数值计算方法

第一章

数值计算引论

1.4 近似数的误差表示法

- 绝对误差
  - 设x^*是准确值x的一个近似值，则

$$e^* = x - x^*$$

称为近似值x^*的绝对误差，简称误差。

- 定义1.1: 如果

$$|e^*| = |x - x^*| \le \varepsilon^*$$

那么ε^* 叫做近似数x^*的绝对误差限，用它反映近似数的精度。

1.4 近似数的误差表示法

- 通常用$x = x^* \pm \varepsilon^*$来表示近似数$x^*$的精确度或准确值所在的范围，$\varepsilon^*$应该取得尽可能小
- 例$x = 4.3762816\ldots$，取近似数$x^* = 4.376$，则$x - x^* = 0.0002816\ldots$，有

  $|x - x^*| = 0.0002816 < 0.0003 = 0.3 \times 10^{-3}$

  同样 $|x - x^*| = 0.0002816 < 0.00029 = 0.29 \times 10^{-3}$
- 绝对误差限不是唯一的， ε^*越小，x^*近似真值x的程度越好，即x^*的精度越高
- 按四舍五入原则取近似值是使用最广泛的取近似值的方法

1.4 近似数的误差表示法

- 例1.4.1：用一把有毫米刻度的尺子，测量桌子的长度，读出来的值x^*＝1235mm，这是桌子实际长度x的一个近似值，由尺子的精度可以知道，这个近似值的误差不会超过1/2mm
  - $|x-x^*|=|x-1235\text{mm}|\leq 1/2\text{mm}$
  - $1234.5\leq x\leq 1235.5\text{mm}$
  - 这里$\varepsilon^*=0.5\text{mm}$ 是绝对误差限

四舍五入的误差限

- 设x为一个实数，其十进制的标准形式(十进制规格化浮点数形式)为：

$$x=\pm 0.x_1x_2\ldots\times 10^m$$

其中m是整数，$x_1,x_2,\ldots$是0,1,…,9中的任一数，但$x_1\neq 0$，经过四舍五入保留n位数字，得到近似值

$$x^*=\begin{cases}\pm 0.x_1x_2\ldots x_n\times 10^m, \text{当} x_{n+1}\leq 4 (\text{四舍})\\ \pm 0.x_1x_2\ldots x_{n-1}(x_n+1)\times 10^m, \text{当} x_{n+1}\geq 5 (\text{五入})\end{cases}$$

四舍五入的误差限

- 四舍时的误差限为

$$
\begin{aligned}
|x - x^*| &= (0.x_1x_2 \ldots x_n x_{n+1} \ldots - 0.x_1x_2 \ldots x_n) \times 10^m \\
&\leq (0.x_1x_2 \ldots x_n 499 \ldots - 0.x_1x_2 \ldots x_n) \times 10^m \\
&= 10^m \times 0.\underbrace{0 \ldots 0}_{\text{n 个 0}} 499 \ldots \leq \frac{1}{2} \times 10^{m-n}
\end{aligned}
$$

- 五入时的误差限为

$$
\begin{aligned}
|x - x^*| &= (0.x_1x_2 \ldots x_{n-1}(x_n + 1) - 0.x_1x_2 \ldots x_n \ldots) \times 10^m \\
&= (0.\underbrace{0 \ldots 0}_{n-1\text{个 0}} 1 - 0.\underbrace{0 \ldots 0}_{n\text{个 0}} x_{n+1} \ldots) \times 10^m \\
&\leq 10^{m-n} \times (1 - 0.x_{n+1})
\end{aligned}
$$

