HIROSHIMA MATH. J.
10 (1980), 469–494

On errors in the numerical solution of ordinary differential equations by step-by-step methods

Hisayoshi SHINTANI
(Received January 18, 1980)

1. Introduction

Consider the initial value problem

$$y' = f(x, y) \qquad (a \leqq x \leqq b), \tag{1.1}$$

$$y(a) = y_0, \tag{1.2}$$

where $f(x, y)$ is sufficiently smooth in $I \times R$, $I=[a, b]$ and $R=(-\infty, \infty)$. Denote by $y(x)$ the solution of this problem and for a positive constant h_0 let

$$x_j = a + jh \ (j = 0, 1,\dots, N), \quad h = (b-a)/N \leqq h_0. \tag{1.3}$$

We consider the case where the approximate values y_m of $y(x_m)$ $(m=k, k+1,\dots, N)$ are obtained by the k-step method [2]

$$\textstyle\sum_{j=0}^{k} \alpha_j y_{n+j} = h\Phi(x_n, y_n,\dots, y_{n+k}; h) \qquad (n = 0, 1,\dots, N-k), \tag{1.4}$$

where α_j $(j=0, 1,\dots, k)$ are real constants and $\alpha_k=1$. The method (1.4) includes one-step methods, linear multistep methods, hybrid methods, pseudo-Runge-Kutta methods and so on.

In Section 3 for sufficiently smooth $\Phi(x, u_0,\dots, u_k; v)$ we study the asymptotic behavior of errors

$$e_j = y_j - y(x_j) \qquad (j = 0, 1,\dots, N) \tag{1.5}$$

as $h \to 0$. In Section 4 the local truncation error is approximated and Milne's device in the predictor-corrector method is justified under certain conditions. In Section 5 we are concerned with the approximate computation of errors and illustrate the method by numerical examples.

2. Preliminaries

2.1. Assumptions

For simplicity the dependence of Φ on f is not expressed explicitly. Let

(2.1) $$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \quad H = [0, h_0]$$

and assume that the following conditions are satisfied.

CONDITION A: $\Phi(x, u_0, \ldots, u_k; v)$ is sufficiently smooth in $I \times R^{k+1} \times H$.

CONDITION B: If $f \equiv 0$, then $\Phi \equiv 0$.

CONDITION R: The modulus of no zero of $\rho(\zeta)$ exceeds 1 and the zeros of modulus 1 are all simple.

For any solution $z(x)$ of (1.1) let

(2.2) $$T(x, z(x); h) = \sum_{j=0}^{k} \alpha_j z(x+jh) - h\Phi(x, z(x), z(x+h), \ldots, z(x+kh); h)$$

and suppose that the method (1.4) is of order p ($p \geqq 1$) and that $y(x)$ exists over I. Then we have

(2.3) $$\rho(1) = 0, \quad \rho'(1) \neq 0,$$

(2.4) $$\Phi(x, y, \ldots, y; 0) = \rho'(1) f(x, y)$$

and the method (1.4) is convergent if $e_i \to 0$ ($i=0, 1, \ldots, k-1$) as $h \to 0$ [2, pp. 410–417].

2.2. Two lemmas

Suppose that $T(x, y; h)$ can be written as

(2.5) $$T(x, y; h) = h^{p+1} \sum_{j=0}^{2} h^j \varphi_j(x, y) + O(h^{p+4})$$

and let

$$\Phi_j = \frac{\partial \Phi}{\partial u_j}, \quad \Phi_{ij} = \frac{\partial^2 \Phi}{\partial u_i \partial u_j}, \quad \Phi_v = \frac{\partial \Phi}{\partial v}, \quad \Phi_{vi} = \frac{\partial^2 \Phi}{\partial v \partial u_i} \qquad (i, j = 0, 1, \ldots, k).$$

We write $\Phi(x, u, \ldots, u; v)$, $\Phi_j(x, u, \ldots, u; v)$, etc. as $\Phi(x, u; v)$, $\Phi_j(x, u; v)$, etc. respectively and denote by δ_{ij} Kronecker's delta. Let

(2.6) $$f^{(j+1)} = f_x^{(j)} + f f_y^{(j)} \ (j = 0, 1, \ldots), \quad f^{(0)} = f,$$

(2.7) $$\alpha = \rho'(1), \quad \omega = \left(\sum_{j=0}^{k} j^2 \alpha_j\right)/2, \quad N_1 = N - k.$$

LEMMA 1.

(2.8) $$\sum_j \Phi_j(x, y; 0) = \alpha f_y(x, y),$$

(2.9) $$\sum_{i,j} \Phi_{ji}(x, y; 0) = \alpha f_{yy}(x, y),$$

(2.10) $$\sum_j j\Phi_j(x, y; 0) f(x, y) + \Phi_v(x, y; 0) + \delta_{p1} \varphi_0(x, y) = \omega f^{(1)}(x, y),$$

(2.11) $$\sum_{i,j} j\Phi_{ji}(x, y; 0) f(x, y) + \sum_j j\Phi_j(x, y; 0) f_y(x, y) + \sum_i \Phi_{vi}(x, y; 0) + \delta_{p1} \varphi_{0y}(x, y) = \omega f_y^{(1)}(x, y),$$

where i and j range from 0 *to* k.

PROOF. For any solution $z(x)$ of (1.1) we have

$$\sum_j \alpha_j z(x+jh) = \rho(1)z(x) + \alpha hz'(x) + \omega h^2 z''(x) + O(h^3),$$

$$\Phi(x, z(x),\dots, z(x+kh); h) = \Phi(x, z(x); 0) + h\sum_j j\Phi_j(x, z(x); 0)z'(x)$$
$$+ h\Phi_v(x, z(x); 0) + O(h^2).$$

Using (2.3) and (2.4) and noting that

$$z'(x) = f(x, z(x)), \quad z''(x) = f^{(1)}(x, z(x)),$$

$$T(x, z(x); h) = \delta_{p1} h^2 \varphi_0(x, z(x)) + O(h^3),$$

we have from (2.2)

$$\sum_j j\Phi_j(x, z(x); 0)f(x, z(x)) + \Phi_v(x, z(x); 0) + \delta_{p1}\varphi_0(x, z(x))$$
$$= \omega f^{(1)}(x, z(x)).$$

Since $z(x)$ is an arbitrary solution, (2.10) is valid for any (x, y) in $I \times R$.

Calculating the partial derivatives of (2.4), (2.8) and (2.10) with respect to y, we find (2.8), (2.9) and (2.11) respectively, and the proof is complete.

Consider the difference equation

$$\sum_{j=0}^k \alpha_j z_{n+j} = h\sum_{j=0}^k \beta_{j,n} z_{n+j} + \lambda_n \qquad (n = 0, 1,\dots, N-k), \tag{2.12}$$

where $\alpha_k = 1$. Then we have the following lemma [1, pp. 243–244].

LEMMA 2. *Under Condition* R *let* B, β *and* Λ *be the constants such that*

$$\sum_{j=0}^k |\beta_{j,n}| \leqq B, \quad |\beta_{k,n}| \leqq \beta, \quad |\lambda_n| \leqq \Lambda \qquad (n = 0, 1,\dots, N-k) \tag{2.13}$$

and let $\beta h < 1$. *Then every solution of* (2.12) *for which*

$$|z_i| \leqq Z \qquad (i = 0, 1,\dots, k-1) \tag{2.14}$$

satisfies

$$|z_n| \leqq K^* e^{nhL^*} \qquad (n = 0, 1,\dots, N), \tag{2.15}$$

where

$$K^* = \Gamma^*(N\Lambda + kAZ), \quad L^* = \Gamma^* B, \quad A = \sum_{j=0}^k |\alpha_j|, \ \Gamma^* = \Gamma/(1-\beta h) \tag{2.16}$$

and Γ *is a positive constant depending on* α_j $(j=0, 1,\dots, k)$.

2.3. Notation

Let $B_M=[-M, M]$ $(M>0)$, choose M large so that

$$y(x)\in B_M \quad \text{for} \quad x\in I, \ y_j\in B_M \quad (j=0, 1,\dots, N) \qquad \text{for} \quad h\leqq h_0$$

and put $\Omega_M=I\times B_M^{k+1}\times H$. Let b_j $(j=0, 1,\dots, k)$ be the positive constants such that

$$|\Phi_j(x, u_0, u_1,\dots, u_k; v)|\leqq b_j \quad (j=0, 1,\dots, k) \quad \text{on} \quad \Omega_M$$

and put

$$B=\textstyle\sum_{j=0}^k b_j, \quad \beta=b_k, \quad h_1=\min(\beta^{-1}, h_0).$$

Let $x_u=a+uh$ $(0\leqq u\leqq N)$, denote by y_u the approximate value of $y(x_u)$ and put $f_u=f(x_u, y_u)$. We write $T(x, y(x); h)$, $\varphi_j(x, y(x))$, etc. simply as $T(x; h)$, $\varphi_j(x)$, etc. respectively. By (1.4) and (2.2) e_j $(j=0, 1,\dots, N)$ satisfy the equation

$$\begin{aligned}(2.17)\quad \textstyle\sum_{j=0}^k \alpha_j e_{n+j} &= h\Phi(x_n, y(x_n)+e_n,\dots, y(x_{n+k})+e_{n+k}; h)\\ &\quad - h\Phi(x_n, y(x_n),\dots, y(x_{n+k}); h) - T(x_n; h) \quad (n=0, 1,\dots, N_1).\end{aligned}$$

