王鹏飞，黄荣辉，李建平. 2011. 数值积分过程中截断误差和舍入误差的分离方法及其效果检验[J]. 大气科学，35 (3)：403－410. Wang Pengfei, Huang Ronghui, Li Jianping. 2011. Separation of truncation error and round-off error in the numerical integration and its validation [J]. Chinese Journal of Atmospheric Sciences (in Chinese), 35 (3): 403－410.

数值积分过程中截断误差和舍入误差的分离方法及其效果检验

王鹏飞[1,2]　黄荣辉[3]　李建平[1]

1 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
2 中国科学院研究生院，北京　100049
3 中国科学院大气物理研究所季风系统研究中心，北京　100190

摘　要　本文讨论数值积分过程中截断误差和舍入误差的分离方法和理论，解析地给出某些数值计算方法的理论截断误差，并以此来分离计算结果中的误差。然后引入参考解的办法，用来分离更为一般的微分方程求解过程中的截断误差和舍入误差。以参考解算法为基础，对一个偏微分方程的数值解进行计算，所得结果与采用理论截断误差得到的结果进行了对比，发现：(1) 当使用迎风差和中央差格式时，理论截断误差和近似截断误差在数值上高度一致，说明了参考解方法的正确性；(2) 对于一阶的波动方程，迎风差和中央差格式的理论截断误差在形式上也具有波动的周期特征，振幅的大小与计算参数有关；(3) 理论截断误差可以适用于任意 t 时刻，而近似截断误差的适用时间范围为一个有限的时间段，不过它可以很容易的获取一般微分方程的截断误差，而不需要复杂的理论推导。

关键词　数值积分　截断误差　舍入误差　参考解

文章编号　1006－9895 (2011) 03－0403－08　　**中图分类号**　P435　　**文献标识码**　A

Separation of Truncation Error and Round-off Error in the Numerical Integration and Its Validation

WANG Pengfei[1, 2], HUANG Ronghui[3], and LI Jianping[1]

1 *State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing*　100029
2 *Graduate University of Chinese Academy of Sciences, Beijing*　100049
3 *Center for Monsoon System Research, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing*　100190

Abstract　The authors propose a method to separate the truncation error and the round-off error from the numerical solution. The analytical truncation error formulas of a partial differential equation are given for the upstream scheme and the centered difference scheme, respectively. The reference solution method is then introduced to separate these two types of errors for more general equations. A scheme based on the reference solution is used to obtain the approximate truncation error. Comparing the results for the upstream scheme and the centered difference scheme, the authors find that: 1) the approximate truncation error is highly consistent with the analytical one. 2) The truncation errors of 1-D wave equations for the two schemes both show wavy periodicities with amplitudes being related to the

收稿日期　2010－05－26，2010－11－19 收修定稿

资助项目　国家自然科学基金资助项目 40730952，国家重点基础研究发展计划项目 2009CB421405、2011CB309704

作者简介　王鹏飞，男，硕士，高级工程师，主要从事数值分析、并行计算、非线性可预报性等方面的研究。E-mail：wpf@mail.iap.ac.cn

parameters of computation. 3) The analytical error is suitable for the analysis of any slice of t, while the approximate one is only suitable for the analysis of a certain time range. However, the approximate error can be more easily obtained for general differential equations without a complex theoretical deduction.

Key words numerical integration, truncation error, round-off error, reference solution

1 引言

根据微分方程的理论，符合 Lipschitz 条件的非线性动力系统局部存在唯一的解。数值计算是一种常用的研究方法，但使用计算模拟来代替物理实验进行研究时，结果不可避免地会受到误差的影响，正如 von Neumann (1960) 所指出的，通常有 4 种误差来源会对模拟结果有影响。这 4 种误差分别为：数学模型的误差、初值的误差、差分格式带来的截断误差（也称为离散化误差）和计算机的舍入误差。

如果不考虑初值误差且假定模型是完美的，那么计算过程的误差将主要由截断误差和舍入误差所构成。数值计算中舍入误差的研究可以追溯到电子计算发明的时期，最早被计算天体运行轨道的天文学家 (Brouwer, 1937) 所注意。von Neumann and Goldstine (1947)、Turing (1948) 对舍入误差的积累进行了最初的研究，Rademacher (1948) 首先研究了数值积分过程中的舍入误差，Wilkinson (1963) 详细地讨论了算术过程中的舍入误差问题，其思想和方法对以后的研究者产生了重要的影响。Henrici (1962, 1963) 的研究表明，微分方程的计算结果受舍入误差的影响可能比一般的代数过程更为严重，而且他引入的一些统计假设成为是研究舍入误差影响的强有力工具。很多使用计算机进行计算研究的学科遇到舍入误差的问题，关于这些研究的更详细的介绍可以参考 Higham (1996) 的专著及其中所列文献。通常认为舍入误差的影响并不是主要的，不会影响计算的主要结果。但是 20 世纪末 Li et al. (2000; 2001) 的研究改变了以往对舍入误差影响重要程度的认识，发现了某些非线性系统即使初值完全准确，由于舍入误差的存在，系统的计算也存在最大有效计算时间。他们首次将舍入误差的影响研究提为“计算不确定性原理”，这无疑是对计算本质的非常重要的新认识。Teixeira et al. (2007) 的工作进一步支持了 Li et al. (2000, 2001) 的发现，他们不仅对 Lorenz 系统进行了时间步长敏感性的研究，而且对准地转模式进行了分析。

