Charles Maher

1.)

let $x = \pm \left(\sum_{\ell=1}^{\infty} b_\ell 2^{-\ell} \right) 2^e$

Case 1:

$b_{p+1} = 0$ $\quad x - \text{round}(x) = (0.0 b_{p+2} b_{p+3} \ldots)_2 \times 2^e 2^{-p}$

$|x - \text{round}(x)| < (.0\overline{11})_2 \times 2^{e-p} = 2^{e-p-1}$

note smallest possible mantissa $\rightarrow (0.b_1 b_2 \ldots)_2 \geq (0.b_1 00\ldots)_2 = (0.1)_2$

Case 2:

$b_{p+1} = 1$ $\quad \text{round}(x) = (0.b_1 \ldots b_p)_2 2^e + 2^{-p} \times 2^e$

$x - \text{round}(x) = (0.1\, b_{p+2} \cdots)_2 2^{e-p} - 2^{e-p} = \frac{1}{2} + \frac{1}{2}(0.b_{p+2} \ldots)_2) 2^{e-p} - 2^{e-p}$

$= \left(\frac{1}{2} + \frac{1}{2}(0.b_{p+2} \ldots) - 1\right) 2^{e-p}$

and $\quad 0 \leq (0.b_{p+2} \cdots)_2 \leq 1$

$|x - \text{round}(x)| = \frac{1}{2}(1 - (0.b_{p+2} \ldots)_2) 2^{e-p} < \frac{1}{2} 2^{e-p} = 2^{e-p-1}$

using the above mantissa lower bound:

$$\frac{|x - \text{round}(x)|}{|x|} < \frac{2^{e-p-1}}{2^{e-1}} = 2^{-p} \quad \square.$$

Charles Maher

2.)

(code used to generate numbers attached separately)

a.)

(first 31 terms in attached notebook)

→ sum of these 31 terms is $\boxed{244.71}$

b.)

(partial sums in notebook)

converges at $k=17$

magnitude of error is $\frac{|e^{5.5} - 244.71|}{e^{5.5}} = 7.3838 \times 10^{-5}$

c.)

(partial sums in notebook)

also converges to 244.71, takes all 31 terms

relative error is the same because the converged value is the same

d.) (partial sums in notebook)

i.) converges to 3.8363×10^{-3} at $k=26$

$$\frac{|e^{-5.5} - 3.8363 \times 10^{-3}|}{e^{-5.5}} = 0.06128$$

ii.) converges to 4×10^{-3} at $k=31$

$$\frac{|e^{-5.5} - 4 \times 10^{-3}|}{e^{-5.5}} = 0.02123$$

← least error

iii.) converges to 0 at $k=18$, relative error is 1 ← fastest converging

iv.) converges to 1×10^{-2} at $k=31$

$$\frac{|e^{-5.5} - 1 \times 10^{-2}|}{e^{-5.5}} = 1.4469$$

- all are less accurate/slower converging than the $e^{5.5}$ case

PERFORATED

e.) to compute $e^{-5.5}$ more accurately:

$$\frac{1}{e^{5.5}} \rightarrow \frac{1}{244.71} = 0.0040865$$

$$\frac{|e^{-5.5} - .0040865|}{e^{-5.5}} = 6.6419 \times 10^{-5}$$

↑ orders of magnitude less than error of other methods explored

Charles Maher

3.)

a.) let $x, n \in \mathbb{R}(p,q)$

i.) $x \to fl(x \cdot x) = x^2(1+\varepsilon) \to fl(x^3(1+\varepsilon)) = x^3(1+\varepsilon)^2 \sim x^3(1+2\varepsilon)$

$\ldots \to x^n(1+(n-1)\varepsilon)$

$$|\varepsilon_{x^n}| \leq (n-1)\,eps$$

ii.) $fl(\ln(x)) = \ln(x)(1+\varepsilon)$

$fl(n\ln(x)(1+\varepsilon)) = n\ln(x)(1+\varepsilon)(1+\varepsilon) \sim n\ln(x)(1+2\varepsilon)$

$fl(e^{n\ln(x)(1+2\varepsilon)}) = e^{n\ln(x)}\, e^{2n\ln(x)\varepsilon}(1+\varepsilon)$

$\sim e^{n\ln(x)}(1+2n\ln(x)\varepsilon)(1+\varepsilon)$

$= e^{n\ln(x)}(1+(2n\ln(x)+1)\varepsilon) + \mathcal{O}(\varepsilon^2)$

$$|\varepsilon_{e^{n\ln(x)}}| \leq (2n\ln(x)+1)\,eps$$

→ if $(n-1) \leq 2n\ln(x)+1$ then repeated multiplication is less error prone than the exp-log method

b.)

i.) $x^{(a+\varepsilon_a)} \to x^a + \varepsilon_a x^a \log(x)$

$= x^a(1+\varepsilon_a \log(x))$

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

Scenario 1 is not likely to ever have substantially high error, but scenario 2 can high error if $x \to 0$

Charles Maher

4.)