Let

$$(2.18)\quad g_j(x)=f_y^{(j)}(x, y(x)) \quad (j=0, 1,\dots), \ g(x)=g_0(x), \ k(x)=f_{yy}(x, y(x))/2,$$

$$(2.19)\quad \beta_{j,n}=\Phi_j(x_n, y(x_n),\dots, y(x_{n+k}); h) \quad (j=0, 1,\dots, k;\ n=0, 1,\dots, N_1),$$

$$(2.20)\quad \gamma_{j,n}=\Phi_j(x_n, y(x_n); 0), \quad \gamma_j=\Phi_j(x_0, y_0; 0),$$

$$(2.21)\quad \phi(\zeta)=\textstyle\sum_{j=0}^k \gamma_j\zeta^j, \quad \varphi(\zeta)=\rho(\zeta)-h\phi(\zeta),$$

$$(2.22)\quad c=1/\alpha, \quad a(x)=\textstyle\sum_{j=0}^k j\Phi_j(x, y(x); 0).$$

Let $e(x)$ and $v(x)$ be the solutions of the initial value problems

$$(2.23)\qquad e'=g(x)e-c\varphi_0(x), \quad e(a)=0,$$

$$(2.24)\qquad v'=g(x)v-ct(x)-\delta_{p1}b(x), \quad v(a)=0$$

respectively, where

$$(2.25)\qquad t(x)=\varphi_1(x)+c(a(x)-\omega g(x))\varphi_0(x)-\omega c\varphi_0'(x),$$

$$(2.26)\qquad b(x)=c\varphi_{0y}(x)e(x)-k(x)e(x)^2.$$

Let ζ_μ $(\mu=1, 2,\dots, l)$ be all the zeros of $\rho(\zeta)$ of modulus 1 and let

$$(2.27)\qquad \zeta_1=1, \ \zeta_\mu=e^{i\varphi_\mu} \qquad (\mu=1, 2,\dots, l).$$

Denote by $e_\mu(x)$ $(\mu=1, 2,\dots, l)$ the solutions of the initial value problems

(2.28) $$e'_\mu = k_\mu(x)e_\mu, \quad e_\mu(a) = 1 \qquad (\mu = 1, 2,\dots, l),$$

where

(2.29) $$k_\mu(x) = \sum_{j=0}^k \zeta_\mu^j \Phi_j(x, y(x); 0)/(\zeta_\mu \rho'(\zeta_\mu)) \qquad (\mu = 1, 2,\dots, l).$$

3. Asymptotic formulas for errors

We introduce the following

CONDITION J: There exists a positive number q such that

$$e_i = O(h^q) \qquad (i = 0, 1,\dots, k-1).$$

THEOREM 1. *Under Condition* J

(3.1) $$e_n = O(h^r) \qquad (n = 0, 1,\dots, N)$$

for sufficiently small h, where $r=\min(p, q)$.

PROOF. By (2.17) we have

$$\sum_{j=0}^k \alpha_j e_{n+j} = h\sum_{j=0}^k \Phi_j(x_n, y(x_n)+\theta e_n,\dots, y(x_{n+k})+\theta e_{n+k}; h)e_{n+j} - T(x_n; h)$$
$$(0<\theta<1).$$

Let K and K_1 be the constants such that

$$|T(x; h)| \leqq Kh^{p+1} \qquad \text{for} \quad x\in I,\ h<h_1,$$
$$|e_i| \leqq K_1 h^q \quad (i = 0, 1,\dots, k-1) \qquad \text{for} \quad h<h_1.$$

Then by Lemma 2 for $h<h_1$

$$|e_n| \leqq [h^p(b-a)K + h^q kAK_1]\Gamma^* e^{(b-a)L^*} \qquad (n = 0, 1,\dots, N).$$

THEOREM 2. *Under Condition* J

(3.2) $$e_n = h^p e(x_n) + O(h^s) \qquad (n = 0, 1,\dots, N)$$

for sufficiently small h, where $s=\min(p+1, q)$.

PROOF. Put $e_n=h^p e(x_n)+v_n$ $(n=0, 1,\dots, N)$. Then by (2.17), (3.1), (2.5), (2.8) and (2.23) we have

$$\sum_{j=0}^k \alpha_j v_{n+j} = h\sum_{j=0}^k \beta_{j,n} v_{n+j} + O(h^{p+2}) + O(h^{2r+1}) \quad (n = 0, 1,\dots, N_1),$$

where $r=\min(p, q)$. Since

$$e(x_i) = ih\int_0^1 e'(a+iht)dt \qquad (i = 0, 1,\dots, k-1)$$

and $e'(x)$ is bounded on $[a, a+kh_0]$, it follows that

$$v_i = e_i - h^p e(x_i) = O(h^s) \qquad (i = 0, 1,\dots, k-1).$$

Hence by Lemma 2 we have $v_n = O(h^s)$ $(n=0, 1,\dots, N)$, because $\min(s, p+1, 2r) = s$.

COROLLARY. *Under Condition* J

$$e_n = h^p e(x_n) + h^{p+1} v(x_n) + O(h^s) \qquad (n = 0, 1,\dots, N) \tag{3.3}$$

for sufficiently small h, where $s = \min(p+1, q)$.

Now we introduce the following conditions.

CONDITION I: There exist constants c_i $(i=0, 1,\dots, k-1)$ and a positive integer q such that

$$e_i = c_i h^q + O(h^{q+1}) \qquad (i = 0, 1,\dots, k-1).$$

CONDITION H: The common factor $d(\zeta)$ of maximum degree of $\rho(\zeta)$ and $\phi(\zeta)$ has no common factor with $\rho(\zeta)/d(\zeta)$.

For instance Condition H is satisfied in the following cases:

Case 1°. $\Phi_j(x, y; 0) = \beta_j f_y(x, y)$ $(j=0, 1,\dots, k)$ and $\rho(\zeta)$ has no common factor with $\sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j$.

Case 2°. Zeros of $\rho(\zeta)$ are all simple.

Let

$$r_n = h^{-s}[e_n - h^p e(x_n) - h^{p+1} v(x_n)] \qquad (n = 0, 1,\dots, N). \tag{3.4}$$

Then by Condition I there exist constants d_i $(i=0, 1,\dots, k-1)$ such that

$$r_i = d_i + O(h) \qquad (i = 0, 1,\dots, k-1). \tag{3.5}$$

Let

$$\rho(\zeta)/(\zeta-\zeta_\mu) = \textstyle\sum_{j=0}^{k-1} \alpha_{\mu j}\zeta^j \qquad (\mu = 1, 2,\dots, l),$$

$$A_\mu = \left(\textstyle\sum_{j=0}^{k-1} \alpha_{\mu j} d_j\right)/\rho'(\zeta_\mu) \qquad (\mu = 1, 2,\dots, l). \tag{3.6}$$

Then we have the following

THEOREM 3. *Under Conditions* I *and* H *there exists a nonnegative integer* J *such that*

(3.7) $$e_n = h^p e(x_n) + h^{p+1}v(x_n) + h^s \sum_{\mu=1}^{l} A_\mu e^{in\varphi_\mu} e_\mu(x_n) + O(h^{s+1})$$
$$(n = J, J+1,\dots, N)$$

for sufficiently small h, where $s=\min(p+1, q)$, $J=0$ if $k=l$, $J=2r-1$ if $k=l+r$ and $\zeta=0$ is a zero of $\rho(\zeta)$ of multiplicity r, and $J=O(|\log h|)$ otherwise.

PROOF. The proof of this theorem follows the line along which Henrici proved his theorem [1, pp. 249–255].

Let ζ_j $(j=1, 2,\dots, k)$ be all the zeros of $\rho(\zeta)$ and let

$$t = (1+\max_{l<j\leqq k}|\zeta_j|)/2, \quad \varphi(\zeta) = d(\zeta)\tilde{\varphi}(\zeta).$$

Then there exists a positive number h_1' $(h_1'<h_1)$ such that for $h\leqq h_1'$ the zeros of $\tilde{\varphi}(\zeta)$ are all distinct. Let $\tilde{\zeta}_j$ $(j=1, 2,\dots, r)$ be all the distinct zeros of $\varphi(\zeta)$ for $h\leqq h_1'$ and p_j be the multiplicity of $\tilde{\zeta}_j$. We may assume that $\tilde{\zeta}_\mu\to\zeta_\mu$ $(\mu=1, 2,\dots, l)$ as $h\to 0$. Let h_2 $(h_2\leqq h_1')$ be a positive number such that

$$|\tilde{\zeta}_j| \leqq t \quad (j = l+1, l+2,\dots, r) \qquad \text{for} \quad h \leqq h_2.$$

Let $q(\zeta)$ be the $(N+1)$-vector defined by $q(\zeta)=(1, \zeta,\dots, \zeta^N)^T$ and denote by

$$z^{(\mu)} = (z_0^{(\mu)}, z_1^{(\mu)},\dots, z_N^{(\mu)})^T \qquad (\mu = 1, 2,\dots, k)$$

the vectors

$$q(\tilde{\zeta}_j),\ q'(\tilde{\zeta}_j),\dots,\ q^{(p_j-1)}(\tilde{\zeta}_j) \qquad (j = 1, 2,\dots, r),$$

where $q^{(i)}(\zeta)$ denotes the vector $q(\zeta)$ differentiated i-times.

