Experimental Study on Round-off Error in Ordinary Differential Equations

Takashi Yoshimura

(Received on 27 October, 1977)

1 Accumulated round-off error

In Adams method the approximate values y_n to the true solution $y(x_n)$ of initial value problem

(1) $y' = f(x, y), \quad y(x_0) = y_0$

are calculated recursively according to the following formula:

(2) $y_{n+k} - y_{n+k-1} = h(\beta_k f_{n+k} + \beta_{k-1} f_{n+k-1} + \cdots\cdots + \beta_0 f_n) \quad n=0, 1, 2\cdots\cdots$

where k is a fixed integer, $f_m = f(x_m, y_m)$ and $x_m = x_0 + mh$ $(m=0, 1, 2, \cdots\cdots)$, and β_j $(j = 0, 1, 2, \cdots\cdots, k)$ denote real constants which do not depend on h. We shall always assume that $\beta_0 \neq 0$.

In most cases, however, the numbers y_n cannot be calculated with infinite precision because of the finite accuracy of any computing equipment. Hence the quantities $\tilde{y}_n$, that are actually calculated in place of y_n, satisfy an equation which we write in the form

(3) $\tilde{y}_{n+k} - \tilde{y}_{n+k-1} = h\{\beta_k f(x_{n+k}, \tilde{y}_{n+k}) + \cdots\cdots + \beta_0 f(x_n, \tilde{y}_n)\} + \varepsilon_{n+k}$
$n = 0, 1, 2, \cdots\cdots$

where the quantity ε_{n+k} is the local round-off error.

In this section, we shall study the influence of these local errors on the accumulated round-off error $r_n = \tilde{y}_n - y_n$, without making any speculations about the nature of the local round-off error. Appropriate assumptions on the local round-off errors are discussed in the next section.

We subtruct from (3) the corresponding equation (2), using

$$f(x_m, \tilde{y}_m) - f(x_m, y_m) = g_m r_m + \Theta_m K_m \varepsilon \quad (|\Theta_m| < 1)$$

where $g_m = g(x_m)$, and $g(x) = f_y(x, y)$, then we obtain

(4) $r_{n+k} - r_{n+k-1} = h(\beta_k g_{n+k} r_{n+k} + \cdots\cdots + \beta_0 g_n r_n) + \varepsilon_{n+k} + \Theta_{n+k} Kh\varepsilon$

We shall write

$$r_n = r_n^{(1)} + r_n^{(2)}$$

where $\{r_n^{(1)}\}$is the solution of (4) for $\Theta_{n+k} = 0$, and $\{r_n^{(2)}\}$ is the solution for $\varepsilon_{n+k} = 0$. We shall call $r_n^{(1)}$ the primary error and $r_n^{(2)}$ the secondary error.

Since the secondary error, which is due to the nonlinearity of the given differntial equation, is $0(\varepsilon)$, whereas the primary error must be expected to be $0(\varepsilon h^{-1})$, we may assume that for $h\to 0$ the behavior r_n is governed by the behavior $r_n^{(1)}$, and the following considerations will be directed exclusively toward the primary error $r_n^{(1)}$.

The round-off errors depend on the number of digits carried, on the number system employed by the machine, on the location of the decimal point (fixed or floating operations), on the precision of the subroutines which may be used in the evaluation of $f(x, y)$, and on other factors. Thus the realistic statements about the size of the round-off error that must actually be expected in numerical integrations can be formed by the statistical methods. We accordingly introduce the hypothesis that the local round-off errors may be treated as random variables. According to Henrici [1], the following result holds:

If the local round-off errors are independent variables whose mean and variance satisfy

