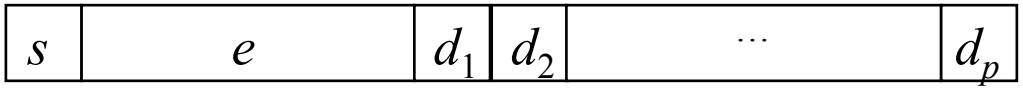
第2回　浮動小数点体系

3. 浮動小数点体系の定義

- 浮動小数点体系の定義
  - 浮動小数点数 ：b 進 q 桁 : q = 1 + (指数部のビット数) + p



符号ビット　指数部　　f　仮数部

2進　　b進

f:仮数(mantissa)
b:基数(exponent base)
e:指数(exponent)

$$s = \begin{cases} 0 & \text{正の数} \\ 1 & \text{負の数} \end{cases}$$

e = 2進整数（2の補数等）　$e_l \le e \le e_u \quad (e_l < 0 < e_u)$

$$f = (0.d_1 d_2 \ldots d_p)_b = \frac{d_1}{b} + \frac{d_2}{b^2} + \cdots + \frac{d_p}{b^p}$$

$d_1 \neq 0$ のとき正規浮動小数点数であるという
（正規化されている）

表現している数 $x = \pm f \times b^e$

3. 浮動小数点体系の定義

- 例
  - 2 進 16(1+4+11)桁

0|0011|11000000000 ---

$$+(0.11000000000)_2 \times 2^{(0011)_2}$$
$$= +(\frac{1}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{0}{2^4} + \cdots + \frac{0}{2^{11}}) \times 2^3$$
$$= (0.5 + 0.25) \times 8 = 6.0$$

1|1101|10010000000 ---

$$-(0.10010000000)_2 \times 2^{(1101)_2}$$
$$= -(\frac{1}{2} + \frac{0}{2^2} + \frac{0}{2^3} + \frac{1}{2^4} + \cdots + \frac{0}{2^{11}}) \times 2^{-3}$$
$$= -(0.5 + 0.0625) \times 0.125 = -0.0703125$$

符号ビット
0:正の数
1:負の数

指数部
2進
2の補数

仮数部
2進

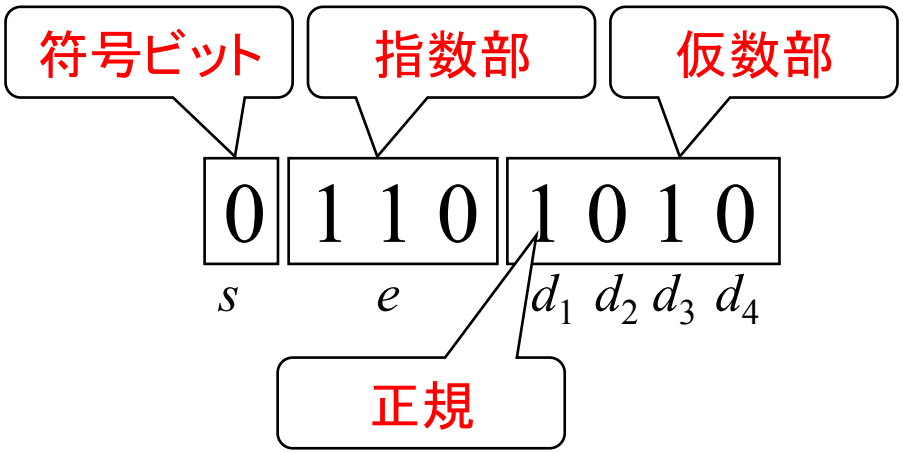
正規化
（ここが常に 1）

- 通常
  - 単精度実数(float) では，符号部 1bit，指数部 8bit，仮数部 23 bit
  - 倍精度実数(double)では，符号部 1bit，指数部 11bit，仮数部 52 bit

ただし，上記の形式とは異なる IEEE754形式 等がよく使われる.

