Floating point representation : x : real number

Note Title 12/23/2010

$$fl(x) = \pm (.d_1 d_2 \cdots d_n)\, \beta^e$$

$$d_1 \neq 0, \quad 0 \leq d_i \leq \beta - 1$$

β : base (or radix)

$.d_1 d_2 \cdots d_n$: significand (or mantissa)

length of significand : precision

e : exponent : $m < e < M$

β : 2, 16, 10

Significant digits : first nonzero digit and all succeeding digits.

Examples : 1.7320 : five significant digits

0.0491 : three significant digits

Note Title 12/24/2010

Cancellation: $f(x) = \dfrac{1 - \cos x}{x^2}$

$x = 1.2 \times 10^{-5}$, $\cos x = 0.9999\ 9999\ 99$

rounded to 10 significant digits.

$1 - \cos x = 0.0000\ 0000\ 01$

$$\frac{1 - \cos x}{x^2} = \frac{10^{-10}}{1.44 \times 10^{-10}} = 0.6944\ldots$$

$0 \leq f(x) < \frac{1}{2}$. $1 - \cos x$: 1 significant digit — wrong

$$x = a - b$$

$$\hat{a} = a(1 + \Delta a), \quad \hat{b} = b(1 + \Delta b)$$

$$\hat{x} = \hat{a} - \hat{b} = a - b + a\,\Delta a - b\,\Delta b$$

$$\left| \frac{x - \hat{x}}{x} \right| = \left| \frac{-a\,\Delta a + b\,\Delta b}{a - b} \right|$$

$$\leq \max(|\Delta a|, |\Delta b|) \frac{|a| + |b|}{|a - b|}$$

$|a - b| << |a| + |b| \Rightarrow$ relative error for $\hat{x}$ is large

Small Pivots

Note Title 11/24/2010

Consider well-conditioned →

$$\begin{bmatrix} .002 & 1.231 & 2.471 \\ 1.196 & 3.165 & 2.543 \\ 1.475 & 4.271 & 2.142 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3.704 \\ 6.904 \\ 7.888 \end{bmatrix}$$

exact solution: $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

small pivot: .002

Small Pivots

Note Title 11/24/2010

well-conditioned →

$$\begin{bmatrix} .002 & 1.231 & 2.471 \\ 1.196 & 3.165 & 2.543 \\ 1.475 & 4.271 & 2.142 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3.704 \\ 6.904 \\ 7.888 \end{bmatrix}$$

$$m_{21} = \frac{1.196}{0.002} = 598.0 \quad , \quad m_{31} = \frac{1.475}{0.002} = 737.5$$

$$R_2 - m_{21} R_1 \quad , \quad R_3 - m_{31} R_1$$

$$3.165 - m_{21}(1.231)$$

$3.165 - (598.0)(1.231) = 3.165 - 736.1 = -732.9$

The last two digits of 3.165 are lost.

Information loss : swamping

At the end of the first step:

$$\begin{bmatrix} .002 & 1.231 & 2.471 \\ 598.0 & -732.9 & -1475. \\ 737.5 & -903.6 & -1820 \end{bmatrix} \tilde{A}$$

The two rows of A are almost linearly dependent: multiples of $[1.231, 2.471]$

$k_\infty(\tilde{A}) \simeq 8400$

$$\begin{bmatrix} .002 & 1.231 & 2.471 \\ 598.0 & -732.9 & -1475. \\ 737.5 & -903.6 & -1820 \end{bmatrix} \tilde{A}$$

Second Step : $m_{32} = \dfrac{-903.6}{-732.9} = 1.233$

$$-1820 - (1.233)(-1475) = -1820. + 1819. = -1.000$$

Severe Cancellation

$Ax = b \longrightarrow Ux = y$

$y_1 = 3.704$,

$y_2 = 6.904 - (598.0)(3.704) = 6.904 - 2215.$

swamping $= -2208$

$y_3 = 7.888 - (737.5)(3.704) - (1.233)(-2208)$

$= 7.888 - 2732. + 2722. = -2724. + 2722$

severe cancellation $= -2.000$

Solution : $\begin{bmatrix} 4.000 \\ -1.012 \\ 2.000 \end{bmatrix}$, Exact solution: $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Small pivot $\Rightarrow$ large multipliers

Very large multiples of pivotal row get subtracted from other rows : swamping

resulting submatrix : ill conditioned

ill conditioning $\Rightarrow$ cancellation

Small Pivots

Consider

well-conditioned →

$$\begin{bmatrix} .002 & 1.231 & 2.471 \\ 1.196 & 3.165 & 2.543 \\ 1.475 & 4.271 & 2.142 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3.704 \\ 6.904 \\ 7.888 \end{bmatrix}$$

Multiply the first equation by 1000

Note Title 12/9/2010

Consider

$$\begin{bmatrix} 2.000 & 1231. & 2471. \\ 1.196 & 3.165 & 2.543 \\ 1.475 & 4.271 & 2.142 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3704. \\ 6.904 \\ 7.888 \end{bmatrix}$$

obtained from (1) by multiply the first equation by 1000.