舍入误差对计算的影响体现在这几个方面：(1) 舍入误差可能造成计算不稳定，是非线性计算不稳定的一个重要原因；(2) 舍入误差还可能造成计算结果不收敛 (Wang and Li, 2008)，从而使计算结果不确定；(3) 舍入误差可能对数值模式的可预报性产生影响，限制了预报时间（王鹏飞和黄刚，2006）。舍入误差与浮点计算精度密切相关，为了研究误差的变化规律，需要一个可行的计算方案分离截断误差和舍入误差，以便对它们的影响进行细致的研究。

2 误差合成公式

为了研究计算中的误差，首先定义一些变量符号。我们以 A 代表一个广义的微分方程（既可以是常微分方程，也可以是偏微分方程）的解析解，D 表示无舍入误差影响时的数值解，N 表示既有截断误差又有舍入误差影响的数值解。E 表示总误差，E_t 表示截断误差，E_r 表示舍入误差。

不同的研究者定义误差的形式稍有差别，如 Anderson (1995) 使用如下公式来定义各种误差：

$$\begin{cases} E_t = A - D, \\ E_r = N - D, \\ E = A - N, \end{cases}$$

因此，E、E_t、E_r 之间存在关系：$E=E_t-E_r$。

周毅等 (2003) 采用的是以准确解做减数的形式来定义误差，有

$$\begin{cases} E_t = A - D, \\ E_r = D - N, \\ E = A - N, \end{cases}$$

E、E_t、E_r 之间的关系为：$E=E_t+E_r$。

在本文的研究中采用加法形式，但以准确解作为被减数来定义各种误差，

$$\begin{cases} E_t = D - A, \\ E_r = N - D, \\ E = N - A, \end{cases}$$

则 E、E_t、E_r 之间的关系为

$$E = E_t + E_r. \tag{1}$$

称公式 (1) 为误差合成公式。这样定义的优点是：几何意义明显。图 1 展示了一个二元微分方程

$$\begin{cases} \dfrac{dx}{dt} = -ay, \\ \dfrac{dy}{dt} = bx, \end{cases} \tag{2}$$

解的分布示意图，其中 a、b 为常数。

参考上面关于广义微分系统的定义，对于变量 X，有

总误差为：

$$E_x = X_r - X, \tag{3}$$

截断误差为：

$$E_{xt} = X_t - X, \tag{4}$$

舍入误差为：

$$E_{xr} = X_r - X_t, \tag{5}$$

误差合成关系为：

$$E_x = E_{xr} + E_{xt}. \tag{6}$$

对变量 Y 也可以定义如上形式的误差。上述各种误差表示方法的差别在于结果的正负号不同，如果我们看重的是误差的绝对值，那么上面各种定义的结果就是完全等价的。

总误差 E 是最终影响计算结果程度的量，其中 E_t 部分由算法决定，随着步长的减小 E_t 的数值逐步减小，但是计算量的增加使舍入误差 E_r 变得越来越大，从而造成总误差的变化不是一致的减小。因此对 E_r 部分的改进和研究越来越重要，且已经成为获得某些方程高精度解的必要条件。

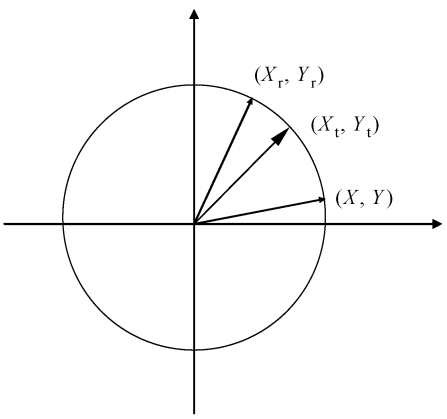

图 1　一个二元微分方程解的分布示意图。(X, Y) 表示解析解；(X_t, Y_t) 表示仅包含截断误差影响的解；(X_r, Y_r) 表示包含舍入误差影响的数值解

Fig. 1　The demonstration of the solution of two-dimensional equation, where (X, Y) is the analytical solution, (X_t, Y_t) is the solution containing truncation error, (X_r, Y_r) is the solution containing truncation error and round-off error simultaneously

