Floating Point Numbers

Sunday, August 23, 2020 1:35 PM

How can we represent numbers on a computer?

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

What about fractions?

5 bits → $\frac{18}{19} + \frac{22}{23} = \frac{18\cdot 23 + 22\cdot 19}{19\cdot 23} = \frac{414+418}{437} = \frac{832}{437}$) 10 bits) 9 bits
3 bits →

10 bits, 10 bits, 19 bits

Problem: memory increases

Could try decimals (fixed pt.)

131.467 (3, 3) or in binary $\ldots 1001\,1\,0.\,1\,0\,0\,1\,1\,0 \ldots$ ($2^2\ 2^1\ 2^0\ 2^{-1}\ 2^{-2}$)

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. $(-1)^s (1+f)\cdot 2^e$ $e = c - 1023$

$\mathbb{R}$ = real numbers

s = sign bit (1 bit)

e = exponent or characteristic, 11 bits ($2^{11} = 2048$)

f = mantissa, 52 bits

(scientific notation)

F also includes 0, NaN ($0/0$, $0\cdot\infty$, ∞/∞), $\pm\infty$

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 $|x| \le 2^{1024} \approx 10^{308}$

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 $|x| < 2^{-1022}$ $\quad x = 2^{-1023}$

$x == 0 \rightarrow$ True

(3) limit to spacing ← relative complicated

$1 \doteq 1 + \varepsilon$, ie., 1 and $1+\varepsilon$ are indistinguishable

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

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

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

$|fl(x) - x| \leq \frac{1}{2} \varepsilon_{machine}$, $fl(x) = x \cdot (1+\varepsilon)$, some $|\varepsilon| \leq \frac{1}{2} \varepsilon_{machine}$

$|fl(x) - x|$ — absolute error

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

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

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

$(1 + \varepsilon_{machine}/2) - 1 \neq 1 + (\varepsilon_{m}/2 - 1)$

$1 + \varepsilon_{machine}/2 = 1$, so $(1 + \varepsilon_{machine}/2) - 1 = 0$

$\varepsilon_{m/2} \approx 1.11 \cdot 10^{-16}$