MATH50003 Numerical Analysis (2022–23)

Problem Sheet 2

This problem sheet explores the bounding of floating point arithmetic errors, and shows how these can be used to bound errors in algorithms.

Please complete the problems using pen-and-paper, though some can be verified using Julia.

Problem 1 Suppose $0 \leq x < \min F_{\sigma,Q,S}^{\rm normal}$ (the *sub-normal range*). Show that rounding has guaranteed *absolute error*:

$$
\begin{aligned}
{\rm fl}^{\rm up}(x) &= x + \delta_x^{\rm up} \\
{\rm fl}^{\rm down}(x) &= x + \delta_x^{\rm down} \\
{\rm fl}^{\rm near}(x) &= x + \delta_x^{\rm near}
\end{aligned}
$$

where

$$
\begin{aligned}
|\delta_x^{\rm up/down}| &\leq 2^{1-\sigma-S} \\
|\delta_x^{\rm near}| &\leq 2^{-\sigma-S}
\end{aligned}
$$

SOLUTION

For

$$x = 2^{1-\sigma} * (0.b_1 b_2 \ldots)_2$$

we have

$$x_- = {\rm fl}^{\rm down}(x) = 2^{1-\sigma} * (0.b_1 \ldots b_S)_2, x_h := x_- + 2^{-\sigma-S}, x_+ = {\rm fl}^{\rm up}(x) = x_- + 2^{1-\sigma-S}$$

Therefore

$$
\begin{aligned}
x - {\rm fl}^{\rm down}(x) = x - x_- \leq x_+ - x_- = 2^{1-\sigma-S} \\
{\rm fl}^{\rm up}(x) - x = x_+ - x \leq x_+ - x_- = 2^{1-\sigma-S}
\end{aligned}
$$

If ${\rm fl}(x) = {\rm fl}(x)^{\rm down}(x)$ then $x_- \leq x \leq x_h$ and

$$x - {\rm fl}(x) \leq x_h - x_- = 2^{-\sigma-S}.$$

If ${\rm fl}(x) = {\rm fl}(x)^{\rm up}(x)$ then $x_h \leq x \leq x_+$ and

$$ {\rm fl}(x) - x \leq x_+ - x_h = 2^{-\sigma-S}.$$

END

Problem 2.1 Suppose $|\epsilon_k| \leq \epsilon$ and $n\epsilon < 1$. Show that

$$\prod_{k=1}^{n}(1+\epsilon_k) = 1+\theta_n$$

for some constant θ_n satisfying

$$|\theta_n| \leq \underbrace{\frac{n\epsilon}{1-n\epsilon}}_{E_{n,\epsilon}}$$

Hint: use induction.

SOLUTION

$$\prod_{k=1}^{n+1}(1+\epsilon_k) = \prod_{k=1}^{n}(1+\epsilon_k)(1+\epsilon_{n+1}) = (1+\theta_n)(1+\epsilon_{n+1}) = 1+\underbrace{\theta_n+\epsilon_{n+1}+\theta_n}_{\theta_{n+1}}$$

where

$$\begin{aligned}
|\theta_{n+1}| &\leq \frac{n\epsilon}{1-n\epsilon}(1+\epsilon)+\epsilon \\
&= \frac{n\epsilon+n\epsilon^2}{1-(n+1)\epsilon}\frac{1-(n+1)\epsilon}{1-n\epsilon}+\frac{\epsilon-(n+1)\epsilon^2}{1-(n+1)\epsilon} \\
&\leq \frac{(n+1)-n\epsilon}{1-(n+1)\epsilon}\epsilon \leq \frac{(n+1)\epsilon}{1-(n+1)\epsilon}
\end{aligned}$$

END

Problem 2.2 Show if $x_1, \ldots, x_n \in F$ then

$$x_1 \otimes \cdots \otimes x_n = x_1 \cdots x_n (1+\theta_{n-1})$$

where $|\theta_n| \leq E_{n,\epsilon_{\mathrm{m}}/2}$, assuming $n\epsilon_{\mathrm{m}} < 2$. You may assume all operations are within the normalised range.

SOLUTION

We can expand out:

$$x_1 \otimes \cdots \otimes x_n = (\cdots((x_1x_2)(1+\delta_1)x_3(1+\delta_2)\cdots x_n(1+\delta_{n-1})) = x_1\cdots x_n(1+\delta_1) \cdots (1+\delta_{n-1})$$

where $|\delta_k| \leq \epsilon_{\mathrm{m}}/2$. The result then follows from the previous result.

