Tuto 1 - Corrigé

Exercise 2 :

(1) Represent 3.14 and -8.625 in IEEE 754 F.P. SP.

$3.14 = 3 + 14$.

$3 = (11)_2$, $0.14 = (0.001)_2 \Rightarrow (3.14) = (11.001)_2$

Normalization = 1.1001×2^1.

$e = 1 \Rightarrow 1 + 127 = 128 = (10000000)_2$.

$S = 0$ (Positive).

So $3.14 =$ | 0 | 10000000 | 100100.....0 |

8 bits | 23 bits

$-8.625 = -8 + 625$, with $S = 1$ (Negative).

$8 = (1000)_2$, $0.625 = (0.101)_2 \Rightarrow 8.625 = (1000.101)_2$

Normalization: 1.000101×2^3.

$e = 3 \Rightarrow 3 + 127 = 130 = (10000010)_2$.

So $-8.625 =$ | 1 | 10000010 | 00010100....0 |

8 bits | 23 bits

(2) 35.5 in IEEE 754 D.P.

$35 = (100011)_2$, $0.5 = (0.1)_2 \Rightarrow 35.5 = (100011.1)_2$

Normalization : 1.000111×2^5. , $S = 0$

$e = 5 \Rightarrow$ ~~5 + 1023 = 1028 = (10000000100)$_2$~~

So $35.5 =$ | 0 | 10000000100 | 00011100...0 |

11 bits | 52 bits

(3) floating-point representation of A:

$A = \underbrace{0}_{s}\ \underbrace{10000100}_{e}\ \underbrace{00011001100000000000000}_{M}$ (8 bits)

$s = 0$: positive

$10000100 = (132)_{10} \Rightarrow e = 132 - 127 = 5$.

Mantissa $= 1.01100110$

so $A = 1.01100110 \times 2^5 = 0.101100110 \times 2^6 = (44.75)_{10}$.

(4) a. $A = 1.11 \times 2^2$ $\quad B = 1.001 \times 2^1 = 0.1001 \times 2^2$

A+B: we add the mantissas:

$$\begin{array}{r} 1.11 \\ +\ 0.1001 \\ \hline 10.0101 \end{array}$$

$A+B = 10.0101 \times 2^2 = 1.00101 \times 2^3$.

A+B | 0 | 10000010 | 00101000...0 |

8 bits, 23 bits

b. $A = 1.11 \times 2^2$, $B = 1.0011 \times 2^3$.

we multiply mantissas, and we add exponents.

$$\begin{array}{r} 1.0011 \\ \times\ 1.11 \\ \hline 10011 \\ 10011\_ \\ +\ 10011\_\_ \\ \hline 10.000101 \end{array}$$

$A \times B = 10.000101 \times 2^5 = 1.0000101 \times 2^6$

A×B = | 0 | 10000101 | 00001010000...0 |

8 bits, 23 bits.

5) Interval of representable Numbers using IEEE 754 S.P.

we know that $-128 \leq e \leq 127$ and $1 \leq 1.M < 2$ ...(1)

So $2^{-128} \leq 2^e \leq 2^{127}$ ...(2) and (1)x(2): $2^{-128} \leq 1.M \times 2^e < 2^{128}$

Exercise 3.

1) (a) 67.1 : 2 , (b) 0.00173 : 3, (c) 2.30×10^{-9} : 3

(d) 3.0054 : 5 (e) 0.0004000 : 4.

2) (a) 62.34 using 2 S.D : 62.

(b) 0.006738 using 3 S.D : 6.74×10^{-3}.

(c) 238.62 using 3 S.D : 239 (with rounding).

(d) 6.345 using 2 S.D : 6.3.

3) $\varepsilon_{machine} = 10^{-5}$, $-20 < e < 20$.

(a) $1 + 10^{26} = (0.1 + 0.000001) \times 10^{7} = 0.1 \times 10^{7} = 10^{6}$ (this is (c))

(b) $1 + 10^{4} = (0.1 + 0.00001) \times 10^{5} = 0.10001 \times 10^{5} = 10001$.

(c) $1 + 10^{-6} = 1 + 0.000001 = 1$. (this is (a))

(d) $10^{3} + 10^{6} = 1001000$

(e) $\frac{10^{6}}{10^{-12}} = 10^{6+12} = 10^{18}$ $(18 < 20)$.

(f) $10^{-9} \times 10^{-16} = 10^{-9-16} = 10^{-25} = 0$ $(-25 < -20)$.

(g) $10^{28} + 10^{4} = Inf$ $(28 > 20$, $10^{28} = Inf)$.

(h) $\frac{10^{5}}{10^{-21}} = 10^{5+21} = 10^{26} = Inf$.

(i) $\sqrt{10^{10} - 10} \Rightarrow 10^{10} - 10 = 10^{10}$, $\sqrt{10^{10}} = 10^{5}$.