(5) $E(\varepsilon_m) = \mu p(x_m), \quad Var(\varepsilon_m) = \sigma^2 q(x_m)$

where $p(x)$ and $q(x)$ are piecewise smooth functions,

then the accumulated round-off error is a random variable, such that

(6) $E(r_n) = \frac{\mu}{h}\{m(x_n) + 0(h)\}$

(7) $Var(r_n) = \frac{\sigma^2}{h}\{v(x_n) + 0(h)\}$

where the function $m(x)$ is defined by

(8) $m'(x) = g(x)m(x) + p(x), \quad m(x) = 0$

and the function $v(x)$ is defined by

(9) $v'(x) = 2g(x)v(x) + q(x), \quad v(x_0) = 0.$

2 Local round-off error

Let x be any real number, then the floating representation of x to base β can be written in the form

$$x = f\beta^e$$

where

$$e = [\log_\beta |x|] + 1, \quad f = \beta^{-e}x.$$

If a floating number is actually represented in a computing machine, the real number f, which be called the mantissa of x, can in general be represented only approximately. That is, let $fl(x)$ denote the floating representation of x, then it takes the form

$$x^* = fl(x) = x(1+\varepsilon)$$

where $-\beta^{1-t} \leqq \varepsilon \leqq 0$ for chopping

and where t is the number of digits in the mantissa.

For the Adams method the polynomial

$$\rho(z) = z^k - z^{k-1}$$

has the only one essential root $z = 1$. Thus the Adams method is strongly stable, and the behavior r_n in the Adams method is likely in a one-step method. Moreover, since $\sum_{j=0}^{k} \beta_j = 1$, $\sum_{j=0}^{k} \beta_j f_{n+j}$ may be considered as the weighted mean of the quantities f_n, f_{n+1}, ……, f_{n+k}. Thus for the sake of simplicity for notation in the following discussion we shall write

$$\Phi(x_n, y_n) = \beta_k f(x_{n+k}, y_{n+k}) + \beta_{k-1} f(x_{n+k-1}, y_{n+k-1}) + \cdots\cdots + \beta_0 f(x_n, y_n)$$

which corresponds to the increment function in a one-step method.

Then the equations (2) and (3) are rewritten in the following form:

(2′) $y_{n+k} = y_{n+k-1} + h\Phi(x_n, y_n) \quad n = 0, 1, 2, \cdots\cdots$

and

(3′) $\tilde{y}_{n+k} = \tilde{y}_{n+k-1} + h\Phi(x_n, \tilde{y}_n) + \varepsilon_{n+k} \quad n = 0, 1, 2, \cdots\cdots$

Whereas, according to Wilkinson [2], $\tilde{y}_{n+k}$ are connected by a relation of the form

$$\begin{aligned}\tilde{y}_{n+k} &= f1\{\tilde{y}_{n+k-1} + f1[h\Phi(x_n, \tilde{y}_n)(1+\rho_{n+k})]\} \\ &= \{\tilde{y}_{n+k-1} + h\Phi(x_n, \tilde{y}_n)(1+\rho_{n+k})(1+\pi_{n+k})\}(1+\alpha_{n+k}).\end{aligned}$$

Expanding the right hand side of the above equation and comparing with (3′), we obtain

$$\varepsilon_{n+k} \approx \tilde{y}_{n+k-1}\alpha_{n+k} + h\Phi(x_n, \tilde{y}_n)(\rho_{n+k} + \pi_{n+k} + \alpha_{n+k})$$

Here the quantities ρ_{n+k}, π_{n+k} and α_{n+k} are considered as random variables uniformly distributed in the range $[-\beta^{1-t}, 0]$

We shall call α_{n+k}, π_{n+k} and ρ_{n+k} by adduced error, produced error and inherent error respectively. If h is sufficiently small we may assume that

$$\varepsilon_{n+k} \approx \tilde{y}_{n+k-1}\alpha_{n+k},$$

and for p(x) anb q(x) in (5) we may take y(x) and $(y(x))^2$ respectively.

