SIAM J. SCI. STAT. COMPUT.
Vol. 10, No. 5, pp. 000-000, September 1989

DEFECT ESTIMATION IN ADAMS PECE CODES[*]

DESMOND J. HIGHAM[†]

Abstract. Many modern codes for solving the nonstiff initial value problem $y'(x)-f(x,y(x))=0$, $y(a)$ given, $a \leqq x \leqq b$, produce, in addition to a discretised solution, a function $p(x)$ that approximates $y(x)$ over $[a,b]$. The associated defect $\delta(x) := p'(x)-f(x,p(x))$ is a natural measure of the error. In this paper the problem of reliably estimating the defect in Adams PECE methods is considered. Attention is focused on the widely used Shampine-Gordon variable order, variable step code fitted with a continuously differentiable interpolant $p(x)$ due to Watts and Shampine (*SIAM J. Sci., Statist. Comput.*, 7 (1986), pp. 334-345]. It is shown that over each step an asymptotically correct estimate of the defect can be obtained by sampling at a single, suitably chosen point. It is also shown that a valid "free" estimate can be formed withour recourse to sampling. Numerical results are given to support the theory.

Key words. Adams PECE method, interpolant, defect

AMS(MOS) subject classification. 65L05

C.R. classification. G.1.7

1. Introduction.

We consider the numerical solution of the nonstiff initial value problem

$$(1.1)\qquad \begin{gathered} y'(x)=f(x,y(x)), \quad y(a)=y_a, \quad a \leqq x \leqq b, \\ f: \mathbf{R}\times\mathbf{R}^N \to \mathbf{R}^N. \end{gathered}$$

In addition to a discretised solution, many modern codes produce a function $p(x)$ that approximates $y(x)$ over $[a,b]$. The associated defect,

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

is a natural measure of the error, and an estimate of the defect can provide valuable information about the accuracy of the solution. Indeed it has recently been suggested that a defect estimate may be used as an alternative to the traditional local error estimate in the stepsize control mechanism [1], [2]. In this paper we are concerned with the problem of estimating the defect when a variable order, variable step Adams PECE method is used. We restrict our attention to the latest DEPAC code [8], [12], but our results can be adapted for other implementations.

In the next section we make use of some analysis from Stetter [9] to obtain an asymptotic expansion for the defect over a step from x_n to x_n+h_{n+1}. We find that, asymptotically, the shape of the defect is determined only by the current order and the local stepsize pattern. This means that a valid estimate of $\delta(x)$ over $[x_n, x_n+h_{n+1}]$ can be formed by sampling at a single point. In particular a technique for approximating $\max_{x\in[x_n,x_n+h_{n+1}]}\|\delta(x)\|$ is given that is a more sophisticated version of that in [7]. Practical details are discussed further in § 3. In § 4 we show that an estimate of the defect can be formed from quantities available in the code. This gives a "free" estimate of the defect—no extra function evaluations are needed. The final section describes the results of some numerical experiments that test the two approaches.

[*] Received by the editors July 11, 1988; accepted for publication (in revised form) October 26, 1988. This work was supported by a Natural Sciences and Engineering Research Council Research Studentship.

[†] Department of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom.

2. The defect. We begin this section with a brief description of the variable order, variable step Adams PECE method implemented, and extensively documented, by Shampine and Gordon [7]. An updated version of the code forms part of the DEPAC ODE solving package [8]; see also [12].

The basic aim is to produce discrete solution vectors $y_n \approx y(x_n)$ by stepping along a mesh $a = x_0 < x_1 < \cdots$. Given $y_n \approx y(x_n)$ and the past f values

$$f_{n+1-j} = f(x_{n+1-j}, y_{n+1-j}), \qquad j = 1, \cdots, k,$$

a kth order step ($1 \leqq k \leqq 12$) is taken as follows. First a predicted approximation to $y(x_{n+1})$ is formed according to the kth order Adams-Bashforth formula,

$$p_{n+1} = y_n + \int_{x_n}^{x_{n+1}} P_{k,n}(t)\, dt,$$

where $P_{k,n}(t)$ denotes the unique interpolating polynomial of degree $\leqq k-1$ that satisfies

$$P_{k,n}(x_{n+1-j}) = f_{n+1-j}, \qquad j = 1, \cdots, k.$$

The new approximation y_{n+1} is then formed using the $(k+1)$st-order Adams-Moulton formula,

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} P^*_{j+1,n}(t)\, dt. \tag{2.1}$$

Here $P^*_{k+1,n}(t)$ denotes the polynomial of degree $\leqq k$ satisfying

$$P^*_{k+1,n}(x_{n+1-j}) = f_{n+1-j}, \qquad j = 1, \cdots, k$$

and

$$P^*_{k+1,n}(x_{n+1}) = f^p_{n+1} := f(x_{n+1}, p_{n+1}).$$

The term PECE is used to indicate the four stages÷predict the value p_{n+1}, evaluate f^p_{n+1}, correct to give y_{n+1}, and then evaluate f_{n+1}, which is needed for the next step.

