Ponto Flutuante e Erros

1.2.2 Aritmética de Ponto Flutuante

IEEE – Institute of Electrical and Electronic Engineers

Representação em ponto flutuante normalizada:
r = sinal x mantissa x base^expoente = +/- 0.ddddd...dd x \beta ^ e

Forma normalizada: um zero antes da vírgula, primeiro dígito depois da vírgula não-nulo.
$0.35 = 0.35x10^0$
$5.47 = 0.547x10^1$
$-123.456 = - 12.3456x10^1 = -1.23456x10^2 = -0.123456x10^3$ (normalizada)

Normalizar
Ex: $2013.54 = 0.201354x10^4$
Ex: $0.000502 = 0.502x10^{-3}$

Sistema de ponto flutuante:
f(beta, t, m, M) , base beta, t dígitos significativos, -m <= e <= M

Considere uma máquina com o sistema f(10,3,5,5)

→Qual é o menor número positivo?

$0.100x10^{-5} = 10^{-6}$ → Representar um número menor que esse: *underflow*

→Qual é o maior número positivo?
$0.999x10^5$ =99900 → Representar um número maior que esse: *overflow*

1.3.1 Erros Absolutos e Relativos (pg 12)

Erro absoluto: diferença entre o valor exato x e o valor aproximado x_
EA = x-x_

Cota superior para erro
\pi \in (3.14,3.15)
|EA| = |pi - pi_| < 0.01

Erros absolutos são sempre iguais? Como comparar essas cotas para erros absolutos?

Seja x_ = 2112.9 tal que |EA|<0.1. Então x_ \in (2112.8,2113).

Seja y_ = 5.3 tal que |EA|<0.1. Então y_ \in (5.2,5.4).

ER = EA/x_ = (x-x_)/x_

|ERx| = 0.1/2112.9 = 0,000047328 = 0.47 x 10^{-4}
|ERy| = 0.1/5.3 = 0,018867925 = 0.19x10^{1}

Portanto x é representado com maior precisão que y.

1.3.2 Erros de arredondamento e truncamento em um sistema de aritmética de ponto flutuante (pg 14)

E quanto o número não pode ser representado na máquina, por falta de precisão?

f(10,4,5,5) e x = 234.57
Normalizando, temos x=0.23457x10^{3} (5 dígitos na mantissa!)

Truncamento: x_=0.2345x10^{3}
Arredondamento x_=0.2346x10^{3}

x = 0.2345x10^{3} + 0.00007x10^{3} = 0.2345x10^{3} + 0.7x10^{-1}

Removidos

Removido 1.2.2

→Quantos números existem numa máquina com o sistema f(2,3,2,2)?
+/- 0.ddd x beta^e = +- 0.ddd x 2^e, onde d1 != 0

→2 possibilidades para o sinal
→1 possibilidade para d1
→2 para d2
→2 para d3
→4 para beta^e

Multiplicando temos 2x1x2x2x4 = 32. Acrescentando o zero, temos 33 números possíveis.

Exemplo: f(10,3,4,4)

x	x_ arredondamento	x_ trucamento
1.25	0.125×10^1	0.125×10^1
10.053	0.101×10^2	0.100×10^2
-238.15	-0.238×10^3	-0.238×10^3
2.71828…	0.272×10^1	0.271×10^1
0.000007	0.700×10^{-5} (underflow)	0.700×10^{-5} (underflow)
718235.82	0.718×10^6 (overflow)	0.718×10^6 (overflow)

Removido 1.3.3 Análise de erros nas operações aritméticas de ponto flutuante (pg 16)

Erro total da operação: erro das parcelas/fatores + erro do resultado

f(10,2,m,M)
$x=0.25 \times 10^3$ $y=0.3 \times 10^1$

Alinhando os pontos decimais:
$x=0.25 \times 10^3$ $y=0.003 \times 10^3$

Resultado exato
$x + y = 0.253 \times 10^3$

Resultado aproximado
$\bar{x} + \bar{y} = 0.25 \times 10^3$
$|EA| = |0.253 - 0.25| \times 10^3 = 0.003 \times 10^3$

Ex: f(10,4,m,M)
$x=0.937x10^4$ $y=0.1272x10^2$
Obter x+y

Alinhando os pontos decimais
$x=0.937x10^4$ $y=0.001272x10^4$

Resultado exato:
$x+y = (0.937 + 0.001272)x10^4 = 0.938272x10^4$

$x\_ + y\_ = 0.9382x10^4$ no truncamento $x\_ + y\_ = 0.9383x10^4$ no arredondamento

Ex: mesmos números x e y, obter xy
Resultado exato:
$xy = (0.937x10^4)(0.1272x10^2) = (0.937 \times 0.1272)x10^6 = 0.1191864 \times 10^6$
$xy\_ = 0.1191x10^6$ no truncamento e $xy\_ = 0.1192x10^6$ no arredondamento