舍入误差分析

教师: 胡俊

北京大学数学科学学院

October 18, 2019

目录

目录

浮点数系统

计算机中浮点数f 表示为:

$$f = \pm w \times \beta^J, \quad L \leqslant J \leqslant U,$$

这里β是机器所用浮点数的基底, J是阶, w是尾数. 尾数w一般可表示为

$$w = 0.d_1 d_2 \cdots d_t,$$

其中 t 是尾数位, 称为字长, $0 \leqslant d_i < \beta$. 若 $d_1 \neq 0$, 则称该浮点数为规格划浮点数.
若用 $\mathcal{F}$ 表示一个系统的浮点数全体构成的集合, 则

$$\mathcal{F} = \{0\} \cup \{f : f = \pm 0.d_1 d_2 \cdots d_t \times \beta^J, 0 \leqslant d_i < \beta,$$
$$d_1 \neq 0, L \leqslant J \leqslant U\}$$

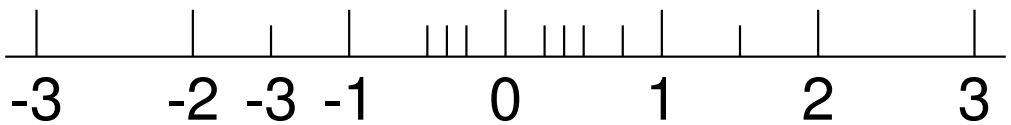
显然, 集合 $\mathcal{F}$ 可用四元数组 (β, t, L, U) 来刻画. 机器不同, 这四个值亦不同, 较典型的值是 $(2, 56, -63, 64)$.

集合 $\mathcal{F}$ 是一个包含 $2(\beta-1)\beta^{t-1}(U-L+1)+1$ 个数的有限集, 这些数对称地分布在区间 $[m,M]$ 和$[-M,-m]$中, 其中

$$m=\beta^{L-1}, \quad M=\beta^{U}(1-\beta^{-t}). \tag{1}$$

值得注意的是, 这些数在 $[m,M]$ 和 $[-M,-m]$ 中的分布是不等距的.例如, 若 $\beta=2,t=2,L=-1$ 和 $U=2$, 则 $\mathcal{F}$ 中17个数的分布如图1所示.

Figure



舍入误差界

记实数x的浮点数表示为$fl(x)$. 若$x = 0$, 则$fl(x)$取为零.
若$m \leqslant |x| \leqslant M$, 则当使用舍入法时, 取$fl(x)$为$\mathcal{F}$中最接近于$x$的数$f$;
当使用截断法时, 取$fl(x)$为满足$|f| \leqslant |x|$且最接近于x的数f. 例如,
对$(\beta, t, L, U) = (10, 3, 0, 2)$和实数$x = 5.45627$, 若用舍入法, 则
有$fl(x) = 0.546 \times 10$; 若用截断法, 则有$fl(x) = 0.545 \times 10$.

定理 2.3.1: 设$m \leqslant |x| \leqslant M$, 其中$m$和$M$由 (1)式定义, 则

$$fl(x) = x(1 + \delta), \quad |\delta| \leqslant \mathbf{u} \tag{2}$$

其中$\mathbf{u}$为 **机器精度**, 即

$$\mathbf{u} = \begin{cases} \frac{1}{2}\beta^{1-t}, & \text{用舍入法} \\ \beta^{1-t}, & \text{用截断法} \end{cases}$$

证明： 现不妨假定$x > 0$ (因若$x < 0$, 证明完全类似). 设α是满足

$$\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \tag{3}$$

的唯一整数. 在$[\beta^{\alpha-1}, \beta^{\alpha})$中浮点数的阶为$\alpha$, 所以在这个区间中所有$t$位的浮点数以间距$\beta^{\alpha-t}$分布. 对于舍入法, 根据(3)式, 有

$$|fl(x) - x| \leqslant \frac{1}{2}\beta^{\alpha-t} = \frac{1}{2}\beta^{\alpha-1}\beta^{1-t} \leqslant \frac{1}{2}x\beta^{1-t},$$

即

$$\frac{|fl(x) - x|}{x} \leqslant \frac{1}{2}\beta^{1-t}; \tag{4}$$

对于截断法, 有

$$|fl(x) - x| \leqslant \beta^{\alpha-1}\beta^{1-t} \leqslant x\beta^{1-t}, \tag{5}$$

即

$$\frac{|fl(x) - x|}{x} \leqslant \beta^{1-t}, \tag{6}$$

这就证明了(2)式.

