第一章 误差

数值计算方法(也称计算方法，数值方法)：是研究科学与工程技术中数学问题的数值解及其理论的一个数学分支，它的涉及面很广，涉及代数、微积分、微分方程数值解等问题。

- **数值计算方法的主要任务**：研究适合于在计算机上使用的数值计算方法及与此相关的理论，如方法的收敛性、稳定性以及误差分析等，此外，还要根据计算机的特点研究计算时间最短、需要计算机内存最少等计算方法问题.
- **数值计算主要过程**：实际问题→建立数学模型→设计高效、可靠的数值计算方法→程序设计→上机计算求出结果。
- **数值计算方法不同于纯数学**：它既具有数学的抽象性与严格性，又具有应用的广泛性与实际试验的技术性，它是一门与计算机紧密结合的实用性很强的有着自身研究方法与理论系统的计算数学课程。
- **数值计算方法的特点**：应提供能让计算机直接处理的，包括加减乘除运算和逻辑运算及具有完整解题步骤的，切实可行的有效算法与程序，它可用框图、算法语言、数学语言或自然语言来描述，并有可靠的理论分析，能逼近且达到精度要求，对近似算法应保证收敛性和数值稳定性、进行必要的误差分析。此外，还要注意算法能否在计算机上实现，应避免因数值方法选用不当、程序设计不合理而导致

超过计算机的存贮能力，或导致计算结果精度不高等.

根据“数值计算”的特点，首先应注意掌握数值计算方法的**基本原理和思想**，注意方法处理的**技巧**及其与计算机的密切结合，重视**误差分析**、**收敛性**及**稳定性**的基本理论；其次还要注意方法的**使用条件**，通过各种方法的比较，了解各种方法的异同及优缺点。

§1　误差的来源

- **误差可分为以下四种：模型误差、观测误差、截断误差、舍入误差。**

1．模型误差

用数值计算方法解决实际问题时，首先必须建立数学模型。由于实际问题的复杂性，在对实际问题进行抽象与简化时，往往为了抓住主要因素而忽略了次要因素，这就会使得建立起来的数学模型只是复杂客观现象的一种近似描述，它与实际问题之间总会存在一定的误差．我们**把数学模型与实际问题之间出现的这种误差称为模型误差。**

2．观测误差

在数学模型中往往包含一些由观测或实验得来的物理量，由于工具精度和测量手段的限制，它们与实际量大小之间必然存在误差，即称为**观测误差**。

3．截断误差

由实际问题建立起来的数学模型，在很多情况下要得到准确解是困难的，通常要用数值方法求出它的近似解。例如常用有限过程逼近无限过程，用能计算的问题代替不能计算的问题。这种**数学模型的精确解与由数值方法求出的近似解之间的误差称为截断误差**，又称为**方法误差。**

例如用函数 $f(x)$ 的泰勒(Taylor)展开式的部分和 $S_n(x)$ 去近似代替 $f(x)$，其余项 R_n 就是真值 $f(x)$ 的截断误差：

$$\boldsymbol{sinx = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \cdots \approx x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \triangleq S_5(x)}$$

截断误差： $\boldsymbol{R_5(x) \triangleq sinx - S_5(x) = -\frac{1}{7!}x^7 + \cdots}$

4．舍入误差

用计算机进行数值计算时，由于计算机的位数有限，计算时只能对超过位数的数字进行四舍五入，由此产生的误差称为**舍人误差**。例如用 2.71828 作为无理数 e 的近似值产生的误差就是舍入误差．应**注意的是，虽然少量的舍入误差是微不足道的，但在计算机上完成了千百万次运算之后，舍入误差的积累却可能是十分惊人的。**

- **截断误差和舍入误差是由计算方法所引起的，是数值计算方法的主要研究对象。**讨论它们在计算过程中的传播和对计算结果的影响，并找出误差的界，对研究误差的渐近特性和改进算法的近似程度具有重要的实际意义。