By Lemma 1, (2.17), (2.5), (2.23) and (2.24) r_n $(n=0, 1,\dots, N)$ satisfy the difference equation

(3.8) $$\sum_{j=0}^{k} \alpha_j r_{n+j} = h \sum_{j=0}^{k} \beta_{j,n} r_{n+j} + h^2\lambda_n \qquad (n = 0, 1,\dots, N_1),$$

where $|\lambda_n|\leqq \Lambda$ $(n=0, 1,\dots, N_1)$ for some constant Λ. Corresponding to (3.8) we consider the homogeneous difference equation

(3.9) $$\sum_{j=0}^{k} \alpha_j u_{n+j} = h \sum_{j=0}^{k} \beta_{j,n} u_{n+j} \qquad (n = 0, 1,\dots, N_1).$$

Let $e_n^{(\mu)}$ $(n=0, 1,\dots, N;\ \mu=1, 2,\dots, k)$ be the solutions of (3.9) satisfying the initial conditions

(3.10) $$e_i^{(\mu)} = z_i^{(\mu)} \qquad (i = 0, 1,\dots, k-1;\ \mu = 1, 2,\dots, k).$$

Since $e_i^{(\mu)}=O(1)$ $(i=0, 1,\dots, k-1)$, by Lemma 2 $e_n^{(\mu)}=O(1)$ $(n=0, 1,\dots, N)$.

Let u_n $(n=0, 1,\dots, N)$ be the solution of (3.9) with $u_i=r_i$ $(i=0, 1,\dots, k-1)$. Then we have

(3.11) $$u_n = \sum_{\mu=1}^{k} B_\mu e_n^{(\mu)} \qquad (n = 0, 1,\dots, N),$$

where B_μ ($\mu=1, 2,\dots, k$) satisfy

$$\sum_{\mu=1}^{k} z_i^{(\mu)} B_\mu = r_i \qquad (i = 0, 1,\dots, k-1). \tag{3.12}$$

Put $w_n = r_n - u_n$ ($n=0, 1,\dots, N$). Then they satisfy (3.8) and $w_i = 0$ ($i=0, 1,\dots, k-1$). By Lemma 2 we have $w_n = O(h)$ ($n=0, 1,\dots, N$), so that

$$r_n = u_n + O(h) \qquad (n = 0, 1,\dots, N). \tag{3.13}$$

From (3.4), (3.11) and (3.13) it follows that

$$e_n = h^p e(x_n) + h^{p+1} v(x_n) + h^s \sum_{\mu=1}^{k} B_\mu e_n^{(\mu)} + O(h^{s+1}) \quad (n = 0, 1,\dots, N). \tag{3.14}$$

Now we study the behavior of $B_\mu e_n^{(\mu)}$.

Case 1. $\mu \leqq l$.

Let

$$\varphi_\mu(\zeta) = \varphi(\zeta)/(\zeta - \tilde{\zeta}_\mu) = \sum_{j=0}^{k-1} \tilde{\alpha}_{\mu j} \zeta^j \qquad (\mu = 1, 2,\dots, l).$$

Then from (3.12) it follows that

$$B_\mu = \left(\sum_{j=0}^{k-1} \tilde{\alpha}_{\mu j} r_j\right)/\varphi_\mu(\tilde{\zeta}_\mu). \tag{3.15}$$

Since

$$\tilde{\zeta}_\mu = \zeta_\mu + h\phi(\zeta_\mu)/\rho'(\zeta_\mu) + O(h^2),$$

by (3.5) and (3.6)

$$B_\mu = A_\mu + O(h) \qquad (\mu = 1, 2,\dots, l). \tag{3.16}$$

Put $f_n^{(\mu)} = \zeta_\mu^{-n} e_n^{(\mu)}$ ($n=0, 1,\dots, N$). Then they satisfy

$$\sum_{j=0}^{k} \alpha_j^{(\mu)} f_{n+j}^{(\mu)} = h \sum_{j=0}^{k} \beta_{j,n}^{(\mu)} f_{n+j}^{(\mu)} \qquad (n = 0, 1,\dots, N_1),$$

$$f_i^{(\mu)} = (\zeta_\mu^{-1} \tilde{\zeta}_\mu)^i = 1 + O(h) \qquad (i = 0, 1,\dots, k-1),$$

where

$$\alpha_j^{(\mu)} = \alpha_j \zeta_\mu^j, \quad \beta_{j,n}^{(\mu)} = \beta_{j,n} \zeta_\mu^j. \tag{3.17}$$

By (2.28) and (2.29) we have

$$\sum_{j=0}^{k} \alpha_j^{(\mu)} e_\mu(x_{n+j}) = h \sum_{j=0}^{k} \gamma_{j,n} \zeta_\mu^j e_\mu(x_{n+j}) + O(h^2).$$

Let $w_n^{(\mu)} = f_n^{(\mu)} - e_\mu(x_n)$ ($n=0, 1,\dots, N$). Then they satisfy

$$\sum_{j=0}^{k} \alpha_j^{(\mu)} w_{n+j}^{(\mu)} = h \sum_{j=0}^{k} \beta_{j,n}^{(\mu)} w_{n+j}^{(\mu)} + O(h^2) \qquad (n = 0, 1,\dots, N_1),$$

$$w_i^{(\mu)} = O(h) \qquad (i = 0, 1,\dots, k-1),$$

because $\beta_{j,n}-\gamma_{j,n}=O(h)$ $(j=0, 1,\dots, k)$. By Lemma 2 $w_n^{(\mu)}=O(h)$ $(n=0, 1,\dots, N)$, so that

$$e_n^{(\mu)} = \zeta_\mu^n[e_\mu(x_n)+O(h)] \qquad (n = 0, 1,\dots, N).$$

Combining this with (3.16) we have

$$\text{(3.18)} \qquad B_\mu e_n^{(\mu)} = A_\mu e^{in\varphi_\mu} e_\mu(x_n) + O(h) \qquad (\mu = 1, 2,\dots, l;\ n = 0, 1,\dots, N).$$

Case 2. $\mu > l$.

(a) Case where $\tilde{\zeta}_\mu$ is not a zero of $d(\zeta)$.

Since $\tilde{\zeta}_\mu$ is a zero of $\tilde{\varphi}(\zeta)$, it is simple. Let $z^{(\mu)}=q(\tilde{\zeta}_\mu)$. Then we show that for any ε $(0<\varepsilon<1)$ and for sufficiently small h there exists a nonnegative integer J such that

$$\text{(3.19)} \qquad e_n^{(\mu)} = O(h^{2-\varepsilon}) \qquad (n = J, J+1,\dots, N).$$

Let

$$e_n^{(\mu)} = z_n^{(\mu)} + w_n^{(\mu)} \qquad (n = 0, 1,\dots, N).$$

Then $w_n^{(\mu)}$ $(n=0, 1,\dots, N)$ satisfy

$$\text{(3.20)} \qquad \textstyle\sum_{j=0}^k \alpha_j w_{n+j}^{(\mu)} = h \sum_{j=0}^k \beta_{j,n} w_{n+j}^{(\mu)} + h\sigma_n \qquad (n = 0, 1,\dots, N_1),$$

$$\text{(3.21)} \qquad w_i^{(\mu)} = 0 \qquad (i = 0, 1,\dots, k-1),$$

where

$$\text{(3.22)} \qquad \sigma_n = \textstyle\sum_{j=0}^k (\beta_{j,n}-\gamma_j) z_{n+j}^{(\mu)} \qquad (n = 0, 1,\dots, N_1).$$

Since there exists a constant K_1 such that

$$|\beta_{j,n}-\gamma_j| \leqq (n+k)K_1 h \qquad (j = 0, 1,\dots, k;\ n = 0, 1,\dots, N_1) \quad \text{for} \quad h \leqq h_2,$$

we have

$$|\sigma_n| \leqq (k+1)K_1(n+k)ht^n \qquad (n = 0, 1,\dots, N_1).$$

Let J be the integer such that $J \leqq 2|\log h/\log t| < J+1$ and let h_3 $(0<h_3\leqq h_2)$ be a number less than 1 such that $J+k<N$ for $h\leqq h_3$. Then for some constants K_2 and K_3

$$|\sigma_n| \leqq K_2(J+k)h \quad (n = 0, 1,\dots, J) \qquad \text{for} \quad h \leqq h_3,$$

$$J+k \leqq K_3|\log h| \qquad \text{for} \quad h \leqq h_3.$$

Applying Lemma 2 to (3.20) for $n\leqq J$, we have for some constant K_4

$$|w_n^{(\mu)}| \leqq e^{nhL^*}\Gamma^* K_2(J+k)^2h^2 \leqq K_4(h\log h)^2 \qquad (n = 0, 1,\dots, J+k)$$

for $h \leqq h_3$.

Since $t^J \geqq h^2 > t^{J+1}$, there exists a constant K_5 such that

$$|z_n^{(\mu)}| = |\tilde{\zeta}_\mu^n| \leqq t^n \leqq K_5 h^2 \qquad \text{for} \quad n \geqq J,\ h \leqq h_3.$$

Hence for some constant C

$$\begin{aligned} (3.23) \qquad |e_n^{(\mu)}| &= |z_n^{(\mu)} + w_n^{(\mu)}| \leqq K_5 h^2 + K_4 (h \log h)^2 \\ &\leqq C h^{2-\varepsilon} \quad (n = J, J+1, \dots, J+k) \qquad \text{for} \quad h \leqq h_3. \end{aligned}$$

Application of Lemma 2 to (3.9) for $n \geqq J$ with the estimate (3.23) yields (3.19).