From (2.1), a natural way to approximate the solution at a point $x \in (x_n, x_{n+1})$ is to use the function

$$\eta(x) = y_n + \int_{x_n}^{x} P^*_{k+1,n}(t)\, dt \tag{2.2a}$$

that may also be written

$$\eta(x) = y_{n+1} + \int_{x_{n+1}}^{x} P^*_{k+1,n}(t)\, dt. \tag{2.2b}$$

The polynomials $\eta(x)$ from each step can be joined together to give a globally continuous approximation to $y(x)$ over $[a, b]$. However, because $\eta'(x_{n+1})$ matches $f(x_{n+1}, p_{n+1})$ and not $f(x_{n+1}, y_{n+1})$, there is a discontinuity in the first derivative at each meshpoint.

An obvious alternative to $P^*_{k+1,n}(t)$ in (2.2) is the polynomial $P_{k+1,n+1}(t)$ that satisfies

$$P_{k+1,n+1}(x_{n+1-j}) = f_{n+1-j}, \qquad j = 0, \cdots, k.$$

This leads to the approximations

$$S(x) = y_n + \int_{x_n}^{x} P_{k+1,n+1}(t)\, dt \tag{2.3}$$

$$y_I(x) = y_{n+1} + \int_{x_{n+1}}^{x} P_{k+1,n+1}(t)\, dt. \tag{2.4}$$

Note that $S'(x)=y_I'(x)$ and hence $S(x)$ and $y_I(x)$ differ only by a constant term. When we use (2.1) and the Lagrangian form of the interpolating polynomials, it is easy to show that

$$(2.5)\qquad S(x)-y_I(x)=\{f_{n+1}-f^p_{n+1}\}\frac{\int_{x_n}^{x_{n+1}}\prod_{i=1}^{k}(t-x_{n+1-i})\,dt}{\prod_{i=1}^{k}(x_{n+1}-x_{n+1-i})}.$$

Since $y_I(x_n)\neq y_n$ and $S(x_{n+1})\neq y_{n+1}$, the corresponding piecewise polynomials over $[a, b]$ both have jumps at the meshpoints, although the first derivatives are continuous.

The interpolant $y_I(x)$ was implemented in the original Shampine–Gordon code and in early versions of the DEPAC Adams code. However, the latest version of DEPAC uses a higher degree polynomial interpolant due to Watts and Shampine [12] that has better continuity properties. This interpolant, which is a polynomial of degree $\leqq k+2$ over $[x_n, x_{n+1}]$, can be written

$$T(x)=S(x)+[y_{n+1}-S(x_{n+1})]\frac{\Phi(x)}{\Phi(x_{n+1})}$$

where

$$\Phi(x)=\int_{x_n}^{x}\prod_{i=0}^{k}(t-x_{n+1-i})\,dt.$$

It is clear that $T(x)$ satisfies the conditions

$$T(x_n)=y_n,\qquad T(x_{n+1})=y_{n+1},$$
$$T'(x_n)=f_n,\qquad T'(x_{n+1})=f_{n+1},$$

and consequently can be used to provide a global approximation to $y(x)$ that is continuously differentiable. We may regard $T(x)$ as a slightly perturbed version of $S(x)$ (or $y_I(x)$) with the small perturbation added in order to force C^1 continuity.

We will denote the defect in $\eta(x)$ by $\delta^\eta(x)$, that is,

$$\delta^\eta(x)=\eta'(x)-f(x, \eta(x)).$$

Similarly, $\delta^{y_I}(x)$, $\delta^S(x)$ and $\delta^T(x)$ will denote the defects in $y_I(x)$, $S(x)$ and $T(x)$, respectively. We are mainly concerned with the defect in $T(x)$, since this is the interpolant currently used in DEPAC, but it will prove helpful and enlightening if we also look at $\delta^\eta(x)$, $\delta^{y_I}(x)$ and $\delta^S(x)$. In performing an asymptotic analysis, we make use of some results of Stetter [9]. Although Stetter's main concern was to assess the quality of the interpolants themselves, his analysis can be adapted in a straightforward manner to give the desired information about the behaviour of the defect.

Consider taking a step of length h_{n+1} at order k from x_n to $x_{n+1}=x_n+h_{n+1}$. The PECE formula uses information from the past k meshpoints $\{x_{n+1-k}, \cdots, x_n\}$. We define the quantities σ_i by

$$(2.6)\qquad x_{n+1-i}=x_n+\sigma_i h_{n+1}.$$

For the step-changing strategy used in [7], the $\{\sigma_i\}_{i=0}^{k}$ can be bounded above and below. This is discussed in more detail in the next section. We define the local solution $z(x)$ by

$$z'(x)=f(x, z(x)),\qquad z(x_n)=y_n,$$

and the local error for the step as $y_{n+1}-z(x_{n+1})$. For clarity we will drop the subscript $n+1$ from h_{n+1} in the succeeding analysis. The basic approach in [9] is to assume that the stepsize and error control mechanism keeps the local error at the same level over $[x_{n+1-k}, x_{n+1}]$. Then, assuming that f is sufficiently smooth, standard interpolation theory is used to show that

$$p_{n+1} - z(x_{n+1}) = O(h^{k+1}) \tag{2.7}$$

and

$$y_{n+1} - z(x_{n+1}) = O(h^{k+2}). \tag{2.8}$$

Furthermore, at a point $x \in [x_n, x_{n+1}]$,

$$\eta(x) - z(x) = O(h^{k+2}) \tag{2.9}$$

and

$$S(x) - z(x) = O(h^{k+2}). \tag{2.10}$$

We can use (2.7)-(2.10) to deduce similar results for $y_I(x)$ and $T(x)$. From (2.5),

$$S(x) - y_I(x) = \{f^*_{n+1} - f^p_{n+1}\} h \frac{\int_0^1 \prod_{i=1}^k (\alpha - \sigma_i)\, d\alpha}{\prod_{i=1}^k (1 - \sigma_i)}.$$

