MAC0210: Relatório EP 1

18 de Setembro de 2016

Nathan Benedetto Proença - 8941276
Victor Sena Molero - 8941317

Sumário

Parte 1: Aritmética de Ponto Flutuante

Questão 1 (3.11)

Suponha que temos um sistema de representação de ponto flutuante com base 2 e,

$$x = \pm S \times 2^E,$$

$$\text{com } S = (0.1b_2b_3b_4 \ldots b_{24}),$$

$$\text{i.e, } \frac{1}{2} \leq S < 1$$

onde o expoente $-128 < E < 127$.

a) Qual é o maior número de ponto flutuante desse sistema?

Resposta. $2^{126} - 2^{101}$, basta preencher todos os bits (de b_2 até b_{24}) e escolher o maior expoente possível. □

b) Qual é o menor número de ponto flutuante positivo desse sistema?

Resposta. 2^{-128}, basta escolher a menor mantissa possível (0.1) e o menor expoente possível (-127). □

c) Qual é o menor inteiro positivo que não é exatamente representável nesse sistema?

Resposta. $2^{24} + 1$, basta escolher a menor mantissa não representável $(0.10 \ldots 01)$ e o menor expoente para o qual ela representa um inteiro (25). □

Questão 2 (5.1)

Qual é a representação do número 1/10 no formato IEEE single para cada um dos quatro modos de arredondamento?

Resposta. A resposta curta é:

$$x = 0.0\overline{0011},$$

$$\text{round\_down}(x) = \text{round\_towards\_zero}(x) = 1.10011001100110011001100 \times 2^{-4} \text{ e}$$

$$\text{round\_up}(x) = \text{round\_to\_nearest}(x) = 1.10011001100110011001101 \times 2^{-4}.$$

Se x é uma representação binária exata de 1/10, $x = 0.0\overline{0011} = 1.\overline{1001} \times 2^{-4}$ (os números abaixo da barra representam uma dízima periódica, se repetem infinitamente).
Calculamos então $x_- = 1.10011001100110011001100 \times 2^{-4}$ e $x_+ = 1.10011001100110011001101 \times 2^{-4}$. Se o modo é *Round Down*, o número será representado por x_- e se for *Round Up*, será x_+, ambos pela definição dos modos.
Já que $x > 0$, o modo *Round towards zero* também usará x_-, porém x é mais próximo de x_+ do que de x_-, basta perceber que o erro relativo entre x e x_- é maior que 1/2, portanto, o modo *Round to nearest* usará a representação x_+. □

E para os números $1 + 2^{-25}$

Resposta.

$$x = 1.0000000000000000000000001 \times 2^0,$$
$$\text{round\_down}(x) = \text{round\_towards\_zero}(x) = \text{round\_to\_nearest}(x) = 1.00000000000000000000000 \times 2^0,$$
$$\text{round\_up}(x) = 1.00000000000000000000001 \times 2^0.$$

Se x é uma representação binária exata de $1 + 2^{-25}$, $x = 1.0000000000000000000000001 \times 2^0$, $x_- = 1.00000000000000000000000 \times 2^0$ e $x_+ = 1.00000000000000000000001 \times 2^0$. O modo *Round down* vai levar para x_- e o modo *Round up* vai levar para x_+, como sempre.
Já que $x > 0$, o modo *Round towards zero* também usará x_-, além disso, o erro relativo entre x e x_- é $1/4$, logo, o modo *Round to nearest* também levará para x_-. □

e 2^{130}?

Resposta.

$$x = 1 \times 2^{130},$$
$$\text{round\_down}(x) = \text{round\_towards\_zero}(x) = \text{round\_to\_nearest}(x) =$$
$$= \text{round\_up}(x) = 1.00000000000000000000000 \times 2^{130}.$$

Se x é uma representação binária exata de 2^{130}, $x = 1 \times 2^{130}$, porém, x é exatamente representável no sistema IEEE single, logo, em todos os modos de arredondamento, será representado como $1.00000000000000000000000 \times 2^{130}$. □

Questão 3 (6.4)

Qual é o maior número de ponto flutuante x tal que $1 \oplus x$ é exatamente 1, assumindo que o formato usado é IEEE single e modo de arredondamento para o mais próximo?

Resposta. $x = 2^{-24}$. $1 + x$ é igualmente próximo de $1 + 2^{-23}$ e 1, porém, por causa do critério de arredondamento em empate (0 menos significativo), ele é arredondado para 1, qualquer x maior do que esse causará um arredondamento para um número maior do que 1. □

E se o formato for IEEE double?

Resposta. $x = 2^{-53}$, seguindo a mesma lógica usada para concluir a resposta do item anterior. □

Questão 4 (6.8)

Em aritmética exata, a soma é um operador comutativo e associativo. O operador de soma de ponto flutuante é comutativo?

