Auto-Correlation Coefficient of Accumulated Round-off Errors on Numerical Solutions to Ordinary Differential Equations

SIGEITI MORIGUTI AND TADASHI YOSHIZAWA*

Introduction

P. Henrici[1] introduced a probabilistic theory of round-off errors on numerical solutions to ordinary differential equations and dealt with the expectations and variances of accumulated round-off errors, assuming that the local round-off errors are random variables which are independent of each other and uniformly distributed. He considered that accumulated round-off errors with small differences in the initial conditions constitute a sample space and that accumulated round-off errors on numerical solutions for the two different initial conditions are independent of each other.

However, S. Kamachi[2] discovered a counter-example (Fig. 1) against the assumption of independence of accumulated round-off errors when he solved the equation $y'=y$ by Euler's method for the 2000 different initial conditions

$$y_0=0.1+q\cdot\Delta\cdot u$$

where $q=0, 1, \cdots, 1999$, and

$$\Delta=5$$

working with a computer of digital unit $u=10^{-12}$.

In this paper, a probabilistic theory of auto-correlation coefficients of a sequence of accumulated round-off errors on numerical solutions by Euler's method will be described and some experimental results will be given.

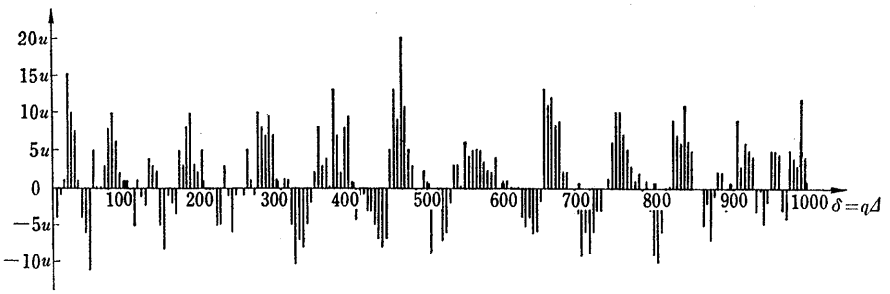

Fig. 1.

This paper first appeared in Japanese in Joho Shori (the Journal of the Information Processing Society of Japan), Vol. 5, No. 2 (1964), pp. 77~82.

* Faculty of Engineering, University of Tokyo.

1. *Euler's method*

In order to solve the initial value problem

$$y'=f(x,\ y),\quad y(a)=\eta \tag{1.1}$$

in Euler's method, the values y_n are calculated successively according to the following formulas:

$$\begin{aligned} &y_0=y(a)=\eta \\ &y_{n+1}=y_n+hf(x_n,\ y_n),\quad n=0,\ 1,\cdots \end{aligned} \tag{1.2}$$

where h is the stepwidth.

If x is any number satisfying $|x|<1$, then x^* will denote a correctly rounded machine representation of x. By "correctly rounded" we mean that

$$|x-x^*|\leqq u/2$$

where u is the basic unit of the machine (in this research OKITAC 5090 was used in fixed point and u is 10^{-12}). It will always be assumed that the step h, the initial point a and the initial value η are integer multiples of u. Under these assumptions, the values y_n actually calculated (values actually calculated are denoted by corresponding symbols with $\tilde{}$) satisfy the following equations:

$$\begin{aligned} &\tilde{y}_0=y_0 \\ &\tilde{y}_{n+1}=\tilde{y}_n+(h\tilde{f}(x_n,\ \tilde{y}_n))^* \end{aligned} \tag{1.3}$$

We replace (1.3) by the relation

$$\tilde{y}_{n+1}=\tilde{y}_n+hf(x_n,\ \tilde{y}_n)+\varepsilon_{n+1} \tag{1.4}$$

The quantity ε_{n+1}, called the local round-off error, is defined by this equation. Thus,

$$\varepsilon_{n+1}=(h\tilde{f}(x_n,\ \tilde{y}_n))^*-hf(x_n,\ \tilde{y}_n) \tag{1.5}$$

