演習問題1.問2の略解

大石進一

May 27, 2003

§1 Dekkerの定理の証明

定義　有理数xがtビットの浮動小数点数とは正の整数tが存在して、

$$x = m2^k$$

と書けることをいう。ただし、mとkは整数で、$|m| < 2^t$を満たしているものとする。また、正規化条件

$$2^{t-1} \leq |m|$$

を仮定する。

以上の正規化の条件の下では2^kが最下位桁($t$bit目)となるので、$u = 2^k$をulp(x)と表す。0は正規化条件を満たすことができないので別に定義し、これも$t$bit浮動小数点数とよぶことにする(例えば$0 = 0 * 2^0$など)。ulp(0)=0と定義しておく。マシンエプシロンは2^{1-t}となる。

浮動小数点数演算がfaithfulであること

実数rから$t$bit浮動小数点数xへの丸めの演算子を

$\bigcirc r = r$に最も近い$t$bit浮動小数点数x,2つあるときは偶数丸め

によって定義する。

このとき、$\cdot \in \{+, -, \times, /\}$として、$t$bit浮動小数点数$x, y$に対して

$$x \odot y = \bigcirc(x \cdot y)$$

が成り立つ浮動小数点演算をfaithfulという。

定理1[Sterbenz]　aとbを$t$bitの浮動小数点数とする。

$$\frac{1}{2} \leq \frac{a}{b} \leq 2$$

が成り立つなら、$a - b$は$t$bit浮動小数点数となる。演算がfaithfulなら

$$a \ominus b = a - b$$

が成り立つ。

証明　簡単のため、$a > b > 0$の場合を考える。ulp(a) $\geq$ ulp(b)であるからaとbはともにulp(b)の整数倍である。したがって、

(I) $a - b$もulp(b)の整数倍である。

仮定から、$a \leq 2b$であるから

(II) $a - b \leq b$が成り立つ。

(I),(II)から$a - b$は$t$bit浮動小数点数であることがわかる。faithfulの定義から

$$a \ominus b = \bigcirc(a - b) = a - b$$

が成り立つ。(QED)

定理2[Dekker]　aとbをt-bit浮動小数点数で$|a| \geq |b|$を満たすものとする。このとき、つぎのアルゴリズムによってnonoverlapping expansion $x + y$が計算でき、$a + b = x + y$ を満たす:

$$\textbf{function}\quad [x, y] = \textbf{FastTwoSum}(a, b)$$
$$x = a \oplus b;$$
$$b_V = x \ominus a;$$
$$y = b \ominus b_V;$$

証明　簡単のため、$a > b > 0$ の場合を考える。関数FastTwoSumの2行目を考える。

$$1 \leq \frac{x}{a} \leq 2$$

であるから、定理1により、

$$b_V = x \ominus a = x - a$$

となる。

ここで

$$x = a \oplus b = a + b + \mathrm{err}(a \oplus b)$$

と書くと、

$$|\mathrm{err}(a \oplus b)| \leq b \tag{1}$$

が示せる (少し後で証明する)。a も b も ulp(b) の整数倍であったから、$a+b$ も ulp(b) の整数倍である。$b < a \leq a \oplus b$ であるので、$a \oplus b$ も ulp(b) の整数倍となる。よって、

$$\mathrm{err}(a \oplus b) = (a+b) - a \oplus b$$

も upl(b) の整数倍になっている。ここで、(1) から、err($a \oplus b$) は $t$bit 浮動小数点数になることがわかる。

(1) の成立することは次の図から説明できるが、これは講義の際に詳しく行う。