Floating pt. numbers

scientific notation $x = \pm S \times 2^{E}$

$1 \le S < 10$

ex: $365.25 = +3.6525 \times 10^{2}$

S = significand , E = exponent

decimal pt 365.25 floats to 3.6525

base 2 $x = \pm S \times 2^{E}$, $1 \le S < 2$,

$\frac{11}{2} = (1.011)_2 \times 2^{2}$ $(x \ne 0)$

$S = b_0.b_1 b_2 \cdots)_2$, $b_0 = 1$ if $x \ne 0$

$= (1.\underbrace{b_1 b_2 \cdots}_{\text{fractional part}})_2$

<u>normalized rep.</u> if $b_0 = 1$

computer word

1 bit	8 bits	23 bits
sign	exp	fraction field

$11/2$ =

0	ebits(2)	011 ...

32 bit word

sign 0 = +, 1 = −

8 bits ~~of~~ for E ∴ $-128 \le \exp \le 127$

such numbers are floating pt. #'s

$71 = (1.000111)_2 \times 2^6$

$\boxed{0 \mid \text{ebits}(6) \mid 000111 0 \ldots}$

ex $1 = (1.000\ldots)_2 \times 2^0$

$\Leftrightarrow \boxed{0 \mid \text{ebits}(0) \mid 0 \ldots}$

$1024 = (1.000)_2 \times 2^{10}$

$\Leftrightarrow \boxed{0 \mid \text{ebits}(10) \mid 0 \ldots}$

can even get $(1.\underbrace{11\cdot 1}_{23 \text{ place}})_2 \times 2^{127}$

$\Leftrightarrow \boxed{0 \mid \text{ebits}(127) \mid \overset{23}{111\ldots 1}}$ (all 1's

$\Updownarrow$

$= 2^{127}$

largest $\# = + 2^{127} \cdot (1.\underbrace{1 \ldots}_{23 \text{ 1's}})$

<u>Double format</u> $\boxed{\pm \mid a_1 a_2 \cdots a_{11} \mid b_1 b_2 \cdots b_{52}}$

$1 + 11 + 52 = 64$ bit word

$(0)_{10} = a_1 \cdots a_{11} = 0 \cdots 0 \Leftrightarrow \pm(0.b_1 \cdots b_{52})_2 \times 2^{-1022}$ ← notice

$= (0 \cdots 01)_2 \Leftrightarrow \pm(1.b_1 \ldots b_{52})_2 \times 2^{-1022}$

$= (0 \cdots 10)_2 \Leftrightarrow \pm(1.b_1 \ldots b_{52})_2 \times 2^{-1021}$

$(1023)_{10} = (01111111)_2 \Leftrightarrow \pm(1.b_1 \ldots b_{52})_2 \times 2^0$

$(1024)_{10} = (10 \cdots 0)_2 \Leftrightarrow \pm(1.b_1 \ldots b_{52})_2 \times 2^1$

$(2046)_{10} = (111111\cdot 0)_2 \Leftrightarrow \pm(1.b_1 \ldots b_{52})_2 \times 2^{1023}$

$(2047)_{10} = (111111)_2 \Leftrightarrow \pm\infty$ if $b_1 = \cdots = b_{52} = 0$, NaN else denormalized

	E_{min}	E_{max}	N_{min}	N_{max}
single	-126	127	2^{-126}	$\sim 2^{128} \sim 3.4 \times 10^{38}$
double	-1022	1023	2^{-1022}	$\sim 2^{1024} \sim 1.8 \times 10^{308}$

machine precision p = # bits in significand
machine epsilon ϵ = gap between 1 & next larger floating pt. #

this differs from definition in book, where ϵ depends on method of rounding. (round up &

$1 + x > 1$ if $x = \epsilon/2$

$\boxed{\text{double precision} = 53 \qquad \text{gap } \epsilon = 2^{-52}}$ (1 is included) confusing??

---

consider only base 2

floating pt. #'s $b_0 . b_1 - b_m \times 2^E$

E in some somewhat symmetric range $E_{min} \le E \le E_{max}$

in single E is is det'd from (not equal to) $a_1 \cdots a_8$, $a_j = 0, 1$

might assume $0 \le E \le 2^8 - 1 = 255$ — 256 numbers
instead $E \Leftrightarrow$ numbers between -126 & 127 — 254 numbers
what happened?

Well, 00000000 and 0 — 01 both correspond to 2^{-126}

+ 111111 corresponds to either $\pm\infty$ or NAN depending on the result of a calculation

then ~~num~~ float number 0 is

$$\pm(0.\underbrace{0 - 0}_{23\ 0's}) \times 2^{-126}$$

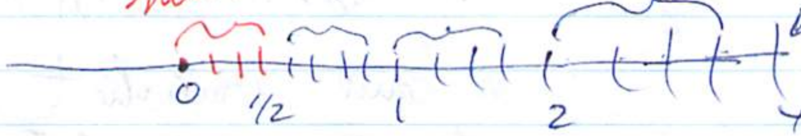
Toy number system not ~~why numbers are~~ the float representation

most computers

$$E = -1, 0, 1, \qquad \pm(b_0.b_1 b_2) \times 2^E$$

normalized ~~iff~~ $b_0 \neq 0$, $(b_0 = 1)$ don't store it

+ $0 \leftrightarrow b_0 = 0 \Rightarrow b_1 = b_2 = 0$



not in the system

$E = -1$ $(1.00, 1.01, 1.10, 1.11) \times 2^{-1}$ $\qquad \epsilon = \frac{1}{4}$

$\left(\frac{1}{2}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}\right)$

$E = 0$ $\left(1, \frac{5}{4}, \frac{6}{4}, 7/4\right)$ $\qquad (\times 2^0)$

$E = +1$ $\left(2, \frac{5}{2}, \frac{6}{2}, 7/2\right)$

add subnormal #'s $(0.00, 0.01, 0.10, 0.11) \times 2^{-1}$