APCS23
A1
Q1
P1

1) prove $\left|\frac{x - rd(x)}{x}\right| \leq 2^{-p}$

note: $rd(x) = x(1+\varepsilon)$ where $\varepsilon \leq eps$

$x = \pm \left(\sum_{l=1}^{\infty} b_l 2^{-l}\right) 2^e$

$x - rd(x) = ?$ ⇒ 2 cases

Case 1: First discarded bit is 0

$x - rd(x) = \pm\left(\sum_{l=p+2}^{\infty} b_{-l} 2^{-l}\right) 2^e$

↑ p+1 digit is correct $2^{-p-2} = 2^{-p}2^{-2}$

$(|x - rd(x)|)_{max} = \left(\sum_{l=1}^{\infty} 2^{-l}\right) 2^{e-p-1}$ → $b_{-l} = 1 \; \forall l = p+2, \ldots$

$\lim_{l \to \infty} \sum_{l=1}^{\infty} 2^{-l} = 1$

∴ $(|x - rd(x)|)_{max} = 2^{e-p-1}$

$(x)_{min} = 2^{e-1}$ → smallest mantissa $= 0.1 = 2^{-1}$

∴ for Case 1: $\left|\frac{x - rd(x)}{x}\right| \leq \left|\frac{2^{e-p-1}}{2^{e-1}}\right| = 2^{-p}$ ✓

Case 2: First discarded bit is 1

$x - rd(x) = \left(2^{-p} - \sum_{l=p+1}^{\infty} b_{-l} 2^{-l}\right) 2^e$ → $b_{-l} = 1 \; \forall p+1$, $b_{-l} = 0 \; \forall l > p+1$

$(|x - rd(x)|)_{max} = \left|2^{-p} - (2^{-p-1})\right| 2^e = (1 - 2^{-1}) 2^{e-p} = (2^{-1}) 2^{e-p} = 2^{e-p-1}$

∴ $\frac{|x - rd(x)|}{x} \leq \frac{2^{e-p-1}}{2^{e-1}} = 2^{-p}$ ✓

↑ x_{min} doesn't change

APC 523
A1
Q3
P1

3ai) → X is a machine # ∴ error in xy comes from the fl(xy) operation

$$fl(x*x) = (1+\varepsilon_f)(xx)$$

ε_f floating error

$$\therefore fl((xx)x) = (xx)x(1+\varepsilon_f)(1+\varepsilon_f) = xxx(1+2\varepsilon_f)$$

$$\therefore fl(x^n) = x^n(1+(n-1)\varepsilon_f)$$

⇒ max error is $x^n \underbrace{(n-1)\,eps}_{\varepsilon_{tot}}$

ii) ① $fl(\ln x)$

② $fl(n\,fl(\ln(x)))$ or $fl(\ln(x)) + fl(\ln(x)) + \dots$

③ $fl(\exp(\text{step}②))$

① $fl(\ln x) = \ln(x)(1+\varepsilon_l)$

② i) $fl(n \cdot fl(\ln(x))) = n\ln(x)(1+\varepsilon_l)(1+\varepsilon_n) = n\ln(x)(1+\varepsilon_l+\varepsilon$

or

$$fl(\underbrace{fl(\ln(x)) + fl(\ln(x))}_{z}) = (\ln x + \ln x)(1+\varepsilon_l+\varepsilon_n)$$

(float of ln, float of ε ln)

$$z + fl(\ln(x)) = (\ln x + \ln x + \ln x)\left(1 + \frac{\ln x}{3\ln x}\varepsilon_l + \frac{2\ln x}{3\ln x}(\varepsilon_l+\varepsilon_n)\right)$$

$$w = 3\ln x\left(1 + \frac{3\varepsilon_l + 2\varepsilon_n}{3}\right) \rightarrow fl(w) = 3\ln x\left(1+\frac{3\varepsilon_l+2\varepsilon}{3}\right.$$

