Nearby Machine Numbers

mantissa = significand

$x = q \cdot 2^m \qquad 1 \le q < 2 \qquad -126 \le m \le 127$

Say $x = (1.a_1 a_2 \cdots a_{23} a_{24} a_{25} \cdots)_2 \cdot 2^m \qquad a_i = 0 \text{ or } 1$

let $x_- = (1.a_1 a_2 \cdots a_{23})_2 \, 2^m$ (chopping) $\qquad (1.a_1 a_2 \cdots)_2$

$0.0\,0\cdots0\,0\,1 = 2^{-23}$ $\qquad 0.0\,1\,0\,0\,0 = 2^{-2}$

$x_+ = ((1.a_1 a_2 \cdots a_{23})_2 + 2^{-23}) \cdot 2^m$

(rounding up)

$x_+ = 5.23197000\,00$

$x = 5.23196824$

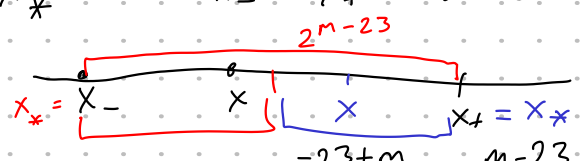
$x_- = 5.23196\,00000$

$x_- \le x \le x_+$

Either $\left| \frac{x - x_-}{x} \right| \le 2^{-24}$ or $\left| \frac{x - x_+}{x} \right| \le 2^{-24}$

Q) why?

A) Let x_* denote x_- or x_+ whichever is closer to x.



Clearly $x_+ - x_- = 2^{-23+m} = 2^{m-23}$

So $|\text{abs. error}| = |x - x_*| \le \frac{1}{2} 2^{m-23} = 2^{m-24}$

$q \ge 1$, $\frac{1}{q} \le 1$

$$\left|\text{Relative error}\right| = \left|\frac{\text{abs. error}}{x}\right| \le \frac{2^{m-24}}{q\,2^m} \le \frac{1}{q} 2^{-24} \le 2^{-24}$$

$$|\text{Relative error}| = \left|\frac{x - x_*}{x}\right| \le 2^{-24}$$: Unit Roundoff Error

Machine Epsilon = 2 · Unit Roundoff Error

64 bits $\qquad$ mach. eps $= 2^{-52}$

Exercise: On a modern machine using double precision, 52 bits are reserved for the mantissa. Try the following on your Python interpreter

```
u = 2**-53
e = 2**-52
1+e == 1
1+u == 1
```

```
for n in range(1,60):
    u = 2**(-(n+1))
    e = 2**(-n)
    if not (1+e==1) and (1+u==1):
        print(n)
        break
```

Notation $fl(x) = x_*$ the floating point machine number closest to x.

Example For $x = \frac{2}{3}$, compute the nearby machine numbers x_- and x_+ (in the architecture we described last time). Find $fl(x)$. What are the absolute round off error and the relative round off error in representing x by $fl(x)$?

$52.879 \times 10 = 528.79$

$\frac{2}{3} < 1$ so $\frac{2}{3} = (0.a_1 a_2 a_3 \cdots)_2$

decimal $\qquad$ binary

$2 \cdot \frac{2}{3} = \frac{4}{3} = a_1.a_2 a_3 \cdots \qquad$ Since $\frac{4}{3} > 1$, $a_1 = 1$.

$-1 \qquad -1.0000\,0$

Subtract 1 from both sides

$\frac{1}{3} = 0.a_2 a_3 a_4 \cdots$

$2 \cdot \frac{1}{3} = \frac{2}{3} = a_2.a_3 a_4 \cdots < 1 \qquad$ so $a_2 = 0$

$2 \cdot \frac{2}{3} = \frac{4}{3} = a_3.a_4 a_5 \cdots \qquad$ so $a_3 = 1$

$\vdots \qquad\qquad \vdots$

Thus, $\frac{2}{3} = (0.1010\,10\cdots)_2 = (0.\overline{10})_2 \qquad x = q \cdot 2^m \quad 1 \le q < 2$

$\frac{2}{3} = (0.\overline{10})_2 = (1.0101\cdots)_2 \, 2^{-1} \qquad m = -1$

$q = (1.0101\cdots)_2 = (1.f)_2 \qquad$ so $f = (0.010101\cdots)_2$

$1^{st}, 3^{rd}, 5^{th} \cdots$ bits (after binary point) are 0. So 23^{rd} bit is also 0 (followed by a 1).