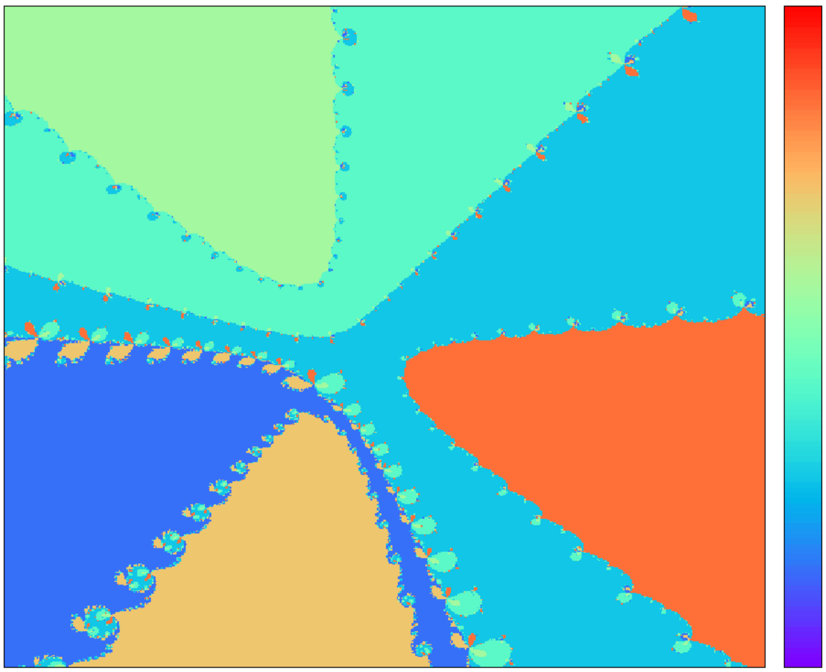
Resposta. Sim, pois para calcular o resultado em soma de ponto flutuante o padrão exige que seja calculado o valor exato e, então, arredondado para o sistema escolhido. Formalmente, denotaremos por $fl(x)$ a representação em ponto flutuante de um real x e por $\oplus$ a operação de soma em ponto flutuante. Temos $fl(x) \oplus fl(y) = fl(fl(x) + fl(y)) = fl(fl(y) + fl(x)) = fl(y) \oplus fl(x)$. □

E associativo?

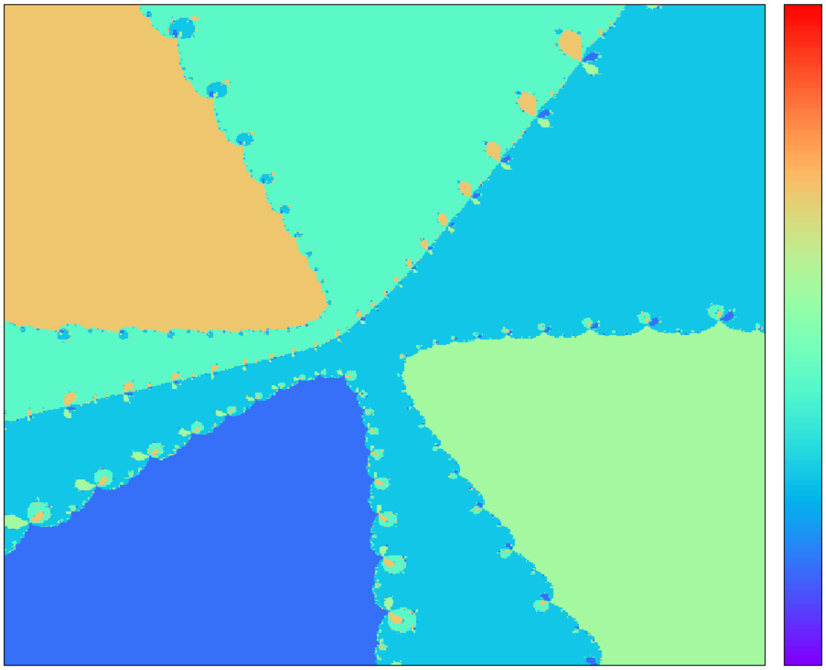
Resposta. Não, os erros de arredondamento podem fazer com que a ordem das somas faça diferença. Por exemplo, considere um sistema com um dígito binário de precisão $(1.b_1)$ e expoentes entre -4 e 4, por exemplo. Escolha os números $x = 1 = 2^0$ e $y = z = 1/4 = 2^{-2}$. Teremos, na notação do sistema (base binária) que $(x \oplus y) \oplus z = (1.0 \times 2^0 \oplus 1.0 \times 2^{-2}) \oplus 1.0 \times 2^{-2} = 1.0 \times 2^0 \oplus 1.0 \times 2^{-2} = 1.0 \times 2^0$, por outro lado, $x \oplus (y \oplus z) = 1.0 \times 2^0 \oplus 1.0 \times 2^{-1} = 1.1 \times 2^0$. □