APC 523
A1
Q3
P2

$$\frac{\ln x}{4\ln x}\varepsilon_1 + \frac{3\ln x}{4\ln x}\left(\frac{3\varepsilon_1 + 5\varepsilon_n}{3}\right)$$

$$\frac{4\varepsilon_1 + 5\varepsilon_n}{4} = \varepsilon_1 + \frac{5}{12}\varepsilon_n$$

$\Rightarrow fl() \Rightarrow \varepsilon_1 + \frac{17}{12}\varepsilon_n$

3aii) using adding technique:

cont'd

$$fl(fl(\ln x) + fl(\ln x) + \cdots) = n\ln x\left(1 + \varepsilon_1 + K\varepsilon_n\right)$$

where $K =$

(B) using $fl(n\,fl(\ln x))$ we get:

$$x^n = \exp(n\ln x\,(1+\varepsilon_1+\varepsilon_n)) = e^{n\ln x} * e^{n\ln x(\varepsilon_1+\varepsilon_n)}$$

$\Downarrow$ taylor

$$\approx e^{n\ln x}\left(1 + n\ln x(\varepsilon_1+\varepsilon_n)\right)$$

now, $fl(x^n) = e^{n\ln x}\left(1 + n\ln(x)(\varepsilon_1+\varepsilon_n) + \varepsilon_f\right)$

$\Rightarrow$ max error is $e^{n\ln x}\underbrace{(2n\ln(x)+1)\,eps}_{\varepsilon_{totii}}$

note:

- if x is small ($\ll 1$), $\ln x$ blows up $\therefore e^{n\ln x}$ is a bad choice as ε_{totii} will blow up
- if $x \sim 1$, $\varepsilon_{totiii} \sim eps$, $e^{n\ln x}$ is a good choice

APC 523
A1
Q3
P3

b) → x, a not machine numbers
→ no floating point error

i) $a = a + \varepsilon_a$, $x = x$

$$x^{a+\varepsilon_a} = x^a x^{\varepsilon_a} = x^a(1 + \varepsilon_a \ln(x))$$

(taylor expand)

$$\varepsilon_{tot} = \varepsilon_a \ln(x)$$

ii) $(x+\varepsilon_x)^a = x^a + a\varepsilon_x x^{a-1} = x^a\left(1 + \frac{a\varepsilon_x}{x}\right)$

(taylor around $\varepsilon_x \simeq 0$)

⇒ for both cases error gets large when $x \to 0$
⇒ Case 1: error will also get large as $x \to \infty$, but asympto
will be very small when $x \sim 1$
⇒ Case 2: error ~ large when $x < 1$, diminishes as $x \to \infty$
↳ also grows with a

Large x → use Case 2
$x \sim 1$ → use Case 1
$x < 1$ → bad all around

APC 523
A1
Q4
P1

4) $f(x) = 1 - e^{-x} = y$, $[0, 1]$

a) $(\text{cond } f)(x) = ?$

$\Delta y = f(x+\Delta x) - f(x) = f'(x)\,\Delta x$

$f'(x) = e^{-x}$, $\Delta y = e^{-x}\,\Delta x$

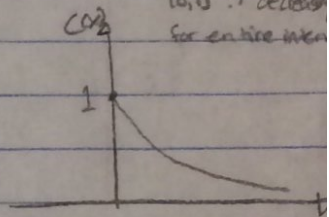
$\varepsilon_y = \left| \underbrace{\frac{x f'(x)}{f(x)}}_{\text{cond \#} = C} \right| \varepsilon_x$

$C(x) = \left| \frac{x f'(x)}{f(x)} \right| = \left| \frac{x e^{-x}}{1 - e^{-x}} \right| = \frac{x e^{-x}}{1 - e^{-x}}$ $(0, 1]$

