SIAM J. SCI. STAT. COMPUT.
Vol. 12, No. 5, pp. 991-999, September 1991

RUNGE-KUTTA DEFECT CONTROL USING HERMITE-BIRKHOFF INTERPOLATION*

DESMOND J. HIGHAM†

Abstract. Two techniques for reliably controlling the defect (residual) in the numerical solution of nonstiff initial value problems were given in [D. J. Higham, *SIAM J. Numer. Anal.*, 26 (1989), pp. 1175-1183]. This work describes an alternative approach based on Hermite-Birkhoff interpolation. The new approach has two main advantages—it is applicable to Runge-Kutta schemes of any order, and it gives rise to a defect of the optimum asymptotic order of accuracy. For a particular Runge-Kutta formula the asymptotic analysis is verified numerically.

Key words. Runge-Kutta, defect, residual, backward error, Hermite-Birkhoff interpolation

AMS(MOS) subject classification. 65L05

1. Introduction. This work deals with the control of errors in the numerical solution of the nonstiff initial value problem

$$y'(x) = f(x, y(x)), \qquad y(a) = y_a \in \mathbb{R}^N, \qquad a \leqq x \leqq b,$$

using an explicit Runge-Kutta method. These methods produce discrete approximations $y_n \approx y(x_n)$ by proceeding in a stepwise fashion; a typical step involves advancing the numerical approximation from x_n to $x_{n+1} := x_n + h_n$. To complement the approximation at the meshpoints $\{x_n\}$, many authors have derived interpolants $p(x)$ which provide approximations $p(x) \approx y(x)$ for other values of x (see, for example, [1], [7], [9], [12], [13]). It is desirable for $p(x)$ to provide efficient, accurate approximations, and to have at least global C^1 continuity. The corresponding defect (residual),

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

may then be used to measure the error in the numerical solution. As a means of error-control, Enright [5] suggests that the defect be sampled at one or more points on each step. By retaking the step with a smaller stepsize if necessary, we could ensure that the sampled values were sufficiently small on every step. The idea of controlling the defect is intuitively reasonable—if the solution has a small defect then it solves a nearby system of differential equations and has a small "backward error." For a discussion of the defect control philosophy, and its relation to the more traditional local error control, see [5] and [6].

When standard Runge-Kutta interpolation schemes are used, the shape of the defect over each step cannot be determined a priori, since it depends on f. In [10] two special classes of interpolant were derived which have the property that asymptotically (as $h_n \to 0$) each component of the defect behaves like a multiple of a known polynomial over each step, and hence the maximum or root-mean-square integral value of the defect can be approximated from one sample. (We mention that it has been shown that in the case of Adams PECE multistep formulas, there are natural interpolation

* Received by the editors October 16, 1989; accepted for publication (in revised form) May 1, 1990. This research was supported by the Information Technology Research Centre of Ontario, and the Natural Sciences and Engineering Research Council of Canada. Part of this manuscript was prepared while the author was visiting the Computer Science Department, Cornell University, Ithaca, New York.

† Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4. Present address, Department of Mathematics and Computer Sciences, University of Dundee, DD1 4HN, Scotland, U.K.

schemes which automatically satisfy this property [11].) Two main drawbacks of the schemes in [10] are:

- They are only viable for low-order Runge–Kutta formulas (say $\leqq 5$); and
- They produce a defect with an asymptotic order of one less than the optimal value.

The purpose of this work is to show that both difficulties can be overcome by the use of Hermite–Birkhoff interpolants.

In the next section we set out some basic definitions and define the particular type of Hermite–Birkhoff interpolant that we need. We then show that the corresponding defect has the desired asymptotic properties. In the final two sections we specialize to fifth-order Runge–Kutta formulas and give some numerical verification of the theoretical results.

