Computational uncertainty principle in nonlinear ordinary differential equations (I)

——Numerical results

LI Jianping (李建平)[1,2], 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@lasgsgi4.iap.ac.cn)

Received September 6, 1999

Abstract In a majority of cases of long-time numerical integration for initial-value problems, round-off error has received little attention. Using twenty-nine numerical methods, the influence of round-off error on numerical solutions is generally studied through a large number of numerical experiments. Here we find that there exists a strong dependence on machine precision (which is a new kind of dependence different from the sensitive dependence on initial conditions), maximally effective computation time (MECT) and optimal stepsize (OS) in solving nonlinear ordinary differential equations (ODEs) in finite machine precision. And an optimal searching method for evaluating MECT and OS under finite machine precision is presented. The relationships between MECT, OS, the order of numerical method and machine precision are found. Numerical results show that round-off error plays a significant role in the above phenomena. Moreover, we find two universal relations which are independent of the types of ODEs, initial values and numerical schemes. Based on the results of numerical experiments, we present a computational uncertainty principle, which is a great challenge to the reliability of long-time numerical integration for nonlinear ODEs.

Keywords: ordinary differential equations (ODEs), computational uncertainty principle, round-off error, discretization error, strong dependence on machine precision, maximally effective computation time (MECT), optimal stepsize (OS), universal relation, nonlinear.

Since most ODEs cannot be analytically solved, it is necessary to get their approximate solutions by numerical methods[1—3]. When solving an initial value problem by any discrete variable method, there are two basic sources of error: discretization error, which is caused by discretizing differential equations[1,2], and round-off error, which is due to the finiteness of machine precision[1,4,5]. For a given initial value problem, all standard discrete variable methods are convergent if the calculations are carried out without round-off error. Therefore, little attention is paid to round-off error in practice. Nevertheless, this does not mean that the influence of round-off error is unimportant in most cases. In fact, the properties of exact solution may be radically changed in long-time numerical integration for nonlinear ODEs because of the inevitability of round-off error. Henrici[1] investigated round-off error on a fixed-point machine using probability theory and illustrated his theory with some linear equations, but the influence of round-off error on long-time numerical integration was unnoticed. However, the floating-point arithmetic is prevalent on modern electronic digital computers, and ODEs are generally nonlinear. We therefore investigated the important effect of round-off error on long-time numerical integration for nonlinear ODEs on floating-point machine using both numerical experiments and theoretical analysis. We find that there exist very serious problems caused by the finiteness of machine precision in numerical calculations, and after the numerical integration of finite steps, numerical solution obtained by any stepsize is unrelated to exact solution. As a result we present a computational uncertainty principle in nonlinear ODEs. This paper is divided into two parts: numerical results and theoretical analysis. The aims of part I is to show some new important phenomena and the main results found in numerical experiments and in providing importantly experimental evidence for the theoretical analysis in part II. And part II gives the theoretical interpretation for the numerical results in part I.

1 Numerical model and numerical methods

The model discussed here is the Lorenz equations[6],

$$\dot{x} = -\sigma x + \sigma y, \tag{1}$$

$$\dot{y} = rx - y - xz, \tag{2}$$

$$\dot{z} = xy - bz, \tag{3}$$

where $\sigma = 10$, $b = 8/3$, $0 < r < +\infty$. The initial problem of Lorenz equations has a unique solution. Here the equations are chosen because of their fine representativeness. When $1 < r < 24.74$, there are two stable fixed points $C\,(\sqrt{b(r-1)},\ \sqrt{b(r-1)},\ r-1)$, $C'(-\sqrt{b(r-1)},\ -\sqrt{b(r-1)},\ r-1)$ and an unstable fixed point $O = (0,0,0)$ in Lorenz equations. When $r > 24.74$, C and C' become unstable, and in this case, there are chaos and a strange attractor[6—8].

Four classes of twenty-nine standard numerical methods[1—4,9] used here are as follows: (i) Explicit one-step methods: Euler's method, Runge-Kutta (RK) methods of orders from 2 to 6, Taylor series methods of orders from 2 to 10; (ii) explicit multistep methods: explicit Adams methods of orders from 2 to 6; (iii) implicit methods: implicit Euler's method, implicit Adams methods of orders from 2 to 6; (iv) modified predictor corrector (PMECME) methods: trapezoidal PMECME method of order 2, Adams PMECME method and Hamming PMECME method of order 4. All computations have been performed on an SGI ORIGIN 2000 computer (its single and double precisions are 7 and 16 significant digits, respectively).

2 The case without chaos

First we study the case of $1 < r < 24.06$, in which there is no chaos in the Lorenz equations[6—8]. Nevertheless, a strange phenomenon results from numerical experiments. Take the classical 4-th order RK method as an instance, we use two stepsizes with very slight difference to compute the same initial value (5,5,10), but we get two essentially different final states (fig. 1(a)—(c)), showing that there is a contradiction between the non-uniqueness of numerical results and the uniquencess of theoretical solution. More importantly, it is a general phenomenon for stepsize. As shown in figs. 2 and 3, the final states of the numerical solutions for this initial value, whether computed in single precision or in double precision, are very sensitive to stepsizes. That is, different stepsizes with slight difference may probably lead to significantly different final states of numerical solution. Obviously, the set of stepsize can be divided into two classes: one (A) corresponds to the final state C and the other (B) to the final state C' (figs. 2 and

