Analysis Report 122, University of Manchester, Manchester, U.K., 1986] implements a Runge-Kutta method for solving the initial value problem using an alternative to the standard local error control scheme. The aim is to control the defect associated with a local interpolant by sampling its value at one or more fixed points within each step. However, in general, the quality of a sample point is problem-dependent and also varies from step to step. Two classes of interpolant are presented for which the asymptotic behaviour of the defect is known a priori, allowing optimal sample points to be chosen.

**Key words.** Runge-Kutta formula, defect, interpolation

**AMS(MOS) subject classification.** 65L05

**1. Introduction.** We consider the numerical solution of a nonstiff system of ordinary differential equations

$$(1.1) \quad \begin{aligned} y'(x) &= f(x, y(x)), & y(x_0) &= y_0, \\ f: \mathbf{R} \times \mathbf{R}^N &\rightarrow \mathbf{R}^N, \end{aligned}$$

using an explicit Runge-Kutta method. Given  $y_n \approx y(x_n)$ , such a method takes a step of length  $h$  ( $=h_n$ ) and produces an approximation  $y_{n+1}$  to  $y(x_n+h)$ . Most popular codes attempt to control the local error  $u(x_n+h) - y_{n+1}$ , where the local solution  $u(x)$  satisfies  $u'(x) = f(x, u(x))$  and  $u(x_n) = y_n$ . Recently several authors have shown that it is possible, at reasonable cost, to produce a function  $p(x)$  that approximates the local solution over the entire step [1], [4], [6], [8]-[10]. It is then natural to consider the defect of  $p(x)$ ,

$$(1.2) \quad \delta(x) := p'(x) - f(x, p(x)),$$

that is, the amount by which  $p(x)$  fails to satisfy the differential equations. Enright [2] suggests that it may be preferable to control the defect on each step rather than the local error. This approach is shown to offer several advantages, particularly from the user's point of view. Further analysis of defect control strategies is given in [3]. We are concerned here with the problem of reliably estimating the defect.

One of the defect control schemes proposed in [2] is motivated by the desire to bound  $\max_{\tau \in [0,1]} \|\delta(x_n + \tau h)\|_\infty$ . The defect is sampled at the points  $\{x_n + \tau_i^* h\}_{i=1}^k$ , where the  $\{\tau_i^*\}_{i=1}^k$  are fixed in  $(0, 1)$ , and the step is accepted if and only if  $\max_{1 \leq i \leq k} \|\delta(x_n + \tau_i^* h)\|_\infty < \text{TOL}$ , for some user-supplied parameter TOL. The simplest case,  $k=1$ , is implemented in [2], and an asymptotic expansion of  $\delta(x)$  is used there to gain insight into the performance of this type of scheme. We now consider this expansion.

To begin, we suppose that the approximation  $p(x)$  has local order  $l+1$ , that is,

$$(1.3) \quad p(x) - u(x) = O(h^{l+1}).$$

\* Received by the editors December 23, 1987; accepted for publication (in revised form) October 14, 1988.

† Department of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom. Present address, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4. This work was supported by a Natural Sciences and Engineering Research Council Research Studentship.

12001  
12002

2

DESMOND J. HIGHAM

12006  
12007  
12008

Assuming that  $f$  satisfies a Lipschitz condition on  $[x_n, x_n + H]$ , where  $h < H$ , we then have

12009  
12010  
12011

$$(1.4) \quad \begin{aligned} \delta(x) &= p'(x) - u'(x) + f(x, u(x)) - f(x, p(x)) \\ &= p'(x) - u'(x) + O(h^{l+1}). \end{aligned}$$

12012  
12013

Using local interpolants derived in [4] for the function  $p(x)$ , Enright notes that for sufficiently smooth  $f$ , (1.4) may be written in the form

12014

$$(1.5) \quad \delta(x_n + \tau h) = h^l \sum_{j=1}^{m_l} q_j'(\tau) F_j + O(h^{l+1}).$$