Employing a Lipschitz condition on f, (2.7) and (2.8), and the fact that $\sigma_1, \sigma_2, \cdots, \sigma_k$ are nonpositive and bounded below it follows that

$$S(x) - y_I(x) = O(h^{k+2}). \tag{2.1}$$

Hence, from (2.10)

$$y_I(x) - z(x) = O(h^{k+2}). \tag{2.12}$$

Also, writing $x = x_n + sh$, we have

$$\begin{aligned} T(x) - S(x) &= [y_{n+1} - S(x_{n+1})] \frac{\int_0^s \prod_{i=0}^k (\alpha - \sigma_i)\, d\alpha}{\int_0^1 \prod_{i=0}^k (\alpha - \sigma_i)\, d\alpha} \\ &= O(\| y_{n+1} - S(x_{n+1}) \|) \\ &= O(h^{k+2}), \end{aligned}$$

from (2.8) and (2.10), and hence

$$T(x) - z(x) = O(h^{k+2}).$$

Since we are assuming that f is Lipschitzian, we have, using (2.9) and (2.2a),

$$\begin{aligned} \delta^\eta(x) &= \eta'(x) - f(x, \eta(x)) \\ &= \eta'(x) - z'(x) + f(x, z(x)) - f(x, \eta(x)) \\ &= P^*_{k+1,n}(x) - z'(x) + O(h^{k+2}). \end{aligned}$$

Using an expansion of $P^*_{k+1,n}(x) - z'(x)$ from [9] we find that

$$\begin{aligned} \delta^\eta(x_n + sh) = -h^{k+1} \Bigg\{ & \pi(s) \frac{z^{(k+2)}(x_{n+1})}{(k+1)!} \\ & + \hat{\pi}(s) \frac{\bar{g}_{n+1} f_y(x_{n+1}, z(x_{n+1})) z^{(k+1)}(x_{n-1})}{\hat{\pi}(1)} \Bigg\} + O(h^{k+2}) \end{aligned} \tag{2.13}$$

where

$$\pi(s) = \prod_{i=0}^{k} (s - \sigma_i), \quad \hat{\pi}(s) = \prod_{i=1}^{k} (s - \sigma_i), \quad \bar{g}_{n+1} = \frac{1}{k!} \int_0^1 \hat{\pi}(\alpha)\, d\alpha,$$

and f_y is the Jacobian of f. Similarly,

$$\begin{aligned}\delta^S(x) &= S'(x) - z'(x) + f(x, z(x)) - f(x, S(x))\\ &= P_{k+1,n+1}(x) - z'(x) + O(h^{k+2}),\end{aligned}$$

and using Stetter's asymptotic result for $P_{k+1,n+1}(x) - z'(x)$ we have

$$\delta^S(x_n + sh) = -h^{k+1}\pi(s)\frac{z^{(k+2)}(x_{n+1})}{(k+1)!} + O(h^{k+2}). \tag{2.14}$$

Asymptotically, the defect in $y_I(x)$ is the same as that in $\delta^S(x)$. This follows from

$$\delta^{y_I}(x) = y_I'(x) - z'(x) + O(h^{k+2})$$

and

$$y_I'(x) = S'(x) = P_{k+1,n+1}(x).$$

Finally, an expression for $\delta^T(x)$ can be deduced from (2.14). We have

$$\begin{aligned}\delta^T(x) &= T'(x) - z'(x) + O(h^{k+2})\\ &= S'(x) + [y_{n+1} - S(x_{n+1})]\frac{\Phi'(x)}{\Phi(x_{n+1})} - z'(x) + O(h^{k+2})\\ &= \delta^S(x) + [y_{n+1} - S(x_{n+1})]\frac{\Phi'(x)}{\Phi(x_{n+1})} + O(h^{k+2}).\end{aligned}$$

Using (2.14) this may be written

$$\delta^T(x_n + sh) = -h^{k+1}(s)\left\{\frac{z^{(k+2)}(x_{n+1})}{(k+1)!} - \frac{[y_{n+1} - S(x_{n+1})]}{\Phi(x_{n+1})}\right\} + O(h^{k+2}). \tag{2.15}$$

The key point arising from this analysis is that the defects in $S(x)$, $y_I(x)$ and $T(x)$ are all of the form

$$\delta(x_n + sh) = A(h)\pi(s) + O(h^{k+2}) \tag{2.16}$$

where $A(h)$ is $O(h^{k+1})$ and independent of s. Recalling that $\pi(s) = \prod_{i=0}^{k}(s - \sigma_i)$, we see that, asymptotically, the shape of the defect over $[x_n, x_{n+1}]$ is determined by the local stepsize pattern only. The practical importance of this result is that by sampling the defect at an internal point $x_n + \hat{s}h$, an approximation $\delta(x_n + \hat{s}h)/\pi(\hat{s}) \approx A(h)$ can be formed. Hence an estimate of $\delta(x_n + sh)$ for $s \in [0, 1]$ is available and so any desired measure of the defect can be estimated. In particular, to approximate $\max_{[0,1]}\|\sigma(x_n + sh)\|$ for any vector norm $\|\cdot\|$, we simply form $\|\delta(x_n + s^*h)\|$ where s^* in $[0, 1]$ maximises $|\pi(s)|$. This approach is discussed in more detail in the next section. We mention that the idea of estimating the maximum defect in $y_I(x)$ by sampling at a single point was proposed in [7, p. 122]. Shampine and Gordon suggest sampling at the midpoint on heuristic grounds; the defect is zero at the two meshpoints. We will show that whilst this is never a bad choice, it is only optimal when the lowest order $k = 1$ is used for the step.