2. The Hermite–Birkhoff interpolant. An s-stage explicit Runge–Kutta formula for advancing the numerical solution over a step of length $h(=h_n)$ can be written

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i,$$

where

$$k_1 = f(x_n, y_n),$$
$$k_i = f\left(x_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} k_j\right), \qquad 2 \leqq i \leqq s.$$

The local solution for the step $u(x)$ is the solution curve which passes through y_n at x_n; that is, $u'(x) = f(x, u(x))$ and $u(x_n) = y_n$. A Runge–Kutta formula is said to be of *order* p if p is the largest integer such that the local error satisfies $y_{n+1} - u(x_n + h) = O(h^{p+1})$. (Here, and in the following analysis, we assume that f is sufficiently differentiable.) We say that a corresponding local interpolant $p(x)$ has *local order* $q+1$ if q is the largest integer such that $p(x_n + \tau h) - u(x_n + \tau h) = O(h^{q+1})$ for any fixed $\tau \in [0, 1]$. The interpolation condition $p(x_n + h) = y_{n+1}$ ensures that $q + 1 \leqq p + 1$.

In deriving the interpolant $p(x)$, a natural approach is to use a polynomial interpolant to solution and derivative approximations. The data $y_n, f(x_n, y_n), y_{n+1}$ and $f(x_{n+1}, y_{n+1})$ is available free of charge, but normally extra data is needed to achieve the desired order, and more stages must be added to the Runge–Kutta process. An automatic bootstrapping technique for generating polynomial interpolants of local order up to $p+1$ was developed by Enright et al. [7]. The technique makes use of a special kind of Hermite–Birkhoff interpolant. We show below that these interpolants can also be used to derive robust defect control schemes.

The Hermite–Birkhoff interpolants that we consider are chosen to match the meshpoint data plus some extra derivative data at intermediate points in the step. If f is Lipschitzian, then we have the result

$$y_{n+1} - u(x_{n+1}) = O(h^{p+1}) \Rightarrow f(x_{n+1}, y_{n+1}) - u'(x_{n+1}) = O(h^{p+1}).$$

We write $\gamma_1 = 0$, $\gamma_2 = 1$, $u_1' := f(x_n, y_n)$ and $u_2' := f(x_{n+1}, y_{n+1})$, and we suppose that there are $r-2$ further derivative approximations $\{u_i'\}_{i=3}^{r}$ available which also satisfy

$$u_i' - u'(x_n + \gamma_i h) = O(h^{p+1}),$$

where $\{\gamma_i\}_{i=3}^{r}$ are distinct points in $(0, 1)$. This data could be generated, for example, by constructing one of the standard locally ordered $p+1$ interpolants, $p(x)$, and fixing $u_i' := f(x_n + \gamma_i h, p(x_n + \gamma_i h))$. We may then consider the Hermite–Birkhoff polynomial interpolant $H(x)$ of degree $\leqq r+1$ which satisfies

$$H(x_n) = y_n, \qquad H(x_{n+1}) = y_{n+1}, \qquad H'(x_n + \gamma_i h) = u_i', \qquad 1 \leqq i \leqq r. \tag{2.1}$$

Note that $H(x)$ does not satisfy the usual Hermite interpolatory conditions, since it matches a derivative value, but not a solution value at $x_n+\gamma_i h$, $i=3,\cdots,r$.

We point out that the existence and uniqueness of $H(x)$ cannot be guaranteed in general. However, for any particular r it is always possible to choose $\{\gamma_i\}_{i=3}^{r}$ so that $H(x)$ exists uniquely [7, p. 197]. In the remainder of this section we assume that this has been done.

Using the normalized variable, $\tau=(x-x_n)/h$, we may write the interpolant in the form

$$H(x_n+\tau h)=d_1(\tau)y_n+d_2(\tau)y_{n+1}+h\sum_{i=1}^{r}e_i(\tau)u_i', \tag{2.2}$$

where $d_1(\tau)$, $d_2(\tau)$, and $\{e_i(\tau)\}_{i=1}^{r}$ are scalar polynomials in τ of degree $\leqq r+1$. To examine the local error and the defect in $H(x)$, we adopt the usual strategy of isolating the interpolation and data errors. To this end, we let $Q(x)$ denote the Hermite-Birkhoff interpolant that matches the exact local solution values:

$$Q(x_n+\tau h)=d_1(\tau)u(x_n)+d_2(\tau)u(x_{n+1})+h\sum_{i=1}^{r}e_i(\tau)u'(x_n+\gamma_i h). \tag{2.3}$$

