3X03 Practice Problem Set 1

Neelkant Teeluckdharry

September 2024

1 Machine Precision and Unit Roundoff

1.1 Problem I

By definition

$$x = \pm(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \frac{d_3}{\beta^3} + ... + \frac{d_{t-1}}{\beta^{t-1}}) \times \beta^e$$

where $0 \leq d_i \leq \beta - 1$ and $e \in [L, U]$. The expression for machine epsilon can be rewritten as:

$$e_{mach} = \min_{x \in FP} |\sum_{i=0}^{t-1}(d_i \beta^{e-i}) - 1|$$

Since $x > 1.0 \Rightarrow$ minimum is given when $d_0 = 1$ and $d_1 = d_2 = ... = d_{t-2} = 0$, $d_{t-1} = 1$ and $e = 0$.

$$\boxed{e_{mach} = \frac{1}{\beta^{t-1}}}$$

1.2 Problem II

By definition

$$fl(x) = x(1+\epsilon)$$

where $|\epsilon| \leq \frac{\epsilon_{mach}}{2}$

$$\frac{fl(x) - x}{x} = \epsilon$$

$$\left|\frac{fl(x) - x}{x}\right| \leq |\epsilon| = \frac{\epsilon_{mach}}{2}$$

As required. ■

1.3 Problem III

$$
\begin{aligned}
fl(fl(x) \circ fl(y)) &= fl(x(1+\delta_x) \circ y(1+\delta_y)) \\
&= (1+\delta_x)(1+\delta_y)(1+\delta_{xy})fl(x \circ y) \\
&\neq fl(x \circ y)
\end{aligned}
$$

As desired. ■

1.4 Problem IV

The error ϵ is given by:

$$|\epsilon| = \left| \frac{fl(fl(x)fl(y)) - xy}{xy} \right| = \left| \frac{fl(1+\delta_x)(1+\delta_y)xy) - xy}{xy} \right|$$

where $|\delta_{xy}| \leq u$ and $|\delta_x| \leq u$ and $|\delta_y| \leq u$.

$$\left| \frac{(1+\delta_{xy})(1+\delta_x+\delta_y)xy - xy}{xy} \right|$$

$$|\delta_x + \delta_y + \delta_{xy}| \leq \boxed{3u}$$

by the triangle inequality.

2 Denormalized Numbers

2.1 Problem V

Part A: The smallest normalized number is given by:

$$(1.0...0)_f \times 2^{-126} \approx 1.2 \times 10^{-38}$$

Part B: The largest denormalized number is given by:

$$(0.11...1)_f \times 2^{-126} \approx 2^{-126} = 1.17 \times 10^{-38}$$

2.2 Problem VI

Part A: The largest positive denormalized number is given by:

$$(0.11..1)_f \times 2^{-127} \approx 2^{-127} = 5.87 \times 10^{-39}$$

Part B: There is a larger gap between the smallest normalized number and largest denormalized number which is problematic as there would be an abrupt change between normalized and denormalized representable numbers.

3 Biased Exponent

3.1 Problem VII

Part A: $2 = 00000010$
Part B: $-2 = 11111101$

3.2 Problem VIII-a

$x = 01000001000000000000000000000000$ $y = 00111110100000000000000000000000$

3.3 Problem VIII-b

$x = 00000001000000000000000000000000$ $y = 01111110100000000000000000000000$

4 Cancellations

4.1 Problem X-c

This can be shown directly

$$
\begin{aligned}
x_1 \cdot x_2 &= \frac{b^2 - (b^2 - 4ac)}{4a^2} \\
&= \frac{c}{a}
\end{aligned}
$$

4.2 Problem X-d

If $b \geq 0$, then determine the negative root. Then $\frac{c}{a \cdot x_1}$ to find the second root. Otherwise $b < 0$, then we should determine the positive root, and use the identity to solve for the other root. This eliminates any potential catastrophic cancellation issues.

4.3 Problem XI-a

The truncated Taylor series for e^x can be used by setting $x = -x$ for some real number. It is more accurate as it wouldn't be an alternating series.