§2　绝对误差、相对误差和有效数字

2.1　绝对误差与绝对误差限

设某一量的精确值为 x，其近似值为x^*，则称：

$$e^*(x) = x - x^* \qquad (1.2)$$

为近似值x^*的**绝对误差**，简称**误差**。

一般地，某一量的精确值 x 是不知道的，因而$e^*(x)$也无法求出，但往往可以估计出$e^*(x)$**的上界：**

即存在正数 ε^*，使得

$$|e^*(x)| = |x - x^*| \leq \mathcal{E}^* \qquad (1.3)$$

称ε^*为近似值x^*的**绝对误差限**，简称**误差限**或**精度**。

ε^*越小，表示近似值x^*的精度越高。

在工程技术中，常将 $x^*-\varepsilon^* \le x \le x^*+\varepsilon^*$，表示为 $x = x^* \pm \varepsilon^*$ 表示近似值x^*的精度或精确值 x 的所在范围，**绝对误差是有量纲的。**

例如，$v = (100 \pm 2)V$ 表示 $v^* = 100V$是电压v的一个近似值，$2V$是近似值 v^* 的一个绝对误差限，即：$|v - v^*| \le 2V$；又如，用毫米刻度的直尺去测量一个长度为 x 的物体，测得其近似值为$x^* = 84mm$，由于直尺以毫米为刻度，所以其误差不超过0.5mm，即 $|x-84| \le 0.5(mm)$。这样，虽然不能得出准确值 x 的长度是多少，但可以知道 x 范围是 $83.5mm \le x \le 84.5mm$，即 x 必在[83.5mm, 84.5mm]内。

例 求 $x^* = 3.14$ 与 π 的绝对误差。

解 由于 $3.1415 < \pi < 3.1416$，得

$$|x^* - \pi| \le |3.14 - 3.1416| = 0.0016$$

即其绝对误差限为 0.0016。

2.2 相对误差与相对误差限

用绝对误差来刻画一个近似值的精确程度是有局限性的，在很多场合无法显示出近似值的准确程度，如测量 100m 和 10m 两个长度若它们的绝对误差都是 1cm，显然前者的测量结果比后者的准确。由此可见，决定一个量的近似值的精确度，除了要看绝对误差的大小外，还必须**考虑该量本身的大小，**为此引入相对误差的概念。

称**绝对误差与精确值之比**，即：

$$e_r(x)=\frac{e^*(x)}{x}=\frac{x-x^*}{x} \tag{1.4}$$

为近似值 x^* 的**相对误差。**

注：在实际计算中，精确值 x 往往是不知道的，所以常将：

$$e_r^*(x)=\frac{e^*(x)}{x^*}=\frac{x-x^*}{x^*} \tag{1.5}$$

作为近似值 x^* 的**相对误差。**

$$e_r(x)-e_r^*(x)=\frac{e^*(x)}{x}-\frac{e^*(x)}{x^*}=-\frac{\left(\frac{e^*(x)}{x^*}\right)^2}{1+\frac{e^*(x)}{x^*}}=-\frac{\left(e_r^*(x)\right)^2}{1+e_r^*(x)}$$

类似于绝对误差，若存在正数 $\mathcal{E}_r^*$，使得

$$|e_r^*(x)|=|\frac{e^*(x)}{x^*}|=|\frac{x-x^*}{x^*}|\le\mathcal{E}_r^* \tag{1.6}$$

则称 $\mathcal{E}_r^*$ 为近似值 x^* 的**相对误差限，相对误差是无量纲的数，**通常用百分比表示，亦称**百分误差**。

例 求 $x^*=3.14$ 与 π 的相对误差限。

解 由于 $3.1415<\pi<3.1416$，因此

$$|e_r^*(\pi)|=|\frac{e^*(\pi)}{x^*}|=|\frac{\pi-3.14}{3.14}|\le|\frac{3.1416-3.14}{3.14}|<0.00051$$