END

Problem 2.3 Show if $x_1, \ldots, x_n \in F$ then

$$x_1 \oplus \cdots \oplus x_n = x_1 + \cdots + x_n + \sigma_n$$

where, for $M = \Sigma_{k=1}^n |x_k|$, $|\sigma_n| \leq M E_{n-1,\epsilon_{\mathrm{m}}/2}$, assuming $n\epsilon_{\mathrm{m}} < 2$. You may assume all operations are within the normalised range. Hint: use Problem 2.1 to first write

$$x_1 \oplus \cdots \oplus x_n = x_1(1+\theta_{n-1}) + \sum\nolimits_{j=2}^n x_j(1+\theta_{n-j+1}).$$

SOLUTION

Using Problem 2.1 we write:

$$(\cdots((x_1 + x_2)(1+\delta_1) + x_3)(1+\delta_2)\cdots + x_n)(1+\delta_{n-1}) = x_1 \prod\nolimits_{k=1}^{n-1}(1+\delta_k) + \sum\nolimits_{j=2}^n x_j \prod\nolimits_{k=j-1}^{n-1}(1+\delta_j) = x_1(1+\theta_{n-1}) + \sum\nolimits_{j=2}^n x_j(1+\theta_{n-j+1})$$

where we have for $j = 2, \ldots, n$

$$|\theta_{n-j+1}| \leq E_{n-j+1,\epsilon_{\mathrm{m}}/2} \leq E_{n-1,\epsilon_{\mathrm{m}}/2}.$$

Thus we have

$$\sum\nolimits_{j=1}^n x_j(1+\theta_{n-j+1}) = \sum\nolimits_{j=1}^n x_j + \underbrace{\sum\nolimits_{j=1}^n x_j \theta_{n-j+1}}_{\sigma_n}$$

where $|\sigma_n| \leq M E_{n-1,\epsilon_{\mathrm{m}}/2}$. ∎

END

Problem 3.1 Consider the algorithm `exp_taylor_fast` from lectures:

In [1]:
```julia
function exp_taylor_fast(x, n)
    ret = zero(x) # 0 of same type as x
    summand = one(x)
    for k = 0:n
        ret += summand
        summand *= x/(k+1)
    end
    ret
end
```

Out[1]: `exp_taylor_fast (generic function with 1 method)`

Write this algorithm as a one-line mathematical function $\exp_n^{\mathrm{t}}(x)$ involving $\oplus$, $\oslash$, and $\otimes$. You may find it convenient to use the notation:

$$\bigoplus\nolimits_{k=1}^n x_k := x_1 \oplus \cdots \oplus x_n = (\cdots((x_1 \oplus x_2) \oplus x_3)\cdots \oplus x_{n-1}) \oplus x_n$$
$$\bigotimes\nolimits_{k=1}^n x_k := x_1 \otimes \cdots \otimes x_n = (\cdots((x_1 \otimes x_2) \otimes x_3)\cdots \otimes x_{n-1}) \otimes x_n$$

SOLUTION

$$\begin{aligned}\exp_n^{\mathrm{t}}(x) &:= 1 \oplus x \oplus (x \otimes (x \oslash 2)) \oplus (x \otimes (x \oslash 2) \otimes (x \oslash 3)) \oplus \cdots \oplus (x \otimes \cdots \otimes (x \oslash \\ &= 1 \oplus \bigoplus\nolimits_{k=1}^{n} \bigotimes\nolimits_{j=1}^{k} (x \oslash j)\end{aligned}$$

END

Problem 3.2 Show that

$$\exp_n^{\mathrm{t}}(x) = \sum\nolimits_{k=0}^{n} \frac{x^k}{k!} + \varepsilon_n$$

where

$$|\varepsilon_n| \leq \exp(|x|)(2E_{2n,\epsilon_{\mathrm{m}}/2} + E_{2n,\epsilon_{\mathrm{m}}/2}^2),$$

assuming $n\epsilon_{\mathrm{m}} < 1$. You may assume all operations are within the normalised range. Hint: combine Problem 2.2 and 2.3 and note that $E_{k,\epsilon_{\mathrm{m}}/2} \leq E_{j,\epsilon_{\mathrm{m}}/2}$ when $k \leq j$.

