Computational uncertainty principle in nonlinear ordinary differential equations

——Ⅱ. Theoretical analysis

LI Jianping (李建平)[1], ZENG Qingcun (曾庆存)[1] & CHOU Jifan (丑纪范)[2]

1. State Key Laboratory of Numerical Modelling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China;
2. Department of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China
Correspondence should be addressed to Li Jianping (email: ljp@lasgsgik.iap.ac.cn)

Received August 30, 1999

Abstract The error propagation for general numerical method in ordinary differential equations ODEs is studied. Three kinds of convergence, theoretical, numerical and actual convergences, are presented. The various components of round-off error occurring in floating-point computation are fully detailed. By introducing a new kind of recurrent inequality, the classical error bounds for linear multistep methods are essentially improved, and joining probabilistic theory the "normal" growth of accumulated round-off error is derived. Moreover, a unified estimate for the total error of general method is given. On the basis of these results, we rationally interpret the various phenomena found in the numerical experiments in part I of this paper and derive two universal relations which are independent of types of ODEs, initial values and numerical schemes and are consistent with the numerical results. Furthermore, we give the explicitly mathematical expression of the computational uncertainty principle and expound the intrinsic relation between two uncertainties which result from the inaccuracies of numerical method and calculating machine.

Keywords: computational uncertainty principle, round-off error, discretization error, universal relation, machine precision, maximally effective computation time (MECT), optimal stepsize (OS), convergence.

In part I of this paper[1], we gave some new phenomena in solving numerically nonlinear ODEs, pointed out the important influences of round-off error due to the finiteness of machine precision on numerical calculations of nonlinear ODEs, and presented the computational uncertainty principle. To explain and prove theoretically the numerical results in ref. [1] and to make the results with general sense and with extensive application value, the error propagation for numerical methods in ODEs must be investigated thoroughly and the influence of round-off errors caused by finite machine precision must be considered. There are many works on the error analyses of numerical methods in ODEs. Some classical results of discretization error can be found in refs. [2—10] and the systematic investigations on round-off error (mainly for fixed point machine) are made by Henrici[2,3]. Examining these classical results thoroughly, however, the error estimates for linear multistep methods but for one-step methods are very coarse and do not apply to the analysis in this paper. Moreover, there is no unified formula of the error estimate for general multistep method, and there are also few works dealing with the round-off error of general multistep method theoretically on floating-point machine. In order to make our results with wide applicability, therefore, we have to improve essentially the classical results and to obtain a unified error estimate for general multistep method, especially, to get the "normal" (or real) accumulated growth of round-off error on floating-point machine. By introducing a new kind of recurrent inequality in this paper, not only the classical error bounds are improved essentially, but also a unified estimate for the total error of general method is given. Specially, the "normal" accumulated growth of round-off error on floating-point machine is derived by using probabilistic theory. Consequently, not only the various phenomena found in the numerical experiments in ref. [1] are rationally explained, but also the explicitly mathematical expression of the computational uncertainty principle is given.

1 Basic description

Consider the following initial-value problem of the first order of m ODEs:

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{y}' = \boldsymbol{f}(t,\boldsymbol{y}), \qquad \boldsymbol{y}(t_0) = \boldsymbol{y}_0, \tag{1}$$

where the vector $\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}, t \in [a, b]$, and $\boldsymbol{f}(t,\boldsymbol{y}) = (f_1(t,\boldsymbol{y}), f_2(t,\boldsymbol{y}), \cdots, f_m(t,\boldsymbol{y}))^{\mathrm{T}}$ is a given continuous vector function. Here the superscript 'T' represents transposition. We always assume that the vector-valued function $\boldsymbol{f}(t,\boldsymbol{y})$ is defined and continuous on the region $S = \{(t,\boldsymbol{y}) \mid a \leqslant t \leqslant b, \boldsymbol{y} \in \mathbb{R}^m\}$ and $\boldsymbol{f}(t,\boldsymbol{y})$ satisfies the Lipschitz condition with respect to $\boldsymbol{y}$, i. e. if there exists a constant L such that for any $t \in [a, b]$ and any two vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$,

$$\| \boldsymbol{f}(t,\boldsymbol{y}_1) - \boldsymbol{f}(t,\boldsymbol{y}_2) \| \leqslant L \| \boldsymbol{y}_1 - \boldsymbol{y}_2 \| . \tag{2}$$

L is called a Lipschitz constant with respect to $\boldsymbol{y}$ for $\boldsymbol{f}(t,\boldsymbol{y})$. Then the initial-value problem (1) has a unique continuously differentiable solution $\boldsymbol{y}(t)$.

The methods and results of initial-value problems and systems of ODEs of the first order, as are well known, are essentially independent of the number m[6]. In the following we often limit ourselves in form to the case of only one ODE of the first order and only one unknown function (i.e. $m = 1$). The results, however, are also valid for systems, provided quantities such as y, $f(t,y)$, Φ, Δ, $e(t;h)$, $r(t;h)$, $E(t;h)$, $T_k(t;y;h)$, $\tau_k(t;y;h)$, $\varepsilon(t;h)$, R and z, etc. are interpreted as vectors. Without loss of generality, we still use norm $\| \cdot \|$ instead of $| \cdot |$ in suitable place. Thus, all the quantities taken by norm $\| \cdot \|$ are regarded as vectors.

A general numerical method can be written in a unified form[7]

$$\begin{cases} \alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = h\Phi(t_n, y_{n+k}, \\ \qquad y_{n+k-1}, \cdots, y_n; h; f), \quad 0 \leqslant n \leqslant N - k; \\ y_j = y(t_0 + jh), \quad 0 \leqslant j \leqslant k - 1, \end{cases} \tag{3}$$

where $\alpha_j (j = 0, 1, \cdots, k)$ are real constants which do not depend on n, $\alpha_k \neq 0$, k is a fixed positive integer, y_0, y_1, $\cdots$ and y_{k-1} are known, and $y_j = y(t_j; h)$ $(j = 0, 1, \cdots, N)$. Eq. (3) is called the general k-step method. This method includes all common numerical methods as special cases. When $k = 1$ we have a one-step method and in this case Φ is called the increment function of the method. If $k > 1$, we have a multistep method. Formula (3) is an explicit method if the function Φ is independent of y_{n+k}; otherwise, it is an implicit method. We obtain different numerical methods for different choices of Φ. For simplicity, the argument f in the function Φ will be omitted. For the sake of convenience, we define the polynomial

$$\rho_k(\xi) = \alpha_k \xi^k + \alpha_{k-1} \xi^{k-1} + \cdots + \alpha_0 \tag{4}$$

which will be called the characteristic polynomials of the k-step method (3). Take $k = 1$ in (3), one obtains a general explicit one-step method

$$y_{n+1} = y_n + h\Phi(t_n, y_n; h). \tag{5}$$

Two particular cases of (5), Taylor series method and Runge-Kutta method, are well known. In (3) let

$$\Phi(t_n, y_{n+k}, y_{n+k-1}, \cdots, y_n; h) = \beta_k f(t_{n+k}, y_{n+k}) + \beta_{k-1} f(t_{n+k-1}, y_{n+k-1}) + \cdots + \beta_0 f(t_n, y_n).$$

Then method (3) becomes

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y_{n+j}). \tag{6}$$

(6) is referred to as a linear multistep method or more precisely a linear k-step method because the function Φ depends linearly on f. (6) is explicit for $\beta_k = 0$ and is implicit for $\beta_k \neq 0$. The polynomials

$$\sigma_k(\xi) = \beta_k \xi^k + \beta_{k-1} \xi^{k-1} + \cdots + \beta_0 \tag{7}$$

are also called the characteristic polynomials of the linear k-step method (6). Two important special cases of linear multistep method are explicit and implicit Adams methods.

The actual error of a numerical method for solving a differential equation or a system of differential equations comes from two basic sources: one source is the method of approximation, that is to say, the numerical method will not yield the exact solution of the given differential equations (even if the calculations are carried out without rounding). The difference

$$e(t;h) = y(t) - y(t;h) \tag{8}$$

where $y(t; h)$ denotes the exact solution produced by the numerical method under the given stepsize h and is also called the theoretical approximation, will be called the global discretization (truncation) error. The error depends on the given initial-value problem, the numerical method used, the stepsize h and the step number n (namely t). The other source of error is the finite accuracy of actual computers which causes the fact that $y(t; h)$ cannot be calculated exactly in practice. Let $\tilde{y}(t; h)$denotes the actually calculated value of $y(t; h)$ and be called the numerical approximation. The difference

$$r(t;h) = y(t;h) - \tilde{y}(t;h) \tag{9}$$

is said to be the (accumulated or global) round-off error. The error depends not only on the given differential equation and the numerical method, but also on the computing machine used, the fixed or floating operations, the number system, details of programming, and especially on the machine precision. We write the total error

$$E(t;h) = y(t) - \tilde{y}(t;h) = e(t;h) + r(t;h). \tag{10}$$

By the triangle inequality we find

$$\| E(t;h) \| \leqslant \| e(t;h) \| + \| r(t;h) \| . \tag{11}$$

Definition 1[5]. The (absolute) local (discretization or truncation) error of the k-step method (3) at $t_{n+k} = a + (n+k)h$ is defined by

$$y(t_{n+k}) - y_{n+k}, \tag{12}$$

where $y(t)$ is the exact solution of eq. (1), and y_{n+k} is the exactly numerical solution obtained from (3) by using the exact k starting values $y_{n+j} = y(t_{n+j}; h)(j = 0, 1, \cdots, k-1)$.