Let $\tilde{\zeta}_\mu \to \eta$ as $h \to 0$ and let η be a zero of $\rho(\zeta)$ of multiplicity r. Then by Condition H η is not a zero of $d(\zeta)$,

$$\tilde{\zeta}_\mu = \eta + \kappa h^{1/r} + O(h^{2/r}),$$

and B_μ is given by (3.15), where κ is one of the r-th roots of $r!\phi(\eta)/\rho^{(r)}(\eta)$. Since

$$\varphi_\mu(\tilde{\zeta}_\mu) = r\phi(\eta) h^{1-1/r}/\kappa + O(h),$$

it follows that $B_\mu = O(h^{-1+1/r})$. The choice $\varepsilon < 1/r$ yields

$$(3.24) \qquad B_\mu e_n^{(\mu)} = O(h) \qquad (n = J, J+1, \dots, N).$$

In the case $\eta = 0$, let $e_n^{(\mu)} = \tilde{\zeta}_\mu^n v_n^{(\mu)}$ $(n = 0, 1, \dots, N)$. Then

$$\begin{aligned} &\textstyle\sum_{j=0}^k \alpha_j^{(\mu)} v_{n+j}^{(\mu)} = h \sum_{j=0}^k \beta_{j,n}^{(\mu)} v_{n+j}^{(\mu)} \qquad (n = 0, 1, \dots, N_1), \\ &v_i^{(\mu)} = 1 \qquad (i = 0, 1, \dots, k-1), \end{aligned}$$

where

$$\alpha_j^{(\mu)} = \alpha_j \tilde{\zeta}_\mu^j, \quad \beta_{j,n}^{(\mu)} = \beta_{j,n} \tilde{\zeta}_\mu^j \qquad (j = 0, 1, \dots, k).$$

By Lemma 2 we have $v_n^{(\mu)} = O(1)$ $(n = 0, 1, \dots, N)$, so that

$$B_\mu e_n^{(\mu)} = O(h) \qquad (n = 2r-1, 2r, \dots, N).$$

(b) Case where $\tilde{\zeta}_\mu$ is a zero of $d(\zeta)$ of multiplicity r.

Since $\tilde{\zeta}_\mu$ is independent of h, we put $\tilde{\zeta}_\mu = \eta$. Let

$$\begin{aligned} &\phi_i(\zeta) = \varphi(\zeta)/(\zeta - \eta)^i = \textstyle\sum_{j=0}^{k-i} \gamma_j^{(i)} \zeta^j \qquad (i = 1, 2, \dots, r), \\ &z^{(\nu+j)} = q^{(j)}(\eta), \quad C_j = B_{\nu+j} \qquad (j = 0, 1, \dots, r-1). \end{aligned}$$

Then we have

$$C_{r-i} = [\textstyle\sum_{j=0}^{k-i} \gamma_j^{(i)} r_j - \sum_{j=1}^{i-1} \phi_i^{(r-j)}(\eta) C_{r-j}]/\phi_i^{(r-i)}(\eta) \qquad (i = 1, 2, \dots, r).$$

As $|\eta| < t$, there exists a constant K such that

$$\left| j!\binom{n}{j}\eta^{n-j} \right| \leqq Kt^n \qquad (j = 0, 1,\ldots, r-1;\; n = j, j+1,\ldots, N),$$

so that

$$|z_n^{(\mu)}| \leqq Kt^n \qquad (n = 0, 1,\ldots, N;\; \mu = \nu, \nu+1,\ldots, \nu+r-1).$$

By the same argument as in the case (a) we have (3.19).

Since η is not a zero of $\rho(\zeta)/d(\zeta)$ by Condition H,

$$\phi_i^{(r-i)}(\eta) = (r-i)!\rho^{(r)}(\eta)/r! + O(h) \qquad (i = 1, 2,\ldots, r),$$

so that $C_j = O(1)$ $(j=0, 1,\ldots, r-1)$ and

$$\text{(3.25)} \quad B_\mu e_n^{(\mu)} = O(h^{2-\varepsilon}) \qquad (\mu = \nu, \nu+1,\ldots, \nu+r-1;\; n = J, J+1,\ldots, N).$$

In the case $\eta=0$, since $z_n^{(\nu+j)} = j!\delta_{jn}$ $(n=0, 1,\ldots, N;\, j=0, 1,\ldots, r-1)$, we have $\sigma_n = O(h)$ $(n=0, 1,\ldots, N_1)$. By Lemma 2 $w_n^{(\mu)} = O(h)$ $(n=0, 1,\ldots, N)$, so that

$$B_\mu e_n^{(\mu)} = O(h) \qquad (n = r, r+1,\ldots, N).$$

This completes the proof.

In the case $k=1$ let $w(x)$ be the solution of the initial value problem

$$w' = g(x)w - \varphi_2(x) - l(x), \quad w(a) = 0,$$

where

$$\text{(3.27)} \quad l(x) = (v'' - g_1 v)/2 + (e''' - g_2 e)/6 + \Phi_1(\Phi_1\varphi_0 + \varphi_1) + (\Phi_{11}f + \Phi_{v1})\varphi_0 + \delta_{p1}m + \delta_{p2}b,$$

$$\text{(3.28)} \quad m(x) = \Phi_1 b + \varphi_{1y}e + (\varphi_{0y} - f_{yy}e)v - f_{yy}^{(1)}e^2/4 - f_{yyy}e^3/6 + (\Phi_1 f_{yy} + \varphi_{0yy})e^2/2 + (\Phi_{10} + \Phi_{11})e\varphi_0,$$

and Φ_1, f, etc. denote $\Phi_1(x, y(x); 0)$, $f(x, y(x))$, etc. respectively. Then we have the following

COROLLARY. *For one-step methods*

$$\text{(3.29)} \quad e_n = h^p e(x_n) + h^{p+1}v(x_n) + h^{p+2}w(x_n) + O(h^{p+3}) \qquad (n = 0, 1,\ldots, N)$$

for sufficiently small h.

For the two-step method of Adams type

$$\text{(3.30)} \qquad y_{n+2} = y_{n+1} + h\Phi(x_n, y_n, y_{n+1}, y_{n+2}; h),$$

(3.7) is valid with $l=1$ and $J=1$.

4. Approximation of local truncation errors

In this section besides Conditions I and H we impose the following

CONDITION L: $\rho(\zeta)$ has only one zero of modulus 1 and $q \geqq p+1$. Hence e_n can be expressed as

(4.1) $e_n = h^p e(x_n) + h^{p+1} v(x_n) + A_1 h^{p+1} e_1(x_n) + O(h^{p+2}) \quad (n = J, J+1, \dots, N)$.

4.1. General results

Let $E(x, u_0, u_1, \dots, u_m; v)$ be a sufficiently smooth function in $I \times R^{m+1} \times H$ and suppose that for any solution $z(x)$ of (1.1)

(4.2) $E(x, z(x), z(x+h), \dots, z(x+mh); h) = h^{p+1+\sigma}[\phi_0(x, z(x)) + O(h)]$

$$(x+jh \in I;\ j = 0, 1, \dots, m;\ m \geqq k),$$

where $\sigma = 0$ if

(4.3) $$\phi_0(x, y) = \gamma\varphi_0(x, y), \quad \gamma \neq 0, \quad 1 + \gamma \neq 0,$$

and $\sigma \geqq 1$ otherwise. Let

$$E_j = \frac{\partial E}{\partial u_j}, \quad E_v = \frac{\partial E}{\partial v}, \quad E_{ij} = \frac{\partial^2 E}{\partial u_i \partial u_j}, \quad E_{vi} = \frac{\partial^2 E}{\partial v \partial u_i} \quad (i, j = 0, 1, \dots, m).$$

We write $E(x, u, \dots, u; v)$, $E_j(x, u, \dots, u; v)$, etc. as $E(x, u; v)$, $E_j(x, u; v)$, etc. respectively. We assume that

(4.4) $$\textstyle\sum_{j=0}^{m} jE_j(x, y; 0) = -\alpha \quad \text{for} \quad (x, y) \in I \times R.$$

LEMMA 3.

(4.5) $E(x, y; 0) = 0$,

(4.6) $\sum_j jE_j(x, y; 0) f(x, y) + E_v(x, y; 0) = 0$,

(4.7) $\sum_j E_j(x, y; 0) = 0$,

(4.8) $\sum_{i,j} jE_{ji}(x, y; 0) f(x, y) + \sum_j jE_j(x, y; 0) f_y(x, y) + \sum_i E_{vi}(x, y; 0) = 0$,

where i and j range from 0 to m.

PROOF. Expanding (4.2) into power series in h and equating to zero the coefficients of h^j $(j=0, 1)$, we have (4.5) and (4.6). Calculation of the partial derivatives of (4.5) and (4.6) with respect to y yields (4.7) and (4.8). This completes the proof.

