Numerical Analysis MATH50003 (2023–24) Problem Sheet 3

Problem 1 What is π to 5 binary places? Hint: recall that $\pi \approx 3.14$.

Problem 2 What are the single precision $F_{32} = F_{127,8,23}$ floating point representations for the following:

$$2, \quad 31, \quad 32, \quad 23/4, \quad (23/4) \times 2^{100}$$

Problem 3 Let $m(y) = \min\{x \in F_{32} : x > y\}$ be the smallest single precision number greater than y. What is $m(2) - 2$ and $m(1024) - 1024$?

Problem 4 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 5 Show that $1/5 = 2^{-3}(1.1001100110011\ldots)_2$. 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 6 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):

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

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

Problem 7 Let $x \in [0,1] \cap F_{\infty,S}$. Assume that $f^{\mathrm{FP}} : F_{\infty,S} \to F_{\infty,S}$ satisfies $f^{\mathrm{FP}}(x) = f(x) + \delta_x$ where $|\delta_x| \leq c\epsilon_{\mathrm{m}}$ for all $x \in [0,1]$. Show that

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

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

Problem 8(a) Suppose $|\epsilon_k| \leq \epsilon$ and $n\epsilon < 1$. Show that $\prod_{k=1}^n (1 + \epsilon_k) = 1 + \theta_n$ for some constant θ_n satisfying

$$|\theta_n| \leq \underbrace{\frac{n\epsilon}{1 - n\epsilon}}_{E_{n,\epsilon}}.$$

Problem 8(b) Show if $x_1, \ldots, x_n \in F_{\infty,S}$ then

$$x_1 \otimes \cdots \otimes x_n = x_1 \cdots x_n (1 + \theta_{n-1})$$

where $|\theta_n| \leq E_{n,\epsilon_{\mathrm{m}}/2}$, assuming $n\epsilon_{\mathrm{m}} < 2$. You may assume all operations are within the normalised range.