Example

x = exact number (often don't know this)
$\bar{x}$ = approx number
$e_x = x - \bar{x}$ = error in $\bar{x}$
$\varepsilon_x = \dfrac{x-\bar{x}}{\bar{x}}$ = relative error in $\bar{x}$.

Let $x = 92.871149$ (in base 10).

(a) Find $\bar{x}$ = binary # used by computer to approximate x with 16 bit precision, with standard rounding [f 10 | e 5 | s 1 | B 15] IEEE 754

Solution: Because our significand will have 11 digits, so we need to calculate x in binary up to 11 signif. digits.

$x = 92.871149$

$92 = \underset{2^6}{64} + 16 + 8 + 4 = \underbrace{1011100_2}_{\text{7 sign digits.}}$

just need 4 more.

$$\begin{array}{rl} .871149 & \\ -.5 & .1_2 \\ \hline .371149 & + .01_2 \\ -.25 & \\ \hline .121149 & + .0001_2 \\ -.0625 & \\ \hline \end{array}$$

$x = 1011100.\overline{1101}1\ldots$

$\bar{x} = 1011100.1110_2$
rounded.

$$\begin{array}{r} .0586149 \\ -.0312500 \\ \hline \vdots \end{array} \quad + .00001_2$$

$$\left.\begin{array}{r} .5 \\ .25 \\ .125 \\ \hline .875 \end{array}\right\} .111_2$$

$$\boxed{\bar{x} = 92.875}$$

$$\boxed{= 1011100.1110_2}$$

(b) Write the 16-bit binary expression (in f|e|s format) that is used to store $\bar{x}$.

f (10)	e (5)	s (1)
0111001110	10101	0

$\uparrow$ + # (1 for neg #s)

$$\bar{x} = 1.0111001110 \times 2^6$$

$$6 = e - B$$

$$6 = e - 15 \Rightarrow e = 21$$

$$e = 21 = 16 + 4 + 1 = 10101$$

Answer: 0111001110101010

(c) Find e_x & $\mathcal{E}_x$:

$$e_x = x - \bar{x} = \begin{array}{r} 92.8711149 \\ -92.875 \end{array} = \boxed{-.0038851}$$

$$\begin{array}{r} 92.8750000 \\ 92.8711149 \\ \hline .0038851 \end{array}$$

$$\varepsilon_x = \frac{e_x}{\bar{x}} = \frac{-.0038851}{92.875} \approx \boxed{-0.0000418315}$$

(d) Suppose that x is instead truncated to binary number with a 400-digit significand (including the leading 1). To how many significant decimal digits is this accurate?

(Note: Accurate to d digits means the number can at most be off by 1 in the d^{th} slot, i.e.

$$|e_x| \le \underset{\uparrow \text{ last slot.}}{1} \cdot b$$

)

Recall $x = 92.871149$

binary: $\bar{x} = \underbrace{\underbrace{1011100}_{7}.\underbrace{11011\ldots\ldots\ldots}_{393 \text{ digits.}}}_{400 \text{ digits}}$

$$|e_x| \le 0.\underbrace{0\ldots\ldots0}_{392}1_2 = 2^{-393}$$

$$\approx 4.957 \times 10^{-119} \leftarrow \text{accurate up to } 10^{-118} \text{ in decimal}$$

ie $|4.957 \times 10^{-119}| \le 10^{-118}$

$\bar{x} = 92.871114$ ——— 10^{-118} slot.

$\Rightarrow$ 120 (significant decimal places of accuracy)

Propagation of error.

Error analysis.

Idea: Using approximate numbers in calculations results in more errors. How bad is it? Can we estimate or bound the error?

Recall x = exact number

$\bar{x}$ = approximation to x

$$e_x = x - \bar{x}$$

$$\varepsilon_x = \frac{x - \bar{x}}{\bar{x}} = \frac{e_x}{\bar{x}}$$

$$\Rightarrow x = \bar{x} + e_x$$

$$e_x = \bar{x}\,\varepsilon_x$$

Example: Use floating point representation

$$\bar{x} = (-1)^s f\, b^{-t}\, b^{e-B} \qquad (f \mid e \mid s)$$

(including leading 1) adding leading 1.

