Aritmética de Ponto Flutuante

Prof. Americo Cunha

Universidade do Estado do Rio de Janeiro – UERJ

americo.cunha@uerj.br

www.americocunha.org

Um experimento computacional

$$a = 4/3,\ \ b = a - 1,\ \ c = 3b,\ \ e = 1 - c$$

Em aritmética exata

$$e = 1 - c = 1 - 3b = 1 - 3(a - 1) = 1 - 3(4/3 - 1) = 0$$

Agora vamos fazer essa conta no GNU Octave:

```
>>  format long
>>  a = 4/3
>>  b = a - 1
>>  c = 3*b
>>  e = 1 - c
```

As operações em ponto flutuante

Sejam $\mathtt{fl}(x)$ e $\mathtt{fl}(y)$ as representações dos reais x e y em $F(\beta, t, m, M)$, respectivamente. As *operações aritméticas em ponto flutuante*, são definidas por:

$$x \oplus y = \mathtt{fl}\left(\mathtt{fl}(x) + \mathtt{fl}(y)\right)$$

$$x \ominus y = \mathtt{fl}\left(\mathtt{fl}(x) - \mathtt{fl}(y)\right)$$

$$x \otimes y = \mathtt{fl}\left(\mathtt{fl}(x) \times \mathtt{fl}(y)\right)$$

$$x \oslash y = \mathtt{fl}\left(\mathtt{fl}(x) / \mathtt{fl}(y)\right)$$

Ao final de cada operação *arredonda-se (ou trunca-se)*.

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \quad x = 9370 \quad y = 12,72$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Soma em ponto flutuante:

$x \oplus y$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \quad x = 9370 \quad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \quad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Soma em ponto flutuante:

$$x \oplus y \;=\; \texttt{fl}\left(\texttt{fl}\,(x) + \texttt{fl}\,(y)\right)$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \quad x = 9370 \quad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \quad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Soma em ponto flutuante:

$$\begin{aligned} x \oplus y &= \texttt{fl}\left(\texttt{fl}\,(x) + \texttt{fl}\,(y)\right) \\ &= \texttt{fl}\left(0,9370 \times 10^4 + 0,1272 \times 10^2\right) \end{aligned}$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Soma em ponto flutuante:

$$\begin{aligned} x \oplus y &= \texttt{fl}\left(\texttt{fl}\,(x) + \texttt{fl}\,(y)\right) \\ &= \texttt{fl}\left(0,9370 \times 10^4 + 0,1272 \times 10^2\right) \\ &= \texttt{fl}\left((0,9370 + 0,001272) \times 10^4\right) \end{aligned}$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Soma em ponto flutuante:

$$\begin{aligned} x \oplus y &= \texttt{fl}\left(\texttt{fl}\,(x) + \texttt{fl}\,(y)\right) \\ &= \texttt{fl}\left(0,9370 \times 10^4 + 0,1272 \times 10^2\right) \\ &= \texttt{fl}\left((0,9370 + 0,001272) \times 10^4\right) \\ &= \texttt{fl}\left(0,938272 \times 10^4\right) \end{aligned}$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \quad x = 9370 \quad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}(x) = 0,9370 \times 10^4 \quad \texttt{fl}(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Soma em ponto flutuante:

$$\begin{aligned} x \oplus y &= \texttt{fl}\left(\texttt{fl}(x) + \texttt{fl}(y)\right) \\ &= \texttt{fl}\left(0,9370 \times 10^4 + 0,1272 \times 10^2\right) \\ &= \texttt{fl}\left((0,9370 + 0,001272) \times 10^4\right) \\ &= \texttt{fl}\left(0,938272 \times 10^4\right) \\ &= 0,9382 \times 10^4 \text{ (truncamento)} \end{aligned}$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \quad x = 9370 \quad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}(x) = 0,9370 \times 10^4 \quad \texttt{fl}(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Soma em ponto flutuante:

$$\begin{aligned}
x \oplus y &= \texttt{fl}\left(\texttt{fl}(x) + \texttt{fl}(y)\right) \\
&= \texttt{fl}\left(0,9370 \times 10^4 + 0,1272 \times 10^2\right) \\
&= \texttt{fl}\left((0,9370 + 0,001272) \times 10^4\right) \\
&= \texttt{fl}\left(0,938272 \times 10^4\right) \\
&= 0,9382 \times 10^4 \text{ (truncamento)} \\
&= 0,9383 \times 10^4 \text{ (arredondamento)}
\end{aligned}$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Produto em ponto flutuante:

$x \otimes y$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \quad x = 9370 \quad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \quad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Produto em ponto flutuante:

$$x \otimes y \;=\; \texttt{fl}\left(\texttt{fl}\,(x) \times \texttt{fl}\,(y)\right)$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Produto em ponto flutuante:

$$\begin{aligned} x \otimes y &= \texttt{fl}\left(\texttt{fl}\,(x) \times \texttt{fl}\,(y)\right) \\ &= \texttt{fl}\left(0,9370 \times 10^4 \times 0,1272 \times 10^2\right) \end{aligned}$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \qquad x = 9370 \qquad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}\,(x) = 0,9370 \times 10^4 \qquad \texttt{fl}\,(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Produto em ponto flutuante:

$$\begin{aligned} x \otimes y &= \texttt{fl}\left(\texttt{fl}\,(x) \times \texttt{fl}\,(y)\right) \\ &= \texttt{fl}\left(0,9370 \times 10^4 \times 0,1272 \times 10^2\right) \\ &= \texttt{fl}\left(0,1191864 \times 10^6\right) \end{aligned}$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \quad x = 9370 \quad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}(x) = 0,9370 \times 10^4 \quad \texttt{fl}(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Produto em ponto flutuante:

$$\begin{aligned} x \otimes y &= \texttt{fl}\left(\texttt{fl}(x) \times \texttt{fl}(y)\right) \\ &= \texttt{fl}\left(0,9370 \times 10^4 \times 0,1272 \times 10^2\right) \\ &= \texttt{fl}\left(0,1191864 \times 10^6\right) \\ &= 0,1191 \times 10^6 \text{ (truncamento)} \end{aligned}$$

Um exemplo simplista em ponto flutuante

Um sistema de ponto flutante e dois reais dados:

$$F(10, 4, -1022, 1023) \quad x = 9370 \quad y = 12,72$$

Representação em ponto flutuante:

$$\texttt{fl}(x) = 0,9370 \times 10^4 \quad \texttt{fl}(y) = 0,1272 \times 10^2$$

$\longrightarrow$ A representação é exata!

Produto em ponto flutuante:

$$\begin{aligned} x \otimes y &= \texttt{fl}\left(\texttt{fl}(x) \times \texttt{fl}(y)\right) \\ &= \texttt{fl}\left(0,9370 \times 10^4 \times 0,1272 \times 10^2\right) \\ &= \texttt{fl}\left(0,1191864 \times 10^6\right) \\ &= 0,1191 \times 10^6 \text{ (truncamento)} \\ &= 0,1192 \times 10^6 \text{ (arredondamento)} \end{aligned}$$

Qual o resultado desses cálculos?

Qual o resultado desses cálculos?

Aritmética exata:

$$x + y = 9382,72 \qquad x \times y = 119186,4$$

Qual o resultado desses cálculos?

Aritmética exata:

$$x + y = 9382,72 \qquad x \times y = 119186,4$$

Truncamento:

$$x \oplus y = 0,9382 \times 10^4 \qquad x \otimes y = 0,1191 \times 10^6$$

Qual o resultado desses cálculos?

Aritmética exata:

$$x + y = 9382,72 \qquad x \times y = 119186,4$$

Truncamento:

$$x \oplus y = 0,9382 \times 10^4 \qquad x \otimes y = 0,1191 \times 10^6$$

$$\text{“}x + y = 9382\text{”} \neq 9382,72 \qquad \text{“}x \times y = 119100\text{”} \neq 119186,4$$

Qual o resultado desses cálculos?