First, we consider the interpolation error $Q(x)-u(x)$. To find the asymptotic order of this expression we follow the approach of Dormand et al. [4, pp. 5-6], which relies only on the uniqueness of the interpolant. (Incidentally, although there is a well-known error expression in the case of Hermite interpolation, the author is not aware of any general expression for the error in $Q(x)$.)

Suppose that the local solution $u(x)$ has at least $r+2$ continuous derivatives. Using the subscript t to denote the tth component of a vector-valued function, we may substitute the truncated Taylor expansions

$$u_t(x_n+h)=\sum_{k=0}^{r+1}\frac{h^k}{k!}u_t^{(k)}(x_n)+\frac{h^{r+2}}{(r+2)!}u_t^{(r+2)}(x_n+\theta_{t,0}h),$$

$$u_t'(x_n+\gamma_i h)=\sum_{k=0}^{r}\frac{(\gamma_i h)^k}{k!}u_t^{(k+1)}(x_n)+\frac{(\gamma_i h)^{r+1}}{(r+1)!}u_t^{(r+2)}(x_n+\theta_{t,i}h),$$

where $0<\theta_{t,0},\ \theta_{t,1},\cdots,\theta_{t,r}<1$, into (2.3) to give

$$Q_t(x_n+\tau h)=d_1(\tau)u_t(x_n)+d_2(\tau)\left\{\sum_{k=0}^{r+1}\frac{h^k}{k!}u_t^{(k)}(x_n)+\frac{h^{r+2}}{(r+2)!}u_t^{(r+2)}(x_n+\theta_{t,0}h)\right\}$$
$$+h\sum_{i=1}^{r}e_i(\tau)\left\{\sum_{k=0}^{r}\frac{(\gamma_i h)^k}{k!}u_t^{(k+1)}(x_n)+\frac{(\gamma_i h)^{r+1}}{(r+1)!}u_t^{(r+2)}(x_n+\theta_{t,i}h)\right\}.$$

This may be rearranged as

$$\begin{aligned}Q_t(x_n+\tau h)&=u_t(x_n)\{d_1(\tau)+d_2(\tau)\}\\&\quad+\sum_{k=1}^{r+1}\frac{h^k}{k!}u_t^{(k)}(x_n)\left\{d_2(\tau)+\sum_{i=1}^{r}e_i(\tau)k\gamma_i^{k-1}\right\}\\&\quad+h^{r+2}\left\{\frac{d_2(\tau)}{(r+2)!}u_t^{(r+2)}(x_n+\theta_{t,0}h)\right.\\&\qquad\left.+\sum_{i=1}^{r}e_i(\tau)\frac{\gamma_i^{r+1}}{(r+1)!}u_t^{(r+2)}(x_n+\theta_{t,i}h)\right\}.\end{aligned} \tag{2.4}$$

Since $d_2(\tau)$, $\{e_i(\tau)\}_{i=1}^r$ and $u^{(r+2)}(x_n+\tau h)$ are bounded on $[0, 1]$, this may be written as

$$Q_t(x_n+\tau h) = u_t(x_n)\{d_1(\tau)+d_2(\tau)\} + \sum_{k=1}^{r+1} \frac{h^k}{k!} u_t^{(k)}(x_n) \cdot \left\{ d_2(\tau) + \sum_{i=1}^{r} e_i(\tau) k \gamma_i^{k-1} \right\} + O(h^{r+2}). \tag{2.5}$$

Now, by uniqueness, $Q_t(x_n+\tau h)$ will be exact for $u_t(x_n+\tau h) \equiv (\tau h)^k$, for $0 \leqq k \leqq r+1$. Hence, in (2.3),

$$d_1(\tau)+d_2(\tau) \equiv 1, \tag{2.6}$$

$$d_2(\tau) + \sum_{i=1}^{r} e_i(\tau) k \gamma_i^{k-1} \equiv \tau^k, \qquad 1 \leqq k \leqq r+1. \tag{2.7}$$