四舍五入的误差限

由于$x_{n+1}\geq 5$,所以$1-0.x_{n+1}\leq 1/2$，因此有

$$|x-x^*|\leq 1/2\times 10^{m-n}$$

- 所以，四舍五入得到近似数的绝对误差限是其末位的半个单位，即

$$\varepsilon^*=\frac{1}{2}\times 10^{m-n}$$

- 例1.4.2：圆周率$\pi=3.14159\ldots$，用四舍五入取小数点后4位时，近似值为3.1416，此时$m=1$，$n=5$，$m-n=1-5=-4$，绝对误差限$\varepsilon^*=1/2\times 10^{-4}$。取小数点后2位时，近似值为3.14，其绝对误差限$\varepsilon^*=1/2\times 10^{-2}$

四舍五入的误差限

- 例1.4.3：取3.141和3.142作为π的近似值的绝对误差限为多少？
  - 解：

$$|\pi - 3.141| = 0.00059... < 0.005 = \frac{1}{2} \times 10^{-2}$$

$$|\pi - 3.142| = 0.0004073... < 0.0005 = \frac{1}{2} \times 10^{-3}$$

相对误差

- 定义1.2： x的近似值x^*的相对误差限为

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

  - 相对误差限可由绝对误差限求出，反之，绝对误差限也可由相对误差限求出，即

$$\varepsilon^* = |x^*|\varepsilon_r^*$$

  - 在实际计算中，相对误差不可能准确的得到，象绝对误差一样，只能估计相对误差的范围

相对误差

- 例1.4.5：取3.14作为π的四舍五入的近似值，试求其相对误差限
  - 解：

$$\varepsilon_r^* = \frac{\varepsilon^*}{|x^*|} = \frac{1/2\times10^{-2}}{3.14} = 0.159\%$$

有效数字

- 定义1. 3：设数x的近似值$x^*=0.x_1x_2\ldots x_n\times10^m$，其中$x_i$是0到9之间的任一个数，但$x_1\neq0$，$i=1,2,3,\ldots,n$正整数，$m$整数，如果

$$|x-x^*|\leq1/2\times10^{m-n}$$

则称x^*为x的具有n位有效数字的近似值，x^*准确到第n位，$x_1x_2\ldots x_n$是x^*的有效数字。

有效数字

- 例1.4.6：$\pi=3.141592\ldots$，当取3.142和3.141作为其近似值时，有效数字分别为多少位？
  - 解：

    $|\pi-3.142|=0.000407<0.0005=1/2\times10^{-3}$

    即$m-n=-3$，$m=1$, $n=4$, 所以3.142作为π的近似值具有4位有效数字

    当取3.141作为π的近似值时

    $|\pi-3.141|=0.00059<0.005=1/2\times10^{-2}$

    即$m-n=-2$, $m=1$, $n=3$, 所以3.141作为π的近似值时有3位有效数字

有效数字

- 例**1.4.7**：下列近似值的有效数字分别多少位

$\pi - 3.14 = 0.0015926$ 有效数位为**3**位

$\pi - 3.1416 = -0.0000074$ 有效数位为**5**位

$\pi - 3.1415 = 0.0000926$ 有效数位为**4**位

$\pi - 3.14159 = 0.0000026$ 有效数位为**6**位

有效数字

- 有效数字的确定
  - 用四舍五入取准确值的前n位x^*作为近似值，则x^*必有n个有效数字
    - 例如，$\pi=3.1415926\ldots$，取3.14作为近似值，则有3位有效数字，取3.142作为近似值，则有4位有效数字
  - 有效数字位数相同的两个近似数，绝对误差不一定相同
    - 例如，设$x_1^*=12345$, $x_2^*=12.345$，二者均有5位有效数字，前者的绝对误差为1/2，后者的绝对误差为$1/2\times10^{-3}$

有效数字

- 把任何数乘以10^p等于移动该数的小数点，这样并不影响其有效数字的位数
  - 例如，$g=9.80\text{m/s}^2$具有3位有效数字，$g=0.00980\times10^3\text{m/s}^2$也具有3位有效数字
  - 如果整数并非全是有效数字，则可用浮点数表示，如$x^*=300\times10^3$或0.300×10^6表示绝对误差限不超过500，而300000则表示其绝对误差限不超过0.5
- 准确值被认为具有无穷多位有效数字
  - 例如，真值0.5，不表示只有1位有效数字，具有无穷多位有效数字