The pivot 2.000 is not small.

$$m_{21} = \frac{1.196}{2.000} = .5980 \,, \quad m_{31} = \frac{1.475}{2.000} = .7375$$

$3.165 - (.5980)(1231.) = 3.165 - 736.1 = -732.9$

Same result as before

Note Title 12/9/2010

$$\begin{bmatrix} 2.000 & 1231. & 2471. \\ 1.196 & 3.165 & 2.543 \\ 1.475 & 4.271 & 2.142 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3704. \\ 6.904 \\ 7.888 \end{bmatrix}$$

$m_{21} = \frac{1.196}{2.000} = .5980$, $m_{31} = \frac{1.475}{2.000} = .7375$

$3.165 - (.5980)(1231.) = 3.165 - 736.1 = -732.9$

Same result as before

Note Title 12/9/2010

After first step:

$$\left[\begin{array}{rrr} 2.000 & 1231. & 2471. \\ .5980 & -732.9 & -1475. \\ .7375 & -903.6 & -1820 \end{array}\right] \tilde{A}$$

Earlier:

$$\left[\begin{array}{rrr} .002 & 1.231 & 2.471 \\ 598.0 & -732.9 & -1475. \\ 737.5 & -903.6 & -1820 \end{array}\right] \tilde{A}$$

Later Computations: same.

Same Solution

Earlier system : small pivot, large multiplier

$$\begin{bmatrix} .002 & 1.231 & 2.471 \\ 1.196 & 3.165 & 2.543 \\ 1.475 & 4.271 & 2.142 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3.704 \\ 6.904 \\ 7.888 \end{bmatrix}$$

New System: First row large

$$\begin{bmatrix} 2.000 & 1231. & 2471. \\ 1.196 & 3.165 & 2.543 \\ 1.475 & 4.271 & 2.142 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3704. \\ 6.904 \\ 7.888 \end{bmatrix}$$

Ill conditioned : rows & columns are out of scale.

Condition number $\geq \dfrac{\| C_j \|}{\| C_i \|}$

loss of accuracy : i) catastrophic cancellation,

ii) gradual accumulation of small errors.

↑ does not happen in practice

If no cancellations occur in an algorithm, then the result will be accurate.

↑ difficult to verify.

Forward or direct approach: Find a bound for each intermediate result

Usually not possible since for each addition/subtraction, one has to prove that: no catastrophic cancellation

Backward Error Analysis

Exact equation : $Ax = b$

$\hat{x}$: computed solution .

We try to find matrix δA such that

$$(A + \delta A)\hat{x} = b .$$

Use perturbation theory to find an estimate for

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \frac{k(A)\,\frac{\|\delta A\|}{\|A\|}}{1 - k(A)\,\frac{\|\delta A\|}{\|A\|}}$$

$$Ax = b : \quad LUx = b$$

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

We show that

$$\|\delta A\| \leq 3n\epsilon\, \|L\|\,\|U\|$$

GEPP : $\|\delta A\|_\infty \leq 3\, g\, n^3 \epsilon\, \|A\|_\infty$ Partial pivoting

$$g = \frac{\max |u_{ij}|}{\max |a_{ij}|} \leq 2^{n-1}$$

GECP : $g \leq n^{\frac{1}{2} + \log_e \frac{n}{4}}$ Complete Pivoting

Cholesky : $\|\delta A\|_\infty \leq 3 n^2 \epsilon\, \|A\|_\infty$

Tutorial 4

Note Title 4/5/2011

Q. Let $A = [a_{ij}]$ be an $n \times n$ matrix. Define

$$\|A\|_{max} = \max_{1 \le i,j \le n} |a_{ij}| .$$

Does the following property hold?

$$\|AB\|_{max} \le \|A\|_{max} \|B\|_{max}$$

Solution: $\|A\|_{max} = \max_{1 \le i,j \le n} |a_{ij}|$.

1) $\|A\|_{max} \ge 0$, $A = 0 \Rightarrow \|A\|_{max} = 0$

$\|A\|_{max} = 0 \Rightarrow a_{ij} = 0, \ 1 \le i,j \le n \Rightarrow A = 0$

2) $\alpha A = [\alpha a_{ij}]$, $\|\alpha A\|_{max} = \max_{i,j} |\alpha a_{ij}|$

$= |\alpha| \, \|A\|_{max}$

3) $A + B = [a_{ij} + b_{ij}]$, $\|A+B\|_{max} = \max_{i,j} |a_{ij} + b_{ij}|$

$\le \max |a_{ij}| + \max |b_{ij}| = \|A\|_{max} + \|B\|_{max}$

$$C = AB$$

$$c_{ij} = \sum_{k=1}^{n} a_{ik} \, b_{kj}$$

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

$$\| A \|_{max} = 1 , \quad \| A^2 \|_{max} = 2$$

$$\| A^2 \|_{max} \not\leq \| A \|_{max}^2$$