Homework/Numerical Analysis tips

① ⓐ What is $x = \underbrace{111\ldots1}_{112}\underbrace{00\ldots0}_{42}{}_2$?

Answer. 154 slots → 2^{153}

$$\Rightarrow x = 2^{153} + 2^{152} + \ldots + 2^{42} = 2^{42}\left(\underbrace{2^{111} + 2^{110} \cdots + 1}_{2^{112}-1}\right)$$

$$= 2^{42}\left(2^{112}-1\right) = \boxed{2^{154} - 2^{42}}$$

ⓑ What is $y = 0.\underbrace{0\ldots0}_{10}\underbrace{111\ldots11}_{92}$?

2^{-1} 2^{-10} 2^{-11} ...

$$y = 2^{-11} + 2^{-12} + \ldots + 2^{-102} = 2^{-102}\left(2^{91} + 2 + \ldots + 1\right)$$

$2^{-102}\,2^{91} = 2^{-11}$ ✓

$$= 2^{-102}\left(2^{92}-1\right) = \boxed{2^{-10} - 2^{-102}}$$

② f/e/s notation.

This means what # ?

$$\pm 1.f \times 2^{e-B}$$

$\begin{pmatrix} s=1 & - \\ s=0 & + \end{pmatrix}$

Given a #, how do you make the f/e/s.

→ First write in scientific notation

$\underbrace{\pm}_{s=0\text{ or }1}\ 1.\underbrace{\ldots\ldots}_{f}\times 2^{\text{exponent}}$ ✓ exponent $= e - B$

$\Rightarrow e = B + \text{exponent}$.

(3) When using significant figures, it's best to use the relative error estimate.

If x has m sign figures:

$$|\varepsilon_x| \le b^{-m+1}$$

Want $|\varepsilon_x| \le 2^{-k+1} \le 10^{-m+1}$ for binary to result in m sign figures in decimal

(4) Example

$z = 843.78\ldots\ldots$ accurate to last number

946 digits

949 sign. figures.

(5) 2nd rounding method — round to closest, ties to even

Key: the even thing applies only if it's an exact tie.

Example Decimal

Round to 2 sign figures

$2.650000001 \rightarrow 2.7$ not a tie

$2.65000000 \rightarrow 2.6$ is a tie

Example Binary

Round to 2 sign figures

$1.0100001_2 \rightarrow 1.1_2$ not a tie

$1.0100000_2 \rightarrow 1.0_2$ is a tie

Back to relative error.

Last time: $|\varepsilon_x| \leq b^{1-m}$

for floating point calculations with m sign. figures

Note: If the numbers are calculated by a computer that uses one of the first two rounding methods, then we get a slightly better estimate $|\varepsilon_x| \leq \frac{1}{2} b^{1-m}$.

Last time:

Relative error in the sum of two approximate numbers:

$$|\varepsilon_{x+y}| \leq \left|\frac{\bar{x}}{\bar{x}+\bar{y}}\right| |\varepsilon_x| + \left|\frac{\bar{y}}{\bar{x}+\bar{y}}\right| |\varepsilon_y|$$

If $\bar{x}, \bar{y} > 0$, then the Right side is a weighted average of $|\varepsilon_x|$ & $|\varepsilon_y|$, so in fact $|\varepsilon_{x+y}| \leq \max\{|\varepsilon_x|, |\varepsilon_y|\}$.

So with positive numbers, the percent accuracy does not go down when adding the numbers. (Good!).

Important point: if $\bar{x}$ and $\bar{y}$ have different signs $\left|\frac{\bar{x}}{\bar{x}+\bar{y}}\right|$ & $\left|\frac{\bar{y}}{\bar{x}+\bar{y}}\right|$

could be huge $\longrightarrow$ definitely can lose accuracy.

Example with sagemath:

$$x = 2^{14}$$

$$y = -2^{14} + 2^{-100}$$

$\bar{x} = x$ to a certain of sign figures

$\bar{y} = y$ ... — — —

$\bar{x} + \bar{y} = 0$ in computer.

$$e_{x+y} = (x+y) - (\bar{x}+\bar{y}) = 2^{-100}$$

$\bar{x}+\bar{y}=0 \Rightarrow \mathcal{E}_{x+y}=\frac{\mathcal{I}^{100}}{0}$ Aaahh...

New challenge: what happens to the relative error when multiplying two numbers?