A practical definition of local error is given as follows.

Definition 2. Let $y(t)$ be the exact solution of eq. (1). Then the quantities

$$T_k(t,y;h) = \sum_{j=0}^{k} a_j y(t+jh) - h\Phi_k(t,y;h) \tag{13}$$

where $\Phi_k(t,y;h) = \Phi(t, y(t+kh), \cdots, y(t); h)$, and

$$\tau_k(t,y;h) = \frac{1}{h} T_k(t,y;h) \tag{14}$$

will be called the (absolute) local discretization (or truncation error) error and relative local discretization (or truncation) error of the k-step method (3) (relative with respect to the stepsize h) at the point $(t+kh, y)$, respectively.

Remark 1. For explicit one-step methods, we write $\Phi(t,y;h) = \Phi_1(t,y;h)$.

Lemma 1. Consider the differential equation (1), let $y(t)$ be its exact solution, and let Φ be a continuously differentiable function. For the local error (3) one has

$$y(t_{n+k}) - y_{n+k} = \left(a_k I - h\frac{\partial}{\partial y}\Phi(t_n, \bar{y}_{n+k}, y(t_{n+k-1}), \cdots, y(t_n); h)\right)^{-1} T_k(t,y;h), \tag{15}$$

where I is an unit matrix, $\bar{y}_{n+k}$ is a value between $y(t_{n+k})$ and y_{n+k} if Φ is a scalar function. In the case of a vector-valued function Φ, $\frac{\partial}{\partial y}\Phi(t_n, \bar{y}_{n+k}, y(t_{n+k-1}), \cdots, y(t_n); h)$ is the Jacobian matrix, whose rows are evaluated at possibly different values lying on the segment joining $y(t_{n+k})$ and y_{n+k}.

Lemma 1 shows that $T_k(t,y;h)$ is essentially equal to the local error. Definitions 2 and 1 are therefore equivalent.

Definition 3. The method given by (3) is called theoretically convergent if, for arbitrarily fixed $t \in [a,b], t = a + nh = t_n$,

$$\lim_{h\to 0} e(t;h) = 0, \text{ or } \lim_{h\to 0} \| y(t) - y(t;h) \| = 0. \tag{16}$$

The theoretical convergence can only ensure that for $h \to 0$ the theoretical approximation $y(t;h)$ will approximate arbitrarily well to the exact solution $y(t)$, but cannot guarantee that the numerical approximation $\tilde{y}(t;h)$ converges to $y(t)$ as $h \to 0$. We therefore give the actual convergence and numerical convergence.

Definition 4. The method given by (3) is called actually (or really) convergent if, for arbitrarily fixed $t \in [a,b], t = a + nh = t_n$,

$$\lim_{h\to 0} E(t;h) = 0, \text{ or } \lim_{h\to 0} \| y(t) - \tilde{y}(t;h) \| = 0. \tag{17}$$

If

$$\lim_{h\to 0} r(t;h) = 0, \text{ or } \lim_{h\to 0} \| y(t;h) - \tilde{y}(t;h) \| = 0, \tag{18}$$

the method given by (3) is called numerically convergent.

Obviously, the theoretical convergence (16) does not guarantee the actual convergence (17), and *vice versa*. A numerical method is actually convergent only if it is not only theoretically convergent and but also numerically convergent. Conversely, the actual convergence implies neither the theoretical convergence nor the numerical convergence.

In the investigation of the accumulated round-off error we need to introduce the following concept of (absolute) local round-off error.

Definition 5. The (absolute) local round-off error of the k-step method (3) is defined by

$$\tilde{y}(t) = y(t) + \varepsilon(t;h), \qquad t = t_0 + jh, \qquad j = 0, 1, \cdots, k-1,$$

$$\sum_{j=0}^{k} a_j \tilde{y}(t+jh;h) = h\Phi(t, \tilde{y}(t+kh;h),$$

$$\cdots,\tilde{y}(t;h);h) + \varepsilon(t;h), \qquad n = 0,1,\cdots, \tag{19}$$

where $\tilde{y}(t;h)$ is the numerical approximation of the theoretical approximation $y(t;h)$ of the exact solution $y(t:h)$. And the quantity

$$\delta(t;h) = \frac{\varepsilon(t;h)}{h} \tag{20}$$

is called the relative local round-off error of the method (with respect to the stepsize h).

2 Improved *a priori* bounds of discretization error

2.1 Linear multistep method

Before our discussions, we first list some classical *a priori* bounds of discretization error for linear multistep methods. For this purpose, we introduce two lemmas form Henrici[2,3].

Lemma 2[2,3]**.** Let the characteristic polynomial $\rho_k(\xi)$ satisfy the root condition, and let the coefficients $\gamma_j (j = 0,1,\cdots)$ be defined by

$$\frac{1}{\alpha_k + \alpha_{k-1}\xi + \cdots + \alpha_0\xi^k} = \gamma_0 + \gamma_1\xi + \gamma_2\xi^2 + \cdots, \tag{21}$$

then

$$\Gamma = \sup_{j=0,1,\cdots} |\gamma_j| < \infty.$$

Lemma 3[2,3]**.** Consider the non-homogeneous linear difference equation

$$\begin{gathered}\alpha_k z_{m+k} + \alpha_{k-1}z_{m+k-1} + \cdots + \alpha_0 z_m \\ h(\beta_{k,m}z_{m+k} + \beta_{k-1,m}z_{m+k-1} + \cdots + \beta_{0,m}z_m) + \lambda_m,\end{gathered} \tag{22}$$

let the characteristic polynomial $\rho_k(\xi)$ satisfy the root condition, and let

$$\sum_{j=0}^{k} |\beta_{j,n}| \leqslant B^*, \qquad |\beta_{k,n}| \leqslant \beta, \qquad \|\lambda_n\| \leqslant \Lambda, \quad N = 0,1,\cdots,N, \tag{23}$$

where B^*, β, Λ are constants, and let $0 \leqslant h < |\alpha_k| \beta^{-1}$, then every solution of (22) for which

$$\|z_j\| \leqslant z_{(0)}, \quad j = 0,1,\cdots,k-1$$

satisfies

$$\|z_n\| \leqslant \Gamma^*(Akz_{(0)} + n\Lambda)e^{nhL^*}, \qquad n = 0,\ 1,\ \cdots,\ N, \tag{24}$$

where

$$L^* = \Gamma^* B^*, \quad \Gamma^* = \frac{\Gamma}{1 - h\beta |\alpha_k|^{-1}}, \quad A = |\alpha_k| + |\alpha_{k-1}| + \cdots + |\alpha_0|. \tag{25}$$

Using the above lemmas, for the linear k-step method (6), one has the classical result:

Theorem 1[2,3]**.** Let the function $f(t,y)$ satisfy the Lipschitz condition, and let the relative discretization error be

$$\|\tau_k(t,y;h)\| \leqslant \tau(h), \qquad t \in [a,b],\ h \leqslant h_0, \tag{26}$$

where $\tau(h)$ depends only on h and some constants. If (6) satisfies the condition of consistency and the root condition, then for $h|\alpha_k^{-1}\beta_k|L < 1$ and $t \in [t_0, b]$, the global discretization error of the linear k-step method (6) is

$$\|e(t;h)\| \leqslant \Gamma^*[Ake_{(0)} + (t - t_0)\tau(h)]e^{L\Gamma^* B(t-t_0)}, \tag{27}$$

where $e_{(0)}$ is the maximum starting-value error defined by $e_{(0)} = \max\limits_{0\leqslant j\leqslant k-1} \|e_j\|$, and

$$A = \sum_{j=0}^{k} | \alpha_j |, \; B = \sum_{j=0}^{k} | \beta_j |, \; \Gamma^* = \frac{\Gamma}{1 - h \,|\, \alpha_k^{-1}\beta_k \,|\, L}, \; t - t_0 = nh, \quad a \leqslant t_0.$$

Bound (27) is very coarse because the terms $(t - t_0)\tau(h)$ depends on the variable t, and it cannot be used to discuss the quesions in this paper. Hence, we have to improve it. As in the bound of discretization error for one-step methods, we hope that the above term for linear multistep methods is independent of the variable t. Moreover, we expect that the error bounds for the general k-step method (3) (not limited to the case of linear multistep methods) can be given. To this end, we first give a key lemma——a new kind of recurrent inequality as follows.

