Sam Moore

(1) $x = \pm\left(\sum_{\ell=1}^{\infty} b_{-\ell} 2^{-\ell}\right) \cdot 2^e \quad , \quad b_i \in (0,1)$

If $b_{p+1} = 0$, the rounding is akin to truncation:

$$x - \text{trunc}(x) = \pm\left(\sum_{p+1}^{\infty} b_{-\ell} 2^{-\ell}\right) \cdot 2^e$$

$$\Rightarrow \max|x - \text{trunc}(x)| = \left|\sum_{\ell=p+1}^{\infty} 2^{-\ell}\right| \cdot 2^e \qquad [b_{p+1}=0, \text{ rest} = 1]$$

$$= \left(2^{-(p+2)} + 2^{-(p+3)} + \dots\right) \cdot 2^e$$

$$\Rightarrow \max\left|\frac{x - \text{trunc}(x)}{x}\right| = \frac{2^{-p-1} \cdot 2^e}{2^{-1} \cdot 2^e} = 2^{-p}$$

$$\Rightarrow \left|\frac{x - \text{trunc}(x)}{x}\right| \leq 2^{-p}$$

If $b_{p+1} = 1$, add 1 to $b_p$, and then truncate.

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

$$\Rightarrow x^* = \pm\left(b_{-1}2^{-1} + b_{-2}2^{-2} + \dots + (b_p+1)2^{-p} + b_{(p+1)}2^{-(p+1)} + \dots\right) \cdot 2^e$$

$$\Rightarrow \max|x^* - \text{trunc}(x^*)| = \left|\sum_{\ell=p+1}^{\infty} 2^{-\ell}\right| \cdot 2^e$$

$$= (2^{-p} - 2^{-p-1})(2^{-1} + 2^{-2} + 2^{-3} + \dots)$$

$$\Rightarrow \max\left|\frac{x^* - \text{trunc}(x^*)}{x^*}\right| = \frac{(2^{-p} - 2^{-p-1})2^e}{2^{-1} \cdot 2^e} = 2^{-p}$$

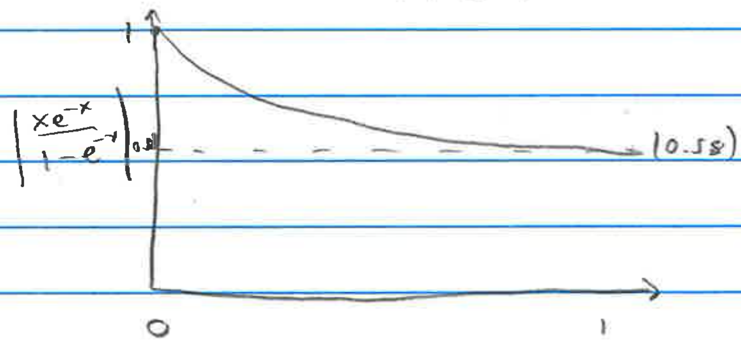
Thus $\left|\frac{x - rd(x)}{x}\right| \leq 2^{-p}$ (by combining both cases)

4 a) $(\text{cond } f)(x) \equiv \left\| \frac{x f'(x)}{f(x)} \right\|$,

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

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

Noting that $\lim_{x \to 0} \left| \frac{x e^{-x}}{1-e^{-x}} \right| = 1$, then it is clear that on the interval $[0,1)$, $\left| \frac{x e^{-x}}{1-e^{-x}} \right| < 1$




$\Rightarrow$ the problem is well-conditioned.

This method is roughly correct, but wasn't working for me

b) First, note that negation yields no error. Then, applying step 2, $f_A(x) = 1.0 - e^{-x}(1 + \varepsilon_{exp})$, where we note the error yielded by the exp function, (and so that as $1.0 \in \mathbb{R}(p,q)$, there is no round-off error introduced here).

From lecture: $|\varepsilon_{x+y}| \le \left| \frac{x}{x+y} \right| |\varepsilon_x| + \left| \frac{y}{x+y} \right| |\varepsilon_y|$

in this case $\Rightarrow |\varepsilon_{x+y}| \le \left| \frac{e^{-x}}{1-e^{-x}} \varepsilon_{exp} \right|$

Then: $\left(1.0 + \frac{e^{-x}}{1-e^{-x}} \varepsilon_{exp}\right)(1 + \varepsilon_{rd}) = 1 - e^{-x_A}$

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

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

$\Rightarrow x_A = \ln(1 + \delta) \simeq \delta + O(\delta^2)$ assuming $\delta$ small

