

## Homework-1

MA579H : Scientific Computing

2025

R. Alam

### Floating Point Arithmetic and Rounding Errors

1. Consider the floating-point system  $F(\beta, t, L, U)$ , where  $L := e_{\min}$  and  $U := e_{\max}$ .
  - (a) Determine the total number of normalized floating-point numbers (including 0) represented by the floating-point system given by  $(\beta, t, L, U) = (10, 8, -20, 20)$ . Also determine machine precision, unit roundoff, the largest and the smallest positive floating point numbers in  $F(\beta, t, L, U)$ .
  - (b) Suppose that  $(\beta, t, L, U) = (10, 4, -2, 2)$ . State (with reason) which of the following calculations are exactly representable as a normalized floating-point number in the system? (i)  $(1/100)^2$  (ii)  $1/3$  (iii)  $\sqrt{121}$ .
  - (c) Given two floating point numbers  $a$  and  $b$ , the midpoint  $c := \text{fl}(\text{fl}(a + b)/2)$  may not lie in  $[a, b]$ . For example, consider  $a = .997$  and  $b = .999$  in three decimal digits arithmetic. Show that this is impossible in IEEE arithmetic ( $\beta = 2$ ), that is, when  $\beta = 2$  we have  $a \leq \text{fl}(\text{fl}(a + b)/2) \leq b$ .  
 [Hint: Assume round to nearest rounding mode and that  $\text{fl}$  is a monotone function, that is,  $a \leq b \implies \text{fl}(a) \leq \text{fl}(b)$ .]
2. Let  $x \in F(\beta, t, L, U)$  be given by  $x = (.d_1 d_2 \cdots d_t)_\beta \times \beta^e$ . Let  $\mathbf{next}(x)$  denote the smallest floating point number larger than  $x$  and  $\mathbf{prev}(x)$  denote the largest floating point number less than  $x$ . Thus  $x$  is the only floating point number such that  $\mathbf{prev}(x) < x < \mathbf{next}(x)$ . Determine  $\mathbf{next}(x)$  and  $\mathbf{prev}(x)$ . Show that the relative gap between  $x$  and  $\mathbf{next}(x)$  is at most  $\mathbf{eps}$ , where  $\mathbf{eps} = \beta^{1-t}$ , that is,  $|\mathbf{next}(x) - x|/|x| \leq \beta^{1-t}$ .
3. Consider  $x = 10^{-9}$  and  $y = 10^{15}$  in  $F(10, 4, -30, 30)$ . Determine  $\text{ulp}(x)$  and  $\text{ulp}(y)$ . Further, determine  $\mathbf{next}(x) - x$  and  $|\mathbf{next}(x) - x|/x$ , and  $\mathbf{next}(y) - y$  and  $|\mathbf{next}(y) - y|/y$ . Also determine  $\mathbf{prev}(x)$  and  $\mathbf{prev}(y)$ .
4. Let  $x \in F(\beta, t, L, U)$  be such that  $\beta^{e-1} < x < \beta^e$ . Let  $y \in \mathbb{R}$ . If  $|y| < \text{ulp}(x)/2$  then show that  $\text{fl}(x \pm y) = x$ . For  $(\beta, t, L, U) = (10, 4, -30, 30)$  and  $x = 10^{-9}$ , what is the smallest  $y > 0$  such that  $\text{fl}(x + y) > x$ ? Next when  $x = 10^{15}$ , what is the smallest  $y > 0$  such that  $\text{fl}(x + y) > x$ .
5. Consider the floating point system  $F(\beta, t, L, U)$ . Let  $x \in \mathbb{R}$  be a positive number. Suppose that  $x = (.d_1 d_2 \cdots d_t d_{t+1} \cdots) \times \beta^e$ , where  $L < e < U$ . Now set

$$x_L := (.d_1 d_2 \cdots d_t) \times \beta^e \text{ and } x_R := (.d_1 d_2 \cdots d_t) \times \beta^e + \text{ulp}(x_L) = (.d_1 d_2 \cdots \hat{d}_t) \times \beta^e,$$

where  $\hat{d}_t := d_t + 1$ . Show that  $x_L \leq x \leq x_R$ .

Let  $x_M := (x_L + x_R)/2$ . Assume that  $\beta$  is even. Show that if  $d_{t+1} < \beta/2$  then  $x_L \leq x \leq x_M$  and if  $d_{t+1} \geq \beta/2$  then  $x_M \leq x \leq x_R$ . Hence or otherwise show that by defining

$$\text{fl}(x) := \begin{cases} x_L, & \text{if } d_{t+1} < \beta/2 \\ x_R, & \text{if } d_{t+1} \geq \beta/2 \end{cases}$$

we obtain round to nearest rounding mode.

\*\*\*\*\*End\*\*\*\*\*