第1章 数值计算方法（数值分析）引论

内容提要：

1.1 数值计算方法研究对象与特点

1.2 数值计算的误差

1.3 误差定性分析与避免误差危害

1.4 增补资料

1.1 研究对象与特点

一、研究对象

计算机解决科学计算问题时经历的过程

实际问题 → 模型设计 → 算法设计 → 程序设计 → 上机计算 → 问题的解

实例

求 $\sqrt{2}$ → 方程求根 $x^2=2$ → 牛顿法 $x_{k+1}=\frac{1}{2}\left(x_k+\frac{2}{x_k}\right)$ → 程序设计 → 上机计算 → 解 $x_0=1,\quad x_1=1.5,\quad x_3=1.417,\quad \cdots$

数值计算方法（**数值分析）的内容**包括**函数的数值逼近与拟合**、**数值微分与数值积分（数值微积分）**、**非线性方程（组）数值解**、**数值线性代数（线性方程组求解）**、**常微分方程**和偏微分方程数值求解等。因为数值分析研究对象以及解决问题方法的广泛适用性，著名流行软件如Matlab等已将其绝大多数内容设计成函数，简单调用之后便可以得到运行结果。

但由于实际问题的具体特征、复杂性，以及算法自身的适用范围决定了应用中必须选择、设计适合于自己特定问题的算法，因而掌握数值方法的思想和内容是至关重要的。

本课程内容包括了微积分（包括常微分方程）和线性代数的一些内容，掌握这些课程的必要基础内容为好。

二、数值计算方法的特点

- 面向计算机，要根据计算机的特点提供切实可行的有效算法。
- 有可靠的理论分析，能任意逼近并达到精度要求，对近似算法要保证收敛性和数值稳定性，还要对误差进行分析。这些都是建立在数学理论的基础上，因此不应片面的将数值分析理解为各种数值方法的简单罗列和堆积。
- 要有好的计算复杂性，时间复杂性好是指节省时间，空间复杂性好是指节省存储量，这也是建立算法要研究的问题，它关系到算法能否在计算机上实现。
- 要有数值实验，即任何一个算法除了从理论上要满足上述三点外，还要通过数值实验证明是行之有效的。

三、学习方法

初学可能会觉得公式多，理论分析复杂。给出如下的五点学习方法。

- 认识建立算法和对每个算法进行基本的理论分析是基本任务，主动适应公式多和讲究理论分析的特点。
- 注重各章节所研究算法的提出，掌握方法的基本原理和基本思想，要注意方法处理的技巧及其与计算机的结合。
- 理解每个算法建立的数学背景、数学原理和基本线索，而且对一些最基本的算法要非常熟悉。
- 要通过算例学习使用各种数值方法解决实际计算问题。
- 为掌握本课的内容，还应做一些理论分析和计算练习。

1.2 数值计算的误差

一、误差的来源与分类

在运用数学方法解决实际问题的过程中，每一步都可能带来误差。

1、模型误差 在建立数学模型时，往往要忽视很多次要因素，把模型“简单化”，“理想化”，这时模型就与真实背景有了差距，即带入了误差。数学模型与实际问题之间出现的误差称为模型误差 。

2、观测误差（测量误差） 数学模型中的已知参数，多数是通过测量得到。而测量过程受工具、方法、观察者的主观因素、不可预料的随机干扰等影响必然带入误差。

3、截断误差（算法误差） **数学模型常难于直接求解，往往要用数值方法求近似解替代，这种简化带入误差称为方法误差或截断误差。**

Taylor展开
很重要！

例如：可微函数 $f(x)$ 用泰勒 (Taylor) 多项式

$$P_n(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n$$

近似代替，则数值方法的截断误差是泰勒余项$R_n(x)$。

4、舍入误差 **计算机只能处理有限数位的小数运算，初始参数或中间结果都必须进行四舍五入运算，这必然产生舍入误差。**

例如：用3.14159 近似代替 π，产生的误差

$$e = \pi - 3.14159 = 0.0000026\cdots$$

误差分析是一门比较艰深的专门学问。在数值计算方法（数值分析）中主要讨论截断误差及舍入误差。但一个训练有素的计算工作者，当发现计算结果与实际不符时，应当能诊断出误差的来源，并采取相应的措施加以改进，直至建议对模型进行修改。