3　理论截断误差

因为计算时 N 是可以得到的；对于某些简单的方程，可以得到解析解 A，而且对于这些方程，在使用迎风差等简单的差分格式时，D 也是可以解析计算的。因此，E_t 是能够得到的（为了与第 4 节中由参考解获得的近似截断误差区别，可称为理论截断误差），E 也是能够得到的，所以根据合成公式 (1)，可以得到 E_r。

常微分方程的截断误差公式可以参考 Henrici (1962) 和 Li et al. (2001) 的文章，本文以偏微分方程为例研究算法的截断误差的公式。考虑如下的一阶双曲型方程：

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \tag{7}$$

在定解条件 $u(x,0)=\sin(x)$ 条件下的解。理论分析表明：所有 $u(x,t)=f(x+t)$ 形式的函数都是方程 (7) 的解。由于 $u(x,0)=\sin(x)$，得到 $u(x,0)=f(x+0)=\sin(x)$，即 $f(x)=\sin(x)$，所以解析解为：$u(x,t)=\sin(x+t)$。现将求解区域限定在 $x\in(0,2\pi)$，可以使用差分方法求解。初条件为：$u(x,0)=\sin(x)$，边条件设为周期边界：$u(2\pi,t)=\sin(t)$。

3.1　迎风差格式的截断误差

迎风差格式：

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^n - u_j^n}{\Delta x} = 0 \tag{8}$$

的稳定计算条件为：

$$\frac{\Delta t}{\Delta x} \leqslant 1.$$

将 (8) 式改写为：

$$u_j^{n+1} = u_j^n + \frac{\Delta t(u_{j+1}^n - u_j^n)}{\Delta x},$$

可以对其进行振幅误差、相位误差等分析。$0<\Delta t/\Delta x<1$ 时，利用关系式 $\sin(x)=(e^{ix}-e^{-ix})/2i$，可得：

$$u_j^0 = \tilde{u}_j^0 = \sin(j\Delta x) = \frac{1}{2i}(e^{ij\Delta x} - e^{-ij\Delta x}),$$

其中，u_j^0 为计算解，$\tilde{u}_j^0$ 为解析解。第一步计算解

为：

$$u_j^1 = \left(1-\frac{\Delta t}{\Delta x}+\frac{\Delta t}{\Delta x}e^{i\Delta x}\right)\frac{1}{2i}e^{ij\Delta x} - \left(1-\frac{\Delta t}{\Delta x}+\frac{\Delta t}{\Delta x}e^{-i\Delta x}\right)\frac{1}{2i}e^{-ij\Delta x},$$

第二步计算解为：

$$u_j^2 = \left(1-\frac{\Delta t}{\Delta x}+\frac{\Delta t}{\Delta x}e^{i\Delta x}\right)^2\frac{1}{2i}e^{ij\Delta x} - \left(1-\frac{\Delta t}{\Delta x}+\frac{\Delta t}{\Delta x}e^{-i\Delta x}\right)^2\frac{1}{2i}e^{-ij\Delta x}。$$

第 $n+1$ 步计算解为：

$$u_j^{n+1} = u_j^n+\frac{\Delta t(u_{j+1}^n-u_j^n)}{\Delta x} = \left(1-\frac{\Delta t}{\Delta x}+\frac{\Delta t}{\Delta x}e^{i\Delta x}\right)^{n+1}\frac{1}{2i}e^{ij\Delta x} - \left(1-\frac{\Delta t}{\Delta x}+\frac{\Delta t}{\Delta x}e^{-i\Delta x}\right)^{n+1}\frac{1}{2i}e^{-ij\Delta x}。$$

而第 $n+1$ 步理论值为：

$$\tilde{u}_j^{n+1} = \sin[(n+1)\Delta t+j\Delta x] = \frac{1}{2i}[e^{i(n+1)\Delta t}e^{ij\Delta x}-e^{-i(n+1)\Delta t}e^{-j\Delta x}],$$

它们之间的误差为：

$$E_t^{n+1} = u_j^{n+1}-\tilde{u}_j^{n+1},$$

这个误差即为理论截断误差。

设 $\Delta t/\Delta x\equiv\lambda$，有如下等式成立，

$$1-\frac{\Delta t}{\Delta x}+\frac{\Delta t}{\Delta x}e^{i\Delta x} = \sqrt{1-4(1-\lambda)\lambda\sin^2\left(\frac{\Delta x}{2}\right)}\cdot \exp\left[i\cdot\mathrm{arctg}\frac{\lambda\sin(\Delta x)}{1-\lambda+\lambda\cos(\Delta x)}\right],$$