Lemma 4[1]**.** If the numbers ξ_n satisfy inequalities of the form

$$| \xi_n | \leqslant A \sum_{j=0}^{n-1} | \xi_j | + mB + C\xi_{(0)}, \quad n = k, k + 1, \cdots, \tag{28}$$

where A, B and C are certain nonnegative constants independent of n, k is a natural number $m = n - K$, $K \leqslant k$ an integer and $\xi_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} | \xi_j |$, and then

$$| \xi_n | \leqslant (1 + A)^n N_1(C)\xi_{(0)} + \begin{cases} \dfrac{B}{A}[(1 + A)^m - 1] & A \neq 0; \\ mB & a = 0 \end{cases} \tag{29}$$

holds for $n = N_0(K), N_0(K) + 1, \cdots$. Here the function $N_a(x)$ is defined by

$$N_a(x) = \begin{cases} a, & \text{if} \quad x \leqslant a; \\ x, & \text{if} \quad x > a. \end{cases} \tag{30}$$

Proof. (30) implies

$$N_a(x) \geqslant a, \tag{31}$$

$$N_a(x) \geqslant x. \tag{32}$$

It is therefore clear that for $A = 0$, (29) is true.

For $A > 0$, we now prove the validity of (29) by induction. If $A > 0$, for $j = 0, 1, \cdots, k - 1$, since $| \xi_j | \leqslant \xi_{(0)}$, $n = N_0(K)$, $K \leqslant k$, then $m = N_0(K) - K$, $N_0(K) \leqslant k$. From (31) and (32) one gets

$$\begin{aligned} | \xi_{N_0(K)} | &\leqslant A \sum_{j=0}^{N_0(K)-1} | \xi_j | + mB + C\xi_{(0)} \\ &\leqslant (1 + AN_0(K))N_1(C)\xi_{(0)} + (N_0(K) - K)B. \end{aligned}$$

By use of the two following facts:

$$(1 + x)^k \geqslant 1 + kx, \tag{33}$$

where the real number $x \geqslant -1$ and k is a nonnegative number, and

$$k \leqslant \frac{(1 + x)^k - 1}{x} \tag{34}$$

[1] More generally, one has: If the numbers ξ_n satisfy inequalities of the form

$$| \xi_n | \leqslant A \sum_{j=0}^{n-1} | \xi_j | + (n - k + 1)B + C\xi_{(0)}, \quad n = k, k + 1, \cdots$$

where $A \geqslant -1$, $B \geqslant 0$ and $C \geqslant 0$ are certain constants independent of n, k is natural number and $\xi_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} | \xi_j |$, then

$$| \xi_n | \leqslant (1 + A)^n N_1(C)\xi_{(0)} + \begin{cases} \dfrac{B}{A}[(1 + A)^{n-k+1} - 1], & A \neq 0; \\ (n - k + 1)B, & A = 0 \end{cases}$$

holds for $n = k, k + 1, \cdots$, where $N_1(C)$ is given by (30).

holds for arbitrary $x > 0$ and nonnegative number k, we have

$$| \xi_{N_0(K)} | \leqslant (1 + A)^{N_0(K)} N_1(C) \xi_{(0)} + \frac{B}{A}[(1 + A)^m - 1].$$

Assume now (29) holds true for n $(n \geqslant N_0(K))$. Sustituting it in the right side of (28), and using (31) and (34), one has

$$\begin{aligned} | \xi_{n+1} | &\leqslant A \sum_{j=0}^{N_0(K)-1} | \xi_n | + A\left\{ N_1(C)\xi_{(0)} \sum_{j=N_0(K)}^{n} (1 + A)^j + \frac{B}{A} \sum_{j=N_0(K)}^{n} [(1 + A)^{j-K} - 1] \right\} \\ &\quad + (m + 1)B + N_1(C)\xi_{(0)} \leqslant N_1(C)\xi_{(0)}\Big[(1 + AN_0(K)) + A \sum_{j=N_0(K)}^{n} (1 + A)^j\Big] \\ &\quad + B\Big[(N_0(K) - K) + \sum_{j=N_0(K)-K}^{m} (1 + A)^j\Big] \leqslant N_1(C)\xi_{(0)}\Big[(1 + A)^{N_0(K)} \\ &\quad + A \sum_{j=N_0(K)}^{n} (1 + A)^j\Big] + \frac{B}{A}\Big[(1 + A)^{N_0(K)-K} + A \sum_{j=N_0(K)-K}^{m} (1 + A)^j - 1\Big]. \end{aligned}$$

By the identical relation

$$(1 + x)^k + x \sum_{j=k}^{n} (1 + x)^j = (1 + x)^{n+1}, \tag{35}$$

where x is an arbitrary real number and k is a nonnegative integer, one gets

$$| \xi_{n+1} | \leqslant (1 + A)^{n+1} N_1(C) \xi_{(0)} + \frac{B}{A}[(1 + A)^{m+1} - 1],$$

and (29) is thus established for $n + 1$. The statement of the lemma follows by induction.

Making use of the inequality $1 + x \leqslant e^x$, where x is arbitrary, we write (29) in the forms

$$| \xi_n | \leqslant N_1(C) \xi_{(0)} e^{nA} + \begin{cases} \frac{B}{A}(e^{mA} - 1), & A \neq 0; \\ mB & A = 0, \end{cases} \tag{36}$$

where $A \geqslant 0$ and $B \geqslant 0$.

Lemma 5. Under the conditions of Lemma 3, every solution of (22) satisfies

$$\| z_n \| \leqslant N_1(\eta) z_{(0)} e^{nhL^*} + \frac{\Lambda}{hB^*}(e^{nhL^*} - 1), \quad n = 0, 1, \cdots, N, \tag{37}$$

where $\eta = A\Gamma^* k$, $A = |\alpha_{k-1}| + \cdots + |\alpha_0|$, and other constants are as in Lemma 3.

Proof. To begin with, setting$\gamma_l = 0$ for negative integer l, and from (21) one can obtain

$$\alpha_k \gamma_l + \alpha_{k-1} \gamma_{l-1} + \cdots + \alpha_0 \gamma_{l-k} = \begin{cases} 1, & l = 0, \\ 0, & l \geqslant 0. \end{cases} \tag{38}$$

For a fixed value n and $l = 0, 1, \cdots, n - k$, multiply eq. (22) corresponding to $m = n - k - l$ by γ_l defined by (21) and add them up. On the left side,

$$\begin{aligned} &(\alpha_k z_n + \alpha_{k-1} z_{n-1} + \cdots + \alpha_0 z_{n-k})\gamma_0 + (\alpha_k z_{n-1} + \alpha_{k-1} z_{n-2} + \cdots + \alpha_0 z_{n-k-1})\gamma_1 + \cdots \\ &\quad + (\alpha_k z_{n-j} + \alpha_{k-1} z_{n-1-j} + \cdots + \alpha_0 z_{n-k-j})\gamma_{n-k} + \cdots + (\alpha_k z_k + \alpha_{k-1} z_{k-1} + \cdots + \alpha_0 z_0)\gamma_{n-k} \\ &= \alpha_k \gamma_0 z_n + (\alpha_k \gamma_1 + \alpha_{k-1} \gamma_0) z_{n-1} + \cdots + (\alpha_k \gamma_{n-k} + \alpha_{k-1} \gamma_{n-k-1} + \cdots + \alpha_0 \gamma_{n-2k}) z_k \\ &\quad + (\alpha_{k-1} \gamma_{n-k} + \alpha_{k-2} \gamma_{n-k-1} + \cdots + \alpha_0 \gamma_{n-2k+1}) z_{k-1} + \cdots + \alpha_0 \gamma_{n-k} z_0 \\ &= z_n + (\alpha_{k-1} \gamma_{n-k} + \alpha_{k-2} \gamma_{n-k-1} + \cdots + \alpha_0 \gamma_{n-2k+1}) z_{k-1} + \cdots + \alpha_0 \gamma_{n-k} z_0, \end{aligned}$$

On the right side, we have

$$\begin{aligned} &h[\beta_{k,n-k} \gamma_0 z_n + (\beta_{k-1,n-k} \gamma_0 + \beta_{k,n-k-1} \gamma_1) z_{n-1} + \cdots + (\beta_{0,n-k} \gamma_0 + \cdots + \beta_{k,n-2k} \gamma_k) z_{n-k} \\ &\quad + \cdots + \beta_{0,0} \gamma_{n-k} z_0] + \lambda_{n-k} \gamma_0 + \lambda_{n-k-1} \gamma_1 + \cdots + \lambda_0 \gamma_{n-k}. \end{aligned}$$

Taking norms, and using (23) and (25), we get

$$\| z_n \| \leqslant h\beta \mid \alpha_k \mid^{-1} \| z_n \| + h\Gamma B^* \sum_{j=0}^{n-1} \| z_j \| + (n-k+1)\Gamma\Lambda + A\Gamma k z_{(0)}.$$

Solving $\| z_n \|$, we get

$$\| z_n \| \leqslant hL^* \sum_{j=0}^{n-1} \| z_j \| + n\Gamma^* \Lambda + A\Gamma^* k z_{(0)}.$$

From Lemma 4, it immediately follows that

$$\| z_n \| \leqslant N_1(\eta) z_{(0)} e^{nhL^*} + \frac{\Lambda}{hB^*}(e^{nhL^*} - 1), \quad n = 0, 1, \cdots, N.$$

The lemma is proved.

