本レジュメは 竹 V 照著「Fortran II」第 3 章, 第 4 章を主に参考にしている.

浮動小数点操作

実数の数体系

実数の数体系の定義

実数の数体系を

$$x = \begin{cases} 0 \\ \text{または} \\ s \times b^e \times \sum_{k=1}^{p} (f_{k10} \times b^{-k}) \end{cases} \tag{1}$$

と定義する. ここで b および p は 2 以上の整数[1)], f_k は b 未満の非負の整数 (ただし, $f_1 \neq 0$[2)]), s は -1 または, $+1$ であり, e はある最小値の整数 $e_{\min}$ からある最大値の整数 $e_{\max}$ までの間の整数である. これは実数 x を指数部 b^e と小数部 $\sum_{k=1}^{p} f_k \times b^{-k}$ に分けて表現する正規化した浮動小数点表現である[3)].

単精度 (32bit 2 進数) での表現

32bit 2 進数の場合 (1) は,

[1)] p は 1 でもいいと思われる.
[2)] 常に小数部を小数点以下 1 桁からの表現にするため
[3)] レジュメ「実数の浮動小数点表現と誤差」(1) 式を, (1) の形で表すと,

$$\begin{aligned} x &= \pm(0.f_1 f_2 \cdots f_m)_\beta \beta^{\pm E} \\ &= 0 \quad or \quad s \times \beta^{\pm E} \times \sum_{k=1}^{m} ((f_k)_{10} \times \beta^{-k}). \end{aligned}$$

となり, 仮数部が小数部と同じで指数部はそのまま指数部として表している.

$$x = \begin{cases} 0 \\ \text{または} \\ s \times 2^e \times \left(\frac{1}{2} + \sum_{k=2}^{24} (f_{k10} \times 2^{-k}) \right) \end{cases} \tag{2}$$

となる[4)]. ここで, $-125 \leq e \leq 128$ である. 指数部 e は 8bit 分の領域があるため, 本来 -127 から 128 までを表現できるはずである. しかしながら, 無限大と非数 (not-a-number, NaN) を表すために 2 つ少なくなっている.

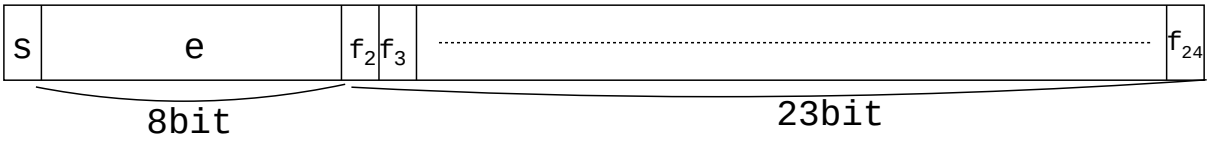

図 1: 単精度浮動小数点数体系の模式図. s は符号, e は指数部, f_k は仮数部をそれぞれ表している.

倍精度 (64bit 2 進数) での表現

倍精度実数の数体系は, (1) によって定義した数体系の p を, 単精度のときの約 2 倍にしたものである. (1) の形で倍精度実数の数体系を表すと,

$$x = \begin{cases} 0 \\ \text{または} \\ s \times 2^e \times \left(\frac{1}{2} + \sum_{k=2}^{53} (f_{k10} \times 2^{-k}) \right) \end{cases} \tag{3}$$

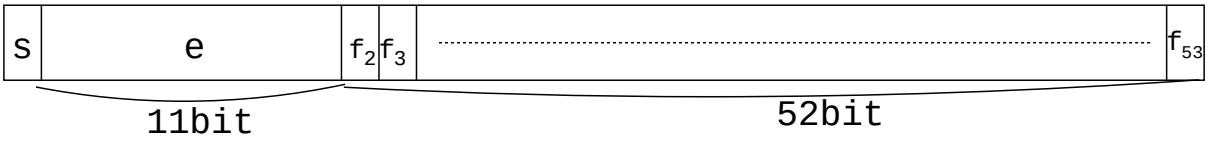

図 2: 倍精度浮動小数点数体系の模式図. s は符号, e は指数部, f_k は仮数部をそれぞれ表している.

