科学计算笔记

任云玮

目录

1 绪论

1.1 计算机数值计算基本原理

1.1.1 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

1.1.2 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 $x \in R$，$x^* = 0.a_1a_2\cdots a_k \times 10^m$ 为其近似值。称 x^* 相对于 x 有 $n\,(n \leq k)$ 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适用的，对于错误的情况，需要再减去一位有效数字。

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

$$\varepsilon_r^* \le \frac{1}{2a_1} \times 10^{1-n}.$$

反之，若

$$\varepsilon_r^* \le \frac{1}{2(1+a_1)} \times 10^{1-n},$$

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 $x \ge 0.a_1$ （仅考虑 $a_1 \ne 0$ 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: $\varepsilon(x^* \pm y^*) \le \varepsilon_x^* + \varepsilon_y^*$
2. 乘法: $\varepsilon(x^*y^*) \le |x^*|\varepsilon_y^* + |y^*|\varepsilon_x^*$
3. 除法: $\varepsilon(\frac{x^*}{y^*}) \le \frac{|x^*|\varepsilon_y^* + |y^*|\varepsilon_x^*}{|y^*|^2}$

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。

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

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 $A = f(\mathbf{x}) = f(x_1, x_2, \dots, x_n)$，$\mathbf{x}^*$ 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

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

利用上式，可知

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

评注 对于定义在 $\mathbf{R}$ 上的函数，即为

$$\varepsilon(f(x^*)) \approx |f'(x^*)|\varepsilon(x^*)$$

1.2.4 数字求和的舍入误差分析

12 命题 n 个浮点数相加，若将它们从小到大排列后相加，则可以减小舍入误差。

证明 考虑浮点数的求和 $S_n = \sum_i^n a_i$，在计算机中的过程表现为

$$S_2^* = fl(a_1 + a_2) = (a_1 + a_2)(1 + \varepsilon_2), \quad |\varepsilon_2| \le \varepsilon = 2^{-t}$$
$$\cdots\cdots$$
$$S_n^* = fl(S_{n-1}^* + a_n)(1 + \varepsilon_n), \quad |\varepsilon_n| \le \varepsilon$$

对于 S_n^* 的误差，若定义 $\varepsilon_1 = 0$，则

$$S_n^* = \sum_{k=1}^{n} a_k \prod_{p=k}^{n}(1 + \varepsilon_p)$$

对误差进行估计，舍去高阶无穷小，有

$$\prod_{i=k}^{n}(1 + \varepsilon_k) \approx 1 + \sum_{i=k}^{n} \varepsilon_k$$

综合上两式，有

$$\begin{aligned} S_n^* &\approx \sum_{k=1}^{n} a_k (1 + \sum_{p=k}^{n} \varepsilon_p) \\ &= S_n + \sum_{k=1}^{n} a_k \sum_{p=k}^{n} \varepsilon_p \end{aligned}$$

进行移项，并取绝对值，再利用三角不等式，以及 $|\varepsilon_i| \le \varepsilon$，得

$$|S_n^* - S_n| \le \sum_{k=1}^{n} |a_k| \sum_{p=k}^{n} |\varepsilon_p| \le \varepsilon \sum_{k=1}^{n} |a_k|(n - k + 1)$$

其中 $n - k + 1$ 关于 k 单调减少，所以根据排序不等式 [引|理23]，即可知命题成立。■

1.3 避免算法失效的基本原则

13 定理 (原则)