Substituting into (2.5) this gives

$$\begin{aligned} Q_t(x_n+\tau h) &= u_t(x_n) + \sum_{k=1}^{r+1} \frac{(\tau h)^k}{k!} u_t^{(k)}(x_n) + O(h^{r+2}) \\ &= u_t(x_n+\tau h) + O(h^{r+2}). \end{aligned} \tag{2.8}$$

Similarly, by differentiating (2.4), (2.6), and (2.7) with respect to τ, we can show that

$$\frac{1}{h}\frac{d}{d\tau} Q_t(x_n+\tau h) = u_t'(x_n+\tau h) + O(h^{r+1}). \tag{2.9}$$

Using (2.2) and (2.3), the data error $H(x)-Q(x)$ may be written

$$\begin{aligned} H(x_n+\tau h) - Q(x_n+\tau h) &= d_2(\tau)[y_{n+1}-u(x_{n+1})] \\ &\quad + h \sum_{i=1}^{r} e_i(\tau)[u_i' - u'(x_n+\gamma_i h)] \\ &= d_2(\tau)[y_{n+1}-u(x_{n+1})] + O(h^{p+2}), \end{aligned} \tag{2.10}$$

since the derivative approximations are accurate to $O(h^{p+1})$. Similarly,

$$\begin{aligned} &\frac{1}{h}\frac{d}{d\tau}\{H(x_n+\tau h) - Q(x_n+\tau h)\} \\ &\quad = \frac{1}{h} d_2'(\tau)[y_{n+1}-u(x_{n+1})] + O(h^{p+1}). \end{aligned} \tag{2.11}$$

If we have $r \geqq p$, that is, if we use a sufficient amount of derivative data, then, combining (2.8) and (2.10) we have

$$H(x_n+\tau h) - u(x_n+\tau h) = d_2(\tau)[y_{n+1}-u(x_{n+1})] + O(h^{p+2}).$$

We thus have an interpolant of local order $p+1$. Furthermore, $H(x)$ has the desirable property that the local error at any point in the step can be directly related to the local error at the next meshpoint $y_{n+1}-u(x_{n+1})$. Some low-order interpolation schemes with this property have been discussed in [9]. The defect in $H(x)$ satisfies

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

which, using (2.9) and (2.11), has the form

$$\delta(x_n+\tau h)=\frac{1}{h}\,d_2'(\tau)[y_{n+1}-u(x_{n+1})]+O(h^{p+1}). \tag{2.12}$$

Hence, for sufficiently small stepsizes each component of the defect behaves like a multiple of a known polynomial over each step.

As we mentioned earlier, the extra derivative data $\{u_i'\}_{i=3}^r$ could be obtained by first constructing a standard interpolant $p(x)$ of local order $p+1$, and then setting $u_i'=f(x_n+\gamma_i h, p(x_n+\gamma_i h))$. With $r=p$, this would mean that the new interpolant requires $p-2$ more f evaluations per step than $p(x)$. However, in the case where $p(x)$ is a Hermite interpolating polynomial it is possible that some $u_i'=f(x_n+\gamma_i h, p(x_n+\gamma_i h))$ data was used in the formation of $p(x)$ and hence will be available "free of charge." An example of this will be seen in the next section.

