MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 140
OCTOBER 1977, PAGES 939–953

Error Estimates for a Stiff Differential Equation Procedure

By R. Sacks-Davis

Abstract. For numerical procedures which solve stiff systems of ordinary differential equations there are problems associated with estimating the local error. In this paper an analysis based on the linear model $y' = Ay$ is carried out for a particular method based on second derivative formulas. It is shown that there exists an error estimate based on a comparison between predicted and corrected values which is both reliable and efficient.

1. Introduction. There are a number of special problems associated with the numerical solution of stiff ordinary differential equations. Only formulas whose stability regions cover large areas of the left half plane may be used. In order to solve the ensuing implicit set of equations a modified Newton-Raphson scheme is used rather than the simple iterative method associated with the nonstiff problems.

For predictor-corrector methods there are difficulties associated with the usual error estimate based on a comparison between the predicted and corrected values. Fast transients can cause this error estimate to severely overestimate the true error. This does not harm the reliability of the method so much as the efficiency since the choices of stepsize are usually based on the estimate of the error.

In this paper the problems associated with error estimation are investigated from the point of view of effectiveness theory using the linear model $y' = Ay$. For a special class of methods, namely those based on second derivative formulas, it is shown that there exists a simple error estimate based on a comparison between predicted and corrected values which is both reliable and efficient even if A has eigenvalues with very negative real parts. A typical effectiveness theorem for this error estimate is proved and some numerical results are given.

2. Second Derivative Methods. We consider the following autonomous system of ordinary differential equations

$$y' = f(y(t)), \qquad y(0) = y_0, \qquad 0 \leqslant t \leqslant b.$$

At previous steps $t = t_{n-1}, t_{n-2}, \ldots, t_0 = 0$, we have approximations y_{n-j} to the true solution $y(t_{n-j})$ as well as approximations $f_{n-j} = f(y_{n-j})$ and $f'_{n-j} = f'(y_{n-j})$ to $f(y(t_{n-j}))$ and $f'(y(t_{n-j}))$, respectively. We are required to advance the solution from t_{n-1} to t_n with stepsize h_n. The basic equations were derived in [4]. Let

Received October 28, 1976.
AMS (MOS) subject classifications (1970). Primary 65L05.

$$q_j(t) = \prod_{i=1}^{j} (t - t_{n-i}), \qquad j \geqslant 1, \qquad q_0(t) = 1,$$

and

$$g_{i,j} = \int_{t_{n-1}}^{t_n} (t - t_n)^i q_j(t)\, dt.$$

The $g_{i,j}$ satisfy the recurrence relation

$$g_{i,j} = (t_n - t_{n-j})g_{i,j-1} + g_{i+1,j-1}. \tag{1}$$

The predicted value $y_{n,0}$ of the solution at $t = t_n$ may be calculated in terms of past values at k previous points using the formula

$$y_{n,0} = y_{n-1} + hf_{n-1} + \sum_{j=0}^{k-1} (g_{1,j} + h_n g_{0,j})[f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-j-1}].$$

Similarly predicted values for the first and second derivatives of the solution may be expressed in terms of previous values. The required formulae are

$$f_{n,0} = f_{n-1} + h_n \sum_{j=0}^{k-1} q_j(t_n)[f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-j-1}] \tag{2}$$

and

$$f'_{n,0} = f'_{n-1} + \sum_{j=1}^{k-1} \{(h_n q'_j(t_n) + q_j(t_n))[f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-j-1}]\}.$$

The divided differences used in the above equations may be determined from the known values $f_{n-1}, f'_{n-1}, f_{n-2}, f_{n-3}, \ldots, f_{n-k}$. The accepted approximation y_n to the solution at $t = t_n$ is found by solving the implicit equation

$$y_n = y_{n,0} + h_n\beta_{n,0}(f(y_n) - f_{n,0}) + h_n^2\gamma_{n,0}(f'(y_n) - f'_{n,0}), \tag{3}$$

where

$$h_n\beta_{n,0} = \frac{1}{q_{k-1}(t_n)} \left[g_{0,k-1} - \frac{q'_{k-1}(t_n)}{q_{k-1}(t_n)}\, g_{1,k-1} \right] \tag{4}$$

and

$$h_n^2\gamma_{n,0} = \frac{g_{1,k-1}}{q_{k-1}(t_n)}. \tag{5}$$

There is a useful alternative expression for y_n. Let $P(t)$ be the Hermite polynomial of least degree uniquely defined by

$$P(t_{n-j}) = f_{n-j}, \qquad j = 1, 2, \ldots, k-1,$$

$$P(t_n) = f(y_n),$$

$$P'(t_n) = f'(y_n).$$

Then

$$y_n = y_{n-1} + \int_{t_{n-1}}^{t_n} P(t)\,dt \tag{6}$$

or

$$y_n = y_{n-1} + h_n \sum_{j=0}^{k-1} \beta_{n,j} f_{n-j} + h_n^2 \gamma_{n,0} f'_n.$$