1. 避免两数相除／相减，否则会严重损失有效数字。
2. 避免两相近数相减。
3. 避免绝对值很小的数做除数。
4. 避免大数与小数相加；
5. 简化计算步骤。

14 算法 (高效计算 e^A) 高效计算 e^A，其中 $A \in \mathbf{R}^{n\times n}$。首先有

$$e^A = e^{(A/2^n)2^n} = B^{2^n}$$

只需要得到 B，即可以利用倍乘的方法快速得到 B^{2^n}。下对于 B 进行估计。当 $x \to 0$ 时，e^x 有 Taylor 展开

$$e^x = 1 + x + \cdots + \frac{x^n}{n!} + \cdots$$

而取足够大的 n，即可以使得 $A/2^n \approx 0$，则可以对它展开得

$$B \approx I + C + \frac{1}{2}C^2, \text{ 其中} C = A/2^n$$

而对于倍乘，考虑 B^2，展开平方得

$$B^2 \approx I + 2(C + \frac{1}{2}C^2) + (C + \frac{1}{2}C^2)^2$$

从右至左相加即可。

15 算法 (秦九韶，多项式估值) 设有多项式 (1)，计算 $p(z), z \in \mathbf{R}$ 的值。

$$p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n \tag{1}$$

定义 b_n 满足

$$b_0 = a_0, \quad b_k = a_k + b_{k-1}z$$

则 b_n 即为所要求的值。并且成立

$$p'(z) = \sum_{k=0}^{n-1} b_k z^{n-1-k}$$

证明 用 $x-z$ 去除 $p(x)$，记所得余数为 $b_n(x)$，即

$$p(x) = (x-z)q(x) + b_n(x),$$

代入 $x = z$，则左侧第一项为 0，可知 $p(z) = b_n(z)$。将两边的式子展开，利用对应系数相等，即可得算法中 b_n 的递推式。

16 定理 (外推法) 设 x_0，x_1 是 x 的两个估计值，且 x_1 相较于 x_0 更接近 x，则可以通过恰当的权值 ω，使得它们的加权平均

$$\bar{x} = x_1 + \omega(x_1 - x_0)$$

更加接近精确值 x。

17 算法 (π 的估计) 考虑单位圆，其面积为 π，设 π_n 为单位圆的内接正 $2n$ 边形的面积，以及

$$\widetilde{\pi}_n = \frac{1}{3}(4\pi_{2n} - \pi_n)$$

则 π_n 与 $\widetilde{\pi}_n$ 与 π 的误差满足

$$|\pi_n - \pi| = O(\frac{1}{n^2}), \quad |\widetilde{\pi}_n - \pi| = O(\frac{1}{n^4})$$

证明 对于 π_n.

$$\pi_n = n\sin\frac{\pi}{n} = \pi - \frac{\pi^3}{3!}\frac{1}{n^2} + \frac{\pi^5}{5!}\frac{1}{n^4} - \cdots \Rightarrow |\pi_n - \pi| = O(\frac{1}{n^2})$$

对于 $\widetilde{\pi}_n$.

$$\begin{aligned}\widetilde{\pi}_n &= \pi_{2n} + k(\pi_{2n} - \pi_n) = (1+k)\pi_{2n} - k\pi_n \\ &= (1+k)(\pi - \frac{\pi^3}{3!}\frac{1}{4n^2} + \cdots) - k(\pi - \frac{\pi^3}{3!}\frac{1}{n^2} + \cdots) \\ &= \pi - (\frac{k+1}{4} - k)\frac{\pi^3}{3!}\frac{1}{n^2} + O(\frac{1}{n^4})\end{aligned}$$

为使式子的第二项为零，取 $k = \frac{1}{3}$，则成立

$$|\widetilde{\pi}_n - \pi| = O(\frac{1}{n^4}) \quad \blacksquare$$

评注 在实际中，π_n 也是没有办法直接计算而得的，但是对于 $n = 3$，即 6 边形的情况，可以知道 $\pi_3 = 3\sqrt{3}/2$。同时有递推公式

$$\pi_{2n} = \sqrt{2n(n - \sqrt{n^2 - \pi_n^2})},$$

而开平方可以通过迭代的方式实现，从而即计算得到足够精确的 π_{2n} 和 π_n。

2 函数的多项式插值与逼近

2.1 问题的提出

18 定义 (插值) 设函数 $y = f(x)$ 在 $[a, b]$ 上有定义，且已知在点 $a \le x_0 < x_1 < \cdots < x_n \le b$ 处的函数值 $y_i = f(x_i)$，若存在一简单函数 $P(x)$，成立

$$P(x_i) = y_i,$$

则称 $P(x)$ 为 $f(x)$ 的插值函数，点 $x_1, x_2, \ldots, x_n$ 称为插值节点，$[a, b]$ 称为插值区间，求 $P(x)$ 的方法被称为插值法。

若 $P(x) \in P_n$ 为次数不超过 n 的多项式，则称为多项式

2.2 多项式插值

19 定理 (唯一性) 给定满足定义18的 $n+1$ 个点上的函数值，则次数不超过 n 的插值多项式 $P_n(x)$ 存在且唯一。

证明 利用待定系数法，设多项式的系数为 $a_0, \ldots a_n$，则有线性方程组

$$\begin{cases} a_0 + a_1 x_0 + \cdots + a_n x_0^n = y_0, \\ a_0 + a_1 x_1 + \cdots + a_n x_1^n = y_1, \\ \qquad\qquad \cdots\cdots \\ a_0 + a_1 x_n + \cdots + a_n x_n^n = y_n, \end{cases}$$