[4)]項に $\frac{1}{2}$ が入るのは $f_1 \neq 0$ であるため. 2 進数の場合は $f_1 = 1$ と固定される.

組み込み関数での表現

使用している計算機の $e_{\max}, e_{\min}, b$ の値, および指数を 10 進数に換算した値は, それぞれ数値問合せ組み込み関数 MAXEXPONENT, MINEXPONENT, RADIX および RANGE によって出力できる. (1) の p の値は数値問合せ組み込み関数 DIGITS と PRECISION によって知ることができる. これらの値から基本実数型の数体系で扱える数の範囲が決まる. また, 最大の数および最小の正の数は, 組み込み関数 HUGE および TINY によって知ることができる.

マシンイプシロン

p の値が有限であることにより,

$$1 + \varepsilon_{\mathrm{M}} \neq 1 \tag{4}$$

となる最小の数 ε_{M} (> 0) が存在する. この ε_{M} を「マシンイプシロン (machine epsilon)」という. マシンイプシロンは組み込み関数 EPSILON を用いて調べることができる. 関数 epsilon(x) での値は x の数体系でのマシンイプシロン b^{1-p} となる[5)].

数値問い合わせ組み込み関数の利用

以下のプログラムを実行することにより, 使用している計算機の実数についての諸量を知ることができる.

単精度実数

次に示すのは単精度実数の諸量を調べるプログラムである.

5)
$$1 \approx (0.1 f_2 \cdots f_p)_b \times b^1.$$
この桁より小さな桁では桁落ちを起こしてしまう. よって, この値に桁落ちを起こさないぎりぎりの数 ε は小数部 $(0.1 f_2 \cdots f_p)_b$ の最小値である. よって,

$$\begin{aligned} \varepsilon &= (0.00\cdots1)_b \times b^1 \\ &= (0.100\cdots0)_b \times b^{1-(p-1)} \\ &= \frac{1}{b} \times b^{1-(p-1)} \\ &= b^{1-p} \end{aligned}$$

である.

```
1  program main
2  implicit none
3  character(8) :: DATE
4  write(*,*)
5  call date_and_time(DATE)
6  write(*,*) DATE(1:4)//'年'//DATE(5:6)//'月'//DATE(7:8)//'
日実施'
7  write(*,*)
8  write(*,*)'単精度実数型種別値        : ', kind(0.0)
9  write(*,*)'単精度実数型基数          : ', radix(0.0)
10 write(*,*)'単精度実数型有効ビット数 : ', digits(0.0)
11 write(*,*)'単精度実数型10進精度      : ', precision(0.0)
12 write(*,*)'単精度実数型イプシロン    : ', epsilon(0.0)
13 write(*,*)'単精度実数型正最小値      : ', tiny(0.0)
14 write(*,*)'単精度実数型最大値        : ', huge(0.0)
15 write(*,*)'単精度実数型最大指数      : ', maxexponent(0.0)
16 write(*,*)'単精度実数型最小指数      : ', minexponent(0.0)
17 write(*,*)'単精度実数型10進指数範囲 : ', range(0.0)
18 end program main
```

実行した結果は,

2011年05月09日実施

単精度実数型種別値	:	4
単精度実数型基数	:	2
単精度実数型有効ビット数	:	24
単精度実数型10進精度	:	6
単精度実数型イプシロン	:	1.19209290E-07
単精度実数型正最小値	:	1.17549435E-38
単精度実数型最大値	:	3.40282347E+38
単精度実数型最大指数	:	128
単精度実数型最小指数	:	-125
単精度実数型10進指数範囲	:	37

となどとなる.