设$a, b \in \mathcal{F}$是两个给定的浮点数, 用$\circ$来表示$+, -, \times, /$中任意一种运算. $fl(a \circ b)$ 的意义是先进行运算, 得到精确的实数, 再按舍入规则表示成浮点数, 在运算中, 若出现$|a \circ b| > M$或$0 < |a \circ b| < m$, 则就是发生了上溢或下溢. 在不发生溢出的情况下, 由定理2.3.1立即得到如下定理:

定理 2.3.2: 设$a, b \in \mathcal{F}$, 则

$$fl(a \circ b) = (a \circ b)(1 + \delta), |\delta| < \mathbf{u}.$$

例2.3.1 设x, y是两个由浮点数构成的n维向量. 试估计

$$|fl(x^T y) - x^T y|$$

的上界.

解: 令

$$S_k = fl\left(\sum_{i=1}^{k} x_i y_i\right),$$

则由定理2.3.2, 可得

$$\begin{aligned} S_1 &= x_1 y_1(1+\gamma_1), \quad |\gamma_1| \leqslant \mathbf{u}, \\ S_k &= fl(S_{k-1} + fl(x_k y_k)) \\ &= [S_{k-1} + x_k y_k(1+\gamma_k)](1+\delta_k), \quad |\delta_k|, |\gamma_k| \leqslant \mathbf{u}, \end{aligned}$$

于是有

$$\begin{aligned} fl(x^T y) = S_n &= \sum_{i=1}^{n} x_i y_i (1+\gamma_i) \prod_{j=i}^{n} (1+\delta_j) \\ &= \sum^{n} (1+\varepsilon_i) x_i y_i, \end{aligned} \tag{7}$$

其中

$$1+\varepsilon_i=(1+\gamma_i)\prod_{j=i}^{n}(1+\delta_j),$$

这里定义$\delta_1=0$. 这样, 就有

$$|fl(x^Ty-x^Ty)|\leqslant\sum_{i=1}^{n}|\varepsilon_i||x_iy_i|\leqslant 1.01n\mathbf{u}\sum_{i=1}^{n}|x_iy_i|,$$

其中最后一个不等式用了下面将要证明的定理2.3.3的结论. 注意上式表明, 若$|x^Ty|\ll\sum_{i=1}^{n}|x_iy_i|$, 则$fl(x^Ty)$的相对误差可能会很大.

向量内积运算舍入误差分析

定理 2.3.3: 若$|\delta_i| \leqslant \mathbf{u}$且$n\mathbf{u} \leqslant 0.01$, 那么

$$1 - n\mathbf{u} \leqslant \prod_{i=1}^{n}(1+\delta_i) \leqslant 1 + 1.01n\mathbf{u},$$

或者写成

$$\prod_{i=1}^{n}(1+\delta_i) = 1 + \delta, \quad |\delta| \leqslant 1.01n\mathbf{u}.$$

证明: 因为$\delta_i \leqslant \mathbf{u}$, 故有

$$(1-\mathbf{u})^n \leqslant \prod_{i=1}^{n}(1+\delta_i) \leqslant (1+\mathbf{u})^n. \tag{8}$$

先估计$(1-\mathbf{u})^n$的下界. 利用函数$(1-x)^n (0 < x < 1)$的Taylor展开

$$(1-x)^n = 1 - nx + \frac{n(n-1)}{2}(1-\theta x)^{n-2}x^2,$$

得到

$$1 - nx \leqslant (1-x)^n.$$

由此立即得到

$$1 - 1.01n\mathbf{u} \leqslant 1 - n\mathbf{u} \leqslant (1 - \mathbf{u})^n \tag{9}$$

下面估计$(1 + \mathbf{u})^n$的上界. 由e^x的幂级数展开

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ &= 1 + x + \frac{x}{2} \cdot x \cdot \left(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots\right) \end{aligned}$$

立即知道, 当$0 \leqslant x \leqslant 0.01$时,有

$$1 + x \leqslant e^x \leqslant 1 + x + \frac{0.01}{2} x e^x \leqslant 1 + 1.01x, \tag{10}$$