其系数矩阵为 Vandermonde 矩阵

$$A = \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix}$$

根据定义18中对于 x_i 的要求，矩阵行列式成立

$$\det A = \prod_{i,j=0, i>j}^{n} (x_i - x_j) \neq 0.$$

所以该方程组有唯一解。

20 定理 (Lagrange 插值法) 定义

$$l_i(x) = \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}$$

$$(L_n f)(x) = \sum_{i=0}^{n} y_i l_i(x),$$

则 $L_n f$ 即为 f 的插值多项式。

证明 考虑构造 $l_i \in P_n$，满足条件 $l_i(x_j) = \delta_{ij}$，这样 $L_n f = \sum y_i l_i$ 满足要求。改写条件为（以 l_0 为例）

$$\begin{aligned} l_0(x) &= \alpha(x - x_1)\cdots(x - x_n) \\ l_0(x_0) &= 1 \end{aligned}$$

解得

$$\alpha = \frac{1}{(x_0 - x_1)\cdots(x_0 - x_n)} \quad \blacksquare$$

评注 这样构造插值多项式的动机在于在取定插值节点后，插值实际上相当于构造一个从 $\mathbf{y} = (y_0, \ldots, y_n) \in \mathbf{R}^{n+1}$ 到 $y^*(x) \in P_n$ 的一个映射 $\mathscr{F}$，并且可以证明，$\mathscr{F}$ 是线性的。因此成立

$$\mathscr{F}(\mathbf{y}) = \mathscr{F}(\sum_{i=0}^{n} y_i \mathbf{e}_i) = \sum_{i=0}^{n} y_i \mathscr{F}(\mathbf{e}_i) = \sum_{i=0}^{n} y_i l_i(x).$$

21 定理 (Lagrange 余项公式) 设符号含义同定理20且 f 充分光滑，则对于每一个固定的 x 成立

$$f(x) - (L_n f)(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)$$

其中 $\xi \in (a, b)$ 且

$$\omega_{n+1}(x) = (x - x_0)\cdots(x - x_n).$$

证明 固定 $x \neq x_i$，定义 $R(x)$ 满足

$$f(x) - (L_n f)(x) = R(x)\omega_{n+1}(x).$$

构造辅助函数 $g(t)$

$$g(t) = f(t) - (L_n f)(t) - R(x)\omega_{n+1}(t).$$

根据插值法与 $R(x)$ 的定义，成立

$$g(x_i) = 0, \quad g(x) = 0,$$

即函数 $g(t)$ 有 $n+2$ 个零点。反复应用 Rolle 定理，可知存在 $\xi \in (a, b)$，成立 $g^{(n+1)} = 0$，即

$$\begin{aligned} & g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - R(x)(n+1)! = 0 \\ \Rightarrow \quad & R(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \end{aligned}$$

结合 $R(x)$ 的定义式即可知命题成立。$\blacksquare$

评注 当已知 $f^{(n+1)}$ 有界时，可以使用此公式进行估计。

2.3 Runge 现象

22 定理 对于复函数 $f(z)$，如果存在 $r_0 > \frac{3}{2}(b-a)$，使得 $f(z)$ 在 $B_{r_0}(\frac{a+b}{2})$ 内解析，则 $P_n(x) = L_n(x)$ 在 $[a,b]$ 上一致收敛与 $f(z)$. 这里 $B_{r_0}(\frac{a+b}{2})$ 为以 $\frac{a+b}{2}$ 为圆心，r_0 为半径的圆。

3 附录

3.1 不等式

23 引理 (排序不等式) 对于满足下述条件的 $\{a_n\}$，$\{b_n\}$，

$$0 \le a_1 \le a_2 \le \cdots \le a_n$$
$$0 \le b_1 \le b_2 \le \cdots \le b_n$$

则同序相乘求和值最大，逆序最小，即

$$\sum_{i=1}^{n} a_i b_i \ge \sum_{i=1}^{n} a_i b_{k_i} \ge \sum_{i=1}^{n} a_i b_{n-i+1}$$

24 引理 (算数-几何均值不等式)

$$(a_1 a_2 \cdots a_n)^{1/n} \le \frac{a_1 + a_2 + \cdots + a_n}{n}$$

当且仅当 $a_1 = a_2 = \cdots = a_n$ 时等号成立。

证明 因为有齐次性，所以不妨设 $\prod a_i = 1$，并令

$$a_1 = \frac{\alpha_1}{\alpha_2}, \quad \ldots, \quad a_{n-1} = \frac{\alpha_{n-1}}{\alpha_n}, \quad a_n = \frac{\alpha_n}{\alpha_1}$$

则只需证明下式即可。

$$\frac{\alpha_1}{\alpha_2} + \cdots + \frac{\alpha_n}{\alpha_1} \ge n$$

不妨设 $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_n$，则根据排序不等式

$$\text{L.H.S} \ge \alpha_1 \frac{1}{\alpha_1} + \cdots + \alpha_n \frac{1}{\alpha_n} = n \quad \blacksquare$$