From the expansion (2.12) we see that in order to have a small defect, it is desirable that the local error per unit step $[y_{n+1}-u(x_{n+1})]/h$ be small. By examining the truncation coefficients in the asymptotic local error expansions, it is possible to derive Runge–Kutta formulas with "minimal" local errors (see [3] for an overview). Such formulas would clearly be useful in our context. The polynomial $d_2'(\tau)$ appearing in (2.12) is determined solely by the choice of $\{\gamma_i\}_{i=3}^r$. (From (2.1) and (2.2), $d_2(\tau)$ is determined by the conditions $d_2(0)=d_2'(\gamma_i)=0$, $1\leqq i\leqq r$, and $d_2(1)=1$.) Hence it is sensible to choose these points so that the relevant measure of $d_2'(\tau)$ is small. In particular, if we are concerned with controlling $\max_{\tau\in[0,1]}\|\delta(x_n+\tau h)\|$ over each step, for some vector norm $\|\cdot\|$, then the most efficient scheme asymptotically is given by choosing $\{\gamma_i\}_{i=3}^r$ to solve the minimax problem

$$\min_{\{\{\gamma_i\}_{i=3}^r\in(0,1)\text{ distinct}\}}\ \max_{\tau\in[0,1]}|d_2'(\tau)|. \tag{2.13}$$

The author does not know whether this problem can be solved analytically. In any case it may be necessary to place other constraints on $\{\gamma_i\}_{i=3}^r$. For example, some γ_i values may be fixed a priori, and the γ_i should be reasonably well spaced out across the step (see the next section).

The following lemma gives a little insight into the problem.

LEMMA 2.1. *Suppose there exists a unique polynomial $d_2(\tau)$ of degree $\leqq r+1$ satisfying*

$$d_2(0)=0, \qquad d_2(1)=1,$$
$$d_2'(\gamma_i)=0, \qquad 1\leqq i\leqq r,$$

where $\gamma_1=0$, $\gamma_2=1$ and $\{\gamma_i\}_{i=3}^r\in(0,1)$ are distinct. Let $d_2^*(\tau)$ be the corresponding polynomial which satisfies

$$d_2^*(0)=0, \qquad d_2^*(1)=1,$$
$$d_2^{*\prime}(1-\gamma_i)=0, \qquad 1\leqq i\leqq r.$$

Then $1-d_2(\tau)\equiv d_2^*(1-\tau)$ and hence $d_2'(\tau)\equiv d_2^{*\prime}(1-\tau)$, so that

$$\max_{[0,1]}|d_2'(\tau)|=\max_{[0,1]}|d_2^{*\prime}(\tau)|.$$

Proof. Given $d_2(\tau)$ as in the lemma, $g(\tau):=1-d_2(\tau)$ is the unique polynomial of degree $\leqq r+1$ which satisfies

$$g(0)=1, \qquad g(1)=0,$$
$$g'(\gamma_i)=0, \qquad 1\leqq i\leqq r.$$

These conditions are precisely those which $d_2^*(1-\tau)$ must satisfy. □

The lemma shows that solutions to the minimax problem (2.13) generally arise in pairs $\{\gamma_i\}_{i=3}^r$ and $\{1-\gamma_i\}_{i=3}^r$. If the problem is to be solved by performing a grid search, then the lemma allows us to reduce the amount of searching. For example, if $r=5$, then rather than considering $\gamma_3, \gamma_4, \gamma_5 \in (0,1)$, we may restrict attention to $\gamma_5 \in (0,1)$ and $\gamma_3, \gamma_4 \in (0, \frac{1}{2}]$. (We may also assume, without loss of generality, the ordering $\gamma_3 < \gamma_4 < \gamma_5$.)

In the remaining sections we focus on the case $p=5$. First we derive the conditions which guarantee the existence of a unique Hermite–Birkhoff interpolant. We then compute the optimal points $\{\gamma_i\}_{i=3}^5$ and test the resulting defect control scheme numerically.

3. The case $p=5$. With a fifth-order Runge–Kutta formula ($p=5$) we must choose $r \geqq 5$ in order to construct a Hermite–Birkhoff interpolation scheme of the required form. Taking $r=5$, we have three free parameters γ_3, γ_4, and γ_5, which must be distinct in $(0,1)$. The interpolation conditions that $H(x)$ must satisfy can be cast as a system of linear equations. The determinant of the system can be shown to be a constant multiple of

$$\det := 3(\gamma_3+\gamma_4+\gamma_5) - 5(\gamma_3\gamma_4+\gamma_3\gamma_5+\gamma_4\gamma_5) + 10\gamma_3\gamma_4\gamma_5 - 2.$$

