Alkalmazott matematika

Baran Ágnes

Lebegőpontos számok 2.

Lebegőpontos számok

Példa.

$a = 10$

$$0.3721 = \frac{3}{10} + \frac{7}{10^2} + \frac{2}{10^3} + \frac{1}{10^4}$$

$$21.65 = 0.2165 \cdot 10^2 = \left(\frac{2}{10} + \frac{1}{10^2} + \frac{6}{10^3} + \frac{5}{10^4} \right) \cdot 10^2$$

$a = 2$

$$0.1101 = \frac{1}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{1}{2^4}$$

$$0.001011 = 0.1011 \cdot 2^{-2} = \left(\frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} \right) \cdot 2^{-2}$$

Lebegőpontos számok

A nemnulla lebegőpontos számok alakja:

$$\pm a^k \left(\frac{m_1}{a} + \frac{m_2}{a^2} + \cdots + \frac{m_t}{a^t} \right)$$

ahol

$a > 1$ egész, a számábrázolás alapja

$t > 1$, egész, a mantissza hossza

$k_- \leq k \leq k_+$ egész, a karakterisztika, ahol $k_- < 0$ és $k_+ > 0$ adott

$1 \leq m_1 \leq a - 1$, egész (a szám normalizált)

$0 \leq m_i \leq a - 1$, egész, ha $i = 2, \ldots, t$

röviden: $[\pm|k|m_1,\ldots,m_t]$
ahol $(m_1,\ldots,m_t)$ a mantissza.

Az a, t, k_-, k_+ értéke egyértelműen leírja az ábrázolható számok halmazát.

Példa

$a=2,\ t=4,\ k_-=-3,\ k_+=2$ esetén mi lesz a 0.1875 lebegőpontos alakja?

Megoldás.

$$
\begin{array}{r|l}
0. & 1875 \\
0 & 375 \\
0 & 75 \\
1 & 5 \\
1 & 0
\end{array}
$$

$0.1875_{10} = 0.0011_2$

Normalizálás után: $2^{-2} \cdot 0.1100$

Feladat

Legyen $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$. Írjuk fel a következő számok lebegőpontos alakját.

$$0.6875, \quad 0.8125, \quad 3.25, \quad 0.875, \quad 0.5625, \quad 1.625, \quad 2.75$$

Példa

Hány pozitív normalizált lebegőpontos szám ábrázolható az előző feladatban adott jellemzők mellett?

Megoldás

A karakterisztika összesen 6-féle értéket vehet fel

A mantissza első helyén csak az 1 állhat (normalizálás)

A mantissza maradék helyeit $2 \cdot 2 \cdot 2 = 8$-féleképpen tölthetjük ki.

Összesen $6 \cdot 8 = 48$ lehetőség.

Adott a, t, k_-, k_+ esetén
a legnagyobb ábrázolható szám:

$$M_\infty = a^{k_+}\left(\frac{a-1}{a} + \frac{a-1}{a^2} + \cdots + \frac{a-1}{a^t}\right)$$

$$= a^{k_+}\left(1 - \frac{1}{a} + \frac{1}{a} - \frac{1}{a^2} + \cdots + \frac{1}{a^{t-1}} - \frac{1}{a^t}\right)$$
$$= a^{k_+}\left(1 - a^{-t}\right)$$

a legkisebb pozitív normalizált ábrázolható szám:

$$\varepsilon_0 = a^{k_-}\left(\frac{1}{a} + 0 + \cdots + 0\right) = a^{k_- -1}$$

Szubnormális számok: ha $k = k_-$, akkor $m_1 = 0$ is lehet.

Az 1 mindig lebegőpontos szám:

$$1 = a^1 \cdot \frac{1}{a}, \qquad \text{vagy röviden: } 1 = [+|1|1, 0, \dots, 0]$$

Az 1 jobboldali szomszédja:

$$1 + \varepsilon_1 = [+|1|1, 0, \dots, 0, 1]$$

másképp:

$$1 + \varepsilon_1 = a\left(\frac{1}{a} + 0 + \cdots + 0 + \frac{1}{a^t}\right) = 1 + a^{1-t}$$

azaz $\varepsilon_1 = a^{1-t}$ (**gépi epszilon**)

5. feladat

(a) Írjon egy kódot a gépi epszilon meghatározására.

