第1章 数值计算中的误差

1.1 基本要求与主要内容

了解误差的概念及分类，知道科学计算中误差的来源，理解有效数字的概念，掌握数值计算中误差的传播规律和分析方法，以及数值计算方法中一般应遵循的原则.

1.1.1 误差

1. 误差的来源

在科学计算中误差来源一般有以下 4 个方面：模型误差、观测误差、截断误差和舍入误差. 这里主要考虑截断误差和舍入误差.

2. 绝对误差（限）、相对误差（限）和有效数字

(1) 绝对误差与绝对误差限

设某一个量的准确值为 x，其近似值为 x^*，则称 $e^*=e(x^*)=x-x^*$ 为近似值 x^* 的绝对误差.

如果可估计出误差绝对值的一个上界 ε，即 $|e(x^*)|=|x-x^*|\leqslant\varepsilon$，则称 ε 为近似值 x^* 的绝对误差限，简称误差限.

(2) 相对误差与相对误差限

称绝对误差与准确值之比 $e_r^*=e_r(x^*)=\dfrac{e(x^*)}{x}=\dfrac{x-x^*}{x}$ 为近似值 x^* 的相对误差.

在实际问题中常取 $e_r^*=\dfrac{e(x^*)}{x^*}$ 作为相对误差的另一定义.

如果存在一个正数 ε_r，使得 $|e_r(x^*)|\leqslant\varepsilon_r$，则称 ε_r 为近似值 x^* 的相对误差限.

注：近似值 x^* 的绝对误差限 ε 和相对误差限 ε_r 都不是唯一的.

(3) 有效数字

如果近似值 x^* 的误差限是它的某一位的半个单位，则称该近似值准确到这一位；设从该位到 x^* 的第一位非零数字共有 n 位，则称 x^* 有 n 位有效数字.

具体地说，设 x 的近似值 x^* 的规格化形式为

$$x^* = \pm 0.\alpha_1\alpha_2\cdots\alpha_n \times 10^m \tag{1.1}$$

其中 $\alpha_1,\alpha_2,\cdots,\alpha_n$ 都是 0～9 中的任一整数，且 $\alpha_1 \neq 0$；n 是正整数，m 是整数. 若 x^* 的误差限为

$$|x - x^*| \leqslant \frac{1}{2} \times 10^{m-l}, \quad 1 \leqslant l \leqslant n \tag{1.2}$$

则称 x^* 为具有 l 位有效数字的有效数，或称它精确到10^{m-l}.

例如，π 的近似值 3.143 和 3.142 分别有 3 位和 4 位有效数字.

(4) 有效数字和相对误差限的关系

定理 1.1 若近似值 $x^* = \pm 0.\alpha_1\alpha_2\cdots\alpha_n \times 10^m$ 具有 n 位有效数字，则其相对误差限为

$$\varepsilon_r \leqslant \frac{1}{2\alpha_1} \times 10^{-n+1} \tag{1.3}$$

定理 1.2 若近似值 $x^* = \pm 0.\alpha_1\alpha_2\cdots\alpha_n \times 10^m$ 的相对误差满足 $\varepsilon_r \leqslant \dfrac{1}{2(\alpha_1+1)} \times 10^{-n+1}$，则 x^* 至少具有 n 位有效数字.

3. 数值计算中误差的传播规律

设近似值 x_1^* 和 x_2^* 的误差限分别是 $\varepsilon(x_1^*)$ 和 $\varepsilon(x_2^*)$，则它们进行加、减、乘、除运算得到的误差限分别为

$$\varepsilon(x_1^* \pm x_2^*) \leqslant \varepsilon(x_1^*) + \varepsilon(x_2^*);$$

$$\varepsilon(x_1^* x_2^*) \leqslant |x_1^*|\varepsilon(x_2^*) + |x_2^*|\varepsilon(x_1^*);$$

$$\varepsilon\left(\frac{x_1^*}{x_2^*}\right) \leqslant \frac{|x_1^*|\varepsilon(x_2^*) + |x_2^*|\varepsilon(x_1^*)}{|x_2^*|^2}, \quad x_2^* \neq 0$$

一般地，当自变量有误差时计算函数值也产生误差，其误差限可利用函数的泰勒展开式进行估计. 函数值的误差如下：

设有 n 元函数 $f(x_1,x_2,\cdots,x_n)$，计算 $A = f(x_1,x_2,\cdots,x_n)$. 如果 $x_1,x_2,\cdots,x_n$ 的近似值为 $x_1^*,x_2^*,\cdots,x_n^*$，则 A 的近似值为 $A^* = f(x_1^*,x_2^*,\cdots,x_n^*)$. 于是函数值 A^* 的误差

$$e(A^*) = f(x_1,x_2,\cdots,x_n) - f(x_1^*,x_2^*,\cdots,x_n^*) \approx \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^* e(x_i^*) \tag{1.4}$$

