数値解析第1回 数値計算の基礎

1．精度と誤差

真値 x，近似値 $x' \cong x$ について，

・**誤差：** $\Delta x = x' - x$， ・**絶対誤差：** $|\Delta x|$，

・**相対誤差：** $\dfrac{|\Delta x|}{|x|}$ $(x \neq 0)$， ・**(10進)有効桁数：** $-\log_{10}\dfrac{|\Delta x|}{|x|}$ $(x \neq 0)$ （≅ 何桁あってるか）．

2．簡易機械実数

・**2進浮動小数：** 固定された桁数 n について，

$$a = \pm 2^e (f_0.f_1 f_2 \cdots f_n)_2 = \pm 2^e \sum_{i=0}^{n} f_i 2^{-i} = \pm 2^e \left(f_0 + \frac{f_1}{2} + \frac{f_2}{4} + \cdots + \frac{f_n}{2^n} \right)$$

ここで， $f_i \in \{0,1\}$ は 2^{-i} の位の数字，指数 $e\,(e_{\min} \le e \le e_{\max})$ は整数．

［例］ $a = -2^2 (1.011)_2 = -2^2 \left(1 + \dfrac{0}{2} + \dfrac{1}{4} + \dfrac{1}{8} \right) = -4 \times \dfrac{11}{8} = -5.5$ ．

・**正規数：** 1の桁の数字 $f_0 = 1$ の数．

・**簡易機械実数：** $\mathbb{F} = \underbrace{\left\{ \pm 2^e (1.f_1 f_2 \cdots f_n)_2 \mid e_{\min} \le e \le e_{\max},\ f_i \in \{0,1\}\ (1 \le i \le n) \right\}}_{\text{正規数全体の集合}} \cup \{0\}$

正規数の**最大絶対値：** $A_{\max} = 2^{e_{\max}} (1.1 \cdots 1)_2 = 2^{e_{\max}} (2 - 2^{-n}) \cong 2^{e_{\max}+1}$

オーバーフロー： 計算値の絶対値が $A_{\max}$ を超えること．

正規数の**最小絶対値：** $A_{\min} = 2^{e_{\min}} (1.0 \cdots 0)_2 = 2^{e_{\min}}$

アンダーフロー： 計算値の絶対値が $A_{\min}$ を下回ること．

3．簡易機械実数への丸め（四捨五入）

一般の実数（無限小数）： $x = 2^e (1.f_1 \cdots f_n f_{n+1} \cdots)_2 \notin \mathbb{F}$

切り捨て： $x_0 = 2^e (1.f_1 \cdots f_n)_2 \in \mathbb{F}$

切り上げ： $x_1 = 2^e \left\{ (1.f_1 \cdots f_n)_2 + 2^{-n} \right\} \in \mathbb{F}$

四捨五入： $\hat{x} = (x_0, x_1$ のうち x に近い方$)$

＜丸め誤差解析＞ $x_0 < x < x_1$，$x_1 - x_0 = 2^{e-n}$． $\therefore |\hat{x} - x| \le \dfrac{1}{2}|x_1 - x_0| = 2^{e-n-1}$．

$$\therefore \frac{|\hat{x} - x|}{|x|} \le \frac{2^{e-n-1}}{2^e (1.f_1 \cdots f_n f_{n+1} \cdots)_2} \le \frac{2^{e-n-1}}{2^e (1.0 \cdots 00 \cdots)_2} = 2^{-n-1}$$

☆ 丸めの相対誤差： $\dfrac{|\hat{x} - x|}{|x|} \le \mathrm{u}$ ，$\mathrm{u} = 2^{-n-1}$ を**丸め誤差単位**という．//

４．実際のシステム(IEEE754)

	ビット長 $m+n+2$	n	$e_{\min}$	$e_{\max}$	u	$A_{\min}$	$A_{\max}$
単精度(float)	32	23	-126	127	6.0×10^{-8}	1.2×10^{-38}	3.4×10^{38}
倍精度(double)	64	52	-1022	1023	1.1×10^{-16}	2.2×10^{-308}	1.8×10^{308}

・ビット配列： $\boxed{s}\boxed{e_m\ \cdots\ e_1\ \ e_0}\boxed{f_1\ \ f_2\ \cdots\ f_n}$　　$m+n+2$ ビット長

符号部　指数部　少数部

・符号： $(-1)^s=\begin{cases}1 & ,s=0,\\ -1 & ,s=1.\end{cases}$　　・指数： $e=(e_m\cdots e_0)_2-2^m+1$

・少数： $f=(0.f_1\cdots f_n)_2$

☆ $e_{\min}=2-2^m, e_{\max}=2^m-1$ として， $e_{\min}-1\le e\le e_{\max}+1$.

○　正規数，副正規数，０，±∞，NaN　の５種類の機械実数を表記.

1) 正規数： $e_{\min}\le e\le e_{\max}$ の場合．普通の大きさ絶対値を持つ非零数.

$$a=(-1)^s 2^e(1+f)=(-1)^s 2^e(1.f_1f_2\cdots f_n)_2$$

2) 副正規数： $e=e_{\min}-1, f\ne0$ の場合．小さい絶対値を持つ非零数.

$$a=(-1)^s 2^{e_{\min}} f=(-1)^s 2^{e_{\min}}(0.f_1f_2\cdots f_n)_2$$

3) ０： $e=e_{\min}-1, f=0$ の場合．零を表す．符号は無視される.

$$a=0$$

4) ±∞ ： $e=e_{\max}+1, f=0$ の場合．正規数より絶対値が大きい数.

$$a=(-1)^s\infty$$

5) NaN： $e=e_{\max}+1, f\ne0$ の場合．Not a Number(無茶苦茶)．不正規な演算($0/0, \infty-\infty, \sqrt{-1}$ など)の結果.

$$a=\mathrm{NaN}$$

［例］ $e=2.718281828\cdots$ を倍精度正規数に丸めた a の絶対誤差を見積もる.

$\dfrac{|a-e|}{|e|}\le \mathrm{u}$ より， $|a-e|\le \mathrm{u}|e|<1.1\times10^{-16}\times2.8=3.08\times10^{-16}<3.1\times10^{-16}$.//

練習問題

(1) $\pi=3.141592\cdots$ を倍精度正規数に丸めた p の絶対誤差を見積もれ.

(2) 4^k が倍精度でオーバーフローする最小の k を求めよ.