Lecture Notes

Section 1.3
Floating Pt Number Systems

I Example: Modulus of Elasticity of Steel

$$E = \frac{4mgl^3}{dab^3}$$

(see Example on page 31)

Types of Errors
Modeling
Discretization/Truncation
Roundoff/Data
Human Error

II Floating Point Number System

a) Definition: $F(B, k, m, M)$

B: the base
k: # of digits in Base expansion
m: minimum exponent
M: max exponent.

i.e. $\pm (.d_1 d_2 \ldots d_k)_B \times B^e$

where $d_1 \neq 0$ & $m \leq B \leq M$

b) Example: $F(10, 2, 0, 2)$

i.e. 2 digits, base 10, $0 \leq e \leq 2$

$= \{.10, .11, \ldots .99\}$ A $e=0$

$\{1.0, 1.1, \ldots 9.9\}$ A $e=1$

$\{10, 11, \ldots 99\}$ $e=2$

III. Round-off Error

a) Definition: Round off error is error induced by converting real # to float pt. equivalent

b) Example: $\sqrt{7.1} = 2.66458\ldots\ldots$

in $F(10,2,0,2)$

$\sqrt{7.1} = 2.6$ ("chopping" the number = "fl_{CHOP}")

$\sqrt{7.1} = 2.7$ (rounding the number = "fl_{ROUND}")

c) Definition: Let p^* be any approximation of a number p

① Absolute Error: $|p^* - p|$

② Relative Error: $|p^* - p| / |p| \Rightarrow p \neq 0$

i.e. Example above using fl_{CHOP}

$p^* = 2.66458 \quad p = 2.6$

$\Rightarrow$ Abs Err $= |2.6 - 2.66458| = 6.458 \times 10^{-2}$

$$\text{Rel Er} = \frac{|2.6 - 2.66458|}{|2.66458|} = .0242 \text{ or } 2.42\%$$

d) Absolute Error General Expressions

① $\frac{|fl_{chop}(y) - y|}{|y|} \leq \beta^{1-k}$ (proof on p 35)

↑ for $fl(\beta, k, m, M)$

② $\frac{|fl_{round}(y) - y|}{|y|} \leq \frac{1}{2}\beta^{1-k}$ (test for proof in HW)

e Definition Machine Precision 'u'

$$u = \begin{cases} \beta^{1-k} & \text{chop} \\ \frac{1}{2}\beta^{1-k} & \text{round} \end{cases}$$

I.E single precision (IEEE standard) is $F(2, 25, -125, 128)$ ← Base 2

$$u = \begin{cases} 2^{1-25} = 2^{-24} & \text{chop} \\ \left(\frac{1}{2}\right)2^{1-25} = 2^{-25} & \text{Round.} \end{cases}$$

I.E. double precision is $F(2, 53, -1021, 1024)$

$\therefore u =$ ____

(5)

Definition: $B^{-(t+1)} < \left|\frac{x-y}{x}\right| < B^{-t}$

Then x & y agree to at least t and at most t+1 "significant" base B "digits"

Example: 1.9 p 36

IV Conditioning

Solve: $x' - x = e^{-2t}$, $x(0) = -1/2$ Integrating Factor $u(t) = e^{\int -dt} = e^{-t}$

$\Rightarrow e^{-t}x' - e^{-t}x = e^{-3t} \Rightarrow \int \frac{d}{dt}[e^{-t}x] = \int e^{-3t}dt$

$\Rightarrow -e^{-t}x = -\frac{1}{3}e^{-3t} + C$

$\Rightarrow x = -\frac{1}{3}x^{-2t} + Ce^{t} \Rightarrow x(0) = -\frac{1}{3} + C = -\frac{1}{3} \Rightarrow C = 0$

$\therefore$ $\boxed{x = -\frac{1}{3}x^{-2t}}$

$\Rightarrow$ Now perturb initial condition i.e. $x(0) = -\frac{1}{3} + \epsilon$ ← a tiny bump

$x(0) = -\frac{1}{3} + C = -\frac{1}{3} + \epsilon \Rightarrow C = \epsilon$

$\therefore$ $\boxed{x = -\frac{1}{3}e^{-2t} + \epsilon e^{t}}$

this tiny 'bump' causes solution to 'blow up'

Problem is "ill-conditioned"

V Appendices

A Proof of General Error Expression for "Chopping"

For $F(B,k,m,AK)$ $\quad \dfrac{|fl_{CHOP}(y)-y|}{|y|} \leq B^{1-k}$

① let $y = (.d_1 d_2 \ldots d_k d_{k+1} d_{k+2} \ldots)_B \times B^e$

② $fl_{CHOP}(y) = (.d_1 d_2 \ldots d_k)_B \times B^e$

③ $|fl_{CHOP}(y)-y| = (.\underbrace{0000\ldots0}_{k \text{ zeros}} d_{k+1} d_{k+2} \ldots)_B \times B^e$

$= (.d_{k+1} d_{k+2} \ldots)_B \times B^{e-k}$

④ $\dfrac{|fl_{CHOP}(y)-y|}{|y|} = \dfrac{(.d_{k+1} d_{k+2} \ldots)_B \times B^{e-k}}{(.d_1 d_2 \ldots)_B \times B^e}$

$= \dfrac{(.d_{k+1} d_{k+2} \ldots)_B \, B^{-k}}{(.d_1 d_2 \ldots)_B}$

⑤ note: $(.d_1 d_2 \ldots)_B \geq (.1)_B = B^{-1}$

and $(.d_{k+1} d_{k+2} \ldots) < 1_B$

$\therefore \dfrac{|fl_{CHOP}(y)-y|}{|y|} < \dfrac{1_B * B^{-k}}{B^{-1}} = B^{1-k}$

QED

B) Matlab eps "smallest # in Matlab

eps = $2.2204\ldots \times 10^{-16}$

$\log_2(\text{eps}) = -52$

$\therefore$ precision = $B^{1-k} = 2^{-52} \Rightarrow k = 53$

Recall Double Precision)

$F(2, 53, -1021, 1024)$

↑ Matlab actually appears to be $-1074, 1023$