# 非线性常微分方程的计算不确定性原理*
## ——Ⅰ. 数值结果

李建平 曾庆存
(中国科学院大气物理研究所大气科学和地球流体力学数值模拟国家重点实验室, 北京 100029)
丑纪范
(北京气象学院, 北京 100081)

**摘要** 在大多数解初值问题的长时间数值积分计算中很少考虑由于机器的有限精度所导致的舍入误差. 利用 29 种标准的数值方法, 通过大量的数值试验深入考察了舍入误差的影响, 发现在有限的机器精度下数值求解非线性常微分方程初值问题存在对机器精度强的依赖性(是与对初值敏感依赖性不同的一种新的依赖性)提出一种计算有限精度下数值方法的最大有效计算时间和最优步长的最优搜索法, 得到最大有效计算时间和最优步长与数值方法的阶数及机器精度之间的关系. 数值分析结果表明舍入误差在上述现象中起着至关重要的作用, 并发现两个与方程、初值及数值格式无关的普适关系. 根据数值试验结果提出一个计算不确定性原理, 这个原理对于非线性常微分方程的长时间数值积分的可靠性提出了挑战.

关键词 常微分方程 计算不确定性原理 舍入误差 离散化误差 非线性

大多数常微分方程是不能用初等函数求解的, 所以需要用数值的方法来求其近似解[1~3]. 用任一离散变量法求一个初值问题的近似解时, 有两个基本的误差来源: 离散误差和舍入误差. 离散误差是由于对微分方程离散化而产生的[1,2], 舍入误差是由于计算机机器字长是有限的这一固有属性所造成的[1,4,5]. 对给定的初值问题, 在假定没有舍入误差的情况下, 因为所有标准的离散变量法都是收敛的, 所以在实际数值求解时舍入误差也就不受重视. 然而, 这并不意味着舍入误差是一个不重要的问题, 事实恰恰相反. 因为计算机的有限精度性, 在实际计算中舍入误差不可避免的存在, 所以在非线性方程的长时间数值积分中就可能改变真解的根本性质. Henrici[1] 利用概率论详细地考察了定点机上的舍入误差, 并以线性方程给以说明, 但他并未注意舍入误差对长时间数值积分的影响. 实际上, 现代计算机普遍采用浮点运算, 同时要解决的模型常常是非线性方程. 因此, 本工作用数值试验和理论分析来研究浮点机

2000-02-25 收稿, 2000-04-25 收修改稿
* 国家重点基础研究发展规划项目、国家自然科学基金资助项目(批准号: 49805006, 49905007)、优秀国家重点实验室资助项目(批准号: 49823002)和中国科学院资源环境领域知识创新工程重要方向项目及中国科学院大气物理研究所创新项目

舍入误差对非线性常微分方程数值计算的重要影响. 结果表明由于机器精度的有限性, 用数值方法求解非线性常微分方程的初值问题时会出现严重的问题, 在有限时间的数值积分后, 任何步长的数值解都与真解无关, 从而导致我们提出计算不确定性原理.

## 1 数值模型和数值方法

本文的数值试验模型是 Lorenz 方程[6],

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

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

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

其中 $\sigma = 10$, $b = 8/3$, $0 < r < +\infty$, 此方程初值问题的解存在且惟一. 这里之所以选择这个方程, 因为它具有很好的代表性. 当 $1 < r < 24.74$ 时, Lorenz 方程有两个稳定的不动点, $C(\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1)$, $C'(-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1)$ 和一个不稳定的不动点 $O = (0, 0, 0)$. 当 $r > 24.74$ 时, $C$ 和 $C'$ 都变成不稳定的, 此时, 存在混沌和一个奇怪吸引子[6~8].

本文采用的标准数值方法共 4 类 29 种[1~4,9]: (ⅰ) 显式单步法: Euler 法, 2 到 6 阶 Runge-Kutta (RK) 型算法, 2 到 10 阶 Taylor 级数法; (ⅱ) 显式多步法: 2 到 6 阶 Adams 算法; (ⅲ) 隐式法: 隐式 Euler 法, 2 阶到 6 阶隐式 Adams 法; (ⅳ) 预测校正法: 2 阶修正的梯形预测校正法, 4 阶修正的 Adams 及 4 阶修正的 Hamming 预测校正法 (PMECME). 全部计算是在 SGI ORIGION 2000 (单精度和双精度分别具有 7, 16 位有效数字) 上完成的.

## 2 无混沌情形

考虑 $1 < r < 24.06$ 的情形, 此时 Lorenz 方程没有混沌现象[6~8]. 然而数值计算会出现奇异的现象. 以经典四阶 RK 方法为例, 同一初值 (5, 5, 10) 用两个相差极其微小的步长积分却可以得到本质上完全不同的两个终态 (图 1(a) ~(c)). 这说明数值解的不惟一性和理论解的惟一性之间产生了矛盾. 更重要的是这并不是个别步长的现象. 正如图 2 所示, 无论是用单精度还是双精度, 数值计算的终态对时间步长非常敏感, 步长的微小变化将导致数值解在本质上的重大差异, 出现截然相反的结果. 很明显, 步长集合分为两类: 一类 (记为 $A$) 对应于终态 $C$, 另一类 (记为 $B$) 对应于 $C'$ (图 2), 即如果一个步长属于 $A$ (或 $B$), 那么用这个步长所得的数值解的终态是 $C$ (或 $C'$). 令人惊奇的是这两类步长集合明显呈现出 Cantor 集[10] 的结构 (图 2(a) ~(c)、(d) 和 (f)). 这表明, 不同步长的数值结果表现出明显的随机性特征. 如果我们考察属于集合 $A$ 的步长个数的百分率, 那么随着所用步长总数的增加这一百分率将接近于 50% (图 3). 上述现象完全在意料之外. 用其他 28 种数值方法都有相同的现象[1] (图略). 因此, 对固定的步长, 数值方法不仅会产生伪解[11], 而且更严重的是对这个初值数值方法无法告诉我们哪个步长的解是真解. 这说明, 像 (5, 5, 10) 这样初值的终态, 在给定的机器精度下是这些数值算法算不准的, 这类初值称为坏的初值. 这些坏的初值组成的点集称为坏的初始点集.

