数理解析研究所講究録
第673巻 1988年 135-138

浜田方式の数表現に対する誤差解析試論

京都大学数理解析研究所／一松　信　*)

0．浜田方式の数表現をめぐって

浜田方式の数表現は、URR(Universal Real Representation)と呼ばれている。この名は必ずしも適切でなく、むしろ Universal Floating Numberというのがよいらしいが、慣用に従うことにする。

これは表現の相対誤差が一定でないために、精度を気にする人が多い。　以下に述べる内容は実質的には知られている事実にすぎず、また決して完全なものではない。　しかし新しい天才的な着想に対して、無視するかまたは叩いて潰すのが、「独創を尊ばない日本の慣習」であったことを反省しつつ、敢えて試論を提唱する。

1．　URRの表現

以下次の記号を使う。

n　語の長さ；32,64 など。　一応一定とする。

$a_0, a_1, \ldots, a_{n-1}$　nビットの語中のビット；純 2進なので 0か1 。

$s = a_0$　符号ビット；　0が正, 1が負。

$a = a_1 = \ldots = a_m, a_{m+1} = \overline{a}$ （aの否定）　主指数部を表す同じ符号の連。

m　上の連の長さ。

$s' = s \oplus a \oplus 1$　　⊕は2進和

$m' = \max(2m, m+2)$　仮数部が始まるビットの位置。

$q(m) = 2\uparrow(2\uparrow m)$　印刷の都合による略記。　但し $q(-1)=1$ と約束する。

$[a_i a_{i+1} \ldots a_j]2$ ； $[0.a_i a_{i+1} \ldots a_j]2$　それぞれそのビットを並べた列が表す通例の2進整数及び小数。

このときnビットの列 $a_0 a_1 \ldots a_{n-1}$ の表す数 $2\uparrow e \times A$　は次のようになる。

---

*)　1989年 4月以後は東京電機大学理工学部（埼玉県鳩山学舎）。

$A = (-2)\uparrow s + [0.a_{m'} \ldots a_{n-1}]2$　仮数部；絶対値は1と2の間。

但し$m' \geqq n$のときには $(-2)\uparrow s$ (s=0なら1, s=1なら-2) とする。

指数部 $e = [e_\ell \ldots e_0]2$ は、以下の算法により2進補数表示で定まる整数である。

(o) a_1 を読んだ段階ですべての $e_i = s'$ と置く。($s'=0$なら正, $s'=1$なら負)

(i) m=1ならそれで完了。　つまり$s'=0$なら e=0; $s'=1$なら e=-1.

(ii) $a_1 = a_2$ すなわちm≧2なら $e_0 = \overline{s'}$ とする。　a_3以後 $a_i = a$ である間、全体を毎回左に1ビットシフトし、右からsと同じビットを補充する。

(iii) 区切り符号 $a_{m+1} = \overline{a}$ がきたときこの操作を終り、以下のm-2ビットをその順序に右から並べてその左に0を詰めた列と2進和(exclusive Or)を取る。

注：　m=2のときには (iii)の操作は省いてよい。　eをまとめて一つの式で書くことは可能だが、かえってわかりにくい。　なお予稿に記述した式は、s=1のときの操作が正しくなかったことに気付いたので、上のように訂正する。

とにかく主指数部の長さがmのとき、表す数の絶対値がつねに

$$\beta(m-2) \text{ と } \beta(m-1) \quad \text{または} \quad 1/\beta(m-2) \text{ と } 1/\beta(m-1)$$

の間にあることに注意する。

2．数の分類

便宜的に URRで表される数を、次のように分類して扱う。

型Ⅰ：最精数　m=1の数 — m=2の数を含めることもある。　副指数部がなく、最大の精度を持つ数である。

型Ⅱ：標準数　$m < n/2$の数、すなわち仮数部が実在する数。

型Ⅲ：指数部数　$n/2 \leqq m < n$ である数、すなわち主指数部は完全に表示されているが、副指数部は一部しかなく、仮数部がまったくない数。　絶対値は極めて大きいかまたは極めて小さく、在来方式ではとっくに溢れて扱えない数。

