Numerical Analysis

GOAL: solve maths/physics problems with a COMPUTER

Problem → $\boxed{\text{MATHEMATICAL MODEL}}$ → (ALGORITHM) $\boxed{\text{NUMERICAL MODEL}}$ → $\boxed{\text{COMPUTER IMPLEMENTATION AND RUN}}$ → Solution

SOURCES of ERRORS

- Physical approximation (MATH MODEL) $\nabla \times \bar{H} = \bar{J} + \cancel{\partial \bar{D}/\partial t} \simeq \bar{J}$
- TRUNCATION ERRORS (NUMERICAL MODE) conversion of analytical operators
- ROUNDOFF ERRORS (COMPUTER IMPLEMENT.) finite # of digits to represent numbers

NUMBER REPRESENTATION

POSITIONAL REPRESENTATION

EX: $(3012)_{10} = 3 \cdot 10^3 + 0 \cdot 10^2 + 1 \cdot 10^1 + 2 \cdot 10^0$

Pos. 2 Pos. 0 digits BASE

INTEGERS

$$q = a_n \beta^n + a_{n-1} \beta^{n-1} + \dots + a_1 \beta^1 + a_0 \beta^0$$

DIGITS, $a_k \in \mathbb{N}$

$0 \leq a_k \leq \beta - 1$

$a_n \neq 0$

BASE, $\beta \in \mathbb{N}$

$\beta \geq 2$

REALS $\quad X = \lfloor X \rfloor + frac(X)$

$\lfloor X \rfloor$: integer part

$frac(X)$: fractional part

EX: $(3012.401)_{10}$

pos = -1   pos: -3

RADIX POINT

$$3\cdot 10^3 + 0\cdot 10^2 + 1\cdot 10^1 + 2\cdot 10^0 + 4\cdot 10^{-1} + 0\cdot 10^{-2} + 1\cdot 10^{-3}$$

$$(X)_\beta = \underbrace{a_n\beta^n + a_{n-1}\beta^{n-1} + \dots + a_0\beta^0}_{\lfloor X \rfloor} + \underbrace{b_1\beta^{-1} + b_2\beta^{-2} + \dots + b_m\beta^{-m}}_{frac(X)}$$

$b_k \in \mathbb{N}$

$0 \leq b_k \leq \beta - 1$

$b_m \neq 0$

FIXED POINT REPRESENTATION

- Positional representation
  - fixed number of digits
  - fixed position RADIX POINT

FIXED POINT SET

$$X(\beta, t, q) = \left\{ x \in \mathbb{R} = sign(x)\left[\sum_{k=0}^{t-(q+1)} a_k\beta^k + \sum_{k=1}^{q} b_k\beta^{-k} \right] \right\}$$

β: base

t: number of digits

q: number of digits for fractional part

$0 \leq q \leq t, \; q \in \mathbb{N}$

$$X(\beta, t, q) = \left\{ x \in \mathbb{R} = \text{sign}(x) \left[\sum_{k=0}^{t-(q+1)} a_k \beta^k + \sum_{k=1}^{q} b_k \beta^{-k} \right] \right\}$$

EX: $X(\beta = 10, \underbrace{t = 4, q = 1}_{})$

3 digits for $\lfloor x \rfloor$
1 " $\text{frac}(x)$

$$\max(x) = 9 \cdot 10^2 + 9 \cdot 10^1 + 9 \cdot 10^0 + 9 \cdot 10^{-1} = 999.9$$

$$\min(x) = 0 \cdot 10^2 + 0 \cdot 10^1 + 0 \cdot 10^0 + 1 \cdot 10^{-1} = 0{,}1$$

$$\begin{array}{ccc}
\lfloor x \rfloor & & \text{frac}(x) \\
9\ 9\ 9 & . & 9 \\
 & . & 8 \\
 & & \vdots \\
9\ 9\ 8 & . & 9 \\
 & . & 8 \\
 & & \vdots \\
9\ 9\ 7 & . & 9 \\
\vdots & & \\
0\ 0\ 0 & . & 9 \\
 & & \vdots \\
0\ 0\ 0 & . & 0{,}1
\end{array}$$