1) 李建平. 微分方程数值积分中的算不准原理及两个普适关系. 中国科学院大气物理研究所博士后研究工作报告, 1999. 174

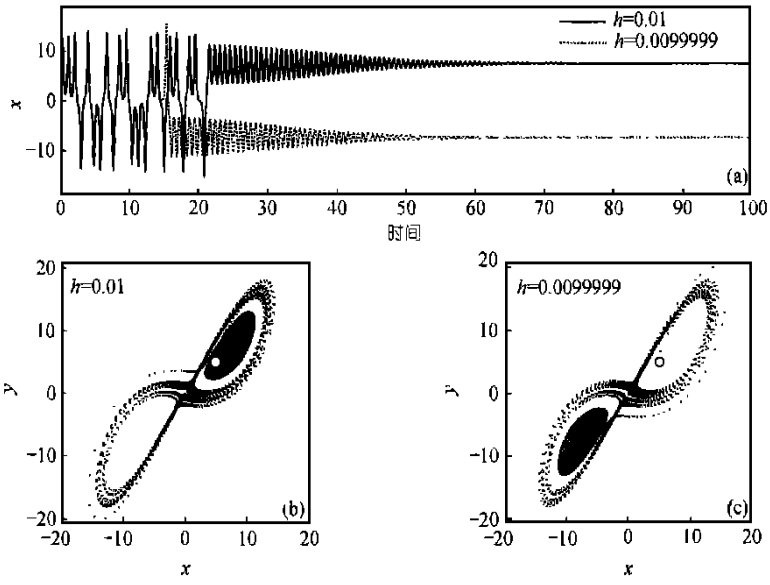



图 1　四阶 RK 方法计算的初值为(5, 5, 10) 和 $r=22$ 的 Lorenz 方程的数值解

(a) 用相差极微小的两个步长所得的 $x$-分量的解; (b) $x$-$y$ 平面上的投影, 步长 $h=0.01$, 图中空心圆是初始点的位置; (c) 同(b), 但步长 $h=0.009\ 999\ 9$

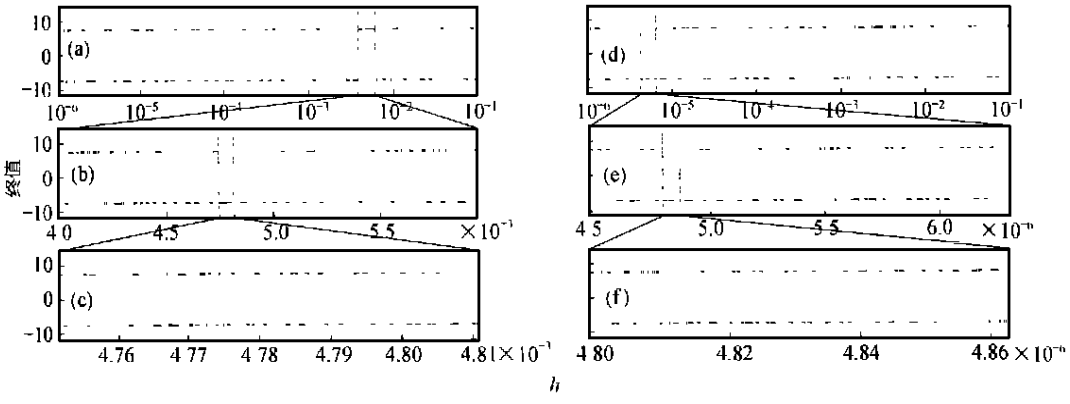



图 2　单精度(a) ~ (c) 和双精度(d)、(f) 时四阶 RK 方法计算得到的 Lorenz 方程 $x$ 分量的终值随步长 $h$ 的变化

步长范围分别为(a) $10^{-6}\sim10^{-1}$, (b) $3.981\ 070\times10^{-2}\sim5.956\ 62\times10^{-2}$, (c) $4.751\ 533\times10^{-2}\sim4.810\ 801\times10^{-2}$, (d) $10^{-6}\sim10^{-1}$, (e) $4.668\ 34\times10^{-6}\sim6.309\ 573\times10^{-6}$, (f) $4.798\ 529\times10^{-6}\sim4.863\ 025\times10^{-6}$. 终值就是终态, 即随着时间的增长解到达并固定在那里. 所用参数 $r$ 和初值同图 1

相应地, 则有好的初值和好的初始点集的概念. 如图 4(a) 和(b) 所示, 好的初值有两种情形.