1．5 数值稳定性和减少运算误差

- 定义1.4：如果在执行算法的过程中，舍入误差在一定条件下能够得到控制（或者说舍入误差的增长不影响产生可靠的结果），则该算法是数值稳定的，否则是数值不稳定的。
- 例1.5.1：计算积分

$$I_n = \int_0^1 x^n e^{x-1} \mathrm{d}x$$

  - 解：根据分部积分法可得

1．5 数值稳定性和减少运算误差

$$I_n = \int_0^1 x^n e^{x-1} \mathrm{d}x = \int_0^1 x^n \mathrm{d}e^{x-1} =$$

$$x^n e^{x-1}\Big|_0^1 - n\int_0^1 x^{n-1} e^{x-1} \mathrm{d}x = 1 - nI_{n-1}$$

当n=0时

$$I_0 = \int_0^1 e^{x-1} \mathrm{d}x = 1 - e^{-1} \approx 0.6321205$$

- 方法一

$$I_n = 1 - nI_{n-1}, \quad I_0 = \int_0^1 e^{x-1} \mathrm{d}x = 1 - e^{-1} \approx 0.6321$$

按照递推关系计算其余的值如下表第二列

1．5 数值稳定性和减少运算误差

- 方法二

  如果将递推公式改成

$$I_{n-1} = (1 - I_n)/n$$

  根据积分估算式

$$e^{-1}\left(\min_{0\le x\le 1} e^x\right)\int_0^1 x^n dx < I_n < e^{-1}\left(\max_{0\le x\le 1} e^x\right)\int_0^1 x^n dx$$

  由此产生估计式

$$\frac{e^{-1}}{n+1} < I_n < \frac{1}{n+1}$$

1．5 数值稳定性和减少运算误差

当n=8时，有0.0409<I_n <0.1111，取初值I_8 =0.1035，按上述递推式倒推计算，计算中四舍五入取小数点后四位，得到的结果如下表第三列所示

I_n	方法一	方法二	准确值
I_0	0.6321	0.6321	0.6321
I_1	0.3679	0.3679	0.3669
I_2	0.2642	0.2642	0.2642
I_3	0.2074	0.2073	0.2073
I_4	0.1704	0.1709	0.1709
I_5	0.1480	0.1455	0.1455
I_6	0.1120	0.1269	0.1268
I_7	0.2160	0.1120	0.1124
I_8	-0.7280	0.1035	0.1008

1．5 数值稳定性和减少运算误差

- 从第二列可以看到采用方法一计算过程中，I_8为负值，显然与$I_n>0$相矛盾，因此这个递推公式就是不稳定的
- 根据第三列的计算结果，方法二的精确度比较高。这样计算的精确度较高的原因是I_8的误差传播到I_7时要乘以1/8，误差是逐渐减少的。因此方法二的递推公式是数值稳定的公式

减少运算误差的若干原则

- 两个相近的数相减，会严重损失有效数字
  - 设$y=x-A$

    其中A和x均为准确值，假设A运算时不发生误差，而x有误差，其近似值为x^*，由此可估计出当用x^*近似代替x时，y的相对误差

$$\varepsilon_r{}^*(y^*)=\frac{\varepsilon^*(y^*)}{|y^*|}=\frac{|(x-A)-(x^*-A)|}{|x^*-A|}=\frac{|x-x^*|}{|x^*-A|}=\frac{\varepsilon^*(x^*)}{|x^*-A|}$$

    由上可以看出，在x的绝对误差$\varepsilon^*(x^*)$不变时，如果x^*越接近A，那么y的相对误差$\varepsilon_r{}^*(y^*)$会变得越大，而相对误差的增大必然会导致有效数字位数的减少