The expression for $\delta^{\eta}(x)$ in (2.13) is of a slightly different form from (2.16). The leading term in (2.13) is a problem-dependent linear combination of $\pi(s)$ and $\hat{\pi}(s)$. In this case two samples would be needed to form an estimate of the defect over a step.

To conclude this section we mention some related work by Hanson and Enright [5]. In its simplest form, the error central mechanism used in [7] ensures that on each step

$$\|y_{n+1} - y_{n+1}(k)\|_2 \leqq \text{TOL} \tag{2.17}$$

for some user-supplied parameter TOL. Here $y_{n+1}(k)$ denotes the result of an Adams-Bashforth predictor of order k followed by an Adams-Moulton corrector of order k. Under certain reasonable assumptions, Hanson and Enright show that this strategy indirectly controls the defect in $S(x)$ and $\eta(x)$, in the sense that

$$\|\delta(x)\| \leqq (C + O(h_{\max}))\text{TOL} \quad \text{for all } x \in [a, b]$$

where C is a constant that depends on f, and $h_{\max} = \max_{n \geq 0} h_{n+1}$. We note that a similar relationship can be shown to hold for the more recent interpolant $T(x)$. This follows because $\delta^T(x)$ is an $O(h^{k+1})$ perturbation of $\delta^S(x)$ and the left-hand side of (2.17) is also $O(h^{k+1})$. The numerical tests of § 5 confirm that the tolerance parameter can be used to control the defect in $T(x)$, albeit in a problem-dependent manner.

3. Location of s^*. In the last section we have shown that an asymptotically valid estimate of $\max_{[0,1]} \|\delta^T(x_n + sh_{n+1})\|$ is given by sampling at $x_n + s^* h_{n+1}$, where s^*, which we will call the optimal sample point, satisfies $|\pi(s^*)| = \max_{[0,1]} |\pi(s)|$. We now examine the practical details of computing s^* and gain some insight into the behaviour of s^* in the course of an integration.

Recall that after a kth order step from x_n to $x_n + h_{n+1}$,

$$\pi(s) = \prod_{i=0}^{k} (s - \sigma_i)$$

where

$$\sigma_i = \frac{x_{n+1-i} - x_n}{h_{n+1}}.$$

Note that $\sigma_k < \sigma_{k-1} < \cdots < \sigma_1 = 0 < \sigma_0 = 1$. By Rolle's Theorem, $\pi'(s)$ has roots in (σ_i, σ_{i-1}), for $i = 1, \cdots, k$, and because $\pi'(s)$ is a polynomial of degree k, each interval must contain exactly one root. Since $\pi(0) = \pi(1) = 0$ the point s^* that we require is the root of $\pi'(s)$ in $(0, 1)$. To compute s^*, given that $\pi'(s)$ and $\pi''(s)$ are inexpensive to evaluate, a straightforward application of Newton's method is sufficient. The fact that s^* is the largest root of a polynomial with purely real roots implies that, providing the initial guess is greater than s^*, the Newton iterates converge monotononically and quadratically to s^* (see [10, p. 272]). In particular, convergence is guaranteed if we start at the point $s^{(0)} = 1$, but later we will show that a better choice is available.

The following theorem gives a useful characterization of s^*.

THEOREM 1. *For $k \geqq 2$, the point s^* defined above is the solution in $(0, 1)$ of*

$$\frac{1-2s}{s(s-1)} = \sum_{i=2}^{k} \frac{1}{s - \sigma_i}, \tag{3.1}$$

and for $k = 1$, $s^* = \frac{1}{2}$.

Proof. When $k = 1$, $\pi(s) = (s-1)s$ and the result is trivial. When $k \geqq 2$ we have

$$\pi(s) = (s-1)s(s-\sigma_2)\cdots(s-\sigma_k),$$

so that

$$\begin{aligned}\pi'(s) &= s(s-\sigma_2)\cdots(s-\sigma_k) + (s-1)(s-\sigma_2)\cdots(s-\sigma_k) \\ &\quad + \sum_{i=2}^{k} \frac{(s-1)s(s-\sigma_2)\cdots(s-\sigma_k)}{s-\sigma_i}\end{aligned} \tag{3.2}$$

in $(0, 1)$. Defining

$$g(s) = \frac{\pi'(s)}{(s-\sigma_2)\cdots(s-\sigma_k)}, \tag{3.3}$$

we note that the zeros of $g(s)$ in $(0, 1)$ are precisely the zeros of $\pi'(s)$ in $(0, 1)$. From (3.2) and (3.3),

$$g(s)=2s-1+\sum_{i=2}^{k}\frac{s(s-1)}{s-\sigma_i}$$

and the result follows. □

The left-hand side of (3.1), which is independent of $\{\sigma_i\}_{i=2}^{k}$, is plotted in Fig. 1. Since the right-hand side is positive in $(0, 1)$, it follows that $s^*>1/2$ for $k\geqq 2$. To obtain more precise bounds for s^* we must use information about the step-changing mechanism. First we give a corollary to the theorem.

