数値計算

大阪大学基礎工学部

永原正章

2012年4月19日

練習問題(前回)

数値計算に偶然誤差は存在するか?

<u>ヒント</u>

- 偶然誤差=「いかに熟練者でも制御しえない偶然的に発生する誤差」
- 確率的な誤差
- 数値計算(コンピュータによる計算)において,確率的な事象とは何か?

練習問題の解答例

偶然誤差は存在する.たとえば

- モンテカルロ法で求めた近似解の誤差
- 偶発的に起こった停電によるプログラムの停止
- クライアント／サーバシステムにおけるデータ通信障害
- 強力な外部ノイズによるメモリに記憶されたビットの反転

など.

- コンピュータを使用する状況で「確率的に起こること」とは何かを考えればよい.

数値計算における誤差

- ここからは,系統誤差のみを扱います.

コンピュータにおける数値の表現

- コンピュータが理解できる数値＝有限桁の2進数
- 2進数による数値の表現:IEEE754: IEEE 標準規格
- なぜ標準化が必要なのか?
  - パソコンの機種(メーカー)ごとに表現が異なると,機種ごとに別のプログラムが必要.
  - プログラムの汎用性がなくなる
- たとえば

$$
\begin{aligned}
8.75 &= 1000.11_2 \\
&= 01000001000011000000000000000000_2
\end{aligned}
$$

  - 添え字 '2' は「2進数表現」を表す.

IEEE754の浮動小数点数表現⋆⋆⋆

$$(-1)^{\mathsf{c}} \times \mathsf{S} \times 2^{\mathsf{e}} = (-1)^{\mathsf{c}} \times 1.\mathsf{d}_1\mathsf{d}_2\cdots\mathsf{d}_{\mathsf{p}-1} \times 2^{\mathsf{e}}$$

- c: 符号ビット.
  - $\mathsf{c} = 0$ または 1.
- $\mathsf{S} = 1.\mathsf{d}_1\mathsf{d}_2\cdots\mathsf{d}_{\mathsf{p}-1}$: 仮数(かすう)
  - $\mathsf{d}_\mathsf{i} = 0$ または 1.
  - p は仮数のビット長.
  - 仮数 S は必ず $1_2 \le \mathsf{S} < 10_2$ を満たす(正規化数).
- e: 指数.
  - $-126 \le \mathsf{e} \le 127$ (単精度, 8ビット)
- たとえば $-1101.1111_2 = (-1)^1 \times 1.1011111_2 \times 2^3$

IEEE754の浮動小数点数表現$\star\star\star$

$$(-1)^c \times S \times 2^e = (-1)^c \times 1.d_1 d_2 \cdots d_{p-1} \times 2^e$$

- c: 符号ビット.
  - $c = 0$ または 1.
- $S = 1.d_1 d_2 \cdots d_{p-1}$: 仮数(かすう)
  - $d_i = 0$ または 1.
  - p は仮数のビット長.
  - 仮数 S は必ず $1_2 \leq S < 10_2$ を満たす(正規化数).
- e: 指数.
  - $-126 \leq e \leq 127$ (単精度, 8ビット)
- たとえば $-1101.1111_2 = (-1)^1 \times 1.1011111_2 \times 2^3$

IEEE754の浮動小数点数表現⋆⋆⋆

$$(-1)^c \times S \times 2^e = (-1)^c \times 1.d_1 d_2 \cdots d_{p-1} \times 2^e$$

- c: 符号ビット.
  - $c = 0$ または 1.
- $S = 1.d_1 d_2 \cdots d_{p-1}$: 仮数(かすう)
  - $d_i = 0$ または 1.
  - p は仮数のビット長.
  - 仮数 S は必ず $1_2 \leq S < 10_2$ を満たす(正規化数).
- e: 指数.
  - $-126 \leq e \leq 127$ (単精度, 8ビット)
- たとえば $-1101.1111_2 = (-1)^1 \times 1.1011111_2 \times 2^3$