减少运算误差的若干原则

- 例1.5.2：当$x=10003$时，计算 $\sqrt{x+1}-\sqrt{x}$ 的近似值

解：如果使用6位十进制浮点运算，运算时取6位有效数字，结果为

$$\sqrt{x+1}-\sqrt{x}=100.020-100.015=0.005$$

只有一位有效数字，损失了5位有效数字

如果改用

$$\sqrt{x+1}-\sqrt{x}=\frac{1}{\sqrt{x+1}+\sqrt{x}}=\frac{1}{100.020+100.015}=0.00499913$$

则其结果有6位有效数字，与精确值

0.00499912523117984…非常接近

减少运算误差的若干原则

- 例1.5.3：x_1=1.99999，x_2=1.99998，求lgx_1-lgx_2
  - 解：如果使用6位十进制浮点运算，运算时取6位有效数字，则lgx_1-lgx_2≈0.301028-0.301026 =0.000002
    只有一位有效数字，损失了5位有效数字。
    如果改用lgx_1-lgx_2= $\lg\frac{x_1}{x_2}$ ≈2.17149×10^{-6}，

    则其结果有6位有效数字，与精确值
    2.171488695634…×10^{-6}非常接近

减少运算误差的若干原则

- 当x接近0时

$$\frac{1-\cos x}{\sin x}=\frac{\sin x}{1+\cos x}$$

- 当x充分大时

$$arctg(x+1)-arctgx=arctg\frac{1}{1+x(x+1)}$$

- 如果计算公式不能改变，可采用增加有效数字位数的方法

减少运算误差的若干原则

- 例1.5.4：已知2.01和2皆为准确数，计算 $u=\sqrt{2.01}-\sqrt{2}$，使具有3位有效数字。

  - 解：当取3位有效数字时，有

$$u=\sqrt{2.01}-\sqrt{2}=1.42-1.41=0.01$$

此时结果只有一位有效数字，如果增加到6位有效数字时，计算结果为

$$u=\sqrt{2.01}-\sqrt{2}=1.41774-1.41421=3.53\times10^{-3}$$

此时有三位有效数字

减少运算误差的若干原则

- 防止大数“吃掉”小数
  - 计算机的位数有限，在进行加减法运算时，要对阶和规格化
  - 例如在四位浮点机上作运算

    $0.7315\times10^{3}+0.4506\times10^{-5}$

    对阶是$0.7315\times10^{3}+0.000000004506\times10^{3}$，计算结果为$0.7315\times10^{3}$，结果是大数“吃掉”了小数
  - 当参加运算的两个数的数量级相差很大时，如果不注意运算次序，就有可能把数量级小的数“吃掉”

减少运算误差的若干原则

- 例如：已知$A=10^5$，$B=5$，$C=-10^5$

  如果按照（A＋B）＋C进行运算，则结果接近于零，结果失真；如果按照（A＋C）＋B进行计算，则结果接近于正确的结果5

- 如果事先大致估计一下计算方案中各数的数量级，编制程序时加以合理的安排，那么重要的小数就可以避免被“吃掉”

减少运算误差的若干原则

- 例1.5.5：求二次方程$x^2-(10^9+1)x+10^9=0$的根。
  - 解：用因式分解$(x-10^9)(x-1)=0$易得方程的二个根为$x_1=10^9, x_2=1$，但用求根公式

$$x_{1,2}=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

编制程序，如果只能在将数表示到小数点后8位的计算机上运算，那么首先要对阶

$$-b=10^9+1=0.10000000\times10^{10}+0.0000000001\times10^{10}$$

$$=10^9$$

减少运算误差的若干原则

$$\sqrt{b^2-4ac}=|b|=10^9$$

因此所得的两个根为$x_1\approx10^9$，$x_2\approx0$。x_2失真的原因是大数吃掉了小数的结果。

- 如果把x_2的计算公式写成

$$x_2=\frac{-b-\sqrt{b^2-4ac}}{2a}=\frac{2c}{-b+\sqrt{b^2-4ac}}$$

可得

$$x_2=\frac{2\times10^9}{10^9+10^9}=1$$

减少运算误差的若干原则

- 在除法运算中要避免出现除数的绝对值远远小于被除数绝对值的情形
  - 在用计算机实现算法的过程中，如果用绝对值很小的数作除数，往往会使舍入误差增大，对计算结果带来严重影响，应该尽量避免。
- 例1.5.6：计算