同理，可得：

$$1-\frac{\Delta t}{\Delta x}+\frac{\Delta t}{\Delta x}e^{-i\Delta x} = \sqrt{1-4(1-\lambda)\lambda\ \sin^2\left(\frac{\Delta x}{2}\right)}\cdot \exp\left[-i\cdot\mathrm{arctg}\frac{\lambda\sin\ (\Delta x)}{1-\lambda+\lambda\cos\ (\Delta x)}\right],$$

所以，

$$u_j^{n+1} = \left[\sqrt{1-4(1-\lambda)\lambda\sin^2\left(\frac{\Delta x}{2}\right)}\right]^{n+1}\cdot \sin\left[(n+1)\mathrm{arctg}\frac{\lambda\sin(\Delta x)}{1-\lambda+\lambda\cos(\Delta x)}+j\Delta x\right],$$

因此，

$$E_t^{n+1} = u_j^{n+1}-\tilde{u}_j^{n+1} = \left[\sqrt{1-4(1-\lambda)\lambda\ \sin^2\left(\frac{\Delta x}{2}\right)}\right]^{n+1}\cdot \sin\left[(n+1)\mathrm{arctg}\frac{\lambda\sin(\Delta x)}{1-\lambda+\lambda\cos(\Delta x)}+j\Delta x\right]- \sin[(n+1)\Delta t+j\Delta x],$$

即

$$E_{tj}^n = \left[\sqrt{1-4(1-\lambda)\lambda\sin^2\left(\frac{\Delta x}{2}\right)}\right]^n\cdot \sin\left[n\cdot\mathrm{arctg}\frac{\lambda\sin(\Delta x)}{1-\lambda+\lambda\cos(\Delta x)}+j\Delta x\right]- \sin(n\Delta t+j\Delta x). \tag{9}$$

公式（9）可以用来计算迎风差的理论截断误差 E_t。

3.2 中央差格式的截断误差

中央差格式（时间、空间都取中央差，也称为蛙跳格式）如下：

$$\frac{u_m^{n+1}-u_m^{n-1}}{2\Delta t}-\frac{u_{m+1}^n-u_{m-1}^n}{2\Delta x} = 0.$$

可改写为

$$u_m^{n+1} = u_m^{n-1}+\frac{\Delta t(u_{m+1}^n-u_{m-1}^n)}{\Delta x},$$

开始时刻，

$$u_m^0 = \tilde{u}_m^0 = \sin(m\Delta x) = \frac{1}{2i}(e^{im\Delta x}-e^{-im\Delta x}),$$

其中，u_m^0 为计算解，$\tilde{u}_m^0$ 为解析解。中央差需要两个初始时间片，第一步用解析给定，

$$u_m^1 = \tilde{u}_m^1 = \sin(\Delta t+m\Delta x) = \frac{1}{2i}(e^{im\Delta x}e^{i\Delta t}-e^{-im\Delta x}e^{-i\Delta t}).$$

设 $\Delta t/\Delta x\equiv\lambda$，有

$$u_m^{n+1} = u_m^{n-1}+\lambda(u_{m+1}^n-u_{m-1}^n),$$
$$u_m^0 = \tilde{u}_m^0 = \sin(m\Delta x) = c_0,$$
$$u_m^1 = \tilde{u}_m^1 = \sin(\Delta t+m\Delta x) = c_1.$$

再设

$$\beta = \frac{\lambda e^{i\Delta x}-\lambda e^{-i\Delta x}}{2i} = \lambda\sin\Delta x,$$

含 $e^{im\Delta x}/2i$ 项的系数为 A_m^n，含 $e^{-im\Delta x}/2i$ 项的系数为 B_m^n，可得：

$$u_m^n = A_m^n\frac{1}{2i}e^{im\Delta x}-B_m^n\frac{1}{2i}e^{-im\Delta x}. \tag{10}$$

计算稳定性条件为：$0\leqslant 1-\beta^2\leqslant 1$，即 $\Delta t/\Delta x\leqslant 1$。

A_m^n 满足如下方程：

$$A_m^{n+1} = A_m^{n-1}+2i\beta A_m^n.$$

由初条件：$A_m^0=1$，$A_m^1=e^{i\Delta t}$。可以求解 A_m^n，得：

$$A_m^n = \frac{1}{2\sqrt{1-\beta^2}}[-A_m^1(\beta i-\sqrt{1-\beta^2})^n+ iA_m^0\beta(\beta i-\sqrt{1-\beta^2})^n+A_m^0\sqrt{1-\beta^2}\cdot$$

$$(\beta\mathrm{i}-\sqrt{1-\beta^2})^n+A_m^1(\beta\mathrm{i}+\sqrt{1-\beta^2})^n-\mathrm{i}A_m^0\beta(\beta\mathrm{i}+\sqrt{1-\beta^2})^n+A_m^0\sqrt{1-\beta^2}\cdot(\beta\mathrm{i}+\sqrt{1-\beta^2})^n],$$