Lemma 5 indicates that bound (24) has been essentially improved, because the term $\Lambda/(hB^*)$ in (37) rather than the term $n\Lambda$ in estimate (24) is independent of the variable $t_n = t_0 + nh$. Bound (37) can still be improved a little further. According to the proof of Lemma 5, we find

$$z_n = \sum_{j=0}^{k-1} A_{j,n} z_j + h \sum_{j=0}^{n} B_{j,n} z_j + \sum_{j=0}^{n-k} \lambda_j \gamma_{n-k-j}, \tag{39}$$

where

$$A_{J,n} = \sum_{j=0}^{J} \alpha_{J-j}\gamma_{n-k-j}, \qquad J = 0, 1, \cdots, k-1, \quad k \leqslant n \leqslant N,$$

$$B_{J,n} = \begin{cases} \sum_{j=0}^{J} \beta_{k-J+j,\, n-k-j}\gamma_j, & J = 0, 1, \cdots, k-1; \\ \sum_{j=0}^{k} \beta_{j,n-J-j}\gamma_{J-k+j}, & J = k+1, k+2, \cdots, n-k; \quad k \leqslant n \leqslant N \\ \sum_{j=0}^{n-J} \beta_{j,n-J+j}\gamma_{J-k-j}, & J = n-k+1, n-k+2, \cdots, n. \end{cases} \tag{40}$$

And then putting

$$a = \max_{\substack{0 \leqslant J \leqslant k-1 \\ k \leqslant n \leqslant N}} \mid A_{J,n} \mid, \qquad b = \max_{\substack{0 \leqslant J \leqslant N \\ k \leqslant n \leqslant N}} \mid B_{J,n} \mid, \qquad \beta = \max_{0 \leqslant n \leqslant N} \mid \beta_{k,n} \mid, \tag{41}$$

one obtains

$$\| z_n \| \leqslant h\beta \mid \alpha_k \mid^{-1} \| z_n \| + hb \sum_{j=0}^{n} z_j + (n-k+1)\Lambda\Gamma + akz_{(0)},$$

where $z_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} \| z_j \|$. From Lemma 4, we have

Lemma 6. Let the polynomial $\rho_k(\xi)$ satisfy the root condition, and let $0 \leqslant h < \beta^{-1} \mid \alpha_k \mid$, then every solution of (22) satisfies

$$\| z_n \| \leqslant N_1(\eta) z_{(0)} \mathrm{e}^{nh\hat{L}} + \frac{\Lambda\Gamma}{hb}(e^{nhb/c} - 1), \quad n = 0, 1, \cdots, N, \tag{42}$$

where $\eta = ak/c$, $c = 1 - h\beta \mid \alpha_k \mid^{-1}$, a, b and β are given by (41).

Furthermore, setting

$$d = \max_{0 \leqslant n \leqslant N} (\alpha_k^{-1}\beta_{k,n}), \tag{43}$$

one can prove

Lemma 7. Let the polynomial $\rho_k(\xi)$ satisfy the root condition, and let $hd < 1$ and $h \geqslant 0$, then every solution of (22) satisfies

$$\| z_n \| \leqslant N_1(\eta) z_{(0)} \mathrm{e}^{nh\hat{L}} + \frac{\Lambda \Gamma}{hb}(\mathrm{e}^{nhb/c^*} - 1) \quad n = 0, 1, \cdots, N, \tag{44}$$

where $\eta = ak/c^*$, $c^* = |1 - dh|$, a and b are given by (41), and d is defined by (43).

On the basis of the above lemmas, we can now make essential improvement for the estimate (27) of linear multistep methods.

Theorem 2. Under the conditions of Theorem 1, if $h|\alpha_k^{-1}\beta_k|L < 1$ and $t \in [t_0, b]$, the global discretization error of linear k-step method (6) satisfies

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{LB\Gamma^*(t-t_0)} + \tau(h) \frac{\mathrm{e}^{LB\Gamma^*(t-t_0)} - 1}{BL} \tag{45}$$

for $h \leqslant h_0$, $t - t_0 = nh$, $n = 0, 1, \cdots$, where $e_0 = \max\limits_{0 \leqslant j \leqslant k-1} \| e_j \|$, $\eta = A\Gamma^* k$, $A = |\alpha_{k-1}| + \cdots + |\alpha_0|$ and other constants are as in Theorem 4.

Proof. We subtract linear k-step method (6) from the corresponding relation

$$\sum_{j=0}^{k} \alpha_j y(t_{n+j}) = h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y(t_{n+j})) + h\tau_k(t_n, y(t_n); h)$$

satisfied by the exact values $y(t_{n+j})$. Writing $e_j = y(t_j) - y_j$, $j = 0, 1, \cdots$, and putting

$$f(t_j, y(t_j)) - f(t_j, y_j) = L_j e_j,$$

one obtains

$$\sum_{j=0}^{k} \alpha_j e_{n+j} = h \sum_{j=0}^{k} \beta_j L_j e_{n+j} + h\tau_k(t_n, y(t_n); h). \tag{46}$$

In view of the Lipschitz condition, $|L_j| \leqslant L (j = 0, 1, \cdots)$. Applying Lemma 5 to (46) with $z_j = e_j$, $z_{(0)} = e_{(0)}$, $\Lambda = h\tau(h)$, $B^* = BL$ and $L^* = B^* \Gamma^*$, we have

$$\| e_n \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{nhLB\Gamma^*} + \tau(h) \frac{\mathrm{e}^{nhLB\Gamma^*} - 1}{BL},$$

where $\eta = A\Gamma^* k$, $A = |\alpha_{k-1}| + \cdots + |\alpha_0|$. Let $t \in [t_0, b]$, $t_n = t_0 + nh = t$, $n \geqslant 0$ is an integer. From $e(t;h) = e_n$, it follows that

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{LB\Gamma^*(t-t_0)} + \tau(h) \frac{\mathrm{e}^{LB\Gamma^*(t-t_0)} - 1}{BL}$$

for $h \leqslant h_0$, $t - t_0 = nh$, $n = 0, 1 \cdots$. The theorem is proved.

Assume that the exact solution $y(t)$ has a continuous derivative of order $p + 1$ for $t \in [a, b]$, for the linear k-step method (6) of order p, one has

$$\| \tau_k(t, y; h) \| \leqslant Ch^p.$$

Hence, (45) becomes

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{LB\Gamma^*(t-t_0)} + Ch^p \frac{\mathrm{e}^{LB\Gamma^*(t-t_0)} - 1}{BL}. \tag{47}$$

Theorem 2 indicates that our aim to remove the variable t from the terms $(t - t_0)\tau(h)$ in (24) has been accomplished. Estimate (45), however, can be improved a little further by use of Lemma 6.

Theorem 3. Under the conditions of Theorem 1, if $h|\alpha_k^{-1}\beta_k|L < 1$ and $t \in [t_0, b_1]$, the global discretization error of linear k-step method (6) satisfies

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{Lb(t-t_0)/c} + \tau(h) \frac{\Gamma(\mathrm{e}^{Lb(t-t_0)/c} - 1)}{bL}, \tag{48}$$

for $h \leqslant h_0$, $t - t_0 = nh$, $n = 0, 1, \cdots$, where $e_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} \| e_j \|$, $\eta = ak/c$, $c = 1 - h |\alpha_k^{-1}\beta_k| L$, a and b are given by (41).

Now we need to determine the numerical values of the constants C, a, b, c and Γ in (47) or (48) for some special methods. For a number of special methods such as the explicit and implicit Adams methods and the methods based on differentiation, we have

$$C = C_{p+1} M_{p+1}, \tag{49}$$

where C_{p+1} is an error constant (for these methods $\sigma_k(1) = 1$). For all methods based on numerical integration, the characteristic polynomial $\rho_k(\xi) = \xi^k - \xi^{k-q}$, where $1 \leqslant q \leqslant k$; therefore

$$\frac{1}{1 - \xi^q} = 1 + \xi^q + \xi^{2q} + \cdots.$$

It follows that $|\gamma_l| \leqslant 1$, showing that $\Gamma = 1$ for those methods. And for those methods, from (39) and (41) one finds $a = 1$. For the explicit and implicit methods, since $\gamma_l = 1$ for $l = 0, 1, \cdots$, the numerical numbers b in table 1 are readily obtained by (40). For explicit methods we have $\beta_k = 0$, and hence $c = 1$. Since $\alpha_k = 1$ for the implicit methods based on numerical integration, $c = 1 - h|\beta_k|L$.

Table 1 Constants b for the Adams methods

p	1	2	3	4	5	6
Explicit Adams methods b	1	$\frac{3}{2}$	$\frac{23}{12}$	$\frac{55}{24}$	$\frac{1901}{720}$	$\frac{6336}{1440}$
Implicit Adams methods b	1	1	$\frac{13}{12}$	$\frac{28}{24}$	$\frac{897}{720}$	$\frac{1902}{1440}$

Writing

$$d = \sup_{(t, y; h) \in S} (\alpha_k^{-1} \beta L(t, y)), \tag{50}$$

where $L(t, y) = \dfrac{\partial}{\partial y} f(t, y)$, we can prove the following theorem by virtue of Lemma 7.