3. 浮動小数点体系の定義（演習）

（問）基数2，指数部3ビット，仮数部4ビットの浮動小数点体系を考える．指数部は2の補数で整数を表す．正規化を行い，零の表現は全てのビットを0とする．



$d_1 \neq 0$ のとき正規浮動小数点数という

表している数値 $\pm\left(\frac{d_1}{2}+\frac{d_2}{2^2}+\frac{d_3}{2^3}+\frac{d_4}{2^4}\right)\times 2^e$

$$s=\begin{cases}0 & \text{正の数}\\ 1 & \text{負の数}\end{cases}$$

3ビットの2の補数	
3	011
2	010
1	001
0	000
–1	111
–2	110
–3	101
–4	100

３．浮動小数点体系の定義（演習）

（問）つぎのビット列は，この浮動小数点体系では値はいくらになるか．

| 0 | 010 | 1101 | ---

| 0 | 101 | 1000 | ---

| 1 | 001 | 1010 | ---

3. 浮動小数点体系の定義（演習）

（問）この浮動小数点体系で次の実数を表すと，どのようなビット列になるか

$$0.125 = (0.25) \times 2^{-1} - (0.5) \times 2^{-2} \quad \text{--- } 0\ 110\ 1000$$

正規化するため，まず，ここを
0.5 以上，1.0 未満にする

$$\begin{aligned} -3.75 &= -(1.875) \times 2^{1} = -(0.9375) \times 2^{2} \\ &= -(0.5+0.4375) \times 2^{2} \\ &= -(0.5+0.25+0.1875) \times 2^{2} \\ &= -(0.5+0.25+0.125+0.0625) \times 2^{2} \end{aligned}$$

--- 1 010 1111

$$\left(\frac{1}{2^1} = 0.5, \quad \frac{1}{2^2} = 0.25, \quad \frac{1}{2^3} = 0.125, \quad \frac{1}{2^4} = 0.0625, \quad \frac{1}{2^5} = 0.03125 \right)$$

3. 浮動小数点体系の定義

- 演習　基数2，指数部3ビット，仮数部4ビットの浮動小数点体系を考える．指数部は2の補数で正規化を行うとする．このとき，次の各問に答えよ．
  - 浮動小数点体系のビット列　⇔ 実数
    - 0 011 1100　⇔
    - 　　⇔　3.75
    - 0 010 1011　⇔
    - 　　⇔　0.4375
    - 0 111 1101　⇔
    - 　　⇔　0.1875
    - 1 110 1000　⇔
    - 　　⇔　-0.15625
    - 1 000 1001　⇔
    - 　　⇔　-1.0
    - 1 001 1010　⇔
    - 　　⇔　-4.5
    - 1 011 1110　⇔

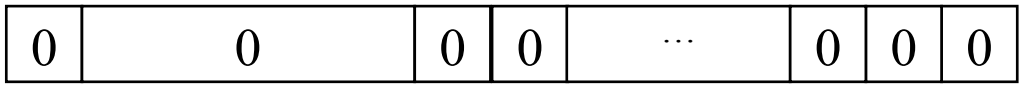
3. 浮動小数点体系の定義

- 表現できる正の数　　2 進 1+4+11 桁の場合

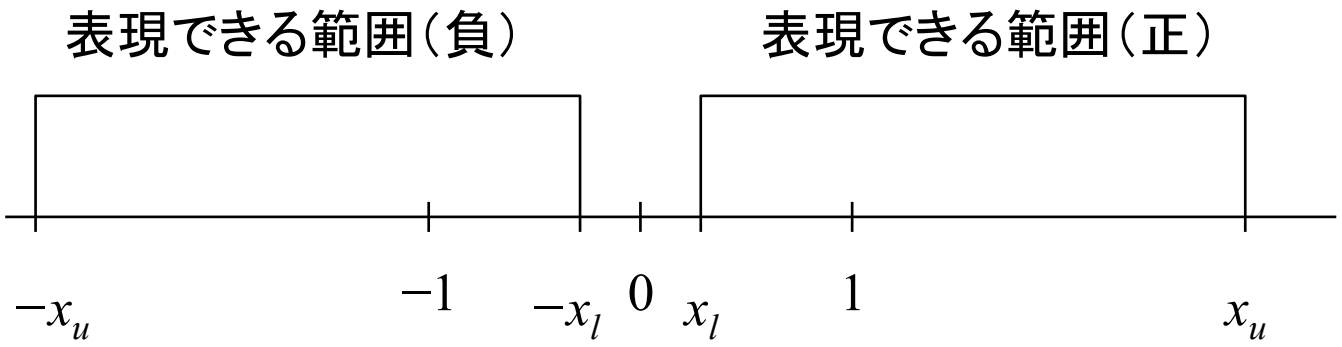
0	0111	11111111111	--- $(0.11111111111)_2\times 2^{7}$
0	0111	11111111110	--- $(0.11111111110)_2\times 2^{7}$
0	0111	11111111101	--- $(0.11111111101)_2\times 2^{7}$
	⋮		
0	0111	10000000000	--- $(0.10000000000)_2\times 2^{7}$
0	0110	11111111111	--- $(0.11111111111)_2\times 2^{6}$
	⋮		
0	0110	10000000000	--- $(0.10000000000)_2\times 2^{6}$
	⋮		
0	1000	11111111111	--- $(0.11111111111)_2\times 2^{-8}$
	⋮		
0	1000	10000000000	--- $(0.10000000000)_2\times 2^{-8}$