ε_{n+1} are assumed to be random variables which are independent of each other and uniformly distributed. The accumulated round-off error at the nth step is defined by the following equation.

$$r_n=\tilde{y}_n-y_n \tag{1.6}$$

Moreover, we shall write

$$g(x)=\left|\frac{\partial f(x,\ y)}{\partial y}\right|_{y=y(x)} \tag{1.7}$$

$$G(x)=\int_a^x g(t)dt \tag{1.8}$$

2. *Correlation coefficient of local round-off errors*

Let $\overset{1}{x}$ and $\overset{2}{x}$ denote two arbitrary real numbers with a definite difference $\xi\cdot u$ and $\overset{1}{x}{}^*$ and $\overset{2}{x}{}^*$ the digital numbers obtained by rounding from $\overset{1}{x}$ and $\overset{2}{x}$ respectively. Then we have

$$\begin{aligned} \overset{1}{x}{}^*&=\overset{1}{x}+\overset{1}{\varepsilon} \\ \overset{2}{x}{}^*&=\overset{2}{x}+\overset{2}{\varepsilon} \end{aligned} \tag{2.1}$$

where $\overset{1}{\varepsilon}$ and $\overset{2}{\varepsilon}$ are the round-off errors. Consequently it follows that

$$\overset{2}{\varepsilon}-\overset{1}{\varepsilon}=j\cdot u-\xi\cdot u \tag{2.2}$$

where j is an integer such that $j\cdot u=\overset{2}{x}{}^*-\overset{1}{x}{}^*$. If we consider $\overset{1}{\varepsilon}$ and $\overset{2}{\varepsilon}$ as random variables which have a degenerated uniform distribution on the pair of lines in Fig. 2 as their joint distribution, we can obtain the correlation coefficient of $\overset{1}{\varepsilon}$ and $\overset{2}{\varepsilon}$

$$\rho_\xi=1-6(\xi-i)(i-1-\xi) \tag{2.3}$$

where i is the smallest integer which is greater than or equal to ξ(Fig. 3).

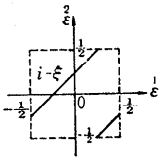

Fig. 2.

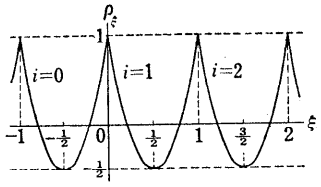

Fig. 3.

Let us use 1 and 2 on top in order to distinguish the two solutions, the two numerical solutions by Euler's method, the two round-off errors at each step, etc. of the ordinary differential equation $y'=f(x, y)$ with the two initial conditions $y(a)=\eta$ and $y(a)=\eta+\delta\cdot u$. We can obtain the following equation in the local round-off errors from (1.5)

$$\begin{aligned}\overset{2}{\varepsilon}_{n+1}-\overset{1}{\varepsilon}_{n+1}&=(h\tilde{f}(x_n,\ \overset{2}{y}_n))^*-hf(x_n,\ \overset{2}{y}_n)-(h\tilde{f}(x_n,\ \overset{1}{y}_n))^*-hf(x_n,\ \overset{1}{y}_n)\\&=j\cdot u-hf(x_n,\ \overset{2}{y}_n)-hf(x_n\ ,\overset{1}{y}_n)\end{aligned} \tag{2.4}$$

where j is an integer.

Corresponding to ξ in (2.2), we may write

$$\xi(x_n)\cdot u=hf(x_n,\ \overset{2}{y}_n)-hf(x_n,\ \overset{1}{y}_n). \tag{2.5}$$