其中第一项：

$$-A_m^1(\beta\mathrm{i}-\sqrt{1-\beta^2})^n+\mathrm{i}A_m^0\beta(\beta\mathrm{i}-\sqrt{1-\beta^2})^n+A_m^0\sqrt{1-\beta^2}(\beta\mathrm{i}-\sqrt{1-\beta^2})^n=(-A_m^1+\mathrm{i}A_m^0\beta+A_m^0\sqrt{1-\beta^2})(-1)^n\cdot\mathrm{e}^{-\mathrm{i}\cdot\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n},$$

第二项：

$$A_m^1(\beta\mathrm{i}+\sqrt{1-\beta^2})^n-\mathrm{i}A_m^0\beta(\beta\mathrm{i}+\sqrt{1-\beta^2})^n+a_m^0\sqrt{1-\beta^2}(\beta\mathrm{i}+\sqrt{1-\beta^2})^n=(A_m^1-\mathrm{i}A_m^0\beta+A_m^0\sqrt{1-\beta^2})\mathrm{e}^{\mathrm{i}\cdot\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n}.$$

因此，

$$A_m^n=\frac{1}{2\sqrt{1-\beta^2}}[(-A_m^1+\mathrm{i}A_m^0\beta+A_m^0\sqrt{1-\beta^2})\cdot(-1)^n\mathrm{e}^{-\mathrm{i}\cdot\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n}+(A_m^1-\mathrm{i}A_m^0\beta+A_m^0\sqrt{1-\beta^2})\mathrm{e}^{\mathrm{i}\cdot\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n}].$$

将 $A_m^0=1$，$A_m^1=\mathrm{e}^{\mathrm{i}\Delta t}$ 代入并化简，有

$$A_m^n=\frac{1}{2\sqrt{1-\beta^2}}[(\sqrt{2-2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t})\cdot(-1)^n\exp\left(-\mathrm{i}\cdot\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n\right)\cdot\exp\left(\mathrm{i}\cdot\mathrm{arctg}\frac{\beta-\sin\Delta t}{\sqrt{1-\beta^2}-\cos\Delta t}\right)+(\sqrt{2+2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t})\cdot\exp\left(\mathrm{i}\cdot\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n\right)\cdot\exp\left(-\mathrm{i}\cdot\mathrm{arctg}\frac{\beta-\sin\Delta t}{\sqrt{1-\beta^2}+\cos\Delta t}\right)].$$

同样，B_m^n 满足如下方程：$B_m^{n+1}=B_m^{n-1}-2\mathrm{i}\beta B_m^n$，由初条件 $B_m^0=1$，$B_m^1=\mathrm{e}^{-\mathrm{i}\Delta t}$，可得：

$$B_m^n=\frac{1}{2\sqrt{1-\beta^2}}[(-B_m^1-\mathrm{i}B_m^0\beta+B_m^0\sqrt{1-\beta^2})\cdot(-1)^n\mathrm{e}^{\mathrm{i}\cdot\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n}+(B_m^1+\mathrm{i}B_m^0\beta+B_m^0\sqrt{1-\beta^2})\cdot\mathrm{e}^{\mathrm{i}\cdot\mathrm{arctg}\frac{-\beta}{\sqrt{1-\beta^2}}\cdot n}],$$

计算步骤同 A_m^n，略去中间结果，最后得到：

$$B_m^n=\frac{1}{2\sqrt{1-\beta^2}}[(\sqrt{2-2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t})\cdot(-1)^n\exp\left(\mathrm{i}\cdot\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n\right)\cdot\exp\left(-\mathrm{i}\cdot\mathrm{arctg}\frac{\beta-\sin\Delta t}{\sqrt{1-\beta^2}-\cos\Delta t}\right)+(\sqrt{2+2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t})\cdot\exp\left(\mathrm{i}\cdot\mathrm{arctg}\frac{-\beta}{\sqrt{1-\beta^2}}\cdot n\right)\cdot\exp\left(\mathrm{i}\cdot\mathrm{arctg}\frac{\beta-\sin\Delta t}{\sqrt{1-\beta^2}+\cos\Delta t}\right)].$$

得到 A_m^n，B_m^n 的计算公式后，利用（10）式可以计算得到 u_m^n 的公式：

$$u_m^n=\frac{1}{2\sqrt{1-\beta^2}}[(\sqrt{2-2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t})\cdot(-1)^n\sin\left(m\Delta x-\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n+\mathrm{arctg}\frac{\beta-\sin\Delta t}{\sqrt{1-\beta^2}-\cos\Delta t}\right)+(\sqrt{2+2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t})\cdot\sin\left(m\Delta x+\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n-\mathrm{arctg}\frac{\beta-\sin\Delta t}{\sqrt{1-\beta^2}+\cos\Delta t}\right)].$$

