科学计算笔记
MA235

任云玮

目录

1 绪论 **2**

1.1 计算机数值计算基本原理 2

1.1.1 计算机基本工作原理 2

1.1.2 实数的存贮方法 2

1.1.3 实数的基本运算原理 3

1.2 误差的来源与估计 4

1.2.1 误差的来源 4

1.2.2 误差与有效数字 4

1.2.3 数值运算的误差估计 5

1 绪论

1.1 计算机数值计算基本原理

1.1.1 计算机基本工作原理

CPU		外存储器
	数据总线	
I/O 接口		I/O 接口		外存接口
输入		输出		外存储器

1.1.2 实数的存贮方法

1 定义 (二进制浮点数系) [1] 实数在计算机内部为**近似存贮**，采用二进制浮点数系

$$F(2, n, L, U) = \{\pm 0.a_1a_2 \dots a_n \times 10^m\} \cup \{0\}$$

其中 $a_1 = 1$，$a_i \in \{0, 1\}$. 指数 m 满足 $L \le m \le U$. 称 n 为其字长，2 表示采用二进制。

2 标准 (IEEE)

1. 单精度: $t = 24, L = -126, U = 127$
2. 双精度: $t = 53, L = -1022, U = 1023$
3. Underflow Limit: $UFL = 0.1 \times 2^L$. 若 $0 < x < UFL$，则 $fl(x) = 0$.
4. Overflow Limit: $OFL = 0.11 \dots 1 * 2^U$. 若 $x > OFL$，则 $fl(x) = \infty$.
5. 舍入: 若 $UFL \le x \le OFL$，则 $fl(x)$ 为舍入所得浮点数。舍入规则如下：设 $x = 0.a_1a_2 \dots a_n \cdots \times 2^m$. 若 $a_{n+1} = 1$，则 $d_t + 1$ 并舍弃其后项；否则直接舍弃其后项。

3 定义 (机器精度) 下仅考虑舍去的情况。

$$\begin{aligned} x - fl(x) &= 2^m \times 0.0 \dots 0a_{n+2} \dots \\ &= 2^m \times [2^{-(t+2)} + 2^{-(t+3)} + \cdots] \\ &= 2^m \times 2^{-(t+1)} \end{aligned}$$

其相对误差满足

$$\frac{x - fl(x)}{x} < \frac{x - fl(x)}{0.5 \times 2^m} = 2^{-t}$$

记为 ε，称之为机器精度。

4 命题

$$fl(x) = x(1 + \delta), \text{ 其中 } |\delta| \le \varepsilon$$

[1] Floating Number System

1.1.3 实数的基本运算原理

加法 + 硬件实现 ⇒ 四则运算。

5 实现 $(x+y)$ 设 x, y 为浮点数，则 $x+y$ 的实现方式如下：

1. 对阶：将指数 m 化为两者中较大者；
2. 尾数相加；
3. 舍入；
4. 溢出分析等……
5. 结果输出。

评注 由 $fl(x)+fl(y)=x(1+\delta_x)+y(1+\delta_y)$ 可知，当一个大数与一个小数相加时，小数有可能被忽略，所以应当避免大数小数间的相加。

1.2 误差的来源与估计

1.2.1 误差的来源

1. 模型问题。例：近似地球为球体来计算。
2. 测量误差。例：测量地球半径时的误差。
3. 方法误差（截断误差）。例：对于 $y = f(x)$，求 $f(x^*)$ 时使用 Taylor 展开。
4. 舍入误差（rounding-off）。例：计算机计算时的误差。

1.2.2 误差与有效数字

6 定义 (绝对误差) 设 x 为给定实数，x^* 为其近似值。定义绝对误差为

$$e(x^*) = x^* - x.$$

称 ε^* 为其误差上界，若

$$|e(x^*)| \leq \varepsilon^*$$

7 定义 (相对误差) 对于同上的 x 和 x^*，定义其相对误差

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

称 ε_r^* 为其相对误差界，若

$$|e_r(x^*)| \leq \varepsilon_r^*$$

评注 在实际应用中，x 通常是未知的，所以会采用

$$\bar{e}_r(x^*) = \frac{x^* - x}{x^*}$$

来代替相对误差。对于分子，使用绝对误差界来替代，有如下不等式

$$|\bar{e}_r(x^*)| \leq \frac{\varepsilon^*}{|x^*|}.$$

这两种相对误差界间的差别，当 $\varepsilon^* \ll 1$ 时，满足

$$|e_r - \bar{e}_r| = O((\varepsilon_r^*)^2)$$

8 定义 (有效数字) 设 $x \in R$，x^* 为其近似值。称 x^* 相对于 x 有 n 位有效数字，若 n 是满足下式的 n 的最大值。

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

评注 在实践中，一般可以采用更加简便的方法，对于归一化以后的 x^*，在尾数部分有 n 位，则称其有 n 位有效数字。注意，此方法对于错误的舍入结果是不实用的。

9 定理 (误差与有效数字) 若 $x = 0.a_1a_2\ldots a_n \times 10^m$ 有 n 位有效数字，则

$$\left|\frac{x^* - x}{x}\right| \leq \frac{1}{2a_1} \times 10^{1-n}.$$

反之，若

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

则 x^* 至少有 n 位有效数字。

证明 对于前者，只需利用有效数字的定义，以及利用 $x \geq 0.a_1$ （仅考虑 $a_1 \neq 0$ 的情况）。对于后者，证明是类似的。

1.2.3 数值运算的误差估计

以下内容都假设运算无误差。

10 定理 (四则运算误差估计)

1. 加/减法: $\varepsilon(x^* \pm y^*) \leq \varepsilon_x^* + \varepsilon_y^*$
2. 乘法: $\varepsilon(x^*y^*) \leq |x^*|\varepsilon_y^* + |y^*|\varepsilon_x^*$
3. 除法: $\varepsilon(\frac{x^*}{y^*}) = \frac{|x^*|\varepsilon_y^* + |y^*|\varepsilon_x^*}{|y^*|^2}$

证明 考虑加法的误差估计。对于 x，y 及其近似值 x^*，y^*，计算 $x^* \pm y^*$ 和 $x \pm y$ 间的误差。

$$\begin{aligned} & |x^* \pm y^* - (x \pm y)| \leq |x^* - x| + |y^* - y| \leq \varepsilon_x^* + \varepsilon_y^* \\ \Rightarrow \quad & \varepsilon(x^* \pm y^*) \leq \varepsilon_x^* + \varepsilon_y^* \end{aligned}$$

对于其他的运算，证明是类似的。（证明中可用 $+1-1$ 技巧）

11 定理 (运算的误差估计) 设 $A = f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$，$\mathbf{x}^*$ 是 $\mathbf{x}$ 的估计值。利用带 Peano 余项的 Taylor 展开，可知 A 的绝对误差满足

$$\begin{aligned} e(A^*) &= f(\mathbf{x}^*) - f(\mathbf{x}) \\ &= \sum_{p=1}^{q} \mathrm{d}^k f(\mathbf{x}^*) + o(||x^* - x||^q) \\ &\text{取 q=1，则} \\ &= \sum_{k=1}^{n} \partial_k f(\mathbf{x}^*)(x^* - x) + + o(||x^* - x||^q) \end{aligned}$$

利用上式，可知

$$\varepsilon(A^*) \approx \sum_{k=1}^{n} \partial_k f(\mathbf{x}^*)\varepsilon(x^*)$$

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