所以相对误差限为 0.00051。

又如测量 100m 和 10m 两个长度，若它们的绝对误差都不超过 1cm，根据上述定义可知，当$|x-x^*|\le 1cm$时，

测量 100m 物体时的相对误差：$|e_r^*(x)|=\frac{1}{10000}=0.01\%$

测量 10m 物体时的相对误差：$|e_r^*(x)|=\frac{1}{1000}=0.1\%$。

可见前者的测量结果要比后者精确。所以，**在分析误差时，相对误差更能刻画误差的特性。**

2.3 有效数字与有效数字位数

为了能给一种数的表示法，使之既能表示其大小，又能表示其精确程度，于是需要引进有效数字的概念。在实际计算中，当准确值 x 有很多位数时，我们常按**四舍五入的原则**得到 x 的近似值 x^*。例如无理数 $e=2.718281828\cdots$, 若按四舍五入原则分别取**二位**和**五位小数**时，则得 $e\approx 2.72,\ \ e\approx 2.71828.$

不管取几位小数得到的近似数，其绝对误差都不超过末位数的半个单位，即 $|e-2.72|\le\frac{1}{2}\times10^{-2}, |e-2.71828|\le\frac{1}{2}\times10^{-5}.$

- **“有效数字”的概念：**若近似值 x^* 的绝对误差限是某一位的半个单位，就称其“准确”到这一位，且从该位直到 x^* 的第一位非零数字共有 n **位有效数字**。

引入有效数字概念后，我们规定所写出的数都应该是有效数字，且在同一计算问题中，参加运算的数，都应有相同的有效数字。

例如，下列各数 358.467，0.00427511，8.000034 的具有 5 位有效数字的近似值分别是 358.47，0.0042751，8.0000。

注：8.000034 的 5 位有效数字是 8.0000，而不是 8，因为 8 只有一位有效数字。前者精确到 0.0001，而后者仅精确到 1，两者相差是很大的。**有效数字尾部的零不能随意省去，**以免损失精度。

2.4 有效数字、绝对误差、相对误差之间的关系

近似值 x^* **具有** n **位有效数字**的科学计数法：

$$x^* = \pm(a_1 \times 10^{-1} + a_2 \times 10^{-2} + \cdots + a_n \times 10^{-n}) \times 10^{m+1}$$

$$= \pm 10^m \times (a_1 + a_2 \times 10^{-1} + \cdots + a_n \times 10^{-(n-1)}) \quad (1.7)$$

其中 $a_1 \neq 0,\ a_2, \cdots, a_n : 0 \sim 9$.

则其绝对误差满足

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

由 $a_1 \times 10^m \leq |x^*| \leq (a_1 + 1) \times 10^m$ ，知

$$|e_r^*(x)| = \left|\frac{x - x^*}{x^*}\right| \leq \frac{\frac{1}{2} \times 10^{m-n+1}}{a_1 \times 10^m} = \frac{1}{2a_1} \times 10^{-n+1}$$

定理 1 若用(1.7)式表示的近似值 x^* 具有 n 位有效数字，则其相对误差满足不等式：

$$|e_r^*(x)| \leq \frac{1}{2a_1} \times 10^{-n+1}$$

其中 a_1 为 x^* 的第一个非零数字。

定理 2 若近似值 x^*的相对误差满足不等式：

$$|e_r^*(x)| \leq \frac{1}{2(a_1+1)} \times 10^{-n+1}$$

其中 a_1 为 x^* 的第一个非零数字，则它至少具有 n 位有效数字。

- **对同一个数的近似值，有效数字位数越多，其绝对误差与相对误差越小；反之，绝对误差或相对误差变小，有效数字的位数则变多。**

§3　数值运算中误差传播规律简析

设 x_1^*, x_2^* 分别是 x_1, x_2 的近似值，$y^* = f(x_1^*, x_2^*)$ 作为 $y = f(x_1, x_2)$ 的近似值，其绝对误差和相对误差有（利用泰勒展开式）：

$$y - y^* = f(x_1, x_2) - f(x_1^*, x_2^*)$$

$$\approx \left(\frac{\partial f}{\partial x_1}\right)^* (x_1 - x_1^*) + \left(\frac{\partial f}{\partial x_2}\right)^* (x_2 - x_2^*)$$