这里用到了$e^{0.01} < 2$这一事实. 令$x = \mathbf{u}$, 由(10)式左端不等式得

$$(1 + \mathbf{u}) \leqslant e^{n\mathbf{u}} \tag{11}$$

令$x = n\mathbf{u}$, 由(11)式右端不等式得

$$e^{n\mathbf{u}} \leqslant 1 + 1.01n\mathbf{u}. \tag{12}$$

综合(11)式和(12)式得

矩阵运算舍入误差分析

下面分析一下矩阵基本运算的舍入误差, 这些结果对以后的讨论是有用的. 为此, 引进记号

$$|E| = [|e_{ij}|], \quad E \in \boldsymbol{R}^{n\times n}$$

并且规定

$$|E| \leqslant |F| \quad \text{当且仅当} \quad |e_{ij}| \leqslant |f_{ij}|, i, j = 1, \cdots, n.$$

设A, B是由$\mathcal{F}$中的元素构成的$n \times n$矩阵, 且$\alpha \in \mathcal{F}$, 则由定理2.3.2易知

$$fl(\alpha A) = \alpha A + E, \quad |E| \leqslant \mathbf{u}|\alpha A|,$$
$$fl(A + B) = (A + B) + E, \quad |E| \leqslant \mathbf{u}|A + B|$$

此外, 应用例2.3.1可得

$$fl(AB) = AB + E, \quad |E| \leqslant 1.01n\mathbf{u}|A||B|.$$

注意, $|AB|$可能比$|A||B|$小得多, 因此矩阵乘积的相对误差未必很小.

基于这个原因, 通常计算内积时是先用双精度(即字长为$2t$)计算, 最后再把计算结果舍入为单精度数(即字长为t).

上述的这三个矩阵基本运算的舍入误差界, 是通过估计计算解与精确解之间的误差得到的, 舍入误差的界与精确解有关. 这种误差分析的方法称为**向前误差分析法**.

实际上常用的是**向后误差分析法**. 为了说明这种误差分析方法, 看一个简单的例子. 假定上面所述的矩阵A, B是2×2的上三角矩阵, 则由定理2.3.1可知

$$fl(AB)=\begin{pmatrix} a_{11}b_{11}(1+\varepsilon_1) & \widetilde{a}_{12} \\ 0 & a_{22}b_{22}(1+\varepsilon_5) \end{pmatrix}$$

其中

$$\widetilde{a}_{12}=(a_{11}b_{12}(1+\varepsilon_2)+a_{12}b_{22}(1+\varepsilon_3))(1+\varepsilon_4),$$
$$|\varepsilon_i|\leqslant \mathbf{u},\quad i=1,2,3,4,5.$$

若令

$$\widetilde{A} = \begin{pmatrix} a_{11} & a_{12}(1+\varepsilon_3)(1+\varepsilon_4) \\ 0 & a_{22}(1+\varepsilon_5) \end{pmatrix},$$

$$\widetilde{B} = \begin{pmatrix} b_{11}(1+\varepsilon_1) & b_{12}(1+\varepsilon_2)(1+\varepsilon_4) \\ 0 & b_{22} \end{pmatrix}$$

则易证

$$fl(AB) = \widetilde{A}\widetilde{B},$$

而且

$$\widetilde{A} = A + E, \qquad |E| \leqslant 3\mathbf{u}|A|,$$

$$\widetilde{B} = B + F, \qquad |F| \leqslant 3\mathbf{u}|B|.$$

换句话说, 计算得到的乘积$fl(AB)$是有了微小扰动的两个矩阵$\widetilde{A}$和$\widetilde{B}$的精确的乘积.

目录

矩阵三角分解的舍入误差分析

引理 2.4.1: 设$n \times n$浮点数矩阵$A = [a_{ij}]$有三角分解, 且$1.01n\mathbf{u} \leqslant \mathbf{0.01}$, 则用Gauss消去法计算得到的单位下三角矩阵$L$和上三角矩阵$U$ 满足

$$\widetilde{L}\widetilde{U} = A + E \tag{14}$$

其中

$$|E| \leqslant 2.05n\mathbf{u}|\widetilde{L}||\widetilde{U}|. \tag{15}$$

证明: 设$\widetilde{U} = [\widetilde{u}_{ij}], \widetilde{L} = [\widetilde{l}_{ij}]$, 由Gauss消去法的具体实现知, $\widetilde{u}_{ij}(i \leqslant j)$是从$a_{ij}$中依次减去$\widetilde{l}_{ik}\widetilde{u}_{kj}(k = 1, \cdots, i-1)$而得到的,即

$$\begin{aligned}
a_{ij}^{(0)} &= a_{ij}, \\
a_{ij}^{(k)} &= fl(a_{ij}^{(k-1)} - fl(\widetilde{l}_{ik}\widetilde{u}_{kj})), \quad k = 1, \cdots, i-2, \\
\widetilde{u}_{ij} &= a_{ij}^{(i-1)} = fl(a_{ij}^{(i-2)} - fl(\widetilde{l}_{i,i-1}\widetilde{u}_{i-1,j}))
\end{aligned}$$