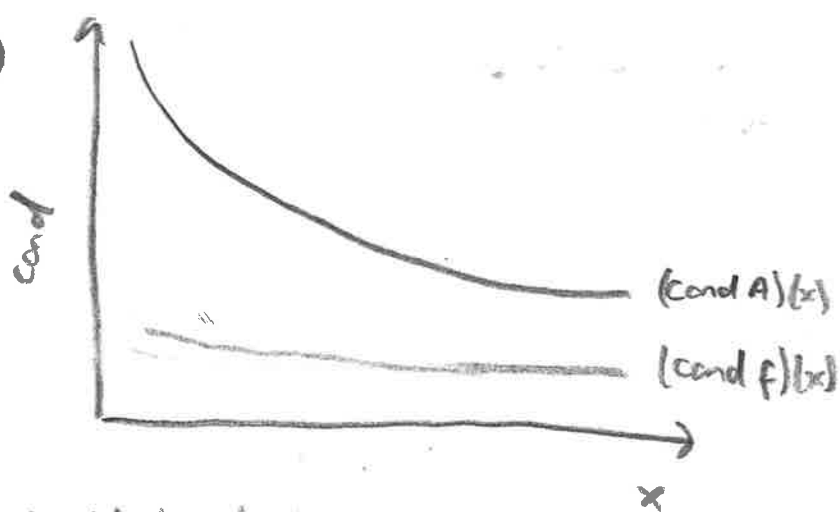
4b) $f_A(x) = [1-e^{-x}(1+\varepsilon)](1+\varepsilon)$

$\simeq (1-e^{-x})\left(1-\frac{e^{-x}}{1-e^{-x}}\varepsilon\right) \simeq (1-e^{-x})\left(1+\frac{\varepsilon}{1-e^{-x}}\right)$

$\Rightarrow |f(x_A)-f(x)| = f(x)\frac{\varepsilon}{1-e^{-x}} \simeq |f'(x)||x-x_A|$

$\Rightarrow \left|\frac{x-x_A}{x}\right| \simeq \left|\frac{f(x)}{xf'(x)}\frac{\varepsilon}{1-e^{-x}}\right| \longleftarrow$ or something close to this

$\Rightarrow (\text{cond } A)(x) \simeq \frac{1}{\varepsilon}\left|\frac{x-x_A}{x}\right| \simeq \frac{e^x}{x}$

c) 


(Using Desmos for convenience)

Note that when x is small, $1-e^{-x}$ gives a large error [as $x \to 0$, $e^{-x} \to 1$, and the subtraction of two numbers that are very close gives a high error].

d) $1-e^{-x} < 2^{-b} \Rightarrow x = \ln(1-2^{-b})$

rel-err $= \frac{\varepsilon}{1-e^{-x}} = 2^b\varepsilon$

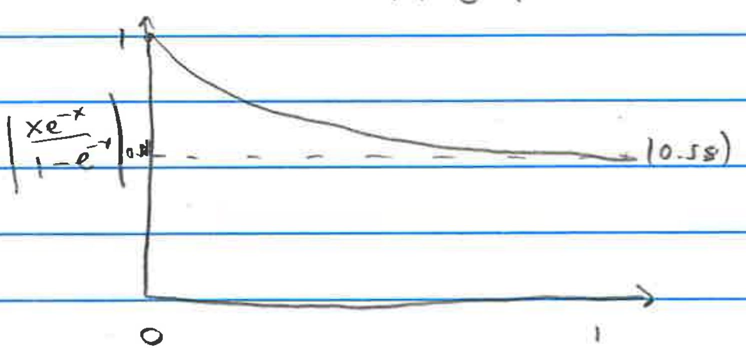
$n=1 \longrightarrow x = 0.693$, rel-err $= 2^{-51}$
$n=2 \longrightarrow x = 0.288$, rel-err $= 2^{-50}$
$n=3 \longrightarrow x = 0.134$, rel-err $= 2^{-49}$
$n=4 \longrightarrow x = 0.065$, rel-err $= 2^{-48}$

4 a) $(\text{cond } f)(x) \equiv \left\| \frac{x f'(x)}{f(x)} \right\|$,

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

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

Noting that $\lim_{x \to 0} \left| \frac{xe^{-x}}{1-e^{-x}} \right| = 1$, then it is clear that on the interval $[0,1)$, $\left| \frac{xe^{-x}}{1-e^{-x}} \right| < 1$




$\Rightarrow$ the problem is well-conditioned.

