CSE330: Numerical Methods

Topic: Rounding, Rounding Error, Normalized and Denormalized System

Prepared by:
Saad Bin Sohan
BRAC University
Email: sohan.academics@gmail.com
GitHub: https://github.com/saad-bin-sohan

IEEE standard

$e_{min} = 0$

$e_{max} = 2047$

$$(1.d_1 d_2 \cdots d_{52}) \times 2^{e-1023}$$

denormalized

$$(0.1 d_1 d_2 \cdots d_{52}) \times 2^{e-1022}$$

$$[-1022, 1025]$$

power 1025 হলে সামনে যাই থাকুক, কম্পিউটার তাকে infinity হিসাবে save করে

$$\cdots \times 2^{1025} \longrightarrow \infty \text{ (infinity)}$$

$$\cdots \times 2^{-1022} \longrightarrow \text{zero}$$

this is how computer represents zero

so,

highest possible positive number (except infinity), $= (0.1\underbrace{11\ldots111}_{52\text{ one}}) \times 2^{1024}$

lowest possible number (except zero),

$$= (0.1\underbrace{00\ldots00}_{52\text{ zero}}) \times 2^{-1021}$$

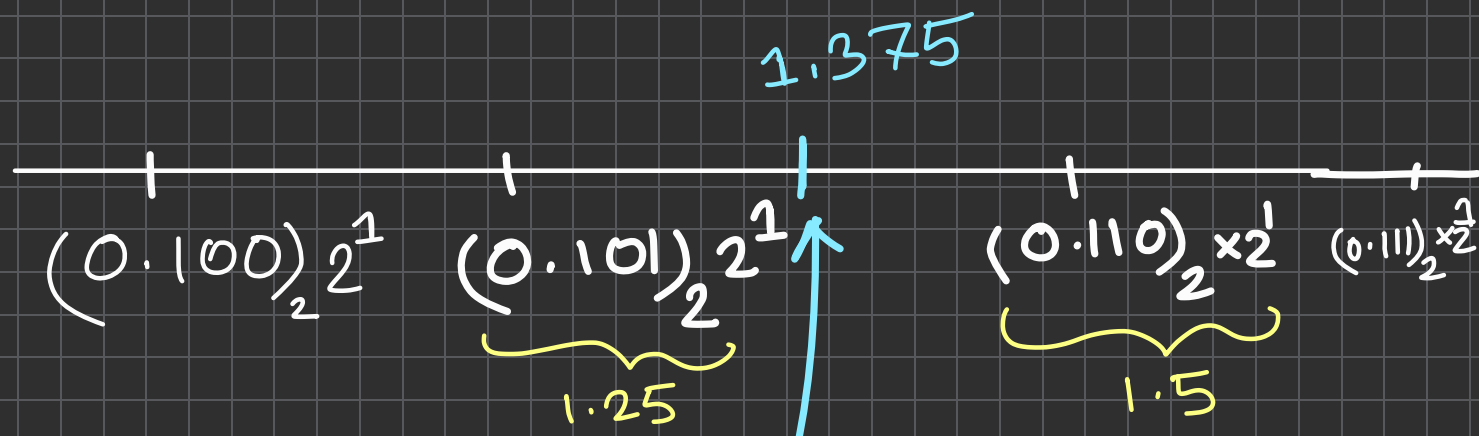
[same things go for negative numbers]

Rounding

let, $\beta = 2$ $\quad e_{min} = -1$

$m = 3$ $\quad e_{max} = 2$

এই constraint এ



$(0.1011)_2 \times 2^1$ exact middle point.

so how to round up then?

[given computer এরে store করতে পারবে না

কারণ $m = 3$, কিন্তু mantissa এ 4 bits.

