Exercise 3 – Solutions

1. (a) $x \approx 0$ is problematic, because the numerator is then a difference of 1 and $(1-x)^3 \approx 1$.

Better multiply out and simplify:

$$\frac{1-(1-x)^3}{x} = \frac{1-(1-3x+3x^2-x^3)}{x}$$

$$= \frac{3x-3x^2+x^3}{x} = 3-3x+x^2$$

$$= 3+x(-3+x) \qquad \text{(Horner's scheme)}$$

<u>Note</u>: $x \approx 1$ is <u>not</u> problematic. Even though $1-x$, in this case, is computed with a big relative error, its absolute error is small, so that $1-(1-x)^3$ has again a small absolute <u>and</u> relative error.

(b) Again, $x \approx 0$ is problematic. Use trick from class:

$$\frac{1-\sqrt{1-x^2}}{x} = \frac{1-\sqrt{1-x^2}}{x} \; \frac{1+\sqrt{1-x^2}}{1+\sqrt{1-x^2}}$$

$$= \frac{1-(1-x^2)}{x\,(1+\sqrt{1-x^2})} = \frac{x}{1+\sqrt{1-x^2}}$$

(c) Problematic values are $\sec x \approx 1$, i.e. $x \approx 2\pi n$, $n \in \mathbb{Z}$.

$$\frac{1-\sec x}{\tan^2 x} = \frac{1-\sec x}{\tan^2 x}\,\frac{1+\sec x}{1+\sec x} = \frac{1-\sec^2 x}{\tan^2 x\,(1+\sec x)} = \frac{-\tan^2 x}{\tan^2 x\,(1+\sec x)}$$

$$= -\frac{1}{1+\sec x}$$ (Note: last expression is bad when $\sec x \approx -1$)

2. No. Try, for example, on the Python console:

$$0.3 \ominus (0.2 \oplus 0.1) \qquad \text{vs.} \qquad 0.3 \odot 0.2 \oplus 0.3 \odot 0.1$$

3. $a = 1.0101 \cdot 2^5$, $b = 1.1101 \cdot 2^3$

(i) computation of $a \oplus b$:

$$\begin{array}{r} 101010 \\ +\ \ 1110.1 \\ \hline (\underbrace{11100}_{\text{size of significant}}0.1)_2 \end{array} = (56.5)_{10}$$

"Round to nearest" and normalization gives $1.1100 \cdot 2^5$

$\Rightarrow$ absolute error: $(1)_2 \cdot 2^{-1} = (0.5)_{10}$

relative error: $\frac{0.5}{56.5} \approx 0.0088\ldots \approx 1\%$

(ii) computation of $a \ominus b$:

$$\begin{array}{r} 101010 \\ -\ \ 1110.1 \\ \hline (011011.1)_2 \end{array} = (27.5)_{10}$$

"Round to even" and normalization gives $(1.1010)_2 \cdot 2^4$

$\Rightarrow$ absolute error: $(1)_2\, 2^{-1} = (0.5)_{10}$

relative error: $\frac{0.5}{27.5} \approx 0.018\ldots \approx 2\%$

4. "Round to nearest" in IEEE-floating point implies that

$$fl(x) \oslash fl(y) = \frac{fl(x)}{fl(y)}(1+\delta_1) \qquad \text{with } |\delta_1| \le \frac{\varepsilon}{2}$$

$$= \frac{x(1+\delta_2)}{y(1+\delta_3)}(1+\delta_1) \qquad \text{with } |\delta_2|, |\delta_3| \le \frac{\varepsilon}{2}$$

$$\approx \frac{x}{y}(1+\delta_2)(1+\delta_1)(1-\delta_3) \qquad \left(\text{linear approximation of } \frac{1}{1+\delta_3}\right)$$

$$\approx \frac{x}{y}(1+\delta_1+\delta_2-\delta_3) \qquad (\text{drop all superlinear terms})$$

$$= \frac{x}{y}(1+\delta) \qquad \text{with } |\delta| \le 3\frac{\varepsilon}{2}$$

$\Rightarrow$ Propagation of relative error works as for floating point multiplication.

It is moderate.

5. $\underbrace{1}_{\text{sign bit}}\ \underbrace{10001010}_{\text{stored exponent}}\ \underbrace{1100001 0 \ldots\ldots 0}_{\text{stored significant}}$

Sign bit 1 means the number is negative.

Exponent e = stored exponent − bias , in 32-bit floating point, bias = 127.

Since $(10001010)_2 = (138)_{10}$, $e = 138 - 127 = 11$.

Number is normal (stored exponent $\neq (FF)_{16}$), so significant is

$$1.1100001 = 1 + 2^{-1} + 2^{-2} + 2^{-7}$$

Altogether: number is

$$-(1+2^{-1}+2^{-2}+2^{-7}) \cdot 2^{11} = -(2^{11}+2^{10}+2^{9}+2^{4})$$

$$= -(2048 + 1024 + 512 + 16)$$

$$= -3600$$

6. $(500.25)_{10} = \cancel{+}(111110100.01)_2 = \cancel{+}(1.\underbrace{1111010001}_{\substack{\text{significant to store}\\ \text{(number is normal)}}})_2 \cdot 2^8$, $e = 8$, bias = 127

$\Rightarrow e_{store} = (135)_{10} = 10000111$

Bit pattern:

0
1 bit

1	0	0	0	0	1	1	1
8 bits							

1	1	1	1	0	1	0	0	0	1	0	0	0	- - -
23 bits													