为了解释数值计算所出现的上述现象, 我们对另一组初值(3, 4, 10) 进行了试验, 对四阶 RK 方法结果(图 5) 表明这个初值在单精度下是坏的初值, 而在双精度下却是好的初值. 对于其他 28 种数值方法也是如此(图略). 这个事实证实机器精度的有限性所导致的舍入误差是造成上述现象的关键因素. 而且, 它表明一个初值好坏的品质是依赖于机器精度的. 一个初

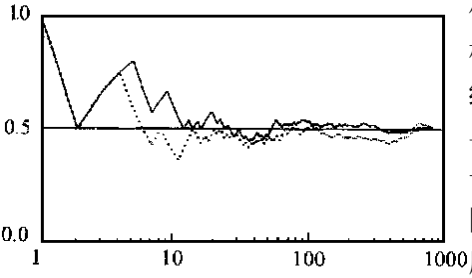



图 3 属于集合 $A$ 的步长百分率随所用步长总数的变化

实线和点线分别代表单精度和双精度的结果. 这里所用数值方法、方程组、$r$ 和初值同图 1, 所有步长属于[ $10^{-6}$, 0. 1]

值在较低的机器精度下是坏的初值, 但它在较高的机器精度下却可以是一个好的初值. 这揭示出数值结果有一种对机器精度强的依赖性, 而这种依赖性与对初值条件的敏感依赖性有本质区别, 原因有二: 一是, 这里讨论的是无混沌情形; 另一个更重要的原因是对初值的敏感依赖性本质上不依赖于机器精度. 因此, 对机器精度强的依赖性是一种新的现象, 在数值计算和模拟中应给予充分的重视.

通常, 在无混沌情形, 给定机器精度下的一个好初值的解可以由数值方法准确的计算. 然而, 由图 4 ( c) 和( d) 可知, 这里所谓的准确性也仅仅是在可接受水平意义上而言的. 此外, 图 4( c) 、( d) 还指出数值

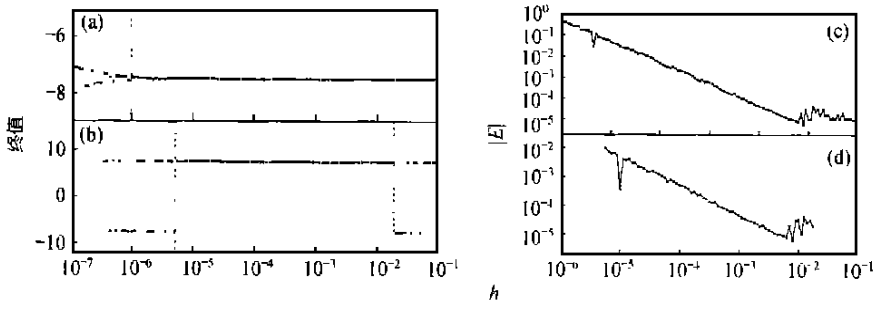



图 4 好初值的两种类型( a) 和( b) 及其计算正确的 $x$ 分量终值的绝对误差 $E$( c) 和( d)

( a) 初值为( 0, 1, 0) 的 $x$-分量的终值随步长的变化. 计算的终值基本上不随步长( 在 $10^{-6}$到 0. 1 之间) 而改变; ( b) 同( a) , 但对初值( 5, 1, 19. 5) . 计算的终值对较大和较小的步长是敏感的, 但对适中的步长是不变的. 这里所用数值方法、方程及 $r$ 同图 1, 且使用单精度. ( c) 和( d) 分别与( a) 和( b) 对应

解的误差不是随着步长的减小而减小, 而是当步长小于某一临界步长时, 随着步长的减小而增加. 这个临界步长与最小误差相对应, 正如下文( 及本文的第 2 部分——理论分析) 所指出的它称为最优步长.

一个重要的问题就是坏的初值点集的 Lebesgue 测度是否为零. 如果在给定的机器精度下一数值算法坏的初值点集的 Lebesgue 测度为零, 则说明在实际计算中碰到坏的初值是零概率事件, 该数值算法就是成功的算法. 否则称该算法在给定精度下不是绝对成功的. 根据四阶 RK 方法对 580, 810 个初值的试验结果( 表 1) , 对 Lorenz 方程没有混沌现象的情况, 坏的初值是很多的, 在 $r=23$ 时占有大多数, 随着 $r$ 的进一步增大, 坏的初值所占比例还会明显的增加. 令人惊奇的是所谓算不准的坏初值并不是通常直觉上所认为的只是在定常吸引子 $C$ 和 $C'$ 的吸引域的交界处( 图版

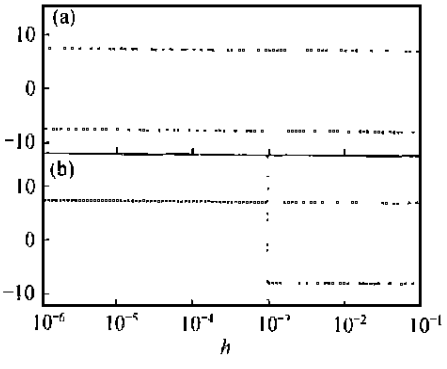



图 5 初值( 3, 1, 10) 的 $x$ 分量的终值随步长的变化

( a) 单精度的结果. 计算的终值对步长非常敏感, 表明在单精度下这是一个坏的初值; ( b) 双精度的结果. 计算的终值对较大和较小的步长敏感, 但对适中的步长却是一个近似常数, 表明在双精度下这是一个好的初值. 这里所用数值方法、方程及 $r$ 同图 1

Ⅰ-1, 附本刊后, 下同)．这些结果表明在这些条件下四阶 RK 方法不是绝对成功的算法．此外, 双精度下坏的初值比单精度的要少, 说明提高机器精度, 坏的初值会有一定的减少．这再次证实舍入误差在数值计算中有着十分重要的影响．以上结论同样适用于其他 28 种数值方法[1)]．