误差限为

$$\varepsilon(A^*) \approx \sum_{i=1}^{n}\left|\left(\frac{\partial f}{\partial x_i}\right)^*\right|\varepsilon(x_i^*) \tag{1.5}$$

而 A^* 的相对误差限为

$$\varepsilon_r(A^*) = \frac{\varepsilon(A^*)}{|A^*|} \approx \sum_{i=1}^{n} \left| \left(\frac{\partial f}{\partial x_i}\right)^* \right| \frac{\varepsilon(x_i^*)}{|A^*|} \tag{1.6}$$

1.1.2　数值计算中应注意的一些问题

为保证数值算法的稳定性,在数值计算中一般应遵循如下一些原则:

(1) 应选用数值稳定的计算方法,避开不稳定的算式.

例如,已知 $a_0=\sqrt{3}$,试利用递推公式 $a_k=10a_{k-1}-1(k=1,2,\cdots)$计算 a_{100}.

由于$\sqrt{3}$是无理数,计算机只能截取前有限位数来计算,设 a_0 经机器舍入得到的近似值为 a_0^*,利用公式计算得到 a_k^*,则 $a_{100}-a_{100}^*=10^{100}(a_0-a_0^*)$. 误差扩大了$10^{100}$倍,计算显然是不稳定的.

(2) 注意简化计算步骤及公式,减少误差的积累;设法减少乘除法运算,节约计算机的机时.

例如,考虑多项式 $a_0+a_1x+\cdots+a_nx^n$ 的算法设计. 仅讨论乘法的计算量. 如果直接计算,需要$\frac{n(n+1)}{2}$次乘法;如果使用下面的递推方法

$$t_0=1,\quad p_0=a_0,\quad t_k=xt_{k-1},\quad p_k=p_{k-1}+a_kt_k\quad (k=1,2,\cdots,n)$$

则只需要 $2n$ 次乘法;如果用秦九韶算法,则只有 n 次乘法了.

(3) 应合理安排运算顺序,防止参与运算的数在数量级相差悬殊时,大数"淹没"小数的现象发生.

一个 p 进制数 x 可以写成

$$x=\pm a\times p^J \tag{1.7}$$

其中 $a=\sum_{k=1}^{t}d_kp^{-k}=0.d_1d_2\cdots d_t$. 称 a 为数 x 的尾数. 自然数 t 为计算机字长,整数 J 称为数 x 的阶. 计算机在进行运算时,首先要把参加运算的数对阶. 例如,计算 $x=10^9+1$,必须改写成

$$x=0.1\times10^{10}+0.0000000001\times10^{10}$$

如果计算机的字长为 8,则算出 $x=0.1\times10^{10}$,大数"淹没"了小数.

(4) 应避免两相近数相减,可用变换公式的方法来解决. 可参考下节的例 1.6.

(5) 绝对值太小的数不宜作为除数,否则产生的误差过大,甚至会在计算机中造成"溢出"错误.

例如,$\varepsilon\left(\frac{x_1^*}{x_2^*}\right)\approx\frac{|x_1^*|\varepsilon(x_2^*)+|x_2^*|\varepsilon(x_1^*)}{|x_2^*|^2}$,当$|x_2^*|\ll|x_1^*|$时,误差可能增大很多.

1.2 例题选讲

例 1.1 若 3.142，3.141，$\frac{22}{7}$分别作为圆周率 π 的近似值，问它们各具有几位有效数字？

解 $\pi=3.14159\cdots$，记 $x_1^*=3.142$，$x_2^*=3.141$，$x_3^*=\frac{22}{7}$.

由 $\pi-x_1^*=-0.00041\cdots$，知

$$\frac{1}{2}\times10^{-4}<|\pi-x_1^*|\leqslant\frac{1}{2}\times10^{-3}$$

所以 $x_1^*=3.142$ 有 4 位有效数字；

由 $\pi-x_2^*=0.00059\cdots$知

$$\frac{1}{2}\times10^{-3}<|\pi-x_2^*|\leqslant\frac{1}{2}\times10^{-2}$$

所以 $x_2^*=3.141$ 有 3 位有效数字；

由 $\pi-x_3^*=3.14159\cdots-3.14285\cdots=-0.00126\cdots$，知

$$\frac{1}{2}\times10^{-3}<|\pi-x_2^*|\leqslant\frac{1}{2}\times10^{-2}$$

所以 $x_2^*=3.141$ 有 3 位有效数字.