3. 浮動小数点体系の定義

- 0 の表現



- 表現できる数（離散的であることに注意）



- 表現できない範囲 ：2種類ある（表現したい数を x^* とする）

$|x^*| > x_u$ ：オーバーフロー　計算を中断

$0 < |x^*| < x_l$ ：アンダーフロー　0 で代用

3. 浮動小数点体系の定義（演習）

（問）基数2，指数部3ビット，仮数部4ビットの浮動小数点体系で表現できる**正の**浮動小数点数のうち，最大のものと最小のものは，それぞれ，値がいくらになり，ビット列はどのようになるか

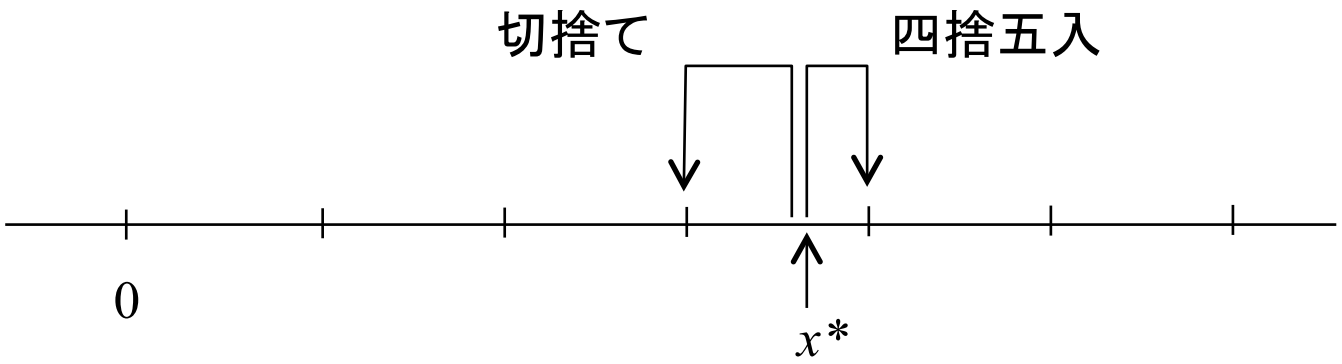
・最大のもの

・最小のもの

4. 丸め誤差

- 丸め -- 実数値 x^* (連続)を浮動小数点数(離散)で近似的に表現すること
- 「丸め」の方法

四捨五入	x^* に最も近い浮動小数点数にする.
切捨て	絶対値に関して $\|x^*\|$ を超えない値のうち, 最も近い値を持つ表現にする.



0.10110101... $\times 2^3$を 仮数部の桁数 p=5 で丸めると

切捨て	0.10110×2^3
四捨五入	0.10111×2^3

0.10110011... $\times 2^3$を 仮数部の桁数 p=5 で丸めると

切捨て, 四捨五入 0.10110×2^3

4. 丸め誤差

（演習）基数2，指数部3ビット，仮数部4ビットの浮動小数点体系で円周率を表すと，どのようなビット列になるか．またその際，どの程度の丸め誤差が生じるか．丸めは切捨てで行う．

$$
\begin{aligned}
\pi &= 3.1415926... \\
&= (1.5707963...) \times 2^1 \\
&= (0.7853981...) \times 2^2 \\
&= (0.5+0.2853981...) \times 2^2 \\
&= (0.5+0.25+0.0353981...) \times 2^2 \\
&= (0.5+0.25+0.125+0.0625+0.0353981...) \times 2^2 \\
&= (0.5+0.25+0.125+0.0625) \times 2^2 \ + \ (0.0353981...) \times 2^2 \\
&= 3 \ + \ (0.0353981...) \times 2^2
\end{aligned}
$$

答　ビット列は　0 010 1100

誤差は $\pi - 3 = (0.0353981...) \times 2^2 = 0.1415926...$ になる

4. 丸め誤差（演習）

（演習）基数2，指数部3ビット，仮数部4ビットの浮動小数点体系で円周率を表すと，どのようなビット列になるか．またその際，どの程度の丸め誤差が生じるか．丸めは四捨五入で行う．