$\updownarrow 0{,}1$

- \# of POSITIVE ELEMENTS
  $\beta^t - 1 = 10000 - 1 = 9999$
- maximum element
  $(\beta^t - 1)\,\beta^{-q} = 9999 \cdot 10^{-1} = 999.9$, with $q = 1$
- minimum element
  $\beta^{-q} = 10^{-1} = 0{,}1$

$\boxed{}$ SPACING: $\Delta = \beta^{-q}$

$\Rightarrow$ CONSTANT

ERRORS of fixed point representations Hp: $X(\beta = 10, t = 4, q = 1)$

ABS. ERROR $E(x) = x - \text{fip}(x)$

$x_1 = 10^3/3 = 333.\bar{3}$ $E(x_1) = 333.\bar{3} - \underset{\text{fip}(x)}{333.3} = 0{,}0\bar{3}$

$x_2 = 10^0/3 = 0.\bar{3}$ $E(x_2) = 0{,}\bar{3} - 0{,}3 = 0{,}0\bar{3}$

$\rightarrow$ ABS ERROR is CONSTANT

RELATIVE ERROR: $e(x) = \left| \frac{E(x)}{x} \right|$

$$e(x_1) = \left| \frac{0,0\bar{3}}{10^3/3} \right| = \left| \frac{10^{-1}/3}{10^3/3} \right| = 10^{-4} \rightarrow \text{"WRONG" by } 1/10000$$

$$e(x_2) = \left| \frac{0,0\bar{3}}{10^0/3} \right| = \left| \frac{10^{-1}/3}{10^0/3} \right| = 10^{-1} \rightarrow \text{"WRONG" by } 1/10$$

→ VARIABLE relative error → relative ACCURACY is smaller for small numbers

PRO:

- SIMPLE → ALLOWS for FAST ARITHMETICS

⇒ VIDEOGAMES

CONS:

- NON-CONSTANT REL. ERROR

FLOATING POINT REPRESENTATION

any x can be represented:

$$x = \text{sign}(x) \underbrace{\left[\sum_{k=0}^{\infty} d_k \beta^{-k} \right]}_{\text{MANTISSA } m} \beta^{+p}$$

EXPONENTIAL PART, $p \in \mathbb{N}$

MANTISSA m

$1 \leq m < \beta$

digits

$0 \leq d_k \leq \beta - 1$

$d_0 \neq 0$

FLOATING POINT SET

$$F(\beta, t, L, U) = \{0\} \cup \left\{ x \in \mathbb{R} = \text{sign}(x) \left[\sum_{k=0}^{t-1} d_k \beta^{-k} \right] \beta^p \right\}$$

t: # of digits of m

L, U: $\in \mathbb{N}$

$p \in [L, U]$

$X(\beta=10, t=4, q=1)$

$\max(X) = 999.9$

$\min(X) = 0,1$

$F(\beta=10, t=3, L=0, U=9)$

$p \in [0, 9]$

$$\max(F) = \underbrace{\left[\overset{9}{d_0}\beta^0 + d_{-1}\beta^{-1} + d_{-2}\beta^{-2}\right]}_{m} \cdot \beta^p$$

$\max(m) = 9,99$

$\max(\beta^p) = 10^9$

$\max(F) = 9,99 \cdot 10^9 \sim 10^{10}$

RANGE $X \rightarrow 0,1 \div 999.9$

RANGE $F \rightarrow 1 \div 999.9 \cdot 10^9$

$$\min(F) = \underbrace{\left[\overset{1}{d_0}\beta^0 + \overset{0}{d_{-1}}\beta^{-1} + \overset{0}{d_{-2}}\beta^{-2}\right]}_{m} \cdot \beta^p$$