由基本运算舍入误差分析的基本结果可得

$$a_{ij}^{(k)} = \left(a_{ij}^{(k-1)} - \left(\widetilde{l}_{ik}\widetilde{u}_{kj}\right)(1+\gamma_k)\right)(1+\varepsilon_k),$$

其中$|\gamma_k| \leqslant \mathbf{u}, |\varepsilon_k| \leqslant |\mathbf{u}| (k = 1, \cdots, i-1)$. 由此出发, 利用与例2.3.1同样的推理方法可得

$$\widetilde{u}_{ij} = a_{ij}(1+\delta_i) - \sum_{k=1}^{i-1}\left(\widetilde{l}_{ik}\widetilde{u}_{kj}\right)(1+\delta_k) \tag{16}$$

其中$|\delta_k| \leqslant 1.01n\mathbf{u} (k = 1, \cdots, i)$. 从(16)中将$a_{ij}$解出来, 便有

$$\begin{aligned} a_{ij} &= \frac{\widetilde{u}_{ij}}{1+\delta_i} + \sum_{k=1}^{i-1}\left(\widetilde{l}_{ik}\widetilde{u}_{kj}\right)\frac{1+\delta_k}{1+\delta_i} \\ &= \sum_{k=1}^{i}\widetilde{l}_{ik}\widetilde{u}_{kj} - e_{ij} \end{aligned} \tag{17}$$

其中$\widetilde{l}_{ii}=1$, 而

$$e_{ij}=\left(\widetilde{l}_{ii}\widetilde{u}_{ij}\right)\frac{\delta_i}{1+\delta_i}+\sum_{k=1}^{i-1}\left(\widetilde{l}_{ik}\widetilde{u}_{kj}\right)\frac{\delta_i-\delta_k}{1+\delta_i} \tag{18}$$

注意到$|\delta_k|\leqslant 1.01n\mathbf{u}<1.01$, 有

$$\left|e_{ij}\right|\leqslant\frac{2.02n\mathbf{u}}{1-0.01}\sum_{k=1}^{i}\left|\widetilde{l}_{ik}\right|\left|\widetilde{u}_{kj}\right|\leqslant 2.05n\mathbf{u}\sum_{k=1}^{i}|\widetilde{l}_{ik}||\widetilde{u}_{kj}|. \tag{19}$$

此外, 由Gauss消去法的具体实现知, $\widetilde{l}_{ij}(i>j)$的计算过程如下

$$\begin{aligned}
&a_{ij}^{(0)}=a_{ij},\\
&a_{ij}^{(k)}=fl\left(a_{ij}^{(k-1)}-fl\left(\widetilde{l}_{ik}\widetilde{u}_{kj}\right)\right),\quad k=1,2,\cdots,j-1\\
&\widetilde{l}_{ij}=fl\left(a_{ij}^{(j-1)}/\widetilde{u}_{jj}\right).
\end{aligned}$$

由基本运算的舍入误差分析的基本结果可得

$$a_{ij}^{(k)} = \left(a_{ij}^{(k-1)} - \left(\widetilde{l}_{ik}\widetilde{u}_{kj}\right)(1+\gamma_k)\right)(1+\varepsilon_k)$$
$$\widetilde{l}_{ij} = a_{ij}^{(j-1)} / \left[\widetilde{u}_{jj}(1+\delta).\right]$$

其中$|\delta| \leqslant \mathbf{u}, |\gamma_k| \leqslant \mathbf{u}, |\varepsilon_k| \leqslant \mathbf{u}$ $(k = 1, \cdots, j-1)$ 由此出发, 完全类似(17)式和(19)式的证明, 可证

$$a_{ij} = \sum_{k=1}^{j} \widetilde{l}_{ik}\widetilde{u}_{kj} - e_{ij} \tag{20}$$

其中

$$\left|e_{ij}\right| \leqslant 2.05n\mathbf{u} \sum_{k=1}^{j} |\widetilde{l}_{ik}||\widetilde{u}_{kj}| \tag{21}$$

综合上面所证, 即知引理的结论成立.