Theorem 4. Under the conditions of Theorem 4, if $hd < 1$, and $t \in [t_0, b]$, the global discretization error of linear k-step method (6) satisfies

$$\| e(t; h) \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{Lb(t-t_0)/c} + \tau(h) \frac{\Gamma(\mathrm{e}^{Lb(t-t_0)/c} - 1)}{bL}, \tag{51}$$

for $h \leqslant h_0$, $t - t_0 = nh$, $n = 0, 1, \cdots$, where $e_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} \| e_j \|$, $\eta = ak/c^*$, $c^* = |1 - dh|$, a and b are given by (41), and d is defined by (50).

2.2 General k-step method

Now we give the unified form of error estimate for general k-step method (3).

Theorem 5. Consider the initial-value problem (1) having the exact solution $y(t)$ for $t \in [a, b]$. Let the function $\Phi_k(t, y; h)$ satisfy the Lipschitz condition, i.e. there exist constants h_0 and L such that

$$\| \Phi(t, y_k, y_{k-1}, \cdots, y_0; h) - \Phi(t, y_k^*, y_{k-1}^*, \cdots, y_0^*; h) \| \leqslant L \sum_{j=0}^{k} \| y_j - y_j^* \| \tag{52}$$

for all $t \in [a, b]$, $0 < h \leqslant h_0$, y_j, $y_j^* \in R$, and $f \in C^1[a, b]$ and let the relative discretization error be

$$\| \tau_k(t,y;h) \| \leqslant \tau(h), \qquad t \in [a,b], h \leqslant h_0, \tag{53}$$

where $\tau(h)$ depends only on h and some constants. If (3) satisfies the condition of consistency and the root condition, then, for arbitrary $h \leqslant h_0$ and all $t \in [t_0, b]$, $a \leqslant t_0$, $t - t_0 = nh$, the global discretization error of general k-step method (3) will be

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{kL\Gamma^*(t-t_0)} + \tau(h) \frac{\mathrm{e}^{kL\Gamma^*(t-t_0)} - 1}{kL}, \tag{54}$$

where $e_{(0)}$ is the maximum starting-value error defined by $e_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} \| e_j \|$, $\eta = A\Gamma^* k$, $A = |\alpha_{k-1}| + \cdots + |\alpha_0|$, $\Gamma^* = \Gamma/(1 - h|\alpha_k^{-1}| L_k)$, $\Gamma = \sup\limits_{j=0,1,\cdots} |\gamma_j| < \infty$, $\gamma_j (j = 0,1,\cdots)$ are given by (59) below, and L_k is defined by (21).

Proof. From (13) and (14) one finds that $y(t_{n+j}) = y(t_n + jh)$ $(j = 0, 1, \cdots, k-1)$ satisfy

$$\begin{aligned} &\alpha_{n+k} y(t_{n+k}) + \alpha_{n+k-1} y(t_{n+k-1}) + \cdots + \alpha_n y(t_n) \\ &\qquad = h\Phi(t_n, y(t_{n+k}), \cdots, y(t_n); h) + h\tau_{k,n}, \end{aligned} \tag{55}$$

where $\tau_{k,n} = \tau_k(t_n, y(t_n); h)$. Subtracting (3) from (55), writing $e_{n+j} = y(t_{n+j}) - y_{n+j}$, $y_{n+j} = y(t_{n+j}; h)(j = 0,1,\cdots,k-1)$, we have

$$\begin{aligned} &\alpha_{n+k} e_{n+k} + \alpha_{n+k-1} e_{n+k-1} + \cdots + \alpha_n e_n \\ &= h(\Phi(t_n, y(t_{n+k}), \cdots, y(t_n); h) - \Phi(t_n, y_{n+k}, \cdots, y_n; h)) + h\tau_{k,n}. \end{aligned} \tag{56}$$

Setting

$$\begin{aligned} &\Phi(t_n, y(t_{n+k}), \cdots, y(t_n); h) - \Phi(t_n, y_{n+k}, \cdots, y_n; h) \\ &= \Phi(t_n, y_{n+k} + e_{n+k}, \cdots, y_n + e_n; h) - \Phi(t_n, y_{n+k}, \cdots, y_n; h) \\ &= L_{k,n} e_{e+k} + L_{k-1,n} e_{n+k-1} + \cdots + L_{0,n} e_n. \end{aligned} \tag{57}$$

Using (57), (56) can be replaced by

$$\begin{aligned} \alpha_{n+k} e_{n+k} + \alpha_{n+k-1} e_{n+k-1} + \cdots + \alpha_n e_n &= h(L_{k,n} e_{n+k} + L_{k-1,n} e_{n+k-1} + \cdots \\ &+ L_{0,n} e_n) + h\tau_{k,n}. \end{aligned} \tag{58}$$

In view of the Lipschitz condition, $|L_{i,j}| \leqslant L(i = 0,1,\cdots,k, j = 0,1,\cdots)$. And put

$$L_k = \max_{0 \leqslant j \leqslant N} | L_{k,j} |, \quad N = (b - t_0)/h. \tag{59}$$

By Lemma 5 we take $z_j = e_j$, $z_{(0)} = e_{(0)}$, $\Lambda = h\tau(h)$, $B^* = kL$ and $L^* = B^* \Gamma^*$. It follows that

$$\| e_n \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{nhkL\Gamma^*} + \tau(h) \frac{\mathrm{e}^{nhkL\Gamma^*} - 1}{kL}, \quad n = 0, 1, \cdots, N.$$

Let $t \in [t_0, b]$, $t_n = t_0 + nh = t$, $n \geqslant 0$ being an integer. Since $e(t;h) = e_n$, we have

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} \mathrm{e}^{kL\Gamma^*(t-t_0)} + \tau(h) \frac{\mathrm{e}^{kL\Gamma^*(t-t_0)} - 1}{kL}$$

for arbitrary $h \leqslant h_0$ and all $t \in [t_0, b]$, $a \leqslant t_0$. The theorem is proved.

It is easy to see that both the *a priori* bound of global discretization error for one-step methods and (45) of linear multistep methods are the special cases of (54). That is to say, estimate (54) is a unified form of the global discretization error in all k-step methods. For one-step methods, $k = 1$, $A = 1$, $L_1 = 0$, $\Gamma = 1$, $\Gamma^* = 1$ and $N_1(\eta) = 1$. So in this case (54) is identical with the *a priori* bound. If we let $kL = BL$ or bL/Γ (note that the two L here have different senses) and $L_k = |\beta_k| L$ then (54) is the same as (45).

Assuming that method (3) is of the order p, we have

$$\| \tau_k(t,y;h) \| \leqslant Ch^p.$$

Hence

$$\| e(t;h) \| \leqslant N_1(\eta) e_{(0)} e^{kL\Gamma^*(t-t_0)} + Ch^p \frac{e^{kL\Gamma^*(t-t_0)} - 1}{kL}. \tag{60}$$

Different methods of the same order have different global discretization errors; they are distinguished by the constant C.

3 Improved bounds for the accumulated round-off error

To discuss round-off error on floating-point machine, let us first investigate the various components of round-off error. Without loss of generality one assumes that the quantities h, t_0, t, α_j are exact in the process of computation (i.e. no round-off error). Then

$$\tilde{y}(t) = \mathrm{fl}(y(t)), \qquad t = t_0 + jh, \qquad j = 0,1,\cdots,k-1,$$

$$\begin{aligned} &\alpha_k \tilde{y}(t+kh;h) + \mathrm{fl}\Big\{ \sum_{j=0}^{k-1} \mathrm{fl}[\alpha_j \tilde{y}(t+jh;h)] \\ &\quad - \mathrm{fl}[h \cdot \mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))]\Big\} = 0 \\ &\quad n = 0,1,\cdots, \end{aligned} \tag{61}$$

where the notation $\mathrm{fl}(x)$ denotes a floating-point rounded value of x. Therefore the local round-off error reads

$$\begin{aligned} \varepsilon(t;h) = {} & \Big[\sum_{j=0}^{k-1} \alpha_j \tilde{y}(t+jh;h) - h\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h) \Big] \\ & - \mathrm{fl}\Big\{ \sum_{j=0}^{k-1} \mathrm{fl}[\alpha_j \tilde{y}(t+jh;h)] - \mathrm{fl}[h \cdot \mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots, \\ & \tilde{y}(t;h);h))]\Big\}, \end{aligned}$$

and can be regarded as the sum of the following components:

$$\varepsilon(t;h) = \sigma(t;h) + \pi_j(t;h) + \pi(t;h) + \mu(t;h), \tag{62}$$

where

$$\begin{aligned} \sigma(t;h) = {} & \Big\{ \sum_{j=0}^{k-1} \mathrm{fl}[\alpha_j \tilde{y}(t+jh;h)] - \mathrm{fl}[h \cdot \mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))]\Big\} \\ & - \mathrm{fl}\Big\{ \sum_{j=0}^{k-1} \mathrm{fl}[\alpha_j \tilde{y}(t+jh;h)] - \mathrm{fl}[h \cdot \mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h), \\ & \cdots,\tilde{y}(t;h);h))]\Big\}, \end{aligned} \tag{63}$$

$$\pi_j(t;h) = \alpha_j \tilde{y}(t+jh;h) - \mathrm{fl}(\alpha_j \tilde{y}(t+jh;h)), \qquad j = 0,1,\cdots,k-1, \tag{64}$$

$$\begin{aligned} \pi(t;h) = {} & \mathrm{fl}[h \cdot \mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))] \\ & - [h \cdot \mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))], \end{aligned} \tag{65}$$

$$\begin{aligned} \mu(t;h) = {} & h \cdot [\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h)) \\ & - \Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h)]. \end{aligned} \tag{66}$$