$\min(F) = 1.00 \cdot 10^0 = 1$

DISCRETIZATION

$\Delta \rightarrow$ NOT CONSTANT

$\left|\frac{E(x)}{x}\right|$

ERRORS:

$$e(x) = \left|\frac{x - flp(x)}{x}\right| = \left|\frac{\overbrace{sign(x)\left[\sum_{k=0}^{\infty} d_k \beta^{-k}\right]\cancel{\beta^p}}^{x} - sign(x)\left[\sum_{k=0}^{t-1} d_k \beta^{-k}\right]\cancel{\beta^p}}{\underbrace{sign(x)\left[\sum_{k=0}^{\infty} d_k \beta^{-k}\right]}_{m(x)}\cancel{\beta^p}}\right|$$

$$= \frac{\sum_{k=t}^{\infty} d_k \beta^{-k}}{m} \leq \frac{\sum_{k=t}^{\infty} d_k \beta^{-k}}{1} \Rightarrow e(x) < \beta^{1-t}$$

$m \geq 1$

$$\sum_{k=t}^{\infty} d_k \beta^{-k} = d_t \beta^{-t} + d_{t+1}\beta^{-(t+1)} + \dots + d_{t+m}\beta^{-(t+m)} + \dots$$

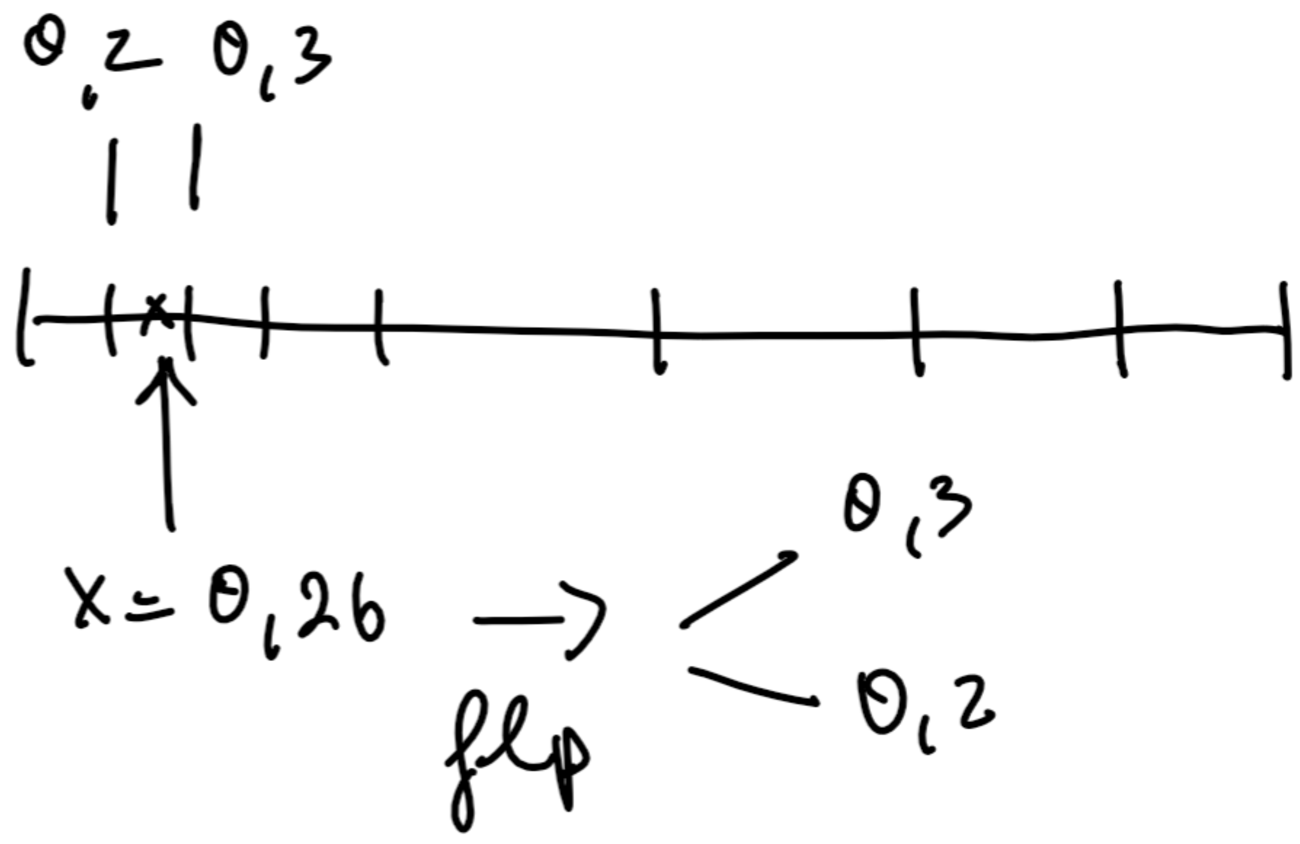
$$= \beta^{-t}\underbrace{\left[\overset{9}{d_t}\beta^0 + \overset{0,9}{d_{t+1}\beta^{-1}} + \dots + \overset{0,09 \;\dots}{d_{t+m}\beta^{-m}} + \dots\right]}_{<\beta} < \beta^{1-t}$$