二、绝对误差、相对误差与有效数字

1、绝对误差与绝对误差限

定义1-1 设 x^* 为准确值，x 为 x^* 的一个近似值，称 $e = x - x^*$ 为近似值 x 的绝对误差，简称误差, 记为 e 。

误差是有量纲的量，量纲同 x^*，它可正可负。误差一般无法准确计算，只能根据测量或计算情况估计出它的误差绝对值的一个上界，这个上界称为近似值 x 的误差限，记为ε。它是正数，有量纲的。如用毫米刻度尺测量长度。误差限是0.5 mm。

误差限的2种表示：

（1）对于一般情形 $\left|x - x^*\right| \leq \varepsilon$

（2）$x^* - \varepsilon \leq x \leq x^* + \varepsilon$

2、相对误差与相对误差限

定义1-2　近似值的误差 e 与准确值 x^* 的比值

$$\frac{e}{x^*} = \frac{x - x^*}{x^*}$$

称为近似值 x 的相对误差，记作 e_r。　相对误差无量纲，可正可负。

在计算中，由于真值 x^* 总是不知道的，通常取

$$e_r = \frac{e}{x} = \frac{x - x^*}{x}。$$

相对误差限　相对误差的绝对值上界叫相对误差限，记作 ε_r，即

$$\varepsilon_r = \frac{\varepsilon}{|x|} \geq \frac{|x - x^*|}{|x|} = |e_r|$$

若 $\frac{\varepsilon_x}{|x|} = 10\%$ 与 $\frac{\varepsilon_y}{|y|} = 0.1\%$ 分别为 x 与 y 的相对误差限，可见 y 近似 y^* 的程度比 x 近似 x^* 的程度好。

3、有效数字

定义1-3　如果近似值*x*的误差限是某一位的半个单位，该位到 *x*的第一位非零数字共有*n*位，就说*x*有*n*位有效数字.

例如　下列数按四舍五入原则写出下列各数的具有5位有效数字的近似数？

187.9325　　0.037 855 51　　8.000 033　　2.718 281 8

解：　187.93　　0.037 856　　8.0000　　2.7183

再例　下列各数都是经过四舍五入原则得到的近似数指出它们有几个有效数字？

$x_1 = 1.1021$　　$x_2 = 56.430$　　$x_3 = 0.031$

解：　x_1 五位　　x_2 五位　　x_3 二位

1.3 误差定性分析与避免误差危害

一、算法的稳定性

用一个算法进行计算，由于初始数据误差在计算中传播使计算结果误差增长很快，就是数值不稳定的。

二、病态问题与条件数

1、病态问题：对一个数值问题本身如果输入数据有微小扰动（即误差），引起输出数据（即问题解）相对误差很大，这就是病态问题。

2、条件数:计算函数值 $f(x)$ 时, 若 x 有扰动 $\Delta x = x - \tilde{x}$ ，其相对误差为 $\dfrac{\Delta x}{x}$, 而函数值的相对误差为

$$\frac{f(x) - f(\tilde{x})}{f(x)}$$

故，利用 $f(\tilde{x}) \approx f(x) + f'(x)(\tilde{x} - x)$ ，则相对误差比值如下：

$$\left|\frac{f(x) - f(\tilde{x})}{f(x)}\right| \Bigg/ \left|\frac{\Delta x}{x}\right| \approx \left|\frac{x f'(x)}{f(x)}\right| = C_{\mathrm{p}}$$

