#1] $f(x) = \ln(1 - x/2)$ $\quad f(0) = 0$

$f'(x) = \dfrac{1}{x-2}, \quad f'(0) = -1/2$

$f''(x) = \dfrac{-1}{(x-2)^2}, \quad f''(0) = -\dfrac{1}{4}$

$f'''(x) = \dfrac{2}{(x-2)^3}$

$$\Rightarrow \ln\left(1 - \frac{x}{2}\right) = -\frac{1}{2}x - \frac{1}{8}x^2 + \frac{2}{6(\xi-2)^3}x^3, \ \xi \text{ between } x \text{ and } 0$$

$$.9998 = 1 - \frac{x}{2} \Rightarrow \boxed{x = .0004}$$

$$-\frac{1}{2}x - \frac{1}{8}x^2 \Big|_{x=.0004} = \boxed{-2.0002 \times 10^{-4}}$$

Error bound: $\quad 0 < \xi < .0004 \Rightarrow |\xi - 2| \sim 2$

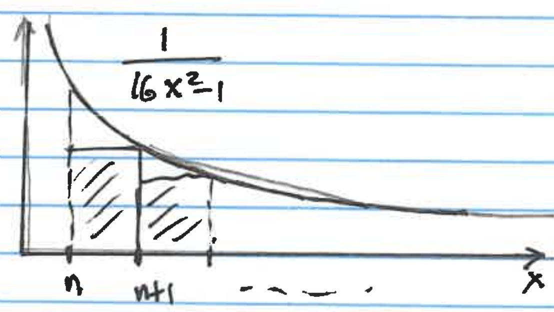
$$\Rightarrow \text{Error} = |\text{Remainder}| \le \frac{(1)(64)}{3(2)^3}(10^{-4})^3 \approx 2.67 \times 10^{-12}$$

#3] Take approximation as $\quad 4 - 8\sum_{k=1}^{n}(16k^2-1)^{-1}$.

$$\text{Error} = |\text{Remainder}| = 8\sum_{k=n+1}^{\infty}(16k^2-1)^{-1}.$$

Error $= 8 \times$ shaded area

$$\le 8\int_n^\infty \frac{dx}{16x^2-1}.$$



To simplify the calculations

$16x^2 - 1 \approx 16x^2$ for x large. Hence

$$\int_n^\infty \frac{dx}{16x^2-1} \approx \int_n^\infty \frac{dx}{16x^2} = -\frac{1}{16}\frac{1}{x}\Big|_n^\infty = -\frac{1}{16n}$$

Obviously not a very good method for approximating π

$$\Rightarrow \boxed{\text{Error} \le \frac{1}{2n}.} \quad \Rightarrow \quad n \ge \frac{1}{2\,\text{error}}$$

So if $\epsilon = 10^{-16}$, then $n \ge 5 \times 10^{15}$!!!

#2) $e^{x} = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!} + \frac{e^{\xi}}{(n+1)!} x^{n+1}$.

Let $\boxed{x=1}$ $\Rightarrow$ $0 < \xi < 1$.

we want to choose n so that $\frac{e^{\xi}}{(n+1)!} 1^{n+1} \leq 10^{-8}$

we take 3 as an upper bound for e^{ξ}.

$$\Rightarrow \quad (n+1)! \geq 3 \times 10^8.$$

Easiest way to solve this inequality is to try various values for n

n	9	10	11
$(n+1)!$	3.6288×10^6	3.99×10^7	4.79×10^8

$\Rightarrow$ $n = 11$ which is to say use the sum of the first $\underline{\underline{12}}$ terms.

#1) Take $x = 1.5$, $y = 9.4$, $z = 5.6$

$fl(1.5 \times 9.4) = fl(14.1) = 14$; $fl(14 \times 5.6) = fl(78.4) = 78$

Thus $fl(fl(xy) \times z) = 78$.

$fl(9.4 \times 5.6) = fl(52.64) = 53$; $fl(1.5 \times 53) = fl(79.5) = 80$

Thus $fl(x\, fl(yz)) = 80 \neq fl(fl(xy) \times z)$.