表 1　由四阶 RK 法得到的 Lorenz 方程在几个剖面上好初值和坏初值个数的百分比(%) [a)]

| | 比率 | 平面 $z=10$ | 平面 $z=r-1$ | 平面 $y=1$ | 平面 $y=10$ |
|---|---|---|---|---|---|
| $r=22$ | $p_1$ | 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 |

a) 括号中的数字是双精度的结果, 其余为单精度的结果．$p_1$, $p_2$ 分别表示终态是 $C$, $C'$ 的好初值的比率, $p_3$ 是坏初值的比率．$z$ 平面中的区域为 $-60\leqslant x\leqslant 60$, $-60\leqslant y\leqslant 60$, $y$ 平面中的区域为 $-60\leqslant x\leqslant 60$, $-60\leqslant z\leqslant 60$, 其中每个区域分成 240 × 240 个网格并以网点(共 241×241=58 081 个)为初值条件

## 3　混沌情形

现在来研究 $r>24.74$ 的情形．这时 Lorenz 方程有混沌现象[6~8], 存在奇怪吸引子, 没有稳定的定态．为了展示混沌情形下数值计算所出现的现象, 我们对不同的初值进行了大量的数值试验, 得到很多数值解的步长-时间演变图．这种图可以非常清晰地显示同一初值用不同的步长所得数值解随时间的演变及其差别．图版Ⅰ-2 是对初值(5, 5, 10)和 $r=28$ 用四阶 RK 方法通过上百种不同步长在单精度下积分得到的数值解随时间变化的步长-时间图．如图所示, 一开始数值解的等值线非常平直, 表明各步长所得的解都比较一致, 但很快较大步长的区域和较小步长区域的等值线开始出现波动, 说明这些步长的解与其他步长的解开始偏离了一致性, 保持一致的步长区间(称为有效步长区间)的宽度开始减小．随着向前积分, 有效步长区间的宽度变得越来越小．终于, 积分时间大约在 17 的时候, 有效步长区间的宽度变成零．超过这个时间, 等值线变得杂乱无章, 所有步长的解之间都失去了一致性, 结果, 所有的数值解都与真解无关．所以, 有效步长区间的宽度变成零的时间称为最大有效计算时间, 而对应于最大有效计算时间的步长称为最优步长．在有效步长区间内不同步长所得的数值解的差异很小, 所以有效步长区间的解把微分方程的解相当好的表示出来．然而, 不幸的是, 对于混沌系统(甚至对于具有暂态过程的系统), 随着积分时间的增加有效步长区间的宽度逐渐减小并很快变成零, 达到最大有效计算时间．超过最大有效计算时间, 所有步长的解之间都有显著的差异, 这时数值解完全失去意义．对于双精度的计算结果(图版Ⅰ-3), 亦有同样的现象, 只是积分时间在大约 35 的时候达到最大有效计算时间, 比单精度时长近两倍, 但最优步长约是单精度时的 1/60．用其他 28 种算法也是如此．上述结果说明：(ⅰ) 由于计算的有限精度, 数值解不会随着时间步长 $h\to 0$ 而收敛到准确解; (ⅱ) 存在最优步长和最大有效计算时间, 超过最大有效计算时间, 微分方程的解是数值方法无法算准的; (ⅲ) 提高机器精度能有效的延缓但不能消除舍入误差的影响.

1) 见 404 页脚注 1)

显然, 在图版Ⅰ-1 和 2 中存在一条廓线, 它是每个步长的有效计算时间的连线, 称为有效计算时间廓线. 有效计算时间廓线把数值解的步长-时间图分成两部分. 在有效计算时间廓线的左侧, 数值解的等值线平直规则, 称为数值层流区, 其中的数值解是把微分方程的真解较好地表现出来; 在有效计算时间廓线的右侧, 数值解的等值线杂乱无章, 漫无规则, 称为数值湍流区, 其中的数值解是虚假的, 与微分方程的准确解无关.

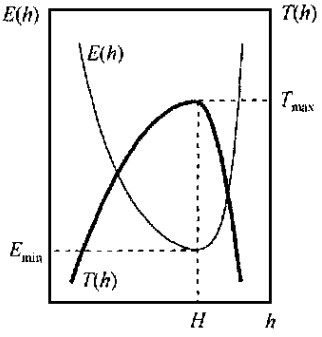



图 6 误差 $E(h)$ 和有效计算时间 $T(h)$ 随步长的变化
图中 $H$ 是最优步长, $E_{min}$ 代表最小误差, $T_{max}$ 代表最大有效计算时间

应当指出, 上述试验所揭示的现象的价值不在于显示对初值条件的敏感依赖性, 而在于确定在有限机器精度下一个数值解的最大有效时间. 根据上述结果(并参考图 4(c)和(d)), 对于计算误差和有效计算时间廓线可以得到如下定性结论(图 6): 一开始, 当步长减小时, 方法误差减小, 总的误差也减小, 使得有效计算时间增加; 然而当步长减小到一定程度时, 由于迭代的步数增多, 机器所带来的舍入误差占主导地位, 从而总误差又开始增大, 有效计算时间开始减小. 这样就必然存在一个步长 $H$(图 6), 此时总误差最小, 有效计算时间最大, 因此这个步长就称为最优步长. 由于方法误差(离散化误差)和舍入误差随步长的这种反向变化关系, 就导致了与量子力学中 Heisenberg 测不准关系[12] 相类似的计算不确定性原理. 具体地说, 如果把离散误差和舍入误差看作是"共轭"量, 那么计算不确定性原理意味着, 若其中之一的不确定性越小, 则它的"共轭"量的不确定性就越大. 因此, 当机器精度给定时, 数值方法所得数值解所能达到的最好准确度就已完全确定. 计算不确定性原理对于对初值极端敏感的混沌系统和一些具有暂态混沌过程的非线性系统的长时间数值积分的计算有效时间加上了确定的限制. 这正是本文数值试验所观察到各种现象的根本. 在机器精度是有限的这一固有属性下, 计算不确定性原理表明, 对于非线性系统, 数值算法的计算能力是有限的.