For simplicity let

(4.9) $$E_n = E(x_n, y_n, y_{n+1},\dots, y_{n+m}; h) \qquad (n = 0, 1,\dots, N-m).$$

LEMMA 4. *Under Conditions* I, H *and* L

(4.10) $$E_n = h^{p+1}[\varphi_0(x_n) + h\varphi_1(x_n) + h^\sigma\phi_0(x_n) + O(h)] \qquad (n = J, J+1,\dots, N-m)$$

for sufficiently small h.

PROOF. Substituting $y_j = y(x_j)+e_j$ $(j=n, n+1,\dots, n+m)$ and (4.1) into E_n and expanding it at $x=x_n$ into power series in h, we have (4.10) by Lemma 3, (4.4), (2.23) and (2.24).

By this lemma and (4.3) we obtain the following

THEOREM 4. *Suppose that Conditions* I, H *and* L *are satisfied. Then*

(4.11) $$E_n = h^{p+1}\varphi_0(x_n) + O(h^{p+2}) \qquad (n = J, J+1,\dots, N-m)$$

for $\sigma \geqq 1$, *and*

(4.12) $$aE_n = h^{p+1}\varphi_0(x_n) + O(h^{p+2}) \qquad (n = J, J+1,\dots, N-m)$$

for $\sigma=0$ *and* $a=1/(1+\gamma)$.

4.2. Construction of the formulas

4.2.1. Formulas without interpolation

Let a_j and b_j $(j=0, 1,\dots, m)$ be the constants such that

(4.13) $$\textstyle\sum_{j=0}^m a_j = 0, \quad \sum_{j=0}^m j a_j = -\alpha,$$

(4.14) $$\textstyle\sum_{j=0}^m j^i a_j = i\sum_{j=0}^m j^{i-1} b_j \qquad (i = 1, 2,\dots, p+\sigma)$$

and let

(4.15) $$E_n = \textstyle\sum_{j=0}^m a_j y_{n+j} - h\sum_{j=0}^m b_j f_{n+j}.$$

Then Theorem 4 is valid, and for $\sigma \geqq 1$

(4.16) $$E_n = T(x_n; h) - c(\omega+\textstyle\sum_{j=0}^m j^2 a_j/2)h^{p+2}\varphi_0'(x_n) + O(h^{p+2}) \qquad (n = J, J+1,\dots, N-m).$$

For the two-step method (3.30), (4.16) is valid for $n \geqq 0$ if $a_0=0$ and $\sigma \geqq 1$. For explicit one-step methods with $p \geqq 2$ and for $\sigma \geqq 2$

(4.17) $$E_n = T(x_n;\, h) - (1+\textstyle\sum_{j=0}^{m} j^2 a_j)h^{p+2}\varphi_0'(x_n)/2$$
$$- h^{p+2}g(x_n)\varphi_0(x_n)/2 + O(h^{p+3}) \qquad (n = 0,\, 1,\ldots,\, N-m).$$

Hence if

(4.18) $$\textstyle\sum_{j=0}^{m} j^2 a_j = -2r - 1 \qquad (r = 0,\, 1,\ldots,\, m-1),$$

then

(4.19) $$E_n = T(x_{n+r};\, h) - h^{p+2}g(x_n)\varphi_0(x_n)/2 + O(h^{p+3}) \qquad (n = 0,\, 1,\ldots,\, N-m);$$

and if

(4.20) $$\textstyle\sum_{j=0}^{m} j^2 a_j = -m,$$

then

(4.21) $$mE_n = \textstyle\sum_{j=0}^{m-1} T(x_{n+j};\, h) - mh^{p+2}g(x_n)\varphi_0(x_n)/2 + O(h^{p+3})$$
$$(n = 0,\, 1,\ldots,\, N-m).$$

EXAMPLE 1. If we impose the condition (4.20) and choose $m=4$, $\alpha=1$ and $p+\sigma=7$, we have

(4.22) $$E_n = [5(y_n - y_{n+4}) + 32(y_{n+1} - y_{n+3})]/84$$
$$+ h(f_n + 16f_{n+1} + 36f_{n+2} + 16f_{n+3} + f_{n+4})/70.$$

There exist also formulas that use the values of f computed already other than f_{n+j} $(j=0,\, 1,\ldots,\, m)$ [4].

4.2.2. Formulas with interpolation

Suppose that there exist constants λ_ν $(m>\lambda_\nu>0)$ that are not integers, and constants $c_{\nu j}$ and $d_{\nu j}$ $(\nu=1,\, 2,\ldots,\, t;\, j=0,\, 1,\ldots,\, m)$ such that

(4.23) $$\textstyle\sum_{j=0}^{m} c_{\nu j} = 1,$$

(4.24) $$\textstyle\sum_{j=0}^{m} j^i c_{\nu j} + i\sum_{j=0}^{m} j^{i-1} d_{\nu j} = \lambda_\nu^i \qquad (i = 1,\, 2,\ldots,\, p+\delta),$$

where δ is a nonnegative integer. Let

(4.25) $$y_{n+\lambda_\nu} = \textstyle\sum_{j=0}^{m} c_{\nu j} y_{n+j} + h\sum_{j=0}^{m} d_{\nu j} f_{n+j} \qquad (\nu = 1,\, 2,\ldots,\, t).$$

Then we have

LEMMA 5. *If $q \geqq p+1$ and $\delta \geqq 0$, then*

(4.26) $$e_{n+\lambda_\nu} = h^p e(x_{n+\lambda_\nu}) + O(h^{p+1}) \qquad (n = J,\, J+1,\ldots,\, N-m;\ \nu = 1,\, 2,\ldots,\, t).$$

Under Conditions I, H *and* L *if*

$$\delta \geqq 1, \quad \textstyle\sum_{j=0}^{m} d_{\nu j} = 0, \tag{4.27}$$

then

$$e_{n+\lambda_\nu} = h^p e(x_{n+\lambda_\nu}) + h^{p+1}v(x_{n+\lambda_\nu}) + A_1 h^{p+1} e_1(x_{n+\lambda_\nu}) + O(h^{p+2}) \tag{4.28}$$

$$(n = J, J+1, \dots, N-m).$$

Proof. Substituting (4.1) into

$$e_{n+\lambda_\nu} = \textstyle\sum_{j=0}^{m} c_{\nu j} e_{n+j} + h \sum_{j=0}^{m} d_{\nu j} g(x_{n+j}) e_{n+j} + O(h^{p+1+\delta}) + O(h^{2p+1})$$

and expanding it at $x = x_n$ into power series in h, we have by (4.23), (4.24) and (2.23)

$$e_{n+\lambda_\nu} = h^p e(x_{n+\lambda_\nu}) + h^{p+1} v(x_n) + A_1 h^{p+1} e_1(x_n)$$
$$+ c(\textstyle\sum_{j=0}^{m} d_{\nu j}) h^{p+1} \varphi_0(x_n) + O(h^{p+2}) + O(h^{p+1+\delta}),$$

which completes the proof.

Let a_j, b_j $(j=0, 1, \dots, m)$ and $b_{m+\nu}$ $(\nu = 1, 2, \dots, t)$ be the constants such that

$$\textstyle\sum_{j=0}^{m} a_j = 0, \quad \sum_{j=0}^{m} j a_j = -\alpha, \tag{4.29}$$

$$\textstyle\sum_{j=0}^{m} j^i a_j = i(\sum_{j=0}^{m} j^{i-1} b_j + \sum_{\nu=1}^{t} \lambda_\nu^{i-1} b_{m+\nu}) \qquad (i = 1, 2, \dots, p+\sigma) \tag{4.30}$$

and let

$$E_n = \textstyle\sum_{j=0}^{m} a_j y_{n+j} - h \sum_{j=0}^{m} b_j f_{n+j} - h \sum_{\nu=1}^{t} b_{m+\nu} f_{n+\lambda_\nu}. \tag{4.31}$$

Then Theorem 4 is valid and (4.16) holds if $\sigma \geqq 1$ and (4.27) is satisfied.

For explicit one-step methods with $p \geqq 2$, (4.17) holds if $\sigma \geqq 2$ and (4.27) is satisfied. For the two-step method (3.30), (4.16) is valid for $n \geqq 0$ if $a_0 = 0$, $\sigma \geqq 1$ and (4.27) is satisfied.

We introduce the following notations:

$$c_{\nu j} = C_{\nu j}/C_\nu, \; d_{\nu j} = D_{\nu j}/D_\nu \qquad (\nu = 1, 2, \dots, t; \, j = 0, 1, \dots, m).$$

Example 2. The choice $m=2$, $t=1$, $p+\delta=5$, $\lambda_1 = 1 + a/3$ and $a = \sqrt{3}$ yields

$$C_1 = 18, \quad C_{10} = 5-2a, \quad C_{11} = 8, \quad C_{12} = 5+2a,$$

$$D_1 = 54, \quad D_{10} = 3-a, \quad D_{11} = 8a, \quad D_{12} = -3-a.$$

The conditions $\alpha = 1$ and $p+\sigma = 6$ lead to

$$(4.32)\qquad E_n = [(15-8a)y_n + 16ay_{n+1} - (15+8a)y_{n+2}]/30$$
$$+ h[(2-a)f_n + 8f_{n+1} + (2+a)f_{n+2} + 18f_{n+\lambda_1}]/30.$$

EXAMPLE 3. If we impose the condition (4.27) and choose $m=t=2$ and $p+\delta=5$, we have

$$\lambda_1 = 1-a/3,\ \lambda_2 = 1+a/3,\ a = \sqrt{6},\ C_1 = C_2 = 18,\ C_{10} = C_{22} = 8+3a,$$
$$C_{11} = C_{21} = 2,\ C_{12} = C_{20} = 8-3a,\ D_1 = D_2 = 54,\ D_{10} = -D_{22} = 3+a,$$
$$D_{21} = -D_{11} = 2a,\ D_{20} = -D_{12} = 3-a.$$

The choice $\alpha=1$ and $p+\sigma=6$ yields

$$(4.33)\qquad E_n = (y_n - y_{n+2})/2 - h(f_n - 14f_{n+1} + f_{n+2} - 9f_{n+\lambda_1} - 9f_{n+\lambda_2})/30,$$

for which (4.20) is satisfied.