#2) $\left(\frac{1}{9}\right)_{10} = (0.000111000111000\text{ - - - -})_2$

$$= 1.\underbrace{110001}_{6}\underbrace{11000111}_{6} \text{ - - - } \underbrace{111.0001}_{6}\,|\,100 \times 2^{-4}$$

43 ⇐ 42nd bit after (·)

$\Rightarrow x_- = 1.11000111000 \text{ - - - } 1110001 \times 2^{-4}$

$x_+ = 1.11000111000 \text{ - - - } 1110010 \times 2^{-4}$

#3) (i) This is always true since if $a_i \neq b_i$ $\forall i \geq 1$, Then we take $m = 1$.

(ii) let m be The largest integer ≥ 1 such That

$$a_k = b_k, \quad k = 1, \dots, m-1 \text{ and } a_m \neq b_m.$$

Assume without loss of generality That $a_m > b_m$. we will show $a_m = b_m + 1$. So assume $\underset{\sim\sim\sim}{a_m \geq b_m + 2}$. we have

$$\frac{a_m}{10^m} + \frac{a_{m+1}}{10^{m+1}} + \dots + \dots = \frac{b_m}{10^m} + \frac{b_{m+1}}{10^{m+1}} + \dots$$

$$\text{left side} \geq \frac{a_m}{10^m} + \frac{0}{10^{m+1}} + \dots = \frac{a_m}{10^m}.$$

$$\text{right side} \leq \frac{b_m}{10^m} + \frac{9}{10^{m+1}} + \frac{9}{10^{m+2}} + \dots + \frac{9}{10^\ell} + \dots.$$

$$\text{Now } \frac{9}{10^{m+1}} + \frac{9}{10^{m+2}} + \dots = \frac{9}{10^{m+1}}\left(1 + \frac{1}{10} + \frac{1}{10^2} + \dots\right)$$
$$= \frac{9}{10^{m+1}} \frac{1}{1-\frac{1}{10}} = \frac{9}{10^{m+1}} \cdot \frac{10}{9} = \frac{1}{10^m}$$

$$\text{Thus, right side} \leq \frac{b_m}{10^m} + \frac{1}{10^m} = \frac{b_m + 1}{10^m}.$$

$$\frac{a_m}{10^m} \leq L.S. = R.S. \leq \frac{b_m + 1}{10^m} \Rightarrow a_m = b_m + 1.$$

(iii) If There is a $k \geq m+1$ such That $a_k > 0$, Then The value of the string $a_m a_{m+1} \dots$ is strictly greater Than $\frac{a_m}{10^m} + \frac{1}{10^k}$. Also, if There is a $\ell \geq m+1$ such That $b_\ell < 9$, Then the value of the string

$$\frac{b_m}{10^m} + \frac{b_{m+1}}{10^{m+1}} + \cdots + \frac{b_e}{10^e} \text{ is } \underline{\text{strictly less}} \text{ than } \frac{b_m}{10^m} + \frac{1}{10^m}$$

(as seen above). Hence, unless $a_{m+1} = a_{m+2} = \cdots = 0$ **and** $b_{m+1} = b_{m+2} = \cdots = 9$, the two values cannot be equal! A contradiction.

(iv) $x = .a_1 a_2 \cdots a_m 0 0 \cdots 0 \cdots$

$10^m x = a_1 a_2 \cdots a_m . 0 0 \cdots 0$ shift

$= \text{integer } L \Rightarrow x = \frac{L}{10^m}$. ✓

#4) $fl(3\pi + 7e) = fl(fl(fl(3) \cdot fl(\pi)) + fl(fl(7) \cdot fl(e)))$

$fl(3) = 3, \; fl(7) = 7$ (integers)

$fl(3\pi) = 3\pi(1+\delta_1), \; fl(7\pi) = 7\pi(1+\delta_2)$

$\Rightarrow \underset{\text{sum}}{fl}(3\pi(1+\delta_1) + 7\pi(1+\delta_2)) = [3\pi(1+\delta_1) + 7\pi(1+\delta_2)](1+\delta_3)$