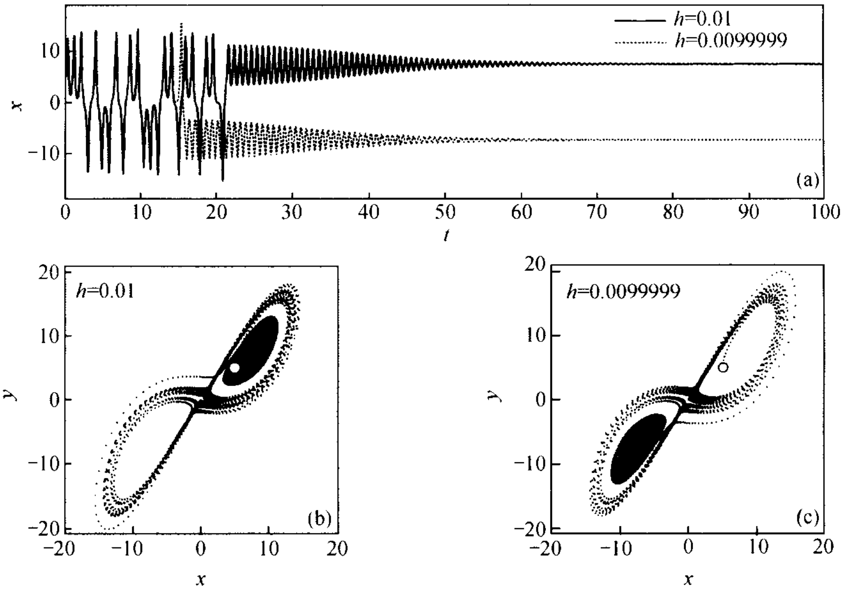

Fig. 1. Numerical solutions of the Lorenz equations computed by the 4-th RK method for $r = 22$ and the initial value (5,5,10). (a) The solutions of x-component obtained by two stepsizes with very slight difference in single precision; (b) the projection on the x-y-plane, the stepsize $h = 0.01$, and the open symbol is the initial point; (c) as in (b), but for the stepsize $h = 0.0099999$.

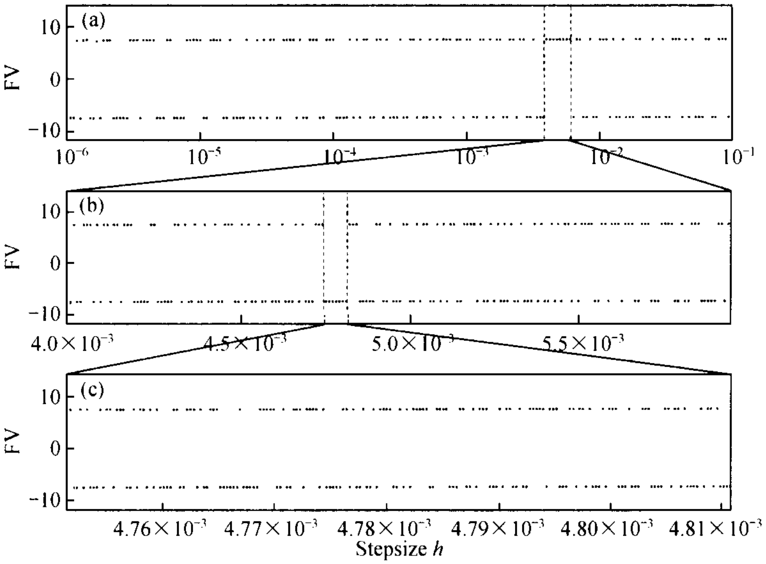

Fig. 2. Final value (FV) of x-component of the Lorenz equations obtained by the 4-th RK method versus stepsize h in single precision. The stepsize h is (a) from 10^{-6} to 10^{-1}, (b) from 3.981070×10^{-2} to 5.956621×10^{-2}, (c) from 4.751533×10^{-2} to 4.810801×10^{-2}, respectively. FV is the value of final state which the solution approaches and remains fixed there as time increases. The used numerical method, initial value and r used here as in fig. 1.

3). That is to say, if a stepsize belongs to A (or B), then the final state of the numerical solution obtained in this stepsize is C (or C'). To our surprise, the two classes of set display the property of the Contor set[10] (fig. 2(b) and (c), and fig. 3(b) and (c)), clearly showing that the numerical solutions for different stepsizes exhibit random property against stepsize. The percentage of the number of stepsizes which belong to A will approach to 50%, as the total number of stepsizes used increases (fig. 4). The above phenomenon is unexpected and holds for the other twenty-eight numerical methods[1)] (figures omitted). Therefore, not only can the numerical methods bring about spurious solutions[11], but more importantly, we are unable to know which solution to stepsize is true. This indicates that the final state of initial value such as (5,5,10) cannot be calculated accurately with these numerical methods under the given machine precision. The initial value of this kind is thus called ill-behaved initial value (IBIV). The point set composed of IBIVs is called the IBIV point set. In contrast, there are well-behaved initial value (WBIV) and the WBIV point set. As shown in fig. 5, there are two types of WBIV.

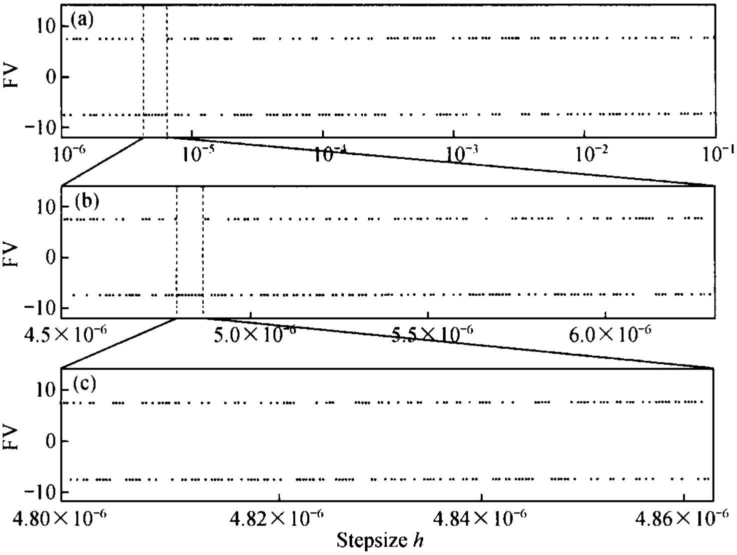

Fig. 3. As in fig. 2, but for double precision, and the stepsize h is (a) as in fig. 2(a), (b) from 4.4466834×10^{-6} to 6.309573×10^{-6}, (c) from 4.798529×10^{-6} to 4.863025×10^{-6}, respectively.