We now determine $\xi(x_n)$. From the mean value theorem

$$hf(x_n,\ \overset{2}{y}_n)-hf(x_n,\ \overset{1}{y}_n)=hf_y(x_n,\ y_n{}^+)(\overset{2}{y}_n-\overset{1}{y}_n) \tag{2.6}$$

where $y_n{}^+$ is a value between $\overset{2}{y}_n$ and $\overset{1}{y}_n$. Then we obtain

$$\xi(x_n)\cdot u\fallingdotseq hg(x_n)(\overset{2}{y}_n-\overset{1}{y}_n) \tag{2.7}$$

since we can approximate $f_y(x_n,\ y_n{}^+)$ by $g(x_n)$ which is defined in (1.7). If we write

$$z_n=\overset{2}{y}_n-\overset{1}{y}_n \tag{2.8}$$

we obtain from (1.5)

$$z_{n+1}=z_n+hf(x_n,\ \overset{2}{y}_n)-hf(x_n,\ \overset{1}{y}_n)+\overset{2}{\varepsilon}_{n+1}-\overset{1}{\varepsilon}_{n+1} \tag{2.9}$$

Since the middle term of (2.9) is equal to $\xi(x_n)u$ we can rewrite (2.9) in the form

$$z_{n+1}=z_n+hg(x_n)z_n+\overset{2}{\varepsilon}_{n+1}-\overset{1}{\varepsilon}_{n+1} \tag{2.10}$$

From the theorem given in Henrici[1] (p. 28) we obtain

$$z_n=z(x_n)+O(h) \tag{2.11}$$

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

$$z'=g(x)z, \quad z_0=\delta\cdot u \tag{2.12}$$

Consequently it follows that

$$z_n=\delta e^{G(x)}\cdot u+O(h)u \tag{2.13}$$

We thus obtain

$$\xi(x_n)=h\delta g(x_n)e^{G(x_n)} \tag{2.14}$$

Theorem 1: If the local round-off errors $\overset{1}{\varepsilon}_{n+1}$ and $\overset{2}{\varepsilon}_{n+1}$ in the numerical solutions of $y'=f(x, y)$ by Euler's method for the two initial conditions with the difference $\delta\cdot u$ are random variables which are uniformly distributed in the interval $[-u/2, u/2]$, then the correlation coefficient $\rho(\overset{1}{\varepsilon}_{n+1}, \overset{2}{\varepsilon}_{n+1})$ is given by

$$\rho(\overset{1}{\varepsilon}_{n+1}, \overset{2}{\varepsilon}_{n+1})=\rho_{\xi(x_n)}=1-6(\xi(x_n)-i)(i-1-\xi(x_n)) \tag{2.15}$$

where $\xi(x_n)$ is given in (2.14) and i is an integer such that $i-1<\xi(x_n)\leqq i$.

3. *Auto-correlation coefficient of accumulated round-off errors*

It is assumed after Henrici[1] that the accumulated round-off errors $\overset{1}{r}_n$ and $\overset{2}{r}_n$ for the initial conditions $\overset{1}{y}(a)=\eta$ and $\overset{2}{y}(a)=\eta+\delta u$, respectively, depend on all local round-off errors $\{\overset{1}{\varepsilon}_i\}$, and $\{\overset{2}{\varepsilon}_i\}$ $(i=1, 2, \cdots\cdots, n)$, in the form

$$\overset{1}{r}_n=\sum_{i=1}^{n} d_{n,\,i}\overset{1}{\varepsilon}_i \qquad \overset{2}{r}_n=\sum_{i=1}^{n} d_{n,\,i}\overset{2}{\varepsilon}_i \tag{3.1}$$

$\overset{1}{\varepsilon}_i$ and $\overset{2}{\varepsilon}_i$ with the same suffix have the correlation given in (2.15). If we assume $\overset{1}{\varepsilon}_i$ and $\overset{2}{\varepsilon}_i$ with different suffixes are independent, then the correlation coefficient of $\overset{1}{r}_n$ and $\overset{2}{r}_n$ is given by

$$\rho(\overset{1}{r}_n, \overset{2}{r}_n)=\frac{\sum_{i=1}^{n} d^2_{n,\,i}\rho(\overset{1}{\varepsilon}_i, \overset{2}{\varepsilon}_i)}{\sum_{i=1}^{n} d^2_{n,\,i}}=\frac{R_n}{V_n} \tag{3.2}$$

where

$$R_n=h\sum_{i=1}^{n} d^2_{n,\,i}\rho(\overset{1}{\varepsilon}_i, \overset{2}{\varepsilon}_i) \tag{3.3}$$

$$V_n=h\sum_{i=1}^{n} d^2_{n,\,i} \tag{3.4}$$

We try to evaluate the sums R_n and V_n by establishing difference equations for them. We use here the following equations (see Henrici[1] p. 38) repeatedly.