12015  
12016  
12017  
12018  
12019  
12020  
12021  
12022  
12023

Here  $F_j$  is an elementary differential that depends only on  $f$ ,  $x_n$ , and  $y_n$ , and  $q_j'(\tau)$  is a polynomial in  $\tau$  whose coefficients depend only on the Runge–Kutta interpolation scheme. Enright recommends that for a particular scheme a sample point  $\tau^*$  should be chosen so that each  $|q_j'(\tau^*)|$  is relatively large. This means that if one of the  $F_j$ 's is dominant, then it will be allowed to make a significant contribution to the leading term. However, he points out that for any fixed  $\tau^*$  there always exists the possibility of cancellation in the sum  $\sum_{j=1}^{m_l} q_j'(\tau^*) F_j$ , caused by the problem-dependent  $F_j$ 's. Hence the size of the defect at  $x_n + \tau^* h$  can be an arbitrarily poor indication of its maximum value over  $[x_n, x_n + h]$ .

12024  
12025  
12026  
12027

In this paper we introduce some alternative interpolants for use with a defect control scheme. For these interpolants the associated defect has the form (1.5), but the important additional feature is that each  $q_j'(\tau)$  is a multiple of a known polynomial  $\Phi(\tau)$  so that

12028

$$(1.6) \quad \delta(x_n + \tau h) = h^l \Phi(\tau) K + O(h^{l+1})$$

12029  
12030  
12031  
12032

where  $K$  is independent of  $\tau$  and  $h$ . (In fact we shall prove this result directly by interpolation theory rather than by using (1.5).) It follows that a sample point is available that is asymptotically optimal for any problem, namely, a  $\tau^*$  that maximises  $|\Phi(\tau)|$  over  $[0, 1]$ .

12033  
12034  
12035  
12036

In the next section we present two classes of interpolant that have this property and give some specific examples. The corresponding defect control schemes incur a higher cost per step than those of [2] and [3]. This is discussed in § 3. The final section describes the results of some numerical experiments that support the theory.

12037  
12038  
12039

**2. The interpolants.** The interpolants considered below fall within the framework of Shampine [9] and Gladwell et al. [6]. We suppose that there are distinct points  $\{\xi_i\}_{i=1}^m$  in  $[x_n, x_n + h]$  with corresponding approximations  $u_i \in \mathbf{R}^N$  satisfying

12040

$$u_i - u(\xi_i) = O(h^{q+1}), \quad i = 1, \dots, m,$$

12041

and that  $u_i' = f(\xi_i, u_i) \in \mathbf{R}^N$  is available for  $i = 1, \dots, r \leq m$ , whence

12042

$$u_i' - u'(\xi_i) = O(h^{q+1}), \quad i = 1, \dots, r$$

12043  
12044  
12045

for Lipschitzian  $f$ . The data  $\{u_i\}_{i=1}^m$  and  $\{u_i'\}_{i=1}^r$  is said to be of local order  $q+1$ . We then take  $p(x): \mathbf{R} \rightarrow \mathbf{R}^N$  to be the unique Hermite interpolating polynomial of degree less than or equal to  $m+r-1$  that satisfies

12046  
12047  
12048

$$(2.1) \quad \begin{aligned} p(\xi_i) &= u_i, & i &= 1, \dots, m, \\ p'(\xi_i) &= u_i', & i &= 1, \dots, r. \end{aligned}$$

12049  
12050  
12051

We insist that  $\xi_1 = x_n$ ,  $u_1 = y_n$ ,  $\xi_r = x_n + h$ , and  $u_r = y_{n+1}$ , ensuring that the piecewise polynomial interpolant is continuously differentiable over the range of integration and hence that the defect (1.2) is properly defined.

