Numerical Analysis MATH50003 (2024–25) Problem Sheet 4

Problem 1 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 $x/n \in [a/n, b/n]$. 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 2(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

$$\begin{aligned} [1, 1] \ominus ([1, 1] \oslash 6) &= [1, 1] \ominus [(0.0010101010101)_2, (0.0010101010110)_2] \\ &= [\mathrm{fl}^{\mathrm{down}}(0.1101010101010111111\ldots)_2, \mathrm{fl}^{\mathrm{up}}(0.11010101010011111\ldots)_2] \\ &= 2^{-1}[(1.1010101010)_2, (1.1010101011)_2] = [0.8330078125, 0.83349609375] \end{aligned}$$

END

Problem 2(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 2(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

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

Noting that in floating point $1 \otimes 1 = 1$ (ie it is exact) we compute

$$\begin{aligned}
\sin 1 &\in [1,1] \ominus [1,1] \oslash 6 \oplus [\mathrm{fl}^{\mathrm{down}}(-1/120), \mathrm{fl}^{\mathrm{up}}(1/120)] \\
&= [(0.11010101010)_2, (0.11010101011)_2] \oplus [-(0.0000001000100010)_2, (0.00000010001000101)_2] \\
&= [\mathrm{fl}^{\mathrm{down}}(0.11010011000{\color{magenta}11101111\ldots})_2, \mathrm{fl}^{\mathrm{up}}(0.11010111100{\color{magenta}000101})_2] \\
&= [(0.11010011000)_2, (0.11010111101)_2] = [0.82421875, 0.84228515625]
\end{aligned}$$

END

Problem 3 For $A \in F_{\infty,S}^{n\times n}$ and $\boldsymbol{x} \in F_{\infty,S}^{n}$ consider the error in approximating matrix multiplication with idealised floating point: for

$$A\boldsymbol{x} = \begin{pmatrix} \bigoplus_{j=1}^{n} A_{1,j} \otimes x_j \\ \vdots \\ \bigoplus_{j=1}^{n} A_{1,j} \otimes x_j \end{pmatrix} + \delta$$

use Problem 8 on PS3 to show that

$$\|\delta\|_\infty \leq 2\|A\|_\infty \|\boldsymbol{x}\|_\infty E_{n,\epsilon_{\mathrm{m}}/2}$$

for $E_{n,\epsilon} := \frac{n\epsilon}{1-n\epsilon}$, where $n\epsilon_{\mathrm{m}} < 2$ and the matrix norm is $\|A\|_\infty := \max_k \sum_{j=1}^{n} |a_{kj}|$.

SOLUTION We have for the k=th row

$$\bigoplus_{j=1}^{n} A_{k,j} \otimes x_j = \bigoplus_{j=1}^{n} A_{k,j} x_j (1+\delta_j) = \sum_{j=1}^{n} A_{k,j} x_j (1+\delta_j) + \sigma_{k,n}$$

where we know $|\sigma_n| \leq M_k E_{n-1,\epsilon_{\mathrm{m}}/2}$, where from 1(b) we have

$$M_k = \Sigma_{j=1}^n |A_{k,j} x_j (1+\delta_j)| = \Sigma_{j=1}^n |A_{k,j}||x_j|(1+|\delta_j|) \leq 2 \max |x_j| \Sigma_{j=1}^n |A_{k,j}| \leq 2\|\boldsymbol{x}\|_\infty \|A\|_\infty$$

Similarly, we also have

$$|\sum_{j=1}^n A_{k,j} x_j \delta_j| \leq \|\boldsymbol{x}\|_\infty \|A\|_\infty \epsilon_{\mathrm{m}}/2$$

and so the result follows from

$$\epsilon_{\mathrm{m}}/2 + 2E_{n-1,\epsilon_{\mathrm{m}}/2} \leq \frac{\epsilon_{\mathrm{m}}/2 + \epsilon_{\mathrm{m}}(n-1)}{1-(n-1)\epsilon_{\mathrm{m}}/2} \leq \frac{\epsilon_{\mathrm{m}} n}{1 - n\epsilon_{\mathrm{m}}/2} = 2E_{n,\epsilon_{\mathrm{m}}/2}.$$

END