为了深入研究最大有效计算时间和最优步长与数值方法的阶数、机器精度及初值之间的内在关系, 就必须给出计算有效计算时间廓线的有效算法. 本文给出一种称之为最优搜索的方法来获得有效计算时间廓线. 这个方法是基于判断多个数值解的差异大小来实现的. 在步长区间[$h_{min}$, $h_{max}$] ($h_{min}$为很小的数, 本文除二阶显式 Adams 法和二阶 Talyor 级数法在双精度下取 $h_{min}=0.3\times10^{-7}$, 其余均取 $h_{min}=10^{-7}$) 选取 $n$ 个步长 $h_i$($i=1, \cdots, n$) (步长的选取要比较均匀且 $n$ 较大). 这 $n$ 个步长同时积分到 $t$ 时的 $n$ 个数值解为 $\tilde{y}_t(n)$, 如果它们的差异 $V_s(t)$ (可用标准差来度量) 较小, 具体地说小于预先给定的容许限 $\delta$, 则说明它们比较接近, 并把真解在 $t$ 时刻的数值相当好的再现出来. 此时有效步长区间为[$h_{min}$, $h_{max}$], 宽度为 $W_h(t)=\lg h_{max}-\lg h_{min}$. 否则, 如果 $V_s(t)$ 大于 $\delta$, 则表明有某些步长的数值解出现较大的偏差, 所以从 $\tilde{y}_t(n)$ 中剔除这些解以使得剩下的 $n_1$($n_1<n$) 个解的差异小于 $\delta$, 以剔除的个数最少为最优. 这样可得到满足给定条件的有效步长区间. 以有效步长区间内的步长继续向前积分, 不断重复上述过程, 直至积分到 $t_1$ 时只剩下相邻两个步长 $h_j(t_1)$, $h_{j+1}(t_1)$, 然后, 将步长区间[$h_j(t_1)$, $h_{j+1}(t_1)$] 划分为 $m$ 份, 以这 $m+1$ 个步长重新开始积分, 继续重复上述过程, 直至没有相邻的两个步长所得数值解的差异小于 $\delta$ 时为止. 此时, 可得有效计算时间廓线、最大有效计算时间及最优步长. 为了能对不同的问题进行比较, 容许限 $\delta$ 最好用相对指标度量. 令

从初始时刻 $t_0$ 到时间 $t$ 时微分方程的真解本身的振动为 $V(t)$，那么给出容许限要求 $V_s(t)/V(t) \leqslant \delta$. 因为微分方程的真解一般是未知的，所以实际中用有效步长区间所得的从初始时刻 $t_0$ 到时间 $t$ 时的数值解的振动 $V^*(t)$ 来代替 $V(t)$. 本文取容许限 $\delta$ 为 1/10. 图版Ⅰ-4 是根据最优搜索法得到的图版Ⅰ-1 和 2 中的有效计算时间廓线. 此方法很好地把有效计算廓线计算出来，这是明显的.

利用最优搜索法用四阶 RK 方法对 116, 160 个初值进行了数值试验，图版Ⅰ-5 和 6 给出单精度和双精度下的结果. 这些试验指出不同的初值其最大有效计算时间是不同的. 平均而言，最大有效计算时间在单精度和双精度下分别为 16.857 和 35.412. 显然，减少舍入误差能有效的增加最大有效计算时间. 通过比较图版Ⅰ-5 和 6，容易发现它们彼此在型式上是极为相似的，且双精度和单精度下最大有效计算时间的差似乎是一个固定的数(就此种情形而言约在 18～19 之间). 这说明虽然在某一机器精度下最大有效计算时间是与初值有关的，但是在两种机器精度之间最大有效计算时间的差是不依赖于初值的. 事实上，这个结果将被本文的第 2 部分的理论分析所证实.

比较图 6 和图版Ⅰ-5 和 6，凡是在图 6 中(没有混沌时)被称为好的初值，在图版Ⅰ-5 和 6 中(混沌时)其有效计算时间相对较长. 对于混沌系统，好初值的 Lebesgue 测度为零，即在有限机器精度下，几乎所有初值的最大有效计算时间都是有限的.

通过对 $r=28$ 时 Lorenz 方程的 10 组初值在不同精度、不同阶数的方法下所得平均最大有效计算时间和平均最优步长的比较(图 7)，得到最大有效计算时间和最优步长与数值方法阶数及机器精度有如下关系：(ⅰ) 最大有效计算时间和最优步长均随数值方法阶数的增大而增加，但增加的幅度却逐渐减少(图 7(a)～(c)). 此外，RK 型、Taylor 级数型以及隐式 Adams 型

