目　录

第一章　近似数及其运算

在实际问题的数值计算中，理想的准确值(或真值)往往得不到，人们常常采用与准确值相近的数值来代替，而两种结果的误差程度便是人们所关心的，因此，要对其误差的程度进行必要的估计。

§1 误　　差

1.1 误差的来源

用数值计算方法解决实际问题时可能会遇到下述两种误差：一种误差是由于在进行具体计算时粗枝大叶、疏忽而产生的，称其为“过失误差”，例如计算中把989误写成998，此外误用公式等等，这种误差只要在实际计算中做到仔细、谨慎即可避免；另一种误差就是下面所要讨论的且按照一般规律计算时所不可避免的误差，即所谓模型、观测、截断、舍入误差等。下面具体介绍这些误差的来源。

首先，在数值计算中必须根据具体问题的自然现象建立数学模型，例如，用公式

$$pV = 常数$$

来计算气体的体积或压力。然而数学模型总是忽略一些次要因素才能获得，故其只是近似地描述所给问题的自然现象，其与自然现象之间有一定的差别，从而出现误差，这通常称为“模型误差”。

其次，在数学模型的建立时所用到的数据往往是通过观测得来的，而观测的结果不可能绝对准确，总是近似的，因而就产生了误差，这种误差通常称之为“观测误差”，由此可知，在数值计算中参与计算的数一般说来都是近似的。

再则，在计算过程中，通常用收敛的无穷级数前几项和作为该级数的近似值，其余项称之为“截断误差”。例如 $\sin x$ 和 $\ln(1+x)$ 能够写成无穷级数展开式：

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \quad (-\infty < x < +\infty),$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \quad (-1 < x \leqslant 1)。$$

当 x 很小时，通常采用 x 来代替 $\sin x$ 以及 $\ln(1+x)$。这就是省去了无穷级数的后段部分，因而引起了“截断误差”。显然，前者的截断误差约为 $\frac{1}{6}x^3$；后者的截断误差约为 $\frac{1}{2}x^2$。

此外，在具体的运算过程中，往往还要利用一些无理数及循环小数，例如 π、e、$\sqrt{2}$、$\frac{1}{3}$ 等，在计算机上或人工手算时这些数只能近似地表示，这样就引起了“舍入误差”。其实，在实际计算中总是按有限位数进行的，所以数值计算的每一步都不可避免有舍入误差的影响。

以上说明，在数值计算过程中常会出现诸如模型、观测、截断、舍入等误差。但是，一旦数学模型建立后人们只考虑后两种误差即截断误差和舍入误差。在计算机上进行计算时，误差的影响非同小可，有时误差会严重泛滥而完全“吃掉”所要求的真值，致使计算结果失去根本的意义。所以，在进行数值计算时，首先必须保证精度，从而应当在满足精度要求的前提下选择、设计好的计算方法。

1.2 绝对误差和绝对误差限

定义 假设某一量的准确值是x，其近似值为x^*，则x与x^*差的绝对值

$$\varepsilon(x)=|x-x^*| \tag{1}$$

叫做近似值x^*的绝对误差，简称误差。

$\varepsilon(x)$的大小标志着x^*的精确度，一般在同一量的不同近似值中，$\varepsilon(x)$越小，x^*的精确度越高。

由于准确值x一般不能算出，误差$\varepsilon(x)$的准确值也不能求出，但根据具体测量或计算的情况，可事先估计出它的大小范围，就是可以指出一个正数η，使

$$\varepsilon(x)=|x-x^*|\leqslant\eta。 \tag{2}$$

称η为近似值x^*的绝对误差限，即为绝对误差的"上界"。显然有

$$x^*-\eta\leqslant x\leqslant x^*+\eta,$$

有时也可用

$$x=x^*\pm\eta \tag{3}$$

表示近似值的精度或准确值的所在范围。例如测量某一物件是5 m长，其误差限为0.01m，通常将准确长度s记为

$$s=5\pm0.01\ (\mathrm{m}),$$

即准确值在5 m左右，但不超过0.01m的误差界限。

1.3 相对误差和相对误差限

绝对误差并不足以表示近似值的好坏，例如测量10m长度时发生了1cm的误差和测量1m长度时发生了1cm的误差是有区别的，前者测量比较精确。可见决定一个量的近似值的精确度除了要看绝对误差的大小之外，还必须考虑到该量本

身的大小。据此，下面引进相对误差的概念。

定义 绝对误差与准确值的绝对值之比

$$\varepsilon=\frac{\varepsilon(x)}{|x|}=\left|\frac{x-x^*}{x}\right|(x\neq 0) \tag{4}$$