To interpret the above phenomenon, we have carried out two groups of numerical experiments on another initial vlaue (3,4,10). As shown in fig. 6, for the 4-th RK method, this initial value is ill-behaved in single precision, but it is well-behaved in double precision. The conclusion holds for the other methods[1)] (figures omitted). This suggests that round-off error due to the finiteness of machine precision plays a decisive role in the above phenomenon. Moreover, it shows that the quality of an initial value depends on machine precision. An initial value is an IBIV in lower machine precision, but it may be a WBIV in higher machine precision. This reveals that the numerical results have a strong dependence on machine precision. And this kind of dependence is quite distinct from sensitive dependence on initial conditions for two reasons.

[1)] Refer to 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.

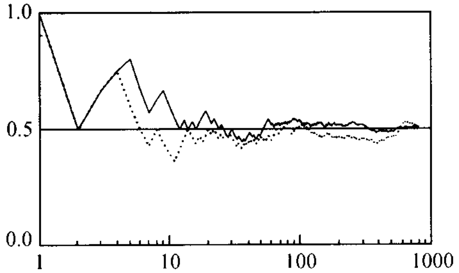

Fig. 4. Percentage of the number of stepsizes which belong to A as a function of the total number of stepsizes used. The soild line is in single precision and the dashed line in double precision. Here the used numerical method, the equations, r and the initial value are the same as in fig. 1 with all stepsizes belonging to $[10^{-6}, 0.1]$.

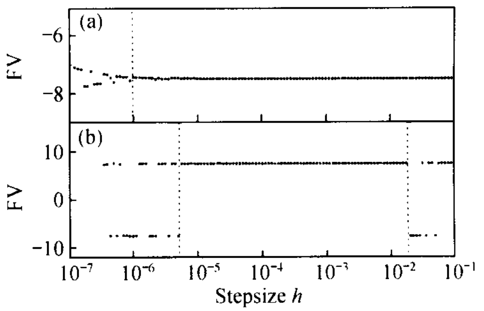

Fig. 5. Two types of WBIV. (a) The computed FV of x-component of the Lorenz equations for the initial value (0, 1, 0) versus stepsize h. The computed FVs are basically equal to the same constant with stepsizes (between 10^{-6} and 0.1); (b) as in (a), but for the initial value (5, 1, 19.5). The computed FVs are sensitive to the bigger and smaller stepsizes, but are unchanged to the moderate stepsizes. The numerical method, equations and r used here are the same as in fig. 1, and using single precision.

In the first place, there is no chaos in the case discussed here. The other, and more important, sensitive dependence on initial conditions is essentially independent of machine precision. Therefore, the strong dependence on machine precision is a new phenomenon to which full attention-should be paid in numerical computation and simulation.

Generally, the solution of a WBIV can be computed accurately by a numerical method. However, as is clear in fig. 7, the so-called accuracy here is only in the sense of acceptable level. In addition, fig. 7 suggests that the absolute error does not decrease as the stepsize decreases, but increases as the stepsize decreases while the stepsize is less than a critical stepsize. This

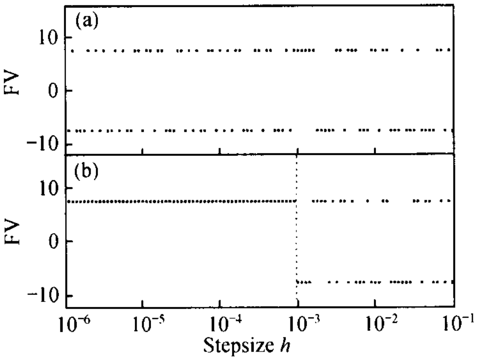

Fig. 6. FV of x-component of the initial value (3, 1, 10) versus stepsize h. (a) Single precision. The computed FVs are sensitive to stepsizes, showing an IBIV in single precision; (b) double precision. The calculated FVs are sensitive to the bigger and smaller stepsizes, but are nearly equal to the same constant for the moderate stepsizes, showing a WBIV in double precision. Here the used numerical method, equations and r are as in fig. 1.

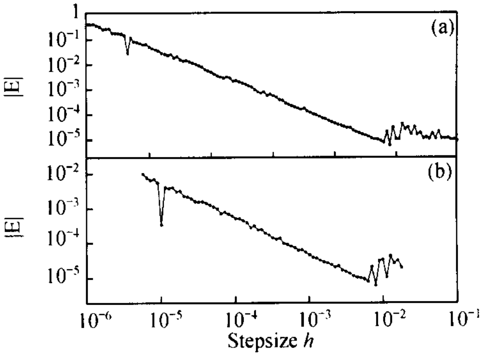

Fig. 7. Absolute error E of correctly computed FV of x-component for two WBIVs in fig. 5. (a) and (b) correspond to fig. 5(a) and (b), respectively.

critical stepsize corresponds to the minimum absolute error and will be called the optimal stepsize as will be pointed out below (and in part II of this paper——theoretical analysis).

Whether the Lebesgue measure of IBIV point set for a numerical method is zero under a given machine precision is an important question. If it is, it is a zero probability event that IBIV is met in practical calculations, and the numerical method is therefore called successful. Otherwise, the method is not absolutely successful. Based on hosts of numerical experiments for 580 810 initial values with the 4-th order RK method (table 1), there are many IBIVs in the Lorenz equations in the case of no chaos. They account for a majority when $r = 23$, and will still markedly grow in number as r increases. To our surprise, the so-called IBIVs are not in the place where two basins of attraction of the two stationary attractors C and C' meet (Plate Ⅰ-1). These results show that the 4-th order RK method is not absolutely successful under these conditions. Additionally, the number of IBIVs in double precision is less than that in single precision, which suggests that the IBIVs would decrease in number if the machine precision increases. This again verifies the importance of round-off error in the numerical calculations. These conclusions are also suitable for the other twenty-eight numerical methods.