C_{p} 称为计算函数值问题的条件数。

课堂推导

分析：自变量相对误差一般不会太大，如果条件数C_p很大，将引起函数值相对误差很大，出现这种情况的问题就是病态问题。

例如　取 $f(x)=x^n$,则有 $C_p=\left|\frac{x\cdot nx^{n-1}}{x^n}\right|=n$ 。它表示相对误差可能放大 n 倍。如 $n=10$, 有 $f(1)=1$, $f(1.02)\approx 1.24$, 若取 $x=1$, $\tilde{x}=1.02$ 自变量相对误差为2%，函数相对误差为24%，这时问题可以认为是病态的。一般情况条件数 $C_p\geq 10$ 就认为是病态，C_p 越大病态越严重。

注意：病态问题不是计算方法引起的，是数值问题自身固有的，因此，对数值问题首先要分清问题是否病态，对病态问题就必须采取相应的特殊方法以减少误差危害。

三、避免误差危害的若干原则

1、要避免除数绝对值远远小于被除数绝对值的除法。

用绝对值小的数作除数舍入误差会增大，如计算 x/y，若$0<|y|<<|x|$，则可能对计算结果带来严重影响，应尽量避免。

2、要避免两相近数相减

在数值中两相近数相减有效数字会严重损失。

例如 x=532.65，y=532.52都具有五位有效数字，但 $x-y$=0.13只有两位有效数字。

通过改变算法可以避免两相近数相减。

例如 $\sqrt{x+1}-\sqrt{x}$ $(x>>1)$ 可改为 $\dfrac{1}{\sqrt{x+1}+\sqrt{x}}$，

$1-\cos x$ $(|x|<<1)$ 可改为 $2\sin^2\left(\dfrac{x}{2}\right)$

例1-8

(1) $\ln\left(x-\sqrt{x^2-1}\right)$ (x 很大) (2) $\dfrac{\sin x}{x-\sqrt{x^2-1}}$ (x 很大)

等等，都可以得到比直接计算好的结果。

答案

(1) $\ln\dfrac{1}{x+\sqrt{x^2-1}}$ (2) $\left(x+\sqrt{x^2-1}\right)\sin x$

（符号数学）等价不一定（数值计算）等效

例8解：

$$(1)\quad \ln\left(x-\sqrt{x^2-1}\right)=\ln\left[\frac{\left(x-\sqrt{x^2-1}\right)\times\left(x+\sqrt{x^2-1}\right)}{\left(x+\sqrt{x^2-1}\right)}\right]=\ln\left(\frac{1}{x+\sqrt{x^2-1}}\right)$$

$$(2)\quad \frac{\sin x}{x-\sqrt{x^2-1}}=\frac{\left(x+\sqrt{x^2-1}\right)\times\sin x}{\left(x-\sqrt{x^2-1}\right)\times\left(x+\sqrt{x^2-1}\right)}=\left(x+\sqrt{x^2-1}\right)\times\sin x$$

3、要防止“大数”吃掉小数

数值运算中参加运算的数有时数量级相差很大，而计算机位数有限，如不注意运算次序就可能出现大数“吃掉”小数的现象，影响计算结果的可靠性。

如用六位浮点数计算某市的工业总产值，原始数据是各企业的工业产值，当加法进行到一定程度，部分和超过100亿元 （0.1×10^{11}），再加产值不足10万元的小企业产值，将再也加不进去。而这部分企业可能为数不少，合计产值相当大. 这种情况应将小数先分别加成大数，然后相加，结果才比较正确。

这个例子告诉我们：在计算机数系中，加法的交换律和结合律可能不成立，这是在大规模数据处理时应注意的问题。

直觉不一定对，要实践（如Matlab的sum函数）：REPORT

4、注意简化计算步骤，减少运算次数

减少算术运算的次数不但可计算机的计算时间，还能减少误差（如舍入误差）的积累效应。使参加运算的数字精度应尽量保持一致，否则那些较高精度的量的精度没有太大意义。

例如　计算多项式值

$$P_n(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots a_1x+a_0$$

解：法一：直接计算 a_kx^k 再逐项相加，一共需要做

$$n+(n-1)+\cdots+2+1=\frac{n(n+1)}{2}$$ 次乘法和 n 次加法。

法二：采用秦九韶算法（西方称Horner 算法）

$$\begin{cases} S_n=a_n \\ S_k=xS_{k+1}+a_k\ (k=n-1,\cdots,2,1,0) \\ P_n(x)=S_0 \end{cases}$$

