

Lista de Exercícios 1
MS211 - 2020/S1
Aritmética de Ponto Flutuante

1. Converta para a base decimal os seguintes números binários:

- | | |
|--------------------|--------------------|
| a) $(0)_2$ | d) $(101)_2$ |
| b) $(10)_2$ | e) $(111111111)_2$ |
| c) $(101010101)_2$ | f) $(1000001)_2$ |

2. Converta para a base binária os seguintes números decimais:

- | | |
|----------------|------------------|
| a) $(0)_{10}$ | d) $(101)_{10}$ |
| b) $(10)_{10}$ | e) 1979 |
| c) 25 | f) $(2615)_{10}$ |

3. Converta para a base decimal os seguintes números binários:

- | | |
|----------------------|-----------------------|
| a) $(1, 1)_2$ | d) $(1100, 01)_2$ |
| b) $(0, 001)_2$ | e) $(11111, 11111)_2$ |
| c) $(11100, 0011)_2$ | f) $(0, 000001)_2$ |

4. Converta para a base binária os seguintes números decimais:

- | | |
|----------------------|---------------------|
| a) $(0, 1)_{10}$ | d) 19, 625 |
| b) $(1100, 01)_{10}$ | e) $-\frac{3}{64}$ |
| c) 25, 12 | f) $(3, 1416)_{10}$ |

5. Um número real na base b em aritmética de ponto flutuante de n dígitos tem a forma geral

$$\pm(d_1d_2 \dots d_n) \times b^e$$

onde $(, d_1d_2 \dots d_n)$ é a mantissa, $0 \leq d_j \leq b - 1$, $j = 1, 2, \dots, n$; e é o expoente, $e \in [e_1, e_2]$, $e_1 \leq 0$ e $e_2 \geq 1$ sendo números inteiros. Se $d_1 \neq 0$, diz-se que o número está normalizado. Escreva os seguintes números decimais em ponto flutuante na forma normalizada:

- | | |
|--------------|-------------------|
| a) -279, 15 | d) 10, 093 |
| b) 1, 35 | e) $\frac{1}{64}$ |
| c) 0, 024712 | f) 2019 |

6. Um sistema de ponto flutuante pode ser expresso pela função

$$F = F(b, n, e_1, e_2).$$

Por exemplo, dado o sistema $F(10, 3, -4, 4)$, o número $x = -279, 15$ é representado como $x = -0, 279 \times 10^3$. Dados os sistemas de aritmética de ponto flutuante a seguir, represente os números (utilize truncamento), indicando possíveis casos de *underflow* e *overflow*.

- | | |
|----------------------|----------------|
| a) $F(10, 3, -4, 4)$ | |
| i) 1, 35 | iv) π |
| ii) 0, 024712 | v) -0, 0000007 |
| iii) -10, 093 | vi) 102983, 65 |

b) $F(2, 4, -2, 2)$

i) $(10, 01)_2$

iv) $(1111, 01)_2$

ii) $(0, 0100)_2$

v) $-(0, 001)_2$

iii) $-(11, 111)_2$

vi) $(1, 0001)_2$

7. Seja o sistema de ponto flutuante $F(b, n, e_1, e_2)$.

a) Qual o menor número, em módulo, diferente de zero que pode ser representado nesse sistema?

b) E o maior?

c) Qual o número de mantissas positivas? Resp.: $M = (b - 1)b^{n-1}$

d) Mostre que o número de números de pontos flutuantes possíveis é dado por

$$\#F = 2(b - 1)b^{n-1}(e_2 - e_1 + 1) + 1$$

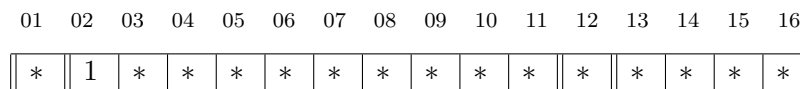
8. Determine (em valores absolutos) o maior e o menor número representado pelos seguintes sistemas:

a) $F(10, 3, -4, 4)$

b) $F(10, 4, -4, 5)$

c) $F(2, 4, -2, 2)$

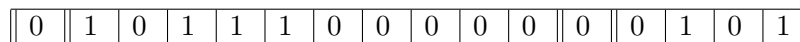
9. O sistema de ponto flutuante $F(2, 10, -15, 15)$ pode ser representado em um computador da seguinte forma, ocupando ao todo 16 bits:



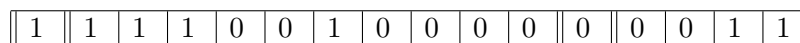
onde

- Posição 1: sinal da mantissa (0 = + ou 1 = -)
- Posições 2 até 11: mantissa (10 dígitos)
- Posição 12: sinal do expoente (0 = + ou 1 = -)
- Posição 13 até 16: representação do expoente

Por exemplo, os números $(23)_{10} = (0, 1011100000)_2 \times 2^5$ (lembrete: $(5)_{10} = (101)_2$) e $(-7, 125)_{10} = -(0, 1110010000)_2 \times 2^3$ (lembrete: $(3)_{10} = (11)_2$) possuem, respectivamente, a seguinte representação:



e



a) Determine o maior e o menor número decimal (em termos absolutos) que podem ser representados nesse sistema.

b) Represente os seguintes números nesse computador:

i) -3,1416

ii) 0,064

iii) $e \times 10^{-5}$

10. Dados os números x e y , efetue as operações

$$x + y, \quad x - y, \quad xy, \quad x/y$$

apresentando o resultado exato obtido, além do resultado truncado e do arredondado, com 4 dígitos:

a) $x = 0,937 \times 10^4$ e $y = 0,1272 \times 10^2$

b) $x = 3.14159$ e $y = 4,0 \times 10^4$

11. Efetue as operações indicadas, com 3 dígitos, utilizando arredondamento:

a) $(11,4 + 3,18) + 5,05$ e $11,4 + (3,18 + 5,05)$

b) $\frac{(3,18 \times 11,4)}{5,05}$ e $\left(\frac{3,18}{5,05}\right) \times 11,4$

c) $3,18 \times (5,05 + 11,4)$ e $3,18 \times 5,05 + 3,18 \times 11,4$

12. Considere uma máquina com sistema de representação de números definido por: base 10 ($\beta = 10$), 4 dígitos na mantissa ($t = 4$) e expoente no intervalo: $[-5; 5]$. Pede-se:

a) Qual o menor e o maior número em módulo representado nessa máquina?

b) Como será representado o número 73758 nesta máquina se for usado o arredondamento? E se for usado o truncamento?

c) Se $a = 42450$ e $b = 3$, qual o resultado de $a + b$ se for usado o arredondamento? E se for usado o truncamento? Justifique o resultado.

d) Considerando ainda $a = 42450$ e $b = 3$, qual o resultado da operação $a + \sum_{i=1}^{10} b$, considerando que está sendo realizado o truncamento?

e) Repetir o item d) para a operação $\sum_{i=1}^{10} b + a$.

f) Considere $a = 4245$, $b = 300$ e $c = 100$. Qual o resultado obtido nesta máquina para d e e , calculados de acordo com: $d = (a * b)/c$ e $e = a * (b/c)$. Justifique!

g) O que podemos concluir sobre a validade das propriedades como: comutativa, associativa, elemento neutro da adição de números em aritmética de ponto flutuante?

13. Seja o polinômio $P(x) = 2,3x^3 - 0,6x^2 + 1,8x - 2,2$. Deseja-se obter o valor de $P(x)$ para $x = 1,61$.

a) Calcule $P(1,61)$ com todos os algarismos da sua calculadora, sem efetuar arredondamento.

b) Calcule $P(1,61)$ considerando o sistema $F(10, 3, -4, 3)$, utilizando arredondamento a cada operação efetuada.

14. Considere $x = 0,5289$, $y = 0,8012$ e $z = 0,6024$ e operações em ponto flutuante numa mantissa com 4 dígitos (os números são sempre arredondados e normalizados após cada operação). Mostre que:

a) $x \times (y + z) \neq x \times y + x \times z$

b) $(x + y) + z \neq x + (y + z)$

15. Seja o número $x = (0,3)_{10}$

a) Escreva sua representação binária.

b) Escreva sua representação em ponto flutuante normalizado $\bar{x} = m \times b^e$ segundo o sistema $F = F(2, 5, -7, 7)$, utilizando truncamento.

c) Transforme a representação truncada da letra b) em decimal $\bar{x} = (?)_{10}$.

d) Calcule o erro absoluto $EA_x = x - \bar{x}$ e o erro relativo $ER_x = EA_x/\bar{x}$.

