

2.1.2

① 浮動小数点数 (floating point number)

例) 10進数の10進数で、実数を近似する:

$$① u = \pm \boxed{d_0} \boxed{d_1} \boxed{d_2} \boxed{d_3} \boxed{d_4} . \boxed{d_5} \boxed{d_6} \boxed{d_7} \boxed{d_8} \boxed{d_9} \quad 0 \leq d_j \leq 9 \quad \text{for } j = 0, \dots, 9$$

↑
符号の情報を別途持つ。

このとき、表現できる数値の範囲は

$$-99999.99999 \leq u \leq 99999.99999$$

②

$$u = \pm \boxed{m_0} . \boxed{m_1} \boxed{m_2} \boxed{m_3} \boxed{m_4} \boxed{m_5} \boxed{m_6} \boxed{m_7} \boxed{m_8} \times 10^{\pm \boxed{e_1}}$$

↑
符号の情報を別途持つ

$$\begin{cases} 0 \leq m_0 \leq 9, \\ 0 \leq m_j \leq 9 \quad \text{for } j = 1, \dots, 8, \\ 0 \leq e_1 \leq 9. \end{cases}$$

$$\text{このとき} \quad u = \pm \left(m_0 + \sum_{j=1}^8 \frac{m_j}{10^j} \right) \times 10^{\pm e_1}$$

すなわち、表現できる数値の範囲は

$$\underline{-9.999999999 \times 10^9 \leq u \leq 9.999999999 \times 10^9}$$

" " " "

$$-9999999990 \quad 9999999990$$

①の表現に比べて、表現できる数値の絶対値の範囲が広がる！
同じ桁数で

但し、絶対値が大きくなるほど、隣どうしの数の間隔も広がる。

②のよう、指数を用いた数値の表現 $\Rightarrow$ 浮動小数点

(浮動小数点)

「浮動小数点」の語源: 表す数値の絶対値の大きさに応じて
小数点の位置が移動するため。 $\Rightarrow$ ①の「固定小数点」。

浮動小数点

実数の表現:

$$(-1)^S \times M \times B^E,$$

$\left\{ \begin{array}{l} S: \text{符号}, M: \text{仮数 (mantissa)}, B: \text{基数 (base)}, \\ E: \text{指数 (exponent)}. \end{array} \right.$

例) 2.9979×10^8 ($m/s = \text{真空中での光速}$)

$$= 29.979 \times 10^7$$

$$= 0.29979 \times 10^9$$

のよう、指数をずらすことで、1つの数値に対していくつかの表現がある。

そこで、仮数の整数部を1以上10未満にする表現が一応に定まる。→ 正規化 (normalization)

計算機上の浮動小数点の例: 実数 r の表現

符号 S : 1ビット, 基数 $B=2$, 指数 E ,
仮数 M : n ビット.

仮数の正規化: $1 \leq M < 2$ とみえらう, E を定める.

仮数の表現

$$m_0, m_1, m_2, \dots, m_{n-1},$$

$$m_0 = 1,$$

$$m_j \in \{0, 1\} \text{ for } j=1, \dots, n-1.$$

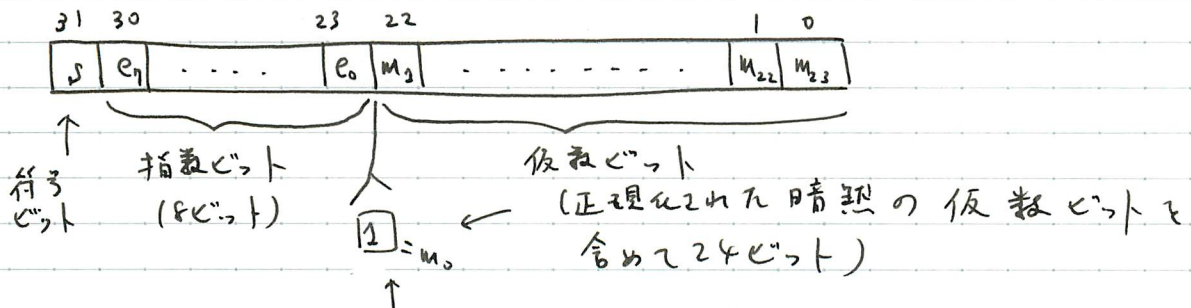
$$\text{s.t. } M = \sum_{j=0}^{n-1} \left(\frac{m_j}{2^j} \right).$$

$$\text{よ} \quad r = (-1)^S \times M \times 2^E \quad \text{と表す.}$$

アイリワイ-

- IEEE 754 浮動小数点規格 ... 現在、幅広く用いられている

① 単精度 (32ビット)



- (1) m₂₂ = 1 とおき, 1 ≤ M < 2 とみなすので正規化する
 (2) m₂₂ = 0 とおき, 非正規化数であり多くの数値を表す.

$$(1) \begin{cases} r = (-1)^s \times M \times 2^E, \\ M = 1 + \sum_{i=1}^{23} \left(\frac{m_i}{2^i} \right), \\ E = \left(\sum_{i=0}^7 2^i e_i \right) - 127. \end{cases}$$