即　$P_n(x)=((a_nx+a_{n-1})x+a_{n-2})x\cdots+a_1)x+a_0$

只要计算 n 次乘法和 n 次加法就可算出 $P_n(x)$ 的值。

Evaluation of a Polynomial

Let the polynomial $P(x)$ of degree n have the form

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0. \tag{20}$$

Horner's method or ***synthetic division*** is a technique for evaluating polynomials. It can be thought of as nested multiplication. For example, a fifth-degree polynomial can be written in the nested multiplication form

$$P_5(x) = ((((a_5 x + a_4)x + a_3)x + a_2)x + a_1)x + a_0.$$

The recursive formula for b_k given in (21) is easy to implement with a computer. A simple algorithm is

$b(n) = a(n);$
for $k = n - 1: -1: 0$
  $b(k) = a(k) + c * b(k+1);$
end

Table 1.2 Horner's Table for the Synthetic Division Process

Input	a_n	a_{n-1}	a_{n-2}	$\cdots$	a_k	$\cdots$	a_2	a_1	a_0
c		xb_n	xb_{n-1}	$\cdots$	xb_{k+1}	$\cdots$	xb_3	xb_2	xb_1
	b_n	b_{n-1}	b_{n-2}	$\cdots$	b_k	$\cdots$	b_2	b_1	$b_0 = P(c)$
									Output

Example 1.9. Use synthetic division (Horner's method) to find $P(3)$ for the polynomial

$$P(x) = x^5 - 6x^4 + 8x^3 + 8x^2 + 4x - 40.$$

	a_5	a_4	a_3	a_2	a_1	a_0
Input	1	-6	8	8	4	-40
$c = 3$		3	-9	-3	15	57
	1	-3	-1	5	19	$17 = P(3) = b_0$
	b_5	b_4	b_3	b_2	b_1	Output

Therefore, $P(3) = 17$. ■

知识结构图

- 误差及算法
  - 误差
    - 分类
      - 舍入误差的产生及定义
      - 截断误差的产生及定义
    - 度量
      - 绝对误差（限）
      - 相对误差（限）
      - 有效数字
    - 传播
      - 一元函数
      - 二元算术运算
      - n元函数
      - 计算函数值问题的条件数
  - 算法
    - 数值稳定性概念
    - 算法设计注意要点

复习与思考题(无需提交)

P19: 2, 3, 5, 6, 10

（REPORT: P19复习与思考题2之计算机实验与报告）

习题(需提交)

P19: 2, 8

1.4 增补资料

- Stability
  - Convergence
  - Divergence

Numerical Errors

- Round-off Errors
    - 3.1415926 → 3.1416
- Truncation Errors - Code dependent

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} + \cdots + \frac{x^{2n}}{n!} + \cdots$$

Matlab - Matrix Laboratory

The Matlab program can be run using command line, batch commands, and programs.

What is a program?

Program consist of three main components:

- Input
- Main Program - Numerical methods and analysis and/or evaluation.
- Output - Results.

Inputs

- Numerical values
- Initialization of the variables
- Conditions
- Equations

Main Program

Using flow charts, the programs can be designed to perform a task. Using:

- Loops (for do while)
- Conditions (if then elseif etc..)
- Error Convergence (while)

Output

Outputs are the results of the program. They can go through a series of post-processing methods.