16. Sejam $EA_x = x - \bar{x}$ o erro absoluto e $ER_x = EA_x/\bar{x}$ o erro relativo. Mostre que o erro relativo na representação de um número em um sistema $F(b, n, e_1, e_2)$, com arredondamento, é limitado por

$$|ER_x| < \frac{1}{2} \times b^{1-n}.$$

(Sugestão: ver livro Ruggiero e Lopes)

17. (Opcional) Mostre que:

- a) $EA_{x+y} = EA_x + EA_y$
- b) $EA_{x-y} = EA_x - EA_y$
- c) $EA_{xy} \approx \bar{x}EA_y + \bar{y}EA_x$
- d) $EA_{x/y} \approx \frac{\bar{x}EA_y - \bar{y}EA_x}{\bar{y}^2}$

(Sugestão: ver livro Ruggiero e Lopes)

18. (Opcional) Mostre que:

- a) $ER_{x+y} = \left(\frac{\bar{x}}{\bar{x} + \bar{y}}\right) ER_x + \left(\frac{\bar{y}}{\bar{x} + \bar{y}}\right) ER_y$
- b) $ER_{x-y} = \left(\frac{\bar{x}}{\bar{x} - \bar{y}}\right) ER_x - \left(\frac{\bar{y}}{\bar{x} - \bar{y}}\right) ER_y$
- c) $ER_{xy} \approx ER_x + ER_y$
- d) $ER_{x/y} \approx ER_x - ER_y$

(Sugestão: ver livro Ruggiero e Lopes)

19. Precisão de máquina. A precisão da máquina é definida como sendo o menor número positivo em aritmética de ponto flutuante, ε , tal que: $(1 + \varepsilon) > 1$.

- a) Dada esta definição, podemos afirmar que a *precisão da máquina* é igual ao menor número representado pela máquina? Por que?
- b) O algoritmo a seguir estima a precisão da máquina:

Passo 1: $A = 1$
 $s = 1 + A$
 $k = 1$

Passo 2: Enquanto $s > 1$, faça:

$A = A/2$
 $s = 1 + A$
 $k = k + 1$

Passo 3: Faça $Prec = A * 2$ e imprime *Prec*

- b.1) Teste este algoritmo usando o MatLab ou uma linguagem a sua escolha. Trabalhe em precisão simples e em precisão dupla. O MatLab trabalha sempre em precisão dupla. Uma forma de trabalhar em precisão simples é declarar as variáveis como `single` e usar `single` em cada expressão do lado direito da igualdade. Exemplo:

```
A = single(1);
s = single(1 + A);
k = 1;
while (s > 1)
    A = single(A/2);
    s = single(1+A);
    k = k + 1;
end
prec = single(A*2);
```

Compare os valores obtidos com os resultados apresentados no MatLab ao se dar os comandos:

`eps` que resulta $2,2204 \times 10^{-16}$ em precisão dupla, e;

`eps('single')` que resulta em $1,1921 \times 10^{-7}$ em precisão simples.

20. Cálculo de $\exp(x)$. O objetivo é calcular $\exp(x)$ pela série de Taylor até ordem n em torno de zero:

$$\exp(x) \simeq 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots + \frac{x^n}{n!}$$

- a) Escreva um programa (no MatLab) para obter uma aproximação para $\exp(x)$ de acordo com a expressão anterior. O valor de x e número de termos da série, n , são dados de entrada do programa. Observe que cálculo do fatorial, $k!$, necessário na série de Taylor, pode ser feito de modo a evitar a ocorrência de *overflow*. Evita-se o *overflow* desde que se observe que o termo (k) pode ser escrito como: $x^k/k! = x^{k-1} * x/(k-1)! * k$, onde o termo $x^{k-1}/(k-1)!$ já está calculado, pois a série está sendo avaliada a partir do primeiro termo. (Um erro comum no uso da fórmula de Taylor para o cálculo de $\exp(x)$ é escrever “procedimentos” para avaliar o fatorial: o valor de k é dado de entrada e a saída é $k!$. Nestes casos, há ocorrência de *overflow*). Evitando o *overflow*, a série de Taylor pode ser calculada com tantos termos quanto se queira. Qual seria um critério de parada para se interromper o cálculo da série, que não seja a comparação com seu valor real de $\exp(x)$?
- b) Teste seu programa para vários valores de x : positivos, negativos, ($x \approx 0$ e x distante de zero) e, para cada valor de x , teste o cálculo da série com vários valores para o número de termos n . Analise os resultados obtidos.
-