a.) $\text{cond}(f(x)) = \left| \frac{x\left(\partial_x(1-e^{-x})\right)}{1-e^{-x}} \right| = \left| \frac{x(e^{-x})}{1-e^{-x}} \right|$

$= \left| \frac{x}{e^x - 1} \right|$

↳ problem condition

$\frac{d}{dx}\left(\frac{x}{e^x-1}\right) = -\frac{e^x(x-1)+1}{(e^x-1)^2}$ → derivative is negative semi-definite

→ $\max \left| \frac{x}{e^x-1} \right|$ on $[0,1]$ is at 0

$\lim_{x\to 0} \left| \frac{x}{e^x-1} \right| = \lim_{x\to 0} \left| \frac{1}{e^x} \right| = 1$

on $[0,1]$ $(\text{cond } f)(x) \leq 1$ □

b.) $-x \in \mathbb{R}(p,q) \to fl(e^{-x}) = e^{-x}(1+\varepsilon_{exp}) \to fl(1-e^{-x}(1+\varepsilon_{exp}))$

↳ $fl\left(1-e^{-x}\left(1+\frac{e^{-x}}{1-e^{-x}}\varepsilon_{exp}\right)\right) = 1-e^{-x}\left(1+\frac{e^{-x}}{1-e^{-x}}\varepsilon_{exp}\right)(1+\varepsilon_{rd})$

$\sim 1-e^{-x}\left(1+e^{-x}/_{1-e^{-x}}\,\varepsilon_{exp}+\varepsilon_{rd}\right)$

$= 1-e^{-x}+(1-e^{-x})\varepsilon_{rd}+e^{-x}\varepsilon_{exp} = f_A(x)$

$f(x_A) = 1-e^{-x_A}$

→ setting equal and solving for x_A:

$(1-e^{-x})(1+\varepsilon_{rd}) + e^{-x}\varepsilon_{exp} = 1-e^{-x_A}$

$e^{-x}+\varepsilon_{rd}(1-e^{-x}) + e^{-x}\varepsilon_{exp} = -e^{-x_A}$

$-\ln\left(e^{-x}(1+\varepsilon_{rd}+\varepsilon_{exp}+\varepsilon_{rd}e^{x})\right| = x_A$

$|x_A - x| = \ln(1+(e^x-1)\varepsilon_{rd}+\varepsilon_{exp}|$

$\leq |(1+e^x)\varepsilon_{rd}| + |\varepsilon_{exp}| \leq |2+e^x|\,eps$

$\text{cond}(A) = \frac{1}{eps}\,\frac{|2+e^x|}{|x|}\,eps = \frac{2+e^x}{x}$

$\frac{2+e^x}{x} > 1 \to 2+e^x > x \to e^x > x-2$ (always true on $[0,1]$) $\Rightarrow \text{cond}(A) > 1$ on $[0,1]$

c.) (plot in separate pdf)
poor conditioning result of subtraction

d.)

I) 1 bit → $x = e^{-x}$; $y = 1$

$2^{-1} < 1 - \frac{x}{4} \rightarrow \frac{1}{2} < 1 - e^{-x}$

$\hookrightarrow e^{-x} < 1/2$ $\quad x > -\ln(1/2)$

$x > .6931$

(at least, 0.6931)

II) 2 bits →

$\frac{1}{4} < 1 - e^{-x} \rightarrow e^{-x} < \frac{3}{4}$

$x > -\ln(3/4)$

$x > .2877$

(at least, 0.2877)

III) 3 bits → $x > -\ln(7/8)$

$x > 0.1335$

(at least, 0.1335)

IV) 4 bits → $x > -\ln(15/16)$

$x > 0.0645$

(at least, 0.0645)

e.) (rel. error in output) $\leq$ (cond f)(x) (eps + (cond A)(x) eps)

I) $\varepsilon_y \leq \left| \frac{x}{e^x - 1} \right| \left(\text{eps} + \frac{2 + e^x}{x} \text{eps} \right)$

$\leq .6931 \text{ eps} + 4 \text{ eps} = 4.6931 \text{ eps}$

$\varepsilon_y \leq 4.6931 \text{ eps}$

II) $\varepsilon_y \leq 10.863 \text{ eps}$

III) $\varepsilon_y \leq 22.9347 \text{ eps}$

IV) $\varepsilon_y \leq 46.9681 \text{ eps}$

f.) $f(x) = 1 - e^{-x}$

$= \frac{e^x - 1}{e^x} \approx \frac{(1 + x + \frac{x^2}{2} + \cdots] - 1}{1 + x + \frac{x^2}{2} + \cdots}$

$= \underbrace{\frac{x + \frac{x^2}{2}}{1 + x + \frac{x^2}{2}}}$

- no subtraction → likely better conditioned but more "expensive"

Charles Maher

5.)

at 10^{16} we observe that the limit becomes one, (which is repeated for 10^{17} (i.e. n_{stop} is 10^{17})) which is because at 10^{16} $\frac{1}{10^{16}}$ becomes too small to store (underflow). Thus; $\frac{1}{10^{16}}$ is stored as 0, so the limit becomes $(1+0)^{10^{16}}$, which is just 1.