Similarly, for the predicted value $y_{n,0}$ we have

$$y_{n,0} = y_{n-1} + \int_{t_{n-1}}^{t_n} P_0(t)\,dt, \tag{7}$$

where $P_0(t)$ is the Hermite polynomial of least degree interpolating the known points

$$P_0(t_{n-j}) = f_{n-j}, \qquad j = 1, 2, \ldots, k,$$

$$P'_0(t_{n-1}) = f'_{n-1}.$$

3. Estimation of the Local Error. In this section we will consider a number of alteratives for estimating the local error. The local error of a multistep method in stepping from t_{n-1} to t_n is defined as

$$le_n = y_{n-1}(t_n) - y_n,$$

where

$$y'_{n-1}(t) = f(y_{n-1}(t)), \qquad y_{n-1}(t_{n-1}) = y_{n-1}.$$

The usual error estimators compare the corrector polynomial $P(t)$ to the polynomial $P^+(t)$ which interpolates one extra point, i.e.

$$P^+(t_{n-j}) = f_{n-j}, \qquad j = 1, 2, \ldots, k,$$

$$P^+(t_n) = f(y_n),$$

$$P^{+'}(t_n) = f'(y_n).$$

The superscript $^+$ will be used to distinguish terms relating to the higher order polynomial. The error estimate E_0 is then of the form

$$E_0 = \int_{t_{n-1}}^{t_n} (P^+(t) - P(t))\,dt$$

$$= \int_{t_{n-1}}^{t_n} (t - t_n)^2 (t - t_{n-1}) \cdots (t - t_{n-k+1})\,[f_n; f_n; f_{n-1}; \ldots; f_{n-k}]\,dt.$$

Thus

$$E_0 = g_{2,k-1}[f_n; f_n; f_{n-1}; \ldots; f_{n-k}]. \tag{8}$$

In [4] it was shown that the local truncation error, T_n, may be written in the form

$$T_n = g_{2,k-1} \frac{y^{(k+2)}(\xi)}{(k+1)!}$$

for some $\xi \in [t_{n-1}; t_n]$. Thus it can be seen that the error estimate E_0 is asymptotically equivalent to the local truncation error.

The error estimate E_0 cannot, however, be calculated from (8), since only differences of the form $[f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-j}]$ are available when the estimate is required. Instead E_0 is calculated as a difference between predicted and corrected values using the following lemma.

LEMMA 1.

$$E_0 = g_{2,k-1}[f_n; f_n; f_{n-1}; \ldots; f_{n-k}]$$

$$= \frac{g_{2,k-1}}{q_k(t_n)} \left\{ (f'_n - f'_{n,0}) - \frac{q'_k(t_n)}{q_k(t_n)} (f_n - f_{n,0}) \right\}.$$

This expression is analogous to the well-known Milne error estimate used in the Adams methods.

Proof.

$$E_0 = \int_{t_{n-1}}^{t_n} (P^+(t) - P(t))\, dt$$

$$= \int_{t_{n-1}}^{t_n} (P^+(t) - P_0(t))\, dt - \int_{t_{n-1}}^{t_n} (P(t) - P_0(t))\, dt$$

$$= h_n(\beta^+_{n,0} - \beta_{n,0})(f_n - f_{n,0}) + h_n^2(\gamma^+_{n,0} - \gamma_{n,0})(f'_n - f'_{n,0})$$

from (3), (6) and (7).

Now by (5)

$$h_n^2\gamma^+_{n,0} - h_n^2\gamma_{n,0} = \frac{g_{1,k}}{q_k(t_n)} - \frac{g_{1,k-1}}{q_{k-1}(t_n)}$$

$$= \frac{1}{q_k(t_n)} [g_{1,k} - (t_n - t_{n-k})g_{1,k-1}]$$

$$= \frac{g_{2,k-1}}{q_k(t_n)} \quad \text{using (1).}$$

Also, from (4)

$$h_n\beta^+_{n,0} - h_n\beta_{n,0}$$

$$= \frac{1}{q_k(t_n)}\left[g_{0,k} - \frac{q'_k(t_n)}{q_k(t_n)}\, g_{1,k}\right] - \frac{1}{q_{k-1}(t_n)}\left[g_{0,k-1} - \frac{q'_{k-1}(t_n)}{q_{k-1}(t_n)}\, g_{1,k-1}\right]$$

$$= \frac{1}{q_k(t_n)}\,[g_{0,k} - (t_n - t_{n-k})g_{0,k-1}]$$

$$- \frac{1}{q_k(t_n)}\left[\frac{q'_k(t_n)}{q_k(t_n)}\, g_{1,k} - \frac{(t_n - t_{n-k})q'_{k-1}(t_n)}{q_{k-1}(t_n)}\, g_{1,k-1}\right]$$

$$= \frac{g_{1,k-1}}{q_k(t_n)} - \frac{1}{q_k(t_n)}\left[\frac{q'_{k-1}(t_n)}{q_{k-1}(t_n)}\, g_{1,k} + \frac{g_{1,k}}{(t_n - t_{n-k})}\right.$$

$$\left. - \frac{q'_{k-1}(t_n)}{q_{k-1}(t_n)}(t_n - t_{n-k})g_{1,k-1}\right]$$

$$= -\frac{1}{q_k(t_n)}\left[\frac{q'_{k-1}(t_n)}{q_{k-1}(t_n)}\, g_{2,k-1} + \frac{g_{2,k-1}}{t_n - t_{n-k}}\right] = -\frac{1}{q_k(t_n)}\left[\frac{q'_k(t_n)}{q_k(t_n)}\right] g_{2,k-1}.$$

The lemma follows.

