

# Numerical Analysis MATH50003 (2023–24) Problem Sheet 4

**Problem 1** Suppose  $x = 1.25$  and consider 16-bit floating point arithmetic ( $F_{16}$ ). What is the error in approximating  $x$  by the nearest float point number  $\text{fl}(x)$ ? What is the error in approximating  $2x$ ,  $x/2$ ,  $x + 2$  and  $x - 2$  by  $2 \otimes x$ ,  $x \oslash 2$ ,  $x \oplus 2$  and  $x \ominus 2$ ?

**Problem 2** Show that  $1/5 = 2^{-3}(1.1001100110011\dots)_2$ . What are the exact bits for  $1 \oslash 5$ ,  $1 \oslash 5 \oplus 1$  computed using half-precision arithmetic ( $F_{16} := F_{15,5,10}$ ) (using default rounding)?

**Problem 3** Prove the following bounds on the *absolute error* of a floating point calculation in idealised floating-point arithmetic  $F_{\infty,S}$  (i.e., you may assume all operations involve normal floating point numbers):

$$\begin{aligned}(\text{fl}(1.1) \otimes \text{fl}(1.2)) \oplus \text{fl}(1.3) &= 2.62 + \varepsilon_1 \\ (\text{fl}(1.1) \ominus 1) \oslash \text{fl}(0.1) &= 1 + \varepsilon_2\end{aligned}$$

such that  $|\varepsilon_1| \leq 11\epsilon_m$  and  $|\varepsilon_2| \leq 40\epsilon_m$ , where  $\epsilon_m$  is machine epsilon.

**Problem 4** Let  $x \in [0, 1] \cap F_{\infty,S}$ . Assume that  $f^{\text{FP}} : F_{\infty,S} \rightarrow F_{\infty,S}$  satisfies  $f^{\text{FP}}(x) = f(x) + \delta_x$  where  $|\delta_x| \leq c\epsilon_m$  for all  $x \in [0, 1]$ . Show that

$$\frac{f^{\text{FP}}(x+h) \ominus f^{\text{FP}}(x-h)}{2h} = f'(x) + \varepsilon$$

where absolute error is bounded by

$$|\varepsilon| \leq \frac{|f'(x)|}{2}\epsilon_m + \frac{M}{3}h^2 + \frac{2c\epsilon_m}{h},$$

where we assume that  $h = 2^{-n}$  for  $n \leq S$ .

**Problem 5** For intervals  $X = [a, b]$  and  $Y = [c, d]$  satisfying  $0 < a < b$  and  $0 < c < d$ , and  $n > 0$  prove that

$$\begin{aligned}X/n &= [a/n, b/n] \\ XY &= [ac, bd]\end{aligned}$$

Generalise (without proof) these formulæ to the case  $n < 0$  and to where there are no restrictions on positivity of  $a, b, c, d$ . You may use the min or max functions.

**Problem 6(a)** Compute the following floating point interval arithmetic expression assuming half-precision  $F_{16}$  arithmetic:

$$[1, 1] \ominus ([1, 1] \oslash 6)$$

Hint: it might help to write  $1 = (0.1111\dots)_2$  when doing subtraction.

**Problem 6(b)** Writing

$$\sin x = \sum_{k=0}^n \frac{(-1)^k x^{2k+1}}{(2k+1)!} + \delta_{x,2n+1}$$

Prove the bound  $|\delta_{x,2n+1}| \leq 1/(2n+3)!$ , assuming  $x \in [0, 1]$ .

**Problem 6(c)** Combine the previous parts to prove that:

$$\sin 1 \in [0.11010011000, 0.11010111101] = [0.82421875, 0.84228515625]$$

You may use without proof that  $1/120 = 2^{-7}(1.000100010001\dots)_2$ .