Table 1 Percentages of WBIVs and IBIVs obtained by the 4-th order RK method for the Lorenz equations in some sections

		$P_{z=10}$	$P_{z=r-1}$	$P_{y=1}$	$P_{y=10}$
$r = 22$,	p_1	31.02 (37.33)	33.31	31.25	29.43
	p_2	31.02 (37.33)	33.31	30.76	28.49
	p_3	37.97 (25.33)	33.38	37.98	42.08
$r = 23$,	p_1	21.70 (23.78)	23.80	20.64	19.90
	p_2	21.70 (23.78)	23.80	20.25	17.90
	p_3	56.59 (52.64)	52.39	59.10	62.20

P, plane. The percentages in parentheses are obtained in double precision, the rest in single precision. p_1 and p_2 represent the percentages of WBIVs whose final states are C and C', respectively. p_3 represents the percentage of IBIVs. The section in the plane z is $-60 \leqslant x \leqslant 60$ and $-60 \leqslant y \leqslant 60$, and in the plane y is $-60 \leqslant x \leqslant 60$ and $-60 \leqslant z \leqslant 60$. Here every section is divided into 240×240 meshes and the initial points are on the mesh points (amount to $241 \times 241 = 58081$).

3 The case with chaos

Now we discuss the case of $r > 24.74$, in which the Lorenz equations have chaos[6—8], a strange attractor and no stable steady state. To display the phenomena from the numerical calculations, we have performed a lot of numerical experiments and have obtained many stepsize-time plots of numerical solution for different initial values. This kind of plot can clearly display the evoluion behavior of numerical solutions obtained by using different stepsizes for the same initial value and the difference between these solutions. Plate Ⅱ-2(a) shows that result of the 4-th RK method with hundreds stepsizes in single precision for the initial value (5,5,10) and $r = 28$. As shown, the isolines of numerical solutions are parallel and straight in the beginning, indicating that the solutions for different stepsizes are very close to each other. However, the isolines in regions of the bigger and smaller stepsizes appear waves after a short run, which implies that the solutions for these stepsizes are not in accord with those for others, and the length of parallel isoline which is called the width of interval of effective stepsizes starts shortening. As integration proceeds, the width of interval of effective stepsize (IES) is becoming smaller and smaller and finally becomes zero at the time of 17 or so. Beyond the time point the isolines become disordered and unsystematic and the numerical solutions for all stepsizes are out of step, meaning that the difference between solutions for different stepsizes is significant. That is to say, all numerical solutions are unrelated to exact solution. Thus, the time when the width of interval of effective stepsize becomes zero is called the maximally effective computation time (MECT), and the corresponding stepsize is called the optimal stepsize (OS). The difference among the solutions in IES is very small, so the exact solution of ODEs is shown quite well by the solution in IES. But, unfortunately, for the chaos system (even for the system with transient chaos), the width of IES decreases as integration time increases and will be zero very soon, and MECT is arrived. Beyond MECT, the difference between solutions obtained by all stepsizes is significant, and at this time numerical solutions make no sense. The same phenomenon can be also observed in double precision (Plate Ⅰ-2(b)), except that its MECT is about 35, or more than two times the above MECT in single precision, and that its OS is about 1/60 the length of OS in single precision. Likewise, this phenomenon can also be found in the other numerical methods, but these methods differ in MECT and OS. The above results suggest the following: (i) In practice numerical solution cannot converge to exact solution as stepsize $h \rightarrow 0$ because of the finite accuracy of calculations; (ii) there exist MECT and OS, and the solution of the nonlinear ODEs cannot be accurately calculated beyond MECT; (iii) increasing the machine precision can efficiently postpone but not eliminate the influence of round-off error.

Additionally, it is obvious in Plate Ⅰ-2(a), (b) that there is a profile which is a line of effective computation time of each stepsize and is called the effective computation time profile (ECTP). It divides a stepsize-time plot of numerical solution into two distinct parts. On the left of ECTP isolines of numerical solution are parallel and regular, and this region is called numerical laminar region where the exact solution of differential equations is well shown. On the right of ECTP they are turbulent and irregular and this region is called numerical turbulent region where numerical solutions are spurious and are unrelated to exact solution.

It must be pointed out that the value of the phenomenon shown above lies not in showing the sensitive dependence on initial values, but in evaluating the maximally effective time of a numerical solutions in finite machine precision. According to the results mentioned above, we can arrive at the following qualitative conclusion for error and ECPT (fig. 8). The total error will initially decreases as stepsize decreases and the discretization error decreases, so that the effective computation time increases; as stepsize decreases to a certain extent, however, the round-off error caused by calculation machine becomes dominate because the iterative times increase, as a results, the total error increases and the effective computation time decreases. The inevitable outcome is that there exists a stepsize H (fig. 8), and so the total error is the smallest and the effective computation time is the largest at this stepsize. The stepsize H is therefore called optimal stepsize. Owing to the inverse variation between discretization error and round-off error against stepsize, a computational uncertainty principle similar to the well-known Heisenberg uncertainty relation of quantum mechanics[12] is led. Specifically, if the discretization error and the round-off error are treated as two "adjoint variables", the computational uncertainty principle reveals that the smaller one of them, the greater will be the other adjoint variable. This means that once the precision of calculation machine used is given, the best degree of accuracy which can be achieved for the numerical solution obtianed by a numerical method is determined entirely.

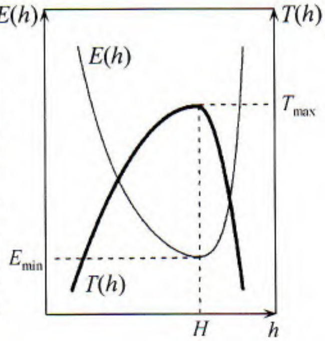

Fig. 8. Total error $E(h)$ and effective computation time $T(h)$ versus stepsize. H denotes OS, $E_{\min}$ minimum error and $T_{\max}$ MECT.