The quantity $\sigma(t;h)$ is called the dominant (or addition) rounding error in the floating-point round-off of the addition $\sum_{j=0}^{k-1} \mathrm{fl}[\alpha_j \tilde{y}(t+jh;h)] - \mathrm{fl}[h \cdot \mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);$

$h))]$, $\pi_j(t;h)$ and $\pi(t;h)$ the rounding errors due to the rounding of the product $\alpha_j\tilde{y}(t+jh;h)$ and $h\cdot\mathrm{fl}(\Phi(t,\tilde{y}(t+kh;h),\cdots,\tilde{y}(t;h);h))$, respectively, $\mu(t;h)$ the rounding errors caused by the inaccuracy of the evaluation of the function Φ. Normally, in practice, the stepsize h is so small that $\pi(t;h)$ and $\mu(t;h) << \sigma(t;h)$. One thus has $\varepsilon(t;h)\approx\sigma(t;h)+\pi_j(t;h)$. And for conventional numerical methods such as all one-step methods and all linear multistep methods based on numerical integration, $\pi_j(t;h)=0$, i. e. the local round-off error is determined primarily by the addition rounding error $\sigma(t;h)$. This is why $\sigma(t;h)$ is called the dominant rounding error.

Even if the quantities h, t_0, t, α_j cannot be represented exactly in computer (i.e. they are rounded to the corresponding machine numbers), we have still the same conclusions1).

Under the sole assumption that

$$\| \varepsilon_{n+k} \| \leqslant \varepsilon \qquad (n=0,1,\cdots), \tag{67}$$

where ε is a constant, one has

Theorem 6[2,3]. Let the function $f(t,y)$ be continuous and continuously differential for $t\in[a,b]$ and satisfy the Lipschitz condition, and if the local round-off errors satisfy (67), then the accumulated round-off error is expressed as

$$\| r(t;h) \| \leqslant \Gamma^*\left[Akr_{(0)}+(t-t_0)\frac{\varepsilon}{h}\right]e^{L\Gamma^*B(t-t_0)}, \tag{68}$$

where $r_{(0)}$ is the maximum starting round-off error defined by $r_{(0)}=\max\limits_{0\leqslant j\leqslant k-1}\| r_j \|$, $t\in[t_0,b]$, $a\leqslant t_0$ and $t-t_0=nh$, and other constants are defined as in Theorem 1.

A priori estimate (68) has the same shortcoming as bound (27) for the global discretization error. Following the way to improve the global discretization error estimate, for the influence of round-off error in linear k-step methods, we have

Theorem 7. Under the conditions of Theorem 6, if $h|\alpha_k^{-1}\beta_k|L<1$ and $t\in[t_0,b]$, then the accumulated round-off error of linear k-step method (3) is

$$\| r(t;h) \| \leqslant N_1(\eta)r_{(0)}e^{Lb\Gamma^*(t-t_0)/c}+\frac{\varepsilon}{h}\frac{e^{LB\Gamma^*(t-t_0)}-1}{BL} \tag{69}$$

for $h\leqslant h_0$, $t-t_0=nh$, $n=0,1,\cdots$, where $r_{(0)}=\max\limits_{0\leqslant j\leqslant k-1}\| r_j \|$, and other constants are as in Theorem 2.

Theorem 8. Under the conditions of Theorem 6, if $h|\alpha_k^{-1}\beta_k|L<1$ and $t\in[t_0,b_1]$, then the accumulated round-off error of linear k- step method (3) is

$$\| r(t;h) \| \leqslant N_1(\eta)r_{(0)}e^{Lb(t-t_0)/c}+\frac{\varepsilon}{h}\frac{\Gamma(e^{Lb(t-t_0)/c}-1)}{bL} \tag{69}$$

for $h\leqslant h_0$, $t-t_0=nh$, $n=0,1,\cdots$, where $r_{(0)}=\max\limits_{0\leqslant j\leqslant k-1}\| r_j \|$, the other constants are as in Theorem 3.

In fact, subtracting (61) from (6), writing $r_j=y_j-\tilde{y}_j$, $j=0,1,\cdots$, and setting

$$f(t_j,y_j)-f(t_j,\tilde{y}_j)=L_jr_j$$

with the Lipschitz condition, $|L_j|\leqslant L$, one gets

1) Li Jianping, Computational uncertainty principle in nonlinear ordinary differential equations and two universal relations, Institute of Atmospheric Sciences, Chinese Academy Sciences, Postdoctor Research Summing-up Report, 1999, 174.

$$\sum_{j=0}^{k} \alpha_j r_{n+j} = h \sum_{j=0}^{k} \beta_j L_j r_{n+j} - \varepsilon_{n+k}. \tag{70}$$

An application of Lemma 6 to this relation with $z_j = r_j$, $\hat{\Lambda} = \varepsilon\Gamma$, $b = bL$, $\eta = ak/c$, $c = 1 - h \mid \alpha_k^{-1}\beta_k \mid L$ and $z_{(0)} = r_{(0)}$ yields Theorem 8.

Theorem 9. Under the conditions of Theorem 6, if $hd < 1$ and $t \in [t_0, b]$, then the accumulated round-off error of linear k-step method (37) is

$$\| r(t;h) \| \leqslant N_1(\eta) r_{(0)} \mathrm{e}^{Lb(t-t_0)/c^*} + \frac{\varepsilon}{h} \frac{\Gamma(\mathrm{e}^{Lb(t-t_0)/c^*} - 1)}{bL} \tag{71}$$

for $h \leqslant h_0$, $t - t_0 = nh$, $n = 0, 1, \cdots$, where $r_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} \| r_j \|$, and other constants are as in Theorem 4.

For the general k-step method, one has a unified result as follows.

Theorem 10. Let $\Phi_k(t, y; h)$ satisfy (52). If the local round-off error satisfies the sole assumption

$$\| \varepsilon(t;h) \| \leqslant \varepsilon, \tag{72}$$

then the accumulated round-off error of the general k-step method is

$$\| r(t;h) \| \leqslant N_1(\eta) r_{(0)} \mathrm{e}^{kL\Gamma^*(t-t_0)} + \frac{\varepsilon}{h} \frac{\mathrm{e}^{kL\Gamma^*(t-t_0)} - 1}{kL}, \tag{73}$$

where $r_{(0)}$ is the maximum starting round-off error defined by $r_{(0)} = \max\limits_{0 \leqslant j \leqslant k-1} \| r_j \|$, and other parameters as in Theorem 5.

Clearly, bound (73) is an extension of (69). The essential result contained in (73) is that the accumulated round-off error $r(t;h)$ is of the order of h^{-1}(i.e. it has the same order as the relative local round-off error $\delta = \varepsilon/h$) and that (73) does not depend on the constants C and p which are typical characteristics for the global discretization error $e(t;h)$ of the method.

The bounds for the accumulated round-off error given above are derived under the sole assumption that all round-off errors accumulate at their maximum values, thus although theoretically valuable (e.g. for the results to be derived below), these estimates vastly overestimate the actual round-off error. In order to get the real estimate for the round-off error, therefore, the appraisal must be carried out according to the "normal" growth of the round-off error. And to this end, we must use the probabilistic theory of round-off errors. Using the probabilistic theory, Henrici[2,3] examined round-off errors on fixed point machine in great detail, and his many results are still suitable for the cases of floating point machine. Before we carry out the statistic treatments for the round-off errors, we must make an important hypothesis that the local round-off errors are independently random variables with the distribution $F(x)$. Additionally, in order to simplify the proof by a large margin and keep the correctness of the final result unchanged, we also assume that the accumulated round-off errors $r_n (n = 1, 2, \cdots)$ are independent (in fact, even if no hypothesis is made, the result are the same except the proofs becomes very complex1)). As mentioned above, the dominant rounding error has most significant contribution to the local round-off error on floating-point machine and if neglecting other round-off errors, then the local round-off error ε_n is of order uy_n, where $u = \gamma/10 = 0.5 \times 10^{-n}$, γ is the machine precision and n the significant digit. Let $E(\xi)$, $D(\xi)$ represent the expected value and the variance of the random

1) see footnote 1) on p.69.