$= 3\pi(1+\delta_1+\delta_3+\delta_1\delta_3) + 7\pi(1+\delta_2+\delta_3+\delta_2\delta_3)$

ignore $\delta_1\delta_3$ and $\delta_2\delta_3 \Rightarrow$

$fl(3\pi + 7e) = 3\pi + 7e + 3\pi(\delta_1+\delta_3) + 7e(\delta_2+\delta_3)$

$\Rightarrow \text{rel. error} = \frac{|3\pi + 7e - fl(3\pi+7e)|}{3\pi + 7e} \le \frac{3\pi|\delta_1+\delta_3| + 7e|\delta_2+\delta_3|}{3\pi+7e}$

$|\delta_1+\delta_3| \le |\delta_1| + |\delta_3| \le 2u$, also $|\delta_2+\delta_3| \le 2u$

$\Rightarrow \text{rel. error} \le \frac{2u(3\pi+7e)}{3\pi+7e} = 2u$

Note that in this special case, the bound is $2u$ whereas we did $\underline{\underline{3}}$ operations.

#5) let $x = \pm 1.\alpha_2 \cdots \alpha_{24} \times 2^p$
$y = \pm 1.\beta_2 \cdots \beta_{24} \times 2^q$

Assume both x, y are positive

The key step is to show that $|y| \le |x| 2^{-25} \Rightarrow \boxed{p \ge q + 25}$

proof.

$$\frac{|y|}{|x|} = \frac{1.\beta_2 \cdots \beta_{24}}{1.\alpha_2 \cdots \alpha_{24}} 2^{q-p} \le 2^{-25}$$

we want to see how small $\frac{1.\beta_2 \cdots \beta_{24}}{1.\alpha_2 \cdots \alpha_{24}}$ can be.

Now

$$\frac{1.\beta_2 \cdots \beta_{24}}{1.\alpha_2 \cdots \alpha_{24}} \ge \frac{1.0 \cdots 0}{1.1 \cdots 1} > \frac{1}{2} \quad \text{since } 1.1 \cdots 1 = 2 - 2^{-23} < 2$$

Thus

$$\frac{1}{2} 2^{q-p} < \frac{1.\beta_2 \cdots \beta_{24}}{1.\alpha_2 \cdots \alpha_{24}} 2^{q-p} \le 2^{-25}$$

$$\Rightarrow 2^{q-p} < 2^{-24} \Rightarrow q - p \underset{\uparrow \text{strict}}{<} -24 \Rightarrow q - p \le -25$$

since the inequality is strict and p, q are integers.

Now before adding, we equalize the exponents of x and y

$$x = 1.\alpha_2 \cdots \alpha_{24} 0 0 \cdots 0 \times 2^p$$
$$y = 0.\underbrace{0 \cdots 0 0 0}1 \beta_2 \cdots \beta_{24} \times 2^p$$

There must be at least 25 zeros in view of $p \ge q + 25$.

Next, we add x and y exactly

$$\Rightarrow x + y = 1.\alpha_2 \cdots \alpha_{24} 0 1 \beta_2 \cdots \beta_{24} \times 2^p$$

Since there is at least one zero in the "gap" between α_{24} and the 1 bit of y; upon rounding, all the bits after α_{24} are discarded $\Rightarrow fl(x+y) = x$. ✓

#6) $(64)_{10} = 2^6$. Also, $.015625 = \frac{1}{64} = \frac{1}{2^6}$

$\Rightarrow (64.015625)_{10} = 1000000.00000100 \cdots$

$= 1.00000000000010000000000000 \times 2^{6}$

$6 = c - 127 \Rightarrow c = (133)_{10} = 10000101$

$$\text{bit pattern } \underset{1\ 2\ \cdots\ 9\ 10\ \cdots\ 32}{\overset{4\quad 2\quad 8\quad 0\quad 0\quad 8\quad 0\quad 0}{|0100|0010|1000|0000|0000|1000|0000|0000|}}$$

$= [42800800]_{Hex}$

#7)

(a) $10^{403} \geq 2^{128}$, hence it is not a machine number

(b) $1 + 2^{-32}$ is a <u>real</u> number in $[1, 2] = [2^0, 2^1]$

The machine number immediately to the right of 1 is $1 + 2^{-24}$. Hence $1 + 2^{-32}$ is not a machine number

(c) $(1/5)_{10} = 0.0011001100 \cdots 001100 \cdots$

Does not terminate, hence it is not a machine #.

