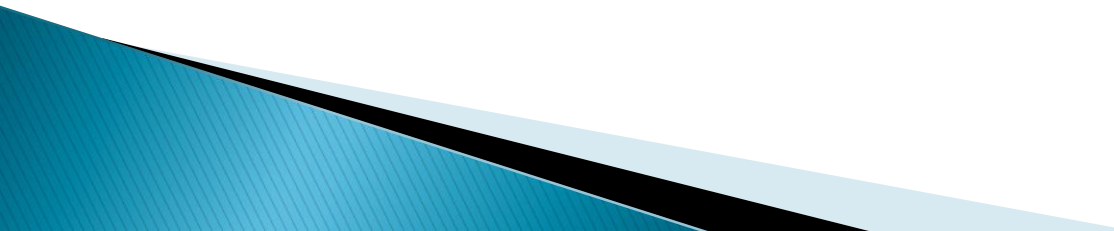


第一章 误差

1. 误差的来源
 2. 浮点数、误差、误差限和有效数字
 3. 相对误差和相对误差限
 4. 误差的传播
 5. 在近似计算中需要注意的一些现象
- 

误差

在计算机中进行如下运算

- ▶ $1 - 0.5$ 是否等于 0.5
- ▶ $1 - 0.9$ 是否等于 0.1
- ▶ 计算 $\sqrt{-1 + (1 - 0.9 - 0.1)i}$
- ▶ 计算 $\sqrt{-1 - (1 - 0.9 - 0.1)i}$

误差

▶ 模型误差

- 实际问题用数学模型刻画时要忽略一些因素, 从而造成数学的量 u_1 和实际的量 u 的误差—模型误差.
- 人口增长模型: $y'(t) = ry$

▶ 数据误差

- 数学模型用到的数据, 可能是观测到的(称观测误差), 也可能是计算得到的, 这种数据误差也造成数学量的近似 u_2 .
- 地球重力加速度 g : 约 $9.8m/s^2$

误差

▶ 方法误差－截断误差

- 解数学问题的方法给出的解也是近似的 u_3 , 它与数学问题的准确解的差叫方法误差, 也叫截断误差.

$$f(x) \approx P_n(x)$$

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

- 截断误差: $R_n(x) = \frac{f^{n+1}(\xi)}{(n+1)!}x^{n+1}$

▶ 舍入误差

- 近似的方法计算数据有误差的数学问题要用有限位数字, 这就要舍入, 计算得 u_4 . 由此引起的误差称舍入误差.

$$R = \pi - 3.141\,592\,653\,589\,793 = 2.3846 \dots \times 10^{-16}$$

浮点数

▶ 规格化浮点数

- $x = \pm 0.\alpha_1\alpha_2\cdots\alpha_t \times \beta^J, \alpha_1 \neq 0, 0 \leq \alpha_j < \beta$
- 阶(亦称指数): J 整数, $L \leq J \leq U$
- 尾数: $w = 0.\alpha_1\alpha_2\cdots\alpha_t$
- 计算机常用的双精度浮点数(64位二进制):
 $\beta = 2, t = 52, -1023 \leq J \leq 1023$

▶ 基-进制

- β 称为基
- 这样表示的数称为 β 进制数

▶ 上溢、下溢

IEEE 单精度浮点数

符号 Sign	指数 Exponent	尾数 Mantissa
1 bit	8 bits	23 bits

IEEE 双精度浮点数

符号 Sign	指数 Exponent	尾数 Mantissa
1 bit	11 bits	52 bits

误差

▶ 误差

- 准确数 x ，近似数 x^*
- 误差 $e^* = x^* - x$ 、误差限 $\varepsilon^* \geq |x^* - x|$
- $x \leq x^* + \varepsilon^*$

准确数	近似数	误差	误差限
x	x^*	$e^* = x^* - x$	$\varepsilon^* \geq x^* - x $
$\pi = 3.141\ 592\ 65\dots$	3	-0.14...	0.15, 0.5, ...
	3.14	-0.001 5...	0.001 6, 0.005, ...
	3.141 6	0.000 007...	0.000 008, ...