Note that from Lemma 1

$$(9)\quad [f_n; f_n; f_{n-1}; \ldots; f_{n-k}] = \frac{1}{q_k(t_n)}\left[(f'_n - f'_{n,0}) - \frac{q'_k(t_n)}{q_k(t_n)}(f_n - f_{n,0})\right].$$

This equation may be used to update the divided differences $[f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-j}]$ for advancement to the next step.

If the equations to be solved contain fast transients, components of the vector f_{n-k} may differ greatly from the corresponding components of the vector f_n causing the error estimate E_0 to overestimate the true error. Consider the estimate

$$E_1 = g_{2,k-1}\,[f_n; f_n; f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-k+1}].$$

Because E_1 only contains those past values of f used in the corrector formula it would be expected that E_1 be a better estimate than E_0. For second derivative methods E_1 may also be expressed in terms of predicted and corrected values.

LEMMA 2. *Let* $p_j(t) = (t - t_{n-1})q_j(t), j \geqslant 1$. *Then*

$$E_1 = g_{2,k-1}\,[f_n; f_n; f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-k+1}]$$

$$= \frac{g_{2,k-1}}{p_{k-1}(t_n)}\left[(f'_n - f'_{n,0}) - \frac{p'_{k-1}(t_n)}{p_{k-1}(t_n)}\,(f_n - f_{n,0})\right].$$

Proof.

$$[f_n; f_n; f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-k+1}](t_n - t_{n-1})$$

$$= [f_n; f_n; f_{n-1}; f_{n-2}; \ldots; f_{n-k+1}; f_{n-1}](t_n - t_{n-1})$$

$$= [f_n; f_n; f_{n-1}; \ldots; f_{n-k+1}] - [f_n; f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-k+1}]$$

$$= (t_n - t_{n-k})[f_n; f_n; f_{n-1}; \ldots; f_{n-k}]$$

$$- (t_{n-1} - t_{n-k})[f_{n-1}; f_n; f_{n-1}; \ldots; f_{n-k}]$$

$$= (t_n - t_{n-k})[f_n; f_n; f_{n-1}; \ldots; f_{n-k}]$$

$$- (t_{n-1} - t_{n-k})[f_n; f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-k}].$$

The first divided difference in the last expression is given by (9). Also, from (2) it is easy to show that

$$[f_n; f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-k}]\, p_k(t_n) = f_n - f_{n,0}.$$

We have then

$$E_1 = \frac{g_{2,k-1}}{(t_n - t_{n-1})} \left\{ \frac{(t_n - t_{n-k})}{q_k(t_n)} \left[(f'_n - f'_{n,0}) - \frac{q'_k(t_n)}{q_k(t_n)} (f_n - f_{n,0}) \right] \right.$$

$$\left. - \frac{(t_{n-1} - t_{n-k})}{p_k(t_n)} (f_n - f_{n,0}) \right\}$$

$$= g_{2,k-1} \left\{ \frac{1}{p_{k-1}(t_n)} (f'_n - f'_{n,0}) \right.$$

$$\left. - \frac{(f_n - f_{n,0})}{p_k(t_n)} \left[(t_n - t_{n-k}) \frac{q'_k(t_n)}{q_k(t_n)} + \frac{(t_{n-1} - t_{n-k})}{(t_n - t_{n-1})} \right] \right\}.$$

The lemma follows since the term in square brackets is equal to

$$\frac{(t_n - t_{n-k}) q'_{k-1}(t_n)}{q_{k-1}(t_n)} + 1 + \frac{(t_{n-1} - t_{n-k})}{(t_n - t_{n-1})}$$

$$= (t_n - t_{n-k}) \left[\frac{q'_{k-1}(t_n)}{q_{k-1}(t_n)} + \frac{1}{t_n - t_{n-1}} \right] = (t_n - t_{n-k}) \frac{p'_{k-1}(t_n)}{p_{k-1}(t_n)}.$$

We will consider one final error estimate E_2. Let

$$W_n = I - h_n\beta_{n,0}\frac{\partial f}{\partial y} - h_n^2\gamma_{n,0}\left(\frac{\partial f}{\partial y}\right)^2$$

and consider

$$E_2 = W_n^{-1}E_1.$$

In view of Lemma 2, we have the following result.

LEMMA 3.

$$E_2 = g_{2,k-1}W_n^{-1}[f_n; f_n; f_{n-1}; f_{n-1}; f_{n-2}; \cdots; f_{n-k+1}]$$

$$= \frac{g_{2,k-1}}{p_{k-1}(t_n)}\,W_n^{-1}\left[(f'_n - f'_{n,0}) - \frac{p'_{k-1}(t_n)}{p_{k-1}(t_n)}(f_n - f_{n,0})\right].$$

Note that the factor W_n^{-1} does not affect the asymptotic behavior (as $h \longrightarrow 0$) of the error estimate so that E_2 is asymptotically equivalent to the local truncation error. The motivation for considering E_2 will be Theorems 1 and 2 of the following section where it will be shown that there is close agreement between E_2 and the true local error even for large values of the stepsize h.