得：$$\boldsymbol{e^*(y) \approx \left(\frac{\partial f}{\partial x_1}\right)^* e^*(x_1) + \left(\frac{\partial f}{\partial x_2}\right)^* e^*(x_2)} \qquad (1.8)$$

$$\boldsymbol{e_r^*(y) \approx \left(\frac{\partial f}{\partial x_1}\right)^* \frac{x_1^*}{y^*} e_r^*(x_1) + \left(\frac{\partial f}{\partial x_2}\right)^* \frac{x_2^*}{y^*} e_r^*(x_2)} \qquad (1.9)$$

一般函数的误差传播规律可通过类似于（1.8）、（1.9）分析的方法得到。

例 4　测得圆环（见教材图 1-3）外径 $D_1 = (10 \pm 0.05)cm$，内径 $D_2 = (5 \pm 0.1)cm$，则其面积 $S = \frac{\pi}{4}(D_1^2 - D_2^2)$ 的近似值为

$$S^* = \frac{\pi}{4}[(D_1^*)^2 - (D_2^*)^2] = \frac{\pi}{4}(10^2 - 5^2) = \frac{\pi}{4} \approx 58.905cm^2.$$

S^* 的绝对误差：

$$|e^*(S)| \approx \left|\left(\frac{\partial S}{\partial D_1}\right)^* e^*(D_1) + \left(\frac{\partial S}{\partial D_2}\right)^* e^*(D_2)\right|$$

$$= \left|\frac{\pi}{2} D_1^* e^*(D_1) - \frac{\pi}{2} D_2^* e^*(D_2)\right|$$

$$\leq \frac{\pi}{2} D_1^* |e^*(D_1)| + \frac{\pi}{2} D_2^* |e^*(D_2)|$$

$$\leq \frac{\pi}{2} \times 10 \times 0.05 + \frac{\pi}{2} \times 5 \times 0.1 = 0.5\pi$$

$$\approx 1.5708cm^2 \ .$$

S^* 的相对误差：

$$|e_r^*(S)| \approx \left|\frac{e^*(S)}{S^*}\right| \leq \frac{1.5708}{58.905} < 0.027 \ .$$

§4　数值运算中应注意的几个原则

利用计算机来求数学模型的数值解时，首先须设计算法。算法的好坏，将会直接影响到计算机的使用效率，也会影响到数值计算结果的精确度与真实性。

- **选用算法时一般应遵循以下原则：算法是否稳定；算法的逻辑结构是否合理简单；算法的运算次数和算法的存储量是否尽量少等等。**

在数值运算过程中，通过对误差传播规律与算法优劣的分析，选用和设计算法时应注意的几个原则如下：

1. 选用数值稳定性好的算法

在算法的执行过程中，会出现舍入误差的积累。舍入误差对计算结果的精确性影响较小的算法，具有较好的数值稳定性，称为**稳定的算法**；反之就认为算法的稳定性较差，称为**不稳定的算法**。

如果算法不稳定，则数值计算的结果就会严重背离数学模型的真实结果。因此，**在选择数值计算公式来进行近似计算时，应特别注意选用那些在数值计算过程中不会导致误差迅速增长的计算公式。**

例 2　计算积分

$$I_n = e^{-1}\int_0^1 x^n e^x dx,\quad n = 0,1,2,\cdots$$

利用分部积分法不难求得I_n的递推关系式为：

$$\begin{cases} I_n = 1 - nI_{n-1} \\ I_0 = 1 - e^{-1} \approx 0.6321 \end{cases}$$