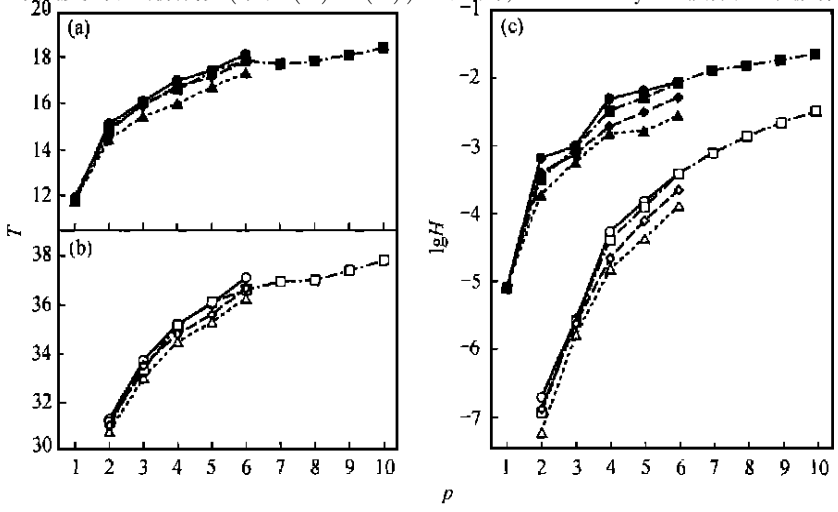



图 7　最大有效计算时间和最优步长随数值方法阶数的变化

(a) $r=28$ 时单精度下 Lorenz 方程 10 个初值的平均最大有效计算时间 $T$. 10 个初值为：(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)和(－6, －7, －8)，实心圆、方块、三角和菱形分别代表 RK 型、Taylor 级数型、显式 Adams 型和隐式 Adams 型方法；(b) 同(a)，除了用双精度和空心符号；(c) 同(a)和(b)，但对平均最优步长 $H$

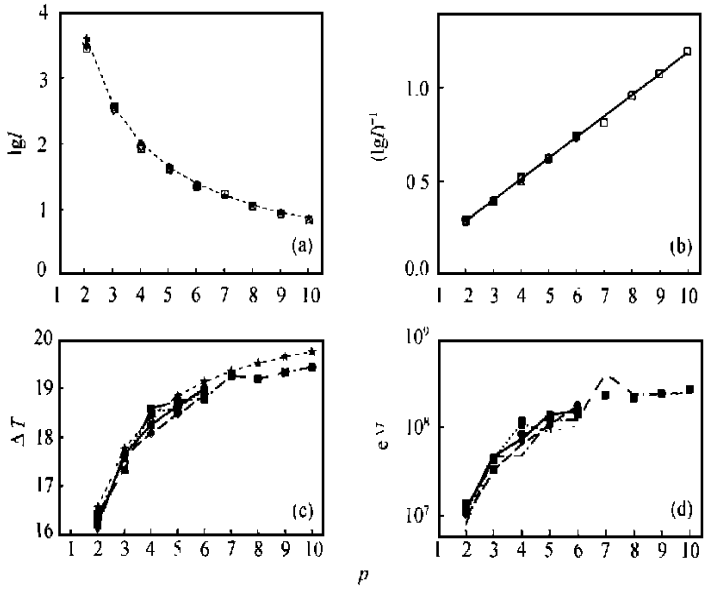



图 8 单精度和双精度下最优步长之比 $l$(a)、(b) 和最大有效计算时间之差 $\Delta T$(c)、(d) 随方法阶数的变化

(a) 同图 7(a), 但对平均 $l$ 值. 星号是根据(5) 式得到, 其余符号同图 7(b); (b) 同(a), 但对$(\lg l)^{-1}$; (c) 星号由(7) 式所得, 其余符号及 10 组初值同图 7(a); (d) $e^{\Delta T}$与 $l^p$ 关系, 符号同(c), 代表$e^{\Delta T}$; 实线、点划线、短虚线及长虚线分别是(a) 中 RK 方法、Taylor 级数法、显式 Adams 法和隐式 Adams 法的 $l$ 的 $p$ 次幂

方法的最大有效计算时间和最优步长均比同阶的显式 Adams 型方法略大一点; (ⅱ) 对于同阶的数值方法, 双精度下的最大有效计算时间比单精度的长很多(图 7(a) 和(b)), 而最优步长却比单精度的小很多(图 7(c)). 可见, 提高机器精度会使最大有效计算时间明显地增大, 但相应的最优步长却要缩小很多.

由以上分析可知, 同阶方法在不同的机器精度下的最优步长和最大有效计算时间显然是不同的, 然而, 同阶的方法在给定的两种机器精度下的两个最优步长或两个最大有效计算时间之间是否存在某种确定的联系呢? 为了回答这个问题, 首先定义如下指标来刻画两种机器精度下最优步长之间的关系:

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

其中 $H_1$, $H_2$ 分别是用同一种 $p$ 阶的数值方法在给定的具有 $n_1$, $n_2$ 位有效数字的两种机器精度 $\gamma_1 = 5\times 10^{-n_1}$, $\gamma_2 = 5\times 10^{-n_2}$ 下的最优步长. 为讨论方便, 以下令 $n_1 \leqslant n_2$(本文中 $n_1 = 7$, $n_2 = 16$). 图 8(a) 是根据图 7 中 10 组初值得到的 $l$ 值随方法阶数的变化. 由图可知, 4 种方法的 $l$ 值是非常一致的, 这表明尽管最优步长与方法有关, 但两种机器精度下的最优步长的比却与方法无关, 即不同的方法其 $l$ 值遵循相同的规律. 对于不同的初值, $l$ 值的变化规律依然如此. 而且, 对于不同的方程(如方程 $y' = y$, $y(0) = 1$ 和 $y' = -y$, $y(0) = 1$ 等)[1], $l$ 值的变化

1) 见 404 页脚注 1)

规律也不改变．因此，在给定的两种机器精度下最优步长的比 $l$ 满足一个普适关系(在本文第 2 部分中将给予理论证明)，且 $l$ 只与数值方法的阶数 $p$ 和机器精度有关，即 $l=l(p, n_1, n_2)$．根据图 8(a)，$\lg l$ 与 $p$ 有类似反比的关系，如果是这样，那么它的倒数一定是 $p$ 的线性函数．果然，图 8(b) 非常清楚的证明了这个结论．利用最小二乘法可确定出线性关系的系数，结果见表 2 中关系式 1．从统计关系式 1 的右端不难看到，其中的分母与 9 很接近．而本文中 $\Delta n=n_2-n_1=9$．这不是偶然的巧合，在本文第 2 部分的理论分析中将证明有如下关系成立：

