数理解析研究所講究録
676巻 1988年 16-32

New Method of A-Posteriori Estimates of Truncation Errors for the Adams Predictor-Corrector Methods

Masatomo FUJII

福岡教育大学
藤 井 正 友

Department of Mathematics, Fukuoka University of Education

1. Introduction

There have been many works which treat multistep methods with variable step/variable order forms (e.g.[4]). While the principal local truncation error has been estimated by the method so called Milne's device which has been discussed in many literatures (e.g.[1,2, 3,5]).

Here we shall consider an accurate method for obtaining better estimates of local truncation errors of the Adams-Bashforth-Moulton methods of p-th order (p=2,3,4,5) with the fixed step/fixed order form in the mode of correcting to convergence. We shall also consider a method for estimating the global truncation errors by using the estimated local truncation errors.

2. Preliminaries

We are concerned with an initial value problem

(2.1) $\quad y' = f(x,y), \quad y(a) = y_0 \quad (a \leqq x \leqq b),$

where we denote by $y(x)$ the solution of this problem and

(2.2) $\quad x_n = a + nh \quad (n=0,1,\ldots,N), \; h = (b - a)/N.$

In what follows, we shall assume that $f(x,y)$ in (2.1) is reasonably smooth on regions in question and numerical operations are carried out with sufficient precision in order to assure the round-off errors are negligible in comparison with global truncation errors.

Let us put

(2.3) $\quad v = n + p - 1.$

Then the formulae of the Adams predictor-corrector method of order p are given by

(2.4) $\quad y_v^* = y_{v-1} + h\sum_{j=1}^{p} a_{pj} f_{v-j},$

(2.5) $\quad y_v = y_{v-1} + h\sum_{j=0}^{p-1} b_{pj} f_{v-j},$

where y_v^* is the value of the predictor, y_v is that of the corrector and $f_{v-j} = f(x_{v-j}, y_{v-j})$. Needless to say, y_μ $(\mu=0,1,\ldots,p-1)$ are starting values. The coefficients in the formulae (2.4) and (2.5) are shown in Henrici [2].

For the formulae (2.4) and (2.5), we define the local truncation errors at x_n by

$$(2.6)\quad T_{p1n} = y(x_v) - y(x_{v-1}) - h\sum_{j=1}^{p} a_{pj} f[x_{v-j}, y(x_{v-j})],$$

$$(2.7)\quad T_{p2n} = y(x_v) - y(x_{v-1}) - h\sum_{j=0}^{p-1} b_{pj} f[x_{v-j}, y(x_{v-j})]$$

respectively, and they are rewritten as follows:

$$(2.8)\quad T_{pin} = \sum_{j=1}^{r} h^{p+j} k_{pij} y^{(p+j)}(x_n) + O(h^{p+r+1}) \quad (i=1,2),\ r \geqq 1.$$

The coefficients k_{pij} are shown in Table 2.1.

3. The behavior of $y_v^* - y_v$

In this section, we shall investigate the behavior of $y_v^* - y_v$. By virtue of the consideration as a quadrature rule of p-th order, the coefficients b_{pj} satisfy the relation

$$(3.1)\quad \sum_{j=0}^{p-1} j^i b_{pj} = 1/(i+1) \quad (i = 0,1,\ldots,p-1),$$

which plays an important role in this paper.

For brevity, let us put

$$(3.2)\quad g(x) = f_y[x, y(x)],$$

(3.3) $$g_v^{(m)} = g^{(m)}(x_v),$$

(3.4) $$g_v = g(x_v).$$

(3.5) $$e_v = y_v - y(x_v),$$

that is, e_v is the global truncation error at x_v.

Suppose that p starting values are chosen so that

(3.6) $$e_\mu = O(h^q) \quad (q \geqq p+1; \mu=0,1,\ldots,p-1).$$

As is well known, e_n can be expressed as follows [2]:

(3.7) $$e_n = h^p e(x_n) + O(h^{p+1}),$$

where $e(x)$ is the solution of the initial value problem

(3.8) $$e' = g(x)e - Cy^{(p+1)}(x), \quad e(a) = 0,$$

where C is the error constant of the formula defined by (2.5).

