① <u>Rounding Error:</u>

For any bases we have to cases: the error for rounding up and the one for rounding down

e.g. $1.23499...$ vs $1.23\cancel{5}0...$ to the 2nd decimal are
$\hookrightarrow 1.23 \rightarrow \epsilon = 0.005$ $\quad$ $\hookrightarrow 1.24 \rightarrow \epsilon = 0.005$
the worst case scenarios for each operation

Meaning, for rounding up the error is largest if the ~~last~~ $(p+1)^{th}$ digit is 5 and the rest is zero and for rounding down it's worst when $(p+1)$ digit is 4 and rest is 9

In general, if β is the base. the round up error is given by

$$\epsilon_{up} = \left| x - x_{rd_{up}} \right| = \left| \frac{\beta}{2} \beta^{-(p+1)} \beta^{q} \right| = \left| \frac{1}{2} \beta^{-p} \beta^{q} \right| = \underline{\underline{\frac{1}{2} \beta^{q-p}}}$$

Round down:

$$\epsilon_{down} = \left| x - x_{rd_{down}} \right| = \left| \left[\left(\tfrac{\beta}{2} - 1\right) \underset{\text{digit}}{\beta^{-(p+1)}} + \sum_{\ell=p+2}^{\infty} (\beta-1) \beta^{-\ell} \right] \beta^{q} \right|$$

e.g. 9 base_{10}, 1 base_2

$$= \left| \left[\frac{1}{2}\beta^{-p} - \beta^{-(p+1)} + \underbrace{\left(\sum_{\ell=p+2}^{\infty} \beta^{-\ell} \right)}_{\text{geometric series}} (\beta-1) \right] \beta^{q} \right|$$

$$= \left| \left[\frac{1}{2}\beta^{-p} - \beta^{-(p+1)} + \frac{1}{\beta^{p+2}} \frac{\cancel{\beta-1}}{\cancel{\beta}(\cancel{\beta-1})} \right] \beta^{q} \right|$$

$$= \left| \left[\frac{1}{2}\beta^{-p} \cancel{- \beta^{-(p+1)}} \cancel{+ \beta^{-(p+1)}} \right] \beta^{q} \right|$$

$$= \frac{1}{2}\beta^{q-p}$$

Since the error for both is the same we can conclude that the max error for rounding $\underline{\underline{\epsilon_{rd} = \frac{1}{2}\epsilon_{tr} \Leftrightarrow \frac{1}{2}\beta^{q-p}}}$

3(a) (i) $fl(x^n) = fl(x \circ fl(x \circ \cdots = x^n(1+\varepsilon)^n$

$$\approx x^n(1+n\epsilon) + O(\epsilon^2)$$

Relative error: $\epsilon_r = \left| \frac{x_n^n - x^n(1+n\epsilon)}{x_n} \right| \leq \underline{\underline{|n\epsilon|}}$

3(a) (ii)

$$e^{n \ln(x)(1+\epsilon_\ell)(1+\epsilon_N)} (1+\epsilon_e)$$

$$e^{n \ln(x)(1+\epsilon_\ell+\epsilon_N)} (1+\epsilon_e)$$

$$e^{n \ln(x)}\, e^{n \ln(x)\epsilon_\ell}\, e^{n \ln(x)\epsilon_N} \quad (1+\epsilon_e)$$

series expansion

$$e^{n \ln(x)}(1+n\ln(x)\epsilon_\ell)(1+n\ln(x)\epsilon_N)(1+\epsilon_e) \quad \left\{ \begin{array}{l}\text{omit}\\ \text{higher}\\ \text{Order}\\ \text{terms}\end{array}\right.$$

$$e^{n \ln(x)}(1+n\ln(x)(\epsilon_\ell+\epsilon_n))(1+\epsilon_e)$$

$$e^{n \ln(x)}(1+\underbrace{n\ln(x)(\epsilon_\ell+\epsilon_n)+\epsilon_e}_{\epsilon})$$

Relative Error:

$$\epsilon = \left| 1 - 1 - n\ln(x)(\epsilon_\ell+\epsilon_n) + \epsilon_e \right|$$

$$= \underline{\underline{\left| n\ln(x)(\epsilon_\ell+\epsilon_n)+\epsilon_e \right|}}$$

3 (b) (i) $fl(x^a) = x^{a(1+\epsilon_a)}(1+\epsilon_p)$

$$= x^a x^{a\epsilon_a}(1+\epsilon_p)$$

$$= x^a e^{a\epsilon_a \ln x}(1+\epsilon_p) \quad \downarrow \text{Taylor}$$

$$= x^a[1+a\epsilon_a\ln(x)](1+\epsilon_p) \quad \text{omit higher order eps}$$

$$= x^a[1+\underbrace{a\epsilon_a\ln(x)+\epsilon_p}_{\epsilon}] \quad \text{omit —"—}$$

Relative Error is

$$\epsilon = a\ln(x)\,\epsilon_a \;(+\epsilon_p)$$

not given/asked for in question but nonnegligible?