▶ 四舍五入

- 十进制数通常取若干位为其近似：若后位为4则舍，为5则进1。

有效数字

▶ 有效数字——准确到该位

- 如果 x^* 的误差限是某位的半个单位，该位到 x^* 的第一位非零数字共 n 位，则称 x^* 有 n 位有效数字或 x^* 准确到该位
- x^* 可表成 $x^* = \pm 0.\alpha_1\alpha_2 \cdots \alpha_t \cdots \times 10^p, (1.1),$

$$\alpha_1 \neq 0, |x^* - x| \leq \frac{1}{2} \times 10^{p-n}$$

则 x^* 有 n 位有效数字

- 例：真空中光速 $c = 299792458m/s$, 若取 $300000000 m/s$ 近似，则有4位有效数字。
- ▶ 四舍五入所得近似数从第一位非零数字到最后一位都是有效数字
- ▶ 若误差为0，则认为有效数字有任意位

相对误差

▶ 相对误差

- $e_r^* = \frac{x^* - x}{x}$, $x \neq 0$, 或 $e_r^* = \frac{x^* - x}{x^*}$

▶ 相对误差限

- $\varepsilon_r^* \geq \left| \frac{x^* - x}{x} \right|$, 或 $\varepsilon_r^* \geq \left| \frac{x^* - x}{x^*} \right|$

准确数	近似数	相对误差	相对误差限
x	x^*	$e_r^* = (x^* - x)/x$ 或 $e_r^* = (x^* - x)/x^*$	$\varepsilon_r^* \geq (x^* - x)/x $ 或 $\varepsilon_r^* \geq (x^* - x)/x^* $
c	$2.997\,925 \times 10^{10}$		$3.3 \dots \times 10^{-7} \approx 3 \times 10^{-7}$
	3.00×10^{10}		$0.00067 \dots \approx 0.0007$
$\pi = 3.14159265 \dots$	3.1416	$2.3 \dots \times 10^{-6}$	$2.4 \times 10^{-6} \approx 2 \times 10^{-6}$

相对误差

- 设数 x^* 可表成 (1.1) ，若 x^* 有 n 位有效数字则有相对误差限 $\frac{1}{2\alpha_1} \times 10^{1-n}$
 - $|x^* - x| \leq \frac{1}{2} \times 10^{p-n}$,
 - $|x^*| \geq \alpha_1 \times 10^{p-1}$ ，相除。
- 若 x^* 相对误差限 $\varepsilon_r^* \leq \frac{1}{2(\alpha_1+1)} \times 10^{1-n}$ ，则 x^* 至少有 n 位有效数字
 - $|x^* - x| \leq \varepsilon_r^* |x^*|$,
 - $|x^*| \leq (\alpha_1 + 1) \times 10^{p-1}$ ，相乘。
- 例： $e = 2.71828 \dots$ ，取三位有效数字 $x^* = 2.72$ ，相对误差限
$$\frac{1}{2 \times 2} \times 10^{1-3} = 0.0025$$
- 由定义，相对误差为 $0.0006\dots$ ，如果再由
$$0.0025 \leq \frac{1}{2 \times 2} \times 10^{1-2}$$
则 x^* 有二位有效数字，估少了一位数字

计算机精度

▶ 计算机表示数的误差

- 数 $x = 0.a_1a_2 \dots a_t \dots \times 10^p$ 引入计算机时四舍五入表示成 t 位尾数 $fl(x) = x^* = 0.a_1a_2 \dots a_t \times 10^q$, $q = p$ 或 $p + 1$
- 令 $\varepsilon = (fl(x) - x)/x$ 则 $|\varepsilon| \leq eps \equiv \frac{1}{2} \times 10^{1-t}$, $fl(x) = x(1 + \varepsilon)$
- 数 x 在 t 位尾数, β 进制, 舍入的计算机系统表为 $fl(x)$ 则有 $fl(x) = x(1 + \varepsilon)$, $|\varepsilon| \leq eps \equiv \frac{1}{2} \times \beta^{1-t}$

▶ 计算机的精度

- 称 $eps \equiv \frac{1}{2} \times \beta^{1-t}$ 为该计算机的精度
- 双精度浮点数的精度为 $2^{-52} \approx 2 \times 10^{-16}$

误差的传播

▶ 用微分来估计误差和误差限

- 误差 $e^* = x^* - x \approx dx$
- 相对误差 $e_r^* \approx d \ln x$, $d_r x = \left| \frac{dx}{x} \right| = |d \ln x|$

▶ 近似数参与运算时结果的误差

- 四则运算时结果误差的估计
 - $d(x \pm y) = dx \pm dy$
 - $d(x \times y) = ydx + xdy$
 - $d\left(\frac{x}{y}\right) = \frac{ydx - xdy}{y^2}$
- 计算函数时误差的估计 $df(x) = f'(x)dx$

误差的传播:相对误差

▶ 近似数参与运算时结果的误差

◦ 四则运算时结果相对误差的估计

- $d_r(x + y) = \left| \frac{dx+dy}{x+y} \right| \leq \left| \frac{dx}{x} \frac{x}{x+y} + \frac{dy}{y} \frac{y}{x+y} \right| \leq \max(d_r x, d_r y)$
- $d_r(x - y) \leq \frac{|x|d_r x + |y|d_r y}{|x-y|}$, x, y 同号 (相近的数相减会放大误差!)

