SME0602 - 2017 - Lista Ponto Flutuante

Representação, arredondamento, Octave

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

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

em que β é a base, t é o número de dígitos significativos na mantissa $0.d_1 d_2 \dots d_t$, $1 \leq d_1 \leq \beta - 1$, $0 \leq d_i \leq \beta - 1$ para $i = 2, \dots, t$ 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: num, base, ndig
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. 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.

4. 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.

5. Considere a definição de $C = F(3, 5, -3, 3)$.

   (a) Indique qual o número de C mais próximo de $\sqrt{2}$. Expressar o resultado em base 3 e em base 10. Quanto vale o erro de arredondamento relativo nesse caso?

   (b) Calcule ϵ_M.

   (c) Calcule o maior e o menor número positivo representável no conjunto C, expressados em base 3 e em base 10.

6. Diga qual é o menor elemento de $F(2, 10, -20, 20)$ que é maior ou igual que $x = 54.11$. Compreende-se que x está escrito em base 10. O resultado é pedido em dois formatos:
   a) Escrever a mantissa $d_1 \dots d_t$ e o expoente e.
   b) Em base 10.

7. Responda com verdadeiro (V) ou Falso (F).

   - Qualquer número real pode ser representado exatamente no computador utilizando o sistema $F(\beta, t, m, M)$.
   - Num sistema $F(\beta, t, m, M)$ a precisão de máquina é β^{1-t}
   - Para o número $\bar{x} \in F(\beta, t, m, M)$, se $|\bar{x}| < \beta^{m-1}$, obtemos erro de underflow.
   - Para o número $\bar{x} \in F(\beta, t, m, M)$, se $|\bar{x}| < \beta^M(1 - \beta^{-t})$, obtemos erro de overflow.
   - O resultado num calculador ao fazer $\bar{x} \otimes (\bar{y} \oplus \bar{z})$ será em geral o mesmo que ao fazer $(\bar{x} \otimes \bar{y}) \oplus (\bar{x} \otimes \bar{z})$.
   - O resultado num calculador ao fazer $\bar{x} \oplus (\bar{y} \oplus \bar{z})$ poderá ser diferente que ao fazer $(\bar{x} \oplus \bar{y}) \oplus \bar{z}$.