$$\frac{x^*}{y^*}=\frac{2.7182}{0.001}$$

的值。

- 解：

$$\frac{x^*}{y^*}=\frac{2.7182}{0.001}=2718.2$$

减少运算误差的若干原则

如果分母变为0.0011，分母的变化只有0.0001，则

$$\frac{x^*}{y^*} = \frac{2.7182}{0.0011} = 2471.1$$

- 商引起了巨大的变化，因此在算法设计时，要尽量避免在算法的计算公式中出现太小的除数

减少运算误差的若干原则

- 例1.5. 7：在4位浮点十进制数下，用消去法解线性方程组

$$\begin{cases} 0.00003x_1 - 3x_2 = 0.6 \\ x_1 + 2x_2 = 1 \end{cases}$$

- 解：仿计算机实际计算，将上述方程组写成

$$\begin{cases} 0.3000\times10^{-4}x_1 - 0.3000\times10^{1}x_2 = 0.6000\times10^{0} \\ 0.1000\times10^{1}x_1 + 0.2000\times10^{1}x_2 = 0.1000\times10^{1} \end{cases}$$

（2）—（1）÷（0.3000×10^{-4}）得到下述方程组

减少运算误差的若干原则

$$\begin{cases} 0.3000\times10^{-4}x_1 - 0.3000\times10^{1}x_2 = 0.6000\times10^{0} \\ 0.1000\times10^{6}x_2 = -0.2000\times10^{5} \end{cases}$$

解得 $x_1=0, x_2=-0.2$

准确解为$x_1=1.399972\ldots$， $x_2=-0.199986\ldots$

- 如果反过来用第二个方程消去第一个方程中含x_1的项，那么就可以避免很小的数作除数的情形，即（1）－（2）×（0.3000×10^{-4}），得

$$\begin{cases} -0.3000\times10^{1}x_2 = 0.6000\times10^{0} \\ 0.1000\times10^{1}x_1 - 0.2000\times10^{1}x_2 = 0.1000\times10^{1} \end{cases}$$

解得 $x_1=1.4, x_2=-0.2$

减少运算误差的若干原则

- 简化计算步骤，减少运算次数
  - 必须要考虑尽量简化计算步骤，这样一方面可以减小计算量，另一方面由于减少了运算次数，从而减少了产生误差的机会，也使误差积累可能减小
- 例1.5.8：计算x^{255}
  - 解：如果逐个相乘，要做254次乘法，但若改成

    $x^{255}=x\ x^2\ x^4\ x^8\ x^{16}\ x^{32}\ x^{64}\ x^{128}$

    只要14次乘法运算即可。

减少运算误差的若干原则

- 例1.5.9：计算和式 $\sum_{n=1}^{1000}\frac{1}{n(n+1)}$ 的值

  - 解：如果直接逐项求和，运算次数多且误差积累，但是可以进行化简：

$$\sum_{n=1}^{1000}\frac{1}{n(n+1)}=\sum_{n=1}^{1000}(\frac{1}{n}-\frac{1}{n+1})=(\frac{1}{1}-\frac{1}{2})+(\frac{1}{2}-\frac{1}{3})+...+(\frac{1}{1000}-\frac{1}{1001})=1-\frac{1}{1001}$$

  那么整个计算只要一个求倒数和一次减法就可以

减少运算误差的若干原则

- 选用数值稳定性好的计算公式
  - 在构造算法时，还要考虑算法的稳定性
  - 只有稳定的数值方法才可能给出可靠的计算结果，不稳定的数值方法毫无实用价值