$$d_{n+1,\,i}=d_{n,\,i}+hg(x_n)d_{n,\,i} \qquad \begin{pmatrix} n=0, 1, \cdots\cdots \\ i=1, 2, \cdots\cdots, n \end{pmatrix}$$

$$d_{n+1,\ n+1}=1 \qquad (n=0,\ 1,\cdots\cdots) \tag{3.5}$$

from (3.3), (3.5) we obtain the following difference equation:

$$\begin{aligned} R_{n+1}-R_n &= h\left\{\sum_{i=1}^{n+1} d^2_{n+1,\ i}\rho(\overset{1}{\varepsilon_i},\overset{2}{\varepsilon_i})-\sum_{i=1}^{n} d^2_{n,\ i}\rho(\overset{1}{\varepsilon_i},\overset{2}{\varepsilon_i})\right\} \\ &= h\{\rho(\overset{1}{\varepsilon}_{n+1},\overset{2}{\varepsilon}_{n+1})+\sum_{i=1}^{n}(d^2_{n+1,\ i}-d^2_{n,\ i})\rho(\overset{1}{\varepsilon_i},\overset{2}{\varepsilon_i})\} \\ &= h\{2g(x_n)R_n+\rho(\overset{1}{\varepsilon}_{n+1},\overset{2}{\varepsilon}_{n+1})\}+O(h^2) \end{aligned} \tag{3.6}$$

This is a difference equation to which the theorem given in Henrici[1] (p. 28) can be applied. Since $R_0=0$, it follows that

$$R_n=R(x_n)+O(h) \tag{3.7}$$

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

$$\begin{aligned} &R(a)=0 \\ &R'(x)=2g(x)R(x)+\rho_{\xi(x)} \end{aligned} \tag{3.8}$$

Similarly it follows that

$$V_n=V(x_n)+O(h) \tag{3.9}$$

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

$$\begin{aligned} &V(a)=0 \\ &V'(a)=2g(x)\,V(x)+1 \end{aligned} \tag{3.10}$$

Theorem 2: The sequence of the accumulated round-off errors at x_n of the numerical solution of $y'=f(x,y)$ by Euler's method for a sequence of initial conditions which differ little from the initial condition $y(a)=\eta$ have the following auto-correlation coefficient

$$\rho(h\delta)=\frac{R(x_n)}{V(x_n)}+O(h) \tag{3.11}$$

where $R(x)$ and $V(x)$ are the solutions of the ordinary differential equations (3.8) and (3.10) respectively. For fixed x_n, $\rho(h\delta)$ is the function of the step h and the difference $\delta\cdot u$ in the initial conditions. We can rewrite (3.11) in the integral form:

$$\rho(h\delta)=\frac{\int_a^{x_n}\rho_{\xi(t)}\exp\{2[G(x_n)-G_{(t)}]\}\,dt}{\int_a^{x_n}\exp\{2[G(x_n)-G(t)]\}\,dt}=\frac{\int_a^{x_n}\rho_{\xi(t)}\exp[-2G(t)]dt}{\int_a^{x_n}\exp[-2G(t)]dt} \tag{3.12}$$

4. *Examples*

We shall apply the results obtained above to the differential equations $y'=\pm y$ and compare them with experimental sample auto-correlation coefficients.

(1) We determine auto-correlation coefficients of accumulated round-off errors at $x=1$ solving $y'=y$ by Euler's method in the interval [0, 1]. In this case,

$$a=0,\ x_n=1,\ g(x)=1,\ G(x)=\int_0^x g(t)dt=x\,.$$

Consequently

$$\xi(x)=h\delta e^{x}\,.$$

From (3. 12) we obtain

$$\rho(h\delta)=\int_0^1 \rho_{\xi(x)}e_{-2x}dx\Big/\int_0^1 e^{-2x}dx=\frac{2}{1-e^{-2}}\int_0^1 \rho_{\xi(x)}e_{-2x}dx$$

where $\rho_{\xi(x)}=1-6(h\delta e^{x}-i)(i-1-h\delta e^{x})$, $i-1<h\delta e^{x}\leqq i$.