COROLLARY 1. *For $k\geqq 2$, if the two sets $\{\hat{\sigma}_i\}_{i=2}^{k}$ and $\{\bar{\sigma}_i\}_{i=2}^{k}$ satisfy*

$$\hat{\sigma}_i\leqq\bar{\sigma}_i<0, \qquad i=2,\cdots,k,$$

then the corresponding optimal sample points, $\hat{s}^*$ and $\bar{s}^*$, satisfy $\hat{s}^*\leqq\bar{s}^*$.

Proof. For $s\in(0, 1)$,

$$\sum_{i=2}^{k}\frac{1}{s-\hat{\sigma}_i}\leqq\sum_{i=2}^{k}\frac{1}{s-\bar{\sigma}_i}. \tag{3.4}$$

The intersection of $\sum_{i=2}^{k}1/(s-\hat{\sigma}_i)$ and $(1-2s)/s(s-1)$ will therefore be to the left of the intersection of $\sum_{i=2}^{k}1/(s-\bar{\sigma}_i)$ and $(1-2s)/s(s-1)$ (see Fig. 1). □

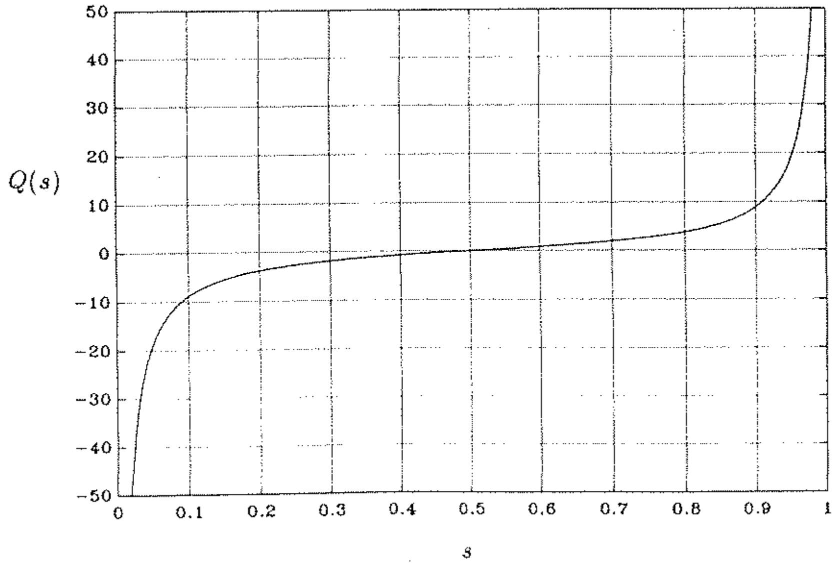

FIG. 1. $Q(s)=(1-2s)/s(s-1)$.

The code we are considering uses a variable order $1\leqq k\leqq 12$ and a variable stepsize. However, for several reasons (an important one being to ensure numerical stability) the ratio of successive stepsizes is bounded in the following way [7, p. 64]:

$$\frac{1}{2}\leqq\frac{h_{i+1}}{h_i}\leqq 8$$

for all i, except, possibly, those for which the step involving h_i is taken at order one. The right-hand inequality is extremely unlikely to be close to equality in practice [7, p. 117]. From the definition of $\{\sigma_i\}_{i=0}^{k}$ it follows that on a general step of order k we have

$$\sigma_i^l \leqq \sigma_i \leqq \sigma_i^u, \qquad i=0, \cdots, k$$

where the upper and lower bounds are defined by

$$\sigma_i^u = \frac{-(2^{i-1}-1)}{2^{i-1}} \quad \text{and} \quad \sigma_i^l = \frac{-(8^i-8)}{7}.$$

Denoting the corresponding optimal sample points by s^{*u} and s^{*l}, it follows from the corollary that

$$s^{*l} \leqq s^* \leqq s^{*u}.$$

The values of s^{*l} and s^{*u} for $k=1, \cdots, 12$ are given in Table 1. Note that since s^{*u} is an upper bound for s^*, it may be used as the initial guess in the Newton iteration that we discussed earlier.

TABLE 1
Values of s^{*l}, s^{*u} and $s^{*\text{const}}$ at order k.

k	1	2	3	4	5	6	7	8	9	10	11	12
s^{*l}	.500	.515	.516	.517	.517	.517	.517	.517	.517	.517	.517	.517
s^{*u}	.500	.608	.672	.717	.750	.777	.799	.817	.832	.844	.855	.865
$s^{*\text{const}}$	.500	.577	.618	.644	.663	.678	.690	.699	.708	.715	.721	.726

In addition to bounding the extreme values of s^*, it is possible to say something about the typical behaviour of s^*. Although the code allows the stepsize to vary, it is strongly biased against doing so. This reduces the overheads involved in the formation of p_{n+1} and y_{n+1}. Another reason for the bias is that the order changing mechanism is set up so that an increase in order is only considered if the past $k+1$ steplengths are equal. Consequently, the overall step pattern can usually be divided into a small number of groups each with a constant steplength. When $h_{n+2-k}=h_{n+3-k}=\cdots=h_{n+1}$, the sigma values become

$$\sigma_i = \sigma_i^{\text{const}} = 1-i, \qquad i=0, \cdots, k,$$

and the corresponding $s^{*\text{const}}$ values appear in Table 1.

