

本レジюмеは竹澤照著「Fortran II」第3章, 第4章を主に参考にしている. またある数 (β 進表記) の k 桁目の数 f_k を α 進表記に変換したものを $({}^\beta f_k)_\alpha$ と表記している.

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

$$x = \begin{cases} 0 \\ \text{または} \\ s \times b^e \times \sum_{k=1}^p (({}^b f_k)_{10} \times b^{-k}) \end{cases} \quad (1)$$

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 $\sum_{k=1}^p ({}^b f_k)_{10} \times b^{-k}$ 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 (IEEE) の場合式 (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юме「実数の浮動小数点表現と誤差」式 (1) を, 本レジюме式 (1) の形で表すと,

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

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.

$$x = \begin{cases} 0 \\ \text{または} \\ s \times 2^e \times \left(\frac{1}{2} + \sum_{k=2}^{24} (({}^2f_k)_{10} \times 2^{-k}) \right) \end{cases} \quad (2)$$

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 1 つ, さらに, 0 と非正規数⁵⁾を表すために 1 つ, 合計 2 つ少なくなっている.

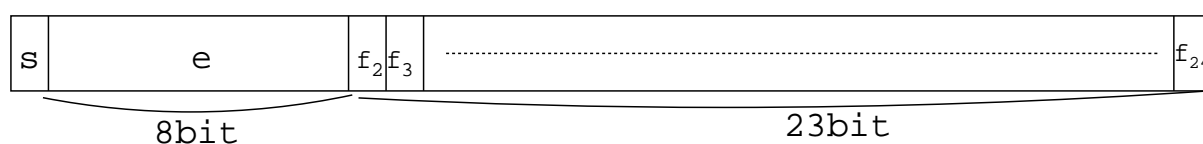


図 1: 単精度浮動小数点数体系の模式図. s は符号, e は指数部, f_k は仮数部をそれぞれ表している.

倍精度 (64bit 2 進数) での表現

倍精度実数の数体系は, 式 (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. 式 (1) の形で倍精度実数の数体系を表すと,

$$x = \begin{cases} 0 \\ \text{または} \\ s \times 2^e \times \left(\frac{1}{2} + \sum_{k=2}^{53} (({}^2f_k)_{10} \times 2^{-k}) \right) \end{cases} \quad (3)$$

となる. ここで, e の範囲は $-1021 \leq e \leq 1024$ である.

Fortran 90 における数値問い合わせ組み込み関数の利用

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 Fortran90 数値問い合わせ組み込み関数 MAXEXPONENT, MINEXPONENT,

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.

⁵⁾仮数部が 0 で始まる数.

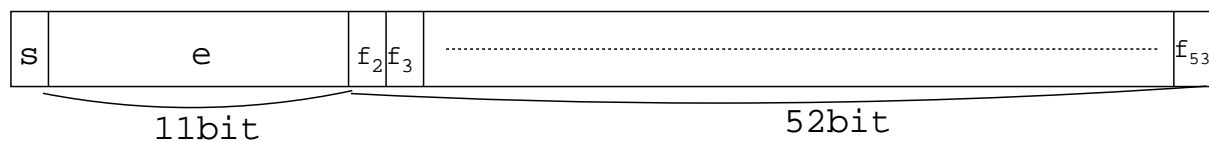


図 2: 倍精度浮動小数点数体系の模式図. s は符号, e は指数部, f_k は仮数部をそれぞれ表している.

RADIX および RANGE によって出力できる. 式 (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,

$$1 + \varepsilon_M \neq 1 \quad (4)$$

となる最小の数 $\varepsilon_M (> 0)$ 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⁶⁾.

付属のプログラム (example1.f90) 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.

⁶⁾1 に加えて桁落ちしない最小の数を求める.

$$1 \approx 0.1({}^b f_2)_{10}({}^b f_3)_{10} \cdots ({}^b f_p)_{10} \times b^1.$$

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 $0.1f_2 \dots f_b$ の最小値である. よって,

$$\begin{aligned} \varepsilon &= 0.00 \cdots 1 \times b^1 \\ &= 0.100 \cdots 0 \times b^{1-(p-1)} \\ &= \frac{1}{b} \times b^{1-(p-1)} \\ &= b^{1-p} \end{aligned}$$

である.

単精度実数

example1.f90 を実行した結果は,

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

example2.f90 を実行した結果は,

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.

参考文献

竹澤 照, 1997, 「Fortran II: 数値計算」, 共立出版, ISBN 4320028686

伊理正夫, 藤野和建, 1985, 「数値計算の常識」 共立出版, ISBN 4320013433

川上一郎, 2009, 「数値計算の基礎」 URL: <http://www7.ocn.ne.jp/kawa1/>