variable ξ. It is easily proved that

Theorem 11. If the local round-off error ε_n is an independent random variable, then

$$E(\varepsilon_{n+1}) = 0, \tag{74}$$

$$D(\varepsilon_{n+1}) = \frac{1}{3}(uy_n)^2. \tag{75}$$

Lemma 8. Let the characteristic polynomial $\rho_k(\xi)$ satisfy the root condition, and let the coefficients $\gamma_j (j = 0, 1, \cdots)$ be defined by

$$\frac{1}{\alpha_k^2 + \alpha_{k-1}^2\xi + \cdots + \alpha_0^2\xi^k} = \tilde{\gamma}_0 + \tilde{\gamma}_1\xi + \tilde{\gamma}_2\xi^2 + \cdots, \tag{76}$$

then

$$\tilde{\Gamma} = \sup_{j=0,1,\cdots} |\tilde{\gamma}_j| < \infty.$$

Following Henrici's treatment, we write the round-off error in the form

$$r_n = \sum_{l=k}^{n} d_{n,l}\varepsilon_l, \tag{77}$$

where $r_i = 0, i = 0, \cdots, k-1, d_{n,l}$ are undetermined. For linear k-step method, $d_{n,l}$ satisfy

$$\sum_{j=0}^{k} \alpha_j d_{n+j,l} = h\sum_{j=0}^{k} \beta_j L_{n+j} d_{n+j,l}, \quad l = k, \cdots, n, \tag{78}$$

$$\sum_{j=J}^{k} \alpha_j d_{n+j,n+j-1} = h\sum_{j=J}^{k} \beta_j L_{n+j} d_{n+j,n+j-1}, \quad J = 1, \cdots, k-1, \tag{79}$$

$$\alpha_k d_{n+k,n+k} = 1 + h\beta_k L_{n+k}. \tag{80}$$

Based on the above assumptions, from (70) it follows that

$$\sum_{j=0}^{k} \alpha_j^2 D(r_{n+j}) = 2h\sum_{j=0}^{k} \beta_j L_j D(r_{n+j}) + \sigma_{n+1}^2 + O(h),$$

where $\sigma_{n+1}^2 = D(\varepsilon_{n+1})$. Thanks to Lipschitz condition, $|L_j| \leqslant L$, and by use of Lemma 5 and $r_i = 0, i = 0, \cdots, k-1$, one has

$$D(r_n) \leqslant (1 + O(h))\frac{\sigma^2}{h}\frac{\mathrm{e}^{2nhLB\tilde{\Gamma}^*} - 1}{2BL},$$

where $\sigma = \max\limits_{k \leqslant j \leqslant n} \sigma_j, \tilde{\Gamma}^* = \tilde{\Gamma}/(1 - h\beta|\alpha_k|^{-1})$. Thereby we have

Theorem 12. If the local round-off errors are independent random variables, for the linear k-step method (6), the accumulated round-off error is a random variable of which the expected value is zero and the variance satisfies

$$D(r(t;h)) \leqslant (1 + O(h))\frac{\sigma^2}{h}\frac{\mathrm{e}^{2LB\tilde{\Gamma}^*(t-t_0)} - 1}{2BL}. \tag{81}$$

In a similar manner, we have

Theorem 13. For the general k-step method (3), let $\Phi_k(t, y; h)$ satisfy (52). If the local round-off errors are independent random variables, the accumulated round-off error is a random variable whose expected value is zero and whose variance satisfies

$$D(r(t;h)) \leqslant (1 + O(h))\frac{\sigma^2}{h}\frac{\mathrm{e}^{2kL\tilde{\Gamma}^*(t-t_0)} - 1}{2kL}, \tag{82}$$

where $\sigma^2 = \max\limits_{k \leqslant j \leqslant n} D(\varepsilon_j) = \max\limits_{k \leqslant j \leqslant n} (uy_j)^2/3, \tilde{\Gamma}^* = \tilde{\Gamma}/(1 - h|\alpha_k|^{-1}L_k), L_k$, as in Theorem 5.

This result shows that the "normal" growth of accumulated round-off error is characterized by the standard deviation of the random variable $r(t;h)$ and is of the order of $h^{-1/2}$, which is by a factor $h^{1/2}$ better than the theoretical upper bound given under the sole assumption. In the following, the higher-order minor term $O(h)$ will be neglected, and then the bound for the accumulated round-off error obtained by use of the probabilistic theory is

$$\| r(t;h) \| \sim \sigma \frac{e^{kL\tilde{\Gamma}^*(t-t_0)}}{\sqrt{2hkL}}.$$

4 Computational uncertainty principle

In the light of the error estimates given above, the various phenomena observed in part I of this paper[1] can be interpreted. If the starting values are exact and there are no initial round-off errors, for the general k-step method of order p, with the above results, we can easily prove that the total error satisfies

$$\| E(t;h) \| = \| e(t;h) + r(t;h) \| \leqslant C(t)\tilde{E}(h) = C(t)\left(Ch^p + \frac{\sigma}{\tilde{C}\sqrt{h}} \right). \quad (83)$$

where $C(t) = e^{C_L\tilde{\Gamma}(t-t_0)}/\sqrt{C_L}$ is the time function, $\hat{\Gamma} = \max(\Gamma^*, \tilde{\Gamma}^*)$, $\tilde{C} = \sqrt{2C_L}$, C is a constant depends on methods, C_L a constant depends on ODEs, $\sigma = \max\limits_{t\in[t_0,t_n]} u \| y(t) \| /\sqrt{3}$, $u = \gamma/10 = 0.5 \times 10^{-n}$ where γ is the machine precision and n the number of significant digit, $C_L = L$ for one-step methods, and $C_L = BL$ or bL/Γ for linear multistep methods. For the general k-step method, $C_L = kL$. For the Taylor series method, the explicit and implicit Adams methods, $C = C_{p+1}M_{p+1}$ where C_{p+1} is the error constant, $M_{p+1} = \max\limits_{t\in[t_0,t_n]} \| y^{(p+1)}(t) \|$. One easily obtains

Theorem 14. $\tilde{E}(h)$ is minimized while

$$h = H = \left(\frac{\sigma}{2pC\tilde{C}} \right)^{1/(p+0.5)}. \quad (84)$$

This indicates that there exists an optimal stepsize H because of the finiteness of machine precision, i.e. there exits a total error (corresponding to H). That is why there exist OSs observed in the numerical experiments in ref. [1]. The round-off error reduces as the machine precision increases, so from (84) OS correspondingly becomes smaller. The smaller the C, the greater the H will be. The higher p, the larger the H. These explain the results in ref. [1]. OS increases as the order of method increases, OS in double precision is smaller than that in single precision, and OSs of the Taylor series and implicit Adams methods are bigger than those of the explicit Adams methods with the same order. Besides, if $y(t) \in \mathbb{C}^\infty[a,b](t\in[a,b])$, then from (84) $H \to 1$ as $p \to \infty$.

Theorem 15. Integrating an ODEs with the same numerical method of order p in two machine precisions γ_1 and γ_2 with n_1 and n_2 being significant digits respectively ($\gamma = 5 \times 10^{-n_1}$, $\gamma_2 = 5 \times 10^{-n_2}$, $n_1 < n_2$), we have the ratio of their OSs H_1 to H_2

$$l = \frac{H_1}{H_2} = 10^{\frac{\Delta n}{p+0.5}}, \quad (85)$$

where $\Delta n = n_2 - n_1$.

This verifies the universal relation of l found in ref. [1]. This relation suggests that l depends only on the order of method and the machine precision (i. e. the significant digits) and is independent of the types of ODEs, initial values and numerical schemes. Therefore OS with any machine precision can be determined in the light of the relation provided that OS under a certain machine precision is known. If n_1 and n_2 are fixed, $l \to 1$ as $p \to \infty$. Thus, once the machine precision is given, the best degree of accuracy which can be achieved for the numerical solution obtained by a numerical method is determined entirely. And the best accuracy is corresponding to the optimal stepsize. From (4) we have the following formula of the total error for the OS H:

$$\| E(t;H) \| \approx C(t) \frac{\sigma}{\tilde{C}\sqrt{H}}\left(1 + \frac{1}{2p}\right). \tag{86}$$

(86) indicates that the error decreases as H or p increases, and decreases as σ(i.e. γ) decreases. If the error tolerance $\delta > 0$ is given, the integration time when the error determined by (86) increases, the error tolerance is just the maximally effective computation time (MECT) T ($T = t - t_0$), namely

$$C(T) = \frac{\delta \tilde{C}\sqrt{H}}{\sigma(1 + 1/2p)}. \tag{87}$$

Obviously, MECT T increases as H or p increases, and increases as σ(i.e. γ) decreases. This explains the results in ref. [1]: MECT in double precision is longer than that in single precision, and MECTs of the RK, Taylor series and implicit Adams methods are longer than those of the explicit Adams methods with the same order.

Theorem 16. With the same numerical method of order p in two machine precisions γ_1 and $\gamma_2(\gamma_1 \geqslant \gamma_2)$, let the corresponding MECTs be T_1 and T_2 respectively, the ratio of the time functions $C(T_1)$ to $C(T_2)$ is

$$k = \frac{C(T_2)}{C(T_1)} = l^p. \tag{88}$$

It is another universal relation. From this relation, we have

$$\Delta T = \hat{C} \cdot p \ln l, \tag{89}$$

where $\Delta T = T_2 - T_1, \hat{C} = (C_L \hat{\Gamma})^{-1}$. As $p \to \infty$, $k \to \gamma_2/\gamma_1 = 10^{\Delta n}$, $\lim\limits_{p\to\infty} \hat{C}^{-1}\Delta T = \Delta n \ln 10$. In two precisions in ref. [1] ($n_1 = 7, n_2 = 16, \Delta n = 9$), $\lim\limits_{p\to\infty} \hat{C}^{-1}\Delta T = 9\ln 10$. This explains the result in ref. [1] that the difference between MECT in double precision and in single precision tends to a fixed value with the increase in order p.