↳ for $x \neq 0$ $\frac{dC}{dx} = \frac{e^x(x-1)+1}{(e^x-1)^2}$, $C'(0) = -\frac{1}{2} < 0$ → decreasing, $C' \neq 0$ on interval $[0,1]$ ∴ decreasing for entire interval

$C(1) = \frac{e^{-1}}{1 - e^{-1}} = \frac{1}{e-1} < 1$

$C(0+\varepsilon) = \left| \frac{\varepsilon e^{\varepsilon}}{1 - e^{\varepsilon}} \right| = \left| \frac{\varepsilon}{e^{\varepsilon} - 1} \right| < 1$ ∴



less than 1 on interval

∴ $C(x) < 0 \quad \forall \; x \in (0, 1]$

for $x = 0$: $C(x) = \left| \frac{e^{-x}}{1 - e^{-x}} \right| = \left| \frac{1}{e^{x} - 1} \right|$

$C(0) = \left| \frac{1}{-1} \right| = 1$

APC 523
A1
Q4
p2

4b) $A(i) : x \times (-1) \rightarrow \varepsilon = 0$

$A(ii) : \exp(A(i)) \rightarrow \exp(-x)(1+\varepsilon_e)$

$A(iii) : 1 - A(ii)$

↳ $(1 - \exp(x)(1+\varepsilon_e))(1+\varepsilon_r) = A$

$$A = (1 - \exp(-x))\left(1 + \left(\frac{e^{-x}}{1-e^{-x}}\right)\varepsilon_e\right)(1+\varepsilon_r)$$

$$A = (1-e^{-x})\left(1 + \frac{e^{-x}}{1-e^{-x}}\varepsilon_e + \varepsilon_r\right)$$

$\varepsilon_e \neq \varepsilon_r$ bounded by eps

$$\Rightarrow \varepsilon_A = \text{eps}\left(\frac{e^{-x}}{1-e^{-x}} + 1\right)$$

$A = (1-e^{-x})(1+\varepsilon_A) = f_A(\hat{x})$

→ now need $f(\hat{x}_A)$

$\Rightarrow$ let $\varepsilon_{xA} = \dfrac{x_A - x}{x}$, $x_A = x(\varepsilon_{xA}+1)$

$f(x_A) = 1 - \exp(-x(\varepsilon_{xA}+1)) = 1 - \exp(-x)\exp(-x\varepsilon_{xA})$

↳ taylor

$f(x_A) = 1 - e^{-x}(1 - x\varepsilon_{xA}) = 1 - e^{-x} + e^{-x}x\varepsilon_{xA}$

$$f(x_A) = (1-e^{-x})\left(1 + \left(\frac{xe^{-x}}{1-e^{-x}}\right)\varepsilon_{xA}\right)$$

↳ $\dfrac{-xf'}{f}$

APC523
A1
Q4
P3

4b) cont'd Now $f(x_A) = f_A(x_A)$

$$(1-e^{-x})\left(1 + \frac{xe^{-x}}{1-e^{-x}}\,\varepsilon_{x_A}\right) = (1-e^{-x})(1+\varepsilon_A)$$

$$\varepsilon_A = \frac{xe^{-x}}{1-e^{-x}}\,\varepsilon_{x_A}$$

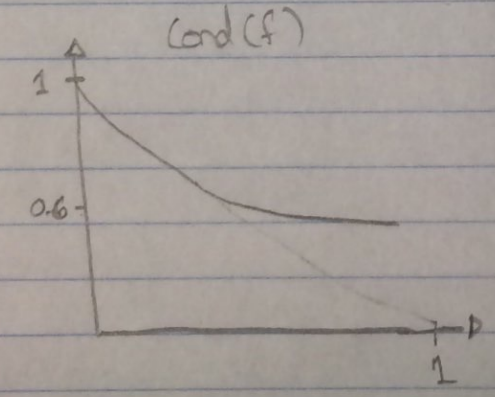
$$\varepsilon_{x_A} = \frac{1-e^{-x}}{xe^{-x}}\left(\text{eps}\left(\frac{e^{-x}}{1-e^{-x}} + 1\right)\right) = \text{eps}\underbrace{\left(\frac{1}{x} + \frac{1-e^{-x}}{xe^{-x}}\right)}_{\text{cond}(A)}$$