$-126 \leq E \leq 127$ の場合と
 (指数ビットのすべてが 0
 の場合を除く)

$$(2) \begin{cases} r = (-1)^s \times M \times 2^E, \\ M = \sum_{i=1}^{23} \left(\frac{m_i}{2^i} \right) \quad (\text{非正規化数}), \\ E = -127. \quad (\text{指数ビットのすべて 0}) \end{cases}$$

特殊な数

- ± 0 : $\begin{cases} M = 0 & (\text{仮数ビットのすべて 0}) \\ E = -127 & (\text{指数ビットのすべて 0}) \\ s \in \{0, 1\} \Rightarrow +0 \text{ or } -0 \text{ を表す.} \end{cases}$

• $\pm\infty$: $\left\{ \begin{array}{l} M = 0 \quad (\text{仮数部はすべて0}) \\ E = \overset{128}{\cancel{127}} \quad (\text{指数ビットはすべて1}) \\ S \in \{0, 1\} \Rightarrow +\infty \text{ or } -\infty \text{ を表す.} \end{array} \right.$

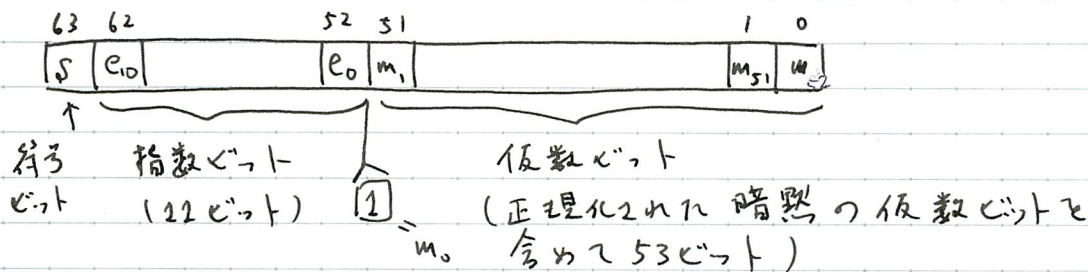
• NaN (Not a Number / 非数). $0 \div 0$ の値.
_{など}

$M \neq 0$ (仮数部の少なくとも1つのビットは1)

$E = \overset{128}{\cancel{127}} \quad (\text{指数ビットはすべて1})$

$S \in \{0, 1\} \quad (\text{符号の値は無関係})$

② 倍精度 (164ビット)



(1) $m_0 = 1$ である正規化数

$$\begin{cases} r = (-1)^S \times M \times 2^E, \\ M = 1 + \sum_{i=1}^{52} \left(\frac{m_i}{2^i} \right), \quad E = \left(\sum_{i=0}^{10} 2^i e_i \right) - 1023. \end{cases}$$

($-1022 \leq E \leq 1022$)

指数ビットはすべて0の場合とすべて1の場合を除く.

(2) $m_0 = 0$ である非正規化数.

$$r = (-1)^S \times M \times 2^E,$$

$$M = \sum_{i=1}^{52} \left(\frac{m_i}{2^i} \right), \quad E = -1023 \quad (\text{指数ビットはすべて0})$$

特殊の数

- $\begin{array}{l} \text{仮数ビットのすべて} \quad \text{指数ビットのすべて} \\ \downarrow \quad \quad \quad \downarrow \\ \bullet \pm 0 : \quad M = 0, \quad E = -1023, \quad S = 0 (+0) \text{ or } 1 (-0) \\ \bullet \pm \infty : \quad M = 0, \quad E = 1023, \quad S = 0 (+\infty) \text{ or } 1 (-\infty) \\ \bullet \text{NaN} : \quad M \neq 0, \quad E = 1023, \quad S \in \{0, 1\} \end{array}$
- $\uparrow$ 指数ビットのすべて 1
 $\downarrow$ 符号の無関係

⑩ 浮動小数点演算における誤差

① 丸め誤差 (rounding error)

(例) $\pi \rightarrow \boxed{3.141592653589793238462643383279502884197169399375105820974941598} \times 10^0$

$\uparrow$ 仮数部: + 通 5 桁. $\uparrow$ 計算機内部に 12 桁 (guard digit/bit) をつけて丸めを行う.

② 桁落ち誤差 (cancellation error)

(例) 仮数部 + 通 9 桁.

$$\begin{array}{r}
 1.23456789 \times 10^0 \\
 -) 1.23450000 \times 10^0 \\
 \hline
 0.00006789 \times 10^0
 \end{array}$$

$\downarrow$ 正規化

有効桁数が 9 桁から $\rightarrow$ 6.78900000 $\times 10^{-5}$
 4 桁へ減少. これらの桁は正確でない.

★ 絶対値の異なる 2 つの数値の計算で生じる誤差がある。

③ 情報落ち

(例) 仮数部 + 通 3 桁.

$$\begin{cases} x = 1.00 \times 10^4 \\ y = -1.00 \times 10^4 \\ z = -1.00 \times 10^0 \quad (= 1.00) \end{cases}$$

271. $x + y + z$ を計算する.