The above theoretical analyses suggest that the phenomena found in the numerical experiments in ref. [1] can be fully explained with round-off error considered. Furthermore, on the basis of them, we will give the mathematical expression of the computational uncertainty principle presented in this paper. To this end, we express $\tilde{E}$ as $\tilde{E} = \tilde{e} + \tilde{r}$, where $\tilde{e} = Ch^p, \tilde{r} = \sigma/\tilde{C}\sqrt{h}$, and express σ in terms of the machine precision γ as $\sigma = C_\sigma \gamma$, here $C_\sigma = \max\limits_{t \in [t_0, t_n]} \| y(t) \| / 10\sqrt{3}$ is a constant. $\tilde{e}$, representing the essential part of the global discretization error $e(t;h)$, is a measure of uncertainty due to the imperfection of numerical methods themselves; $\tilde{r}$, representing the essential part of the accumulated round-off error $r(t;h)$, is a measure of uncertainty due to the inherent inaccuracy of digital computers, and $\tilde{E}$ is the sum of the two uncertainties.

Theorem 17. If the machine precision is finite, $\tilde{E}$ will not tend to zero

$$\tilde{e} + \tilde{r} \geqslant C_{\min}, \tag{90}$$

where $C_{\min} = (1+2p)[C(C_\sigma \gamma / 2p\widetilde{C})^{-2p}]^{1/(2p+1)}$.

This shows that no matter how small the stepsize is, the total error will not be arbitrarily small unless the machine precision $\gamma \to 0$. The global discretization error trends to zero as the stepsize $h \to 0$ without taking the round-off error into account, and in this case, the numerical solution is theoretically convergent. In practice, however, the round-off error due to finite accuracy of calculation machine is not avoidable, so that the total error will initially decrease as stepsize decreases and the discretization error decreases, and then increases as stepsize decreases further and the round-off error becomes more and more significant (the turning point where the total error becomes increasing is simply corresponding to OS). Therefore, the numerical solution is theoretically convergent as $h \to 0$, but not numerically convergent if we use finite accuracy in calculation. In other words, the theoretical convergence and the numerical convergence, generally, cannot take place simultaneously at the finite machine precision. That is to say, the numerical solution is not actually convergent in practice as $h \to 0$. In order to get more precisely numerical solution, we have to add the machine precision in the computations as stepsize decreases. Further, we get

Theorem 18. Let $\tilde{e}^* = \tilde{e}^{1/2p}$, then

$$\tilde{e}^* \cdot \tilde{r} = \hbar_p, \tag{91}$$

where $\hbar_p = \gamma C_\sigma C^{1/2p} / \widetilde{C}$.

This is simply the computational uncertainty relation, which is the expression of the well-known uncertainty relation[11, 12] of quantum mechanics in numerical calculation. It indicates that the global discretization error due to numerical method and the accumulated round-off error due to calculation machine are two "adjoint" variables; they cannot decrease to zero simultaneously, and the smaller one of the two uncertainties, the greater will be the uncertainty of the other adjoint variable. As there exists an inherent relationship between the two uncertainties, it naturally causes limitation in the width of interval of effectively numerical solution. This is at the bottom of the inexorable existence of MECT. That is to say, if one fixes on the error tolerance $\delta > 0$(i.e. the numerical solutions which are less than the tolerance are acceptable), there is surely MECT T, so the numerical solutions in the interval [0, T] satisfy the requirement of the tolerance and present the exact solutions in the interval very well, and the exact solutions beyond the interval cannot be determined by numerical methods. The computational uncertainty principle therefore gives a certain limitation to the calculated ability under the finite machine precision.

5 Conclusion and discussion

On the basis of the classical results of error analysis for numerical methods in ODEs, the error propagation for the general numerical method in ODEs is dealt with, and the shortcomings in the classical results are pointed out. According to the properties and characteristics of the problems discussed, apart from the traditional concept of convergence (namely the theoretical convergence in this paper) two new concepts of convergence are presented: the numerical convergence and the actual convergence. And the various components of round-off error of general numerical method on floating-point machine are fully detailed. By introducing a new kind of recurrent inequality and using the probabilistic theory, not only the classical results on linear multistep methods are improved essentially and the "normal" growth of the accumulated round-off error on floating-point machine is derived, but also a unified estimate for the total error of general multistep method is given. Based on the results of error analysis, we rationally interpret the various phenomena found in the numerical experiments in ref. [1], derive two universal relations which are independent of types of ODEs, initial values and numerical schemes and are in agreement with the results in numerical experiments, and point out that the theoretical convergence and the numerical convergence of numerical solution cannot take place simultaneously, i.e. the numerical solution is not actually convergent. Further theoretical analyses give the mathematical expression of the computational uncertainty principle. In the light of it, we expound that the two uncertainties due to numerical method and calculation machine are two "adjoint" variables, and they cannot decrease simultaneously to zero, and so we explain the root cause that there surly exist OS and ECMT for numerical method under finite machine precision. In order to obtain more precisely numerical solution with longer valid range, from the computational principle, there must be added machine precision in the computations.

Besides, by the results of this paper, there are still some additional discussions about using computer to study the problems of numerical simulation and prediction: (i) In practice, the computational uncertainty principle points out that there is a limit to the ability of effective simulation of computers. We must recognize this point. There is a limit to the ability of effective simulation because computation errors are completely inevitable except for a zero measure set. The existence of this limit is inherent and is independent of the objects simulated (more precisely except a zero measure set). The size of the limit, however, usually depends on the objects simulated. Once the object studied and the machine precision used are given, the best ability of simulation which can be achieved is determined. This limitation is also inherent and cannot be overcome by improving the model describing the object or the data; improving the model describing the object or the data only can make the ability of simulation gradually approach the best degree. (ii) Using the computational uncertainty principle we can make simulations to the best. The computational uncertainty principle on the one hand points out the limit of simulation ability, and on the other hand, points out an optimal relation. The optimal relation gives a way to come up the best ability of simulation. According to the relation, we must affirm which computational results are valid and can be made sure, and which computational results are invalid and cannot be made certain, the correct parts in numerical prediction results are thereby determined. (iii) Developing computers with higher precision is a way to enhance the ability of effective computation. At present for the various numerical methods in differential equations, all their kenels are the recurrent processes step by step. There is surely MECT for this class of methods, and the integration results beyond the time will be invalid, so the long-time behavior of system cannot be properly analyzed. According to the computational uncertainty principle, as long as machine precision is added, MECT can be extended, so the ability of effective computation is raised. In a word, we are up against the change of idea from the idealization of infinite precision to the reality of finite precision. In the course of the change how to break through the computational uncertainty principle and to raise the ability of long time numerical integration for ODEs is an important problem to be solved.

Acknowledgements This work was supported by the Knowledge Innovation Key Project of Chinese Academy of Sciences in the Resource Environment Field (KZCX1-203), Outstanding State Key Laboratory Project (Grant No. 49823002), the National Natural Science Foundation of China (Grant No. 49905007), National Key Foundation Research Project on the Formation Mechanism and Prediction Theory of Heavy Climate Disasters in China (G1998040900, Part 1) and Innovation Project of IAPICAS (8-1301)

References

1. Li Jianping, Zeng Qingcun, Chou Jifan, Computational Uncertainty Principle in Nonlinear Ordinary Differential Equations I.

Numerical Results, Science in China, Ser. E, 2000, 43(5): 449

2. Henrici, P., Discrete Variable Methods in Ordinary Differential Equations, New York: John Wiley, 1962, 1; 187.
3. Henrici, P., Error Propagation for Difference Methods, New York: John Whiley, 1963.
4. Gear, C. W., Numerical Initial Value Problems in Ordinary Differential Equations, Englewood Cliffs, NJ: Prentice-Hall, 1971, 1; 72.
5. Hairer, E., Nørsett, S. P., Wanner, G., Solving Ordinary Differential Equations I. Nonstiff Problems, 2nd ed., Berlin-Heidelberg-New York: Springer-Verlag, 1993, 130.
6. Stoer, J., Bulirsch, R., Introduction to Numerical Analysis, 2nd ed., Vol. 1, Berlin-Heidelberg-New York: Springer-Verlag (reprinted in China by Beijing Wold Publishing Corporation), 1998, 428.
7. Li Qingyang, Numerical Methods in Ordinary Differential Equations (Stiff Problems and Boundary Value Problems), in Chinese Beijing: Higher Education Press, 1991, 1.
8. Li Ronghua, Weng Guochen, Numerical Methods in Differential Equations (in Chinese), 3rd ed., Beijing: Higher Education Press, 1996, 1.
9. Dahlquist, G., Convergence and stability in the numerical integration of ordinary differential equations, Math. Scandinavica, 1956, 4: 33.
10. Dahlquist, G., 33 years of numerical instability, Part I, BIT, 1985, 25: 188.
11. Heisenberg, W., The Physical Principles of Quantum Theory, Chicago: University of Chicago Press, 1930.
12. McMurry, S. M., Quantum Mechanics, London: Addison-Wesley Longman Ltd (reprined in China by Beijing World Publishing Corporation), 1998.