(d) $\dfrac{1}{256} = \dfrac{1}{2^8} = 0.00000001\underset{\uparrow 8}{}0 \cdots 0 \cdots$

hence it is a machine number.

#8) a)

$$[CA3F2900]_{16} = \overset{C\quad A\quad 3\quad F\quad 2\quad 9\quad 0\quad 0}{\underbrace{1\,10010100}\,0111\,1111\,0010\,1001\,0000\,0000}$$

sign → − . $c = 10010100 = (148)_{10} \Rightarrow \text{exp.} = c - 127 = \boxed{21}$

$(1.f)_2 = (1.01111111001010010000000000)_2$

$f \leftrightarrow \dfrac{1}{2^2} + \dfrac{1}{2^3} + \dfrac{1}{2^4} + \dfrac{1}{2^5} + \dfrac{1}{2^6} + \dfrac{1}{2^7} + \dfrac{1}{2^{10}} + \dfrac{1}{2^{12}} + \dfrac{1}{2^{15}}$

$= \dfrac{1}{2^{15}}(2^{13} + 2^{12} + 2^{11} + 2^{10} + 2^9 + 2^8 + 2^5 + 2^3 + 1) = \dfrac{16,169}{2^{15}}$

$\Rightarrow -(1.f)_2 = -\dfrac{16,169}{2^{15}} \times 2^{21} = -2^6(16,169) = \boxed{-3,131,968}$

#9) $[7F7FFFFF]_{16} =$

$$\underset{\text{sign}=+}{0}\,\underbrace{111\ 1111\ 0}_{c}111\ 1111\ 1111\ 1111\ 1111\ 1111$$

sign = + ; $c = (1111\,1110)_2 = 254 \Rightarrow \text{exp.} = 254 - 127 = 127$

$(1.f)_2 = 1.11\text{---}1$.

Thus, Number $= +\,1.1\text{---}1 \times 2^{127}$ which is the largest machine number.

$[00800000]_{16} = [\underset{\text{sign}}{0}\underbrace{000\ 0000\ 1}_{c}000\ 0000\ 0000\ 0000\ 0000\ 0000]$

sign is + ; $c = (1)_2 \Rightarrow \text{exp.} = 1 - 127 = -126$

$\Rightarrow$ number $= +\,1.0 \times 2^{-126}$, smallest positive machine number.

#10) Any positive real number can be expressed as

$$x = 0.b_1 b_2 \text{---} b_m \text{---} \quad \times 2^K \qquad \text{(binary expansion)}$$

with K some integer and $b_1 = 1$.

If k is <u>even</u>, we take $k = 2n$.

We also ~~take~~ $r = .b_1 b_2 \text{-----}$. It is clear that

$\frac{1}{2} = .100\text{---}0\text{---} \leq r \leq .111\text{---} \quad = 1$

If k is <u>odd</u>, we let $k = 2n+1$ and

$r = .0b_1 b_2 \text{----}$. Thus $x = .0b_1 b_2 \text{-----} \times 2^{2n}$.

Also

$\frac{1}{4} = .0100\text{----} \quad \leq r \leq .011\text{-----} \quad = \frac{1}{2}$.

we see that in ~~all~~ both cases $x = r \times 2^{2n}$ where

$\frac{1}{2} \leq r \leq 1$ or $\frac{1}{4} \leq r \leq \frac{1}{2}$ i.e. $\boxed{\frac{1}{4} \leq r \leq 1}$ ✓

Exercises on loss of significance

#1) $\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \cdots$

$\sin x = x - \frac{x^3}{6} + \frac{x^5}{120} - \cdots$

$\sqrt{1+x^2} = 1 + \frac{x^2}{2} - \frac{x^4}{8} + \cdots$

} from Tables or using Taylor's Thm.