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

对本文中的情形($\Delta n=9$)这个关系可以通过表 2 中统计关系式 2 得到证实．由图 8(a) 知，按这个关系所得到的 $l$ 值与试验值之间是极为一致的．所以，只要知道了某一机器精度下的最优步长，那么由上述关系立得任何机器精度下的最优步长．

表 2　试验的 $l$ 值与 $p$ 的统计关系

| 方法 | 统计关系式 1 | 统计关系式 2 |
|---|---|---|
| Runge-Kutta 方法 | $(\lg l)^{-1}=0.112\,494\,7p+0.056\,175\,4=\frac{p+0.4993607}{8.889\,310\,8}$ | $(\lg l)^{-1}\approx\frac{1}{9}(p+0.555\,386\,7)$ |
| Taylor 级数法 | $(\lg l)^{-1}=0.112\,041\,8p+0.061\,110\,0=\frac{p+0.545\,421\,4}{8.925\,237\,7}$ | $(\lg l)^{-1}\approx\frac{1}{9}(p+0.600\,249\,2)$ |
| 显式 Adams 法 | $(\lg l)^{-1}=0.116\,258\,7p+0.045\,283\,2=\frac{p+0.389\,503\,7}{8.601\,507\,2}$ | $(\lg l)^{-1}\approx\frac{1}{9}(p+0.592\,862\,0)$ |
| 隐式 Adams 法 | $(\lg l)^{-1}=0.110\,767\,1p+0.066\,947\,5=\frac{p+0.604\,399\,2}{9.027\,955\,1}$ | $(\lg l)^{-1}\approx\frac{1}{9}(p+0.590\,141\,7)$ |

对于最大有效计算时间，定义如下指标：

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

其中 $T_1$, $T_2$ 分别是用同一种 $p$ 阶的数值方法在给定的具有 $n_1$, $n_2$($n_1\leqslant n_2$) 位有效数字的两种机器精度下的最大有效计算时间．图 8(c) 是根据图 7(a) 和(b) 中的结果而得到的 $\Delta T$ 随方法阶数的变化．由图可知，4 种方法的 $\Delta T$ 也是比较一致的，随着阶数的增加，双精度和单精度下的最大有效计算时间的差 $\Delta T$ 趋近于一定值．$\Delta T$ 是否也存在一个普遍的规律？它与 $l$ 是否有某种必然的联系？图 8(d) 回答了这个问题．由图可知，$e^{\Delta T}$ 与 $l^p$ 有着很好的对应关系，就是说存在如下关系：

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

从图 8(c) 可知，由这个关系所得到的 $\Delta T$ 值与试验值之间也是一致的．所以，根据这个关系式，只要知道了某一机器精度下的最大有效计算时间，就可得任何机器精度下的最大有效计算时间．当 $p\to\infty$ 时，$\Delta T\to\Delta n\ln 10$；对于本文中的两种精度，有 $\Delta T\to 9\ln 10\approx 20.723\,3$($p\to\infty$)．

## 4　结论

虽然舍入误差很小，但在非线性常微分方程的长时间数值积分中有着不可忽视的作用．我们通过大量的数值试验，发现在无混沌情形数值解有对机器精度强的依赖性，这种依赖性与对初值条件的敏感依赖性有本质不同；在混沌情形，在有限的机器精度下数值方法存在最大有效计算时间和最优步长，超过最大有效计算时间，微分方程的解是数值方法无法算准的．而