$\pi = 3.1415926...$

$=$

5. 浮動小数点体系の算術演算

- 以後　基数 $b = 10$　仮数部の桁数 $p = 4$　丸めは四捨五入とする
- 乗算　$0.9000 \cdot 10^3 \times 0.9000 \cdot 10^2 = (0.9000 \times 0.9000) \cdot 10^5 = 0.8100 \cdot 10^5$

$$0.1111 \cdot 10^3 \times 0.1111 \cdot 10^2 = (0.1111 \times 0.1111) \cdot 10^5 = 0.01234321 \cdot 10^5$$
$$= 0.1234321 \cdot 10^4 = 0.1234 \cdot 10^4$$

正規化　　丸め

- 除算　$0.1000 \cdot 10^3 / 0.2000 \cdot 10^5 = (0.1000 / 0.2000) \cdot 10^{-2} = 0.5000 \cdot 10^{-2}$

$$0.4000 \cdot 10^3 / 0.3000 \cdot 10^5 = (0.4000 / 0.3000) \cdot 10^{-2} = 1.29218... \cdot 10^{-2}$$
$$= 0.129218... \cdot 10^{-1} = 0.1292 \cdot 10^{-1}$$

正規化　　丸め

5. 浮動小数点体系の算術演算

- 加減算

指数部の大きい方で桁合わせ

$$0.1234\cdot 10^4 + 0.5678\cdot 10^2 = (0.1234 + 0.005678)\cdot 10^4 = 0.129078\cdot 10^4$$
$$= 0.1291\cdot 10^4$$

丸め

$$0.3333\cdot 10^4 - 0.3321\cdot 10^4 = (0.3333 - 0.3321)\cdot 10^4 = 0.0012\cdot 10^4$$
$$= 0.1200\cdot 10^2$$

正規化

$$0.5670\cdot 10^3 + 0.5430\cdot 10^3 = (0.5670 + 0.5430)\cdot 10^3 = 1.1100\cdot 10^3$$
$$= 0.1110\cdot 10^4$$

正規化

5. 浮動小数点体系の算術演算

- 誤差限界
  - z の誤差を ε_z で表す. ε_z はさまざまな理由で生じ, その値もさまざま.
  - $|\varepsilon_z|$ の最大値を誤差限界という.

例　z を10進仮数 p 桁で, 指数部の値が e_z の浮動小点数とする.
その場合, 四捨五入による丸めの誤差限界は

$$|e_z| \leq 0.5 \times 10^{e_z - p}$$

- 絶対誤差 / 相対誤差
  - 絶対誤差　:　近似値と真値の差

    絶対誤差 = 近似値 - 真値
  - 相対誤差　:　近似値に含まれる誤差の割合

    相対誤差 = 絶対誤差 ／ 真値

6. 情報落ち，桁落ち

- 有効数字

  位取りの “0” を除いた意味のある数値を有効数字という

  1.23 m・・・有効数字3桁
  0.12 m・・・有効数字2桁
  1.2 m ・・・有効数字2桁
  1.20 m・・・有効数字3桁
  120 m・・ 有効数字2桁 あるいは3桁

  → 明確にするには、
  1.2×10^2 m ・・・有効数字2桁
  1.20×10^2 m・・・有効数字3桁

- 桁落ち

  指数部が同じで，仮数部の上位 n 桁が同じとき，減算をすると有効数字が n 桁減る

$$0.1234\cdot10^4-0.1233\cdot10^4=0.0001\cdot10^4=0.1000\cdot10^1$$

有効数字 4桁とする

有効数字 1桁

便宜的な 0

対策 －近接した2数の減算は避ける

6. 情報落ち，桁落ち

- 情報落ち

  加減算時，指数部の小さい方の数の仮数部の一部が失われる

$$0.1234\cdot 10^4 + 0.5678\cdot 10^2 = (0.1234 + 0.005678)\cdot 10^4$$

$$= (0.1234 + 0.005678)\cdot 10^4 = 0.129078\cdot 10^4 = 0.1291\cdot 10^4$$

$$\begin{array}{rl} & 0.1234 \\ + & 0.005678 \\ \hline & 0.129078 \end{array}$$

この部分の情報がなくなる

7. 誤差の種類

- 本来の誤差　数値そのものが持つ誤差
  - 有効数字，絶対誤差 / 相対誤差，誤差限界
- 丸め誤差
  - 桁落ち，情報落ち
- 打ち切り誤差　計算法に起因する誤差