

# CI202 - Lista 1

## Representação Numérica

Prof. Ricardo Oliveira

**Atenção:** Apenas os resultados finais dos exercícios são apresentados, para fins de conferência. Os cálculos que levam a estes resultados devem ser realizados.

**Obs:** Caso encontre algum erro em algum exercício ou resposta, por favor avise o professor.

1. (a)  $EA = 25, ER \approx 0.37313$   
(b)  $EA = -25, ER \approx -0.27174$   
(c)  $EA = ER = 0$   
(d)  $EA = -0.001, ER \approx -0.000359$   
(e)  $EA = 0.12, ER \approx 0.000352$   
(f)  $EA = 0.12, ER \approx 0.27907$   
(g)  $EA = -0.001, ER = -1$   
(h)  $EA = 0.001, ER$  não é definido
2. (a)  $|EA| \leq 6, |ER| \leq 0.1364$   
(b)  $|EA| \leq 4, |ER| \leq 0.08696$   
(c)  $|EA| \leq 6, |ER| \leq 0.125$   
(d)  $|EA| \leq 0.087, |ER| \leq 0.0676$   
(e)  $|EA| = |ER| = 0$
3. (a)  $|EA| \leq 8$   
(b)  $|EA| \leq 0.001$   
(c)  $|EA| \leq 0.18$
4. (a) 0  
(b) 7  
(c) 16  
(d) 42  
(e) 914
5. (a) 1010  
(b) 100101  
(c) 1010101111  
(d) 1111111  
(e) 10000000
6. (a)  $0.5 \times 10^0$   
(b)  $0.9 \times 10^1$   
(c)  $0.203125 \times 10^0$   
(d)  $0.256 \times 10^3$   
(e)  $0.14648438 \times 10^{-1}$   
(f)  $0.4 \times 10^1$
7. (a)  $0.101 \times 2^0$   
(b)  $0.10101 \times 2^6$   
(c)  $0.101 \times 2^4$

- (d)  $0.1 \times 2^1$
  - (e)  $0.11001100110011\dots \times 2^{-3}$
8. (a)  $0.3802 \times 10^2$ ;  $EA = 0$
- (b)  $0.1021 \times 10^{-3}$ ;  $EA = 0.00000007$
- (c)  $0.1372 \times 10^6$ ; (*overflow*)
- (d)  $0.3958 \times 10^{-2}$ ;  $EA = 0.0000004545$
- (e)  $0.1012 \times 10^{-7}$ ; (*underflow*)
9. (a)  $0.11 \times 2^1$ ;  $EA = -0.02734375$
- (b)  $0.11001 \times 2^9$ ;  $EA = 0$
- (c)  $0.11 \times 2^{-12}$ ; (*underflow*)
- (d)  $0.1 \times 2^{11}$ ; (*overflow*)
10. O tipo `float` da linguagem C/C++ representa números reais utilizando o padrão IEEE 754 de 32 bits ( $\beta = 2, t = 24, I = -127, S = 128$ ). O fenômeno ocorre devido à representação finita e principalmente ao acúmulo de erros na aritmética de ponto flutuante utilizada pelo computador. O tipo `double`, por sua vez, utiliza o padrão de 64 bits ( $\beta = 2, t = 53, I = -1023, S = 1024$ ), que pode representar (muito) mais números e números mais precisos que o tipo `float`.