E_2 is readily calculated. In order to solve (3) for y_n the LU decomposition of W_n must be determined. Consequently, the evaluation of E_2 simply requires one extra back substitution.

4. Effectiveness Theory. The concept of the effectiveness theorem, first introduced by Hull [2], [3] describes the ability of a particular method to solve certain problems. The theory takes into account the way the method estimates error and chooses stepsize as well as its basic formula. Results for Adams and Runge-Kutta methods have been proved by Hull and by Sedgwick [5]. For stiff linear problems Enright [1] proved an effectiveness theorem for a class of methods which used a one-step-two-half-step error estimate. In this section results for second derivative methods using error estimates based on a comparison between predicted and corrected values will be proved.

We begin by considering a class, C_0, of linear problems. Using the notation of Hull [3], this class may be represented by the 5-tuple $\langle A_0 y, t_0, y_0, t_f, a(\tau)\rangle$. An approximation to $y(t_f)$ is required where the exact solution $y(t)$ satisfies

$$y' = A_0 y, \qquad y(t_0) = y_0. \tag{10}$$

A_0 is a diagonalizable matrix whose eigenvalues lie on the negative real axis. The acceptability criterion, $a(\tau)$, will be defined as follows: For some $t_0 < t_1 < \cdots < t_M = t_f$ it is required that $\|y_n - y_{n-1}(t_n)\| \leqslant \kappa(H)\tau$, $1 \leqslant n \leqslant M$, where the condition number $\kappa(H) = \|H^{-1}\|\,\|H\|$ and $H^{-1}A_0H = D$ for diagonal D. The ∞-norm will be used throughout this section.

Results will be proved for three second derivative methods of the form

$$y_n = y_{n-1} + h \sum_{j=0}^{k-1} \beta_{n-j} y'_{n-j} + h^2 \gamma_0 y''_n. \tag{11}$$

Formulas of orders three, four and five will be considered in conjunction with the error estimates E_1 and E_2 considered in the previous section.

The essence of effectiveness theory is the relation between the true error T and the error estimate E. Consistent with the notation of Sedgwick [5], we may write

$$E = R(hA)y_{n-k+1} + U(hA), \qquad T = S(hA)y_{n-k+1} + V(hA)$$

for functions R, S, U and V. E and T represent the errors in stepping from t_{n-1} to t_n and U and V depend linearly on the local errors between t_{n-k+1} and t_{n-1}. Thus

$$T = S(hA)R^{-1}(hA)E + W(hA), \tag{12}$$

where $W = V - SR^{-1}U$.

Equation (12) expresses the relation between the true error and the error estimate. This relationship is characterized by SR^{-1} while W represents the errors made in previous steps and may be bounded by a multiple of τ. From (12),

$$\|T\| \leqslant \|S(hA)R^{-1}(hA)\| \, \|E\| + \|W(hA)\|$$

so that

$$\|T\| \leqslant \kappa(H)[\|S(hD)R^{-1}(hD)\| \, \|E\| + \|W(hD)\|]. \tag{13}$$

The elements of the diagonal matrix $S(hD)R^{-1}(hD)$ are of the form $S(h\lambda_i)R^{-1}(h\lambda_i)$ for an eigenvalue λ_i of A. For stiff problems from the class C_0 the relation between T and E is illustrated in Figures 1, 2 and 3 where $|S(z)R^{-1}(z)|$ is plotted against z. For the error estimate E_1, $|S(z)R^{-1}(z)|$ becomes very small even for moderately negative values of z showing that E_1 can severely overestimate components of the true error. However, there is no such problem associated with the error estimate E_2.

The situation on the negative half line reflects the situation throughout the negative half plane as shown by the following theorem about the error estimate E_2.

THEOREM 1. *Consider the linear problem*

$$y' = \lambda y, \tag{14}$$

where $\lambda = r_1 e^{i\theta}$, $\pi/2 < |\theta| \leqslant \pi$. *Let* $z = h\lambda = re^{i\theta}$ *for some* $h > 0$. *For the second derivative methods* (11) *and the error estimate* E_2 *the following results hold:*

(i)

$$S(z)R^{-1}(z) \sim 1/z \qquad \text{as } r \longrightarrow \infty \text{ (order 3)},$$

$$S(z)R^{-1}(z) \longrightarrow 1\frac{1}{14} \qquad \text{as } r \longrightarrow \infty \text{ (order 4)},$$

$$S(z)R^{-1}(z) \longrightarrow 1\frac{31}{102} \qquad \text{as } r \longrightarrow \infty \text{ (order 5)}.$$