□

注意到交换矩阵的行或列并不引进舍入误差，由引理2.4.1立即得到如下结论：

推论2.4.1 设$n \times n$浮点数矩阵A是非奇异的, 且$1.01n\mathbf{u} \leq 0.01$, 则用列主元Gauss消去法计算得到的单位下三角阵$\widetilde{L}$, 上三角阵$\widetilde{U}$以及排列方阵$\widetilde{P}$满足

$$\widetilde{L}\widetilde{U} = \widetilde{P}A + E,$$

其中

$$|E| \leqslant 2.05n\mathbf{u}|\widetilde{L}||\widetilde{U}|.$$

当对A完成LU分解之后, 求解线性方程组$Ax = b$的问题就归为求解两个三角形方程组

$$\widetilde{L}y = \widetilde{P}b \quad \text{和} \quad \widetilde{U}x = y$$

的问题. 所以现在来估计求解三角形方程组的舍入误差.

求解三角方程舍入误差分析

引理2.4.2: 设$n \times n$浮点数三角阵S是非奇异的, 并且假定$1.01n\mathbf{u} \leq 0.01$, 则用第一章所介绍的解三角方程组的方法求解$Sx = b$所得到的计算解$\widetilde{x}$满足

$$(S + H)\widetilde{x} = b, \tag{22}$$

其中

$$|H| \leqslant 1.01n\mathbf{u}|S|, \tag{23}$$

证明: 对n用数学归纳法. 不失一般性, 设$S = L$是下三角阵. 当$n = 1$时, 引理显然成立. 假设对所有的$n-1$阶下三角形方程组已证引理成立. 考虑n阶的情形. 假定用前代法解$Lx = b$的计算解为$\widetilde{x}$,并将L, b和$\widetilde{x}$分块如下:

$$b = \begin{bmatrix} b_1 \\ c \end{bmatrix} \begin{matrix} 1 \\ n-1 \end{matrix}, \quad \widetilde{x} = \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{y} \end{bmatrix} \begin{matrix} 1 \\ n-1 \end{matrix}.$$

$$L = \begin{bmatrix} l_{11} & 0 \\ l_1 & L_1 \end{bmatrix} \begin{matrix} 1 \\ n-1 \end{matrix}$$

$$\begin{matrix} 1 & n-1 \end{matrix}$$

由定理2.3.2有

$$\widetilde{x}_1 = fl\left(\frac{b_1}{l_{11}}\right) = \frac{b_1}{l_{11}(1+\delta_1)}, \quad |\delta_1| \leqslant \mathbf{u} \tag{24}$$

此外，注意到是用前代法求解$n-1$阶方程组

$$L_1 y = fl\left(c - \widetilde{x}_1 l_1\right)$$

所得到的计算解, 由归纳法假设即有

$$(L_1 + H_1)\widetilde{y} = fl\left(c - \widetilde{x}_1 l_1\right),$$

其中

$$|H_1| \leqslant 1.01(n-1)\mathbf{u}\,|L_1| \tag{25}$$

$$fl\left(c - \widetilde{x}_1 l_1\right) = fl\left(c - fl\left(\widetilde{x}_1 l_1\right)\right) = \left(I + D_\gamma\right)^{-1}\left(c - \widetilde{x}_1 l_1 - \widetilde{x}_1 D_\delta l_1\right),$$

其中

$$D_\gamma = \operatorname{diag}\left(\gamma_2, \cdots, \gamma_n\right), \quad D_\delta = \operatorname{diag}\left(\delta_2, \cdots, \delta_n\right)$$
$$\left|\gamma_i\right| \leqslant \mathbf{u}, \quad |\delta_i| \leqslant \mathbf{u}, \quad i = 2, \cdots, n.$$

于是,

$$\widetilde{x}_1 l_1 + \widetilde{x}_1 D_\delta l_1 + \left(I + D_\gamma\right)(L_1 + H_1)\widetilde{y} = c,$$

从而有

$$(L + H)\widetilde{x} = b,$$

其中

$$H = \begin{bmatrix} \delta_1 l_{11} & 0 \\ D_\delta l_1 & H_1 + D_\gamma (L_1 + H_1) \end{bmatrix}.$$