12053

0405''03310

13001  
13003

3

13006  
13007  
13008  
13009

Shampine [9] (see also [6]) has examined the accuracy of this interpolant, and its derivatives, by splitting the error into two components in the following way. Let  $Q(x)$  denote the Hermite polynomial that interpolates to the *exact* local values:

13010  
13011

$$\begin{aligned} Q(\xi_i) &= u(\xi_i), & i = 1, \dots, m, \\ Q'(\xi_i) &= u'(\xi_i), & i = 1, \dots, r. \end{aligned}$$

13012

We assume that  $u$  has  $m+r$  continuous derivatives and write

13013

$$(2.2) \quad p^{(k)}(x) - u^{(k)}(x) = [p^{(k)}(x) - Q^{(k)}(x)] + [Q^{(k)}(x) - u^{(k)}(x)].$$

13014  
13015  
13016  
13017

Shampine shows that the first term on the right-hand side of (2.2), the "data error," is  $O(h^{q+1-k})$  while the second term, the "interpolation error," is  $O(h^{m+r-k})$ . We make use of this result with  $k=0$  and  $k=1$  to examine the asymptotic behaviour of the defect in two special cases.

13018  
13019  
13020

**2.1. Case I.** We suppose that the interpolation scheme has been set up so that the interpolation error is dominant in (2.2), that is,  $m+r < q+1$ . It then follows from (1.3) and (1.4) that

13021

$$(2.3) \quad \delta(x) = Q'(x) - u'(x) + O(h^{m+r})$$

13022

with

13023

$$Q'(x) - u'(x) = O(h^{m+r-1}).$$

13024  
13025  
13026  
13027

The precise form of  $Q'(x) - u'(x)$  can be found by applying classical interpolation theory. We denote by  $u_t(x)$  and  $Q_t(x)$  the  $t$ th components of  $Q(x)$  and  $u(x)$ , respectively, and examine the term  $Q_t'(x) - u_t'(x)$ . The following theorem summarises some results from [12, pp. 1-5].

13028

**THEOREM 1.** *If  $u_t \in C^{m+r}[x_n, x_n+h]$ , then for  $x_n \leq x \leq x_n+h$*

13029

$$(2.4) \quad Q_t(x) - u_t(x) = -\pi(x)G_t(x)$$

13030

where

13031

$$\pi(x) = \prod_{i=1}^r (x - \xi_i)^2 \prod_{i=r+1}^m (x - \xi_i)$$

13032

and

13033

$$G_t(x) = \frac{u_t^{(m+r)}(\theta(x))}{(m+r)!}$$

13034

for some  $x_n \leq \theta(x) \leq x_n+h$ . Furthermore,  $G_t'(x)$  is continuous on  $[x_n, x_n+h]$ .

13035

Now write  $\xi_i = x_n + \sigma_i h$  and let  $x = x_n + \tau h$ , so that

13036

$$\pi(x) = \prod_{i=1}^r (\tau - \sigma_i)^2 h^2 \prod_{i=r+1}^m (\tau - \sigma_i) h$$

13037

$$\equiv h^{m+r} r(\tau).$$

13038

Differentiating (2.4) and using the additional fact that

13039  
13040

$$G_t(x) = \frac{u_t^{(m+r)}(x_n)}{(m+r)!} + O(h),$$

13041

0217''02066

14001  
14003

4

DESMOND J. HIGHAM

14006  
14007

we obtain

14008

$$Q'_i(x) - u'_i(x) = -h^{m+r-1} \frac{dr(\tau)}{d\tau} \frac{u_i^{(m+r)}(x_n)}{(m+r)!} + O(h^{m+r}),$$

14009

for each  $1 \leq i \leq N$ . Substituting this expression into (2.3), we find

14010

$$(2.5) \quad \delta(x_n + \tau h) = -h^{m+r-1} \frac{dr(\tau)}{d\tau} \frac{u^{(m+r)}(x_n)}{(m+r)!} + O(h^{m+r}),$$

14011  
14012

which has the required form (1.6), and shows that, asymptotically, the defect behaves as does a multiple of  $dr(\tau)/d\tau$  over the step.

