# 数値解析（松尾）　配付資料　　2017年10月16日

## ■浮動小数点数 (floating point number)

現在最も普及している表現形式「IEEE754」の概略：

$$x_{\mathrm{f}} = \pm\left(\frac{1}{2^0} + \frac{x_2}{2} + \frac{x_3}{2^2} + \cdots + \frac{x_n}{2^{n-1}}\right) \times 2^m.$$

- 単精度 (single precision)=4byte=1+23+8bit：$n = 23 + 1, -126 \le m \le 127$.
  表現できる最小の正数 $= 2^{-126} \simeq 1.175 \times 10^{-38}$，最大の正数 $\simeq 2^{127} \times 2 \simeq 3.403 \times 10^{38}$.
  有効数字桁数：$2^{24} = 10^{7.224}$.

- 倍精度 (double precision)=8byte=1+52+11bit：$n = 52 + 1, -1022 \le m \le 1023$.
  表現できる最小の正数 $= 2^{-1022} \simeq 2.225 \times 10^{-308}$，最大の正数 $\simeq 2^{127} \times 2 \simeq 1.798 \times 10^{308}$.
  有効数字桁数：$2^{53} = 10^{15.95}$.

(注) 実際にはもっと複雑．たとえば指数部の空き 2 （例：$2^8 = 256 = (127 - (-126) + 1) + 2$）は，0 や特殊な数の表現に予約されている．

## ■丸め (rounding)

浮動小数点数で表現可能な区間内の数 $x$ に対して，

$$x_{\mathrm{f}} = x(1+\epsilon_x), \qquad |\epsilon_x| \le \epsilon_{\mathrm{M}} \quad 「マシンイプシロン」.$$

$\epsilon_{\mathrm{M}} \sim 10^{-7}$（単精度），$\sim 10^{-16}$（倍精度）．

## ■誤差 (error) の種類

- 丸め誤差 (rounding error)：あらゆる演算で発生
- 桁落ち：近い数の引き算で発生
- 情報落ち（積み残し）: 級数の計算などで発生

## ■区間演算 (interval arithmetic)

浮動小数点数をひとつ持つ代わりに，それの含まれる区間

$$x_{\mathrm{f}} \in X = [\underline{x}, \overline{x}]$$

を持ち，区間同士の演算を定義する．例：$X + Y = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$.

参考：「精度保証付き数値計算」＝浮動小数点演算で通常の数値計算をしつつ，同時にそれの含まれる区間も計算して，数値計算結果の精度を保証する．