倍精度実数

次に示すのは倍精度実数の諸量を調べるプログラムである.

```
1  program main
2  implicit none
3  !integer, parameter :: DBL = selected_real_kind(15,99)
4  character(8) :: DATE
5  write(*,*)
6  call date_and_time(DATE)
7  write(*,*) DATE(1:4)//'年'//DATE(5:6)//'月'//DATE(7:8)//'
日実施'
8  write(*,*)
9  write(*,*)'倍精度実数型種別値        : ', kind(0.0d0)
10 write(*,*)'倍精度実数型基数          : ', radix(0.0d0)
11 write(*,*)'倍精度実数型有効ビット数 : ', digits(0.0d0)
12 write(*,*)'倍精度実数型10進精度      : ', precision(0.0d0)
13 write(*,*)'倍精度実数型イプシロン    : ', epsilon(0.0d0)
14 write(*,*)'倍精度実数型正最小値      : ', tiny(0.0d0)
15 write(*,*)'倍精度実数型最大値        : ', huge(0.0d0)
16 write(*,*)'倍精度実数型最大指数      : ', maxexponent(0.0d0)
17 write(*,*)'倍精度実数型最小指数      : ', minexponent(0.0d0)
18 write(*,*)'倍精度実数型10進指数範囲 : ', range(0.0d0)
19 end program main
```

実行した結果は,

2011 年 05 月 09 日実施

倍精度実数型種別値	:	8
倍精度実数型基数	:	2
倍精度実数型有効ビット数	:	53
倍精度実数型 10 進精度	:	15
倍精度実数型イプシロン	:	2.22044604925031308E-016
倍精度実数型正最小値	:	2.22507385850720138E-308
倍精度実数型最大値	:	1.79769313486231571E+308
倍精度実数型最大指数	:	1024
倍精度実数型最小指数	:	-1021
倍精度実数型 10 進指数範囲	:	307

などとなる.

付録: オイラー・マクローリンの公式の導出

台形則近似の (5) 式はオイラーマクローリンの公式である. 以下では, 長田直樹著雑誌「理系への数学」連載「お話:数値解析第3回」を参考にオイラー・マクローリンの公式を導く. なお、連載記事は http://www.cis.twcu.ac.jp/ osada/rikei/rikei2008-7.pdf にて PDF 形式で閲覧することができる.

命題

関数 $f(x)$ は区間 $[a,b]$ で C^{2m+2} 級であるとする. この時,

$$I_N - I = \sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} h^{2k} \left[f^{(2k-1)}(b) - f^{(2k-1)}(a) \right] + O(h^{2m+2}), \ \ (h \to +0) \quad \text{(A. 1)}$$

が成り立つ. 但し, $x_j = a + jh$ である. また, $B_i(t)$ はベルヌーイ多項式[a], B_i はベルヌーイ数である.

[a] ベルヌーイ多項式 $B_i(t)$ の定義式は,

$$B_n(t) = \sum_{k=0}^{n} \binom{n}{k} B_k t^{n-k}.$$

ここで $\binom{n}{k}$ は二項係数で

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k)}{k!}$$

である. また, B_k はベルヌーイ数と呼ばれ,

$$B_0 = 1 \quad or \quad \sum_{k=0}^{n-1} \binom{n}{k} B_k = 0 (n = 2, 3 \cdots)$$

と定義される.

今回の証明ではベルヌーイ多項式, ベルヌーイ数ともに $i = 2$ の場合

$$B_2(t) = t^2 - t + \frac{1}{6}, \tag{A. 2}$$

$$B_2 = \frac{1}{6} \tag{A. 3}$$

を用いる.

証明

$j = 0, ..., n-1$: $k = 1, ..., m+1$ に対し, $I_{j,k}$ を

$$I_{j,k} = \frac{1}{(2k)!} \int_0^h B_{2k}\left(\frac{t}{h}\right) f^{(2k)}(x_j + t)dt \tag{A. 4}$$