4. The free estimate. The cost of taking a step with a discrete ODE method is often measured in terms of the number of f evaluations required. Sampling the defect at one intermediate point clearly increases this quantity by one. In the case of Runge–Kutta-interpolation schemes, where a large number (usually $\geqq 6$) of f evaluations are needed for each step, this increase in cost is relatively small. However, the PECE method considered here uses only two evaluations and so the cost of sampling on each step is more significant. In this section we show that it is possible to form an estimate of $\delta^T(x)$ without using any additional f evaluations. The basic idea is to approximate the leading term in the expansion (2.15) of $\delta^T(x_n+sh)$. In particular, to obtain a valid estimate of $\max_{[0,1]} \|\delta^T(x_n+sh)\|$ it is sufficient to form an $O(h)$ accurate approximation to

$$h^{k+1}\pi(s^*)\left\{\frac{z^{(k+2)}(x_{n+1})}{(k+1)!}-\frac{[y_{n+1}-S(x_{n+1})]}{\Phi(x_{n+1})}\right\}, \tag{4.1}$$

and take its norm. To this end we consider below the computation of $\Phi(x_{n+1})$, $y_{n+1}-S(x_{n+1})$ and an approximation to $z^{(k+2)}(x_{n+1})$. To keep the overall technique as simple as possible we will attempt to make use of quantities that have already been computed during the course of a step.

From Watts and Shampine [12, p. 341], $\Phi(x_{n+1})$ may be written

$$\Phi(x_{n+1}) = -h^2 g_{k+1,2}(1) \prod_{i=1}^{k} (x_{n+1}-x_{n+1-i})$$

where $g_{k+1,2}(1)$ is a certain double integral. The values of $\{x_{n+1}-x_{n+1-i}\}_{i=1}^{k}$ are immediately available, and the $g_{k+1,2}(1)$ term can be readily computed from quantities that are used in the code. In fact $g_{k+1,2(1)}$ is formed in the interpolation routine, SINTRP, where it is called GDI (for more details see [11]).

The vector y_{n+1} is, of course, already available and $S(x_{n+1})$ can be formed in a straightforward way (see [11]). However, the computed difference $y_{n+1}-S(x_{n+1})$ is liable to suffer from catastrophic cancellation at stringent tolerances. To see this, we observe from (2.8) and (2.10) that $y_{n+1}-S(x_{n+1})$ is $O(h^{k+2})$, which is one order higher than the quantity being controlled by TOL in (2.17). Putting $x=x_{n+1}$ in (2.5) we obtain the alternative representation

$$y_{n+1}-S(x_{n+1}) = \{f_{n+1}^{p}-f_{n+1}\}\frac{\int_{x_n}^{x_{n+1}} \prod_{i=1}^{k} (t-x_{n+1-i})\,dt}{\prod_{i=1}^{k} (x_{n+1}-x_{n+1-i})},$$

which may be written

$$y_{n+1}-S(x_{n+1}) = \{f_{n+1}^{p}-f_{n+1}\}hG(k+1)$$

where $G(k+1)$ is computed in the code. The difference $f_{n+1}^{p}-f_{n+1}$ is $O(h^{k+1})$ and so this representation should be less susceptible to cancellation.

To produce an estimate for $z^{(k+2)}(x_{n+1})$ we turn again to Stetter's analysis [9]. Under the assumption that the stepsize and error control mechanism has kept the local error at the same level over $[x_{n+1-k}, x_{n+1}]$, he concludes that

$$f_{n+1-i} = z'(x_{n+1-i}) + O(h^{k+2}), \qquad i=0,\cdots,k. \tag{4.2}$$

We need to extend the assumption to cover $[x_{n-k}, x_{n+1}]$ so that

$$f_{n-k} = z'(x_{n-k}) + O(h^{k+2}) \tag{4.3}$$

also holds. We use the standard notation $f[\;]$ and $z'[\;]$ for divided differences based on $\{f_{n+1-i}\}_{i=0}^{k+1}$ and $\{z'(x_{n+1-i})\}_{i=0}^{k+1}$, respectively; that is, $f[x_{n-k}]=f_{n-k}$ and

$$f[x_{n-k},\cdots,x_{n-k+i}] = \frac{f[x_{n-k},\cdots,x_{n-k+i-1}]-f[x_{n-k+1},\cdots,x_{n-k+i}]}{x_{n-k}-x_{n-k+i}},$$

$$i=1,\cdots,k+1.$$

Using classical interpolation theory (see, for example, [7, p. 39]), we have

$$z'[x_{n-k},\cdots,x_{n+1}] = \frac{z^{(k+2)}(x_{n+1})}{(k+1)!}+O(h),$$

for sufficiently smooth z. Now, from (4.2) and (4.3), it can be proved by induction that

$$f[x_{n-k},\cdots,x_{n+1}] - z'[x_{n-k},\cdots,x_{n+1}] = O(h).$$

Hence

$$f[x_{n-k},\cdots,x_{n+1}] = \frac{z^{(k+2)}(x_{n+1})}{(k+1)!}+O(h),$$

and so it is reasonable to use the divided difference $f[x_{n-k}, \cdots, x_{n+1}]$ to approximate $z^{(k+2)}(x_{n+1})/(k+1)!$ in (4.1). The DEPAC routine requires $f[x_{n-k}, \cdots, x_{n+1}]$ when deciding whether to increase the order. An order increase is only considered when the step size is locally constant, that is when $h_{n+1-k} = h_{n+2-k} \cdots = h_{n+1}$, so the divided difference is only available from the code in this situation. However, a scaled version of $f[x_{n+1-k}, \cdots, x_{n+1}]$ is always available and hence, by storing divided difference information from the previous step, the required quantity can be computed at little cost.