由此可依次算得结果如下：

$$I_0 = 0.6321, \quad I_1 = 0.3680, \quad I_2 = 0.2640.$$
$$I_3 = 0.2080, \quad I_4 = 0.1680, \quad I_5 = 0.1600,$$
$$I_6 = 0.0400, \quad I_7 = 0.7200, \quad I_8 = -0.7280.$$

由于

$$0 < I_n < e^{-1}\max_{0\le x\le 1}(e^x)\int_0^1 x^n dx = \frac{1}{n+1}$$

则由以上I_n的不等式可看出

$$I_7 < \frac{1}{8} = 0.1250\ 。$$

可见**按递推关系式算出的I_7, I_8的结果是错误的，错误产生的原因**是因为I_0本身有不超过$(1/2)\times 10^{-4}$的舍入误差，此误差在运算中传播，积累误差很快，传播到I_7与I_8时，已使得I_7, I_8的结果面目全非。

即，若I_0的近似值I_0^*有误差ε_0，由此引起I_n的近似值I_n^*的误差为：

$$I_n - I_n^* = (-1)^n n!\,\varepsilon_0$$

但若将递推公式改写为

$$I_{n-1} = \frac{1}{n}(1 - I_n) \quad (n = N, N-1, \cdots, 1, 0) \qquad (1.10)$$

可得：

$$I_{n-1}-I_{n-1}^{*}=\frac{1}{n}(1-I_n)-\frac{1}{n}(1-I_n^{*})=\frac{1}{n}(I_n^{*}-I_n)=-\frac{1}{n}\varepsilon_n$$

就是一个稳定的算法。

此时，由：

$$\frac{1}{n+1}>I_n>e^{-1}\min_{0\le x\le 1}(e^x)\int_0^1 x^n dx=\frac{e^{-1}}{n+1}$$

需首先估计I_N的近似值，例如当 $N=7$ 时，由上面的估计式可取 $I_7=0.1124$，作初始值，依次计算，有如下结果：

$$I_7=0.1124 \qquad I_6=0.1269, \qquad I_5=0.1455,$$

$$I_4=0.1708, \qquad I_3=0.2073, \qquad I_2=0.2643,$$

$$I_1=0.3680, \qquad I_0=0.6320.$$

此时，由于因 I_7 引起的初始误差在以后的计算过程中逐渐减小，最后得到了与 $I_0=1-e^{-1}\approx 0.6321$ 相差无几的精确结果。

2. 相近两数避免相减

在数值计算中，**两个相近的数相减将会造成有效数字的严重损失。**因此，遇到这种情况，应当多保留这两个有效数字，尽量避免减法运算。改变计算方法，根据不同情况对公式进行处理，如可通过**因式分解、分子分母有理化、三角函数恒等式、其他恒等式、Taylor 展开式**等计算公式，防止减法运算的出现。