3 Numerical examples

We consider the numerical solutions of the following differential equations

(i) $y' = y, \quad y(0) = 1$

(ii) $y' = -y, \quad y(0) = 1$

(iii) $y' = xy, \quad y(0) = 1$

(iv) $y' = -xy, \quad y(0) = 1$

(v) $y' = \frac{2}{x}y, \quad y(1) = 1$

(vi) $y' = -\frac{y}{x}, \quad y(1) = 1$

(vii) $y' = y(1-y), \quad y(0) = 0.5$

These numerical experiments were performed with a HITAC 8250 computer, using floating point arithmetic. For this computer, $\beta = 16$, $t = 6$, and hence we assume that

$\mu = -2^{-21}$ and $\sigma^2 = \frac{1}{12} 2^{-40}$

The above equations were solved by the third order Adams method with $h = 2^{-4}$. The "exact" values y_n were obtained by numerical computation using double precision, and then "actual" round-off errors r_n were evaluated with

$$r_n = \tilde{y}_n - y_n.$$

The "experimental" round-off errors $\bar{r}_n$ were calculated from the equation (4), and the "predicted" round-off errors $E(r_n) \pm \sigma(r_n)$ were calculated from the equations (6) and (7). The values of $m(x_n)$ and $v(x_n)$ were found by numerical integration (8) and (9) respectively.

The numerical results are given in the following tables (unit 10^{-7}).

Table. 1 $y' = y$, $y(0) = 1$; true solution $y = e^x$

x	0.50	1.00	1.50	2.00
r_n	−6	−19	−44	−82
$\bar{r}_n$	−5	−19	−49	−110
$E(r_n)-\sigma$	−8	−24	−57	−124
$E(r_n)+\sigma$	−5	−18	−45	−101

Table. 2 $y' = -y$, $y(0) = 1$; true solution $y = e^{-x}$

0.50	1.00	1.50	2.00
−2	−5	−7	−6
−17	−24	−23	−19
−28	−32	−29	−23
−18	−24	−23	−19

Table. 3 $y' = xy$, $y(0) = 1$; true solution $y = e^{\frac{1}{2}x^2}$

x	0.50	1.00	1.50	2.00
r_n	−6	−17	−43	−120
$\bar{r}_n$	−3	−11	−33	−110
$E(r_n)-\sigma$	−5	−14	−39	−124
$E(r_n)+\sigma$	−3	−11	−31	−101

Table. 4 $y' = xy$, $y(0) = 1$; true solution $y = e^{-\frac{1}{2}x^2}$

0.50	1.00	1.50	2.00
−3	−6	−6	−5
−25	−40	−33	−19

Table. 5 $y' = \frac{2y}{x}$, $y(1) = 1$; true solution $y = x^2$

x	1.50	2.00	2.50	3.00
r_n	−8	−24	−47	−79
$\bar{r}_n$	−8	−29	−70	−136
$E(r_n)-\sigma$	−10	−35	−80	−151
$E(r_n)+\sigma$	−7	−26	−63	−123

Table. 6 $y' = -\frac{y}{x}$, $y(1) = 1$; true solution $y = \frac{1}{x}$

1.50	2.00	2.50	3.00
−3	−6	−8	−11
−19	−33	−41	−47
−31	−44	−51	−56
−20	−33	−40	−46

Table. 7 $y' = y(1-y)$, $y(0) = 0.5$;
true solution $y = \frac{1}{1+e^{-x}}$

x	0.50	1.00	1.50	2.00
r_n	−4	−9	−10	−12
$\overline{r_n}$	−16	−37	−54	−66
$E(r_n)-\sigma$	−25	−47	−64	−75
$E(r_n)+\sigma$	−16	−35	−50	−60

The agreement between the experimental and the predicted values is good, whereas, in the cases (ii), (iv), (vi) and (vii) the agreement between the actual and the experimental (and hence predicted) values is not very good. This disagreement appears to be due to the neglect of produced and inherent errors $h\Phi(x_n, y_n)(\rho_{n+k} + \pi_{n+k} + \alpha_{n+k})$.

References

1) P. Henrici : Discrete variable Methods in Ordinary Differential lEquations. John Wily & Sons (1968).
2) J. H. Wilkinson : Rouding Errors in Algebraic Processes. Her Britannic Majesty's Stationery Office (1963).