Hence $H(x)$ exists uniquely if and only if $\{\gamma_3, \gamma_4, \gamma_5\}$ are chosen so that $\det \neq 0$. For example, $\{0.4, 0.5, 0.6\}$, $\{0.1, 0.5, 0.9\}$, and $\{0.28, 0.75, 0.8\}$ are invalid parameters. Further straightforward but tedious algebra shows that $d_2(\tau)$ has the form

$$d_2(\tau) = \tau^2(a\tau^4 + b\tau^3 + c\tau^2 + d\tau + e),$$

where

$$a = \frac{10}{\det}, \qquad b = \frac{-12}{\det}(\gamma_3+\gamma_4+\gamma_5+1),$$

$$c = \frac{15}{\det}(\gamma_3+\gamma_4+\gamma_5+\gamma_3\gamma_4+\gamma_3\gamma_5+\gamma_4\gamma_5),$$

$$d = -2-2c-3b-4a, \qquad e = 3+3a+2b+c.$$

As discussed earlier, one reasonable way to choose $\{\gamma_3, \gamma_4, \gamma_5\}$ is to minimize the quantity $\mathrm{d2max} := \max_{[0,1]} |d_2'(\tau)|$. Using a simple three-dimensional grid search, we found that as the grid spacing decreased, the optimal γ_5 value seemed to approach one. To avoid approaching this pathological case, we chose to restrict the parameters to $[0.1, 0.9]$. Varying the parameters in steps of 0.1 gave optimal values of $\{0.3, 0.4, 0.9\}$ for which $\mathrm{d2max} = 2.35$. Although finer grid spacing slightly reduced d2max, it appeared that γ_3 and γ_4 were converging to the same value.

In the numerical experiments described in the next section we base our Hermite–Birkhoff interpolant on the six-stage, fifth-order formula from the RK5(4)7FM pair derived by Dormand and Prince [2]. For this formula, Shampine [13] showed that by using $f(x_{n+1}, y_{n+1})$ and adding one extra stage, it is possible to construct an approximation $y_{n+1/2}$ which satisfies $y_{n+1/2} - u(x_n + h/2) = O(h^6)$. It follows that the Hermite interpolant to $y_n, f(x_n, y_n), y_{n+1/2}, f(x_n + h/2, y_{n+1/2}), y_{n+1}$, and $f(x_{n+1}, y_{n+1})$ is locally sixth order. While the general Hermite–Birkhoff interpolant derived above requires three extra f evaluations per step, if we set $\gamma_3 = \frac{1}{2}$ then we can save one evaluation by using $f(x_n + h/2, y_{n+1/2})$ as derivative data. With $\gamma_3 = \frac{1}{2}$ and $\{\gamma_4, \gamma_5\}$ varying in $[0.1, 0.9]$, a grid search with spacing of 0.0001 gave $\{0.1, 0.7051\}$ as the optimal values. The corresponding d2max value is 3.46, and the optimal sample point is $\tau^* = 0.89994049343102$.

4. Numerical results. We now describe some numerical testing of the scheme outlined above. On each step we sampled the defect in $H(x)$ at the asymptotically optimal point and accepted the step if and only if $\|\delta(x_n+\tau^* h)\|_\infty < \mathrm{TOL}$, where TOL is an absolute error tolerance. Stepsize selection was based on the usual asymptotically motivated formula

$$\frac{h_{\text{new}}}{h_{\text{old}}} = 0.9\left(\frac{\mathrm{TOL}}{\|\delta(x_n+\tau^* h)\|_\infty}\right)^{1/5}.$$

For comparison, we also implemented a defect control scheme using the locally sixth-order Hermite interpolant based on $y_n, f(x_n, y_n), y_{n+1/2}, f(x_n+h/2, y_{n+1/2}), y_{n+1}$, and $f(x_{n+1}, y_{n+1})$. In this case the shape of the defect is not known a priori—it is problem dependent and varies from step to step. The construction of this interpolant requires two fewer f evaluations than that of the Hermite-Birkhoff interpolant; hence we chose to sample the defect at three points on each step in order to give schemes with the same overall cost per step. The maximum observed defect at the three sample points was used as the defect estimate. Noting that the defect is zero at the midpoint of the step, we used $\tau = \frac{1}{4}, \frac{2}{3}$, and $\frac{5}{6}$ as the three sample points.