由(24)式和(25)式得

$$\begin{aligned} |H| &\leqslant \begin{bmatrix} |\delta_1|\,|l_{11}| & 0 \\ |D_\delta|\,|l_1| & |H_1| + \left|D_\gamma\right|(|L_1| + |H_1|) \end{bmatrix} \\ &\leqslant \begin{bmatrix} u\,|l_{11}| & 0 \\ \mathbf{u}\,|l_1| & |H_1| + \mathbf{u}\,(|L_1| + |H_1|) \end{bmatrix} \\ &\leqslant u \begin{bmatrix} |l_{11}| & 0 \\ |l_1| & (1.01(n-1) + 1 + 1.01(n-1)\mathbf{u})\,|L_1| \end{bmatrix} \\ &\leqslant 1.01n\mathbf{u}|L|, \end{aligned}$$

其中最后一个不等式用到了假设条件$1.01n\mathbf{u} \leqslant 0.01$.

列主元高斯消去法舍入误差分析

应用引理2.4.2得到三角形方程组

$$\widetilde{L}y = \widetilde{P}b \quad \text{和} \quad \widetilde{U}x = y$$

即知最后得到的计算解$\widetilde{x}$应满足

$$(\widetilde{L} + F)(\widetilde{U} + G)\widetilde{x} = \widetilde{P}b$$

即

$$(\widetilde{L}\widetilde{U} + F\widetilde{U} + \widetilde{L}G + FG)\widetilde{x} = \widetilde{P}b \tag{26}$$

其中

$$|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|, \quad |G| \leqslant 1.01n\mathbf{u}|\widetilde{U}|. \tag{27}$$

再将$\tilde{L}\widetilde{U} = \widetilde{P}A + E$带入(26)式, 得

$$(A + \delta A)\widetilde{x} = b,$$

这里

$$\delta A = \widetilde{P}^T(E + F\widetilde{U} + \tilde{L}G + FG).$$

由(27)式和推论2.4.1 得

$$|\delta A| \leqslant 4.09n\mathbf{u}\widetilde{P}^T|\tilde{L}||\widetilde{U}|.$$

注意到$\tilde{L}$的元素的绝对值均不超过1, 故有

$$\|\widetilde{L}\|_\infty \leqslant n.$$

为了给出$\|\widetilde{U}\|_\infty$的估计, 定义

$$\rho = \max_{i,j} \left|\widetilde{u}_{ij}\right| / \max_{i,j} \left|a_{ij}\right|,$$

通常称之为列主元 Gauss 消去法的**增长因子**. 于是,

$$\|\widetilde{U}\|_\infty \leqslant n \max_{i,j} \left|\widetilde{U}_{ij}\right| = n\rho \max_{i,j} \left|a_{ij}\right| \leqslant n\rho\|A\|_\infty.$$

这样, 就得到了本节的主要定理:

定理2.4.1: 设$n \times n$浮点数矩阵A是非奇异的, 且$1.01n\mathbf{u} \leqslant 0.01$, 则用列主元 Gauss 消去法解线性方程组$Ax = b$ 所得到的计算解$\widetilde{x}$ 满足

$$(A + \delta A)\widetilde{x} = b \tag{28}$$

其中

$$\|\delta A\|_\infty / \|A\|_\infty \leqslant 4.09n^3\rho\mathbf{u}. \tag{29}$$

定理2.4.1表明, 由于消去法求解过程中引进舍入误差而产生的计算解相当于系数矩阵作某些扰动而得到的扰动方程组的精确解. 一般来说, δA的元素比起A的元素的初始误差（数据的测量误差、数学模型误差等）来是很小的. 在这个意义上来讲, 列主元Gauss消去法是数值稳定的.

最后需指出的是, 理论上可以证明$\rho \leq 2^{n-1}$, 而且上界可以达到. 但在实际运算时, 常遇到的问题其ρ很小, 一般来讲不会超过n. 此外, 定理2.4.1中所给出的上界$4.09n^3\rho\mathbf{u}$一般要比真正的$\|\delta A\|_\infty/\|A\|_\infty$大很多.在实际计算时,常遇到的问题几乎都有$\|\delta A\|_\infty/\|A\|_\infty \approx \mathbf{u}$.