Finally we point out that there will be a small number of steps at the start of the integration on which $f[x_{n-k}, \cdots, x_{n+1}]$ is not directly obtainable. On the first step from x_0 to x_1, for which the order k is always equal to one, we require $f[x_{-1}, x_0, x_1]$. This term involves a function evaluation f_{-1} that has not actually be made. Similarly if the order is increased to two for the next step, then at the end of that step we need $f[x_{-1}, x_0, x_1, x_2]$. Generally, f_{-1} remains until we step from x_n to x_{n+1} with order $k \leqq n$. In the numerical testing described in the next section, rather than computing an extra f value to act as f_{-1} we refrained from forming an estimate of the defect until $n \geqq k$.

5. Numerical results. The techniques we have described produce estimates of the defect that are valid asymptotically. In this section we investigate the accuracy of the estimates in practical computations. To do this we used the DEPAC Adams package in "single-step" mode by making repeated calls to the routines STEPS and SINTRP. STEPS is the routine that advances the solution by a single step and, once a step has been taken, SINTRP can be called to evaluate $T(x)$ and $T'(x)$ at some point within the step. After advancing from x_n to $x_n + h$ we formed

$$\mathrm{DEF}_n := \max_{0 \leqq j \leqq 100} \|\delta^T(x_n + jh/100)\|_\infty$$

as a discrete approximation to $\max_{s\in[0,1]} \|\delta^T(x_n + sh)\|_\infty$, along with

$$\mathrm{SAM}_n := \|\delta^T(x_n + s^* h)\|_\infty,$$

and

$$(5.1)\,\mathrm{APPROX}_n := \left\| h^{k+1}\pi(s^*)\left[f[x_{n-k}, \cdots, x_{n+1}] + \frac{\{f^p_{n+1} - f_{n+1}\}G(k+1)}{hg_{k+1,2}(1)\prod_{i=1}^{k}(x_{n+1} - x_{n+1-i})}\right]\right\|_\infty,$$

as discussed in §§ 3 and 4. We recorded the quantities

$$\mathrm{DEFMAX} := \max_{n \geqq 0} \{\mathrm{DEF}_n\},$$

$$\mathrm{SAMMAX} := \max_{n \geqq 0} \left\{\frac{\mathrm{DEF}_n}{\mathrm{SAM}_n}\right\},$$

$$\mathrm{AUNDERMAX} := \max_{n \geqq 0} \left\{\frac{\mathrm{DEF}_n}{\mathrm{APPROX}_n}\right\},$$

$$\mathrm{AOVERMAX} := \max_{n \geqq 0} \left\{\frac{\mathrm{APPROX}_n}{\mathrm{DEF}_n}\right\}.$$

Hence SAMMAX measures the worst case of the sampled value underestimating DEF_n. Similarly, AUNDERMAX and AOVERMAX measure the extreme cases of the free approximation under- and overestimating DEF_n. We would like all three values to be as close to one as possible.

We mention that in order to compute APPROX_n using (5.1) an extra parameter, used to return f^p_{n+1} to the driver program, was added to the call list of STEPS. No other changes were made to STEPS.

The following test problems were used:

(i) The orbit equations [3, Class D]:

$$\begin{aligned} y_1' &= y_3, & y_1(0) &= 1-\varepsilon, \\ y_2' &= y_4, & y_2(0) &= 0, \\ y_3' &= \frac{-y_1}{(y_1^2+y_2^2)^{3/2}}, & y_3(0) &= 0, \\ y_4' &= \frac{-y_2}{(y_1^2+y_2^2)^{3/2}}, & y_4(0) &= \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^{1/2}, \end{aligned}$$

$$0 \leqq x \leqq 20,$$

where values of .1, .5 and .9 were chosen for the eccentricity parameter ε.

(ii) A logistic curve [3, Problem A4]:

$$y' = \frac{y}{4}\left(1-\frac{y}{20}\right), \quad y(0)=1, \quad 0 \leqq x \leqq 20.$$

(iii) A problem due to Fehlberg that is used in [4, p. 174]:

$$\begin{aligned} y_1' &= 2xy_1 \log(\max(y_2, 10^{-3})), & y_1(0) &= 1, \\ y_2' &= -2xy_2 \log(\max(y_1, 10^{-3})), & y_2(0) &= e, \end{aligned}$$

$$0 \leqq x \leqq 5.$$

In each case an absolute error criterion was specified with tolerances $\text{TOL} = 10^{-2}, 10^{-3}, \cdots, 10^{-8}$. Results are presented in Tables 2–6. We see that the SAMMAX values are extremely good with the maximum over all problems and tolerances being 1.34. In fact on the vast majority of steps the ratio $\text{DEF}_n/\text{SAM}_n$ was less than 1.01. The sample

TABLE 2
Orbit problem, $\varepsilon = .1$.

TOL	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
DEFMAX	9E−2	1E−2	2E−3	2E−4	2E−5	2E−6	4E−7
SAMMAX	1.01	1.00	1.01	1.01	1.02	1.02	1.01
AUNDERMAX	1.50	2.18	2.36	2.24	4.32	3.28	5.03
AOVERMAX	1.78	2.77	2.48	5.44	3.97	2.81	2.67

TABLE 3
Orbit problem, $\varepsilon = .5$.