Thus, $x_- = (1.0101\cdots010)_2 \cdot 2^{-1}$

$x_+ = (1.0101\cdots011)_2 \cdot 2^{-1}$

23 bits

To find $fl(x)$ we need to compute $x_+ - x$ and $x - x_-$ (and choose the smaller one).

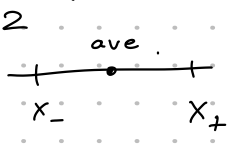
Here is another way: the number that is exactly in the middle of x_+ and x_- is $\frac{x_+ + x_-}{2} = \text{ave}$

23 bits

$x_- = (1.0101\cdots010\,|\,000\cdots)_2 \, 2^{-1}$

$\text{ave} = (1.0101\cdots010\,|\,100\cdots)_2 \, 2^{-1}$

$x_+ = (1.0101\cdots011\,|\,000\cdots)_2 \, 2^{-1}$



2.34|20
2.34|25
2.34|30

since $x > \text{ave}$, $fl(x) = x_+$.

$\frac{2}{3} = 1.0101\cdots0101\,|\,0101\cdots > 1.0101\cdots0101\,|\,0000$

ave

x_+ is closer to x.

Now $|\text{abs. error}| = |fl(x) - x| = x_+ - x$

(We can compute the difference in binary directly but it is easier to compute the following equivalent expression.)

$$x_+ - x = (x_+ - x_-) - (x - x_-)$$

$$x_+ - x_- = 2^{-23} \cdot 2^{-1} = 2^{-24} = 2^{m-23}$$

$$x - x_- = (0.\underbrace{0000\cdots000}_{23 \text{ bits}}\,1\,01010\,10\cdots)_2 \cdot 2^{-1}$$

So $(x - x_-) \cdot 2^{24} = 0.101010\cdots = \frac{2}{3}$

Thus $x - x_- = \frac{2}{3} \cdot 2^{-24}$

and $x_+ - x = 2^{-24} - \frac{2}{3}(2^{-24}) = \frac{1}{3} 2^{-24}$

$$|\text{abs. error}| = \frac{1}{3} \cdot 2^{-24}$$

$$\left|\text{Relative error}\right| = \frac{|\text{abs. error}|}{|x|} = \frac{\frac{1}{3} \cdot 2^{-24}}{\frac{2}{3}} = 2^{-25}$$

Recall $m = -1$ and $e = m + 127 = 126 = (01\ 111\ 110)_2$

So x is stored as

$$fl(x) = [\underbrace{0}_{s}\,|\,\underbrace{01\ 111\ 110}_{e}\,|\,\underbrace{010101\ 010101\ 010101\ 01011}_{f}]_2$$

in memory.

2.2 Loss of Significance

Let $\delta = \frac{fl(x) - x}{x}$ ($|\delta| \le u$ ← unit roundoff error)

then $fl(x) = x(1+\delta)$ with $|\delta| \le u$

We can assume $fl(x \square y) = (x \square y)(1+\delta)$ with $|\delta| \le u$ for machine numbers x and y where $\square$ is one of $+, -, \times, \div$.

Example (in a 5-digit decimal architecture)

$x = 0.31426 \cdot 10^3$ $\quad y = 0.92577 \cdot 10^5$

$x + y = 0.92891\ 26000 \cdot 10^5$

$x - y = -0.92262\ 74000 \cdot 10^5$

$x \cdot y = 0.29093\ 24802 \cdot 10^8$

$\frac{x}{y} = 0.33945\ 79647 \cdot 10^{-2}$

Then $fl(x+y) = 0.92891 \cdot 10^5$

$fl(x-y) = -0.92263 \cdot 10^5$

$fl(x \cdot y) = 0.29093 \cdot 10^8$

$fl(\frac{x}{y}) = 0.33946 \cdot 10^{-2}$

Caveat! Say $x = 0.372147\ 8693$

$y = 0.372023\ 0572$

Then $x - y = 0.000\underline{12\ 48\ 121}$

$fl(x) = 0.37215$

$fl(y) = 0.37202$

$fl(x) - fl(y) = 0.00013$

Rel. Error is $\left|\frac{(x-y) - (fl(x) - fl(y))}{x - y}\right| \approx 4\% \gg 5 \cdot 10^{-6}$

$\lim_{x \to a} f(x) = L$

$\exists L\ \forall \varepsilon > 0\ \exists \delta > 0$ if $0 \neq |x - a| < \delta$ then $|f(x) - L| < \varepsilon$

negate →

$\forall L\ \exists \varepsilon > 0\ \forall \delta > 0$

$\exists x$ s.t. $0 \neq |x - a| < \delta$ but $|f(x) - L| > \varepsilon$