- Numerical Values
- Decisions
- Graphs and Plots

MatLab

Variable Types

- Integers
- Real Values (float and double)
- Complex Numbers (a + ib)
  - a - real value
  - b - imaginary value ("i" is the square root of -1)

Matlab里面用i和j要小心（误入复数域）

Matlab

Data types

- Numerical
  - Scalars
  - Vectors
  - Matrices
- Logic Types

Matlab

- A scalar value is the simple number, a, 2, 3.14157…,
- A vector is a union of scalars: $\mathrm{x} = [x_1, x_2, x_3, x_4]$
- Transpose vector: $\mathrm{x}^{\mathrm{T}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$

Matlab

- Matrix is a combination of vectors and scalars. Scalar and vectors are subsets of matrices.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

- Matlab uses matrix to do mathematical methods.
- A=[a11 a12 a13; a21 a22 a23; a31 a32 a33];

Matlab

- Set of computer functions
  - Trigonometric functions - sin(x),cos(x), tan(x), asin(x), acos(x), atan(x)
  - Hyperbolic functions - sinh(x), cosh(x), tanh(x)
  - Logarithmic functions - ln(x), log(x), exp(x)
  - Complex functions - abs(x), real(x), imag(x)

Matlab

- Simple commands
  - clc - clears window
  - clg - clear graphic window
  - clear - clears the workspace
  - who - variable list
  - whos - variable list with size
  - help - when doubt use it!

Matlab

- Simple commands and symbols
  - ^C - an escape from a loop
  - inf - infinity
  - NaN - No numerical value

Matlab - Scalar Operations

- Addition - a + b
- Subtraction - a - b
- Multiplication - a * b
- Division - a / b
- Exponential - a^b

Matlab - Vector Operations

Matlab - Matrix Operations

Order of Precedence of Arithmetic Operations

Precedence

(1) - Parenthesis

(2) - Exponential from left to right

(3) - Multiplication and division from left to right.

(4) - Addition and subtraction from left to right.

Taylor Expansion

Theorem 1.12 (Taylor's Theorem). Assume that $f \in C^{n+1}[a,b]$ and let $x_0 \in [a,b]$. Then, for every $x \in (a,b)$, there exists a number $c = c(x)$ (the value of c depends on the value of x) that lies between x_0 and x such that

$$f(x) = P_n(x) + R_n(x), \tag{16}$$

where

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{17}$$

and

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}. \tag{18}$$

Big O(h)

Definition 1.11. Assume that $f(h)$ is approximated by the function $p(h)$ and that there exist a real constant $M > 0$ and a positive integer n so that

$$(9) \qquad \frac{|f(h) - p(h)|}{|h^n|} \leq M \quad \text{for sufficiently small } h.$$

We say that $p(h)$ ***approximates*** $f(h)$ with order of approximation $\boldsymbol{O}(h^n)$ and write

$$(10) \qquad f(h) = p(h) + \boldsymbol{O}(h^n). \qquad \blacktriangle$$

When relation (9) is rewritten in the form $|f(h) - p(h)| \leq M|h^n|$, we see that the notation $\boldsymbol{O}(h^n)$ stands in place of the error bound $M|h^n|$. The following results show how to apply the definition to simple combinations of two functions.

注意小o(h)和大O(h)之不同

$$\frac{1}{1-h} = 1 + h + h^2 + h^3 + \boldsymbol{O}(h^4)$$

Big O(h)

Theorem 1.15. Assume that $f(h) = p(h) + \boldsymbol{O}(h^n)$, $g(h) = q(h) + \boldsymbol{O}(h^m)$, and $r = \min\{m, n\}$. Then

$$f(h) + g(h) = p(h) + q(h) + \boldsymbol{O}(h^r), \tag{11}$$

$$f(h)g(h) = p(h)q(h) + \boldsymbol{O}(h^r), \tag{12}$$

and

$$\frac{f(h)}{g(h)} = \frac{p(h)}{q(h)} + \boldsymbol{O}(h^r) \quad \text{provided that } g(h) \neq 0 \text{ and } q(h) \neq 0. \tag{13}$$

Big O(h)

$$\boldsymbol{O}(h^{n+1}) \approx M h^{n+1} \approx \frac{f^{(n+1)}(c)}{(n+1)!} h^{n+1} \tag{14}$$

for sufficiently small h. Hence the notation $\boldsymbol{O}(h^{n+1})$ stands in place of the quantity Mh^{n+1}, where M is a constant or "behaves like a constant."