SOLUTION

From Problem 2.2 we have

$$x_k := \bigotimes\nolimits_{j=1}^{k} (x \oslash j) = \bigotimes\nolimits_{j=1}^{k} (x/j)(1 + \delta_j) = (1 + \theta_{2k-1}) \frac{x^k}{k!}$$

where θ_{2k-1} are some constants (the subscript denotes the number of terms in the expansion) satisfying

$$|\theta_{2k-1}| \leq E_{2k-1,\epsilon_{\mathrm{m}}/2} \leq E_{2n,\epsilon_{\mathrm{m}}/2}.$$

(We use the convention $x_0 := 1$ and $\theta_{-1} := 0$.) Note that

$$\begin{aligned}M := \sum\nolimits_{k=0}^{n} |x_k| \leq \sum\nolimits_{k=0}^{n} (1 + |\theta_{2k-1}|) \frac{|x|^k}{k!} &\leq \max_k (1 + |\theta_{2k-1}|) \sum\nolimits_{k=0}^{n} \frac{|x|^k}{k!} \\ &\leq (1 + E_{2n,\epsilon_{\mathrm{m}}/2}) \exp|x|.\end{aligned}$$

Then from Problem 2.3 we have

$$\oplus_{k=0}^{n} x_k = \sum\nolimits_{k=0}^{n} x_k + \sigma_{n+1} = \sum\nolimits_{k=0}^{n} \frac{x^k}{k!} + \underbrace{\sum\nolimits_{k=0}^{n} \theta_{2k-1} \frac{x^k}{k!} + \sigma_{n+1}}_{\varepsilon_n}$$

for

$$|\sigma_{n+1}| \leq M E_{n,\epsilon_{\mathrm{m}}/2} \leq (E_{2n,\epsilon_{\mathrm{m}}/2} + E_{2n,\epsilon_{\mathrm{m}}/2}^2) \exp|x|.$$

Thus

$$|\varepsilon_n| \leq (E_{2n,\epsilon_\mathrm{m}/2}^2 + 2E_{2n,\epsilon_\mathrm{m}/2}) \exp |x|$$

END

Problem 3.3 For $x > 0$, find a bound on the relative error $|\rho_n|$ where

$$\exp_n^{\mathrm{t}}(x) = (1 + \rho_n) \exp x.$$

Why does the bound break down when $x < 0$?

SOLUTION

We use Taylor's remainder theorem to write:

$$\sum_{k=0}^{n-1} \frac{x^k}{k!} = \exp x - \exp t \frac{x^n}{n!}$$

for $t \in [0, x]$. Thus we have

$$\exp_n^{\mathrm{t}}(x) = \exp x + \underbrace{\exp(x)(2E_{n,\epsilon_\mathrm{m}/2} + E_{n,\epsilon_\mathrm{m}/2}^2) - \exp(t)\frac{x^n}{n!}}_{\tilde{\varepsilon}_n}$$

where

$$|\tilde{\varepsilon}_n| \leq (2E_{n,\epsilon_\mathrm{m}/2} + E_{n,\epsilon_\mathrm{m}/2}^2 + \frac{|x|^{n+1}}{(n+1)!}) \exp x$$

Dividing through by $\exp x$ we have

$$|\rho_n| = \exp(-x)|\tilde{\varepsilon}_n| \leq 2E_{n,\epsilon_\mathrm{m}/2} + E_{n,\epsilon_\mathrm{m}/2}^2 + \frac{|x|^{n+1}}{(n+1)!}.$$

The expression breaks down because $\exp |x| \neq \exp x$ hence one gets an error bound that grows exponentially as $x \to -\infty$.

END

Problem 3.4 Give two reasons why the above error bound is not valid as $n \to \infty$ if $F_{\sigma,Q,S}$ is fixed. If S and Q are allowed to depend on n can we guarantee convergence to $\exp x$?

SOLUTION

1. $x^k/k!$ is eventually a sub-normal number so the assumption on normalised range is not valid.
2. If $n > 2/\epsilon_\mathrm{m}$ the conditions of the theorem are not met.

Yes: if we grow Q sufficiently fast then we will never reach a sub-normal number, and $\epsilon_\mathrm{m} = S^{-n}$ will become suffiently small that $\epsilon_\mathrm{m} n < 2$. In this case our error estimate goes to 0 as $n \to \infty$.

END