The computational uncertainty principle gives a certain limitation to the MECT of long-time numerical integration for chaos system which is sensitive to initial values and for some systems with transient chaos. This is just the root cause of the various phenomena observed in our numerical experiments. This principle indicates that the computational capacity of numerical methods is very limited for nonlinear ODEs under the inherent property of finite machine precision.

To reveal the intrinsic relationships between MECT, OS, the order of numerical method and machine precision, it is necessary to design and efficient algorithm for ECTP. Here we give an optimal searching algorithm to obtain ECTP. The optimal searching method is carried out by comparing the difference among numerical solutions obtained by a sequence of stepsizes. In the initial interval of stepsize $[h_{\min}, h_{\max}]$ ($h_{\min}$ should be sufficiently small, here $h_{\min}=10^{-7}$, but $h_{\min}=0.3\times10^{-7}$ for the explicit Adams method of order 2 and the Talyor series method of order 2), choose n stepsizes $h_i(i=1,\cdots,n)$ (as well-distributed as possible and n should be larger). Let n numerical solutions obtained by these n stepsizes be $\tilde{y}_t(n)$ at the integration time t. If their difference $V_s(t)$ (may be measured in standard deviation) is less than a given tolerance δ, they express well the value of exact solution at t, and so the interval of effective stepsize is $[h_{\min}, h_{\max}]$ and its width is $W_h(t)=\lg h_{\max}-\lg h_{\min}$. Otherwise, if $V_s(t)$ is greater than δ, it indicates that numerical solutions obtained by certain stepsizes deviate from exact solution, and so reject them from $\tilde{y}_t(n)$ such that the difference among the rest is less than δ. Take the minimum rejection numbers as the optimal principle. From this we have an interval of effective stepsize at t. Go on integrating with the stepsizes in the interval of effective stepsize and repeat continually the above processes until there leave only two adjacent stepsizes $h_j(t_1)$ and $h_{j+1}(t_1)$. Then choose m stepsizes in the interval $[h_j(t_1), h_{j+1}(t_1)]$ again and use them to compute and to compare anew, until finally there is no difference between two adjacent stepsizes remaining below δ. At this time we get ECTP, MECT and OS. For the purpose of comparing different problems, the tolerance δ should be measured in the relative index. Let the oscillation of exact solution from the initial time t_0 to the time t be $V(t)$. Then the tolerance δ requires $V_s(t)/V(t)\leqslant\delta$. Exact solution is generally unknown, so in practice we use the oscillation $V^*(t)$ of numerical solutions obtained by stepsize in the interval of effective stepsize from t_0 to t instead of $V(t)$. In our experiments the tolerance δ is 1/10. Plate Ⅰ-2(c) plots the ECTPs of Plate Ⅰ-2(a), (b) from this optimal searching algorithm. Obviously, ECTP can be obtained accurately.

Using the optimal searching algorithm we have carried out many experiments on 116,160 initial values with the 4th order RK method. Fig. 9(a) and (b) show the distribution of MECT on one section in single precision and in double precision, respectively. These experiments indicate that different initial values have different MECT, and that for the same initial value the MECT in double precision is longer than that in single precision. On average, MECT in single precision and in double precision are 16.857 and 35.412, respectively. Evidently reducing round-error can efficiently increase MECT. Comparing fig. 9(a) with fig. 9(b), we easily find that in pattern they are very similar to each other, and that the difference between MECT in double precision and in single precision seems to approximate a certain fixed number (between 18 and 19 in this case). Therefore, although MECT under one machine precision is dependent on the initial value, the difference in MECT between two machine precision may be independent of the initial value. In fact, the results will be verified by our theoretical analyses in part II of this paper.

Comparing Plate Ⅰ-1 with Plate Ⅰ-3, we can find that for each WBIV in Plate Ⅰ-1 (without chaos) its MECT is relatively long in Plate Ⅰ-3 (with chaos). For the cases with chaos the Lebesgue measure of WBIVs is zero, i.e. the MECTs of almost all initial values are finite under the finite machine precision.

Based on the comparison between mean MECT and mean OS of ten initial values of the Lorenz equations for $r = 28$ under different precisions and different orders (fig. 9), the relationships between MECT, OS, order of numerical method and machine precision are as follows: (i) MECT and OS increase as the order of method increases, but their increments gradually decrease (fig. 9(a)—(c)). Additionally, the MECT and the OS of RK, Taylor series and implicit Adams methods are a little bigger than those of explicit Adams methods with the same order; (ii) in double precision MECT is two times longer than that in single precision (fig. 9(a),(b)), but OS is much smaller than that in single precision (fig. 9(c)). It follows that MECT can be remarkably enhanced by an increase in machine precision, but the corresponding OS is reduced distinctly.

