Math 340 • Fall 2026 • Homework #2 • Victor Matveev

Problem 1 (60pts) Find the range of positive values of x for which the following expressions computed in double precision have no correct significant digits, i.e. relative error reaches 100% (assume that $x < 1$). You can easily double-check your answers using MATLAB, but do provide an explanation for each:

$$(a)\ \ f(x)=\frac{\sqrt{100+x^2}-10}{x^2} \qquad (b)\ \ g(x)=\frac{x^8}{1-x^8}-\frac{x^8}{1+x^8} \qquad (c)\ \ h(x)=\frac{x^3}{x-\sin x}$$

NOTE: The only information needed is that the roundoff $A+a \Rightarrow A$ happens when $\left|\frac{a}{A}\right| \le \underbrace{10^{-16}}_{\approx 2^{-53}}$

The **true answer is irrelevant**: it only matters that this answer is **lost completely**

$$(a)\ \ f(x)=\frac{\sqrt{100+x^2}-10}{x^2}$$

- Round off to $\frac{10-10}{x^2}=0$ (100% error) when $100+x^2 \Rightarrow 100$ which happens when $\frac{x^2}{100} \le 10^{-16} \ \Rightarrow\ x^2 \le 10^{-14}$
- Round-off also occurs if $\underbrace{\sqrt{100+x^2}}_{\approx 10+\frac{x^2}{20}} \Rightarrow 10$ which happens at **larger** x: $\frac{x^2/20}{10}=\frac{x^2}{200} \le 10^{-16} \ \Rightarrow\ x^2 \le 2\cdot 10^{-14}$
- Complete loss of significance starts at larger of these two values, so $\boxed{0<|x| \le \sqrt{2}\cdot 10^{-7} \approx 1.4\cdot 10^{-7}}$

$$(b)\ \ g(x)=\frac{x^8}{1-x^8}-\frac{x^8}{1+x^8}$$

- Round off to $\frac{x^8}{1}-\frac{x^8}{1}=0$ (100% error) when $1 \pm x^8 \Rightarrow 1$ which we expect when $x^8 \le 10^{-16}$
- Therefore, complete loss of significance expected if $\boxed{0<|x| \le 10^{-2}}$

NOTE: Roundoff $1-x^8 \Rightarrow x$ actually happens for smaller $|x| \le 2^{-54}$, as compared to $1+x^8 \Rightarrow x$, but the round-off in the 2nd term still leads to non-zero answer with **no** correct digits at $x=0.01$

$$(c)\ \ h(x)=\frac{x^3}{x-\sin x}=\frac{x^3}{x-\left(x-\frac{x^3}{3!}+\frac{x^5}{5!}+O(x^7)\right)}=\underbrace{\frac{x^2}{1-\left(1-\frac{x^2}{3!}+\frac{x^4}{5!}+O(x^6)\right)}}_{\text{Not needed for correct conclusion!}}=6+\frac{3x^2}{5}+O(x^4)$$

- Round off to $\frac{x^3}{x-x}=\frac{x^3}{0}=$ **Inf** (all digits lost) when $x-\frac{x^3}{3!} \Rightarrow x$ which happens when $\underbrace{\frac{x^3/3!}{x}}_{x^2/6} \le 10^{-16}$
- Therefore, complete loss of significance if $\boxed{0<|x| \le \sqrt{6}\cdot 10^{-8} \approx 2.45\cdot 10^{-8}}$

Problem 2 (20pts) Modify functions ***f*(*x*)** and ***g*(*x*)** above so that they don't suffer from loss of significance for small *x*.

$$(a)\quad f(x)=\frac{\sqrt{100+x^2}-10}{x^2}\cdot\frac{\sqrt{100+x^2}+10}{\sqrt{100+x^2}+10}=\frac{\left(\cancel{100}+x^2\right)-\cancel{100}}{x^2\left(\sqrt{100+x^2}+10\right)}=\boxed{\frac{1}{\sqrt{100+x^2}+10}}$$

$$(b)\quad g(x)=\frac{x^8}{1-x^8}-\frac{x^8}{1+x^8}=\frac{x^8\left(1+x^8-1+x^8\right)}{\left(1-x^8\right)\left(1+x^8\right)}=\boxed{\frac{2x^{16}}{1-x^{16}}}$$

Note: at $x\approx 0.1$, x^{16} is rounded off to zero when subtracted from 1 in denominator, but this will not impact the result much, since the numerator provides all signifianct digits and is computed accurately (you may use MATLAB to verify)

Problem 3 (20pts) Convert this decimal number to a binary number, keeping only the first 10 bits after the binary point (i.e. in the fractional part): $\mathbf{119.7_{10}}$.

Let's convert the integer part first
(we don't need an algorithm since the number is so simple: just show expansion in powers of 2):

$$119 = \underset{2^6}{64} + \underset{2^5}{32} + \underset{2^4}{16} + \underset{2^2}{4} + \underset{2^1}{2} + \underset{2^0}{1} = 1110111_2$$

Now, the 10 bits in the fractional part, using our simple algorithm:

$$\begin{pmatrix} 0.7*2=1+0.4\\ 0.4*2=0+0.8\\ 0.8*2=1+0.6\\ 0.6*2=1+0.2\\ 0.2*2=0+0.4\\ 0.4*2=0+0.8\\ 0.8*2=1+0.6\\ 0.6*2=1+0.2\\ 0.2*2=0+0.4\\ 0.4*2=0+0.8 \end{pmatrix}$$

Thus, we obtain: $\text{Binary}(119.7) = 1110111.\overline{10110}$

$$= 1.1101111\overline{0110}\cdot 2^6$$

OPTIONAL: therefore, double-precision representation would be:

- Sign bit: $\mathbf{S} = 0\ (\text{1 bit only})$
- Significand: $\mathbf{F} = \underbrace{11011110110011001100110011001100110011001100110011001100}_{52\text{ bits}}$
- Exponent field: $6=\mathbf{E}-1023 \Rightarrow \mathbf{E} = 6+1023=1029=\underbrace{1024+4+1}_{2^{10}\ 2^2\ 2^0} = \underbrace{10000000101}_{11\text{ bits}}$