(b) Olvassa el az Octave/Matlab eps függvényének help-jét. Nézze meg az eps (azaz az eps(1)) értékét.

6. feladat

(a) Írjon egy kódot az ε_0 meghatározására.

(b) Nézze meg az `eps(0)` értékét!

7. feladat

Írassa ki gépén a `realmin` és `realmax` értékét. Vizsgálja meg a `realmin('single')` és `realmax('single')` értékeket is.

Az IEEE lebegőpontos aritmetikai szabvány:

	egyszeres pontosság	dupla pontosság
méret	32 bit	64 bit
mantissza	23+1 bit	52+1 bit
karakterisztika	8 bit	11 bit
ε_1	$\approx 1.19 \cdot 10^{-7}$	$\approx 2.22 \cdot 10^{-16}$
M_∞	$\approx 10^{38}$	$\approx 10^{308}$

mivel m_1 mindig 1, ezért nem ábrázoljuk
az előjel ábrázolására 1 bit

Adott a, t, k_+, k_- mellett az ábrázolható lebegőpontos számok a $[-M_\infty, M_\infty]$ intervallum egy megszámlálható részhalmazát alkotják.

8. feladat

(a) Ábrázoljuk számegyenesen az $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$ jellemzők mellett felírható összes pozitív normalizált lebegőpontos számot.

(b) A fenti számábrázolási jellemzők mellett mennyi lesz M_∞, ε_0 és ε_1 értéke?

(c) Mit mondhatunk két szomszédos szám távolságáról?

(d) Mit mondhatunk a szomszédos számok távolságáról, ha k_+ értékét 4-re módosítjuk?

(e) Mi lenne, ha $k_+ > 4$ teljesülne?

Példa.

A pozitív normalizált lebegőpontos számok $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$ esetén.

	$k = 0$	$k = 1$	$k = 2$	$k = -1$	$k = -2$	$k = -3$
0.1000	$\frac{8}{16}$	$\frac{8}{8}$	$\frac{8}{4}$	$\frac{8}{32}$	$\frac{8}{64}$	$\frac{8}{128}$
0.1001	$\frac{9}{16}$	$\frac{9}{8}$	$\frac{9}{4}$	$\frac{9}{32}$	$\frac{9}{64}$	$\frac{9}{128}$
0.1010	$\frac{10}{16}$	$\frac{10}{8}$	$\frac{10}{4}$	$\frac{10}{32}$	$\frac{10}{64}$	$\frac{10}{128}$
0.1011	$\frac{11}{16}$	$\frac{11}{8}$	$\frac{11}{4}$	$\frac{11}{32}$	$\frac{11}{64}$	$\frac{11}{128}$
0.1100	$\frac{12}{16}$	$\frac{12}{8}$	$\frac{12}{4}$	$\frac{12}{32}$	$\frac{12}{64}$	$\frac{12}{128}$
0.1101	$\frac{13}{16}$	$\frac{13}{8}$	$\frac{13}{4}$	$\frac{13}{32}$	$\frac{13}{64}$	$\frac{13}{128}$
0.1110	$\frac{14}{16}$	$\frac{14}{8}$	$\frac{14}{4}$	$\frac{14}{32}$	$\frac{14}{64}$	$\frac{14}{128}$
0.1111	$\frac{15}{16}$	$\frac{15}{8}$	$\frac{15}{4}$	$\frac{15}{32}$	$\frac{15}{64}$	$\frac{15}{128}$

$M_\infty = 2^2(1 - 2^{-4}) = \frac{15}{4}$ és $\varepsilon_0 = 2^{-3-1} = \frac{1}{16}$ $\left(= \frac{8}{128}\right)$

Legyen $y = a^k \cdot 0.m_1 m_2 \ldots m_t$.