For the analysation of the behavior of $y_v^* - y_v$, the following lemma plays an important role:

LEMMA 3.1. For the Adams-Bashforth-Moulton pair of p-th order

(3.9) $$y_v^* = y_{v-1} + h \sum_{j=0}^{p-1} \gamma_j \nabla^j f_{v-1},$$

$$(3.10) \qquad y_v = y_{v-1} + h \sum_{j=0}^{p-1} \hat{\gamma}_j \nabla^j f_v$$

the identity

$$(3.11) \qquad y_v^* - y_v = - \gamma_{p-1} h \nabla^p f_v$$

holds, Here the rational numbers γ_j and $\hat{\gamma}_j$ are given by

$$(3.12) \qquad \gamma_j = \frac{1}{j!}\int_0^1 s(s+1)\cdots(s+j-1)ds,$$

$$(3.13) \qquad \hat{\gamma}_j = \frac{1}{j!}\int_0^1 (s-1)s\cdots(s+j-2)ds.$$

Simirarly, we see that

$$(3.14) \qquad T_{p2n} - T_{p1n} = - \gamma_{p-1} h \nabla^p y'(x_v).$$

Thus we have the relation

$$(3.15) \qquad y_v^* - y_v = - \gamma_{p-1} h \nabla^p [f_v - y'(x_v)] + T_{p2n} - T_{p1n}.$$

We also have the following

LEMMA 3.2. It holds that

$$(3.16) \quad \sum_{j=0}^{p} (-1)^j P_i(j)\binom{p}{j} = 0 \quad \text{for } p-1 \geqq i \geqq 0,$$

where $P_i(j)$ is a polynomial of i-th degree in j.

From (2.3), (2.5) and (2.7), we have

$$(3.17) \quad e_v = e_{v-1} + h\sum_{j=0}^{p-1} b_{pj} g_{v-j} e_{v-j} - T_{p2n} + O(h^{2p+1}).$$

Then from (3.6) and (3.7), it is readily seen that

$$(3.18) \quad e_{v-j} = e_v + O(h^{p+1}) \text{ for } v - j \geqq 0;\ p \geqq j \geqq 1.$$

Since

$$(3.19) \quad \nabla^p[f_v - f(x_v, y(x_v))] = \nabla^p(g_v e_v) + O(h^{2p})$$

and

$$(3.20) \quad \nabla^p(g_v e_v) = \sum_{j=0}^{p} (-1)^j \binom{p}{j} g_{v-j} e_{v-j}.$$

From (3.18) and applying Lemma 3.2, we see that

$$(3.21) \quad \nabla^p(g_v e_v) = \sum_{j=0}^{p} (-1)^j \binom{p}{j} [g_v e_v + O(h^{p+1})]$$

$$= O(h^{p+1}) \qquad \text{for } v \geqq p.$$

From (3.15),(3.19) and (3.21), we have

$$(3.22)\quad y_v^* - y_v = T_{p2n} - T_{p1n} + O(h^{p+2}) + O(h^{2p+1}) \quad \text{for } v \geqq p.$$

Furthermore from (3.1), (3.18) and (3.17), it is seen that

$$(3.23)\quad e_{v-j} = e_v - jhg_v e_v + jh^{p+1}k_{p21}y^{(p+1)}(x_n)$$

$$+ O(h^{p+2}) + O(h^{2p+1}) \quad \text{for } v-j \geqq p-1;\ p \geqq j \geqq 1.$$

Substituting (3.23) into (3.20) and applying Lemma 3.2, we obtain

$$(3.24)\quad \nabla^p(g_v e_v) = \textstyle\sum_{j=0}^{p}(-1)^j\binom{p}{j}[g_v e_v - jh(g_v^{(1)} + g_v^2)e_v$$

$$+ jh^{p+1}k_{p21}g_v y^{(p+1)}(x_n) + O(h^{p+2}) + O(h^{2p+1})]$$

$$= O(h^{p+2}) + O(h^{2p+1}) \quad \text{for } v \geqq 2(p-1)+1.$$

From (3.15), (3.19) and (3.24), we have

$$(3.25)\quad y_v^* - y_v = T_{p2n} - T_{p1n} + O(h^{p+3}) + O(h^{2p+1})$$

$$\text{for } v \geqq 2(p-1)+1.$$

If $p=2$, then we cannot obtain a better relation than (3.25) even when n is large, because of (3.1) and $p+3=2p+1$. But when $p > 2$, the above process can be repeated further and thus, summerizing the whole results including (3.22) and (3.25) we have the following