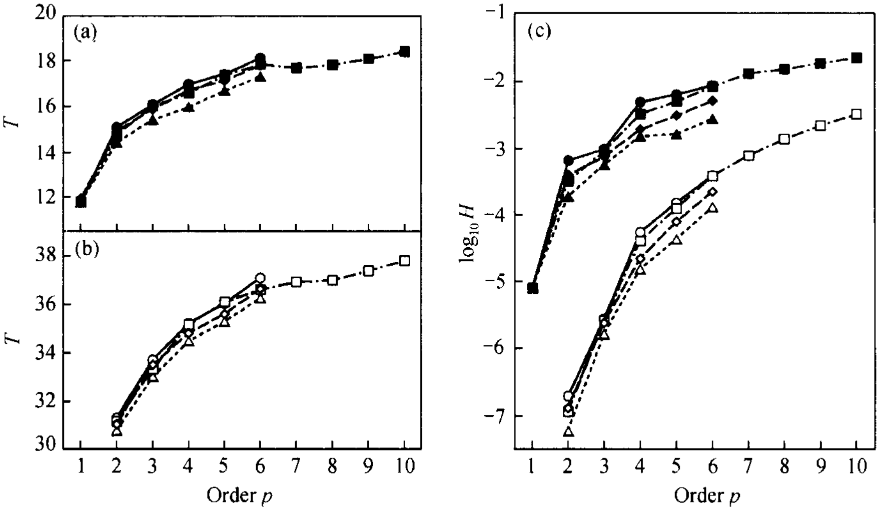

Fig. 9. MECT and OS versus the orders of numerical methods. (a) The mean MECT of ten initial values in single precision for the Lorenz equations when $r = 28$. The ten initial values are (0,1,0), (5,5,10), (−13,4,28), (−15,5,20), (10, −8, −20), (−20, −15,11), (−6,8,13), (11,10,15), (2, −3,16) and (−6, −7, −8), respectively. Solid circles, squares, triangles and diamonds denote RK, Taylor series, explicit Adams and implicit Adams methods, respectively; (b) as in (a), except using double precision and open symbols; (c) as in (a) and (b), but for the mean OS H.

The above analyses show that under different machine precisions MECTs or OSs of numerical methods with the same order are different. However, are there any certain relations between two OSs or between two MECTs for the same order methods under two machine precision? To answer this question, the following index for characterizing the relationship between OSs under double machine precision is defined as

$$l = \frac{H_1}{H_2}, \tag{4}$$

where H_1 and H_2 are two OSs of the same numerical method of order p in double given machine precision $\gamma_1 = 5 \times 10^{-n_1}$, $\gamma_2 = 5 \times 10^{-n_2}$ with n_1 and n_2 significant digits respectively. For convenience, let $n_1 < n_2$ in the following (in this paper $n_1 = 7$, $n_2 = 16$). Fig. 10(a) shows the

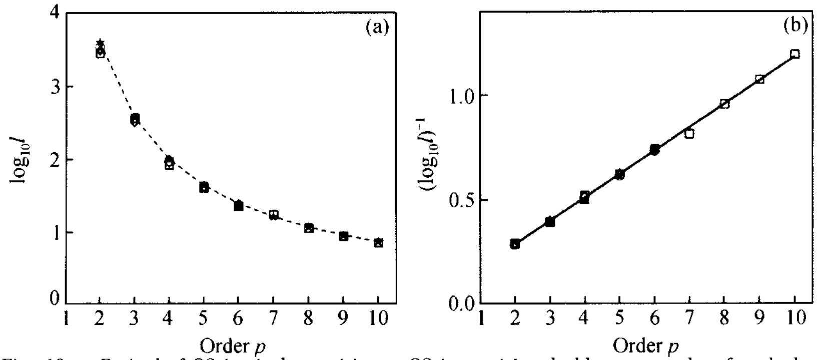

Fig. 10. Ratio l of OS in single precision to OS in precision double versus order of method. (a) As in fig. 9(a), but for the mean l. Stars are theoretical from formula (5), and the rest of the symbols are as in fig. 9(b); (b) as in (a) except for $(\log_{10} l)^{-1}$.

values l from ten initial values in fig. 9 versus the orders. As indicated in the graph, the l values of four classes of numerical methods agree well with each other. This suggests that the ratio of OSs under double machine precision is independent of the numerical method although OS depends on numerical method, i.e. the values l for different methods obey the same law. For different initial values, the law of l remains unchanged. Moreover, the law of l is still the same for different equations[1)] (for example, $y' = y$, $y(0) = 1$ and $y' = -y$, $y(0) = 1$, etc). Thus, the ratio l satisfies universal relation under double machine precision (The conclusion will be proved theoretically in part II of this paper), and l depends only on the order p of method and the machine precision, namely $l = l(p, n_1, n_2)$. By fig. 10(a), the relation between $\log_{10} l$ and p is analogous to the inverse ratio relation. If it is so, its reciprocal is certainly a linear function of p. Just as we expected, this conclusion is verified clearly in fig. 10(b). The coefficients of the linear relation can be determined by the least square method. The results are shown in the column of relation 1 in table 2. From the right side of the statistic relations 1, it is easy to see that their denominators are close to 9. And $\Delta n = n_2 - n_1 = 9$ in this paper. This is not any accidental coincidence. The theoretical analysis in part II of this paper will prove the relation as follows:

$$l = 10^{\frac{\Delta n}{p+0.5}}. \tag{5}$$

In the case of this paper ($\Delta n = 9$) this relation is confirmed by the statistic relation 2 in table 2. Fig. 10(a) shows that the theoretical l values from formula (5) are in close accordance with the experimental values. Hence, OS under any machine precision can be determined in the light of this formula provided that OS at certain machine precision is known.

Table 2 Statistic relations between the experimental l values and p

Methods	Statistic relation 1	Statistic relation 2
RK	$(\log_{10} l)^{-1} = 0.1124947p + 0.0561754 = \dfrac{p + 0.4993607}{8.8893108}$	$(\log_{10} l)^{-1} = \dfrac{1}{9}(p + 0.5553867)$
Taylor series	$(\log_{10} l)^{-1} = 0.1120418p + 0.0611100 = \dfrac{p + 0.5454214}{8.9252377}$	$(\log_{10} l)^{-1} = \dfrac{1}{9}(p + 0.6002492)$
Explicit Adams	$(\log_{10} l)^{-1} = 0.1162587p + 0.0452832 = \dfrac{p + 0.3895037}{8.6015072}$	$(\log_{10} l)^{-1} = \dfrac{1}{9}(p + 0.5928620)$
Implicit Adams	$(\log_{10} l)^{-1} = 0.1107671p + 0.0669475 = \dfrac{p + 0.6043992}{9.0279551}$	$(\log_{10} l)^{-1} = \dfrac{1}{9}(p + 0.5901417)$

[1)] Same footnote 1) on page 452.