Note: assume $e \neq 0$, $e \neq 1111..$

$$\underbrace{|e_x| \leq 1 \cdot b^{-t}\, b^{e-B}}_{\text{1 in the last significant place}}$$

$$|\varepsilon_x| = \left|\frac{e_x}{\bar{x}}\right| \leq \left|\frac{b^{-t}\, b^{e-B}}{(-1)^s f\, b^{-t}\, b^{e-B}}\right| = \frac{|b^{-t} b^{e-B}|}{|(-1)^s f\, b^{-t}\, b^{e-B}|}$$

$$= \frac{b^{-t}\, b^{e-B}}{f\, b^{-t}\, b^{e-B}} = \frac{1}{f}$$

$$|\varepsilon_x| \leq \frac{1}{f} \leq \frac{1}{10_b^{m-1}} = \frac{1}{b^{m-1}} = b^{1-m}$$

lowest possible denom.

sign digits

example ①: approximating a decimal # to 15 sign digits $|\varepsilon_x| \leq 10^{1-15} = 10^{-14}$

② approx. a binary # to 39 sign digits $\Rightarrow |\varepsilon_x| \le 2^{1-39} = 2^{-38}$.

---

think of as $(100\% - 2^{-38})$ accurate.

These give error estimates & relative error estimates for floating point representation.

$\boxed{\text{Operations}}$

<u>Addition</u>: Given $\bar{x}, \bar{y}$ as approximations to x, y (so we know $e_x, \varepsilon_x, e_y, \varepsilon_y$, or we know bounds on $|e_x|, |e_y|, |\varepsilon_x|, |\varepsilon_y|$).
How close is $\bar{x}+\bar{y}$ to $x+y$?

i.e. What are e_{x+y}, ε_{x+y} in terms of $e_x, e_y, \varepsilon_x, \varepsilon_y$?

We use $\overline{x+y} = \bar{x} + \bar{y}$ as our estimate of $x+y$.

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

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

$$\Rightarrow \boxed{e_{x+y} = e_x + e_y}$$

TRIANGLE INEQUALITY

(i) $|A+B| \leq |A| + |B|$

(ii) $|A+B| \geq |A| - |B|$

$|A+B| \geq |B| - |A|$

$|A-B| \geq |A| - |B|$

$|A-B| \geq |B| - |A|$.

$$|e_{x+y}| = |e_x + e_y|$$

$$\Rightarrow \boxed{|e_{x+y}| \leq |e_x| + |e_y|}$$

Relative error:

$$e_x = \bar{x}\,\varepsilon_x$$

$$e_y = \bar{y}\,\varepsilon_y$$

$$e_{x+y} = (\bar{x}+\bar{y})\,\varepsilon_{x+y}$$

Plug in: $$|(\bar{x}+\bar{y})\mathcal{E}_{x+y}| \leq |\bar{x}\mathcal{E}_x| + |\bar{y}\mathcal{E}_y|$$

$|AB| = |A|\,|B| \Rightarrow$ $$|\bar{x}+\bar{y}|\,|\mathcal{E}_{x+y}| \leq |\bar{x}|\,|\mathcal{E}_x| + |\bar{y}|\,|\mathcal{E}_y|$$

If $\bar{x},\bar{y} > 0$: $|\bar{x}| = \bar{x}$, $|\bar{y}| = \bar{y}$ etc.

$$\overbrace{(\bar{x}+\bar{y})}^{>0}\,|\mathcal{E}_{x+y}| \leq \bar{x}|\mathcal{E}_x| + \bar{y}|\mathcal{E}_y|$$

$$\Rightarrow |\mathcal{E}_{x+y}| \leq \frac{\bar{x}|\mathcal{E}_x| + \bar{y}|\mathcal{E}_y|}{\bar{x}+\bar{y}}$$

If $A \leq B$

C positive $\Rightarrow \frac{A}{C} \leq \frac{B}{C}$

C negative $\Rightarrow \frac{A}{C} \geq \frac{B}{C}$

$\bar{x},\bar{y} > 0$

$$\boxed{\begin{aligned}&\Rightarrow |\mathcal{E}_{x+y}| \leq \frac{\bar{x}}{\bar{x}+\bar{y}}|\mathcal{E}_x| + \frac{\bar{y}}{\bar{x}+\bar{y}}|\mathcal{E}_y| \;\star\\ &\Rightarrow |\mathcal{E}_{x+y}| \leq |\mathcal{E}_x| + |\mathcal{E}_y|\end{aligned}}$$

$\star$ better estimate.