IEEE754の浮動小数点数表現⋆⋆⋆

$$(-1)^{\mathsf{c}} \times \mathsf{S} \times 2^{\mathsf{e}} = (-1)^{\mathsf{c}} \times 1.\mathsf{d}_1\mathsf{d}_2\cdots\mathsf{d}_{\mathsf{p}-1} \times 2^{\mathsf{e}}$$

- c: 符号ビット.
  - $\mathsf{c} = 0$ または 1.
- $\mathsf{S} = 1.\mathsf{d}_1\mathsf{d}_2\cdots\mathsf{d}_{\mathsf{p}-1}$: 仮数(かすう)
  - $\mathsf{d}_i = 0$ または 1.
  - p は仮数のビット長.
  - 仮数 S は必ず $1_2 \leq \mathsf{S} < 10_2$ を満たす(正規化数).
- e: 指数.
  - $-126 \leq \mathsf{e} \leq 127$ (単精度, 8ビット)
- たとえば $-1101.1111_2 = (-1)^1 \times 1.1011111_2 \times 2^3$

IEEE754の浮動小数点数表現

$$(-1)^{\mathsf{c}} \times \mathsf{S} \times 2^{\mathsf{e}} = (-1)^{\mathsf{c}} \times 1.\mathsf{d}_1\mathsf{d}_2\cdots\mathsf{d}_{\mathsf{p}-1} \times 2^{\mathsf{e}}$$

IEEE 754 の浮動小数点数表現の精度.数字の単位はビット.

	仮数S	指数e	符号c	合計	C言語
単精度	23	8	1	32	`float`
倍精度	52	11	1	64	`double`

IEEE754の浮動小数点数表現

- 絶対値が最小の数(単精度): $\pm 1.0\ldots0 \times 2^{\mathsf{e}_{\mathsf{min}}}$

$$\mathsf{m} = 2^{\mathsf{e}_{\mathsf{min}}} = 2^{-126} \approx 1.2 \times 10^{-38}$$

  $\mathsf{c} = 0$ or 1, $\mathsf{S} = 1.00\ldots0$, $\mathsf{e} = -126$

- 絶対値が最大の数(単精度): $\pm 1.1\ldots1 \times 2^{\mathsf{e}_{\mathsf{max}}}$

$$\mathsf{m} = 2^{\mathsf{e}_{\mathsf{max}}}(2 - 2^{-23}) \approx 2^{\mathsf{e}_{\mathsf{max}}+1} = 2^{128} \approx 3.4 \times 10^{38}$$

  $\mathsf{c} = 0$ or 1, $\mathsf{S} = 1.11\ldots1$, $\mathsf{e} = 127$

- 実は,± 0 や $\pm\infty$ の表現もある(後述).

IEEE754の2進数表現⋆⋆⋆

c	e	$d_1d_2\dots d_{23}$

3130　　2322　　0

- 先頭は符号 c の1ビット
- 次は指数 e の8ビット(単精度).ただし,バイアス表現(実際の値に127を足す)
  - $00000001_2 \leftrightarrow -126$
  - $00000010_2 \leftrightarrow -125$
  - $11111110_2 \leftrightarrow 127$
  - 00000000_2 および 11111111_2 には特別な意味がある(後述)
- 最後に仮数部の $d_1, \dots, d_{23}$ の23ビット
- 合計32ビット(単精度)

IEEE754の2進数表現⋆⋆⋆

[例題] 10進数 8.75 を IEEE 754 形式の2進数に変換せよ.
まず,

$$\mathbf{8.75 = 8 + 0.75 = 8 + 0.5 + 0.25 = 2^3 + \frac{1}{2} + \frac{1}{2^2}}$$

であるので,

$$\mathbf{8 = 2^3 = 1000_2, \quad 0.5 = 1/2 = 0.1_2, \quad 0.25 = 1/2^2 = 0.01_2}$$

$$\therefore \mathbf{8.75 = 1000.11_2 = 1.00011_2 \times 2^3}$$

と表現できる.