例 1.2 设有三个近似数 $x_1^*=2.31$，$x_2^*=1.93$，$x_3^*=2.24$，它们都有三位有效数字，试计算 $y=x_1+x_2x_3$ 及其绝对误差限，并说明 y 的计算结果有多少位有效数字？

解 $y=x_1+x_2x_3=2.31+1.93\times2.24=6.6332$

$$\begin{aligned}\varepsilon(y)&=\varepsilon(x_1)+\varepsilon(x_2x_3)\approx\varepsilon(x_1)+|x_2^*|\varepsilon(x_3)+|x_3^*|\varepsilon(x_2)\\&=0.005+0.005(1.93+2.24)=0.02585\end{aligned}$$

因为 $\varepsilon(y)\approx0.02585<\frac{1}{2}\times10^{-1}$，$m-l=-1$，$m=1$，所以 $l=m+1=2$，即 y 的计算结果有 2 位有效数字.

例 1.3 已知近似数 x^* 具有 2 位有效数字，试求其相对误差限.

解 依题意 $l=2$，并考虑到 α_1 是 1～9 之间的数字，利用有效数字与相对误差的关系得

$$\varepsilon_r\leqslant\frac{1}{2\alpha_1}\times10^{-l+1}\leqslant\frac{1}{2\times1}\times10^{-2+1}=5\%$$

例 1.4 设计算球体积允许其相对误差限为 1%，问测量球半径的相对误差限最大为多少？

解 记球半径为 R，球体积为 V，则由 $V=\frac{4}{3}\pi R^3$ 得 $\mathrm{d}V=4\pi R^2\mathrm{d}R$，从而$\frac{\mathrm{d}V}{V}=3\frac{\mathrm{d}R}{R}$. 于

是，$\varepsilon_r(R)=\frac{1}{3}\varepsilon_r(V)\leqslant\frac{1}{3}\times1\%=0.33\%$.

例 1.5　若$|x|\ll1$，利用等价变换使下列表达式的计算结果比较精确：

(1) $\frac{1}{1+2x}-\frac{1-x}{1+x}$；(2) e^x-1.

解　(1) $\frac{1}{1+2x}-\frac{1-x}{1+x}=\frac{1+x-(1-x)(1+2x)}{(1+2x)(1+x)}=\frac{2x^2}{(1+2x)(1+x)}$；

(2) $e^x-1=x+\frac{1}{2}x^2+\frac{1}{6}x^3+\frac{1}{24}x^4+\cdots\approx x+\frac{1}{2}x^2+\frac{1}{6}x^3+\frac{1}{24}x^4$.

例 1.6　当 N 充分大时，如何计算 $\int_N^{N+1}\frac{1}{1+x^2}dx$？

解　令 $\tan\theta_1=N$，$\tan\theta_2=N+1$，则由

$$\tan(\theta_2-\theta_1)=\frac{\tan\theta_2-\tan\theta_1}{1+\tan\theta_2\cdot\tan\theta_1}=\frac{N+1-N}{1+(N+1)N}=\frac{1}{1+(N+1)N}$$

得

$$\theta_2-\theta_1=\arctan(N+1)-\arctan N=\arctan\frac{1}{1+(N+1)N}$$

于是

$$\int_N^{N+1}\frac{1}{1+x^2}dx=\arctan\frac{1}{1+(N+1)N}.$$

1.3　练习题及解答

1.3.1　练习题

1. 若 $x^*=3587.64$ 是 x 的有 6 位有效数字的近似值，求 x 的绝对误差限.

2. 为使 $\sqrt{70}$ 的近似值的相对误差小于 0.1%，问查开方表时，要取几位有效数字？

3. 设 $x>0$，x 的相对误差限为 δ，求 $f(x)=\ln x$ 的相对误差限.

4. 求方程 $x^2-56x+1=0$ 的两个根，使它至少具有 4 位有效数字(已知 $\sqrt{783}\approx27.982$).

5. 为了使计算 $y=10+\frac{3}{x-1}+\frac{4}{(x-1)^2}-\frac{6}{(x-1)^3}$ 的乘除法运算次数尽量少，应将该表达式改写为________.

6. 为了减少舍入误差的影响，应将表达式 $\sqrt{2012}-\sqrt{2010}$ 改写为________.

7. 若$|x|\ll1$，利用等价变换使表达式 $1-\cos x$ 的计算结果比较精确.