A legközelebbi nála nagyobb szám

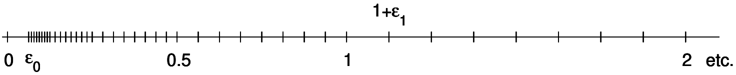
$$a^k \cdot \frac{1}{a^t} = a^{k-t}$$

távolságra van tőle.

Nagyobb karakterisztika $\rightarrow$ nagyobb lépésköz.

Ha $k > t$, akkor a lépésköz nagyobb mint 1.

$a = 2,\ t = 4,\ k_- = -3$ esetén



$\varepsilon_0 = a^{k_- - 1} = 2^{-4} = \frac{1}{16}$,

$\varepsilon_1 = a^{1-t} = 2^{-3} = \frac{1}{8}$

9. feladat

Vizsgálja meg számítógépén a $2^{66} + 1 == 2^{66}$, $2^{66} + 10 == 2^{66}$, $2^{66} + 100 == 2^{66}$, $2^{66} + 1000 == 2^{66}$ és $2^{66} + 10000 == 2^{66}$ logikai kifejezések értékét! Keresse meg azt a legkisebb $n > 0$ számot, melyre a $2^{66} + n == 2^{66}$ logikai kifejezés értéke hamis. Mennyi az `eps(2^66)` értéke?

Dupla pontosság esetén ($t = 53$):

y	a jobboldali szomszéd távolsága
1	$\approx 2.22 \cdot 10^{-16}$
16	$\approx 3.5527 \cdot 10^{-15}$
1024	$\approx 2.27 \cdot 10^{-13}$
$2^{20} \approx 10^6$	$\approx 2.33 \cdot 10^{-10}$
$2^{52} \approx 4.5 \cdot 10^{15}$	1
$2^{60} \approx 1.15 \cdot 10^{18}$	256
$2^{66} \approx 7.38 \cdot 10^{19}$	16384

Kerekítés

A $[-M_\infty, M_\infty]$ intervallumból nem minden szám írható fel lebegőpontos alakban.

Példa

A 0.1 kettes számrendszerbeli alakja:

$$0.0001100110011001100....$$

Az $\frac{1}{3}$ kettes számrendszerbeli alakja:

$$0.0101010101010....$$

Kerekítés

Legyen $x \in [-M_\infty, M_\infty]$ egy valós szám, $fl(x)$ pedig a hozzárendelt lebegőpontos szám.

Szabályos kerekítés esetén:

$$fl(x) = \begin{cases} 0, & \text{ha } |x| < \varepsilon_0 \\ \text{az } x\text{-hez legközelebbi lebegőpontos számok} & \text{ha } |x| \geq \varepsilon_0 \\ \text{közül a nagyobb abszolút értékű,} & \end{cases}$$

Levágás esetén:

$$fl(x) = \begin{cases} 0, & \text{ha } |x| < \varepsilon_0 \\ \text{az } x\text{-hez legközelebbi lebegőpontos szám a 0 felé, ha } |x| \geq \varepsilon_0 & \end{cases}$$

Megjegyzés

Ha az ábrázolni kívánt szám két szomszédos lebegőpontos szám között félúton helyezkedik el, akkor a valóságban az előzőnél bonyolultabb kerekítési szabály alapján történik a kerekítés.

Példa

Legyen $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$. Mi lesz a 0.1-hez rendelt lebegőpontos szám szabályos kerekítés, illetve levágás esetén?

A 0.1 kettes számrendszerben, normalizálva:

$$2^{-3} \cdot 0.1100\textcolor{red}{110011001100\ldots.}$$

Szabályos kerekítés:

$$fl(0.1) = 2^{-3} \cdot 0.1101$$

Levágás:

$$fl(0.1) = 2^{-3} \cdot 0.1100$$

10. feladat.