目录

设用某种计算方法求解线性代数方程组$Ax = b$得到的计算解为$\hat{x}$.
令

$$r = b - A\hat{x},$$

则有

$$r = Ax - A\hat{x} = A(x - \hat{x}).$$

于是

$$\|x - \hat{x}\| = \left\|A^{-1}r\right\| \leqslant \left\|A^{-1}\right\| \|r\|.$$

再注意到

$$\|b\| \leqslant \|A\|\|x\|,$$

即有

$$\frac{\|x - \hat{x}\|}{\|x\|} \leqslant \left\|A^{-1}\right\| \|A\| \frac{\|r\|}{\|b\|}.$$

特别地, 在上式中取∞范数便有

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \leqslant \kappa_\infty(A) \frac{\|r\|_\infty}{\|b\|_\infty}. \tag{30}$$

这样可在实际计算时通过计算

$$\kappa_\infty(A)\frac{\|r\|_\infty}{\|b\|_\infty}$$

来给出计算解的精度估计, 而上式中除了$\kappa_\infty(A)$外, 其余的量都是容易计算的. 注意到

$$\kappa_\infty(A) = \left\|A^{-1}\right\|_\infty \|A\|_\infty,$$

而$\|A\|_\infty$又是易于计算的, 便知用(30)式来估计计算解精度的关键在于如何估计$\left\|A^{-1}\right\|_\infty$. 现在已有不少的实用方法可以给出这一估计. 这里介绍LAPACK所采用的一种优化方法, 该方法就是著名的“盲人爬山法”的一个具体应用.

设$B \in \mathbf{R}^{n\times n}$, 下面估计$\|B\|_1$. 定义

$$\begin{aligned} f(x) &= \|Bx\|_1 = \sum_{i=1}^{n} \left|\sum_{j=1}^{n} b_{ij}x_j\right|, \\ \mathcal{D} &= \{x \in \mathbf{R}^n : \|x\|_1 \leqslant 1\} \end{aligned},$$

则易证f是凸函数, $\mathcal{D}$是凸集, 而且求$\|B\|_1$的问题就等价于求凸函数f在凸集$\mathcal{D}$上的最大值问题.
设f在x点的梯度$\nabla f(x)$存在, 则由凸函数的性质可知

$$f(y) \geqslant f(x) + \nabla f(x)^T(y-x), \quad y \in \mathbf{R}^n.$$

现假定$x_0 = (x_j^{(0)}) \in \mathbf{R}^n$满足$\|x_0\|_1 = 1$, 使得

$$\sum_{j=1}^{n} b_{ij}x_j^{(0)} \neq 0, \quad i = 1, \cdots, n.$$

令

$$\xi_i = \operatorname{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right),$$

则在x_0附近有

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i b_{ij} x_j.$$

因此, 有

$$\begin{aligned}
\nabla f(x_0) &= \left(\frac{\partial f(x_0)}{\partial x_1}, \cdots, \frac{\partial f(x_0)}{\partial x_n}\right)^T \\
&= \left(\sum_{i=1}^{n} \xi_i b_{i1}, \cdots, \sum_{i=1}^{n} \xi_i b_{ij}, \cdots, \sum_{i=1}^{n} \xi_i b_{in}\right)^T. \\
&= B^T v
\end{aligned}$$

其中$v = (\xi_1, \cdots, \xi_n)^T$.

再令

$$w = Bx_0, \quad z = B^T v.$$

于是, 有下面结论成立:

定理 2.5.1: 假定B, x_0, v, w和z 如上所述, 则有

1. 若$\|z\|_\infty = z^T x_0$ (**Holder不等式成立**),
   则$\|w\|_1 = \|Bx_0\|_1$是$f(x) = \|Bx\|_1$ 在$\mathcal{D}$ 中的局部极大值;

2. 若$\|z\|_\infty > z^T x_0$, 则$\left\|Be_j\right\|_1 > \|Bx_0\|_1$, 其中$j$ 满足

$$\left|z_j\right| = \|z\|_\infty.$$

证明： (1) 由于在x_0附近$f(x)$是x的线性函数, 因此有

$$f(x) = f(x_0) + \nabla f(x_0)^T (x - x_0).$$

这样, 只需证在x_0附近有

$$\nabla f(x_0)^T (x - x_0) = z^T (x - x_0) \leqslant 0$$

即可. 事实上, 对于$\|x\|_1 \leqslant 1$, 有

$$\begin{aligned} z^T (x - x_0) &= z^T x - z^T x_0 \\ &\leqslant \|z\|_\infty \|x\|_1 - z^T x_0 \\ &\leqslant \|z\|_\infty - z^T x_0 \leqslant 0. \end{aligned}$$