当 $\Delta x\to 0$ 时，

$\beta=\lambda\sin\Delta x\to 0$，

$$\sqrt{2-2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t}\to 0,$$

含 $(-1)^n$ 项的影响会减小，

$$\sqrt{2+2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t}\to 2,$$

u_m^n 趋向于真解。

此格式的理论截断误差公式为：

$$E_{tm}^n=u_m^n-\tilde{u}_m^n=\frac{1}{2\sqrt{1-\beta^2}}[(\sqrt{2-2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t})\cdot(-1)^n\sin\left(m\Delta x-\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n+\mathrm{arctg}\frac{\beta-\sin\Delta t}{\sqrt{1-\beta^2}-\cos\Delta t}\right)+(\sqrt{2+2\sqrt{1-\beta^2}\cos\Delta t-2\beta\sin\Delta t})\cdot\sin\left(m\Delta x+\mathrm{arctg}\frac{\beta}{\sqrt{1-\beta^2}}\cdot n-\mathrm{arctg}\frac{\beta-\sin\Delta t}{\sqrt{1-\beta^2}+\cos\Delta t}\right)-\sin(n\Delta t+j\Delta x).\tag{11}$$

公式（11）与周毅等（2003）介绍的使用谐波

来推导截断误差的相位差的所得结果相似，但不同的是，本文采用的是实数初条件 $u_m^0=\sin(m\Delta x)$，这样做的好处是可以方便地利用二维图形来显示变量 u、截断误差 E_t 和舍入误差 E_r 在 x 方向的分布，而且避免了使用复数进行计算。

4 近似截断误差

并不是所有的微分方程都能得到解析解 A，因此有必要发展一种方法来分离那些无法得到解析解的方程在积分过程中的截断误差和舍入误差。

计算机的浮点计算精度直接影响到最终计算结果的舍入误差，以 K 表示浮点计算精度的话，对应的计算结果的舍入误差为 E_r；如果计算精度为 K_1，我们记最终结果的舍入误差为 E_{r1}，其它精度时的记号以此类推。当算法是稳定（不发散）的情况下，对于有限时间段的数值积分有如下引理：

引理 1：当计算机浮点计算精度无限高时，所得到的计算结果 $N_{K\to\infty}\to D$，即 $E_{r(K\to\infty)}\to 0$。

证明：由浮点计算的定义知，浮点精度为 K 时，单步积分计算引入的舍入误差可以记为 $E_{rK}\propto c\cdot 10^{-K}$，其中 c 为常数。可见当 $K\to\infty$ 时，$c\cdot 10^{-K}\to 0$，得证。

引理 2：若某次计算时使用的精度为 K_1，对应的舍入误差为 E_{r1}，那么一定存在一个计算精度 K_2，使所有的 $K_i\geqslant K_2\gg K_1$ 时得到的 $|E_{ri}|\ll|E_{r1}|$，特别是 $|E_{r2}|\ll|E_{r1}|$。

证明：假定不存在 K_2 使得 $|E_{r2}|\ll|E_{r1}|$，则必有 $\lim\limits_{K\to\infty}|E_{rK}|\neq 0$，与引理 1 矛盾，所以得证。

引理 3：若某次计算时使用的精度为 K_1，对应的舍入误差为 E_{r1}，对于某个计算精度 K_2，满足 $K_2\geqslant K_1$ 时不一定得到 $|E_{r2}|<|E_{r1}|$。

证明：只要举出 $|E_{r2}|>|E_{r1}|$ 的例子即得证，此例参见 Higham (1996) 图 1.3。

需要说明的是，引理 3 与引理 2 并不矛盾，引理 3 说明的是当 $K_2\geqslant K_1$ 但是两者的差别不大时的情况，而引理 2 说明的是 K_2 比 K_1 大得多时的情况。

计算近似的截断误差需要分两步进行，以变量 X 为例：

$$X_N = X + E_t(h_1) + E_r(K_1),$$

其中，X_N 为数值解，$E_t(h_1)$ 表示使用的步长为 h_1 时得到的截断误差，$E_r(K_1)$表示使用精度为 K_1 时的舍入误差。第一步需要计算 $X_N^{ref}=X+E_t(h_2)+E_r(K_2)$，其中 $E_t(h_2)$和$E_r(K_2)$ 的含义同上，但 $h_2\ll h_1$，$K_1\ll K_2$，以至于 $E_t(h_2)\to 0$，$E_r(K_2)\to 0$，X_N^{ref} 称为参考解，且 $X_N^{ref}\to X$。第二步计算 $X_N^t=X+E_t(h_1)+E_r(K_2)$，其中 $K_1\ll K_2$ 使得 $E_r(K_2)\to 0$，因此 $X_N^t\to X+E_t(h_1)$。最后，由得到的 X_N^{ref} 和 X_N^t 可以计算 $E_t(h_1)$的近似值：$E_t(h_1)\approx X_N^t-X_N^{ref}$，即近似截断误差。