4.3. Milne's device

Let

$$(4.34)\qquad \alpha_k^* y_{n+k}^* + \sum_{j=-r}^{k-1} \alpha_j^* y_{n+j} = h\Theta(x_n, y_{n-r},\dots, y_{n+k-1}; h)$$

be a predictor of order p which satisfies the conditions analogous to Conditions A, B and R, where $\alpha_k^*=1$ and $r\geqq 0$. Put $\tilde\rho(\zeta)=\sum_{j=-r}^{k}\alpha_j^*\zeta^j$, $\alpha^*=\tilde\rho'(1)$, and for any solution $z(x)$ of (1.1) let

$$\sum_{j=-r}^{k}\alpha_j^* z(x+jh) = h\Theta(x, z(x-rh),\dots, z(x+(k-1)h); h) + T^*(x, z(x); h).$$

Assume that $T^*(x, y; h)$ can be expressed as

$$T^*(x, y; h) = h^{p+1}\varphi_0^*(x, y) + O(h^{p+2}).$$

Then we have the following

THEOREM 5. *Suppose that*

$$(4.35)\qquad \varphi_0^*(x, y) = \gamma\varphi_0(x, y),\quad \gamma \neq 0,\quad \alpha^* \neq \alpha\gamma.$$

Then, for the predictor-corrector method (4.34)-(1.4), *under Conditions* I, H *and* L

$$(4.36)\quad C(y_{n+k} - y_{n+k}^*) = T(x_n; h) + O(h^{p+2})\qquad (n = J+r, J+r+1,\dots, N-k)$$

for sufficiently small h, where

$$(4.37)\qquad C = \alpha/(\alpha\gamma - \alpha^*).$$

PROOF. From (4.34) and the assumptions it follows that

$$\tilde{\rho}(1) = 0, \quad \Theta(x, y; 0) = \alpha^* f(x, y). \tag{4.38}$$

By (1.4) and (4.34) we have

$$\begin{aligned} y_{n+k} - y^*_{n+k} &= \sum_{j=-r}^{k} \alpha^*_j y_{n+j} - h\Theta(x_n, y_{n-r}, \ldots, y_{n+k-1}; h) \\ &= \sum_{j=-r}^{k} \alpha^*_j e_{n+j} - h \sum_{j=-r}^{k-1} \Theta_j(x_n, y(x_{n-r}), \ldots, y(x_{n+k-1}); h) e_{n+j} \\ &\quad + h^{p+1}\varphi^*_0(x_n) + O(h^{p+2}). \end{aligned}$$

Substituting (4.1) into the right side, expanding it at $x = x_n$ into power series in h and using (4.38), we have by (2.23)

$$y_{n+k} - y^*_{n+k} = h^{p+1}\varphi^*_0(x_n) - \alpha^* c h^{p+1}\varphi_0(x_n) + O(h^{p+2}),$$

from which (4.36) follows.

This theorem justifies Milne's device with C defined by (4.37) for sufficiently small h and large n.

Numerical examples

We use the following predictor and correctors:

$$y^*_{n+3} = 9(y_{n+1} - y_{n+2}) + y_n + 6h(f_{n+2} + f_{n+1}), \tag{4.39}$$

I. $y_{n+3} = y_{n+2} + h(9f_{n+3} + 19f_{n+2} - 5f_{n+1} + f_n)/24,$

II. $y_{n+3} = (2y_{n+1} + y_n)/3 + h(25f_{n+3} + 91f_{n+2} + 43f_{n+1} + 9f_n)/72,$

III. $y_{n+3} = y_n + 3h(f_{n+3} + 3f_{n+2} + 3f_{n+1} + f_n)/8,$

IV. $y_{n+3} = y_{n+1} + h(f_{n+3} + 4f_{n+2} + f_{n+1})/3.$

The following problems are solved by these formulas with $h = 2^{-5}$.

Problem 1. $y' = 2y, \quad y(0) = 1.$

Problem 2. $y' = -y^2, \quad y(0) = 1.$

Problem 3. $y' = 1 - y^2, \quad y(0) = 0.$

Problem 4. $y' = -5y, \quad y(0) = 1.$

Starting values are computed by the Runge-Kutta method. The local truncation error T and the value M of (4.36) at the step where the approximate value of $y(3)$ is computed are listed in Table 1. It is to be noted that the correctors III and IV do not satisfy the first part of Condition L.

Table 1.

Prob \ Form		I	II	III	IV
1	T	$-9.36-06$	$-7.05-06$	$-1.31-05$	$-4.02-06$
	M	$-8.90-06$	$-6.90-06$	$-1.31-05$	$-3.89-06$
2	T	$2.45-11$	$2.06-11$	$3.47-11$	$7.34-12$
	M	$2.87-11$	$2.38-11$	$-6.48-08$	$2.06-08$
3	T	$-1.24-10$	$-9.33-11$	$-1.89-10$	$-4.75-11$
	M	$-1.16-10$	$-8.80-11$	$1.43-08$	$-6.92-09$
4	T	$9.22-13$	$6.98-13$	$1.36-12$	$3.71-13$
	M	$1.05-12$	$7.23-13$	$-1.12-05$	$7.42-05$

REMARK. For the linear method $\Phi=\sum_{j=0}^{k}\beta_j f(x_{n+j}, y_{n+j})$ Condition H is satisfied if $\rho(\zeta)$ has no common factor with $\sigma(\zeta)=\sum_{j=0}^{k}\beta_j\zeta^j$.

5. Approximate computation of errors

In this section we assume that

$$e_0 = 0, \quad e_i = O(h^{p+1}) \qquad (i = 1, 2,\dots, k-1) \tag{5.1}$$

and approximate the errors e_{jm} ($j=0, 1,\dots, P$; $Pm\leqq N$) for a fixed positive integer m.

5.1. Method for approximation

Let $\Delta(x, y; h)$ be the function such that for any solution $z(x)$ of (1.1)

$$z(x+h) = z(x) + h\Delta(x, z(x); h). \tag{5.2}$$

Then it can be written as

$$\Delta(x, y; h) = \textstyle\sum_{j=0}^{r} h^j f^{(j)}(x, y)/(j+1)! + O(h^{r+1}) \qquad (r \geqq p).$$

From (2.6) and (2.18) it follows that

$$g_{j+1}(x) = g_j'(x) + g_j(x)g(x) \qquad (j = 0, 1,\dots).$$

Hence $g_j(x)$ can be written as a sum of products of $g(x)$ and its derivatives in the form

$$g_j(x) = \textstyle\sum_{k=0}^{j} g_{jk}(x) \qquad (j = 0, 1,\dots).$$

For instance $g_{00}=g$, $g_{10}=g'$ and $g_{11}=g^2$.

LEMMA 6. *For any integer s $(1\leqq s\leqq p+1)$ there exist an integer r $(r\geqq m)$*

and functions $A(x_n, y_n,\dots, y_{n+r}; h)$, $A_{jk}(x_n, y_n,\dots, y_{n+r}; h)$ $(j, k=0, 1,\dots, M)$ *and* $S(x_n, y_n,\dots, y_{n+r}; e_n, h)$ *such that*

$$(5.3)\qquad e_{n+m} = e_n + mh[\Delta(x_n, y_n; mh) - \Delta(x_n, y(x_n); mh)] + A + h \textstyle\sum_{j=0}^{M} h^j \sum_{k=0}^{j} A_{jk} g_{jk}(x_n) + h^{p+s+1} S,$$

where

$$(5.4)\qquad A = O(h^{p+1}),\quad A_{jk} = O(h^{p+1})\qquad (j, k = 0, 1,\dots, M;\ M \geqq s-2).$$

PROOF. Let D be the differential operator and Δ be the forward difference operator. Then there exists an integer r $(r \geqq m)$ such that

$$(5.5)\qquad y(x+jh) = y(x) + j[\textstyle\sum_{k=0}^{r} (jhD)^k/(k+1)!]hy'(x) + O(h^{p+s+\delta})\qquad (j = 1, 2,\dots, r),$$

where $\delta = \delta_{jm}$. Substituting

$$hD = \log(1+\Delta) = \Delta - \Delta^2/2 + \Delta^3/3 - \cdots$$

into (5.5), we have

$$y(x+jh) = y(x) + h \textstyle\sum_{k=0}^{r} \tilde{c}_{jk} \Delta^k y'(x) + O(h^{p+s+\delta}),$$

which can be rewritten as

$$(5.6)\qquad y(x+jh) = y(x) + h \textstyle\sum_{k=0}^{r} c_{jk} y'(x+kh) + O(h^{p+s+\delta})\qquad (j = 1, 2,\dots, r).$$

