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 $x \ominus 2$?

SOLUTION None of these computations have errors since they are all exactly representable as floating point numbers. **END**

Problem 2 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)?

SOLUTION

For the first part we use Geometric series:

$$\begin{aligned} 2^{-3}(1.1001100110011\ldots)_2 &= 2^{-3}\left(\sum_{k=0}^{\infty} \frac{1}{2^{4k}} + \frac{1}{2}\sum_{k=0}^{\infty} \frac{1}{2^{4k}}\right) \\ &= \frac{3}{2^4}\frac{1}{1-1/2^4} = \frac{3}{2^4-1} = \frac{1}{5} \end{aligned}$$

Write $-3 = 12 - 15$ hence we have $q = 12 = (01100)_2$. Since $1/5$ is below the midpoint (the midpoint would have been the first magenta bit was 1 and all other bits are 0) we round down and hence have the bits:

0 01100 1001100110

Adding 1 we get:

$$1 + 2^{-3} * (1.1001100110)_2 = (1.00110011001 1)_2 \approx (1.0011001101)_2$$

Here we write the exponent as $0 = 15 - 15$ where $q = 15 = (01111)_2$. Thus we have the bits:

0 01111 0011001101

END

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):

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

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

SOLUTION

The first problem is very similar to what we saw in lecture. Write

$$(\mathrm{fl}(1.1) \otimes \mathrm{fl}(1.2)) \oplus \mathrm{fl}(1.3) = (1.1(1+\delta_1)1.2(1+\delta_2)(1+\delta_3) + 1.3(1+\delta_4))(1+\delta_5)$$

where we have $|\delta_1|, \ldots, |\delta_5| \leq \epsilon_{\mathrm{m}}/2$. We first write

$$1.1(1+\delta_1)1.2(1+\delta_2)(1+\delta_3) = 1.32(1+\varepsilon_1)$$

where, using the bounds:

$$|\delta_1\delta_2|, |\delta_1\delta_3|, |\delta_2\delta_3| \leq \epsilon_{\mathrm{m}}/4, |\delta_1\delta_2\delta_3| \leq \epsilon_{\mathrm{m}}/8$$

we find that

$$|\varepsilon_1| \leq |\delta_1| + |\delta_2| + |\delta_3| + |\delta_1\delta_2| + |\delta_1\delta_3| + |\delta_2\delta_3| + |\delta_1\delta_2\delta_3| \leq (3/2 + 3/4 + 1/8) \leq 3\epsilon_{\rm m}$$

Then we have

$$1.32(1 + \varepsilon_1) + 1.3(1 + \delta_4) = 2.62 + \underbrace{1.32\varepsilon_1 + 1.3\delta_4}_{\varepsilon_2}$$

where

$$|\varepsilon_2| \leq (6 + 1)\epsilon_{\rm m} = 7\epsilon_{\rm m}.$$

Finally,

$$(2.62 + \varepsilon_2)(1 + \delta_5) = 2.62 + \underbrace{\varepsilon_2 + 2.62\delta_5 + \varepsilon_2\delta_5}_{\varepsilon_3}$$

where, using $|\delta_7\delta_5| \leq \epsilon_{\rm m}/4$,

$$|\varepsilon_3| \leq (7 + 3/2 + 1/4)\epsilon_{\rm m} \leq 9\epsilon_{\rm m}.$$

(Note this our answer is sharper than required as the question gave some breathing room in case the pessimistic bounds you used were more pessimistic than mine).

For the second part, we do:

$$(\mathrm{fl}(1.1) \ominus 1) \oslash \mathrm{fl}(0.1) = \frac{(1.1(1 + \delta_1) - 1)(1 + \delta_2)}{0.1(1 + \delta_3)}(1 + \delta_4)$$

where we have $|\delta_1|, \ldots, |\delta_4| \leq \epsilon_{\rm m}/2$. Write

$$\frac{1}{1 + \delta_3} = 1 + \varepsilon_1$$

where, using that $|\delta_3| \leq \epsilon_{\rm m}/2 \leq 1/2$, we have

$$|\varepsilon_1| \leq \left|\frac{\delta_3}{1 + \delta_3}\right| \leq \frac{\epsilon_{\rm m}}{2}\frac{1}{1 - 1/2} \leq \epsilon_{\rm m}.$$

Further write

$$(1 + \varepsilon_1)(1 + \delta_4) = 1 + \varepsilon_2$$

where

$$|\varepsilon_2| \leq |\varepsilon_1| + |\delta_4| + |\varepsilon_1||\delta_4| \leq (1 + 1/2 + 1/2)\epsilon_{\rm m} = 2\epsilon_{\rm m}.$$

