Numerical Analysis MATH50003 (2025–26) Problem Sheet 4

We investigate how interval arithmetic can be used for computing rigorous bounds on calculations. This includes both simple arithmetic expressions as well as more complicated mathematical expressions like computing functions from their Taylor series. In particular, we show that one can compute fairly sharp bounds on $\sin 1$. The lab takes this further and implements these computations on a computer, allowing for much higher accuracy rigrouous computations.

In the lectures/notes we also began discussing numerical linear algebra, beginning with a study of structured matrices. This material is primarily studied in the lab, which investigates the practical implementation. In this problem sheet we investigate bounding errors in matrix multiplication, tying in with last weeks content on floating point arithmetic.

We begin by completing the proofs of interval multiplication and division that were omitted in the notes, and ask for the generalisation to cases not considered in the notes:

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

The next problem computes rigorous bounds on $\sin 1$ using interval arithmetic. First consider a simple arithmetic computation with floating point intervals, corresponding to the first two terms of the Taylor series of $\sin x$:

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

$$\begin{aligned}
[1,1] \ominus ([1,1] \oslash 6) &= [1,1] \ominus [(0.0010101010101)_2, (0.0010101010110)_2] \\
&= [\mathrm{fl}^{\mathrm{down}}(0.11010101010\textcolor{magenta}{011111\ldots})_2, \mathrm{fl}^{\mathrm{up}}(0.11010101010\textcolor{magenta}{10111111\ldots})_2] \\
&= 2^{-1}[(1.1010101010)_2, (1.1010101011)_2] = [0.8330078125, 0.83349609375]
\end{aligned}$$

END

When using interval arithmetic with a Taylor series we need to not only compute the errors from floating point rounding, but also the error due to truncating the Taylor series. The following asks to bound this tail (a computation that does not involve floating point arithmetic at all):

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

We can finally combine the two sources of error to get a rigorous bound on $\sin 1$. This builds understanding on what the computer is doing in the related questions in the lab:

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

$$\begin{aligned}
\sin 1 &\in [1, 1] \ominus [1, 1] \oslash 6 \oplus [\mathrm{fl}^{\mathrm{down}}(-1/120), \mathrm{fl}^{\mathrm{up}}(1/120)] \\
&= [(0.11010101010)_2, (0.11010101011)_2] \oplus [-(0.00000010001000101)_2, (0.00000010001000101)_2] \\
&= [\mathrm{fl}^{\mathrm{down}}(0.11010011000\,11101011\ldots)_2, \mathrm{fl}^{\mathrm{up}}(0.11010111100\,000101)_2] \\
&= [(0.11010011000)_2, (0.11010111101)_2] = [0.82421875, 0.84228515625]
\end{aligned}$$

END

Our last problem considers the implementation of matrix multiplication in (idealised) floating point. In the previous PS we saw we could bound errors in multiplication and addition. Here we apply these results to relate the error of matrix multiplication to a matrix norm. This demonstrates how one can move away from the technical details of floating point arithmetic and instead understand errors in terms of more mathematical notions like norms that do not depend on the particulars of floating point:

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

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

0.1 END