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 $\mathrm{fl}(x)$? 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 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):

$$(1.1 \otimes 1.2) \oplus 1.3 = 2.62 + \varepsilon_1$$
$$(1.1 \ominus 1) \oslash 0.1 = 1 + \varepsilon_2$$

such that $|\varepsilon_1| \leq 11\epsilon_{\mathrm{m}}$ and $|\varepsilon_2| \leq 40\epsilon_{\mathrm{m}}$, where ϵ_{m} is machine epsilon.

Problem 4 Assume that

$$f^{\mathrm{FP}}(x) = f(x) + \delta_x$$

where $|\delta_x| \leq c\epsilon_{\mathrm{m}}$ for all x. Using idealised floating point arithmetic $F_{\infty,S}$, for

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

show the absolute error is bounded by

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

where we assume $x \in [0,1]$, $h = 2^{-n}$ for $n \leq S$ so that $x \oplus h = x + h$ and $x \ominus h = x - h$.

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

$$X/n = [a/n, b/n]$$
$$XY = [ac, bd]$$

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

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 $1 = (0.1111\ldots)_2$ 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 $1/120 = 2^{-7}(1.000100010001\ldots)_2$.