Floating Point Numbers

Sunday, August 23, 2020 1:35 PM

Updated 8/22/25

How can we represent numbers on a computer?

Integers are easy. $6 = 1\cdot 4 + 1\cdot 2 + 0\cdot 1$, ie., 110 (binary)

What about fractions?

$$\frac{18}{19} + \frac{22}{23} = \frac{18\cdot 23 + 22\cdot 19}{19\cdot 23} = \frac{414+418}{437} = \frac{832}{437}$$

5 bits → 18; 5 bits → 19; 10 bits, 10 bits; 832) 10 bits; 437) 9 bits; 19 bits

Problem: memory increases

Could try decimals (fixed pt.)

131.467 (3, 3) or in binary ... 1 0 0 1 1 0 . 1 0 0 1 1 0 ... ($2^2\ 2^1\ 2^0\ 2^{-1}\ 2^{-2}$)

Done in embedded system

Instead, the standard for numerical computing, is floating pt.

Usually use IEEE 754 "double precision" → 64 bits = 8 bytes

(single precision = 32 bits)

Store numbers

$\mathbb{F}$ = Floating pt. $\quad (-1)^s (1+f)\cdot 2^e \qquad e = c - 1023$

$\mathbb{R}$ = real numbers

s = sign bit (1 bit)

e = exponent or characteristic, 11 bits $\quad 2^{11} = 2048$

f = mantissa, 52 bits

(scientific notation)

F also includes 0, NaN (0/0, 0·∞, ∞/∞), ±∞

Rule of thumb: precision $\approx 2^{52} = 4.5\cdot 10^{15}$

15 digits of precision in double

8 digits ... in single

Implications

① We can't represent very large (or very negative) numbers

$x \in \mathbb{F}$, then $|x| \leq 2^{1024} = 10^{308}$

ie., $x = -10^{400}$ is not in $\mathbb{F}$ (it is $-\infty$)

Overflow if not in range

② we can't get too small in magnitude (close to 0)

$x \in \mathbb{F}$, $|x| \geq 2^{-1022} \approx 10^{-308}$

underflow if $|x| < 2^{-1022}$ $\qquad x = 2^{-1023}$

$x == 0$ → True

③ limit to spacing ← relative $\qquad$ "Machine epsilon", SILENT ERROR

$1 \stackrel{?}{==} 1+\varepsilon$, ie., 1 and $1+\varepsilon$ are indistinguishable

true if $\varepsilon < \varepsilon_{machine} = 2^{-52} \approx 2.2\cdot 10^{-16}$

NOTE: youtube video is wrong, it is relative

2 vs $2+\varepsilon$, $\varepsilon < 2\cdot\varepsilon_{machine}$

Notation: $x \in \mathbb{R}$, $fl(x)$ is nearest number to x that's in $\mathbb{F}$

Notation: $x \in \mathbb{R}$, $fl(x)$ is nearest number to x that's in $\mathbb{F}$

$$\frac{|fl(x) - x|}{|x|} \leq \underbrace{\tfrac{1}{2}\varepsilon_{machine}}_{\text{depends on precise definition}}, \quad fl(x) = x \cdot (1+u), \text{ some } |u| \leq \tfrac{1}{2}\varepsilon_{machine}$$

$\frac{|fl(x)-x|}{|x|}$ } relative error = accuracy

$|fl(x) - x|$ = absolute error

digits of accuracy $\approx -\log_{10}$ (rel. error)

Precision : # digits, need not be correct!

Accuracy : relative error, limited by precision

More implications: lose associative rule

$$(a+b)+c = a+(b+c)$$

$$\underbrace{\left(\underbrace{1}_{1} + \varepsilon_{machine}/2\right)}_{0} - 1 \overset{?}{\neq} 1 + \underbrace{(\varepsilon_m/2 - 1)}_{\varepsilon_{m/2} \approx 1.11 \cdot 10^{-16}}$$

Catastrophic Cancellation

Pretend we have 3 digits of precision.

$x = 1.23\underbrace{587325\ldots}_{\text{garbage}}$

$y = 1.22358126\ldots$

$x - y = 0.01\underbrace{\text{mm mm mm mm mm}}_{\text{garbage}}$

"Silent error"!

$= 1.\boxed{\text{mm mm}}\text{mm mm mm} \cdot 10^{-2}$

you might think these are accurate since within our precision.