SME0305 - Lista 2

Representação, arredondamento, Octave

Lembremos que $F(\beta, t, m, M)$ é o conjunto de números da forma:

$$\bar{x} = \pm\ (0.d_1 d_2 \ldots d_t)_\beta \times \beta^e$$

em que β é a base, t é o número de digitos significativos na mantissa $0.d_1 d_2 \ldots d_t$, $1 \leq d_1 \leq \beta - 1$, $0 \leq d_i \leq \beta - 1$ para $i = 2, \ldots, t$ e finalmente o número enteiro $e \in [m, M]$ é o expoente.

---

1. Faça um programa em Octave que, dado um número x, retorne a representação de x em base b. Em seguida,

   (a) Compare com o seguinte programa:

```
% Dados a serem fornecidos pelo
% usuario: num, base, ndig
%clear;
pe = floor(num);
pf = num - pe;
i = 1;
while (pe ~= 0)
  quo = floor(pe/base);
  rest = pe - quo*base;
  de(i) = rest;
  i = i + 1;
  pe = quo;
end
for i=1:ndig
  aux = pf*base;
  df(i) = floor(aux);
  m = aux - df(i);
  pf = m;
end
```

   (b) Explique de maneira bem suscinta qual é o resultado esperado nos arrays `de(:)` e `df(:)`. Qual será a dimensão desses arrays?

   (c) Aplique o algoritmo na mão para `num = 3.125`, `base = 2`, `ndig = 3`.

   (d) Execute o algoritmo em Octave e imprima os arrays `de(:)` e `df(:)` quando `num = 23.48`, `base = 2` e `base = 10` e `ndig = 5`.

2. Considere a definição de $F(\beta, t, m, M)$.

   (a) Determine quantos números distintos há no conjunto.

   (b) Particularize a resposta do item anterior para $F(2, 4, -2, 1)$.

3. Faça um programa que implemente a função `fl(x)`, que retorna o valor numérico da representação de ponto flutuante de `x`. Comparar com a programada no código `fl.m` disponibilizada pela cátedra.

4. A precisão de maquina ϵ_M é definida como a distancia de 1 ao menor número maior que 1 dentro do conjunto.

   (a) Mostre que $\epsilon_M = \beta^{1-t}$.

   (b) Calcule ϵ_M para $F(2, 4, -2, 1)$

   (c) Calcule ϵ_M para $F(2, 53, -1021, 1024)$. Verifique a resposta usando o comando `eps` em Octave ou Matlab, que utilizam tal sistema.

5. Para o sistema $F(\beta, t, m, M)$

   (a) Provar que o menor número real positivo $x_{\min}$ e o maior número real positivo $x_{\max}$ que podem ser representados no sistema são:

   $$x_{\min} = \beta^{m-1}, \quad x_{\max} = \beta^M (1 - \beta^{-t})$$

   (b) Particularize a resposta para $F(10, 3, -2, 1)$.

   (c) Particularize a resposta para $F(2, 53, -1021, 1024)$. Neste caso verifique a resposta usando os comandos `realmin` e `realmax` em Octave.

   (d) Que acontecerá durante a execução de um programa em que uma variável recebe um valor menor que $x_{\min}$ ou um valor maior que $x_{\max}$? Faça um teste em Octave.

6. Usando a função `fl.m` determine o valor ótimo de δ (i.e., o que gera menor error) para o cálculo da derivada de $f(x) = 1/(x - \pi)^2$ no ponto $x = \pi + 100$ utilizando a aproximação

   $$f'(x) \simeq \frac{f(x + \delta) - f(x)}{\delta} .$$

   Calcule o δ ótimo para três valores de t (a base β é sempre igual a 2, assim como $m = -1021$ e $M = 1024$), sendo estes 53 (double), 24 (single) e 11 (half).

7. Usando a precisão por defeito do Octave, calcule numéricamente o δ ótimo para o cálculo de $f''(x) = (f')'(x)$ quando toda derivada é aproximada como no item precedente, $f(x) = 1/x$, e $x = 100$.

8. Seja v um vetor arbitrário em dimensão 3. Construa uma base ortogonal de $\mathbb{R}^3$ cujo primeiro vetor seja v. Implemente essa função em Octave.

9. O plano definido pelos pontos O, P e Q é o teto de um prédio. Um raio de luz que incide verticalmente sobre ele, em que direção é refletido? Responda em Octave.

10. Seja `A` uma matriz $n \times m$ e seja V o espaço vetorial de todas as combinações lineares de suas colunas (que supomos l.i.). A matriz `M=calcbase2(A)` deveria ter por colunas m vetores que são base ortogonal de V. Como checar que

    (a) as colunas de `M` pertencen mesmo a V?

    (b) as colunas de `M` são mesmo ortogonais?

    Responda em Octave.