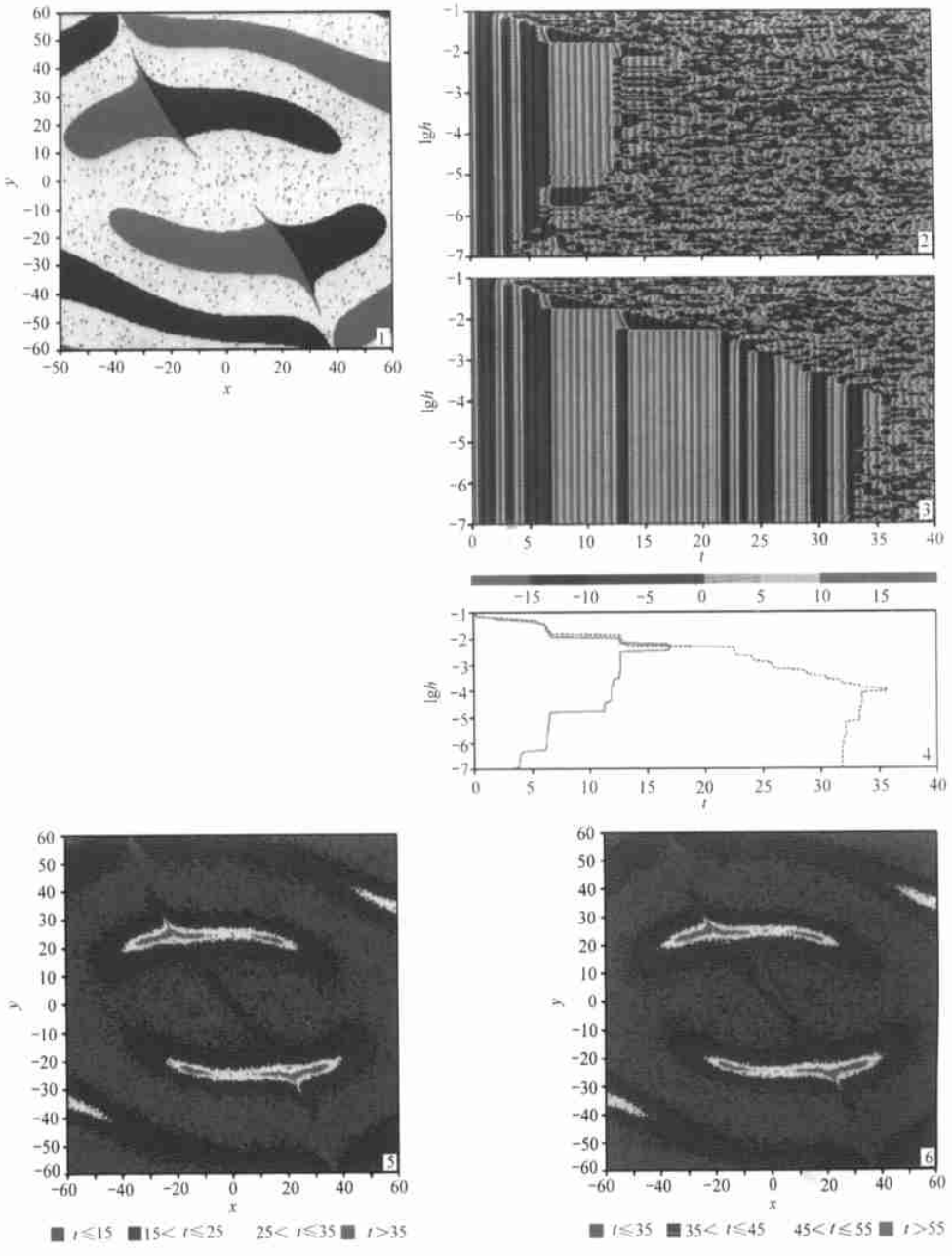
且,本文提出一种能很好计算出有限精度下数值方法的最大有效计算时间和最优步长的最优搜索法,利用该法得到最大有效计算时间和最优步长与数值方法阶数及机器精度之间的本质关系. 同时发现两个普适关系,这两个关系揭示出任何两种机器精度下的两个最优步长之间和两个最大有效计算时间之间的本质联系. 根据这两个关系,只要知道某一机器精度下数值方法的最优步长和最大有效计算时间,就可以推得任何机器精度下的最优步长和最大有效计算时间. 在有限机器精度下数值方法之所以存在最大有效计算时间,一方面是由于差分方法求解的稳定性条件要求时间步长有一个上界限制,而另一方面是实际执行这种计算的计算机又有精度限制从而为使误差不致累积过大而必然对计算步数带来一个上界限制. 这两者是相互矛盾的,其结果就得相互补充而得到计算不确定性原理. 由于篇幅所限,这里仅以 Lorenz 方程为例进行了试验,其实我们还有其他微分方程的例子[1)](包括线性方程). 事实上,根据本文第 2 部分的理论分析可知这里的结果同样适用于其他常微分方程. 本文的结果甚至也适用于偏微分方程,尽管那里还涉及到时空步长的匹配问题. 目前看来要减少舍入误差的影响,只有不断提高机器的精度. 但无论如何,机器精度都不可能是无限的. 所以,计算不确定性原理对于非线性微分方程的长时间数值积分的有效性和可靠性以及如何改变现有计算方式以有效地延长最大有效计算时间提出了挑战.

**致谢** 本文的部分内容在北京计算流体力学 1998 年秋季讨论会和 1999 年春季讨论会上做过报告,得到北京应用物理与计算数学研究所李德元、蔚喜军研究员及中国科学院计算数学所邬华谟研究员的有益讨论和指教;此外,中国科学院大气物理研究所季仲贞、穆穆及王斌研究员与本文第一作者进行过有益讨论并提出宝贵意见,在此一并表示感谢.

## 参 考 文 献

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: Prentice-Hall, 1971. 1~14, 72~86
3 Hairer E, Nþrsett 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, 1998. 1~36, 428~569
5 Sterbenz P H. Floating Point Computation. Englewood Cliffs: Prentice-Hall, 1974
6 Lorenz E N. Deterministic nonperiodic flow. J Atmos Sci, 1963, 20(130): 130~141
7 Kaplan J L, Yorke J A. Chaotic behavior of multidimensional difference equations. In: Peitgen H O, Walther H O, ed. Functional Differential Equations and Approximation of Fixed Points. 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: 545~591
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: 1 723~1 751
12 Heisenberg W. The Physical Principles of Quantum Theory. Chicago: University of Chicago Press, 1930

1) 见 404 页脚注 1)




1．平面 $z=10$ 上好和坏初值的分布，此区域分成 $240\times240$ 个网格并以网点（共 $241\times241=58\,081$ 个）为初值条件，蓝和红方块分别代表终态是 $C$，$C'$ 的好的初值，黄方块代表坏的初值，空心方块代表(0,0,0)点，这里所用数值方法、方程和 $r$ 同图1，且用单精度；2．单精度下用121种步长所得的初值(5,5,10)和 $r=28$ 的 $x$ 分量的步长-时间图．这里 $h$ 取常用对数，时间是无量纲的，所用数值方法和方程同图1；3．同2，但用双精度；4．由最优搜索法所得的2和3的有效计算时间廓线，其中实线和点线分别是2和3的有效计算时间廓线；5，6．平面 $z=10$ 上最大有效计算时间 $T$ 的分布，5为单精度，6为双精度，空心方块代表(0,0,0)点，这里所用数值方法、方程、区域及初始条件同图8，且 $r=28$。