例如，当 $x=1000$ 时，计算 $\sqrt{x+1}-\sqrt{x}$ 的值，若取 4 位有效数字计算：

$$\sqrt{x+1}-\sqrt{x}=\sqrt{1001}-\sqrt{1000}\approx 31.64-31.62=0.02.$$

这个结果只有一位有效数字，损失了三位有效数字，从而绝

对误差和相对误差都变得很大，严重影响了计算结果的精度，但若将公式改变为

$$\sqrt{x+1}-\sqrt{x}=\frac{1}{\sqrt{x+1}+\sqrt{x}}\approx 0.01581.$$

它仍有四位有效数字，可见**改变计算公式可以避免两个相近数相减而引起的有效数字的损失，从而可以得到比较精确的结果。**

又如（见教材例 1）：

$$y=arctan5430-arctan5429=arctan\frac{1}{1+5429\times 5430}$$

$$=arctan\frac{1}{29479471}\approx 3.392191\times 10^{-8}$$

又如，计算 $A=10^7(1-\cos 2^0)$ 的值，若将 $\cos 2^0\approx 0.9994$（具有四位有效数字）代入直接计算：

$$A\approx 10^7(1-0.9994)=6\times 10^3,$$

这个结果只有一位有效数字，但若利用公式：

$$1-\cos x=2\sin^2\frac{x}{2},$$

则有 $A=10^7(1-\cos 2^\circ)=2\times(\sin 1^\circ)^2\times 10^7\approx 2\times 0.01745^2\times 10^7\approx 6.09\times 10^3.$

从而可得到**具有三位有效数字的比较精确的结果。**

3. 绝对值相对太小的数不宜作除数

在数值计算中，用绝对值很小的数作除数，将会使商数量级增加，甚至会在计算机造成“溢出”停机，而且当很小的除数稍有一点误差时，会对计算结果影响很大。

例如，$\frac{3.1416}{0.001}=3141.6$，当分母变为 0.0011，即分母只有

0.0001 的变化时，$\frac{3.1416}{0.0011}=2856$，**商却引起了巨大变化**。又如，计算$\frac{1.23}{\sqrt{x+1}-\sqrt{x}}$，当$x$很大时，如果直接计算，有效数字会严重损失，因此需把公式变换为$1.23(\sqrt{x+1}+\sqrt{x})$。因此，**在计算过程中，不仅要避免两个相近的数相减，还应特别注意避免再用这个差作除数。**

4. 警惕大数“吃掉”小数造成的危害

在数值计算中，参与运算的数有时数量级相差很大，而计算机的位数是有限的。在编制程序时，如不注意运算次序，就很可能出现小数加不到大数中而产生大数吃掉小数的现象。因此，**两数相加时，应尽量避免将小数加到大数中所引起的这种严重后果。**

例如，对a,b,c三数作加法运算，其中

$$a=10^{12},\ b=10,\ c=-a$$

若按$(a+b)+c$的顺序编制程序，在八位的计算机上计算，则a吃掉b，且a与c互相抵消，其结果接近于零，但若按$(a+c)+b$的顺序编制程序，则可得到接近于10的真实结果。

又如：**计算（五位有效数字）**

$$12345+\sum_{i=1}^{100}\frac{1}{500}=12345 \quad \text{而} \quad \sum_{i=1}^{100}\frac{1}{500}+12345=12355$$

。

在实际计算中，我们还要特别注意保护重要的物理参数，防止一些重要的物理量在计算中被吃掉。例如考察物体在阻尼介质

中的运动时，阻尼系数k是一个重要的物理参数，若在动力学方程离散过程中将k置于一个很大的数a的加减运算中，则k就会被数a吃掉，将会使结果严重失真。因此，为了避免大数吃掉小数，我们必须事先分析计算方案的数量级，在编制程序时，加以合理安排，这样，一些重要的处理参数才不至于在计算中被吃掉，以免造成有效数字不必要的损失。

5. 简化计算步骤，减少运算次数

同样一个计算问题，**若能选用更为简单的计算公式，减少运算次数，不但可以节省计算量，提高计算速度，还能增加逻辑结构，减少误差积累。**这也是数值计算必须遵循的原则与计算方法研究的一项主要内容。

例如，计算多项式

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

的值，若采用逐项计算然后相加的算法，计算a_kx^k要做k次乘法，而$P_n(x)$共有$n+1$项，所以需做：

$$1+2+\cdots+(n-1)+n=\frac{1}{2}n(n+1)$$

次乘法和n次加法，但若采用**递推算法**（又称**秦九韶算法**）：

$$\begin{cases}u_0=a_n\\ u_k=u_{k-1}x+a_{n-k}\end{cases}$$

对$k=1,2,\cdots,n$反复执行上式算式，则只需n次乘法和n次加法，即可算了$P_n(x)$的值。