$d_t \leq \beta - 1$

$< \beta^{-t} \cdot \beta$

$e(x) < \beta^{1-t}$

$\rightarrow$ ACCOUNT for ROUNDING



$e(x) < K\beta^{1-t}$

$\downarrow$

$1/2$ for ROUNDING TO NEAREST floating point

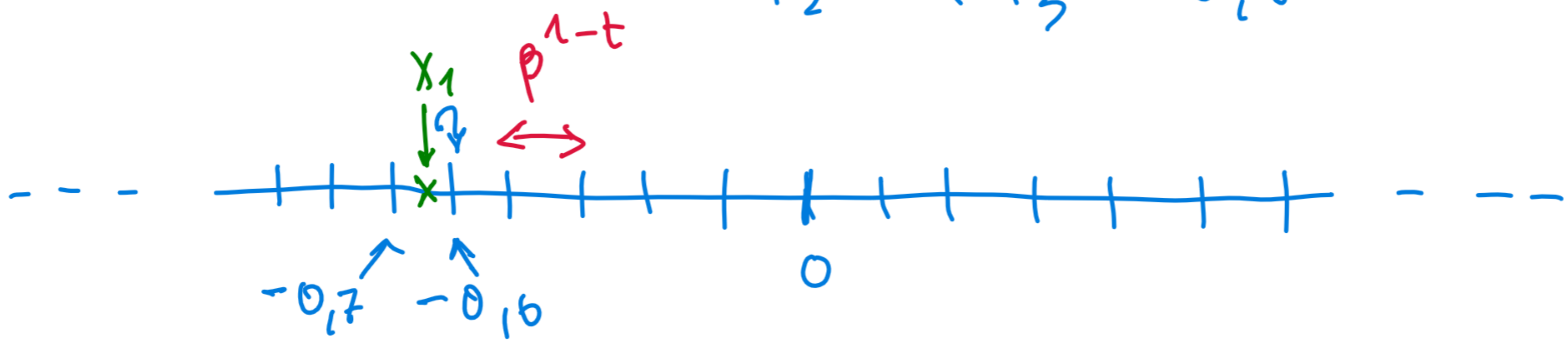
$\boxed{K\beta^{1-t} \Rightarrow \varepsilon \quad \text{MACHINE PRECISION}}$

$flp(x) = x(1 \pm \varepsilon)$ in MATLAB $\varepsilon \sim 10^{-16}$

ROUNDING TOWARD ZERO

$x_1 = -2/3 = -0,\bar{6}$

$x_2 = +2/3 = 0,\bar{6}$



ROUND T. 0 : $flp(x_1) = -0,6$ ROUND TO NEAREST : $flp(x_1) = -0,7$