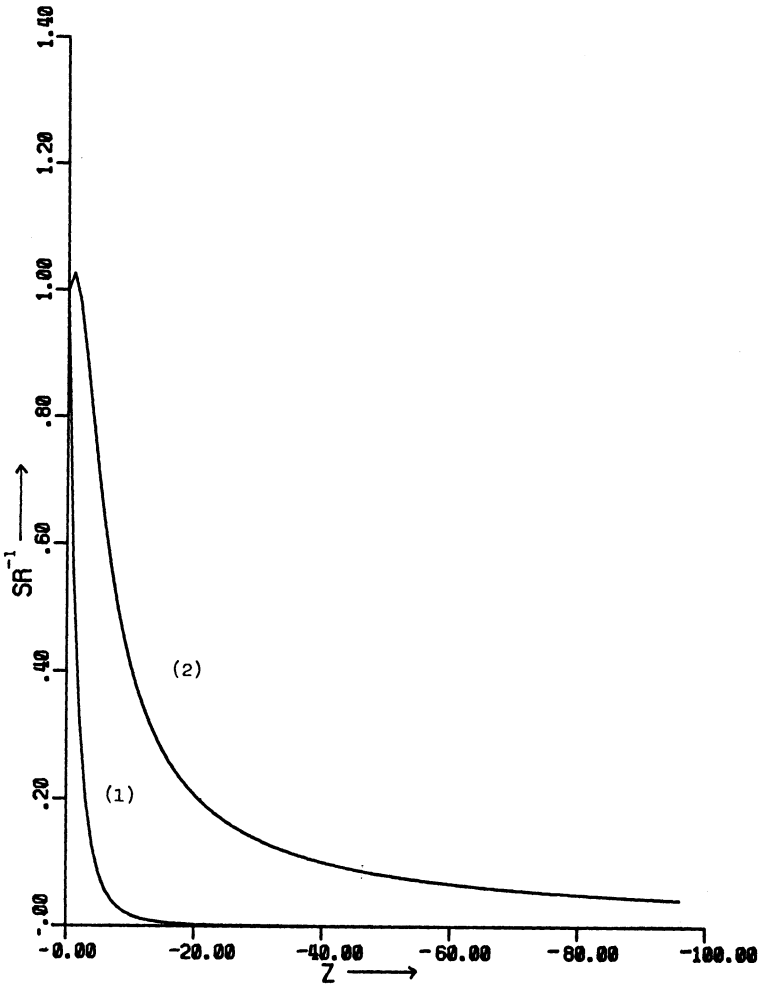

FIGURE 1

Second derivative methods of order three

(1) $|S(z)R^{-1}(z)|$ *for error estimate* E_1

(2) $|S(z)R^{-1}(z)|$ *for error estimate* E_2

(ii) *For each of the second derivative methods* E_2 *may be expressed in the form*

$$E_2 = r_1(z)y_{n-1} + r_2(z)y_{n-2} + \cdots + r_{k-1}(z)y_{n-k+1},$$

where the $r_i(z)$ *are rational functions of* z. *For the second derivative methods of orders* 3 *or more, each of the rational functions is bounded for all values of* $r > 0$.

Proof. (i) The true error, T, is given by

$$T = y_{n-1}(t_n) - y_n = e^z y_{n-1} - y_n,$$

where $z = h\lambda$.

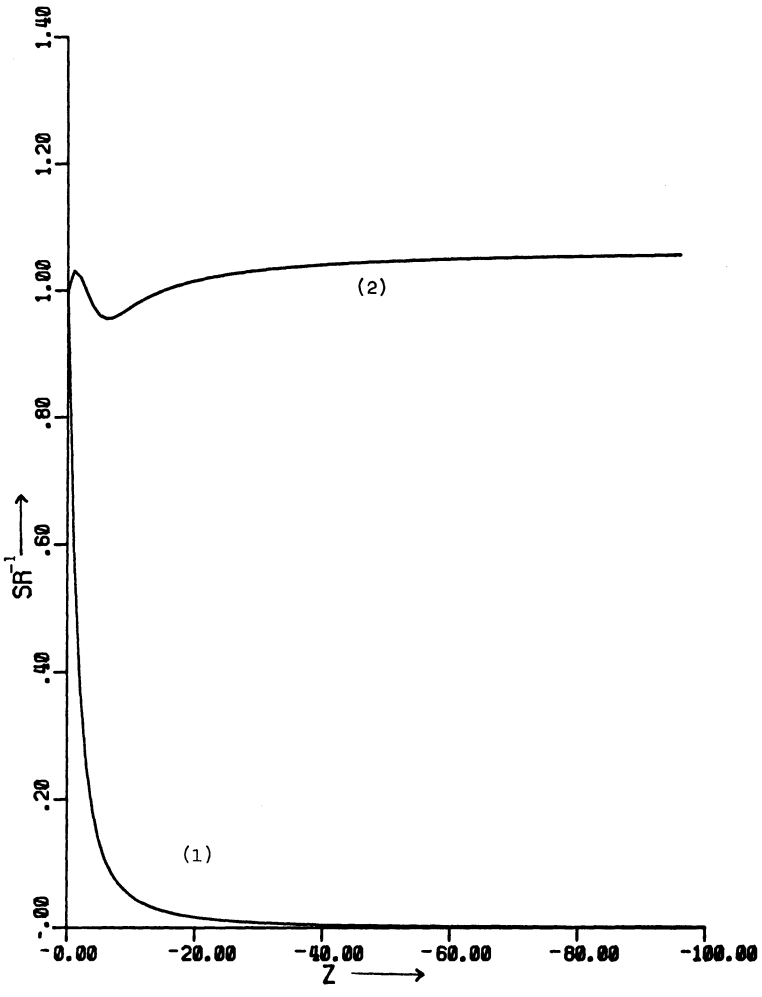

FIGURE 2

Second derivative methods of order four

(1) $|S(z)R^{-1}(z)|$ *for error estimate* E_1

(2) $|S(z)R^{-1}(z)|$ *for error estimate* E_2

From (11) and (14)