THEOREM 3.1. It holds that

$$(3.26) \quad y^*_v - y_v = T_{p2n} - T_{p1n} + \varepsilon_{p,v} + O(h^{2p+1}),$$

where

$$(3.27) \quad \varepsilon_{p,v} = O(h^{p+i+1}) \quad \text{for} \quad v \geqq i(p-1)+1;\ p \geqq i \geqq 1.$$

REMARK. If the solution $y(x)$ of (2.1) a polynomial of which degree is less than $p+1$ (p is the order of the integration method), then the local truncation error is vanish. Thus the exact solution is obtained.

4. Estimation of the local truncation error

For simplicity, let us put

$$(4.1) \quad d_{v+i} = y^*_{v+i} - y_{v+i} \quad (i=0,1,\ldots,p-1).$$

Then, taking the results of Theorem 3.1 into consideration, we have the following

THEOREM 4.1. If $n \geqq \alpha_{pr}$(the values of α_{pr} are given below in this Theorem for ($p=2,3,4,5$; $r=1,2,\ldots,P$)), then the best estimate A_{prn} ($r=1,2,\ldots,p$) of T_{p2n} ($p=2,3,4,5$) among $\binom{p}{r}$ linear combinations of the values of d_{v+i} $(i=0,1,\ldots,p-1)$ are given as follows:

For $p=2$,

$$A_{21n} = d_{v+1}/6 \quad \text{or} \quad d_v/6 \qquad (\alpha_{21}=2),$$

$$A_{22n} = (d_{v+1} + d_v)/12 \qquad (\alpha_{22}=1).$$

For $p=3$,

$$A_{31n} = d_{v+1}/10 \qquad (\alpha_{31}= 3),$$

$$A_{32n} = (-11d_{v+2} + 41d_{v+1})/300$$

$$\text{or } (19d_{v+1} + 11d_v)/300 \qquad (\alpha_{32}=5).$$

$$A_{33n} = (-11d_{v+2} + 60d_{v+1} + 11d_v)/600 \qquad (\alpha_{33}= 1).$$

For $p=4$,

$$A_{41n} = 19d_{v+1}/270 \qquad (\alpha_{41}= 4),$$

$$A_{42n} = (-11d_{v+2} + 49d_{v+1})/540 \qquad (\alpha_{42}= 7),$$

$$A_{43n} = (191d_{v+3} - 844d_{v+2} + 2249d_{v+1})/22680$$

or $(-271d_{v+2} + 1676d_{v+1} + 191d_v)/22680 \quad (\alpha_{43}=10)$,

$A_{44n} = (191d_{v+1} - 1115d_{v+2} + 3925d_{v+1} + 191d_v)/45360 \ (\alpha_{44}= 1)$.

For p=5,

$A_{51n} = 27d_{v+1}/502 \quad (\alpha_{51}= 5)$,

$A_{52n} = (-271d_{v+2} + 1405d_{v+1})/21084 \quad (\alpha_{52}= 9)$,

$A_{53n} = (191d_{v+3} - 924d_{v+2} + 3001d_{v+1})/42168 \quad (\alpha_{53}=13)$,

$A_{54n} = (-2497d_{v+4} + 13221d_{v+3} - 35211d_{v+2} + 92527d_{v+1})/1265040$

or $(3233d_{v+3} - 20229d_{v+2} + 82539d_{v+1} + 2497d_v)/1265040$

$(\alpha_{54}=17)$,

$A_{55n} = (-2497d_{v+4} + 16454d_{v+3} - 55440d_{v+2} + 175066d_{v+1}$

$+ 2497d_v)/2530080 \quad (\alpha_{55}= 1)$.