Let $u(x)$ be the solution of (1.1) with $u(x_n) = y_n$ and let

$$(5.7)\qquad u_{n+j} = u(x_{n+j}),\quad d_{n+j} = y_{n+j} - u_{n+j}\qquad (j = 0, 1,\dots, r),$$

$$(5.8)\qquad w_{n+j} = y_{n+j} - y_n - h \textstyle\sum_{k=0}^{r} c_{jk} f_{n+k}\qquad (j = 1, 2,\dots, r).$$

Since by (5.1) and Theorem 2

$$(5.9)\qquad e_{n+j} = h^p e(x_{n+j}) + O(h^{p+1})\qquad (j = 0, 1,\dots, r)$$

and by Gronwall's inequality

$$u_{n+j} - y(x_{n+j}) = e_n + O(h^{p+1})\qquad (j = 0, 1,\dots, r),$$

we have

$$(5.10)\qquad d_{n+j} = e_{n+j} + y(x_{n+j}) - u_{n+j} = O(h^{p+1})\qquad (j = 1, 2,\dots, r).$$

By (5.6)–(5.10)

$$(5.11)\qquad d_{n+j} = h \textstyle\sum_{k=1}^{r} c_{jk} g(x_{n+k}) d_{n+k} + w_{n+j} + O(h^{p+s+\delta})\qquad (j = 1, 2,\dots, r),$$

from which it follows that

(5.12) $$w_{n+j} = O(h^{p+1}) \qquad (j = 1, 2,\dots, r).$$

By (5.2) we have

(5.13) $$e_{n+m} = e_n + mh[\Delta(x_n, y_n; mh) - \Delta(x_n, y(x_n); mh)] + y_{n+m} - y_n - mh\Delta(x_n, u(x_n); mh).$$

From (5.6) it follows that

$$mh\Delta(x_n, u(x_n); mh) = h\sum_{k=0}^{r} c_{mk} f(x_{n+k}, u(x_{n+k})) + O(h^{p+s+1}) = h\sum_{k=0}^{r} c_{mk}[f_{n+k} - g(x_{n+k})d_{n+k}] + O(h^{p+s+1}).$$

By this and (5.13)

(5.14) $$e_{n+m} = e_n + mh[\Delta(x_n, y_n; mh) - \Delta(x_n, y(x_n); mh)] + w_{n+m} + h\sum_{k=1}^{r} c_{mk} g(x_{n+k})d_{n+k} + O(h^{p+s+1}).$$

Substituting (5.11) repeatedly into (5.14) and expanding the functions at $x=x_n$ into power series in h, we have (5.3) with

(5.15) $$A = w_{n+m},\quad A_{00} = \sum_{j=1}^{r} c_{mj} w_{n+j},\quad A_{10} = \sum_{j=1}^{r} j c_{mj} w_{n+j},\quad A_{11} = \sum_{j=1}^{r} c_{mj} \sum_{i=1}^{r} c_{ji} w_{n+i}$$

and so on. From (5.12) and this (5.4) follows. Thus the proof is complete.

In some cases we may take $r=m$ by using the interpolation.

Suppose that there exist a method of explicit one-step type for approximating e_{n+m} and constants K_1, K_2 and L such that

(5.16) $$e_{n+m} = e_n + mh\Psi(x_n, y_n,\dots, y_{n+r}; e_n, h) + h^{p+d+1}R(x_n, y_n,\dots, y_{n+r}; e_n, h) + h^{p+s+1}S(x_n, y_n,\dots, y_{n+r}; e_n, h),$$

(5.17) $$|R(x, u_0,\dots, u_r; w, v)| \leqq K_1,$$

(5.18) $$|S(x, u_0,\dots, u_r; w, v)| \leqq K_2,$$

(5.19) $$|\Psi(x, u_0,\dots, u_r; w, v) - \Psi(x, u_0,\dots, u_r; \tilde{w}, v)| \leqq L|w - \tilde{w}|$$

$$\text{for} \quad v\in H,\ x,\ x+rh\in I,\ u_i,\ u_i-w,\ u_i-\tilde{w}\in B_M \qquad (i = 0, 1,\dots, r).$$

Let P be an integer such that $(P-1)m+r\leqq N$ and define $\tilde{e}_{jm}$ $(j=0, 1,\dots, P)$ by

(5.20) $$\tilde{e}_{n+m} = \tilde{e}_n + mh\Psi(x_n, y_n,\dots, y_{n+r}; \tilde{e}_n, h)\ (n = jm; j = 0, 1,\dots, P),\ \tilde{e}_0 = 0.$$

Then we have the following

THEOREM 6. *Under the condition* (5.1) *suppose that there exist functions* Ψ, R *and* S *satisfying* (5.16)–(5.19) *and let* $\tilde{e}_{jm}$ $(j=0, 1,\dots, P)$ *be defined by* (5.20). *Then*

$$e_{jm} = \tilde{e}_{jm} + O(h^{p+t}) \qquad (j = 0, 1,\dots, P) \tag{5.21}$$

for sufficiently small h, *where* $t=\min(s, d)$.

PROOF. Let $v_k=e_k-\tilde{e}_k$ $(k=jm; j=0, 1,\dots, P)$. Then for $n=jm$ $(0\leqq j\leqq P-1)$ we have

$$\begin{aligned} v_{n+m} = v_n &+ mh[\Psi(x_n, y_n,\dots, y_{n+r}; e_n, h) - \Psi(x_n, y_n,\dots, y_{n+r}; \tilde{e}_n, h)] \\ &+ h^{p+d+1}R + h^{p+s+1}S. \end{aligned}$$

Let $u=1+mLh$ and K be a constant such that

$$K_1h^d + K_2h^s \leqq Kh^t \qquad \text{for} \quad h\in H.$$

Then

$$|v_{n+m}| \leqq u|v_n| + h^{p+t+1}K,$$

so that

$$\begin{aligned} |v_{jm}| &\leqq (1 + u + \cdots + u^{j-1})Kh^{p+t+1} \leqq jhe^{L(j-1)mh}Kh^{p+t} \\ &\leqq m^{-1}(b-a)e^{L(b-a)}Kh^{p+t} \qquad (j = 0, 1,\dots, P). \end{aligned}$$

This completes the proof.

In the case of variable stepsize where

$$x_{(j+1)m} = x_{jm} + mh_j \quad (j = 0, 1,\dots, P-1), \quad x_{Pm} + (r-m)h_{P-1} \leqq b,$$

if y_{n+i} $(i=0, 1,\dots, r)$ in (5.16) denote the approximate values of $y(x_n+ih_j)$ $(n=jm)$, then (5.21) is valid with $h=\max_{0\leqq j<P} h_j$.

5.2. Examples

In this subsection we consider the case $m=4$.

5.2.1. Formulas (5.6)

We use the notation $c_{jk}=C_{jk}/C_j$ $(j=1, 2,\dots, r;\ k=0, 1,\dots, r)$.

EXAMPLE 4. In the case $r=4$ we have $p+s=6$ and

$$C_1 = 720,\ C_{10} = 251,\ C_{11} = 646,\ C_{12} = -264,\ C_{13} = 106,\ C_{14} = -19; \tag{5.22}$$

$$C_2 = 90,\ C_{20} = 29,\ C_{21} = 124,\ C_{22} = 24,\ C_{23} = 4,\ C_{24} = -1;$$

$$C_3 = 80,\ C_{30} = 27,\ C_{31} = 102,\ C_{32} = 72,\ C_{33} = 42,\ C_{34} = -3;$$

$$C_4 = 90,\ C_{40} = C_{44} = 28,\ C_{41} = C_{43} = 128,\ C_{42} = 48.$$

EXAMPLE 5. In the case $r=6$ we have $p+s=7$ and

$$\begin{aligned}(5.23)\quad & C_1 = 60480,\ C_{10} = 19087,\ C_{11} = 65112,\ C_{12} = -46461,\ C_{13} = 37504,\\ & C_{14} = -20211,\ C_{15} = 6312,\ C_{16} = -863;\ C_2 = 3780,\ C_{20} = 1139,\\ & C_{21} = 5640,\ C_{22} = 33,\ C_{23} = 1328,\ C_{24} = -807,\ C_{25} = 264,\ C_{26} = -37;\\ & C_3 = 2240,\ C_{30} = 685,\ C_{31} = 3240,\ C_{32} = 1161,\ C_{33} = 2176,\\ & C_{34} = -729,\ C_{35} = 216,\ C_{36} = -29;\ C_4 = 945,\ C_{40} = 286,\ C_{41} = 1392,\\ & C_{42} = 384,\ C_{43} = 1504,\ C_{44} = 174,\ C_{45} = 48,\ C_{46} = -8;\ C_5 = 12096,\\ & C_{50} = 3715,\ C_{51} = 17400,\ C_{52} = 6375,\ C_{53} = 16000,\ C_{54} = 11625,\\ & C_{55} = 5640,\ C_{56} = -275;\ C_6 = 140,\ C_{60} = C_{66} = 41,\ C_{61} = C_{65} = 216,\\ & C_{62} = C_{64} = 27,\ C_{63} = 272.\end{aligned}$$