- $d_r(xy) \leq d_r x + d_r y$
- $d_r(x/y) \leq d_r x + d_r y$

◦ 计算函数时结果相对误差的估计

- $d \ln f(x) = f'(x)dx/f(x)$
- $d_r f(x) = \left| \frac{xf'(x)}{f(x)} \right| d_r x$

误差的传播:例

- ▶ 例 设 $a = 1.21 \times 3.65 + 9.81$, 其中每个数据的绝对误差限为0.005, 求 a 的绝对误差限和相对误差限

$$da = d(1.21 \times 3.65) + d9.81$$

$$|da| \leq 1.21 \times 0.005 + 3.65 \times 0.005 + 0.005$$

$$\approx 0.0293 \leq 0.03$$

$$d_r a \approx \max(d_r (1.21 \times 3.65), d_r 9.81)$$

$$\approx \max(d_r 1.21 + d_r 3.65, d_r 9.81)$$

$$= \max(d \ 1.21/1.21 + d \ 3.65/3.65, d \ 9.81/9.81)$$

$$= \max(0.005/1.21 + 0.005/3.65, \ 0.005/9.81)$$

$$\approx \max(0.0055, \ 0.0005) = 0.0055$$

- 设 $y = x^n$, y 的相对误差与 x 的相对误差之间的关系:

$$d_r y = |d(\ln y)| = |nd(\ln x)| = nd_r x$$

浮点运算的误差

- ▶ 计算机中浮点运算的误差

在一个 t 位尾数舍入的计算机系统中，两个机器数 x, y 作四则运算 $\circ (+、-、\times、\div)$ ，记为 $x \circ y$ ，结果舍入得到 $fl(x \circ y)$ 则有 $fl(x \circ y) = (x \circ y)(1 + \varepsilon)$ ， $|\varepsilon| \leq eps$
(计算机精度)

- ▶ 向前误差分析: 直接比对最后的计算结果，得出误差
- ▶ 向后误差分析: 把舍入误差归结为数据引出的误差

经验谈

- ▶ 在近似计算中需要注意的一些事项
 - 避免相近数相减
 - 防止大数‘吃’小数
 - 避免分母为零或比分子小得多
 - 注意简化计算步骤, 减少运算次数
 - 秦九韶算法
 - 选用稳定的公式
- ▶ 计算问题的敏感性与算法的稳定性

例

- 当 x 接近于零时 $\frac{1-\cos x}{\sin x}$ 应变换为 $\frac{\sin x}{1+\cos x}$
- 当 x 充分大时 $\sqrt{1+x} - \sqrt{x}$ 应变换为 $\frac{1}{\sqrt{1+x}+\sqrt{x}}$
- 计算多项式 $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$ 的秦九韶算法

$$P(x) = (\cdots ((a_n x + a_{n-1})x + a_{n-2})x + \cdots + a_1)x + a_0$$

- 更一般的有

$$P(x)$$

$$= (\cdots ((c_n(x - b_n) + c_{n-1})(x - b_{n-1}) + c_{n-2})(x - b_{n-2}) + \cdots + c_1)(x - b_1) + c_0$$

其中 b_1, b_2, \cdots, b_n 被称为基点

例

- 尾数是3位十进制数字的浮点系统中运算求解二元一次方程组

$$\begin{cases} 0.000\ 1x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}$$

- 顺序消元 $-10\ 000x_2 = -10\ 000, x_2 = 1, x_1 = 0$
- 方程交换再消元 $\begin{cases} x_1 + x_2 = 2 \\ 0.000\ 1x_1 + x_2 = 1 \end{cases}, x_2 = 1, x_1 = 1$

例

- ▶ 求 $x^2 - 56x + 1 = 0$ 的根. 取5位数字
 - 准确解 $x_1 = 55.982\ 137\ 159\ \dots$, $x_2 = 0.017\ 862\ 840\ \dots$
 - $x_1 = 28 + \frac{28^2 - 1}{2} = 28 + 27.982 = 55.982$,
 - $x_2 = 28 - \frac{28^2 - 1}{2} = 28 - 27.982 = 0.018$
 - x_1 同上, $x_2 = \frac{1}{x_1} = \frac{1}{55.972} = 0.01786288$
- ▶ 在三位尾数的计算机上计算 $x = a_0 + a_1 + a_2 + \dots + a_{100}$, $a_0 = 0.1$, $a_1 = a_2 = \dots = a_{100} = 0.0001$
 - $a_0 + a_1$ 得 0.1, 再加 a_2 还是 0.1, \dots , $x = 0.1$
 - 如果从后往前加, $0.0001 + 0.0001 = 0.0002$, \dots 最后 $x = 0.1 + 0.01 = 0.11$