14013

14014

14015

14016

14017

14018

14019

14020

14021

14022

We now consider some specific interpolation schemes of this type. It is traditional to measure the cost of a Runge-Kutta-interpolation scheme in terms of the number of  $f$  evaluations required per step. We will use the notation  $s(+1)$  to mean that a scheme requires  $(s+1)f$  evaluations to form  $y_{n+1}$ ,  $\{u_i\}_{i=1}^m$ , and  $\{u'_i\}_{i=1}^r$  and sample the defect at a single point; the "+1" appears in brackets because, after a successful step, one of the  $f$  values,  $f(x_{n+1}, y_{n+1})$ , can be reused at the start of the next step.

The simplest example is the cubic Hermite interpolant to  $y_n$ ,  $f(x_n, y_n)$ ,  $y_{n+1}$ , and  $f(x_n + h, y_{n+1})$ , corresponding to  $m = r = 2$  in (2.1). A fourth-order, four-stage Runge-Kutta formula could be used to generate a  $y_{n+1}$  of local order five. This leads to an overall cost of  $5(+1)f$  evaluations. The polynomial  $dr(\tau)/d\tau$  in (2.5) becomes

14023

$$(2.6) \quad \frac{dr(\tau)}{d\tau} = 2\tau(\tau-1)(2\tau-1),$$

14024

14025

14026

14027

14028

as plotted in Fig. 1. There are two local extrema of equal magnitude at  $\tau = \frac{1}{2} \pm \sqrt{3}/6$ .

Next we consider the case  $m = 3$ ,  $r = 2$ . Using a result of Horn [8] we could employ the higher-order formula of the well-known Fehlberg pair to form  $y_{n+1}$  and then generate a data point  $u_2$  of the correct local order at any point  $\xi_2 = x_n + \sigma_2 h$  for a total of  $9(+1)$  evaluations. In this case

14029

$$(2.7) \quad \frac{dr(\tau)}{d\tau} = \tau(\tau-1)(5\tau^2 - [3 + 4\sigma_2]\tau + 2\sigma_2).$$

14030

14031

14032

From (2.5) we see that the most efficient scheme, asymptotically, is found by choosing  $\sigma_2$  in  $(0, 1)$  to minimise  $\max_{\tau \in [0,1]} |dr(\tau)/d\tau|$ . A computer search revealed that this is achieved at  $\sigma_2 = \frac{1}{2}$ . The graph resulting from (2.7) is plotted in Fig. 2 and has a pleasing