5.2.2. Formulas (5.16)

Let

$$F(x, y, u) = f(x, y) - f(x, y-u). \tag{5.24}$$

EXAMPLE 6. In the case $s=2$ and $M=0$ let

$$F_1 = F(x_n, y_n, e_n),\ F_2 = F(x_{n+2}, y_{n+2}, e_n + 2hF_1 + b),\ b = A_{00}/4.$$

Then we have

$$e_{n+4} = e_n + A + 4hF_2 + O(h^{p+3}), \tag{5.25}$$

$$e_{n+4} = e_n + A + 2h(F_1 + 4F_2 + F_3)/3 + O(h^{p+3}), \tag{5.26}$$

where

$$F_3 = F(x_{n+4}, y_{n+4}, e_n - 4hF_1 + 8hF_2 + 2b).$$

There exists a 4-stage method

$$e_{n+4} = e_n + A + 2h(F_1 + 2F_2 + 2F_3 + F_4)/3 + O(h^{p+3}), \tag{5.27}$$

where

$$F_3 = F(x_{n+2}, y_{n+2}, e_n + 2hF_2 + b),$$

$$F_4 = F(x_{n+4}, y_{n+4}, e_n + 4hF_3 + 2b).$$

EXAMPLE 7. In the case $s \geqq 2$ we have

$$e_{n+4} = e_n + A + 2h(F_1 + F_2) + O(h^{p+t+1}), \tag{5.28}$$

where

$$F_1 = F(x_n, y_n, e_n + b_1), \quad F_2 = F(x_{n+3}, y_{n+3}, e_n + 4hF_1 + b_2),$$

$$b_1 = (3A_{00} - A_{10})/4, \quad b_2 = (A_{10} - A_{00})/4, \quad t = \min(2, s).$$

There is also a 3-stage method

$$e_{n+4} = e_n + A + 4h(2F_1 + 3F_2 + 4F_3)/9 + O(h^{p+t+1}), \tag{5.29}$$

where

$$F_1 = F(x_n, y_n, e_n + b_1), \quad F_2 = F(x_{n+2}, y_{n+2}, e_n + 2hF_1 + b_2),$$

$$F_3 = F(x_{n+3}, y_{n+3}, e_n + 3hF_2 + b_3), \quad b_1 = (12A_{00} - 4A_{10} - A_{11})/8,$$

$$b_2 = (A_{10} + A_{11} - 3A_{00})/4, \quad b_3 = (6A_{00} + A_{10} - 2A_{11})/16, \quad t = \min(3, s).$$

5.2.3. Numerical examples

The predictor (4.39) and correctors I–IV are used to solve Problem 3 and the following problems with $h = 2^{-5}$.

Table 2.

Prob \ Form		I	II	III	IV
	e	1.96−09	6.34−10	1.21−08	−6.21−09
3	$\bar{e}$	1.97−09	6.36−10	1.21−08	−6.20−09
	$\hat{e}$	1.97−09	6.35−10	1.21−08	−6.12−09
	e	3.38−05	1.14−05	1.75−05	6.99−06
5	$\bar{e}$	3.33−05	1.10−05	1.70−05	6.54−06
	$\hat{e}$	3.37−05	1.13−05	1.74−05	6.91−06
	e	1.34+00	4.92−01	7.33−01	3.28−01
6	$\bar{e}$	1.32+00	5.01−01	7.36−01	3.41−01
	$\hat{e}$	1.30+00	4.81−01	7.16−01	3.21−01
	e	1.02−10	2.56−11	1.21−05	−7.49−05
7	$\bar{e}$	9.52−11	2.42−11	1.20−05	−7.46−05
	$\hat{e}$	9.45−11	2.34−11	1.18−05	−6.77−05

Problem 5. $y' = y - 2x/y, \quad y(0) = 1.$

Problem 6. $y' = 2xy, \quad y(0) = 1.$

Problem 7. $y' = 5(1 - y), \quad y(0) = 0.$

Starting values are computed by the Runge-Kutta method. Formula (5.29) is used with quantities in (5.15) whose coefficients are given by (5.22) and (5.23). The error e at $x=3$ and the values $\bar{e}$ and $\hat{e}$ obtained respectively by using (5.22) and (5.23) are listed in Table 2.

For $\bar{e}$ we have $r=4$, $s=t=2$ and $M=1>s-2$, while for $\hat{e}$ we have $r=6$, $s=t=3$ and $M=1=s-2$.

5.3 Explicit one-step methods

We show the following

THEOREM 7. *Let E_n be given by* (4.15) *or by* (4.31) *satisfying* (4.27) *and suppose that $\sigma \geqq 2$ and* (4.20) *is satisfied. Then for explicit one-step methods with $p \geqq 2$*

$$A = -mE_n, \quad A_{00} = -m^2E_n/2, \quad s = 2. \tag{5.30}$$

PROOF. Let u_{n+j} and d_{n+j} $(j=0, 1,\dots, m)$ be defined by (5.7). Since

$$u_{n+j+1} = u_{n+j} + h\Delta(x_{n+j}, u_{n+j}; h) \qquad (j = 0, 1,\dots, m-1),$$

$$y_{n+j+1} = y_{n+j} + h\Delta(x_{n+j}, y_{n+j}; h) - T(x_{n+j}, y_{n+j}; h), \tag{5.31}$$

by (5.10) we have $d_n=0$,

$$d_{n+j+1} = d_{n+j} - h^{p+1}\varphi_0(x_n) + O(h^{p+2}) \qquad (j = 0, 1,\dots, m-1).$$

From this it follows that

$$d_{n+j} = -jh^{p+1}\varphi_0(x_n) + O(h^{p+2}) \qquad (j = 0, 1,\dots, m). \tag{5.32}$$

By (5.31)

$$\begin{aligned} e_{n+j+1} = e_{n+j} &+ h[\Delta(x_{n+j}, u(x_{n+j}); h) - \Delta(x_{n+j}, y(x_{n+j}); h)] \\ &+ h\Delta_y(x_{n+j}, y_{n+j}; h)d_{n+j} - T(x_{n+j}; h) + O(h^{2p+1}) \\ &\qquad (j = 0, 1,\dots, m-1). \end{aligned} \tag{5.33}$$

Since for any solution $z(x)$ of (1.1)

$$h\textstyle\sum_{j=0}^{m-1} \Delta(x_{n+j}, z(x_{n+j}); h) = mh\Delta(x_n, z(x_n); mh),$$

by (5.32) and (5.33) we have

$$e_{n+m} = e_n + mh[\Delta(x_n, y_n; mh) - \Delta(x_n, y(x_n); mh)] - \textstyle\sum_{j=0}^{m-1} T(x_{n+j}; h)$$
$$- m(m-1)h^{p+2}g(x_n)\varphi_0(x_n)/2 + O(h^{p+3}).$$

Substitution of (4.21) into this yields (5.30).

Numerical examples

Problem 5 and the following problem are solved by the Runge-Kutta method and Kutta's method for $m=4$.

Problem 8. $y' = 2xe^{4x^2}/y^3, \quad y(0) = 1.$

E_n is computed by means of (4.22). Formulas (5.26) and (5.27) are used when $p=3$ and 4 respectively.

The same problems are solved by the Runge-Kutta method for $m=2$ with the aid of (4.33) and the formula

$$e_{n+2} = e_n - 2E_n + h(F_1 + F_2) + O(h^{p+3}), \tag{5.34}$$

where

$$F_1 = F(x_n, y_n, e_n - b), \quad F_2 = F(x_{n+2}, y_{n+2}, e_n + 2hF_1 - 2b), \quad b = 2E_n/3.$$

Computation is carried out by the following program:

(1) Compute y_i ($i=1, 2,\ldots, m$) and $\tilde{e}_m$.

(2) If $|mE_0| > 10^{-8}\max(|y_m|, 1)$, halve the stepsize and go to (1). (Initially $h=2^{-3}$.)

(3) Replace x_0, y_0 and $\tilde{e}_0$ by x_m, y_m and $\tilde{e}_m$ respectively.

The error e and the computed value $\tilde{e}$ are listed in Table 3.

Table 3.

Formula		(5.26)		(5.27)		(5.34)	
Prob	x	$\tilde{e}$	e	$\tilde{e}$	e	$\tilde{e}$	e
	3.0	5.90−06	5.85−06	1.96−06	1.97−06	2.15−06	2.18−06
5	4.0	3.85−05	3.82−05	1.29−05	1.30−05	1.40−05	1.43−05
	5.0	2.57−04	2.55−04	8.65−05	8.71−05	9.20−05	9.59−05
	3.0	−1.60−04	−1.58−04	3.70−05	3.83−05	2.49−04	2.49−04
8	4.0	−4.07−01	−4.06−01	5.14−02	5.26−02	5.24−02	5.26−02
	5.0	−8.15+02	−7.96+02	1.03+03	1.05+03	1.05+03	1.05+03

References

[1] P. Henrici, *Discrete variable methods in ordinary differential equations,* John Wiley and Sons, New York, 1962.

[2] E. Isaacson and H. B. Keller, *Analysis of numerical methods,* John Wiley and Sons, New York, 1966.

[3] H. Shintani, *Approximate computation of errors in numerical integration of ordinary differential equations by one-step methods,* J. Sci. Hiroshima Univ., Ser. A-I, **29** (1965), 97–120.

[4] H. Shintani, *On one-step methods for ordinary differential equations,* Hiroshima Math. J., **7** (1977), 769–786.

Department of Mathematics,
Faculty of School Education,
Hiroshima University