To measure the accuracy of the defect sample in the Hermite-Birkhoff scheme we formed

$$D := \max_n \left\{\frac{\max_{0\le j\le 100}\|\delta(x_n+0.01jh)\|_\infty}{\|\delta(x_n+\tau^* h)\|_\infty}\right\}, \tag{4.1}$$

which is the worst case of the sample value underestimating the "true" maximum defect over all steps. Similarly, we measured the accuracy of the underlying interpolant by computing

$$G := \max_n \left\{\frac{\max_{0\le j\le 100}\|H(x_n+0.01jh)-y(x_n+0.01jh)\|_\infty}{\|y_{n+1}-y(x_{n+1})\|_\infty}\right\}. \tag{4.2}$$

The ratio G compares the global error in $H(x)$ with that of the basic Runge-Kutta formula. Corresponding results were obtained for the Hermite scheme, with $\max_{\tau\in\{1/4,2/3,5/6\}}\|\delta(x_n+\tau h)\|_\infty$ in the denominator of (4.1).

We used the following nonstiff systems:

(i) A problem due to Fehlberg [see 11]:

$$y_1' = 2xy_1 \log(\max(y_2, 10^{-3})), \qquad y_1(1) = \exp(\sin(1)),$$
$$y_2' = -2xy_2 \log(\max(y_1, 10^{-3})), \qquad y_2(1) = \exp(\cos(1)), \qquad 1 \le x \le 5.$$

(ii)–(iv) The orbit equations [8, Class D]:

$$y_1' = y_3, \qquad y_1(0) = 1-\varepsilon,$$
$$y_2' = y_4, \qquad y_2(0) = 0,$$
$$y_3' = \frac{-y_1}{(y_1^2+y_2^2)^{3/2}}, \qquad y_3(0) = 0,$$
$$y_4' = \frac{-y_2}{(y_1^2+y_2^2)^{3/2}}, \qquad y_4(0) = \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{1/2}, \qquad 0 \le x \le 20,$$

with values of 0.1, 0.5, and 0.9, respectively, for the eccentricity parameter ε.

The results can be found in Tables 1 and 2.

We see that for the Hermite-Birkhoff scheme, the defect sample is generally very reliable, more so than for the original robust schemes in [10]. It is noticeable that the

TABLE 1
Defect ratios D for the Hermite–Birkhoff (HB) *and Hermite* (H) *schemes.*

		TOL 10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}
(i)	HB	1.002	1.002	1.000	1.001	1.071
	H	1.946	2.570	2.417	2.413	2.405
(ii)	HB	1.000	1.000	1.000	1.000	1.004
	H	1.001	1.001	1.003	1.198	1.383
(iii)	HB	1.000	1.001	1.000	1.000	1.012
	H	1.076	1.115	1.746	2.125	2.078
(iv)	HB	1.025	1.032	1.706	1.032	1.463
	H	1.887	2.718	3.242	2.565	2.500

TABLE 2
Global error ratios G for the Hermite–Birkhoff (HB) *and Hermite* (H) *schemes.*