$$y_n = p_0^{-1}(z)\,[(1 + \beta_1 z)y_{n-1} + \beta_2 z y_{n-2} + \cdots + \beta_{k-1} z y_{n-k+1}],$$

where

$$p_0(z) = 1 - \beta_0 z - \gamma_0 z^2. \tag{15}$$

Hence

$$\begin{aligned} T = {} & [e^z - p_0^{-1}(z)(1 + \beta_1 z)]y_{n-1} \\ & - p_0^{-1}(z)\beta_2 z y_{n-2} - \cdots - p_0^{-1}(z)\beta_{k-1} z y_{n-k+1}. \end{aligned} \tag{16}$$

Now

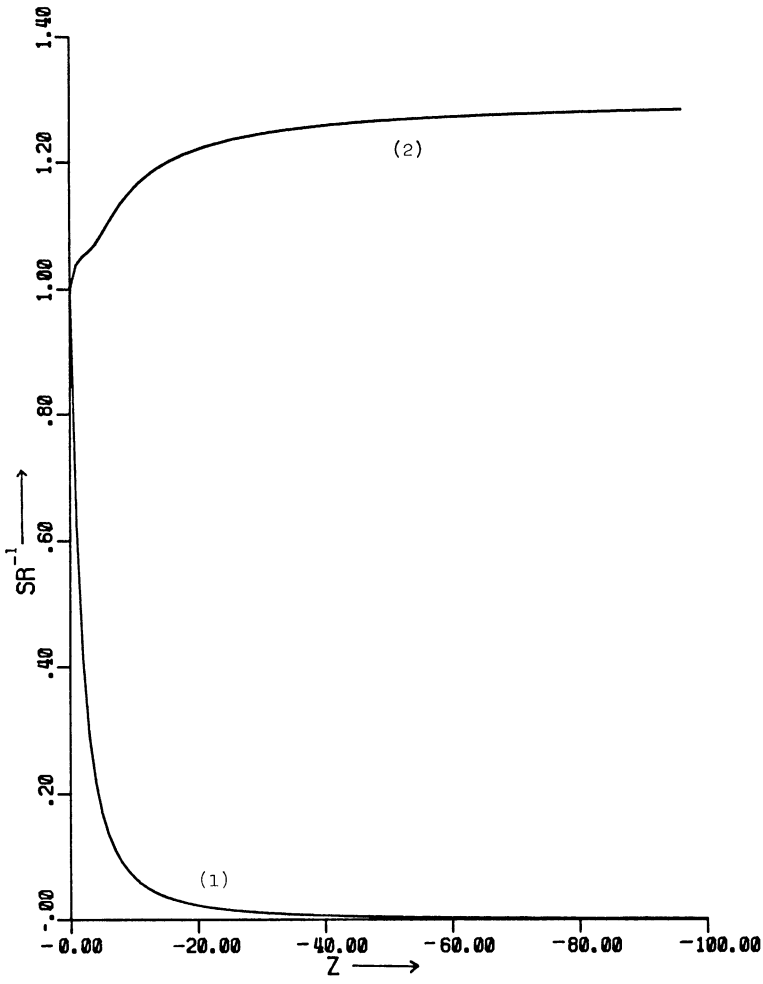

FIGURE 3

Second derivative methods of order five

(1) $|S(z)R^{-1}(z)|$ *for error estimate* E_1

(2) $|S(z)R^{-1}(z)|$ *for error estimate* E_2

$$E_2 = g_{2,k-1}p_0^{-1}(z)\,[f_n; f_n; f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-k+1}].$$

The term $g_{2,k-1}[f_n; f_n; f_{n-1}; f_{n-1}; f_{n-2}; \ldots; f_{n-k+1}]$ is a linear combination of the terms hf_n, h^2f_n', hf_{n-1}, h^2f_{n-1}', hf_{n-2}, $\ldots$, hf_{n-k+1}. So for constants α_1, $\alpha_2, \ldots, \alpha_{k+2}$ we may write

$$\begin{aligned} E_2 &= \alpha_1 z p_0^{-2}(z)\,[(1+\beta_1 z)y_{n-1} + \beta_2 z y_{n-2} + \cdots + \beta_{k-1} z y_{n-k+1}] \\ &\quad + \alpha_2 z^2 p_0^{-2}(z)\,[(1+\beta_1 z)y_{n-1} + \beta_2 z y_{n-2} + \cdots + \beta_{k-1} z y_{n-k+1}] \\ &\quad + \alpha_3 z p_0^{-1}(z) y_{n-1} + \alpha_4 z^2 p_0^{-1}(z) y_{n-1} \\ &\quad + \alpha_5 z p_0^{-1}(z) y_{n-2} + \cdots + \alpha_{k+2} z p_0^{-1}(z) y_{n-k+1}. \end{aligned} \tag{17}$$

