ERRORS of FIXED POINT REPRESENTATION (Hp. $X(\beta=10, t=4, q=1)$)

ABSOLUTE ERROR (ROUNDOFF)

$$E(x) = x - fl_P(x)$$

error associated to rounding of given real number x

"FIXED POINT"

Example. $x_1 = \frac{10^3}{3} = 333.\bar{3}$

$x_2 = \frac{1}{3} = 0.\bar{3}$

closest member of X

$$E(x_1) = 333.\bar{3} - 333.3 = 0.0\bar{3}$$

$$E(x_2) = 0.\bar{3} - 0.3 = 0.0\bar{3}$$

CONSTANT ABSOLUTE ERROR

RELATIVE ERROR

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

$$e(x_1) = \frac{0.0\bar{3}}{333.\bar{3}} = \frac{\cancel{3.\bar{3}} \cdot 10^{-2}}{\cancel{3.\bar{3}} \cdot 10^2} = 10^{-4} \Rightarrow \text{FOUR DIGITS of ACCURACY}$$

$$e(x_2) = \frac{0.0\bar{3}}{0.3} = \frac{\cancel{3.\bar{3}} \cdot 10^{-2}}{\cancel{3.\bar{3}} \cdot 10^{-1}} = 10^{-1} \Rightarrow \text{ONE DIGIT of ACCURACY}$$

Pros.

- Simple ⇒ EASY TO implement
- FAST (OLD VIDEOGAMES, MICROCONTROLLERS)

Cons.

- VARIABLE relative error (IDEALLY: CONSTANT relative error)

FLOATING POINT REPRESENTATION

⇒ GOAL: set of numbers → WIDER RANGE; → CONSTANT rel. error } for SAME amount of memory

$\Rightarrow$ GOAL: set of numbers $\nearrow$ WIDER RANGE / $\searrow$ CONSTANT rel. error } for SAME amount of memory

NORMALIZED REPRESENTATION:

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

$\downarrow$

for any number $\neq 0$

d_k : DIGITS

$0 \le d_k \le \beta - 1$

$d_0 \neq 0$

FIRST DIGIT of mantissa cannot be zero

$1.12 = 01.12 \ldots$

FLOATING POINT SET of COMPUTER NUMBERS

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

β: BASE

t: DIGITS for MANTISSA

LOWER (L)
UPPER (U)
boundaries for P

$P \in [L, U]$

EXAMPLE: $F(\underbrace{\beta = 10, t = 3, L = 0, U = 9}_{4 \text{ DIGITS}})$

$0 \le d_k \le \beta - 1$

$$\text{MAX}(F): [d_0 \beta^0 + d_1 \beta^{-1} + d_2 \beta^{-2}] \beta^U = 9.99 \cdot 10^9$$

$9 \cdot 10^0 \quad 9 \cdot 10^{-1} \quad 9 \cdot 10^{-2} \quad 10^9$

RANGE $\sim 10^{10}$

$\updownarrow$

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

$d_0 \neq 0$

$$\text{MIN}(F) \; [1 \cdot 10^0 + 0 \cdot 10^{-1} + 0 \cdot 10^{-2}] \beta^L = 1$$

10^0

$$\frac{m \qquad \beta^p}{\quad}$$

m	β^r
9.99	10^9
9.98	10^9
⋮	⋮
1.01	10^9
1.00	10^9
9.99	10^8
9.98	10^8
⋮	⋮
1.00	10^8
⋮	⋮
9.99	10^0
9.98	10^0
⋮	
1.00	10^0
0	

$\Delta = 10^7$

$\Delta = 10^6$

$\Delta = 10^{-2}$

ABSOLUTE ERROR NOT CONSTANT

$$E(x) = x - flp(x)$$

FLOATING POIN REPRESENT. of x

RELATIVE ROUNDOFF ERROR

$x - flp(x)$

$+1$ // -1

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

$$= \frac{\sum_{k=t}^{\infty} d_k \beta^{-k}}{\sum_{k=0}^{\infty} d_k \beta^{-k}} < \frac{\beta^{1-t}}{1} = \beta^{1-t}$$

$1 \leq m \leq \beta - 1$

CONSTANT // BOUNDED

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

$1 \le m \le \beta - 1$

RECALL: $d_k \le \beta - 1$

$$\sum_{k=t}^{\infty} d_k \beta^{-k} = d_t \beta^{-t} + d_{t+1} \beta^{-(t+1)} + d_{t+2} \beta^{-(t+2)} + \dots$$

$$= \beta^{-t} \underbrace{\left(d_t + d_{t+1} \beta^{-1} + d_{t+2} \beta^{-2} + \dots \right)}_{< \beta} < \beta^{-t} \cdot \beta = \beta^{1-t}$$