Aritmética exata:

$$x + y = 9382,72 \qquad x \times y = 119186,4$$

Truncamento:

$$x \oplus y = 0,9382 \times 10^4 \qquad x \otimes y = 0,1191 \times 10^6$$

$$\text{“}x + y = 9382\text{”} \neq 9382,72 \qquad \text{“}x \times y = 119100\text{”} \neq 119186,4$$

Arredondamento:

$$x \oplus y = 0,9383 \times 10^4 \qquad x \otimes y = 0,1192 \times 10^6$$

Qual o resultado desses cálculos?

Aritmética exata:

$$x + y = 9382,72 \qquad x \times y = 119186,4$$

Truncamento:

$$x \oplus y = 0,9382 \times 10^4 \qquad x \otimes y = 0,1191 \times 10^6$$

$$\text{“}x + y = 9382\text{”} \neq 9382,72 \qquad \text{“}x \times y = 119100\text{”} \neq 119186,4$$

Arredondamento:

$$x \oplus y = 0,9383 \times 10^4 \qquad x \otimes y = 0,1192 \times 10^6$$

$$\text{“}x + y = 9383\text{”} \neq 9382,72 \qquad \text{“}x \times y = 119200\text{”} \neq 119186,4$$

Qual o resultado desses cálculos?

Aritmética exata:

$$x + y = 9382,72 \qquad x \times y = 119186,4$$

Truncamento:

$$x \oplus y = 0,9382 \times 10^4 \qquad x \otimes y = 0,1191 \times 10^6$$

$$\text{“}x + y = 9382\text{”} \neq 9382,72 \qquad \text{“}x \times y = 119100\text{”} \neq 119186,4$$

Arredondamento:

$$x \oplus y = 0,9383 \times 10^4 \qquad x \otimes y = 0,1192 \times 10^6$$

$$\text{“}x + y = 9383\text{”} \neq 9382,72 \qquad \text{“}x \times y = 119200\text{”} \neq 119186,4$$

Em geral, operações em ponto flutuante não são exatas.

Qual o resultado desses cálculos?

Aritmética exata:

$$x + y = 9382,72 \qquad x \times y = 119186,4$$

Truncamento:

$$x \oplus y = 0,9382 \times 10^4 \qquad x \otimes y = 0,1191 \times 10^6$$

$$\text{“}x + y = 9382\text{”} \neq 9382,72 \qquad \text{“}x \times y = 119100\text{”} \neq 119186,4$$

Arredondamento:

$$x \oplus y = 0,9383 \times 10^4 \qquad x \otimes y = 0,1192 \times 10^6$$

$$\text{“}x + y = 9383\text{”} \neq 9382,72 \qquad \text{“}x \times y = 119200\text{”} \neq 119186,4$$

Em geral, operações em ponto flutuante não são exatas.

Até quando x e y têm representação exata!

Erros nas operações em ponto flutuante

Se $\mathtt{fl}(x), \mathtt{fl}(y) \in F(\beta, t, m, M)$ então

$$x \oplus y = \left(\mathtt{fl}(x) + \mathtt{fl}(y)\right)(1 + \varepsilon_1)$$

$$x \ominus y = \left(\mathtt{fl}(x) - \mathtt{fl}(y)\right)(1 + \varepsilon_2)$$

$$x \otimes y = \left(\mathtt{fl}(x) \times \mathtt{fl}(y)\right)(1 + \varepsilon_3)$$

$$x \oslash y = \left(\mathtt{fl}(x) \,/\, \mathtt{fl}(y)\right)(1 + \varepsilon_1)$$

onde $|\varepsilon_j| \leq \epsilon_M$, sendo ϵ_M é a precisão da máquina.