This method is roughly correct, but wasn't working for me.

b) First, note that negation yields no error. Then, applying step 2, $f_A(x) = 1.0 - e^{-x}(1+\varepsilon_{exp})$, where we note the error yielded by the exp function, (and so that as $1.0 \in \mathbb{R}(p,q)$, there is no round-off error introduced here).

From lecture: $|\varepsilon_{x+y}| \leq \left|\frac{x}{x+y}\right| |\varepsilon_x| + \left|\frac{y}{x+y}\right| |\varepsilon_y|$

in this case $\Rightarrow |\varepsilon_{x+y}| \leq \left|\frac{e^{-x}}{1-e^{-x}}\right| \varepsilon_{exp}$

Then: $\left(1.0 + \frac{e^{-x}}{1-e^{-x}} \varepsilon_{exp}\right)(1 + \varepsilon_{rd}) \approx 1 - e^{-x_A}$

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

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

$\Rightarrow x_A = \ln(1+\delta) \simeq \delta + O(\delta^2)$ assuming $\delta$ small

4b) $f_A(x) = [1 - e^{-x}(1+\varepsilon))](1+\varepsilon)$

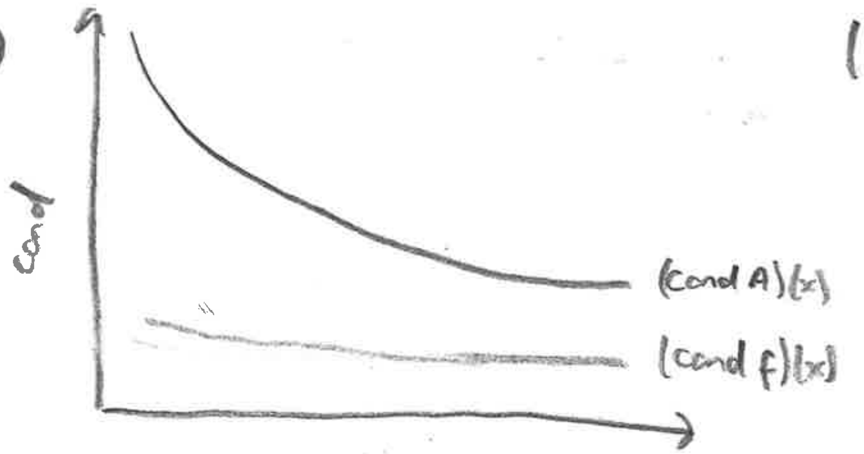
$\simeq (1-e^{-x})\left(1 - \frac{e^{-x}}{1-e^{-x}}\varepsilon\right) \simeq (1-e^{-x})\left(1 + \frac{\varepsilon}{1-e^{-x}}\right)$

$\Rightarrow |f(x_A) - f(x)| \simeq f(x)\frac{\varepsilon}{1-e^{-x}} \simeq |f'(x)||x - x_A|$

$\Rightarrow \left|\frac{x - x_A}{x}\right| \simeq \left|\frac{f(x)}{x f'(x)} \frac{\varepsilon}{1-e^{-x}}\right|$ ⟵ or something close to this

$\Rightarrow (\text{Cond } A)(x) \simeq \frac{1}{\varepsilon}\left|\frac{x - x_A}{x}\right| \simeq \frac{e^x}{x}$

c)




(Using Desmos for convenience)

Note that when x is small, $1-e^{-x}$ gives a large error [as $x \to 0$, $e^{-x} \to 1$, and the subtraction of two numbers that are very close gives a high error].

d) $1 - e^{-x} < 2^{-b} \Rightarrow x = \ln(1 - 2^{-b})$

$\text{rel-err} = \frac{\varepsilon}{1-e^{-x}} = 2^b \varepsilon$

$n=1 \longrightarrow x = 0.693$, rel-err $= 2^{-51}$

$n=2 \longrightarrow x = 0.288$, rel-err $= 2^{-50}$

$n=3 \longrightarrow x = 0.134$, rel-err $= 2^{-49}$

$n=4 \longrightarrow x = 0.065$, rel-err $= 2^{-45}$

b) $x^a = e^{a \ln x}$

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

$\simeq x^a(1 + a\varepsilon_a \ln x)$

$\Rightarrow \varepsilon = (a \ln x)\,\varepsilon_a \longrightarrow$ if $a \ln x \gg 1$, $\varepsilon$ is significant.

ii) $\ln x = \ln(x(1+\varepsilon_x)) = \ln x + \varepsilon_x$

$\Rightarrow e^{a \ln x} = \exp[a\ln x + a\varepsilon_x] = x^a(1 + a\varepsilon_x)$

$\Rightarrow \varepsilon = a\varepsilon_x \longrightarrow$ if $a \gg 1$, $\varepsilon$ is significant.