$R(z)$ and $S(z)$ are determined from E_2 and T, respectively, by replacing the terms

$$y_{n-j} \quad \text{by} \quad y_{n-k+1}(t_{n-j}) = e^{(k-j-1)z}y_{n-k+1}.$$

As $r \to \infty$, $e^{(k-j-1)z} \to 0, j = 1, 2, \ldots, k-2$, hence the terms in y_{n-k+1} dominate. For the third order formula ($k = 2$)

$$SR^{-1} \to \frac{-(1+\beta_1 z)}{(\alpha_1 + \alpha_2 z)p_0^{-1}(z)(1+\beta_1 z)z + (\alpha_3 + \alpha_4 z)z}$$

$$\sim 1/z \quad \text{as } r \to \infty.$$

Similarly, for the higher order formulas ($k > 2$)

$$SR^{-1} \to \frac{\beta_{k-1}\gamma_0}{\alpha_2\beta_{k-1} - \alpha_{k+2}\gamma_0}.$$

Part (i) now follows by determining the constants $\alpha_1, \alpha_2, \ldots, \alpha_{k+2}$. Part (ii) follows from (17) and the fact that for second derivative methods $\beta_0 > 0$ and $\gamma_0 < 0$, so the polynomial $p_0(z)$ is of the form $a_0 + a_1 z + a_2 z^2$ where $a_0 > 0$, $a_1 < 0$, and $a_2 > 0$ and, therefore, has no zeros in the left half plane.

The undesirable behavior of the error estimate E_1 is reflected by the following theorem.

THEOREM 2. *Consider the linear problem* (14) *where* $\lambda = r_1 e^{i\theta}$, $\pi/2 < |\theta| \leqslant \pi$. *Let* $z = h\lambda = re^{i\theta}$ *for some* $h > 0$. *Then for the second derivative methods* (11) *and the error estimate* E_1 *the following results hold*:

(i)

$$S(z)R^{-1}(z) \sim 1/z^3 \quad \text{as } r \to \infty \text{ (order 3)},$$

$$S(z)R^{-1}(z) \sim 1/z^2 \quad \text{as } r \to \infty \text{ (order 4)},$$

$$S(z)R^{-1}(z) \sim 1/z^2 \quad \text{as } r \to \infty \text{ (order 5)}.$$

(ii) *For each of the second derivative methods* E_1 *may be expressed in the form*

$$E_1 = r_1(z)y_{n-1} + r_2(z)y_{n-2} + \cdots + r_{k-1}(z)y_{n-k+1},$$

where the $r_i(z)$ *are rational functions of* z *which become unbounded as* $r \to \infty$.

We conclude this section by proving a typical effectiveness theorem for a second derivative method used in conjunction with the error estimate E_2.

THEOREM 3. (i) *The fourth order second derivative method*

$$y_n = y_{n-1} + \frac{29}{48}hy'_n + \frac{5}{12}hy'_{n-1} - \frac{1}{48}hy'_{n-2} - \frac{1}{8}h^2y''_n$$

is effective for the class of problems C_0 *provided a step is accepted only if*

$$\|E_2\| \leqslant \tau/2 \tag{18}$$

and if the stepsizes are chosen as given below.

(ii) *The stepsize need only by restricted for a finite time* t_N.

Proof. The theorem is proved by induction. Suppose that at previous steps it holds that

$$\|y_{n-j} - y_{n-j-1}(t_{n-j})\| \leqslant \kappa(H)\tau, \quad j = 1, 2, \ldots. \tag{19}$$

For the fourth order second derivative method the relationship between the true error and the error estimate is characterized by (13) and Figure 2. For the error estimate E_2 it holds that

$$\|S(hD)R^{-1}(hD)\| \leqslant 3/2.$$

Thus from (13) and (18) it holds that $\|T\| < \kappa(H)\tau$ if

$$\|W(hD)\| \leqslant \tau/4. \tag{20}$$

In order to study $W(hA)$, recall that $W = V - SR^{-1}U$. $V(hA)$ and $U(hA)$ can be determined from (16) and (17) respectively by replacing the terms y_{n-j} by $y_{n-j} - y_{n-k+1}(t_{n-j})$ and z by hA. Thus it can be shown that for $k = 3$, $W(hA)$ is of the form

$$g(hA)(y_{n-1} - y_{n-2}(t_{n-1})),$$

where

$$\begin{aligned} g(z) = {} & [e^z - p_0^{-1}(z)(1 + \beta_1 z)] \\ & - S(z)R^{-1}(z)\,[\alpha_1 z p_0^{-2}(z)(1 + \beta_1 z) + \alpha_2 z^2 p_0^{-2}(z)(1 + \beta_1 z) \\ & \qquad + \alpha_3 z p_0^{-1}(z) + \alpha_4 z^2 p_0^{-1}(z)]. \end{aligned}$$

Thus (20) holds if

$$|g(h\lambda_j)|\,\|y_{n-1} - y_{n-2}(t_{n-1})\| \leqslant \tau/4 \tag{21}$$

for each component j. But the function $g(z) \longrightarrow 0$ as $z \longrightarrow 0$ and is bounded throughout the negative half plane. By the inductive hypothesis (19), inequality (21) is true provided that the stepsize is restricted so that

$$|g(h\lambda_j)| \leqslant \kappa^{-1}(H)/4$$

for each component j.