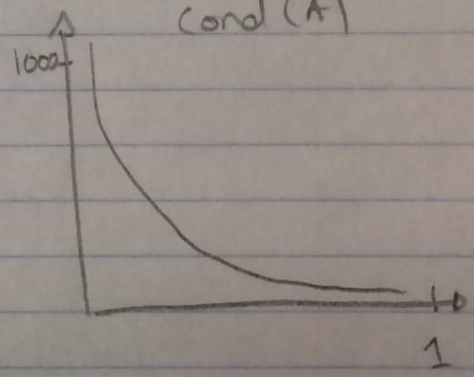
$$\varepsilon_{x_A} = \frac{1}{x}\left(\frac{e^{-x}}{e^{-x}} + \frac{(1-e^{-x})}{e^{-x}}\right)\text{eps} = \frac{e^x}{x}\,\text{eps}$$

$$\Rightarrow \text{cond}(A) = \frac{e^x}{x}$$

See code for nice plot

c)





$$\lim_{x\to 0} \frac{e^x}{x} = \frac{e^0}{0} = \frac{1}{0} = \infty$$

→ $1/x$ causes cond(A) to explode on $[0,1]$

APCS23
A1
Q4
P4

4d) $f = 1 - e^{-x}$

$y \quad x$

$2^{-b} \leq 1 - e^{-x}$

→ $b = 1$, $2^{-1} \leq 1 - e^{-x}$

$e^{-x} \leq 1/2$

$-x \geq \ln(1/2)$

$x \geq -\ln(1/2) = 0.693$

→ $b = 2$, $e^{-x} \leq -(2^{-2} - 1) = 3/4$

$x \geq 0.287682 \quad = -\ln(0.75)$

→ $b = 3$, $e^{-x} \leq -(2^{-3} - 1) = 7/8$

$x \geq -\ln(7/8) = 0.1335$

→ $b = 4$, $e^{-x} \leq -(2^{-4} - 1) = 15/16$

$x \geq -\ln(15/16) = 0.06454$

e) $\frac{\|y_x^* - y\|}{\|y\|} \leq (\text{cond } f)(x) \{ \varepsilon + (\text{cond } A)(x^*) \text{ eps} \}$

$\frac{\|x^* - x\|}{\|x\|} = 0$ since x is a machine#

APCS23
A1
Q4
PS

4e) cont'd

$$\frac{\|y_h^* - y\|}{\|y\|} \leq \left(\frac{xe^{-x}}{1-e^{-x}}\right)\left(\frac{e^x}{x}\right) eps$$

$\|y\|$: err

eps : 2^{-53}

$$err \leq \frac{1}{(1-e^{-x})} eps$$

for $b=1$, $e^{-x} = 1/2$, $err_{max} = 2\,eps$

$b=2$, $e^{-x} = 3/4$, $err_{max} = 4\,eps$

$b=3$, $e^{-x} = 7/8$, $err_{max} = 8\,eps$

$b=4$, $e^{-x} = 16/16$, $err_{max} = 16\,eps$

4f) $1-e^{-x}\left(\frac{1+e^{-x}}{1+e^{-x}}\right) = \frac{1-(e^{-x})^2}{1+e^{-x}}$

$e^{-x} = \cosh x - \sinh x$

$(e^{-x})^2 = c^2 + s^2 - 2sc$

$1 = \cosh^2 x - \sinh^2 x$ } plug in

$= \frac{1}{1+\frac{1}{e^x}}\left(e^{-x} - s^2 - c^2 - s^2 + 2sc\right) = \frac{1}{1+\frac{1}{e^x}}\left(+2s(\underbrace{c-s}_{e^{-x}})\right)$

$= \frac{1}{1+\frac{1}{e^x}}\left(2\sinh(x)e^{-x}\right) = \frac{2\sinh(x)}{e^x+1}$

where $\sinh(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!}$ → no subtractions!