7e) i) $p(x) = \sum_{k=0}^{n} a_k x^k = a_0 + a_1 x + a_2 x^2 + \dots + a_{n-1} x^{n-1} + x^n$

Let $p(x)$ be perturbed s.t. $p_{per}(x) = a_0 + \dots + (a_i + \delta a_i)x^i + \dots + a_{n-1}x^{n-1} + x^n$

$\Rightarrow p_{per}(x+\delta x) = a_0 + \dots + (a_i + \delta a_i)(x+\delta x)^i + \dots + (a_{n-1})(x+\delta x)^{n+1} + (x+\delta x)^n$

$\Rightarrow p_{per}(x+\delta x) = p(x+\delta x) + \delta a_i (x+\delta x)^i$

$\Rightarrow p_{per}(x-\delta x) - p(x+\delta x) = \delta a_i (x+\delta x)^i$

$$\Rightarrow \left| \frac{p(x+\delta x) - p(x)}{\delta x} \right| = \left| \frac{a_i (x+\delta x)^i}{\delta x} \right|$$

$\delta x \to 0$

$$\Rightarrow |p'(x)| = \left| \frac{a_i x^i}{\delta x} \right|$$

From $K \left| \frac{\delta a_i}{a_i} \right| = \left| \frac{\delta x}{x} \right| \quad \Rightarrow K_i = \left| \frac{a_i x^{i-1}}{p'(x)} \right|$

Now $(\text{Cond } \Omega_k)(a_i) = \left| \frac{a_i \Omega_k^{i-1}}{p'(\Omega_k)} \right|$

$$\Rightarrow (\text{cond } \Omega_k)(\vec{a}) = \sum_{i=0}^{n-1} \left| \frac{a_i \Omega_k^{i-1}}{p'(\Omega_n)} \right|$$

ii) See code.

iii) I doubt that it could, as these condition numbers are very large, so the problem isn't well-conditioned

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

$$\Rightarrow \boxed{y_n = \frac{e - y_{n+1}}{n+1}}$$

Then for the mapping $g_k(y_N) = y_k$ :

$$g_N(y_N) = y_N$$
$$g_{N-1}(y_N) = \frac{e - y_N}{N}$$
$$g_{N-2}(y_N) = \frac{e(N-1) + y_N}{N(N-1)}$$

One could formally prove by induction that

$$g_{N-a}(y_N) = \frac{e(N-1)^{a-1} + y_N}{(N-a-1)!}$$

[though the pattern is rather obvious]

Furthermore, $g_k'(y_N) = \frac{(-1)^{N-k} k!}{N!}$

$$\Rightarrow (\mathrm{Cond}\, g_k)(y_N) = \left| \frac{y_N k!}{y_k N!} \right|$$

As $k < N$, $\max \left| \frac{y_N}{y_n} \right| = 1$

$$\Rightarrow \boxed{(\mathrm{Cond}\, g_k)(y_N) = \frac{k!}{N!}}$$

b) ~~Let $\varepsilon_N$~~

b) $\varepsilon_N = 1$, $\left|\frac{\varepsilon_k}{\varepsilon_N}\right| = \left|\frac{y_N g'_k}{y_k}\right| \Rightarrow \varepsilon_k = \frac{g_N g'_k}{y_k}$

This is well-conditioned if $\varepsilon_k (= \varepsilon) \leq 1$

$\Rightarrow \frac{y_N g'_k}{y_k} = 1$ for minimum

$\Rightarrow \frac{k!\, y_N}{N!\, y_k} = 1$ Then noting that $\frac{y_N}{y_k} \leq 1$, for min. $\frac{y_N}{y_k} = 1$

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

c) $\varepsilon = \varepsilon ps$, $k = 20$

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

By inspection: $31! = 8.22 \times 10^{33}$
$32! = 2.63 \times 10^{35}$

$\Rightarrow \boxed{N = 32}$

d) $x = 20$, $N = 32$

yields $y_{20} = 0.1238038\ldots$ [see code]

Using Mathematica, $y_{20} = \int_0^1 e^x x^{20}\, dx \simeq 0.112380$
(not accurate enough to see relative error)

Using quad function, see code

rel_error = $4.48 \times 10^{-16}$

Sam Moore