		TOL 10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}
(i)	HB	1.070	1.001	1.000	1.000	1.000
	H	1.037	1.000	1.000	1.000	1.000
(ii)	HB	1.033	1.010	1.000	1.000	1.000
	H	1.003	1.004	1.001	1.000	1.000
(iii)	HB	1.023	1.011	1.004	1.001	1.000
	H	1.000	1.025	1.000	1.000	1.000
(iv)	HB	1.038	1.004	1.002	1.000	1.000
	H	1.018	1.007	1.001	1.000	1.000

ratio D worsens as TOL decreases from 10^{-8} to 10^{-10}. This behavior can be attributed to rounding errors—the defect is formed as a numerical difference which can involve a significant amount of cancellation. (For these computations, the unit roundoff was $\approx 2 \times 10^{-16}$.) Such rounding errors are especially prone to occur on the first few steps of an integration. Here a code will typically take conservatively small steps until it finds the scale of the problem. On these steps the defect may be considerably less than TOL. Overall the scheme performed worst on the $\varepsilon = 0.9$ orbit problem. Here the solution is known to change rapidly in certain regions, and hence it is likely that there are some steps where the stepsize-selection scheme has not adequately taken account of these rapid changes and the higher-order terms in the defect expansion cause the "optimum" sample to be slightly less accurate than normal. The Hermite scheme performs less reliably than the Hermite–Birkhoff scheme on all four problems. It is liable to underestimate the defect by a factor of around two. This behavior is to be expected given the problem-dependent nature of the defect. In fact it is perhaps surprising that the defect ratio D remains reasonably small (in theory it can be arbitrarily large). The original experiments of Enright [5] with such "nonrobust" defect control schemes gave similar results. In terms of global errors, Table 2 shows that both interpolants deliver almost exactly the same accuracy as the Runge–Kutta formula on each test problem.

In summary, we have given a general technique for constructing robust defect control schemes, and the particular case that we implemented performed reliably in practice. Two important questions, which lie beyond the scope of this paper, are the following:

- How to compare numerically two different defect control schemes, using such criteria as efficiency, reliability, and the proportionality of the global error to the tolerance.
- How to compare numerically schemes which use different types of error control, such as defect control and the various types of local error control (see [6] for more details).

It is hoped that these issues will be addressed in the near future using a modified version of the DETEST package [8] which is currently under development at the University of Toronto.

Acknowledgments. I am grateful to Wayne Enright, Nick Higham, and Ken Jackson for helpful comments.

REFERENCES

[1] J. R. CASH, *A block 6(4) Runge-Kutta formula for nonstiff initial value problems*, ACM Trans. Math. Software, 15 (1989), pp. 15–28.
[2] J. R. DORMAND AND P. J. PRINCE, *A family of embedded Runge-Kutta formulae*, J. Comput. Appl. Math., 6 (1980), pp. 19–26.
[3] ———, *Practical Runge-Kutta processes*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 977–989.
[4] J. R. DORMAND, M. A. LOCKYER, N. E. MCGORRIGAN, AND P. J. PRINCE, *Global error estimation with Runge-Kutta triples*, Tech. Report TP-CS-88-01, Department of Computer Science, Teesside Polytechnic, Middlesbrough, England, 1988.
[5] W. H. ENRIGHT, *A new error-control for initial value solvers*, Appl. Math. Comput., 31 (1989), pp. 288–301.
[6] ———, *Analysis of error control strategies for continuous Runge-Kutta methods*, SIAM J. Numer. Anal., 26 (1989), pp. 588–599.
[7] W. H. ENRIGHT, K. R. JACKSON, S. P. NØRSETT, AND P. G. THOMSEN, *Interpolants for Runge-Kutta formulas*, ACM Trans. Math. Software, 12 (1986), pp. 193–218.
[8] W. H. ENRIGHT AND J. D. PRYCE, *Two* FORTRAN *packages for assessing initial value methods*, ACM Trans. Math. Software, 13 (1987), pp. 1–27.
[9] 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.
[10] D. J. HIGHAM, *Robust defect control with Runge-Kutta schemes*, SIAM J. Numer. Anal., 26 (1989), pp. 1175–1183.
[11] ———, *Defect estimation in Adams* PECE *codes*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 964–976.
[12] ———, *Highly continuous Runge-Kutta interpolants*, Tech. Report No. 220/89, Department of Computer Science, University of Toronto, Canada, 1989; ACM Trans. Math. Software, to appear.
[13] L. F. SHAMPINE, *Some practical Runge-Kutta formulas*, Math. Comp., 46 (1986), pp. 135–150.