Essas operações:

- tem erro relativo menor que ϵ_M;
- não são associativas e nem distributivas.

Experimento computacional 1

$$10^{-17}\,x^2 + x - 1 = 0$$

$$x_1 \approx -1.000000000000000 \times 10^{17}$$

$$x_2 \approx \quad 1.000000000000000$$

(16 dígitos de precisão)

```
1  function [x1,x2] = Eq2Classic(a,b,c)
2    x1 = (-b - sqrt(b^2 - 4*a*c))/(2*a);
3    x2 = (-b + sqrt(b^2 - 4*a*c))/(2*a);
4  end
```

```
1  format long
2  a = 1.0e-17; b = 1.0; c = -1.0;
3  [x1,x2] = Eq2Classic(a,b,c)
4  a*x1*x1 + b*x1 + c
5  a*x2*x2 + b*x2 + c
```

Experimento computacional 2

$$10^{-17}\,x^2 + x - 1 = 0$$

$$x_1 \approx -1.000000000000000 \times 10^{17}$$
$$x_2 \approx \quad 1.000000000000000$$

(16 dígitos de precisão)

```
function [x1,x2] = Eq2Citardauq(a,b,c)
  if b >= 0.0
    x1 = (-b - sqrt(b^2 - 4*a*c))/(2*a);
    x2 = (2*c)/(-b - sqrt(b^2 - 4*a*c));
  else
    x1 = (2*c)/(-b + sqrt(b^2 - 4*a*c));
    x2 = (-b + sqrt(b^2 - 4*a*c))/(2*a);
  end
end
```

```
format long
a = 1.0e-17; b = 1.0; c = -1.0;
[x1,x2] = Eq2Citardauq(a,b,c)
a*x1*x1 + b*x1 + c
a*x2*x2 + b*x2 + c
```

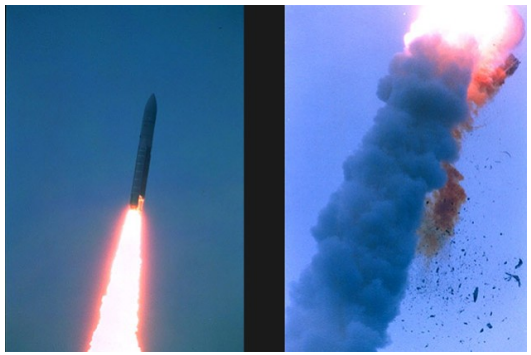
Desastres envolvendo ponto flutuante

- Bolsa de Vancouver
  Jan. 1982: ↑ 1000,000 pontos
  Nov. 1983: ↓ 524,811 pontos
  Corrigido: ↑ 1098,892 pontos

- Míssil Patriota (25/2/1991)
  28 vítimas fatais e +100 feridos

- Foguete Ariane 5 (4/6/1996)
  U$ 500M em prejuízo

Patriota ®



Ariane 5*

Detalhes em:

http://shodor.org/succeed-1.0/programs/sspa2002/pages/math_disasters.html

* Figura obtida em https://5g.security/timeline/ariane-5-rocket-explosion/

Para pensar em casa ...

Exercício teórico:

Considere o sistema de ponto flutuante $F(\beta, t, m, M) = F(10, 3, -3, 4)$, que trunca a parcela que não pode ser incorporada à mantissa. Nesse sistema de ponto flutuante, qual o resultado de $10^3 + 1 - 10^3$?

Exercício teórico:

Sabendo que o sistema de ponto flutuante IEEE 754 de precisão dupla é da forma $F(\beta, t, m, M) = F(2, 53, -1022, 1023)$, explique porque a maioria dos dispositivos eletrônicos calcula $(1 + 2^{53}) - 2^{53} = 0$.

Como citar esse material?

A. Cunha, *Aritmética de Ponto Flutuante*,
Universidade do Estado do Rio de Janeiro – UERJ, 2020.

@AmericoCunhaJr

@AmericoCunhaJr @AmericoCunhaJr