本レジュメは 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

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

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 と仮数部 $\sum_{k=1}^{p} ({}^{b}f_k)_{10} \times b^{-k}$ 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 進数の場合 (**??**) 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, 本レジュメ (**??**) 式 の形で表すと,

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

となる. 竹澤照著「Fortran II」第 3 章, 第 4 章」では, 仮数部を小数部と表現しているが, このレジュメでは前のレジュメにあわせて仮数部としている.

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

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, (**??**) 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. (**??**) の形で倍精度実数の数体系を表すと,

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

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. (**??**) 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

[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.

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.

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.

---

5)

$$1 \approx^{b} 0.1 f_2 \cdots f_p \times b^1.$$

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,

$$\begin{aligned} \varepsilon &= {}^{b}0.00\cdots1 \times b^1 \\ &= b0.100\cdots0 \times b^{1-(p-1)} \\ &= \frac{1}{b} \times b^{1-(p-1)} \\ &= b^{1-p} \end{aligned}$$

である.