とおく. $k = 1$ のとき (A. 2), (A. 3) を用いると, $I_{j,1}$ は部分積分を用いて,

$$\begin{aligned}
I_{j,1} &= \frac{1}{2!}\int_0^h \left(\frac{t^2}{h^2} - \frac{t}{h} + B_2\right) f''(x_j + t)dt \\
&= \frac{1}{2!}\left[\left(\frac{t^2}{h^2} - \frac{t}{h} + B_2\right) f'(x_j + t)\right]_0^h - \frac{1}{2!}\int_0^h \left(\frac{2t}{h^2} - \frac{1}{h}\right) f'(x_j + t)dt \\
&= \frac{B_2}{2!}[f'(x_j + h) - f'(x_j)] - \frac{1}{2!}\left[\left(\frac{2t}{h^2} - \frac{1}{h}\right) f(x_j + t)\right]_0^h + \frac{1}{2!}\int_0^h \frac{2}{h^2} f(x_j + t)dt \\
&= \frac{B_2}{2!}[f'(x_j + h) - f'(x_j)] - \frac{1}{2h}[f(x_{j+1}) + f(x_j)] + \frac{1}{h^2}\int_0^h f(x_j + t)dt \\
&= \frac{B_2}{2!}[f'(x_j + h) - f'(x_j)] - \frac{1}{2h}[f(x_{j+1}) + f(x_j)] + \frac{1}{h^2}\int_{x_j}^{x_{j+1}} f(t)dt
\end{aligned} \tag{A. 5}$$

となる. (A. 5) を $j = 0, ..., n-1$ について加えると,

$$\begin{aligned}
\sum_{j=0}^{n-1} I_{j,1} &= \frac{B_2}{2!}[f'(a + h) + f'(a + 2h) + \cdots + f'(x_{n-1}) + f'(b) - f'(a) - \cdots - f'(x_{n-1})] \\
&\quad - \frac{1}{2h}[f(a) + \cdots + 2f(x_{n-1}) + f(b)] + \frac{1}{h^2}\left(\int_a^{x_1} f(t)dt + \cdots + \int_{x_{n-1}}^b f(t)dt\right) \\
&= \frac{B_2}{2!}[f'(b) - f'(a)] - \frac{1}{h^2} I_N + \frac{1}{h^2} I.
\end{aligned} \tag{A. 6}$$

$k = 2, ..., m+1$ のとき, ベルヌーイ多項式の性質,

$$B_k'(t) = kB_{k-1}(t) \tag{A. 7}$$

$$B_{2k}(1) = B_{2k}(0) = B_{2k} \tag{A. 8}$$

$$B_{2k-1}(1) = B_{2k-1}(0) = 0 \tag{A. 9}$$

に注意すると,

$$
\begin{aligned}
I_{j,k} &= \frac{1}{(2k)!}\left[B_{2k}\left(\frac{t}{h}\right)f^{(2k-1)}(x_j+t)\right]_0^h - \frac{1}{(2k)!}\int_0^h \frac{1}{h}B'_{2k}\left(\frac{t}{h}\right)f^{(2k-1)}(x_j+t)dt \\
&= \frac{1}{(2k)!}\left[B_{2k}(1)f^{(2k-1)}(x_j+h) - B_{2k}(0)f^{(2k-1)}(x_j)\right] \\
&\quad - \frac{1}{(2k)!h}\int_0^h 2kB_{2k-1}\left(\frac{t}{h}\right)f^{(2k-1)}(x_j+t)dt \\
&= \frac{1}{(2k)!}\left[B_{2k}(1)f^{(2k-1)}(x_j+h) - B_{2k}(0)f^{(2k-1)}(x_j)\right] \\
&\quad - \frac{1}{(2k-1)!h}\left[B_{2k-1}\left(\frac{t}{h}\right)f^{(2(k-1))}(x_j+t)\right]_0^h \\
&\quad + \frac{1}{(2k-1)!h}\int_0^h \frac{1}{h}B'_{2k-1}\left(\frac{t}{h}\right)f^{(2(k-1))}(x_j+t)dt.
\end{aligned}
$$

ここで (A. 9) 式より第 2 項目が零になるので

$$
\begin{aligned}
I_{j,k} &= \frac{1}{(2k)!}\left[B_{2k}(1)f^{(2k-1)}(x_j+h) - B_{2k}(0)f^{(2k-1)}(x_j)\right] \\
&\quad + \frac{1}{(2k-1)!h^2}\int_0^h (2k-1)B_{2(k-1)}\left(\frac{t}{h}\right)f^{(2(k-1))}(x_j+t)dt \\
&= \frac{B_{2k}}{(2k)!}\left[f^{(2k-1)}(x_j+1) - f^{(2k-1)}(x_j)\right] + \frac{1}{h^2}I_{j,k-1}. \qquad \text{(A. 10)}
\end{aligned}
$$