(2) 取$\widetilde{x}=e_j \operatorname{sgn}\left(z_j\right)$, 则有

$$\begin{aligned}
\left\|Be_j\right\|_1 &= \|B\widetilde{x}\|_1 = f(\widetilde{x}) \\
&\geqslant f\left(x_0\right)+\nabla f\left(x_0\right)^T\left(\widetilde{x}-x_0\right) \\
&= \|Bx_0\|_1 + z^T\widetilde{x} - z^T x_0 \\
&= \|Bx_0\|_1 + \left|z_j\right| - z^T x_0 \\
&= \|Bx_0\|_1 + \|z\|_\infty - z^T x_0 \\
&> \|Bx_0\|_1
\end{aligned}$$

即(2)成立.

盲人爬山法

基于这一定理可设计算法如下:

算法2.5.1: 估计矩阵的1 范数的优化法

设$k = 1$

while $k = 1$

$w = Bx;\ v = \mathbf{sign}(w);\ z = B^T v$

if $\|z\|_\infty = z^T x$

$v = \|w\|_1$

$k = 0$

else

$x = e_j$, 其中$|z_j| = \|z\|_\infty$

$k = 1$

end

end

该算法初始的x可选任意满足$\|x\|_1 = 1$的向量, 例如可取

$$x_i = 1/n, \quad i = 1, \cdots, n.$$

假如已经计算好矩阵A 的列主元三角分解: $PA = LU$，则利用上述算法仅用$O\left(n^2\right)$的运算量就可给出$\left\|A^{-1}\right\|_\infty$的一个估计值.

由于$\left\|A^{-1}\right\|_\infty = \left\|A^{-T}\right\|_1$, 因此只需应用算法2.5.1 于矩阵$B = A^{-T}$上即可, 此时计算$w = Bx$和$z = B^T v$就相当于解方程组$A^T w = x$和$Az = v$, 利用$A$的三角分解这两个方程组是很容易求解的.

综合上述讨论, 可以按如下的步骤来估计一个计算解$\hat{x}$的精度:

1. 应用算法2.5.1于$B = A^{-T}$上得到$\left\|A^{-1}\right\|_\infty$的一个估计值$\tilde{\nu}$;
2. 分别计算$\|r\|_\infty$, $\|b\|_\infty$ 和$\|A\|_\infty$得到它们的计算值$\widetilde{\gamma}$, $\tilde{\beta}$, 和$\tilde{\mu}$;
3. 计算$\tilde{\rho} = \dfrac{\widetilde{\nu\mu\gamma}}{\widetilde{\beta}}$, 则数$\tilde{\rho}$就可以作为计算解$\hat{x}$的相对误差的一个估计.

对于绝大多数问题这一方法常常可以给出计算解相对误差的相当好的估计, 但是也有一些特殊的问题利用该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差. 这主要是由两个方面的原因引起的:一是由于舍入误差的影响使得计算得到的$\tilde{\gamma}$远远小于$\|r\|_\infty$的真值; 二是当A十分病态时计算得到的三角分解已经相当不准确, 以至于应用它去估计出的$\tilde{\nu}$要比$\left\|A^{-1}\right\|_\infty$的真值小得多. 虽然现在已经有了一些部分解决这些问题的方法, 但怎样才能使给出的估计更好仍然是一个值得进一步深入研究的问题.

迭代改进法

若计算的解$\hat{x}$的精度太低, 可将$\hat{x}$作为初值, 应用Newton迭代法于函数$f(x) = Ax - b$上, 来改进其精度. 具体计算过程可按如下步骤进行:

(1) 计算$r = b - A\hat{x}$ (用双精度和原始矩阵A).

(2) 求解$Az = r$(利用A的三角分解).

(3) 计算$x = \hat{x} + z$.

(4) 若$\frac{\|x-\hat{x}\|_\infty}{\|x\|_\infty} \leqslant \varepsilon$, 则结束; 否则, 令$\hat{x} = x$, 转步(1).

实际计算的经验表明, 当A的病态并不是十分严重时, 利用这一方法最终可使其解的计算精度达到机器精度. 可是，当A十分病态时，这样做对解的精度并不会有太大的改进.