为了验证参考解方法的有效性，进行如下的数值试验。图 2 为对比通过理论公式得到的截断误差与通过参考解算法得到的近似截断误差之间的差别，所用差分方案为迎风差。图 2a 所用的计算参数为 $\Delta x=2\pi/1000$，$\Delta t=0.005$，图 2b 中第一个点所用的计算参数为 $\Delta x=2\pi/1000$，$\Delta t=0.005$，第二个点的 Δx 和 Δt 为前一点的一半，以此类推。计算平台为 IBM-P690，计算所用精度为双精度，参

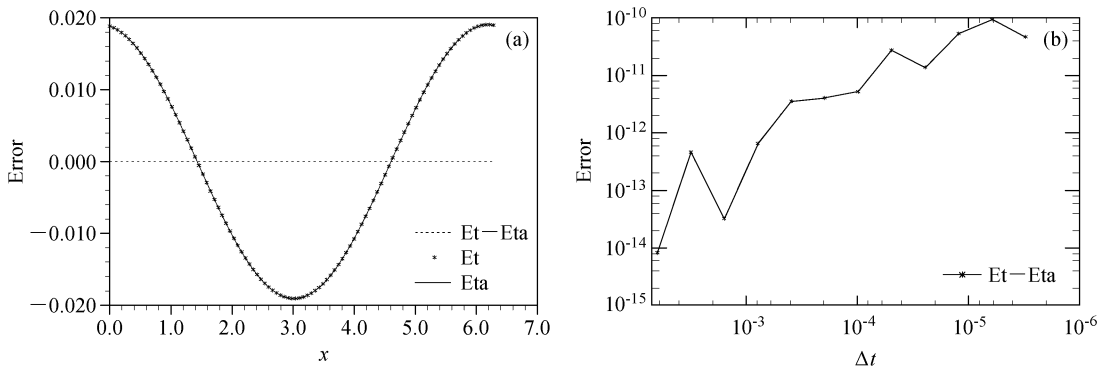

图 2 (a) 迎风差格式的理论截断误差 (Eta) 和近似截断误差 (Et) 随 x 的变化 ($t=30$)；(b) 理论截断误差和近似截断误差的差别随计算步长的变化

Fig. 2 (a) The variations of analytical truncation error (Eta) and approximate truncation error (Et) with x for the upstream scheme ($t=30$)；(b) the analytical truncation error and approximate truncation error versus step size

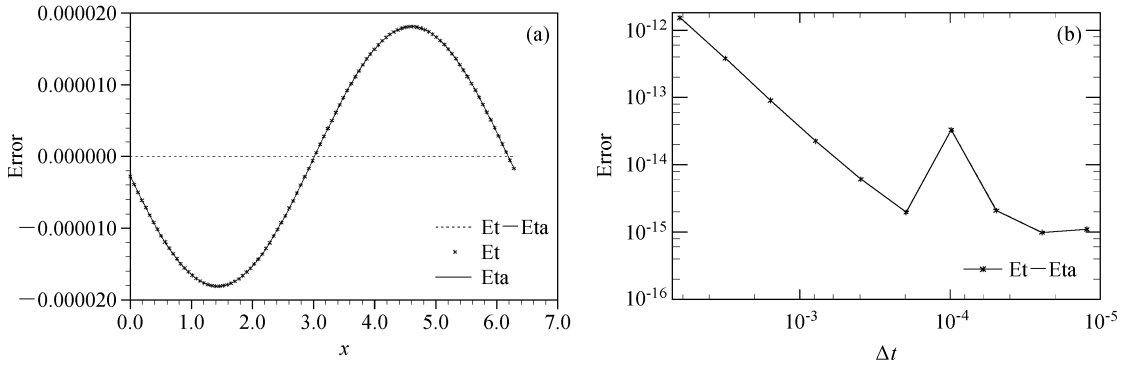

图 3　同图 2，但为中央差格式

Fig. 3　Same as Fig. 2, but for the centered difference scheme

考解的计算使用 4 倍精度。

从图 2a 中可以看出，理论截断误差（由实线表示）和近似截断误差（由星号表示）非常一致，说明了参考解算法得到的截断误差和舍入误差非常精确。虚线为理论截断误差和近似截断误差之间的差别，在10^{-10}以内，完全达到一般数值分析的需求。如果算法中不断提高计算精度 p_2 同时减小计算步长 h_2，完全能够达到更高的误差分离精度。从图 2b 可以看出，随着空间分辨率和时间分辨率的提高（即 Δx 和 Δt 逐渐减小的过程），理论截断误差和近似截断误差之间的差别有增大的趋势，但在$10^{-14}\sim10^{-10}$之内，这进一步说明了参考解的误差分离方法在多种不同 Δx 计算条件下都是相当精确的。