Legyen $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$. Mi lesz az alábbi számokhoz rendelt lebegőpontos szám szabályos kerekítés, illetve levágás esetén?

$$0.4, \quad 0.3, \quad \frac{1}{3}, \quad 0.7, \quad \frac{1}{32}$$

11. feladat

Vizsgálja meg számítógépén a $0.4 - 0.5 + 0.1 == 0$ logikai kifejezés értékét! Magyarázza meg a tapasztalt jelenséget! Mi lesz a $0.1 - 0.5 + 0.4 == 0$ logikai kifejezés értéke? Vizsgálja meg a $0.4 - 0.5 + 0.1$ és $0.1 - 0.5 + 0.4$ kifejezések értékét is!

12. feladat

Az alábbi algoritmus végrehajtása után mennyi az x elméleti, illetve a gépi számítás után adódó értéke?

```
x=1/3;
for i=1:40
   x=4*x-1;
end
```

Magyarázza meg a tapasztalt jelenséget! (Duplapontosságú számábrázolás ($t = 53$) esetén mennyi az $\frac{1}{3}$ ábrázolásakor bekövetkező hiba? Miért nő ez a ciklus lefuttatása során olyan nagyra?)

13. feladat

Az alábbi algoritmus elméletileg minden $x \geq 0$ esetén az x eredeti értékét adja vissza. Vizsgálja meg mi történik a gyakorlatban, ha az algoritmust $x = 1000$, $x = 100$ kezdőértékkel futtatja! Mi az oka a tapasztalt jelenségnek?

```
for i=1:60
   x=sqrt(x);
end
for i=1:60
   x=x^2;
end
```

Kerekítés

Az **abszolút hiba** becslése

szabályos kerekítésnél:

$$|fl(x) - x| \leq \begin{cases} \varepsilon_0, & \text{ha } |x| < \varepsilon_0 \\ \frac{1}{2}\varepsilon_1 |x|, & \text{ha } |x| \geq \varepsilon_0 \end{cases}$$

levágásnál:

$$|fl(x) - x| \leq \begin{cases} \varepsilon_0, & \text{ha } |x| < \varepsilon_0 \\ \varepsilon_1 |x|, & \text{ha } |x| \geq \varepsilon_0 \end{cases}$$

Kerekítés

A **relatív hiba** becslése, ha $|x| \geq \varepsilon_0$

szabályos kerekítésnél:

$$\frac{|fl(x) - x|}{|x|} \leq \frac{1}{2}\varepsilon_1$$

levágásnál:

$$\frac{|fl(x) - x|}{|x|} \leq \varepsilon_1$$

Gépi epszilon (ε_1)

Adott számábrázolási jellemzők mellett az 1 és a jobboldali lebegőpontos szomszédjának a távolsága.

Alapműveleteknél:

Jelölje $\triangle$ a négy alapművelet valamelyikét, legyen x és y lebegőpontos szám. Tfh a gép a műveletet pontosan végrehajtja és az eredményhez hozzárendel egy lebegőpontos számot. Ekkor

szabályos kerekítés esetén:

$$|fl(x\triangle y) - x\triangle y| \leq \begin{cases} \varepsilon_0, & \text{ha } |x\triangle y| < \varepsilon_0 \\ \frac{1}{2}\varepsilon_1 |x\triangle y|, & \text{ha } |x\triangle y| \geq \varepsilon_0 \end{cases}$$

levágás esetén:

$$|fl(x\triangle y) - x\triangle y| \leq \begin{cases} \varepsilon_0, & \text{ha } |x\triangle y| < \varepsilon_0 \\ \varepsilon_1 |x\triangle y|, & \text{ha } |x\triangle y| \geq \varepsilon_0 \end{cases}$$

Összefoglalva:

ha $|x \triangle y| > M_\infty$, akkor **túlcsordulás**,

ha $|x \triangle y| < \varepsilon_0$, akkor **alulcsordulás** ($fl(x \triangle y) = 0$)

ha $\varepsilon_0 \leq |x \triangle y| \leq M_\infty$, akkor az előző reláció átírható:

$$fl(x \triangle y) = (x \triangle y) \cdot (1 + \varepsilon_\triangle), \quad \text{ahol } |\varepsilon_\triangle| \leq \varepsilon_1 \begin{cases} 1, & \text{levágás} \\ \frac{1}{2}, & \text{szabályos kerekítés} \end{cases}$$