Problems could occur if either a is large or x is large. Also, x can ~~not~~ be negative but → Complex and $x=0$ is also an issue.

3(b) ii) $fl(x^a) = [x(1+\epsilon_x)]^a \; (1+\epsilon_p)$? → Power Error

$$[= [x + x\epsilon_x]^a]$$

$$= x^a (1+\epsilon_x)^a$$

$$= x^a e^{a \ln(1+\epsilon_x)} \quad \text{(Logarithmic Series } \cancel{O(\epsilon_x^2)} \approx 0\text{)}$$

$$\approx x^a e^{a\epsilon_x} \quad \text{(Exponential Series } \cancel{O(\epsilon_x^2)} \approx 0\text{)}$$

$$\underline{\underline{= x^a [1 + a\epsilon_x]}}$$

$$= x^a [1 + \underbrace{a\epsilon_x + \epsilon_p}_{\epsilon}] \quad \text{if power error is regarded.}$$

Relative error:

$$\underline{\underline{\epsilon = a\epsilon_x + \epsilon_p}}$$

Here, the error could only become substantial if a is large. Otherwise this is fine, I guess.

4 (a) Find Condition of $f(x) = 1 - e^{-x}$

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

Expand to series

$$= \left| \frac{x}{e^{x}-1} \right| = \left| \frac{x}{\cancel{1}+\cancel{1}+\frac{x}{1}+ \cdots +} \right| \quad \Big| : \text{cancel } x$$

$$(\text{cond } f)(x) = \left| \frac{1}{1+\frac{x}{1}+\frac{x^2}{2!}+\cdots} \right| \geq 1$$

I.e., on $x \in [0,1]$ $\underline{\underline{(\text{cond } f)(x) \leq 1}}$

4(b) $f(x_A) = f_A(x)$, Where $f(x) = 1 - e^{-x_A}$

$$f_A(x) = \left[1 - e^{-x}(1+\epsilon_1) \right](1+\epsilon_1)$$

(c) Find ϵ error in f_A

↳ See Code

$$f_A(x) = \left[1 - e^{-x}(1-\epsilon_1) \right](1-\epsilon_2)$$
$$= 1 - e^{-x} - \epsilon_1 e^{-x} + \epsilon_2 - \epsilon_2 e^{-x}$$
$$= (1-e^{-x}) \Big\{ 1 + \underbrace{\frac{\epsilon_2 - e^{-x}(\epsilon_1 + \epsilon_2)}{1-e^{x}}}_{\epsilon_A} \Big\}$$

Now, find condition of A

$$f(x_A) = f_A(x)$$
$$1 - e^{-x_A} = \left[1 - e^{-x} \right](1+\epsilon_A)$$
$$\cancel{1} - e^{-x_A} = \cancel{1} - e^{-x} + \epsilon_A - \epsilon e^{-x} \quad | \cdot e^{x}$$
$$-e^{x - x_A} = -1 + \epsilon_A e^{+x} - \epsilon \quad | \wedge(-1)$$
$$e^{x_A - x} = (1 + \epsilon_A - \epsilon_A e^{x})^{-1} \quad | \ln$$
$$x_A - x = \ln \left\{ \frac{1}{1 + \epsilon_A (1 - e^{x})} \right\}$$
$$x_A - x = -\ln(1 + \epsilon_A - e^{x}) \qquad \text{Log series} \cdot \ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} \cdots$$
$$= -\epsilon(1 - e^{x}) + O(\epsilon_A^2)$$
$$x_A - x = \epsilon(e^{x} - 1)$$

Then, we need to find the max
smallest sensible number

$$\max_{x\in[0,1]} \left| \frac{x_A - x}{x} \right| = \left\{ \max(x_A - x)\Big|_{x \le 1} = \frac{\epsilon_A(e^{x}-1)}{x} \rightarrow \underline{\underline{x=0}} \right.$$ unbounded at

$$\Rightarrow (\text{cond } A)(x) = \frac{\epsilon_A \left| \frac{x_A - x}{x} \right|}{\epsilon} = \underline{\underline{\frac{e^{x}-1}{x} > 1}} \qquad x \in [0,1]$$

Hence, (cond A)(x) is always larger than 1 on [0,1]

4 d We have $(1+e^{-y})$

$x \quad y$

$$2^{-b} \leq 1 - \frac{y}{x} \leq 2^{-a}$$

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

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

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

$$+x \geq -\ln\left(1 - 2^{-b}\right)$$

$$x \geq \ln\left((1-2^{-b})^{-1}\right)$$

Willing to lose one bit $b = 1$

$$x \geq \ln\left(\left(1 - \frac{1}{2}\right)^{-1}\right) = \ln 2$$

2 bits $b = 2$

$$x \geq \ln\left((1 - 1/4)^{-1}\right) = \ln \frac{4}{3}$$

$\cancel{x \geq \ln(}$

~~[illegible]~~ 3 bits $b = 3$

$$x \geq \ln\left((1 - 1/8)^{-1}\right) = \ln \frac{8}{7}$$

4 bits $b = 4$

$$x \geq \ln\left((1 - 1/16)^{-1}\right) = \ln \frac{16}{15}$$

4 (e) The upper bound for the relative Error is given by the $(\mathrm{cond}\, A)(x)\Big|_{x=x_0}$, where x_0 is/are the value/s from from (d)