Proof. We shall prove this theorem by investigating all possible cases. Let us denote $Qh^i y^{(i)}(x_n)$ and $O(h^i)$ simply by $[Q]_i$ and O_i respectively. Then applying Theorem 3.1, we see that the followingresults hold:

For p=2,

---

r=1

$$\# \quad d_v/6 - T_{22n} = [1/24]_4 + \begin{cases} O_4 & n = 1 \\ O_5 & n \geqq 2, \end{cases}$$

$$\# \quad d_{v+1}/6 - T_{22n} = -[1/24]_4 + O_5 \qquad n \geqq 1,$$

$(\alpha_{21}=2)$.

(When n=1, the right-hand side of the upper relation is not valid. Thus we cannot say which of $d_v/6$ and $d_{v+1}/6$ is better. But when $n \geqq 2$, the values of both principal parts are same in magunitude. Hence we take 2 as α_{21}.)

---

r=2

$$\# \quad (d_{v+1} + d_v)/12 - T_{22n} = \begin{cases} O_4 & n = 1 \\ O_5 & n \geqq 2, \end{cases}$$

$(\alpha_{22}=1)$.

(In this block, there is only one object. Hence we take 1 as α_{22}.)

(By the same consideration as in the case p=2, we can determine the value of α_{pr}.)

For p=3,

r=1

$$d_v/10 - T_{32n} = [19/720]_5 + \begin{cases} 0_5 & 2 \geqq n \geqq 1 \\ 0_6 & n \geqq 3, \end{cases}$$

$$\# \quad d_{v+1}/10 - T_{32n} = - [11/720]_5 + \begin{cases} 0_5 & n = 1 \\ 0_6 & n \geqq 2 \end{cases}$$

$$d_{v+2}/10 - T_{32n} = -[41/720]_5 + 0_6 \qquad n \geqq 1,$$

$(\alpha_{31}=3)$,

---

r=2

$$\# \quad (19d_{v+1} + 11d_v)/300 - T_{32n} = - [11/1440]_6 + \begin{cases} 0_5 & 2 \geqq n \geqq 1 \\ 0_6 & 4 \geqq n \geqq 3 \\ 0_7 & \underline{\underline{n}} > 5, \end{cases}$$

$$(19d_{v+2} + 41d_v)/600 - T_{32n} = - [30/1440]_6 +$$

+ { The same as the above },

$$\# \quad (-11d_{v+2} + 41d_{v+1})/300 - T_{32n} = [11/1440]_6 + \begin{cases} 0_5 & n = 1 \\ 0_6 & 3 \geqq n \geqq 2 \\ 0_7 & n \geqq 4, \end{cases}$$

$(\alpha_{32}=5)$,

---

r=3

$$\# \quad (-11d_{v+2} + 60d_{v+1} + 11d_v)/600 - T_{32n} = \begin{cases} 0_5 & 2 \geqq n \geqq 1 \\ 0_6 & 4 \geqq n \geqq 3 \\ 0_7 & n \geqq 5, \end{cases}$$

$(\alpha_{33}=1)$.

....

As is readily seen from the above statements, for each p, p blocks have been constructed by $\binom{p}{r}$ rows. When $n \geqq \alpha_{pr}$, the rows with # show that the quantity A_{prn} has the smallest tolerance from T_{p2n} in each block. Therefore, they correspond to the best quantity A_{prn} for estimating the local truncation errors in each block. Thus the proof is completed.

5. Estimation of the global truncation error

Let us put

$$(5.1) \qquad \hat{g}_m = f_y(x_m, y_m) \quad (m=0,1,\ldots,N),$$

and by $\hat{e}_{prv}$ we denote the solution of the following equation:

(5.2) $\quad \hat{e}_{prv} = \hat{e}_{pr,v-1} + h\, \Sigma_{j=0}^{p-1} b_{pj} \hat{g}_{v-j} \hat{e}_{pr,v-j} - A_{prn},$

(5.3) $\quad \hat{e}_{pr\mu} = 0 \quad (\mu=0,1,\ldots,p-1).$

We also put

(5.4) $\quad G_{prm} = \hat{e}_{prm} - e_m \quad (m=0,1,\ldots,N).$

Since

$$\hat{g}_m \hat{e}_m - g_m e_m = \hat{g}_m G_{prm} + O(e_m^2) \quad (m=0,1,\ldots,N),$$

from (3.15), (5.2) and the assumption (3.6), G_{prv} satisfies the following relation:

(5.5) $\quad G_{prv} = G_{pr,v-1} + h\, \Sigma_{j=0}^{p-1} b_{pj} \hat{g}_{v-j} G_{pr,v-j} + \lambda_{prn},$

(5.6) $\quad G_{pr\mu} = O(h^q) \quad (q \geqq p+1;\ \mu = 0,1,\ldots,p-1),$

where

(5.7) $\quad \lambda_{prn} = \Sigma_{i=0}^{r-1} O(\varepsilon_{p,v+i+\delta}) + O(h^{p+r+1}) + O(h^{2p+1}),$

where δ is 0 or 1 depending on the form of A_{prn} (cf. Theorem 4.1).

Let L_1, L_2 and $K \geqq 0$ be constants such that

$$|g_m| \leqq L_1, \ |f_{yy}(x_m, \xi_m)| \leqq L_2 \ \text{ and } \ e_m = \theta_m h^p K, \quad |\theta_m| < 1,$$

where ξ_m is a value between y_m and $y(x_m)$. Then we see that

$$|\hat{g}_m| \leqq L_1 + h^p K L_2$$

Furthermore, let B, L, Z, b, L^* be constants such that

$$B = \textstyle\sum_{j=0}^{p-1} |b_{pj}|, \quad L_1 + h^p K L_2 = L,$$
$$|G_{pr\mu}| \leqq Z \qquad (\mu=0,1,\ldots,p-1),$$
$$L|b_{p0}| \leqq b, \quad L^* = (1 - hb)^{-1} LB.$$

About the global truncation error, we have the following

THEOREM 5.1. If $0 \leqq h < b^{-1}$, then it holds that

$$(5.8) \qquad |G_{prv}| \leqq (1 - hb)^{-1}(Z + \textstyle\sum_{k=1}^{n} |\lambda_{prk}|) \cdot \exp[(x_v - a)L^*].$$

Till now, we considered the initial value problem (2.1) for a single equation. But here we consider (2.1) for a system of equations. Then we have the following

ALGORITHM 5.1. Starting from

$$(5.9) \qquad \hat{e}_{pr\mu} = 0 \qquad (\mu=0,1,\ldots,p-1),$$

and solving the system of linear equations

(5.10) $(I - hb_{p0}\hat{g}_v)\hat{e}_{prv} = \hat{e}_{pr,v-1} + h \sum_{j=1}^{p-1} b_{pj}\hat{g}_{v-j}\hat{e}_{pr,v-j} - A_{prn}$,

we can obtain the estimate $\hat{e}_{prv}$ of the global truncation error e_v, provided that $1 > |hb_{p0}|\|\hat{g}_v\|$, where $\|\cdot\|$ is a suitable norm and I is the unit matrix.

As is readily seen from Theorem 3.1, the first term of the right hand side in (5.7) is not small in early steps and affects on G_{prv} in the later ones.

References

[1] G. Hall and J. M. Watt, Modern numerical methods for ordinary differential equations, Clarendon Press, Oxford,1976.

[2] P. Henrici, Discrete variable methods in ordinary differential equations, Wiley, New York, 1962.

[3] J. D. Lambert, Computational methods in ordinary differential equations, Wiley, New York, 1973.

[4] L. F. Shampine and N. K. Gordon, Computer solution of ordinary differential equations, The initial value problem, W. H. Freeman and Campany, San Francisco, 1975.

[5] H. Shintani, On errors in the numerical solutions of ordinary differential equations by step-by-step methods, Hiroshima Math. J. **10** No. 2, 1980, 469-494.

Table 2.1

The coeficients k_{pij}

j	1	2	3	4	5
k_{21j}	5/12	-1/24			
k_{22j}	-1/12	-1/24			
k_{31j}	3/8	29/180	3/40		
k_{32j}	-1/24	-17/360	-7/240		
k_{41j}	251/720	95/288	6313/30240	265/2688	
k_{42j}	-19/720	-13/288	-1247/30240	-71/2688	
k_{51j}	95/288	14531/30240	7157/17280	476981/1814400	139867/1036800
k_{52j}	-3/160	-641/15120	-175/3456	-38237/907200	-28303/1036800