$$\Rightarrow \boxed{f(x) = \frac{\tan x - \sin x}{x - \sqrt{1+x^2}}} = \frac{\left(x + \frac{x^3}{3} + \frac{2x^5}{15} + \cdots\right) - \left(x - \frac{x^3}{6} + \frac{x^5}{120} - \cdots\right)}{x - \left(1 + \frac{x^2}{2} - \frac{x^4}{8} + \cdots\right)}$$

$$= \frac{\frac{x^3}{2} + \frac{x^5}{8} + \cdots}{-1 + x - \frac{x^2}{2} + \frac{x^4}{8} + \cdots} \equiv C_3 x^3 + C_4 x^4 + \cdots$$

Note There are no C_0, C_1x, C_2x^2 terms

$$\Rightarrow \frac{x^3}{2} + \frac{x^5}{8} + \cdots = x^3(-C_3) + x^4(C_3 - C_4) + x^5\left(-1 + C_4 - \frac{C_3}{2}\right) + \cdots$$

$$\Rightarrow \boxed{C_3 = -\frac{1}{2}}, \quad C_3 - C_4 = 0 \Rightarrow \boxed{C_4 = -\frac{1}{2}}$$

$$\Rightarrow \boxed{f(x) = -\frac{1}{2}x^3 - \frac{1}{2}x^4 + \cdots} = -\frac{1}{2}x^3(1+x) \quad (*)$$

Use This formula

$(.0125)^3 = 1.953125 \times 10^{-6} \xrightarrow{fl} 1.953 \times 10^{-6}$

$1.953 \times 10^{-6}/2 = 9.765 \times 10^{-7}$ No error here.

$1 + .0125 = 1.0125 \xrightarrow{fl} 1.013$

$9.765 \times 10^{-7} \times 1.013 = 9.891945 \times 10^{-7} \xrightarrow{fl} 9.892 \times 10^{-7}$

$\Rightarrow \boxed{-9.892 \times 10^{-7}}$ is value calculated by "improved" formula (*) using 4-decimal digit arithmetic with rounding

Now the value calculated by using the f(x) as is on a calculator using 12-13 decimal digits

$\boxed{-9.888 \times 10^{-7}}$

$\Rightarrow$ rel. Error for improved formula is : $\dfrac{.004 \times 10^{-7}}{9.888 \times 10^{-7}} \approx 4 \times 10^{-4}$

Not too bad!.

Now let us calculate $f(.0125)$ directly using 4-decimal digit arithmetic:

$$\tan(.0125) = .012500651\text{--} \xrightarrow{fl} .0125$$
$$\sin(.0125) = .01249967\text{---} \xrightarrow{fl} .0125.$$

Thus, The calculated value will be zero, causing a relative error of $\dfrac{|9.888\text{--} \times 10^{-7} - 0|}{9.888\text{--} \times 10^{-7}} = 1.$!!!

#2)

$$f(x) = \frac{\sqrt{1+x^2}-1}{x^2} - \frac{x^2 \sin x}{x - \tan x}$$

we use Maclaurin series for all functions!

$$f(x) = \frac{1 + \frac{x^2}{2} - \frac{x^4}{8} + \text{--} - 1}{x^2} - \frac{x^2\left(x - \frac{x^3}{6} + \frac{x^5}{120} \text{--}\right)}{x - \left(x + \frac{x^3}{3} + \frac{2}{15}x^5 + \text{--}\right)}$$

After some simplifications

$$f(x) = \frac{1}{2} - \frac{x^2}{8} + \cdots + 3 - \frac{17}{10}x^2 + \text{--}.$$
$$= \boxed{\frac{7}{2} - \frac{73}{40}x^2 + \text{---}}$$

$$\Rightarrow \boxed{\lim_{x \to 0} f(x) = 7/2}$$

#13) $f(x) = \dfrac{\cos x - e^{-x}}{\sin x}$.

using Maclaurin series for all 3 functions and simplifying, we get

$$f(x) \simeq \frac{1 - x + x^2/6 - \frac{x^4}{120}}{1 - \frac{x^2}{6} + \frac{x^4}{5!}}.$$

Since the terms of the series are decreasing rapidly we can argue that the errors in both the numerator and denominator do not exceed $\dfrac{(.008)^5}{6!} \approx \dfrac{3.28 \times 10^{-11}}{720} \approx 10^{-13}$

Hence, an approximation to $f(.008)$ accurate to 10 decimal digits is $9.920212481 9 \times 10^{-1}$.