Rule: if the number is in

so কমিটিং নাম্বার টাকে round up করে represent করতে।

the middle, then round it off to the nearest (binary) even number

so, actual value, $x = (0.1011)_2 \times 2^1$

round value, $fl(x) = (0.110) \times 2^1$

Rounding Error

$$\text{Error} = \left| fl(x) - x \right| \rightarrow \text{modulus}$$

$$\underbrace{\text{relative rounding error}}_{\text{scale-invariant error}(\delta)}, \delta = \frac{|fl(x) - x|}{|x|}$$

max value of δ = Machine Epsilon (ϵ_m)

[3ঢ়া convention-এ ϵ_m কে 3ঢ়া value বলা হয়]

Que Let, $\beta = 2$ $e_{min} = -1$

$m = 3$ $e_{max} = 2$

convention-1 ও ϵ_m এর value কত?

soln: ϵ_m হচ্ছে rel. rounding এর max value.

so, $|fl(x) - x|$ has to be maximum

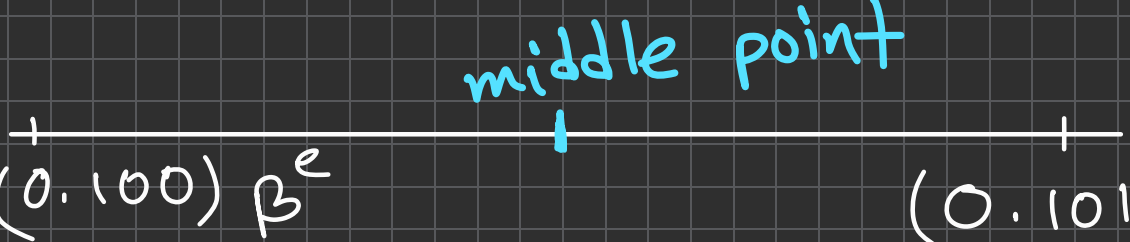
and $|x|$ has to be minimum.

we will find

the $|fl(x) - x|$ to be maximum, when

actual given number টি দুইটা 3 bit এর representable number এর middle এ হবে।

so, for any e,



Maximum difference,

$$= \frac{1}{2}\left[(0.101)\beta^e - (0.100)\beta^e\right]$$

$$= \frac{1}{2}(0.001)\beta^e$$

$$= \frac{1}{2} \times 2^{-3}\beta^e$$

$$= \frac{1}{2} \times \beta^{-3}\beta^e$$

$$= \frac{1}{2}\beta^{-m}\beta^e$$

min of $x = (0.100)\beta^e$

$$= 2^{-1}\beta^e$$

$$= \beta^{-1} \beta^{e}$$

and now,

$$\epsilon_m = \frac{|fl(x) - x|}{|x|}$$

$$\Rightarrow \epsilon_m = \frac{\frac{1}{2} \beta^{-m} \beta^{e}}{\beta^{-1} \beta^{e}}$$

$$\Rightarrow \boxed{\epsilon_m = \frac{1}{2} \beta^{1-m}}$$

for convention-1 system

[N.B: ϵ_m ক always decimal এ floating point এ থাকতে হবে else - 1 marks gone.

যেমন: 0.625 ✓ correct

$\frac{1}{2} \times 2^{-3}$ ✗ wrong

for both normalized and denormalised system,

$$\epsilon_m = \frac{1}{2} \beta^{-m}$$

arithmetic operation

Que let,

$\beta = 2$

$m = 3$

$e_{min} = -1$

$e_{max} = 2$

$x = \frac{5}{8}$

$y = \frac{7}{8}$

$x * y = ?$

Soln:

$x * y = fl(x) * fl(y)$

$x = \frac{5}{8} = (0.101)_2 \times 2^0 = fl(x)$

$y = \frac{7}{8} = (0.111)_2 \times 2^0 = fl(y)$

m limit-এর মধ্যে আছে x, y এর

rounding not needed

so,

$$x * y = fl(x) * fl(y)$$

$$= (0.101)_2 \times 2^0 \times (0.111) \times 2^0$$

$$= \underbrace{(0.100011)_2}_{} \times 2^0$$

0.100 | 011

m = 3

↓

division

common sense

↓

rounding ↓

left number ↓

কারণ zero, because of zero

so, $fl(xy) = \underbrace{(0.100)_2 \times 2^0}$

thats how the computer will save it, as a rounded value, instead of actual number

(NB:

if $(m+1)$th digit is zero, then round it to previous number

elif thats 1, round it up-to the next number)

[δ ক্ষুদ্র হলে $\Rightarrow \left|\frac{fl(xy)-xy}{xy}\right|$]