We have a simple algorithm for calculating the theoretical value of $\rho(h\delta)$ (Fig. 4).

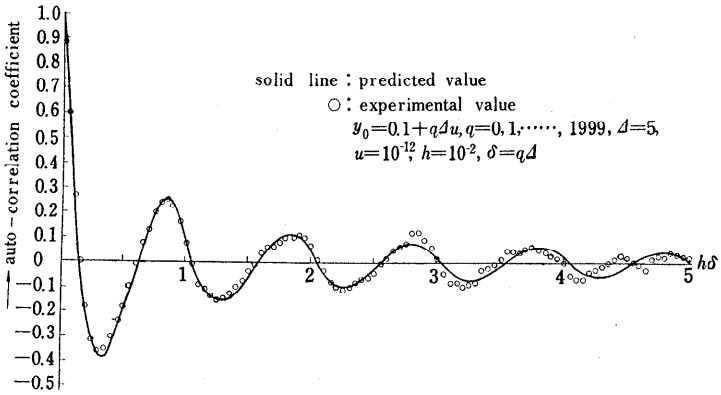

Fig. 4.

The experimental value of $\rho(h\delta)$ was calculated as follows: The equation $y'=y$ was solved by Euler's method for the 2000 different initial conditions

$$y_0=0.1+q\cdot\Delta\cdot u,\qquad q=0,\ 1,\cdots,1999,\qquad \Delta=5\,.$$

The numerical end values y_n were compared with the theoretical values

$$y_n=y_0(1+h)^{1/h}$$

calculated with the formula

$$(1+h)^{1/h}=e\left[1-\frac{1}{2}h+\frac{11}{24}h^2-\frac{21}{48}h^3+\frac{2447}{5760}h^4-\frac{959}{2304}h^5+\cdots.\right].$$

From the sequence of the 2000 round-off errors

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

sample auto-correlation coefficients were calculated. See Fig. 4.

The excellent agreement between the predicted and the experimental values is evident. The slight difference appears to be well within the range of sampling errors.

(2) Similar results were obtained for the differential equation $y'=-y$. We find

$$\xi(x)=-h\delta e^{-x}$$

$$\rho(h\delta)=\frac{2}{e^{2}-1}\int_0^1 \rho_{\xi(x)}e^{2x}dx$$

where $\rho_{\xi(x)}=1-6(-h\delta e^{-x}-i)(i-1+h\delta e^{-x})$, $i-1<-h\delta e^{-x}\leqq i$.

In this case the accumulated round-off errors were calculated for 1000 different initial conditions

$$y_0=0.1+q\cdot\Delta\cdot u, \qquad q=0, 1,\cdots, 999, \qquad \Delta=20.$$

The predicted and the experimental values are shown in Fig. 5.

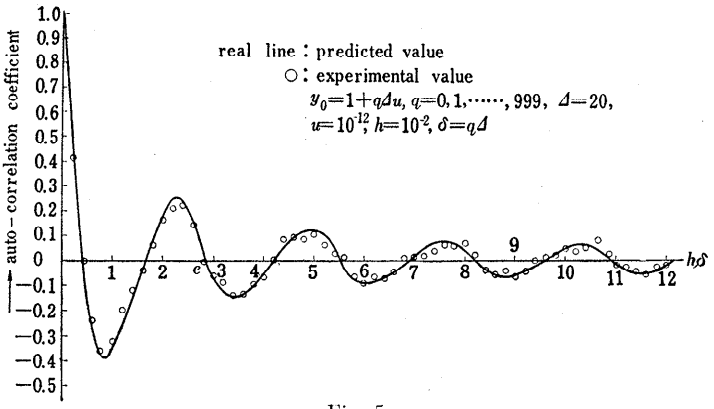

Fig. 5.