IEEE754の2進数表現$\star\star\star$

符号はプラスであるので $\mathsf{c} = 0$. 指数は 3 であるが,バイアス表現にするために 127 を足して

$$\begin{aligned} 3 + 127 = 130 = 128 + 2 = 2^7 + 2^1 &= 1000\ 0000_2 + 10_2 \\ &= 1000\ 0010_2 \end{aligned}$$

となる. 仮数は $\mathsf{S} = 1.00011$ であるので, 小数点以下を23 ビットで表して,

$$\mathsf{d}_1\mathsf{d}_2\ldots\mathsf{d}_{23} = 0001\ 1000\ 0000\ 0000\ 0000\ 000_2$$

である.以上を並べると

$$0 \mid 1000\ 0010 \mid 0001\ 1000\ 0000\ 0000\ 0000\ 000_2$$

が得られる(縦棒は符号部・指数部・仮数部の区切りを表す).

練習問題

10進数 -15.125 をIEEE754 の形式 にしたがって単精度の2進浮動小数点数に変換せよ.

練習問題の解答例

まず,

$$
\begin{aligned}
-15.125 &= -(8+4+2+1+0.125) \\
&= -\left(2^3+2^2+2^1+2^0+\frac{1}{2^3}\right)
\end{aligned}
$$

であるので,

$$
\begin{aligned}
&8 = 2^3 = 1000_2, \quad 4 = 2^2 = 0100_2, \quad 2 = 2^1 = 0010_2, \\
&1 = 2^0 = 0001_2, \quad 0.125 = 1/2^3 = 0.001_2
\end{aligned}
$$

$$
\therefore -15.125 = -1111.001_2 = -1.111001_2 \times 2^3
$$

と表現できる.

練習問題の解答

符号はマイナスであるので $\mathsf{c} = 1$. 指数は 3 であるが,バイアス表現にするために 127 を足して

$$
\begin{aligned}
3 + 127 = 130 = 128 + 2 = 2^7 + 2^1 &= 1000\,0000_2 + 10_2 \\
&= 1000\,0010_2
\end{aligned}
$$

となる. 仮数は $\mathsf{S} = 1.111001_2$ であるので, 小数点以下を23 ビットで表して,

$$
\mathsf{d}_1\mathsf{d}_2\ldots\mathsf{d}_{23} = 1110\,0100\,0000\,0000\,0000\,000_2
$$

である.以上を並べると

$$
1 \mid 1000\,0010 \mid 1110\,0100\,0000\,0000\,0000\,000_2
$$

が得られる(縦棒は符号部・指数部・仮数部の区切りを表す).

IEEE754の例外処理

1. オーバーフロー: 浮動小数点数演算結果の絶対値が最大数 M を上回る現象,$\pm\infty$
2. アンダーフロー: 演算結果の絶対値が最小数 m を下回る現象,非正規化数
3. ゼロ割り: 数値を 0 で割ること,$\pm\infty$
4. 不正: $\sqrt{-1}$ や $0 \times \infty$,$0/0$,∞/∞, $\infty - \infty$ などの演算で生じる,`NaN` (Not a Number)
5. 不正確: 通常の桁数で表現しきれないとき(オーバーフロー,アンダーフロー,丸め誤差)

IEEE 754 における特殊な数

$f = d_1 d_2 \dots d_{p-1}$ とする.$\pm$ の符号は符号ビット c により定義する.

指数	指数(bias表現)	仮数と符号	意味
$e_{min} - 1$	$0000\ 0000_2$	$\pm 1.f,\ f = 0_2$	± 0
$e_{min} - 1$	$0000\ 0000_2$	$\pm 1.f,\ f \neq 0_2$	$\pm 0.f \times 2^{e_{min}}$
$e = e_{max} + 1$	$1111\ 1111_2$	$\pm 1.f,\ f = 0_2$	$\pm\infty$
$e = e_{max} + 1$	$1111\ 1111_2$	$\pm 1.f,\ f \neq 0_2$	NaN