Theorem 1.16 (Taylor's Theorem). Assume that $f \in C^{n+1}[a, b]$. If both x_0 and $x = x_0 + h$ lie in $[a, b]$, then

$$f(x_0 + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} h^k + \boldsymbol{O}(h^{n+1}). \tag{15}$$

The following example illustrates the theorems above. The computations use the addition properties (i) $\boldsymbol{O}(h^p) + \boldsymbol{O}(h^p) = \boldsymbol{O}(h^p)$, (ii) $\boldsymbol{O}(h^p) + \boldsymbol{O}(h^q) = \boldsymbol{O}(h^r)$, where $r = \min\{p, q\}$, and the multiplicative property (iii) $\boldsymbol{O}(h^p)\boldsymbol{O}(h^q) = \boldsymbol{O}(h^s)$, where $s = p + q$.

Big O(h)

Example 1.22. Consider the Taylor polynomial expansions

$$e^h = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \boldsymbol{O}(h^4) \quad \text{and} \quad \cos(h) = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + \boldsymbol{O}(h^6).$$

Determine the order of approximation for their sum and product.

For the sum we have

$$\begin{aligned} e^h + \cos(h) &= 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \boldsymbol{O}(h^4) + 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + \boldsymbol{O}(h^6) \\ &= 2 + h + \frac{h^3}{3!} + \boldsymbol{O}(h^4) + \frac{h^4}{4!} + \boldsymbol{O}(h^6). \end{aligned}$$

Since $\boldsymbol{O}(h^4) + \frac{h^4}{4!} = \boldsymbol{O}(h^4)$ and $\boldsymbol{O}(h^4) + \boldsymbol{O}(h^6) = \boldsymbol{O}(h^4)$, this reduces to

$$e^h + \cos(h) = 2 + h + \frac{h^3}{3!} + \boldsymbol{O}(h^4),$$

and the order of approximation is $\boldsymbol{O}(h^4)$.

The product is treated similarly:

$$
\begin{aligned}
e^h \cos(h) &= \left(1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \boldsymbol{O}(h^4)\right)\left(1 - \frac{h^2}{2!} + \frac{h^4}{4!} + \boldsymbol{O}(h^6)\right) \\
&= \left(1 + h + \frac{h^2}{2!} + \frac{h^3}{3!}\right)\left(1 - \frac{h^2}{2!} + \frac{h^4}{4!}\right) \\
&\quad + \left(1 + h + \frac{h^2}{2!} + \frac{h^3}{3!}\right)\boldsymbol{O}(h^6) + \left(1 - \frac{h^2}{2!} + \frac{h^4}{4!}\right)\boldsymbol{O}(h^4) \\
&\quad + \boldsymbol{O}(h^4)\boldsymbol{O}(h^6) \\
&= 1 + h - \frac{h^3}{3} - \frac{5h^4}{24} - \frac{h^5}{24} + \frac{h^6}{48} + \frac{h^7}{144} \\
&\quad + \boldsymbol{O}(h^6) + \boldsymbol{O}(h^4) + \boldsymbol{O}(h^4)\boldsymbol{O}(h^6).
\end{aligned}
$$

Since $\boldsymbol{O}(h^4)\boldsymbol{O}(h^6) = \boldsymbol{O}(h^{10})$ and

$$
-\frac{5h^4}{24} - \frac{h^5}{24} + \frac{h^6}{48} + \frac{h^7}{144} + \boldsymbol{O}(h^6) + \boldsymbol{O}(h^4) + \boldsymbol{O}(h^{10}) = \boldsymbol{O}(h^4),
$$

the preceding equation is simplified to yield

$$
e^h \cos(h) = 1 + h - \frac{h^3}{3} + \boldsymbol{O}(h^4),
$$

and the order of approximation is $\boldsymbol{O}(h^4)$. ■