5. *Analysis of auto-correlation coefficients by Fourier expansion*

Since ρ_ξ in (2.6) has period 1 as a function of ξ, we can easily expand ρ_ξ into the form:

$$r(\xi)=\sum_{n=1}^{\infty}\frac{6}{n^2\pi^2}\cos(2n\pi\xi) \tag{5.1}$$

From (2.14)

$$\xi(x)=h\delta g(x)e^{G(x)} \tag{5.2}$$

We change the variable x into θ in the following form:

$$\theta=g(x)e^{G(x)} \tag{5.3}$$

and let $x(\theta)$ denote the inverse function. Since

$$d\theta=\{g'(x)e^{G(x)}+(g(x))^2e^{G(x)}\}dx \tag{5.4}$$

it follows that

$$e^{-2G(x)}dx=\varphi(\theta)d\theta \tag{5.5}$$

where

$$\varphi(\theta)=\left[\frac{e^{-3G(x)}}{g'(x)+(g(x))^2}\right]_{x=x(\theta)} \tag{5.6}$$

If we write

$$A=g(a)e^{G(a)}, \qquad B=g(b)e^{G(b)} \tag{5.7}$$

we obtain, for the errors at the end b of the interval $[a, b]$,

$$\rho(h\delta)=\frac{\int_a^b r(h\delta g(x)e^{G(x)})e^{-2G(x)}dx}{\int_a^b e^{-2G(x)}dx}=\frac{\int_A^B r(h\delta\theta)\varphi(\theta)d\theta}{\int_A^B\varphi(\theta)d\theta} \tag{5.8}$$

Since the denominator is a constant, let it be denoted by c. Substituting (5.1) into (5.8), we obtain

$$c\cdot\rho(h\delta)=\int_A^B \sum_{n=1}^{\infty} \frac{6}{n^2\pi^2}\cos(2n\pi h\delta\theta)\varphi(\theta)d\theta \tag{5.9}$$

If we exchange the order of integration and summation, integrate by parts, and take the first order terms in $1/h\delta$, we obtain

$$\begin{aligned} c\cdot\rho(h\delta) &\doteqdot \sum_{n=1}^{\infty} \frac{6}{n^2\pi^2}\left\{\frac{\sin(2n\pi h\delta B)}{2n\pi h\delta}\varphi(B)-\frac{\sin(2n\pi h\delta A)}{2n\pi h\delta}\varphi(A)\right\} \\ &= \frac{\varphi(B)}{4h\delta}\{1-2(h\delta B-n)+[2(h\delta B-n)-1]^3\} \\ &\quad -\frac{\varphi(A)}{4h\delta}\{1-2(h\delta A-m)+[2(h\delta A-m)-1]^3\} \end{aligned} \tag{5.10}$$

where $m=[h\delta A]$, $n=[h\delta B]$ ([ ] is the Gaussian symbol). When $h\delta$ is sufficiently large, the period of $\rho(h\delta)$ becomes approximately equal to

$$\begin{aligned} &1/|B|, \text{ if } |\varphi(B)|\gg|\varphi(A)| \\ &1/|A|, \text{ if } |\varphi(B)|\ll|\varphi(A)| \end{aligned} \tag{5.11}$$

and the amplitude is damped in proportion to $1/h\delta$. If $\varphi(A)=\varphi(B)$, damping is proportional to $(1/h\delta)^2$.

In the case of $y'=y$, $\xi(x)=h\delta e^x$, $\theta=e^x$, $\varphi(\theta)=\theta^{-3}$, $A=1$, $B=e$, $\varphi(A)=1$, $\varphi(B)=e^{-3}$. Consequently, in this case the period is about 1. In the case of $y'=-y$ the period is similarly determined to be about e. The excellent agreement with the curves in Fig. 4 and Fig. 5 is evident.

References

[1] HENRICI, P., *Discrete Variable Methods in Ordinary Differential Equations,* John Wiley & Sons (1962), 407 pp.

[2] KAMACHI, S., On the rounding error in the numerical solution of ordinary differential equations, Graduation Thesis, Faculty of Engineering, University of Tokyo (1962), 35 pp.