14033

14034

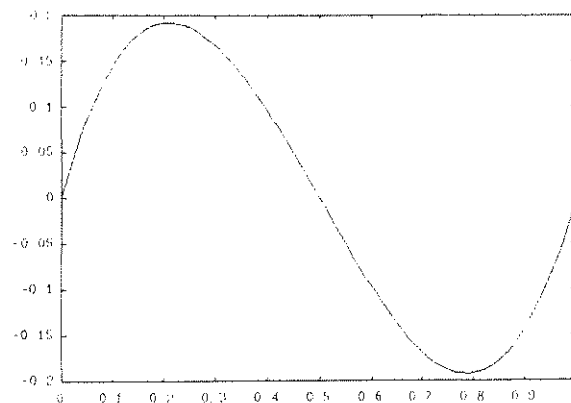


FIG. 1

14038

0292''02312

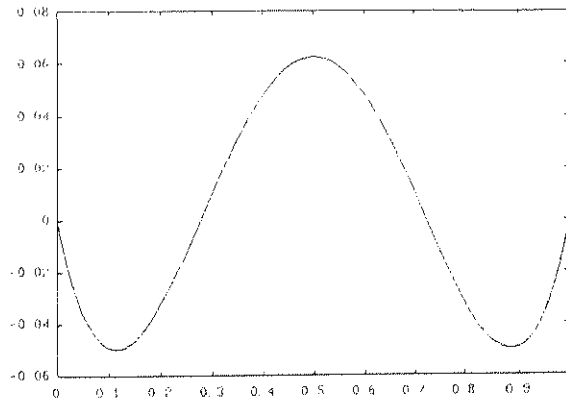


FIG. 2

shape; the peak at  $\tau = \frac{1}{2}$  is 1.25 times as big as the other local extrema at  $\frac{1}{2} \pm \sqrt{15}/10$ . Also, for the choice  $\sigma_2 = \frac{1}{2}$  a similar scheme requiring one fewer function evaluation is available using the Dormand-Prince-Shampine (DPS) triple [10].

For larger values of  $m+r$  the problem of finding enough data of the correct local accuracy, using a reasonable number of function evaluations, becomes more challenging [9, p. 1020]. Enright et al. [4] give an alternative to Shampine's method for constructing interpolants. Their method is perfectly general: given any Runge-Kutta formula an interpolant can be produced that has the same local accuracy. It would therefore be possible to choose a high-order Runge-Kutta formula, construct a high-order interpolant using the technique of [4] and then use this interpolant to provide the necessary data for an interpolant of the form considered here. Due to the high cost (in terms of  $f$  evaluations per step) we do not pursue such an approach in this paper.

**2.2. Case II.** We now examine the case where the data error dominates in (2.2). This situation has been analysed by Gladwell et al. [6] with a view to relating  $u^{(k)}(x) - p^{(k)}(x)$  to the local error at  $x_n + h$ . We shall use their results to reveal the asymptotic behavior of the defect.

It has been shown [6, p. 325] that for  $m+r > q+1$ , (2.2) takes the form

$$p^{(k)}(x) - u^{(k)}(x) = \sum_{i=2}^m A_i^{(k)}(x)[u_i - u(\xi_i)] + O(h^{q+2-k})$$

where  $A_i(x)$  is a fundamental interpolating polynomial that depends only on the abscissae  $\{\xi_i\}_{i=1}^m$ . From (1.3) and (1.4) we then have

$$(2.8) \quad \delta(x) = \sum_{i=2}^m A_i'(x)[u_i - u(\xi_i)] + O(h^{q+1}).$$

To recover a result of the form (1.6) we must reduce the summation in (2.8) to a single term. For example, the standard cubic Hermite interpolant ( $m=r=2$ ) could be used with a second-order, two-stage Runge-Kutta formula ( $q+1=3$ ) to give

$$\delta(x) = A_2'(x)[y_{n+1} - u(x_n + h)] + O(h^3).$$

This may be written

$$\delta(x_n + \tau h) = 6\tau(1-\tau) \frac{[y_{n+1} - u(x_n + h)]}{h} + O(h^3),$$

16001  
16003

6

DESMOND J. HIGHAM

16004  
16007  
16008  
16009

which is essentially the same form as (1.6). Figure 3 presents a plot of  $6\tau(1-\tau)$ ; the single extremum in  $[0, 1]$  occurs at  $\tau = \frac{1}{2}$ . The cost of this scheme is  $(3+1)$  evaluations per step.

16010  
16011  
16012  
16013  
16014  
16015  
16016  
16017

For higher-order schemes we look at  $m = r = 3$ . Interpolants of the desired form can be obtained by applying a single formula of order less than or equal to 4 over steps of length  $h$  and  $h/2$  or over two steps of length  $h/2$  [6, p. 326], but the resulting defect control schemes are more expensive than those of § 2.1. An alternative approach suggested in [6] that turns out to be useful in our context, is to use a fifth-order Runge-Kutta formula to generate  $y_{n+1}$ , making  $u_3 - u(\xi_3) = y_{n+1} - u(x_n + h) = O(h^6)$ , and then to form an approximation at  $\xi_2$  that is locally  $O(h^5)$ . In this way the  $i = 2$  term dominates the right-hand side of (2.8) to give