For MECT, we define the following index:

$$\Delta T = T_2 - T_1, \tag{6}$$

where T_1, T_2 are the MECTs of the same numerical method of order p under two given machine precisions with n_1 and $n_2(n_1 \leqslant n_2)$ significant digits respectively. Fig. 11(a) shows ΔT from the results in figs. 9(a) and (b) versus order, in which, the values ΔT of four classes of methods are also in good agreement with each other, and the difference ΔT between double precision and single precision tends to a fixed value with the increase in order. Is there also a universal law for ΔT? Is it linked certainly to l? These questions are answered in fig. 11(b). As shown in fig. 11(b), there is a good relation between $e^{\Delta T}$ and l^p, i.e. there is the following formula:

$$\Delta T = p \ln l. \tag{7}$$

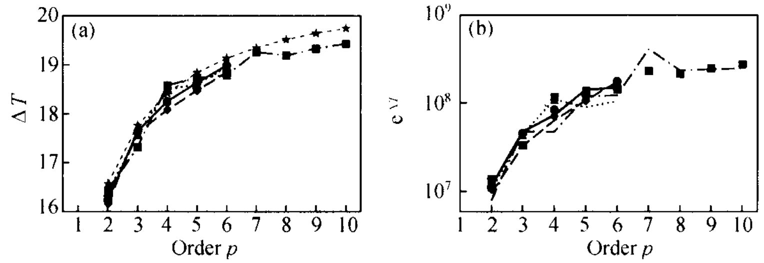

Fig. 11. (a) The difference ΔT between MECTs in double precision and in single precision versus order of method. Stars are from formula (7), and all other symbols are same as in fig. 9 (a); (b) relationship between $e^{\Delta T}$ and l^p. The symbols are as in (a) and represent $e^{\Delta T}$. The solid, dot-dashed, short dashed and long dashed lines denote l^p where l are from RK, Taylor series, explicit and implicit Adams methods in fig. 10(a), respectively.

From fig. 11(a), the ΔT values from this relation also accord with the experimental values. Therefore, MECT at any machine precision can be determined by use of this relation on condition that MECT at some machine precision is known. As $p \to \infty$, $\Delta T \to \Delta n \ln 10$. In two precisions in this paper, one has $\Delta T \to 9 \ln 10 \approx 20.7233$ $(p \to \infty)$.

4 Conclusions

Although round-off error is very small, its effect on the long-time numerical integration for nonlinear ODEs cannot be neglected. Based on a great number of numerical experiments, we find that in the case without chaos numerical solution has a strong dependence on machine precision which essentially differs from sensitive dependence on initial conditions, and that in the case with chaos there exist MECT and OS under a finite machine precision. The solution of ODEs cannot be accurately calculated beyond MECT. Moreover, an optimal seraching method which can evaluate MECT and OS well under a finite machine precision is presented. By use of this method, the essential relationships between MECT, OS, the order of numerical method and machine precision are found. Besides, we have found two universal relations which reveal the intrinsic relationships between two OSs and between two MECTs under any two machine precisions. According to the two relations, OS and MECT can be determined provided that OS and MECT under a certain machine precision are known. There is MECT in a numerical method under a finite machine precision because on the one hand, there is an upper bound limitation for the magnitude of stepsize due to the stability condition of difference method, on the other hand, there must be another limitation of upper bound for the number of integration steps because of the limitations of finite accuracy due to computing on actual machines. The two aspects are contradictory to each other, with the results that they are complementary to each other, leading to the computational uncertainty principle. Because of space limitation the numerical experiments are carried out only for the Lorenz equations in this paper. In fact, we have other ODEs[1] (including linear equations). Based on the theoretical analysis in part II of this paper, the results here are still useful for partial differential equations, though they involve the match between space stepsize and time stepsize. Thus far in our findings, in order to reduce the influence of round-off error, the machine precision must be constantly improved. The machine precision, however, will not be infinite. Therefore the computational uncertainty principle presents a great challenge not only to the effectiveness and reliability of long-time numerical integration for nonlinear ODEs, but also to how to change the current calculation fashions to generalize efficiently the MECT of numerical methods.

Acknowledgements Part of this paper was reported at the Autumn Discussion Meeting 1998 and the Spring Discussion Meeting 1999 of Beijing Computational Fluid Dynamics. The authors are grateful to Profs. Li Deyuan, Yu Xijun and Wu Huamo for their useful discussion and suggestions. The first author is also indebted to Profs. Ji Zhongzhen, Mu Mu and Wang Bin for their helpful advices. This work was supported by the Outstanding State Key Laboratory Project (Grant Nos. 49905007, 49823002), the National Natural Science Foundation of China (Grant No. 49805006), and the National Key Foundation Research Project on the Formation Mechanism and Prediction Theory of Heavy Climate Disasters in China (G1998040900, Part I).

References

1. Henrici, P., Discrete Variable Methods in Ordinary Differential Equations, New York: John Wiley, 1962, 1—165; 187—288.
2. Gear, C. W., Numerical Initial Value Problems in Ordinary Differential Equations, Englewood Cliffs, NJ: Prentice-Hall, 1971, 1—14; 72—86.
3. Hairer, E., Nφrsett, S. P., Wanner, G., Solving Ordinary Differential Equations I. Nonstiff Problems, 2nd ed, Berlin-Heidelberg-New York: Springer-verlag, 1993, 130—430.
4. Stoer, J., Bulirsch, R., Introduction to Numerical Analysis, 2nd ed., Berlin-Heidelberg-New York: Springer-verlag, 1996, 1—36; 428—569.
5. Sterbenz, P. H., Floating Point Computation, Englewood Cliffs, NJ: Prentice-Hall, 1974.
6. Lorenz, E. N., Deterministic nonperiodic flow, J. Atmos. Sci., 1963, 20(130): 130.
7. Kaplan, J. L., Yorke, J. A., Chaotic behavior of multidimensional difference equatins, in Functional Differential Equations and Approximation of Fixed Points (eds. Peitgen, H. O., Walther, H. O.), Berlin: Springer-Verlag, 1979, 204—227.
8. Guckenheimer, J., Holmes, P., Nonlinear Oscillation, Dynamical System and Bifurcations of Vector Fields, New York: Springer-Verlag, 1983.
9. Ralston, A., Rabinowitz, P., A First Course in Numerical Analysis, New York: McGraw-Hill, 1978.
10. Contor, G., Grundlagen einer allgemeinen mannichfältigkeitslehre, Math. Annalen., 1883, 21: 545.
11. Iserles, A., Peplow, A. T., Stuart, A. M., A unified approach to spurious solutions introduced by time discretization, Part I. Basic theory, SIAM J. Numer. Anal., 1991, 28: 1723.
12. Heisenberg, W., The Physical Principles of Quantum Theory, Chicago: University of Chicago Press, 1930.

