

実数

1. 正規化

実数は正規化してあらわされる。m 進法を使った場合，正規化された実数は

$$\pm (0.f_1f_2f_3\dots f_n)_m \times m^{\pm E} \quad (1)$$

ここで，

$$\pm (0.f_1f_2f_3\dots f_n) = \pm (f_1 \times m^{-1} + f_2 \times m^{-2} + f_3 \times m^{-3} + \dots + f_n \times m^{-n}) \quad (2)$$

は，仮数部で f_i は 1 から n までの整数， $f_1 \neq 0$ である。また，E は指数部で 0 または正の整数である。

2. IEEE 形式

表現

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
±	E+126				f ₂ ⋯f ₄				f ₅ ⋯f ₈				f ₉ ⋯f ₁₂				f ₁₃ ⋯f ₁₆				f ₁₇ ⋯f ₂₀				f ₂₁ ⋯f ₂₄					

1. 指数部は E に 126 を足した「ゲタばき表現」で表され，E は -125 から 128 の値をとる。
2. 仮数部は 2 進の 24 桁で， f_1 は常に 1 で省略。
3. 指数部が 0 (E=-126) の時は $f_1=0$ と見なし，小さな数を表現する。
4. 指数部が 255 (E=129) の時は，無限大など特殊な数字を表す。実数演算はできない。
5. 0 は全てのビットが 0 である
6. $f_1(=1).f_2(=0)f_3(=0)f_4(=0)\dots$ のときは，0 01111111 0000⋯⋯
7. $f_1(=1)f_2(=0).f_3(=0)f_4(=0)\dots$ のときは，0 10000000 0000⋯⋯

最大値、最小値

1. 正規数 ($f_1 \neq 0$) で絶対値最大のものは

$$(1-2^{-24}) \times 2^{128} = 3.40282347 \times 10^{38} \quad (4)$$

2. 絶対値最小のものは

$$2^{-1} \times 2^{-125} = 1.17549435 \times 10^{-38} \quad (5)$$

3. 非正規数 ($f_1=0$) まで含めて, 絶対値最小のものは,

$$2^{-24} \times 2^{-125} = 1.40129846 \times 10^{-45} \quad (6)$$

誤差

1. 25桁目を0捨1入(四捨五入の2進数版)した場合.

1) 正規数の相対誤差は $f_1=f_2=\dots=f_{24}=1$ の時, 最も小さくて

$$\text{約 } 2^{-25} \approx 3 \times 10^{-8} \quad (7)$$

2) $f_1=1, f_2=\dots=f_6=0$ の時, 最も大きくて

$$\text{約 } 2^{-24} \approx 6 \times 10^{-8} \quad (8)$$

2. IEEE規格ではその他の丸めも可.

3. 切捨ての場合はこの2倍.

4. 非正規数はこの限りではない.