称为x^*的相对误差。

显见，上面前者测量的相对误差为1/1000；而后者测量的相对误差是1/100，其是前者的10倍。一般地，在同一量或不同量的几个近似值中，ε小者为精确度高。

由上述可知，相对误差可以从绝对误差求出；反之，绝对误差也可由相对误差求得，其关系是

$$\varepsilon(x)=|x|\varepsilon。$$

相对误差不仅可表示绝对误差，而且在讨论对近似数进行运算结果的误差分析时，其更能反映出误差的特征，因此在误差分析中相对误差比绝对误差显得更为重要。由于$\varepsilon(x)$与x都不能准确地求得，那么相对误差ε也不可能准确地求出，但也像绝对误差那样，可以估计它的大小范围，即指出一个正数δ，使

$$\varepsilon=\frac{\varepsilon(x)}{|x|}\leqslant\delta, \tag{5}$$

称δ为x^*的相对误差限，也即为相对误差的"上界"。但在实际计算中，由于准确值x为未知数，所以取

$$\varepsilon=\frac{\varepsilon(x)}{|x^*|} \tag{6}$$

作为x^*的相对误差的另一定义。

1.4 有效数字

对于一个近似值，我们还希望知道它的准确程度，为此

再引进有效数字的概念。

定义 若近似值x^*的绝对误差限是某一位上的半个单位，且该位直到x^*的第一位非零数字一共有n位，则称近似值x^*有n位**有效数字**，或说x^*精确到该位。

例1 设$x=\pi=3.141\,592\,6\cdots$，那么

$x_1^*=3$，$\varepsilon_1(x)=0.141\,5\cdots\leqslant 0.5\times 10^0$，

x_1^*的有效数字为1位，或说x_1^*精确到个位；

$x_2^*=3.14$，$\varepsilon_2(x)=0.001\,59\cdots\leqslant 0.5\times 10^{-2}$，

x_2^*的有效数字为3位，或说x_2^*精确到0.01；

$x_3^*=3.141\,6$，$\varepsilon_3(x)=0.000\,007\,34\cdots\leqslant 0.5\times 10^{-4}$，

x_3^*的有效数字为5位，或说x_3^*精确到0.000 1；

$x_4^*=3.141\,5$，$\varepsilon_4(x)=0.000\,092\,6\cdots\leqslant 0.5\times 10^{-3}$，

x_4^*的有效数字为4位，或说x_4^*精确到0.001。

用四舍五入方法取准确值x的前n位作为其的近似值x^*，则x^*有n位有效数字，其中每一位数字(包括后面的零)都叫做x^*的有效数字。

例2 设$x=4.269\,72$，那么

取2位，$x_1^*=4.3$，有效数字为2位；

取3位，$x_2^*=4.27$，有效数字为3位；

取4位，$x_3^*=4.270$，有效数字为4位；

取5位，$x_4^*=4.269\,7$，有效数字为5位。

值得注意的是，近似值后面的零不能随便省去。例如，2.18和2.180 0，前者精确到0.01，其有3位有效数字；而后者精确到0.000 1，其有5位有效数字。可见，它们的近似程度完全不同。

此外，下面再给出可靠数字与可疑数字的概念。

定义 若近似值x^*的绝对误差限为某一位上的一个单

位，且该位直到第一位非零的数字共有n位，则称x^*为具有n位可靠数字的近似值。

例3 设$x=5.382$，那么

$x_1^*=5.3$，$\varepsilon_1(x)=0.082\leqslant 1\times 10^{-1}$，

x_1^*的可靠数字位数为2；

$x_2^*=5.38$，$\varepsilon_2(x)=0.002\leqslant 1\times 10^{-2}$，

x_2^*的可靠数字位数为3。

上例中x_1^*具有1位有效数字，x_2^*具有3位有效数字。由此可知有效数字一定为可靠数字，它要比可靠数字精确，因此平常一般都采用有效数字的概念。如果近似值x^*有k位有效数字(或可靠数字)，则左起第k位数字称为**可疑数字**。

特别指出，两个相近的近似数相减可能会造成有效数字的严重损失，实际计算时要尽量避免这种情况的发生。

例4 设要对$a=1000$取四位有效数字计算

$$x=\sqrt{a+1}-\sqrt{a}。$$

由于

$$\sqrt{a+1}=31.64,\ \sqrt{a}=31.62,$$

两者相减得

$$x=0.02,$$

其结果为一位有效数字（x的实际值为0.015 807 437），但若先考虑计算公式

$$x=\sqrt{a+1}-\sqrt{a}=\frac{1}{\sqrt{a+1}+\sqrt{a}},\tag{7}$$

则它的计算结果

$$x=0.015\,81$$

有四位有效数字。可见计算公式写成怎样的形状直接影响到计算结果的精度，且其影响可能很大。

在计算机上进行计算时，计算机要求输入的数有一定多位数字，其计算结果也只保留一定多位数字，因此在计算的结果中，所保留下来的就不一定都是有效数字，同时也不是所有的有效数字都保留下来，于是不能把所得的数字都视为有效数字，这些数字中有些是有效数字，而有些则不是，甚至有可能一个有效数字也没有，实际计算时要引为注意。