型Ⅳ：非数　ここでは　0，±ε，±∞，∞ (s=1; 他のビットはすべて0)をこれに入れる。

この試論で主に扱うのは、普通の計算に現れる型Ⅱの数である。

3. 標準数の精度

$x = a_0 a_1 \ldots a_{n-1}$ が標準数とする。 そのとき仮数部は n-2m ビットの長さを持つ。 したがってxが近似している数の真値 $\tilde{x}$ に対して、相対誤差 δ は

$$\tilde{x} = x(1+\delta), \quad |\delta| \leq 2^{n-2m}$$

を満たす。

いま便宜上、累乗の記号 a↑bを a<0のときにも拡張して

$$a\uparrow b = \mathrm{sgn}(a) \times (|a| \uparrow b)$$

と約束する。 また2を底とする対数を lb で表すと

$$2\uparrow(m-2) \leq |\mathrm{lb}\ x| < 2\uparrow(m-1)$$

なので、 $\mathrm{lb}\ \tilde{x} = \mathrm{lb}\ x + \mathrm{lb}\ (1+\delta) = \mathrm{lb}\ x \times [1 + \mathrm{lb}(1+\delta)/\mathrm{lb}\ x]$

となる。 すなわち $\tilde{x} = x\uparrow(1+\epsilon)$, $|\epsilon| \leq 2\uparrow(n-2m)/2\uparrow(m-2) = 2\uparrow(n-m+2)$ ---- (*)

という評価ができる。

この評価は型Ⅲの数にも成立する。 型Ⅲの数では仮数部が空なので、その相対誤差は100% (!)であり、さらに副指数部の不足分が

$$m-2-(n-m-2) = 2m-n$$

ビットある。 したがって $\tilde{x}/x$ は $2\uparrow(n-2m)$ と $2\uparrow(2m-n+1)$ の間にある。 他方 $|\mathrm{lb}\ x|$ はほぼ $2\uparrow(m-1)$ なので、その比について上と同じ評価が成立する。

4. 実際の評価

前節の評価から、少なくとも乗法・除法についてはWillkinson流の後退誤差解析ができる。

不幸にして前節の評価は、もっとも精度の高い型Ⅰの数については、lb xが0に近いために成立しない。 それをも含めるために、lb xを x=1の近くで修正した関数を使うことができるが、それは評価には有用でない。 したがって実用的には、型Ⅱまでの数を使って少なくとも n/2ビットの精度が保証されれば満足するか、あるいは型Ⅰの数の部分だけを普通に評価して(n-3または n-4ビット)、他の部分と組合せる必要がある。

この評価はURRにおいて、1に近い数を扱うようにスケーリングすれば精度が高いという常識と矛盾するものではない。　逆に指数部数でさえも十分の情報をもっていて、無意味ではないと解釈すべきものであろう。

もっとも実際のシステムでは仮数部がない数が現れたら、念のために警告を出すような注意が必要かもしれない。

5. 指数部の桁落ちについて

URRのように指数部の絶対値が極めて大きい数が扱える体系では、指数部の絶対値が正と負とで大きい2個の数 x,yを掛けたときに「指数部の桁落ち」が生じる。　実際の計算ではこれをそれ程恐れることはない。　そのようにして得られる数自体が答になることは珍らしく、多くはそれが小さいけれども必要な量だからである。　その場合には型IIIの数同志の積でも意味がある。

主指数部の長さmが大きい型IIの2数 x,yの積を考える。　それらの主指数部の長さをm(同じ)とすると、仮数部の長さはともに　n－2m　ビットだから、その積はレジスタ内では2(n－2m)ビットの精度を持つ。　(もとの副指数部)

他方指数部が正と負とで打ち消しあったとしても、少なくとも log(2m) ビットの指数部が残り、その分仮数部の積の下位が保存される。　この効果は精度を保つのに有効である。　(が長くなるため、そ)

∞. 結び

以上は入口にすぎない。　次に必要な作業は一般論を展開することではなく、指数部の絶対値が極めて大きい数の現れる典型的な計算例について、具体的に誤差解析を進めることであろう。　今回はその第一歩を述べた次第である。

[参]　浜田穂積：ＵＲＲ－溢れのない数値表現、早稲田大学学位論文　1986