This completes the first part of the theorem. To see how the restriction (20) on the stepsize behaves as we proceed along the integration interval, observe that since $\{y_n\} \longrightarrow 0$ as $n \longrightarrow \infty$,

$$\|y_{n-1} - y_{n-2}(t_{n-1})\| \longrightarrow 0 \quad \text{as } n \longrightarrow \infty.$$

Hence, if $|g(z)| \leqslant M$, then there exists an integer $N_1(M, \tau)$ such that $\|W(hD)\| \leqslant \tau/4$ for all $n > N_1$.

TABLE 1
Numerical results

METHOD	PROBLEM	STEPS	MAXIMUM STEPSIZE	FUNCTION CALLS	JACOBIAN CALLS	LU DECOMP.	GLOBAL ERROR AT t = 100.
	i=2	32	$.14\times10^{2}$	66	66	44	$.14\times10^{-6}$
Error	i=3	175	$.13\times10^{1}$	353	353	266	$.21\times10^{-11}$
Estimate	i=4	1621	$.13\times10^{0}$	3244	3244	3237	$.11\times10^{-9}$
E_1	i=5	16001	$.13\times10^{-1}$	32005	32005	31998	$.89\times10^{-9}$
	i=2	13	$.60\times10^{2}$	27	27	17	$.59\times10^{-3}$
Error	i=3	15	$.48\times10^{2}$	31	31	21	$.82\times10^{-7}$
Estimate	i=4	13	$.98\times10^{2}$	26	26	18	$.13\times10^{-9}$
E_2	i=5	14	$.82\times10^{2}$	28	28	19	$.13\times10^{-12}$

At this stage the only restriction on the stepsize comes from (18). However,

$$E_2 = r_1(hA)y_{n-1} + r_2(hA)y_{n-2}$$

$$= H^{-1}r_1(hD)Hy_{n-1} + H^{-1}r_2(hD)Hy_{n-2};$$

and by Theorem 1(ii) each term $r_j(hD)$ is bounded by R, say. Hence, there exists an integer $N_2(R, \tau, \kappa(H))$ such that $\|E_2\| \leqslant \tau/2$ for all $n > N_2$. Let $N = \max(N_1, N_2)$. For $n > N$ there is no restriction on the stepsize and hence the theorem.

Note that the stepsize strategy will have to take both the function $g(z)$ and the size of the components y_n^j into consideration. By Theorem 2(ii) there is no analogous result to Theorem 3 for the error estimate E_1.

Similar theorems can be proved for the class of problems $\langle A_\alpha y, t_0, y_0, t_f, a(\tau)\rangle$ where the eigenvalues of A_α lie in the stability regions, S_α, of the second derivative methods. By Theorem 1(i) there is a close agreement between the true error and the error estimate in large areas of S_α, and by Theorem 1(ii) for formulae of order three or more the stepsize is only restricted at the beginning of the interval.

5. Numerical Results. In order to illustrate the problems associated with error estimates based on a comparison between predicted and corrected values, E_1 and E_2 were incorporated into the fourth order variable-step second derivative method. Consistent with Section 4, an error per step criterion

$$\|\text{Error Estimate}\| \leqslant \text{tolerance}/2$$

was used to determine whether the current solution was acceptable. A conservative choice of stepsize for the following step, 0.9 × [tolerance/(4. error estimate)]$^{1/5}$ × h_{n-1} was used.

The methods were used to test the following problems:

$$y = \begin{bmatrix} -10^{-i} & \\ & -10^{i} \end{bmatrix} y, \qquad y(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix};$$

range [0,100], initial stepsize 10^{-i}, tolerance 10^{-2}, $i = 2(1)5$.

The results were obtained on the CYBER 73 (University of Melbourne) which has a 60-bit word. As can be seen from Table 1, the stepsize is prohibitively restricted by E_1 and the value of $|\max(h\lambda_i)|$ is restricted to approximately $.13 \times 10^4$ for each of the four problems. However, no such problem exists for the error estimate E_2 and $|\max(h\lambda_i)|$ increases rapidly.

Computer Science Department
University of Melbourne
Parkville, Victoria 3052, Australia

1. W. H. ENRIGHT, *Studies in the Numerical Solution of Stiff Ordinary Differential Equations,* Tech. Report No. 46, Dept. of Computer Science, University of Toronto, 1972.

2. T. E. HULL, "The numerical integration of ordinary differential equations," *Information Processing,* 68, Vol. 1 (Proc. IFIP Congress, Edinburgh, 1968), North-Holland, Amsterdam, 1969, pp. 40–53. MR **41** #7850.

3. T. E. HULL, The effectiveness of numerical methods for ordinary differential equations," *Studies in Numerical Analysis.* 2 (J. N. Ortega & W. C. Rheinboldt, Editors), SIAM, Philadelphia, Pa., 1970, pp. 114–121. MR **42** #2671.

4. R. SACKS-DAVIS, "Solution to stiff ordinary differential equations by a second derivative method," *SIAM J. Numer. Anal.* (To appear.)

5. A. SEDGWICK, *An Efficient Variable Order Variable Step Adams Method,* Tech. Report No. 53, Dept. of Computer Science, University of Toronto, 1973.