8. 在计算函数 $f(x)=\ln(x+\sqrt{x^2+1})$ 的值时，应如何计算才能避免有效数字的损失？

9. 按四舍五入原则写出下列各数具有 5 位有效数字的近似数：

(1) 187.9325；(2) 0.03785551；(3) 8.000033.

10. 计算 $A=10^7(1-\cos 2°)$(用四位数学用表).

11. 设 $y=\ln x$，当 $x\approx a(a>0)$时，已知对数 $\ln a$ 的绝对误差限为$\frac{1}{2}\times 10^{-n}$，试估计真值 a 的相对误差限.

12. 已知 $I_n=\int_0^1 \frac{x^n}{4x+1}\,\mathrm{d}x$，试建立一具有较好数值稳定性的求 I_n 的递推公式.

13. 改变表达式$\int_N^{N+1}\ln x\mathrm{d}x=(N+1)\ln(N+1)-N\ln N-1$ (N 充分大)，以提高计算精度.

1.3.2 提示与解答

1. 因为 $x^*=0.358764\times 10^4$ 有 6 位有效数字，即 $m=4, l=n=6$，所以误差限 $\varepsilon(x^*)\leqslant \frac{1}{2}\times 10^{4-6}=0.005$.

2. 设需取 n 位有效数字，又 $8<\sqrt{70}<9$，可取 $\alpha_1=8$，则由

$$\frac{1}{2\alpha_1}\times 10^{-n+1}=\frac{1}{2\times 8}\times 10^{-n+1}\leqslant 0.1\%=10^{-3}$$

得 $n=3$.

所以至少应取 3 位有效数字.

3. 设 $\varepsilon(x^*)$是与 δ 对应的绝对误差限，则 $\delta=\frac{\varepsilon(x^*)}{x}$，于是 $\ln x$ 的相对误差限 $\tilde{\delta}\leqslant |f'(x)|\frac{\varepsilon(x^*)}{|\ln x|}=\frac{\varepsilon(x^*)}{x|\ln x|}=\frac{\delta}{|\ln x|}$.

4. 由求根公式得 $x_1=\frac{56+\sqrt{56^2-4}}{2}\approx 55.982$，再由韦达定理得

$$x_2=\frac{c}{ax_1}=\frac{1}{28+\sqrt{783}}\approx\frac{1}{55.982}\approx 0.017863.$$

5. $t=\frac{1}{x-1}$，$y=10+(3+(4-6t)t)t$.

6. $\frac{2}{\sqrt{2012}+\sqrt{2010}}$.

7. $1-\cos x=2\sin^2\frac{x}{2}$，$1-\cos x\approx 1-\left(1-\frac{x^2}{2}+\frac{x^4}{24}\right)=x^2\left(\frac{1}{2}-\frac{x^2}{24}\right)$.

8. (1) 当 $x\geqslant 0$ 时，直接计算 $f(x)=\ln(x+\sqrt{x^2+1})$；

(2) 当 $x<0$ 时,做恒等变形后计算: $f(x)=\ln(x+\sqrt{x^2+1})=-\ln(\sqrt{x^2+1}-x)$.

9. (1) 187.93; (2) 0.037856; (3) 8.0000.

10. 利用 $1-\cos x=2\sin^2\dfrac{x}{2}$,则 $A=2\times(\sin 1°)^2\times10^7=6.13\times10^3$(取 $\sin 1°=0.0175$).

11. 因为 $x^*=a>0$, $f(x)=\ln x$,故绝对误差 $\varepsilon(\ln a)\approx|f'(a)|\varepsilon(a)=\dfrac{1}{|a|}\varepsilon(a)$,于是 a 的相对误差限 $\varepsilon_r^*(a)\leqslant\varepsilon(\ln a)=\dfrac{1}{2}\times10^{-n}$.

12. 由 $4I_n+I_{n-1}=\displaystyle\int_0^1\frac{4x^n+x^{n-1}}{4x+1}\mathrm{d}x=\frac{1}{n}$,得 $I_n=\dfrac{1}{4n}-\dfrac{1}{4}I_{n-1}$, $n=1,2,\cdots$

13.
$$\begin{aligned}\int_N^{N+1}\ln x\mathrm{d}x&=(N+1)\ln(N+1)-N\ln N-1=\ln\frac{(N+1)^{N+1}}{N^N}-1\\&=\ln\left(1+\frac{1}{N}\right)^N+\ln(N+1)-1\approx\ln(N+1).\end{aligned}$$

1.4 数值实验

1.4.1 实验目的

了解 MATLAB 基本操作;了解数值计算过程中的误差种类和传递,以及避免误差的几种方法.