$x = \bar{x} + e_x$ | $e_x = \bar{x}\mathcal{E}_x$

$y = \bar{y} + e_y$ | $e_y = \bar{y}\mathcal{E}_y$

$xy = \bar{x}\bar{y} + e_{xy}$ | $e_{xy} = \bar{x}\bar{y}\mathcal{E}_{xy}$

$x, \bar{x}$

$y, \bar{y}$

$\Rightarrow \bar{x}\bar{y}$ with be the approximation for xy.

$$e_{xy} = xy - \bar{x}\bar{y}$$

$$\Rightarrow e_{xy} = (\bar{x}+e_x)(\bar{y}+e_y) - \bar{x}\bar{y}$$

$$\Rightarrow e_{xy} = \cancel{\bar{x}\bar{y}} + \bar{x}e_y + \bar{y}e_x + e_xe_y - \cancel{\bar{x}\bar{y}}$$

$$\Rightarrow e_{xy} = \bar{x}e_y + \bar{y}e_x + e_xe_y$$

$$\Rightarrow (\bar{x}\bar{y})\mathcal{E}_{xy} = \bar{x}(\bar{y}\mathcal{E}_y) + \bar{y}(\bar{x}\mathcal{E}_x) + (\bar{x}\mathcal{E}_x)(\bar{y}\mathcal{E}_y)$$

Assuming $\bar{x}, \bar{y}$ are non zero

$$\Rightarrow \boxed{\varepsilon_{xy} = \varepsilon_y + \varepsilon_x + \varepsilon_x \varepsilon_y}$$

Usually $\varepsilon_x, \varepsilon_y$ are very small (like .001 for 3 sign figures)

$\Rightarrow$ Usually $\varepsilon_x \varepsilon_y$ is much smaller than either ε_x or ε_y.

$$\Rightarrow \quad \varepsilon_{xy} \approx \varepsilon_y + \varepsilon_x$$

$$|\varepsilon_{xy}| \approx |\varepsilon_y + \varepsilon_x| \leq |\varepsilon_y| + |\varepsilon_x|$$

good estimate for the error in a product.

Example to check:

$x = 3.1416$ ← round down to 4 sig. fig.

$y = 2.7183$ ← "

$\bar{x} = 3.141 \quad e_x = x - \bar{x} = .0006$

$\bar{y} = 2.718 \quad e_y = y - \bar{y} = .0003$

$$\Rightarrow \varepsilon_x = \frac{e_x}{\bar{x}} = \frac{.0006}{3.141} = .0001910$$

$$\varepsilon_y = \frac{e_y}{\bar{y}} = \frac{.0003}{2.718} = .0001104$$

Compare to

$$e_{xy} = xy - \bar{x}\bar{y}$$

$$= (3.1416)(2.7183) - (3.141)(2.718)$$

$$= .002573\ldots$$

$$\Rightarrow \varepsilon_{xy} = \frac{e_{xy}}{\bar{x}\bar{y}} = \frac{.002573}{(3.141)(2.718)} = .0003014$$

Let see if $|\varepsilon_{xy}| \approx |\varepsilon_x| + |\varepsilon_y|$

Left side $= .0003014$

Right side $= .0001910 + .0001104$

$= \underline{.0003014}$

Indeed $|\varepsilon_{xy}| \approx |\varepsilon_x| + |\varepsilon_y|$

$\bar{x} \leftarrow$ accurate within $.02\%$

$\bar{y} \leftarrow$ accurate within $.01\%$

$\bar{x}\bar{y}$ ← accurate within .03%.

Tool from Calculus we will need to find rel error in $f(\bar{x})$, given f & $\bar{x}$ rel error.

Taylor polynomial with Taylor-Lagrange Remainder:

$f(x)$ has Taylor series

$$T(x) = f(a) + f'(a)(x-a)^1 + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

Taylor series of $f(x)$ — close to $f(x)$ when x is near a.

$a = 0$

$\Rightarrow$

$$T(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots$$

Taylor polynomial of degree 3

$$T_3(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3.$$

(The 3rd degree polynomial that fits $f(x)$ best when $x \approx 0$.

at $x=a$

$$T_3(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3.$$

Example.

$f(x)=e^x \Rightarrow f'(x)=e^x,\ f'''(x)=e^x.$

at $x=0$

$$T_3(x) = e^0 + e^0 x + \frac{e^0 x^2}{2!} + \frac{e^0 x^3}{3!}.$$

$$T_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}.$$

$e^x \approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$ for x near 0.

eg $x = .1$

$$e^{.1} \approx 1 + .1 + \frac{(.1)^2}{2} + \frac{(.1)^3}{6}$$

$$= 1 + .1 + .005 + .000\overline{166}$$

$e^{.1} \approx 1.10516\overline{6}.$

pretty close!

$e^{.1} = 1.1051709\ldots$