6.) Charles Maher
(plot is in notebook)

remaining value for S2 are e^0, e^1, e^2

S3 e^0, e^2

S4 e^0

S5 $e^{1/4}, e^{1/2}, e^{3/4}, e^1, e^{5/4}, e^{3/2}$

S0 $e^{7/4}, e^2, e^{9/4}$

→ powers are those representable with $\overbrace{S3-n}^{\text{from IEEE double precision}}$ bits of precision, where n is the number of square roots / squares taken

This is because each $\sqrt{x}$ removes a bit of precision for each iteration, and so any # above something representable with S3−n bits will get floored because the less significant bits wind up getting lost.

We get e to these powers of this because when we square these functions back up to their "original value" we can think of the following:

we now have numbers written like:

$1 + (\text{something tiny}^{5} \text{ with } S3-n \text{ bits of mantissa})$

after the sqrts we have something like:

$e^{1+\epsilon} \xrightarrow{\text{squaring}} e^x$ → but now x has very little precision, i.e. can only take the values listed above

□.

Charles Maher

7.)

a.) (coeff in attached notebook)

b.) I) Newton largest root: 19.999999495714 18

II) Numpy "roots": 20.00054209

c.) $\delta = 10^{-8}$

I) 9.585389646516598

II) 20.6475355 ± 1.18691242 i

$\delta = 10^{-6}$

I) 7.752713003402644

II) 23.1496169 ± 2.7409845 i

$\delta = 10^{-4}$

I) 5.969334849605957

II) 28.40021241 ± 6.51043422 i

$\delta = 10^{-2}$

I) 5.469597915093453

II) 38.47818362 + 20.83432359 i

(all roots also in attached notebook)

(Newton based method only detects real roots, so largest real root is taken)

d.) roots 16, 17 have collapsed to the complex pair:

16.73096403 ± 2.81265595 i

e.)

i.) $(\text{cond}\,\Omega_k)(\vec{a}) = \sum_{l=0}^{n-1} (\Gamma_{kl})(\vec{a})$

$= \left| \frac{x f'(x)}{f(x)} \right| = \left| \frac{a_l \frac{\partial \rho_k(\vec{a})}{\partial a_l}}{\rho_k(\vec{a})} \right| = \left| \frac{a_l \Omega_k^{l-1}}{\rho'(\Omega_k)} \right|$

ii.) (calculations in mathematica notebook attached)

14: 1.33297×10^9

16: 2.40751×10^9

17: 1.9044×10^9

20: 4.30999×10^{17}

extremely poorly conditioned, seems to be worse conditioned for intermediate roots (<20), so the problem itself is prone to significant error

iii.) unlikely because the problem itself is poorly conditioned, so the framework itself is flawed

Charles Maher

8.)

a.) $y_{n+1} = e - (n+1)y_n \rightarrow y_n = \frac{e - y_{n-1}}{n+1} \rightarrow \boxed{y_{n-1} = \frac{e - y_n}{n}}$

$y_{n-1} = \frac{e - y_n}{n}$; $y_{n-2} = \frac{e - y_{n-1}}{n-1} = \frac{e}{n-1} - \frac{e}{n(n-1)} - \frac{y_n}{n(n-1)}$,

$y_{n-3} = \frac{e}{n-2} - \frac{e}{(n-2)(n-1)} + \frac{e}{n(n-1)(n-2)} + \frac{y_n}{n(n-1)(n-2)}$

$y_{N-k} = \ldots \pm \frac{y_N}{N(N-1)\cdots(N-k)}$

$\rightarrow \pm \frac{y_N\, k!}{N!}$

$(\text{cond } g_k) = \left|\frac{y_N}{y_k} g'(y_N)\right| = \left|\frac{y_N}{y_k}\frac{k!}{N!}\right|$

$\frac{y_N}{y_k}$ is at greatest ~ 2 if $N > k$

$\rightarrow \boxed{|\text{cond } g_k| < \left|\frac{k!}{N!}\right|}$

b.) $\varepsilon_y = |\text{cond}| \, \varepsilon_x$

$\varepsilon_y = 2$, $\varepsilon_x = 1$

$\varepsilon = |\text{cond}| \rightarrow \varepsilon \geq \frac{k!}{N!}$ $\boxed{\frac{k!}{\varepsilon} \leq N!}$

c.) Let $k = 20$, $\varepsilon = 2^{-53}$

$20!\, 2^{53} \approx 2.2 \times 10^{34}$ $\underbrace{2.2 \times 10^{34} \leq N!}$

(closest two, but less than 32!)

$\boxed{N \geq 32}$

d.) (notebook w/ code attached)

↳ result = 0.1238038 Wolfram Alpha (to same amount of sig figs): 0.1238038

error: $\frac{|1.1238038 - .1238038|}{.1238038} = 0$