(j) $\ln(10^{-25})$, $-25 < -20$ the $10^{-25} = 0$, so $\ln(0) = Nan$.

Exercise 4:

(1) Absolute True error and Percent Relative error when approximating π with (we take 5 sign. digits in the results),

(a) 3 : $e_A = |\pi - 3| = 0.14159$, $e_R\% = \frac{0.14159}{\pi} \times 100 = 4.5\%$

(b) 3.14 : $e_A = |\pi - 3.14| = 0.0015926$, $e_R\% = \frac{e_A}{\pi} \times 100 = 0.05\%$

(c) $\frac{22}{7}$: $e_A = |\pi - \frac{22}{7}| = 0.0012645$, $e_R\% = 0.043\%$.

We observe that $\frac{22}{7}$ is the best approximation because it gives the smallest error.

(2) $A = 4\pi r^2$. The impact means $e_A = \Delta A$

(a) Five approximations used to calculate A in a machine:

(1) the earth is supposed Spherical, Impact: Difficult to find
(2) the value of r is based on measures, so it's approximated
Impact: $\Delta A = 2\pi 2r\Delta r$.
(3) the value of π is not exact, Impact: $\Delta A = 4r^2\Delta\pi$
(4) the Numbers $4, \pi, r$ are stored with errors (it depends on the machine properties). Impact: Difficult to find.
(5) floating-point Arithmetic errors (Rounding and chopping).
Impact: Difficult to find.

(b) the radius is known with precision 2% means that the relative percent error is 2% and this means precise relative error $e_{r\%} = \frac{e_r}{r} \times 100 = 2$.

we find $e_A = \Delta r$ (the relative error in the radius).

$$\Delta r = \frac{2r}{100}$$

and we want to find the relative percent error in $A = \frac{\Delta A}{A} \times 100$

we know that:

$$\Delta A = 4\pi 2r\Delta r \Rightarrow 4\pi \cdot 2r \cdot \frac{2r}{100} = \frac{4\pi r^2}{25}$$

so $\frac{\Delta A}{A} \times 100 = \frac{4\pi r^2}{25} \times \frac{1}{4\pi r^2} \times 100 = 4\%$.

Exercise(s)

(1) $122 \times (333 \times 695) = (0.122 \times 10^3 \times (0.333 \times 10^3 + 0.695 \times 10^3))$

$= (0.122 \times 10^3) \times (1.028 \times 10^3)$ → 1.03×10^3

$= 0.122 \times 10^3 \times 0.103 \times 10^4$

$= 0.012566 \times 10^7 = \boxed{0.126 \times 10^6}$

$(122 \times 333) + (122 \times 695) = (0.122 \times 10^3 \times 0.333 \times 10^3) + (0.122 \times 10^3 \times 0.695 \times 10^3$

$= (0.040626 \times 10^6) + (0.08479 \times 10^6)$

$= 0.0406 \times 10^6 + 0.0848 \times 10^6)$

$= (0.406 \times 10^5) + 0.848 \times 10^5)$

$= 1.254 \times 10^5$

$= 1.254 \times 10^5 = \boxed{0.125 \times 10^6}$

we observe a slight difference in the result.

Ex5: (2)

(a). $\cos(1.57079) \simeq 0.000000632679 = x_{ex}$.

$\cos(1.57078) \simeq 0.000016326779 = x_{app}$.

$\text{error}\% = \left|\frac{x_{app} - x_{ex}}{x_{ex}}\right| \times 100 = 100\%$.

(2) $e_{abs} = \cos(\overbrace{x+h}^{u}) - \cos(\overbrace{x}^{v}) = -2\sin\left(\frac{x+v}{2}\right)\sin\left(\frac{x-v}{2}\right)$.

$= -2\sin\left(\frac{2x+h}{2}\right)\sin\left(\frac{h}{2}\right) = -2\sin\left(x+\frac{h}{2}\right)\sin\left(\frac{h}{2}\right)$

$x \simeq \frac{\pi}{2} \rightarrow \sin\left(x+\frac{h}{2}\right) \simeq 1$, $h \ll 1 \rightarrow \sin\left(\frac{h}{2}\right) \simeq \frac{h}{2}$.

$e_{abs} \simeq -2 \cdot 1 \cdot \frac{h}{2} = -h$.

(3) $e_{rel} = \frac{e_{abs}}{x_{ex}} \simeq \frac{-h}{\cos x} \simeq \frac{-h\sin x}{\cos x}$ ($\sin x \simeq 1$) $= -h\tan x \gg 1 \simeq \infty$.

We conclude that the relative error is large and at the same time the absolute error is small, so we conclude that rel error is large as when the ex value is small.

relative error has two aims.

relative error: How accurate the measurment is compared to the actual value (or exact value).

Or the measur of precision.

Absolute: The distance between ex and app.

Conclusion: Relative error only makes sense where the exact value not zero or close to zero.