Parte 3: Exemplos

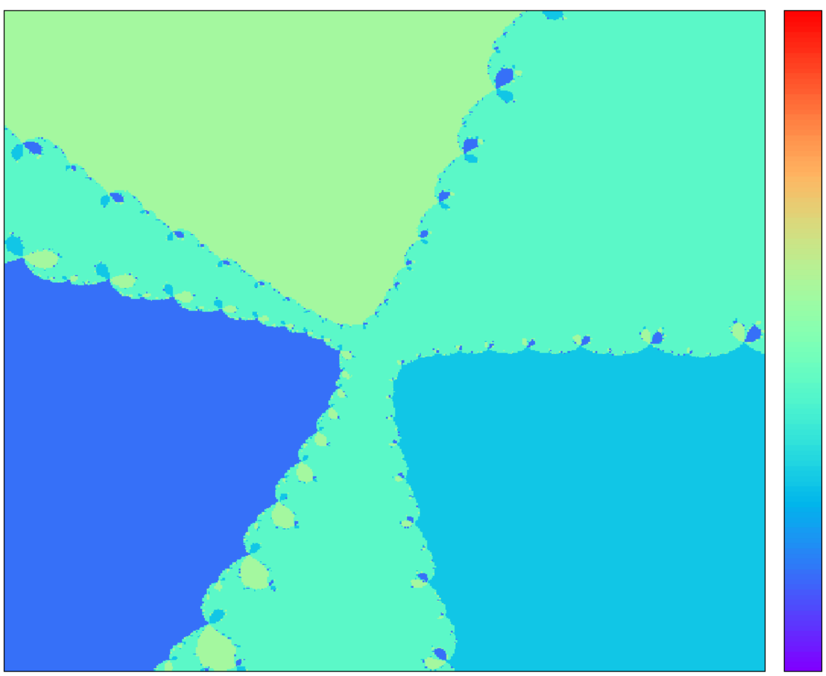
Para ilustrar o uso do método desenvolvido na parte 3 do EP. Escolhemos gerar as *Newton Basins* para um polinômio e as expansões de Taylor deste polinômio em torno do ponto 0. Para isso, escolhemos a função $f(x) = x^6 + (3 - 7i)x^5 + (-20 - 15i)x^4 + (-40 + 20i)x^3 + (-26 + 40i)x^2 + (52 + 52i)x$ e rodamos nosso EP nela.



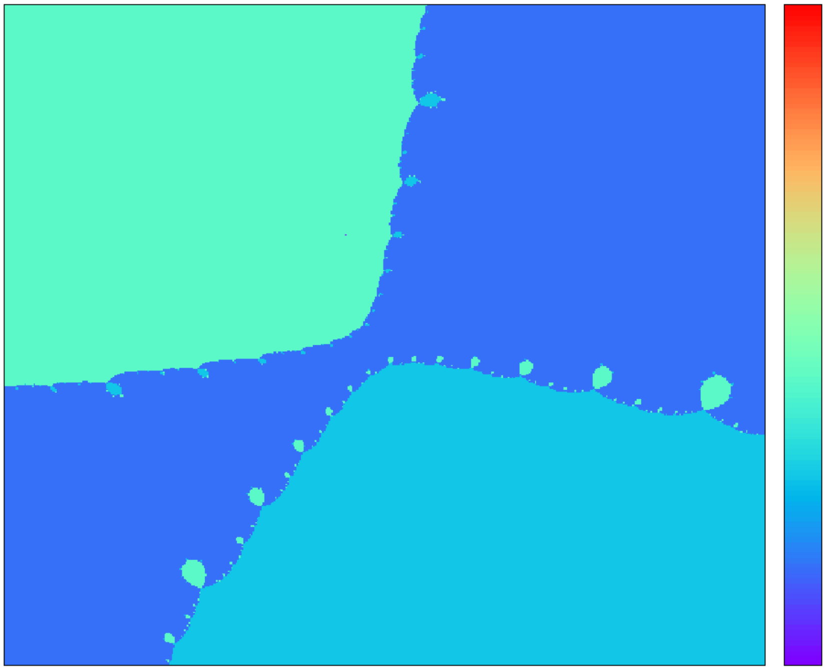
E, como planejado, repetimos o processo para as expansões de Taylor de diversos graus deste polinômio. Denotaremos a i-ésima expansão por $f_i(x)$. Primeiro, expandimos em grau 5, ou seja $f_5(x) = (3 - 7i)x^5 + (-20 - 15i)x^4 + (-40 + 20i)x^3 + (-26 + 40i)x^2 + (52 + 52i)x$.



Depois $f_4(x) = (-20 - 15i)x^4 + (-40 + 20i)x^3 + (-26 + 40i)x^2 + (52 + 52i)x$.



Depois $f_3(x) = (-40 + 20i)x^3 + (-26 + 40i)x^2 + (52 + 52i)x$.



Seguido de $f_2(x) = (-26 + 40i)x^2 + (52 + 52i)x$.

E, finalmente $f_1(x) = (52 + 52i)x$.