TOL	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
DEFMAX	2E−1	3E−2	6E−3	5E−4	5E−5	6E−6	4E−7
SAMMAX	1.04	1.02	1.02	1.01	1.03	1.02	1.02
AUNDERMAX	3.47	2.66	8.59	6.41	5.05	8.18	5.75
AOVERMAX	2.64	3.36	3.82	3.95	5.19	4.51	5.63

TABLE 4
Orbit problem, $\epsilon = .9$.

TOL	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
DEFMAX	3E+0	3E−1	9E−2	1E−2	8E−4	7E−5	7E−6
SAMMAX	1.09	1.02	1.02	1.01	1.02	1.02	1.15
AUNDERMAX	5.23	21.33	12.58	19.21	18.38	15.30	8.69
AOVERMAX	6.11	5.46	9.15	7.16	13.03	38.85	10.51

TABLE 5
Logistic curve.

TOL	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
DEFMAX	7E−3	5E−4	5E−5	7E−6	6E−7	7E−8	7E−9
SAMMAX	1.01	1.00	1.28	1.01	1.05	1.06	1.34
AUNDERMAX	11.05	1.58	9.62	2.20	5.77	9.85	4.11
AOVERMAX	1.32	7.36	9.29	1.84	1.61	8.72	5.64

TABLE 6
Fehlberg problem.

TOL	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
DEFMAX	8E−1	1E−1	2E−2	1E−3	2E−4	2E−5	1E−6
SAMMAX	1.09	1.03	1.03	1.11	1.25	1.04	1.21
AUNDERMAX	5.35	6.89	14.07	26.06	11.89	9.32	47.38
AOVERMAX	3.12	2.63	2.44	9.35	10.83	7.79	13.74

point s^* for each step was found to full accuracy (unit roundoff $\approx 2\times 10^{-16}$) in the manner described in § 3. The average number of Newton iterations required was five and the maximum was six. The free approximation, whilst being less accurate, could usually be relied on to give the correct order of magnitude of the defect over a step. At more stringent tolerances of 10^{-9} and 10^{-10}, however, we found that the AUNDERMAX and AOVERMAX values began to worsen considerably. This was caused by severe cancellation in the formation of $f_{n+1}^{p}-f_{n+1}$. The problem only arose during the first few steps of an integration; here the code appeared to use conservatively small stepsizes.

Throughout the testing we also sampled the defect at the midpoint of each step, as suggested in [7], and computed

$$\text{MIDPTMAX} = \max_{n\geq 0}\left\{\frac{\text{DEF}_n}{\|\delta^T(x_n+.5h)\|_\infty}\right\}.$$

We found that although the midpoint sample never gave a very bad estimate of DEF_n, MIDPTMAX was always greater than the corresponding SAMMAX value and had an average of 1.9. Generally, sampling at $x_n+.5h$ was least successful on steps where the order k was high. This is to be expected from Table 1; at higher orders s^* is likely to be further from .5. Finally we mention that DEFMAX is seen to be controlled by TOL as predicted at the end of § 2.

In summary, we have given an asymptotic analysis of the defect associated with some commonly used Adams-interpolation procedures and have presented two techniques for estimating the defect in the DEPAC implementation. A reasonable order-of-magnitude estimate can be formed at little cost, and for the price of one f evaluation per step an extremely accurate alternative is available. As well as being of interest in their own right, these estimates may prove useful in the development of an error-control scheme of the type discussed in [1], [2], [6] were the defect, rather than the local error, is directly controlled.

Acknowledgments. Helpful comments from Nick Higham and George Hall improved this manuscript.

REFERENCES

[1] W. H. ENRIGHT, *A new error-control for initial value solvers*, Numerical Analysis Report No. 122, University of Manchester, U.K., 1986.

[2] ———, *Analysis of error control strategies for continuous Runge–Kutta methods*, Technical Report No. 205/87, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, 1987.

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

[4] E. HAIRER, S. P. NORSETT, AND G. WANNER, *Solving Ordinary Differential Equations* I, Springer-Verlag, Berlin, 1987.

[5] P. M. HANSON AND W. H. ENRIGHT, *Controlling the defect in existing variable-order Adams codes for initial-value problems*, ACM Trans. Math. Software, 9 (1983), pp. 71–97.

[6] D. J. HIGHAM, *Robust defect control with Runge–Kutta schemes*, Numerical Analysis Report No. 150, University of Manchester, U.K., 1987.

[7] L. F. SHAMPINE AND M. K. GORDON, *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, W. H. Freeman, San Francisco, CA, 1975.

[8] L. F. SHAMPINE AND H. A. WATTS, DEPAC—*Design of a user oriented package of* ODE *solvers*, Report SAND79-2374, Sandia National Laboratories, Albuquerque, NM, 1979.

[9] H. J. STETTER, *Interpolation and error estimation in Adams* PC-*Codes*, SIAM J. Numer. Anal., 16 (1979), pp. 311–323.

[10] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[11] H. A. WATTS, *A smoother interpolant for* DE/STEP, INTRP *and* DEABM: II, Report SAND84.0293, Sandia National Laboratories, Albuquerque, NM, 1984.

[12] H. A. WATTS AND L. F. SHAMPINE, *Smoother interpolants for Adams codes*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 334–345.