16018

$$\delta(x) = A_2'(x)[u_2 - u(\xi_2)] + O(h^5),$$

16019

which may be written

16020

$$\delta(x_n + \tau h) = \frac{2\tau(\tau-1)(\tau^2[5-10\sigma_2] + \tau[10\sigma_2^2-3] - 5\sigma_2^2 + 3\sigma_2)}{\sigma_2^3(\sigma_2-1)^3} \frac{[u_2 - u(\xi_2)]}{h} + O(h^5)$$

16021

$$\equiv g_{\sigma_2}(\tau) \frac{[u_2 - u(\xi_2)]}{h} + O(h^5)$$

16022

where we recall that  $\xi_2 = x_n + \sigma_2 h$ . With the fifth-order formula of the DPS triple it has been shown [1] that a suitable approximation  $u_2$  can be found for any  $\sigma_2$  using no extra function evaluations. The overall cost of the defect control scheme is then  $8(+1)$  evaluations. One reasonable way of choosing  $\sigma_2 \in (0, 1)$  is to minimise  $\max_{\tau \in [0,1]} |g_{\sigma_2}(\tau)|$ . The minimum was found by a computer search to occur at  $\sigma_2 = \frac{1}{3}$ . The corresponding polynomial,  $g_{1/2}(\tau)$ , is an exact multiple of the right-hand side of (2.6). Similar schemes can be constructed using Horn's results for the Fehlberg fifth-order formula. With  $\sigma_2 = 3/5$  a scheme costing  $8(+1)$  evaluations is available while for any other  $\sigma_2$  the cost increases by 1.

16023  
16024  
16025  
16026  
16027  
16028  
16029  
1603016031  
16032

The idea of making one of the  $[u_i - u(\xi_i)]$  terms in (2.8) dominate can be used to construct higher-order schemes but, as in Case I, this may not be practical.

16033  
16034  
16035

**3. Discussion.** Apart from the low-order version in § 2.2, the schemes presented here are more expensive than their competitors. The first example in § 2.1 gives an  $O(h^3)$  defect for  $5(+1)$  evaluations per step, while the same order of accuracy could

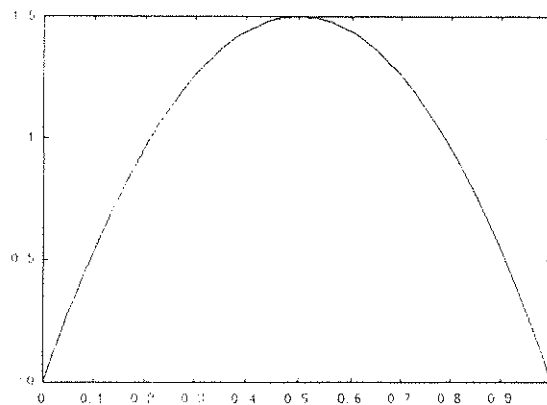


FIG. 3

16036

16037

16041

0329"02486



be achieved with 4(+1) evaluations by a third-order, three-stage Runge-Kutta formula and a cubic Hermite interpolant. Similarly, the 8(+1) evaluations required by the DPS based schemes of §§ 2.1 and 2.2 to obtain an  $O(h^4)$  defect can be compared with the 6(+1) evaluations used by the most efficient of those in [3]. However, the extra cost is countered by the asymptotic validity of the new schemes.

It is shown in [3] that standard nonextrapolated error-per-unit-step control can also be regarded as asymptotically valid defect control. The schemes proposed in this paper, which require fewer function evaluations per step, are closely related—they are asymptotically equivalent to controlling the local error-per-unit-step in the interpolant at some method dependent point,  $x_n + \hat{\tau}h$ .

**4. Numerical results.** We have implemented the following schemes:

3/8: Case I,  $m = r = 2$ , with the "3/8-Rule" [7, p. 137] as the fourth-order Runge-Kutta formula.