↳ Maclaurin series

A1 APC523
Q6
P1

a) for all these cases, first we do:

$x^{1/2^n}$ → can write as $e^{\frac{1}{2^n}\ln(x)}$

↳ taylor expand

$x^{1/2^n} = 1 + \ln x \; 1/2^n$

now, this is the same as moving the decimal place of $\ln(x)$ n times ∴ when $n = 52$ we only have 1 sigfig left in the mantissa since for double precision it can hold 53 digits. This means when $n = 52$ the only numbers that match $y = x$ are those which are integer powers of e. Similar logic applies for other values of n.

APCS23
A1
Q7
P1

$\Gamma_{k,\ell}$, $\vec{a}$

7e) $\Omega = \Omega(a_0, a_1, \ldots, a_{n-1})$, $\Omega_k = k^{th}$ root of p_k

i) $\Gamma_{k\ell}$, cond # for k^{th} root changes ℓ^{th} coeff

for 1D, $\text{cond}(f) = f(x + \Delta x) - f(x)$

for 2D we get $p((\Omega_k + \delta\Omega_k), (a_\ell + \delta a_\ell)) = p_\Delta$

$p_\Delta = a_0 + a_1(\Omega_k + \delta\Omega_k) + (a_\ell + \delta a_\ell)(\Omega_k + \delta\Omega_k)^2 + \ldots$

note: $(\Omega_k + \delta\Omega_k)^\ell = \Omega_k^\ell \left(1 + \frac{\delta\Omega_k}{\Omega_k}\right)^\ell$

$\frac{\delta\Omega_k}{\Omega_k}$ small, taylor expand

$= \Omega_k^\ell (1 + \ell\, \delta\Omega_k/\Omega_k)$

$= \Omega_k^\ell + \ell\Omega_k^{\ell-1}\delta\Omega_k$

$d\Omega_k^\ell$, $\frac{dp}{d\Omega_\ell}\Big|_{\Omega_k}$

$p_\Delta = \underbrace{a_0 + a_1\Omega_k + \ldots + a_\ell\Omega_k^\ell}_{P(\Omega_k) = 0} + a_1\delta\Omega_k + \ldots + a_\ell(\ell\Omega_k^{\ell-1}\delta\Omega_k) + \delta a_\ell\Omega_k^\ell + \underbrace{\delta a_\ell(\ell\Omega_k^{\ell-1}\delta\Omega_k)}_{\text{second order, ignore}}$

$p_\Delta = \left(\frac{dp}{d\Omega_k}\Big|_{\Omega_k}\right)\delta\Omega_k + \delta a_\ell\Omega_k^\ell$

$\frac{dp}{da_\ell}\Big|_{\Omega_k}$

$\Rightarrow$ note $\text{cond}_{\Omega_k}(a) = p[(\Omega_k + \delta\Omega_k), (a_\ell + \delta a_\ell)] - p(\Omega_k, a_\ell)$

but $p(\Omega_k, a_\ell) = 0$

APC 523
A1
07
P2

W[0] Ω_k^{20}

7e) ∴ in the limit that $\delta\Omega_k \neq \delta a_l$ are small, $p_\Delta \approx 0$

control

$$\Rightarrow \frac{\delta\Omega_k}{\delta a_l} = -\left(\frac{\partial p}{\partial a_l}\Big|_{\Omega_k}\right)\left(\frac{\partial p}{\partial \Omega_k}\Big|_{\Omega_k}\right)^{-1}$$