图 3 试验同图 2，但所用差分方案为中央差。从图 3a 中可以看出，理论截断误差（由实线表示）和近似截断误差（由星号表示）非常一致，但是位相与图 2a 不同，而且由于中央差为 2 阶计算精度，迎风差为 1 阶计算精度，因此截断误差的振幅比图 2a 中的要小。从图 3b 可以看出，随着空间分辨率和时间分辨率的提高，理论截断误差和近似截断误差之间的差别有减小的趋势，这也与迎风差的情况（图 2b）不同。差别的数值在 $10^{-16}\sim10^{-12}$ 之内，仍然保持了精确性。

5　结论和讨论

本文讨论数值积分过程中截断误差和舍入误差的分离方法，给出某些数值差分格式的理论截断误差，并以此来分离计算结果中的误差。介绍了参考解的办法来分离更为一般的微分系统中的截断误差和舍入误差，以此算法为基础，对一个偏微分方程的数值解进行计算，所得结果与采用理论截断误差得到的结果高度一致，说明了参考解方法的正确性。

对于文中所列的一阶波动方程，迎风差和中央差格式的理论截断误差在形式上也具有波动的周期特征，但是振幅的大小与计算参数有关，这一点无论从公式（9）、（11）或从图 2、3 都可以清楚地看到。参考解方法在使用的过程中也有需要注意的地方，首先，参考解方法可以应用的时间 t 是有限的一个时间段，而理论截断误差公式可以适用于任意 t 时刻。其次，参考解方法所用的计算精度要尽量高，如本文所用的 IBM-P690 计算机的四倍精度比双精度要高一倍左右。

误差分离方法的理论和试验方法具有广泛的应用价值，例如，可以用来验证偏微分方程数值解中 E_t 和 E_r 之间是否存在反向关系，讨论它们之间是否存在不确定性原理等。这些工作有待进一步研究。

参考文献（References）

Anderson J D Jr. 1995. Computational Fluid Dynamics: The Basics with Applications [M]. New York: McGraw-Hill Inc, 547pp.

Brouwer D. 1937. On the accumulation of errors in numerical integration [J]. Astronomical Journal, 46: 149－153.

Henrici P. 1962. Discrete Variable Methods in Ordinary Differential Equations [M]. New York: John Wiley, 187 pp.

Henrici P. 1963. Error Propagation for Difference Methods [M]. New York: John Wiley, 73pp.

Higham N J. 1996. Accuracy and Stability of Numerical Algorithms [M]. Philadelphia: SIAM, 688pp.

Li J P, Zeng Q C, Chou J F. 2000. Computational uncertainty principle in nonlinear ordinary differential equations I. Numerical re-

sults [J]. Science in China (Ser. E), 43: 449-461.

Li J P, Zeng Q C, Chou J F. 2001. Computational uncertainty principle in nonlinear ordinary differential equations II. Theoretical analysis [J]. Science in China (Ser. E), 44: 55-74.

Rademacher H A. 1948. On the accumulation of errors in processes of integration on high speed calculating machines [M]// The Annals of the Computation Laboratory of Harvard University. Cambridge: Harvard University Press, 16: 176-187.

Teixeira J, Reynolds C A, Judd K. 2007. Time step sensitivity of nonlinear atmospheric models: Numerical convergence, truncation error growth, and ensemble design [J]. J. Atmos. Sci., 64 (1): 175-189.

Turing A M. 1948. Rounding-off errors in matrix processes [J]. Quart. J. Mech. Appl. Math., 1: 287-308.

王鹏飞, 黄刚. 数值模式预报时效对计算精度和时间步长的依赖关系 [J]. 气候与环境研究, 2006, 11 (3): 395-403. Wang P F, Huang G. 2006. A study on the dependency of maximum prediction time on computation precision and time step-size in numerical model [J]. Climatic and Environmental Research (in Chinese), 11 (3): 395-403.

von Neumann J, Goldstine H H. 1947. Numerical inverting of matrices of high order [J]. Bull. Amer. Math. Soc., 53: 1021-1099.

von Neumann J. 1960. Some remarks on the problem of forecasting climatic fluctuations [M]// Preffer R L. Dynamics of Climate. New York: Pergamon Press, 9-12.

Wang P F, Li J P. 2008. The finite precision computation and the nonconvergence of difference scheme. http://arxiv.org/abs/0806.0421 [2010-07-01].

Wilkinson J H. 1963. Rounding Errors in Algebraic Processes [M]. London: Her Majesty's Stationery Office, 161pp.

周毅, 侯志明, 刘宇迪. 2003. 数值天气预报基础 [M]. 北京: 气象出版社. 222pp. Zhou Y, Hou Z M, Liu Y D. 2003. The Foundmental of Numerical Weather Prediction (in Chinese) [M]. Beijing: China Meteorological Press, 222pp.