

### Numerical Analysis MATH50003 (2023–24) Problem Sheet 3

**Problem 1** What is  $\pi$  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  $\text{fl}(x)$ ? 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 5** 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 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):

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

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

**Problem 7** 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  $n \leq S$ .