§2 近似数的算术运算及算法的数值稳定性

2.1 算术运算法则

1. 加减运算

近似数进行加减时，把小数位数较多的四舍五入，使其比小数位数最少的数多一位小数，计算结果保留的小数位数与原近似数中小数位数最少者相同。

例 $x=2.718$，$y=3.408\ 23$，则

$$x+y\approx 2.718+3.408\ 2\approx 6.126,$$

$$x-y\approx 2.718-3.408\ 2\approx -0.690.$$

2. 乘除运算

近似数相乘除时，各因子保留的位数应比有效数字位数最少者的位数多一位，所得结果的可靠数字位数与原近似值中有效数字位数最少者的位数相等。

例 $x=2.501$，$y=1.2$，则

$$x\times y\approx 2.50\times 1.2\approx 3.0,$$

$$x/y\approx 2.50/1.2\approx 2.1.$$

3. 乘方与开方运算

近似数在进行乘方与开方时，原来近似值有几位有效数

字，计算结果仍保留几位数字。

例 $x=1.2$，则

$x^2\approx1.4$；$\sqrt{x}\approx1.1$。

4．对数运算

在进行对数运算时，所取对数的位数应与其真数的有效数字的位数相等。

例 $\lg2.718\approx0.4343$，

$\lg0.0618\approx\bar{2}.791\approx-1.21$。

在进行实际计算过程中，中间的结果应比上述各法则所提及的位数多取一位，在进入最后一次计算时这一位要舍入。另外，如果计算结果是由加减法求得，则原始数据的小数位数应比结果所要求的多一位；若计算结果是由乘除、开方、乘方求得，则原始数据的有效数字位数应比结果所要求的位数多一位。

以上谈及的是在人工手算时所必须遵循的运算法则，但用电子计算机计算时根本用不着这样繁琐的操心和关注，这是因为机器是自动进行有限位运算的。

2.2 算法的数值稳定性

实际计算中，对于同一问题选用不同的算法得到的结果往往大不相同，这是因为初始数据的误差或计算中的舍入误差在计算过程中的传播因算法不同而相异，于是就产生了算法的数值稳定性问题。一个算法如果计算结果受误差的影响小，便称之具有较好的稳定性，否则反之。为了说明算法稳定性这个概念，下面列举几例。

例5 二次方程

$$x^2+2px-q=0$$

的两个根分别为

$$x_1 = -p + \sqrt{p^2 + q},$$

$$x_2 = -p - \sqrt{p^2 + q}。$$

当$p = -0.5 \times 10^5$及$q = -1$时，方程的两个根取 11 位有效数字为

$$x_1 = 99\,999.999\,990,$$

$$x_2 = 0.000\,010\,000\,000\,001,$$

而在字长为 8 、基底为10的计算机上直接用上述公式计算的结果为

$$x_1 = 100\,000.00,$$

$$x_2 = 0,$$

可见x_1很好，而x_2很坏。这说明直接利用上述公式计算第二个根是不稳定的，但是根据根与系数的关系可知

$$x_1 x_2 = -q = 1,$$

则

$$x_2 = 1/x_1。$$

因此，若用前述方法算出 x_1，然后用公式 $x_2 = 1/x_1$ 来计算x_2便得到

$$x_1 = 100\,000.00,$$

$$x_2 = 0.000\,010\,000\,000,$$

其结果均非常好。

此上说明第二种算法是稳定的或说其具有较好的数值稳定性。还得说明，在使用根与系数的关系式求根时，我们要先算出绝对值较大的根，然后再计算另一根，这样可得到精度较高的结果。

例6 计算积分

$$E_n=\int_0^1 x^n e^{x-1}dx \quad (n=1,2,\cdots)。$$

利用分部积分法得

$$E_n=x^n e^{x-1}\Big|_0^1-n\int_0^1 x^{n-1}e^{x-1}dx,$$

从而有递推公式

$$E_n=1-nE_{n-1} \quad (n=2,3,\cdots), \tag{8}$$

$$E_1=1/e。$$

应用上面递推公式在字长为 6 ，基底为10的计算机上从E_1出发计算前几个积分值，其结果如下表：

k	E_k
1	0.367 879
2	0.264 242
3	0.207 274
4	0.170 904
5	0.145 480
6	0.127 120
7	0.110 160
8	0.118 720
9	−0.068 480

然而被积函数 $x^n e^{x-1}$在整个区间(0，1)内部是正的，其E_9取三位有效数字的精确结果为 0.091 6 ，但这里却得到了负的积分值$E_9=-0.068\ 480$。

为什么会发生这种情况？在上述计算中唯一舍入误差为计算E_1时因舍入(取六位有效数字)而产生的误差，其大小约为

$$\varepsilon=4.412\times10^{-7},$$