よって,

$$
I_{j,k} = \frac{1}{h^2}I_{j-1,k} + \frac{B_{2k}}{(2k)!}\left[f^{(2k-1)}(x_{j+1}) - f^{(2k-1)}(x_j)\right] \qquad \text{(A. 11)}
$$

となる. (A. 11) を $k = 2, \cdots m+1$ まで計算する. $k = 2$ のときは

$$
I_{j,2} = \frac{1}{h^2}I_{j,1} + \frac{B_4}{(4)!}\left[f^{(3)}(x_{j+1}) - f^{(3)}(x_j)\right]
$$

となる. この $I_{j,2}$ を使って $k = 3$ のときの $I_{j,3}$ を求める.

$$
\begin{aligned}
I_{j,3} &= \frac{1}{h^2}I_{j,2} + \frac{B_6}{(6)!}\left[f^{(5)}(x_{j+1}) - f^{(5)}(x_j)\right] \\
&= \frac{1}{h^2}\left(\frac{1}{h^2}I_{j,1} + \frac{B_4}{(4)!}\left[f^{(3)}(x_{j+1}) - f^{(3)}(x_j)\right]\right) + \frac{B_6}{(6)!}\left[f^{(5)}(x_{j+1}) - f^{(5)}(x_j)\right]
\end{aligned}
$$

よって, $k = m+1$ のときは

$$
I_{j,m+1} = \frac{1}{h^{2m}}I_{j,1} + \sum_{k=2}^{m+1}\frac{B_{2k}}{(2k)!h^{2k-1}}\left[f^{(2k-1)}(x_{j+1}) - f^{(2k-1)}(x_j)\right].
$$

さらに $j = 0, \cdots, n-1$ まで足し合わせると

$$\sum_{j=0}^{n-1} I_{j,1} = h^{2m} \sum_{j=0}^{n-1} I_{j,m+1} + \sum_{k=2}^{m+1} \frac{B_{2k}}{(2k)! h^{2k-1}} \left[f^{(2k-1)}(b) - f^{(2k-1)}(a) \right]. \qquad \text{(A. 12)}$$

(A. 6), (A. 12) より,

$$I_N - I = \sum_{k=1}^{m+1} \frac{B_{2k} h^{2k}}{(2k)!} \left[f^{(2k-1)}(b) - f^{(2k-1)}(a) \right] + R_{m+1}, \qquad \text{(A. 13)}$$

但し,

$$R_{m+1} = -h^{2m+2} \sum_{j=0}^{n-1} I_{j,m+1} \qquad \text{(A. 14)}$$

が言える. [0, 1] において $|B_{2n}(t)| \leq |B_{2n}|$ が成り立つ [6] ので,

$$R_{m+1} = -\frac{h^{2m+2}}{(2m+2)!} \int_0^h B_{2(m+1)} \left(\frac{t}{h} \right) \sum_{j=0}^{n-1} f^{(2m+2)}(x_j + t) dt \qquad \text{(A. 15)}$$

より,

$$|R_{m+1}| \leq \frac{h^{2m+2} |B_{2m+2}|}{(2m+2)!} \int_a^b |f^{(2m+2)}(t)| dt. \qquad \text{(A. 16)}$$

$f^{(2m+2)}(x)$ は区間 $[a, b]$ で連続だから

$$R_{m+1} = O(h^{(2m+2)}), \ \ (h \to +0). \qquad \text{(A. 17)}$$

したがって,

$$I_N - I = \sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} h^{2k} \left[f^{(2k-1)}(b) - f^{(2k-1)}(a) \right] + O(h^{2m+2}), \ \ (h \to +0). \qquad \text{(A. 18)}$$

証明終わり.

[6] 証明は割愛