1 bits: $(\mathrm{cond}\, A)(\ln 2) = \frac{e-1}{\ln 2} \approx \underline{\underline{2{,}47896}}$

2 bits: $(\mathrm{cond}\, A)(\ln \frac{4}{3}) = \frac{e-1}{\ln 4/3} \approx \underline{\underline{5.97285}}$

3 bits: $(\mathrm{cond}\, A)(\ln \frac{8}{7}) = \frac{e-1}{\ln 8/7} \approx \underline{\underline{12.868}}$

4 bits: $(\mathrm{cond}\, A)(\ln \frac{16}{15}) = \frac{e-1}{\ln \frac{16}{15}} \approx \underline{\underline{26.62413}}$

4 f) $f(x) = 1 - e^{-x}$ can be rewritten into

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

This again can be rewritten into Another function, i.e, ~~the~~ using Taylor expansion

$$f(x) = \frac{x + \frac{x^2}{2!} + \cdots}{e^x}$$

this way we omit the subtraction and hence the source of largest error.

(6.) Explanation

Mantissa

$1.\boxed{0\,0 \cdots - \quad \cdots \quad 0\,d}$

Double Precision has 52-bits in mantissa

When we take the square root of a binary number say

$(4)_{10} = (100)_2 = 2^2 \Rightarrow \sqrt{2^2} = 2^{2/2}$ Same again 52 times

↳ bit shifts to right in mantissa

Meaning, we will be left with with $1.\,0 \cdots 0d$, ie.

1 significant bit

Meaning, when taking the square root and the squaring 52 times we are left with:

$$(1+2^{-52})^{2^{52}}$$

Looking back at exercise 5, there is striking similarity and we find that $(1+2^{-52})^{2^{52}} \approx \lim_{n \to \infty} \left(1+\frac{1}{n}\right)^n$

Inspecting more closely we find that values larger that 2^2 say 2^3 result in e^2 eg.

The reason is simply that their significant digit is on exponent larger

giving $(1+2^{-51})^{2^{52}} = (1+2^{-51})^{2^{51} \cdot 2} \approx \lim_{n \to \infty} \left(1+\frac{1}{n}\right)^2 = e^2$

For comparison, square-rooting and squaring 51 times leaves one more significant digit that is $1\boxed{.00 \cdots dd}$

and the result is $(1+2^{-50}+2^{-52})$. Hence, the plot will output more intermediate values between $e^0\, e^1\, e^2 \ldots$

7(c) The $(\text{cond}\ \Omega_k)(\underline{a}) = \sum_{l}^{n-1} (\Gamma_{kl})(\vec{a})$

$$\Gamma_{kl} = \left| \frac{a_l \frac{\partial \Omega_k}{\partial a_l}}{\Omega_k} \right| = \left| \frac{a_l \frac{\partial \Omega_k}{\partial p(\Omega_k)} \frac{\partial p(\Omega_k)}{a_l}}{\Omega_k} \right|$$

$$= \left| \frac{a_l \frac{1}{p'(\Omega_k)} (\Omega_k)^l}{\Omega_k} \right|$$

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

So, the condition is given by

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

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

$$(\text{cond } g_k)(y_N) = \left| \frac{y_N \frac{\partial g_k}{\partial y_N}}{g_k} \right|$$

$$= \left| \frac{y_N}{y_k} \frac{\partial y_k}{\partial y_{k+1}} \frac{\partial y_{k+1}}{\partial y_{k+2}} \cdots \frac{\partial y_{N-1}}{\partial y_N} \right|$$

$$= \left| \frac{g_N}{g_k} \frac{(-1)}{k+1} \frac{(-1)}{k+2} \cdots \frac{(-1)}{N} \right|$$

$$= \left| \frac{y_N}{y_k} (-1)^{N-k+1} \frac{k!}{N!} \right|$$

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

(b) Now, ~~y_N is always $y_N < y_k$, so the~~ we have that y_N is always smaller than y_k, to evaluate an upper bound we "need" to say that y_N/y_k is largest when $\frac{y_N}{y_k} = 1$. Hence, we get the expression

$$\underline{\underline{\epsilon_k \leq \left| \frac{k!}{N!} \right|}}$$

(c) <u>See in the code!</u>

$$N! \leq \frac{k!}{\epsilon}$$

Stirling's approx for guessing $N \ln(N) - N \Rightarrow N! = 32$

$\underline{\underline{N! \leq 32!}}$ To get a good approximation of $y_k |_{k=20}$.

(d) Even though we start at the wrong initial value there is a recurrence relation which makes up for it an converges to the value of $0.1238\ldots$ See Code!