We also write

$$\frac{(1.1(1 + \delta_1) - 1)(1 + \delta_2)}{0.1} = 1 + \underbrace{11\delta_1 + \delta_2 + 11\delta_1\delta_2}_{\varepsilon_3}$$

where

$$|\varepsilon_3| \leq (11/2 + 1/2 + 11/4) \leq 9\epsilon_{\rm m}$$

Then we get

$$(\mathrm{fl}(1.1) \ominus 1) \oslash \mathrm{fl}(0.1) = (1 + \varepsilon_3)(1 + \varepsilon_2) = 1 + \underbrace{\varepsilon_3 + \varepsilon_2 + \varepsilon_2\varepsilon_3}_{\varepsilon_4}$$

and the error is bounded by:

$$|\varepsilon_4| \leq (9 + 2 + 18)\epsilon_{\rm m} \leq 29\epsilon_{\rm m}.$$

END

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

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

where absolute error is bounded by

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

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

SOLUTION

In floating point we have

$$\begin{aligned}
\frac{f^{\rm FP}(x+2h) \ominus f^{\rm FP}(x-2h)}{2h} &= \frac{f(x+h) + \delta_{x+h} - f(x-h) - \delta_{x-h}}{2h}(1+\delta_1) \\
&= \frac{f(x+h) - f(x-h)}{2h}(1+\delta_1) + \frac{\delta_{x+h} - \delta_{x-h}}{2h}(1+\delta_1)
\end{aligned}$$

Applying Taylor's theorem we get

$$(f^{\rm FP}(x+h) \ominus f^{\rm FP}(x-h))/(2h) = f'(x) + \underbrace{f'(x)\delta_1 + \delta^{\rm T}_{x,h}(1+\delta_1) + \frac{\delta_{x+h} - \delta_{x-h}}{2h}(1+\delta_1)}_{\delta^{\rm CD}_{x,h}}$$

where

$$|\delta^{\rm CD}_{x,h}| \leq \frac{|f'(x)|}{2}\epsilon_{\rm m} + \frac{M}{3}h^2 + \frac{2c\epsilon_{\rm m}}{h}$$

END

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.

SOLUTION

For X/n: if $x \in X$ then $a/n \leq x/n \leq b/n$ means $x \in [a/n, b/n]$. Similarly, if $z \in [a/n, b/n]$ then $a \leq nz \leq b$ hence $nz \in X$ and therefore $z \in X/n$.

For XY: if $x \in X$ and $y \in Y$ then $ac \leq xy \leq bd$ means $xy \in [ac, bd]$. Note $ac, bd \in XY$. To employ convexity we take logarithms. In particular if $z \in [ac, bd]$ then $\log a + \log c \leq \log z \leq \log b + \log d$. Hence write

$$\log z = (1-t)(\log a + \log c) + t(\log b + \log d) = \underbrace{(1-t)\log a + t\log b}_{\log x} + \underbrace{(1-t)\log c + t\log d}_{\log y}$$

i.e. we have $z = xy$ where

$$\begin{aligned} x &= \exp((1-t)\log a + t\log b) = a^{1-t}b^t \in X \\ y &= \exp((1-t)\log c + t\log d) = c^{1-t}d^t \in Y. \end{aligned}$$

The generalisation to negative cases proceeds by being a bit careful with the signs. Eg if $n < 0$ we need to swap the order hence we get:

$$A/n = \begin{cases} [a/n, b/n] & n > 0 \\ [b/n, a/n] & n < 0 \end{cases}$$

For multiplication we just use min and max in a naive fashion:

$$AB = [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)].$$

END

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.

SOLUTION Note that

$$\frac{1}{6} = \frac{1}{2}\frac{1}{3} = 2^{-3}(1.010101\ldots)_2$$

Thus

$$[1, 1] \oslash 6 = 2^{-3}[(1.0101010101)_2, (1.0101010110)_2]$$

And hence

$$\begin{aligned} [1, 1] \ominus ([1, 1] \oslash 6) &= [1, 1] \ominus [(0.0010101010101)_2, (0.0010101010110)_2] \\ &= [\mathrm{fl}^{\mathrm{down}}(0.110101010100111111\ldots)_2, \mathrm{fl}^{\mathrm{up}}(0.110101010100111111\ldots)_2] \\ &= 2^{-1}[(1.1010101010)_2, (1.1010101011)_2] = [0.8330078125, 0.83349609375] \end{aligned}$$

END

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 $x \in [0, 1]$.

SOLUTION

We have from Taylor's theorem up to order x^{2n+2}:

$$\sin\ x = \sum_{k=0}^{n} \frac{(-1)^k x^{2k+1}}{(2k+1)!} + \underbrace{\frac{\sin^{2n+3}(t) x^{2n+3}}{(2n+3)!}}_{\delta_{x,2n+1}}.$$

The bound follows since all derivatives of sin are bounded by 1 and we have assumed $|x| \leq 1$.

END

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

$$\sin 1 \in [(0.11010011000)_2, (0.11010111101)_2] = [0.82421875, 0.84228515625]$$

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

SOLUTION Using $n = 1$ we have

$$\sum_{k=0}^{1} \frac{(-1)^k x^{2k+1}}{(2k+1)!} = x - \frac{x}{3!} \in x \ominus (x \oslash 6).$$

Thus we compute

$$\begin{aligned}
\sin 1 &\in [1,1] \ominus [1,1] \oslash 6 \oplus [\mathrm{fl}^{\mathrm{down}}(-1/120), \mathrm{fl}^{\mathrm{up}}(1/120)] \\
&= [(0.11010101010)_2, (0.11010101011)_2] \oplus [-(0.0000001000100010)_2, (0.00000010001000101)_2] \\
&= [\mathrm{fl}^{\mathrm{down}}(0.110100011000{\color{magenta}11101111\ldots})_2, \mathrm{fl}^{\mathrm{up}}(0.110101111100{\color{magenta}000101})_2] \\
&= [(0.11010011000)_2, (0.11010111101)_2] = [0.82421875, 0.84228515625]
\end{aligned}$$

END