DPS#1: Case I,  $m = 3$ ,  $r = 2$ , with the fifth-order formula of RK5(4)7FM and the  $O(h^6)$  midpoint approximation from [10],

DPS#2: Case II,  $m = r = 3$ , with the fifth-order formula of RK5(4)7FM and the  $O(h^5)$  midpoint approximation from [10] (also derived in [1]).

The defect was controlled by sampling at a single point:  $\tau^* = \frac{1}{2} + \sqrt{3}/6$  for the 3/8 and DPS#2 schemes, and  $\tau^* = \frac{1}{2}$  for the DPS#1 scheme. After each step, whether successful or not, the new stepsize was chosen according to

$$\frac{h_{\text{new}}}{h_{\text{old}}} = .9 \left( \frac{\text{TOL}}{\|\delta(x_n + \tau^*h)\|_{\infty}} \right)^{1/p}$$

for a defect of  $O(h^p)$ . As a safety precaution we imposed the restriction

$$\frac{1}{10} \leq \frac{h_{\text{new}}}{h_{\text{old}}} \leq 5,$$

discussed in [11]. Following [2] the quantities

$$R1 = \frac{\max_{j=1, \dots, 100} \|\delta(x_n + .01jh)\|_{\infty}}{\|\delta(x_n + \tau^*h)\|_{\infty}},$$

$$R2 = \frac{\max_{j=1, \dots, 100} \|\delta(x_n + .01jh)\|_{\infty}}{\text{TOL}}$$

were computed on each step; R1MAX and R2MAX denote their respective maximum values over the range of integration. Note that R1MAX measures the quality of the sample point and R2MAX indicates how successful the code was in keeping  $\|\delta(x)\|_{\infty} < \text{TOL}$ . Ideally we would like R1MAX = 1 and R2MAX  $\approx < 1$ .

The methods were tested on the orbit equations [5, Class D]:

$$y'_1 = y_3, \quad y_1(0) = 1 - \varepsilon,$$

$$y'_2 = y_4, \quad y_2(0) = 0,$$

$$y'_3 = \frac{-y_1}{(y_1^2 + y_2^2)^{3/2}}, \quad y_3(0) = 0,$$

$$y'_4 = \frac{-y_2}{(y_1^2 + y_2^2)^{3/2}}, \quad y_4(0) = \left( \frac{1 + \varepsilon}{1 - \varepsilon} \right)^{1/2},$$

$$0 \leq x \leq 20$$



TABLE 1  
R1MAX, R2MAX pairs for 3/8 scheme.

TOL	$10^{-2}$	$10^{-4}$	$10^{-6}$	$10^{-8}$
$\epsilon = .1$	3.3, 2.7	2.1, 1.6	1.4, 1.1	1.1, 0.8
$\epsilon = .5$	1.9, 1.7	1.3, 1.1	1.1, 0.8	1.0, 0.8
$\epsilon = .9$	1.5, 1.1	1.1, 1.0	1.0, 0.9	1.0*, 0.8*

\* The integration was halted prematurely after 5,000 steps.

TABLE 2  
R1MAX, R2MAX pairs for DPS#1 scheme.

TOL	$10^{-2}$	$10^{-4}$	$10^{-6}$	$10^{-8}$
$\epsilon = .1$	3.2, 1.5	3.4, 2.2	1.3, 1.1	1.0, 0.8
$\epsilon = .5$	4.9, 0.9	2.0, 1.0	1.0, 1.0	1.0, 0.8
$\epsilon = .9$	1.2, 0.8	1.0, 1.0	1.0, 1.0	1.0, 0.8

TABLE 3  
R1MAX, R2MAX pairs for DPS#2 scheme.