1.4.2 MATLAB 基本操作

1. 矩阵运算

MATLAB 软件最基本的语句是矩阵与数组运算.它可以非常方便地完成向量、矩阵的各种运算,如向量的加法与数乘,矩阵的加法、数乘、乘法、除法(求逆)和乘方等运算以及数组的点乘、点除等运算.

(1) 行向量

```
  >> x = [ - 3 1 7]          % 中间用空格将数据分开,也可以用逗号分开
x =
   - 3     1     7           % 显示输入结果,末尾输入分号则不显示输入结果
```

(2) 列向量

```
>> y = [0 2 9] '             % '表示转置
```

```
y =
     0
     2
     9
```

(3) 矩阵

```
>> A = [1 2 3;4 5 6;7 8 9]  % ; 表示换行
A =
     1    2    3
     4    5    6
     7    8    9
```

(4) 转置

```
>> B = A'
B =
     1    4    7
     2    5    8
     3    6    9
```

(5) 矩阵的加(减)法运算

加(减)法运算的基本要求：两矩阵(或向量)是同型的.例如

```
>> C = [ - 3 0 1; - 1 2 5;7 2 1];
>> D = A + C
D =
    - 2     2    4
      3     7   11
     14    10   10
```

但矩阵 $\boldsymbol{D}$ 可以与一个数字相加,又如

```
>> D - 5
ans =
     - 7     - 3     - 1
     - 2       2       6
       9       5       5
```

相当于每个元素加上－5.

(6) 乘法运算

矩阵乘法(*)运算,有数乘运算,即纯量与矩阵相乘,如

```
>> 2 * A
ans =
      2     4     6
      8    10    12
     14    16    18
```

矩阵与矩阵相乘，如

```
>> E = A * C
E =
    16    10    14
    25    22    35
    34    34    56
```

(7) 矩阵求逆

inv()函数是矩阵求逆运算函数，如

```
>> A = [1 2;3 5];
>> inv(A)
ans =
     -5.0000    2.0000
      3.0000   -1.0000
```

(8) 矩阵的除法运算

```
>> A = [1 2 3;4 5 6;7 8 0]; % 输入矩阵 A
>> b = [1 1 1]';            % 输入向量 b
>> x = A\b                  % 左除运算
x =
     -1.0000
      1.0000
     -0.0000
```

表示方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的解，即 $\boldsymbol{x}=\boldsymbol{A}^{-1}\boldsymbol{b}$. 因此

```
>> x = inv(A) * b
```

将得到同样的结果.

2. 数组运算

通常将向量或矩阵的点运算称为数组运算(加减运算无点运算)，其点运算有

```
.* 点乘    ./点右除    .\点左除    .^点乘幂
```

(1) 数组的输入

有两种方法构造一维数组. 第一种方法，用“：”构造一维等差数组，其格式为

数组＝初值：增量：终值，如

```
>> a = 0:3:10
a =
    0    3    6    9
```

当增量为 1 时，增量值可缺省，如

```
>> b = 1:6
b =
    1    2    3    4    5    6
```

第二种方法,用 linspace()函数构造一维等差数组,其格式为

数组= linspace(初值,终值,等分点数),如

```
>> c = linspace(1,12,5)
c =
    1.0000    3.7500    6.5000    9.2500    12.0000
```

(2) 数组加减法运算

数组的加减法运算与矩阵的加减法运算相同,就是通常意义下的加减法运算.

(3) 数组乘法运算

数组的乘法运算,也叫点乘(.*)运算,它表示两数组在同一位置上的元素相乘,如

```
>> x = [1 2 4];
>> y = [ - 2 1 8];
>> x. * y
ans =
    - 2    2    32
```

注:在数组的乘法中,点乘(.*)必须看成一个整体的运算符号.

(4) 数组除法运算

数组的除法运算,也叫点除(./或 .\)运算,两者都表示在同一位置上的元素相除,但意义上有差别.如

```
>> z = x.\y                %.\ 表示 z = y/x
z =
       - 2.0000    0.5000    2.0000
>> z = x./y                %./ 表示 z = x/y
z =
       - 0.5000    2.0000    0.5000
```

(5) 数组乘幂运算

数组的乘幂运算,也叫点乘幂(.^)运算,如

```
>> z = x.^2                %.^对每个数组中的元素作乘幂运算
z =
     1    4    16
```

3. 逻辑运算

在条件语句中有 &(与),|(或),~(非)等逻辑运算.

```
A&B                        %条件 A 与条件 B 同时成立,则为真(返回值为 1).
```