[1] As the footnote in p452.

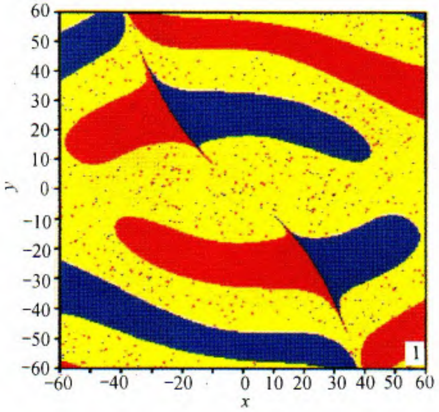

Plate I-1 Distribution of WBIVs and IBIVs on the plane $z = 10$. The section is divided into 240×240 meshes and the initial points are on the mesh points (amount to $241 \times 241 = 58081$). Blue and red solid squares denote the WBIVs whose final states are C and C', respectively. Yellow solid squares denote IBIVs and open one $(0, 0, 0)$. Here the used numerical method, equations and r are as in fig. 1, and using single precision.

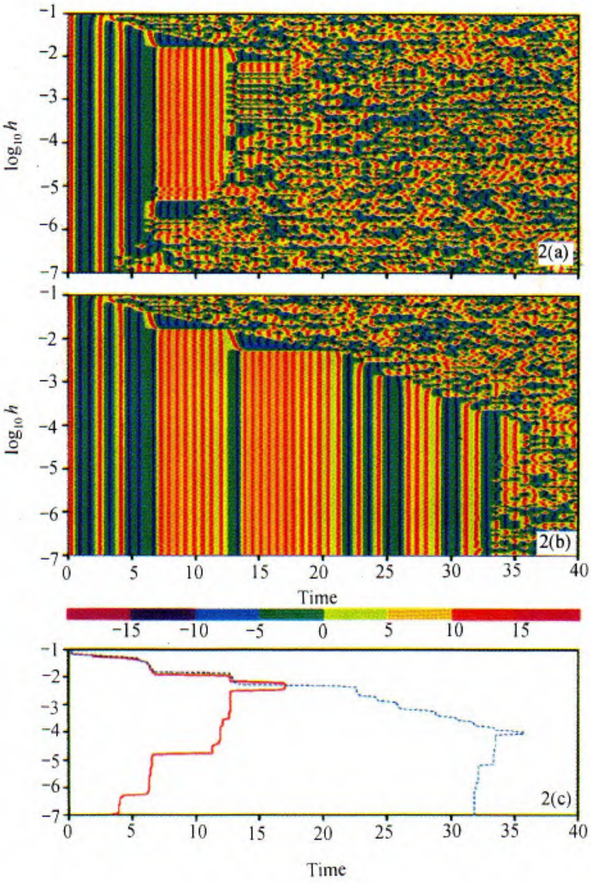

Plate I-2 Stepsize-time plots of numerical solution and ECTPs. (a) The stepsize-time plot of x-component obtained by 121 stepsizes in single precision for the initial value $(5, 5, 10)$ and $r = 28$. Here stepsize h takes logarithm to the base of 10, time is non-dimensional, and the used numerical method and equations are as in fig. 1; (b) as in (a), but for double precision; (c) the ECTPs of (a) and (b) obtained by the optimal searching method. The solid line and the dotted line are the ECTPs of (a) and (b), respectively.

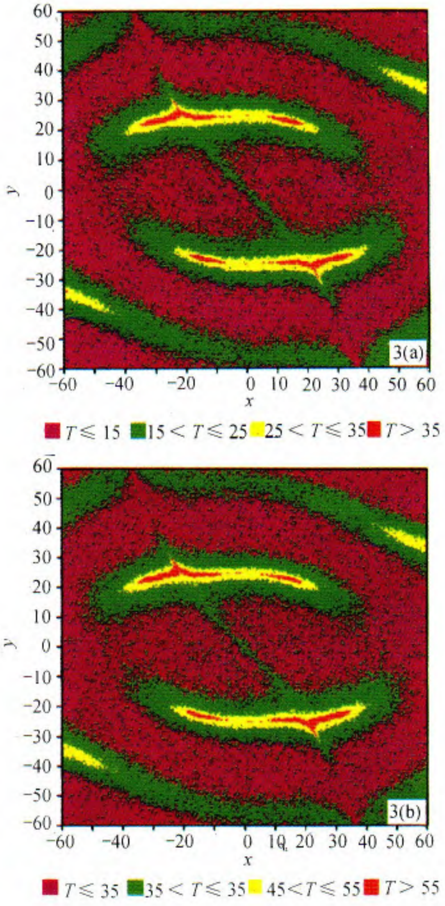

Plate I-3 Distribution of MECT T on the plane $z = 10$. (a) Single precision; (b) double precision. Open one denotes $(0, 0, 0)$. Here the used numerical method, equations, precision, section and initial conditions are as in Plate I-1, and $r = 28$.