TOL	$10^{-2}$	$10^{-4}$	$10^{-6}$	$10^{-8}$
$\epsilon = .1$	2.3, 0.9	5.1, 2.0	1.2, 0.8	1.0, 0.7
$\epsilon = .5$	3.0, 0.9	2.2, 1.3	1.2, 1.0	1.1, 0.8
$\epsilon = .9$	4.2, 1.0	2.1, 1.0	1.1, 1.0	1.0, 0.9

where values of .1, .5, and .9 were chosen for the eccentricity parameter  $\epsilon$ . Tables 1-3 record the R1MAX and R2MAX values. As the analysis of § 2 predicts, we see that the performance becomes extremely good as TOL decreases. The results can be compared with those of the original defect control schemes in [2] for the orbit problems; while both sets of results are satisfactory, the new schemes are clearly more reliable at stringent tolerances. We emphasise that the price to be paid for this improvement is a higher cost per step.

To conclude, we outline some possible extensions to this work. The schemes described in § 2 allow some freedom in the choice of Runge-Kutta formula and interpolation points. By examining higher-order terms in the expansion (1.5), it may be possible to use this freedom to produce schemes for which the leading term is more likely to dominate. Finally, in some cases it may be preferable to estimate and control a measure of the defect other than  $\max_{\tau \in [0,1]} \|\delta(x_n + \tau h)\|_\infty$  (see [2]). Whatever measure is used, the interpolants presented here should prove extremely useful, since an asymptotically correct approximation to the defect over the entire step can be constructed from a single sample value.

**Acknowledgments.** I thank Nick Higham and George Hall, whose comments improved this manuscript.

#### REFERENCES

- [1] J. R. DORMAND AND P. J. PRINCE, *Runge-Kutta triples*, *Comput. Math. Appl.*, 12 (1986), pp. 1007-1017.



19001  
19002

## DEFECT CONTROL WITH RUNGE-KUTTA SCHEMES

9

19006  
19007

[2] W. H. ENRIGHT, *A new error control for initial value solvers*, Numerical Analysis Report 122, University of Manchester, Manchester, U.K., 1986.

19008

19009

[3] ———, *Analysis of error control strategies for continuous Runge-Kutta methods*, Tech. Report 205/87, Dept. of Computer Science, University of Toronto, Toronto, Ontario, Canada, 1987.

19010

19011

[4] W. H. ENRIGHT, K. R. JACKSON, S. P. NORSETT, AND P. G. THOMSEN, *Interpolants for Runge-Kutta formulas*, ACM Trans. Math. Software, 12 (1986), pp. 193-218.

19012

19013

[5] W. H. ENRIGHT AND J. D. PRYCE, *Two FORTRAN packages for assessing initial value methods*, ACM Trans. Math. Software, 13 (1987), pp. 1-27.

19014

19015

[6] I. GLADWELL, L. F. SHAMPINE, L. S. BACA, AND R. W. BRANKIN, *Practical aspects of interpolation in Runge-Kutta codes*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 322-341.

19016

19017

[7] E. HAIRER, S. P. NORSETT, AND G. WANNER, *Solving Ordinary Differential Equations 1*, Springer-Verlag, Berlin, New York, 1987.

19018

19019

[8] M. K. HORN, *Fourth- and fifth-order, scaled Runge-Kutta algorithms for treating dense output*, SIAM J. Numer. Anal., 20 (1983), pp. 558-568.

19020

19021

[9] L. F. SHAMPINE, *Interpolation for Runge-Kutta methods*, SIAM J. Numer. Anal., 22 (1985), pp. 1014-1027.

19022

19023

[10] ———, *Some practical Runge-Kutta formulas*, Math. Comput., 46 (1986), pp. 135-150.

19024

19025

[11] L. F. SHAMPINE AND H. A. WATTS, *The art of writing a Runge-Kutta code. II*, Appl. Math. Comput., 5 (1979), pp. 93-121.

19026

[12] B. WENDROFF, *Theoretical Numerical Analysis*, Academic Press, New York, 1966.

19028

19029

19029

0218''01796''23325