$$= \Omega_k^l \left(p'(\Omega_k)\right)^{-1}$$

$$\therefore \text{cond}\,\Omega_k(a) = \sum_{l=0}^{n-1} \left|\frac{a_l \Omega_k^l}{\Omega_k p'(\Omega_k)}\right| = \sum_{l=0}^{n-1} \left|\frac{a_l \Omega_k^{l-1}}{p'(\Omega_k)}\right|$$

ii) $r = 14, 16, 17, 20$

$\Omega_k = 15 \quad 17 \quad 18 \quad 21$

$C_{14} = 5.88 \times 10^{13}$

$C_{16} = 3.854 \times 10^{13}$

$C_{17} = 1.723 \times 10^{13}$

$C_{20} = 1.3798 \times 10^{11}$

→ condition number is lower for higher roots

→ all condition #s are $\gg 1$

→ small perturbations in coeffs lead to large changes in roots

iii) → Problem itself is ill-conditioned ∴ no clever algorithm will help ∴

AFC523
A1
Q8
P1

8) $y_n = \int_0^1 x^n e^x \, dx \qquad n \geq 0$

$y_{n+1} = e - (n+1) y_n$

a) g_K is a map from y_N to y_K, $K < N$

$(\text{cond } g_K)(y_N)$ wrt $K \neq N$ = ?

$y_n = \frac{e - y_{n+1}}{n+1} \Rightarrow y_K = \frac{e - y_N}{N}$ → $N = K+1$ (special case)

$(\text{cond } g_K)(y_N) = \left| \frac{y_N \, g'_K}{g_K} \right|$ → need $g'(K)$

↳ figured this out before

$g_K(y_N) = \left(e \cdots \left(e - \left(e - \left(\frac{e - y_N}{N}\right)\frac{1}{(N-1)}\right)\frac{1}{(N-2)}\right) \cdots \right)\frac{1}{(K+1)}$

$g'_K(y_N) = \prod_{i=0}^{N-(K+1)} \frac{1}{(N-i)} = \frac{K!}{N!}$

$\Rightarrow (\text{cond } g_K)(y_N) = \left| \frac{y_N \, K!}{y_K \, N!} \right|$

APC 523
A1
Q8
P2

8b) now assume $\varepsilon_N = 1$

$$\frac{\varepsilon_K}{\varepsilon_N} = \left| \frac{y_N g'_K}{y_K} \right|$$

$$\Rightarrow \varepsilon_K = \varepsilon = \frac{y_N g'_K}{y_K} \quad \Rightarrow \text{well conditioned if } \varepsilon \leq 1$$

$$\therefore \quad 1 \geq \frac{y_N g'_K}{y_K} = \frac{y_N}{y_K} \frac{K!}{N!}$$

need max of this

$$y_N = \int_0^1 e^x x^N \, dx \Rightarrow \text{for } N > K, \; y_N < y_K$$

$$\therefore \frac{y_N}{y_K} \leq 1 \text{, let it equal 1}$$

$$\varepsilon \geq \frac{K!}{N!} \Rightarrow \boxed{N! \geq \frac{K!}{\varepsilon}}$$

8c) $\varepsilon = eps$, $K = 20$, $N = ?$

$\varepsilon = 2^{-53}$

$$N! = \frac{20!}{2^{-53}} = 2.19 \times 10^{34}$$

$30! = 2.6 \times 10^{32}$, $31! = 8.22 \times 10^{33}$, $32! = 2.63 \times 10^{35} > 2.19 \times 10^{34}$

$\boxed{N = 32}$

APC 523
A1
Q8
P3

d) using $K = 20$, $N = 32$, $y_N = 0$:

$y_{20} = 0.1238\ldots$ → python

$\int_0^1 e^x x^{20} dx = 0.1238\ldots$ → wolfram

rel error $= \frac{y_w - y_{20}}{y_w} = 2.24 \times 10^{-16} \sim 2^{-53} = 1 \times 10^{-1}$

✓

WOW !!