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

if $t = 3$, $\beta = 10$ $\Rightarrow \beta^{1-t} = \underbrace{10^{-2}}_{\text{RELATIVE ACCURACY}} \Rightarrow$ computer numbers accurate to the SECOND DIGIT

IEEE 754 STANDARD → FOR FLOATING POINT IMPLEMENTATION

GOAL: reproducibility of results

(1) BINARY FORMAT + HIDDEN BIT TECHNIQUE
$\beta = 2$

Example $F(\beta = 2, t = 2, L = -1, U = +2)$

mantissa $0 \le d_k \le \underbrace{\beta - 1}_{=1}$; $d_0 \ne 0$ $\xRightarrow{\text{if } \beta = 2}$ d_0 ALWAYS $= 1$

without hidden bit: $t = 2$

d_0	d_1
1	0

with hidden bit: $t = 2 \rightarrow 2 + \text{"1"}$

d_0	d_1	d_2
1	0	0

$\overset{2}{1} \beta^0 + 0 \beta^{-1} + 0 \beta^{-2} = (1)_{10}$

a_0	a_1
1	0
1	1

a_0	a_1	a_2	
1	0	0	$1\beta^0 + 0\beta^{-1} + 0\beta^{-2} = (1)_{10}$
1	0	1	$1 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} = (1.25)_{10}$
1	1	0	$1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} = (1.5)_{10}$
1	1	1	$1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2} = (1.75)_{10}$

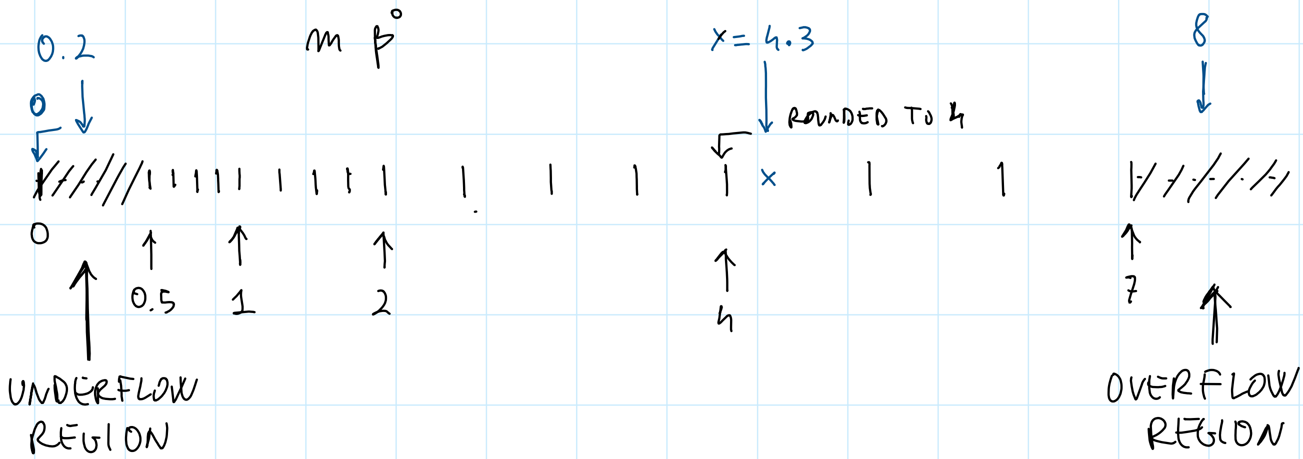
$(m)_2 \Rightarrow (m)_{10}$

SPACING MANTISSA $\Delta = (0.25)_{10}$

$L = -1, \quad U = 2$

$p=-1$ $2^{-1} = \frac{1}{2}$	$p=0$ $2^0 = 1$	$p=1$ $2^1 = 2$	$p=2$ $2^2 = 4$
0.5	1	2	4
0,625	1.25	2.5	5
0.75	1.5	3	6
0.875	1.75	3.5	7
$\Delta = 0{,}125$	0,25	0,5	1

$m\beta^0$



② PRECISION LEVELS

[FP32] SINGLE PRECISION	DOUBLE PRECISION
32 BITS	64 BITS

SIGN	1 BIT	1 BIT
EXPONENT	8 BIT $(2^8 = 256)$ $\Rightarrow p \in [-126 ; 127]$	11 BIT $(2^{11} = 2048)$ $\Rightarrow p \in [-1022 ; 1023]$
MANTISSA	23 BIT	52 BIT
ε	$\beta^{1-(t+1)} \sim \beta^{-23}$ $\sim 10^{-7}$ numbers accurate to the 7th DIGIT!	$\beta^{1-(t+1)} \sim \beta^{-52}$ $\sim 10^{-16}$ accurate to 16th digits

FP 16 $\rightarrow$ Floating Point 16 BIT "QUANTIZATION"
(HALF)