Cálculo Numérico - IME/UERJ

Trabalho extra 1 - Gabarito

1. **(Trabalho extra 1 - Valendo 1,0 ponto)** Seja um computador binário de precisão simples, ou seja, de 32 bits, cujo sistema de ponto flutuante armazena 1 bit para o sinal do número, 8 bits para o expoente e 23 bits para a mantissa. Responda **justificando cada item**:

   (a) Qual o maior número positivo nele representável?

   Resposta:

   Na forma normalizada, a gama de variação do expoente é de 00000001 a 11111110, isto é, em decimal, de 1 a 254. No padrão IEEE-754, tomando 01111111 (127, em decimal) para representar o zero, 00000001 (1, em decimal) será $1 - 127 = -126$ e 11111110 (254, em decimal) será $254 - 127 = +127$. Nos oito bits reservados para o expoente, representaremos o expoente desejado mais 127.

   Maior número positivo:

   Bit de sinal: 0

   Expoente: $11111110 \Rightarrow 254 - 127 = 127$

   Mantissa: 11111111111111111111111 (23 bits)

   Forma normalizada:

   $1,11111111111111111111111 \times 2^{127} = \mathbf{2^{128} - 2^{104} \approx 3.4028235 \times 10^{38}}$.

   (b) Qual o menor número positivo nele representável?

   Resposta:

   Estou assumindo que se trata do menor número **não normalizado**, para podermos chegar ainda mais próximo a zero (underflow gradual).

   Bit de sinal: 0

   Expoente: $00000000 \Rightarrow 1 - 127 = -126$ (na forma desnormalizada)

   Mantissa: 00000000000000000000001 (23 bits)

   Forma desnormalizada:

   $0,00000000000000000000001 \times 2^{-126} = 2^{-23} \times 2^{-126}$

   $= \mathbf{2^{-149} \approx 1.4012985 \times 10^{-45}}$.

   (c) Qual o erro relativo máximo considerando que houve truncamento ao aproximar um certo número?
   Dica: Primeiro calcule, por exemplo, o erro relativo do número 3.6, onde ocorrerá truncamento na aproximação, e depois calcule o erro relativo máximo para este computador.

Resposta:

Na forma normalizada,

$(3.6)_{real} = 1,1100110011001100110011001 \cdots \times 2^1.$

Na máquina com 23 bits para a mantissa, ele recebe um truncamento e fica representado por:

$(3.6)_{ap} = 1,11001100110011001100110 \times 2^1 \approx 3.59999990463.$

Assim, o erro absoluto neste caso é dado por:

$$
\begin{array}{rrlr}
 & (3,6)_{real} & = 1,{\color{red}11001100110011001100110}0110011001\ldots & \times\ 2^1 \\
- & (3,6)_{ap} & = 1,{\color{red}11001100110011001100110} & \times\ 2^1 \\
\hline
 & & \ 0,{\color{red}00000000000000000000000}0110011001\ldots & \times\ 2^1
\end{array}
$$

Ou seja,

$E_{abs} = 0,000000000000000000000000110011001\ldots\ \times\ 2^1$

$E_{abs} = 0,0110011001\ldots\ \times\ 2^1 \times 2^{-23}.$

Na forma normalizada,

$E_{abs} = 1,10011001\ldots\ \times\ 2^1 \times 2^{-23} \times 2^{-2}.$

Assim, o erro relativo na forma normalizada é dado por:

$E_{rel} = \dfrac{E_{abs}}{(3.6)_{real}} = \dfrac{1,10011001\ldots\ \times\ 2^1 \times 2^{-23} \times 2^{-2}}{1,110011001 \cdots \times 2^1}$

Portanto, o erro relativo máximo é dado pelo maior erro absoluto dividido pelo menor valor real na forma normalizada. Ou seja,

$(E_{rel})_{MAX} = \dfrac{1,11111111\ldots\ \times\ 2^1 \times 2^{-23} \times 2^{-2}}{1,000000000 \cdots \times 2^1}$

$(E_{rel})_{MAX} = 1,11111111\ldots\ \times\ 2^{-2} \times 2^{-23}$

$(E_{rel})_{MAX} = 0,0111111111 \cdots \times 2^{-23} < 2^{-1} \times 2^{-23} = 2^{-24}.$

(d) Qual o valor representado por 12.8 neste computador?

Resposta:

Primeira maneira:

Na forma normalizada,

$(12.8)_{real} = 1,100110011001100110011001100 \cdots \times 2^3.$

Na máquina com 23 bits para a mantissa, ele recebe um arredondamento e fica representado por:

$(12.8)_{ap} = 1,1001100110011001100110{\color{red}1} \times 2^3 \approx \mathbf{12.8000017166}.$

Segunda maneira:

Calcular o erro absoluto:

$$
\begin{array}{rl}
 & (12.8)_{ap} \ = 1,10011001100110011001101 \qquad\qquad \times 2^3 \\
- & \underline{(12.8)_{real} = 1,1001100110011001100110011001100\ldots \times 2^3}
\end{array}
$$

é o mesmo que calcular:

$$
\begin{array}{rl}
 & 0,00000000000000000000001 \qquad\qquad \times 2^3 \ = 1,00000000\cdots \times 2^{-23} \times 2^3 \\
- & \underline{0,0000000000000000000000001100\ldots \times 2^3 \ = 0,11001100\cdots \times 2^{-23} \times 2^3}
\end{array}
$$

Ou seja,

$E_{abs} = (1 - \underbrace{0,11001100\ldots}_{0.8}) \times 2^{-23} \times 2^3 = 0.2 \times 2^{-20} \approx 1.907348633 \times 10^{-7}$.

Assim,

$(12.8)_{ap} = (12.8)_{real} + 1.907348633 \times 10^{-7} \approx \mathbf{12.80000019}$.

(e) Qual o valor representado por 28.8 neste computador?

Resposta:

Primeira maneira:

Na forma normalizada,

$(28.8)_{real} = 1,1100110011001100110011001100\cdots \times 2^4$.

Na máquina com 23 bits para a mantissa, ele recebe um truncamento e fica representado por:

$(28.8)_{ap} = 1,11001100110011001100110 \times 2^4 \approx \mathbf{28.7999992371}$.

Segunda maneira:

Calcular o erro absoluto:

$$
\begin{array}{rl}
 & (28.8)_{real} = 1,1100110011001100110011001100\ldots \times 2^4 \\
- & \underline{(28.8)_{ap} \ = 1,11001100110011001100110 \qquad\qquad \times 2^4}
\end{array}
$$

é o mesmo que calcular:

$$
\begin{array}{rl}
 & 0,00000000000000000000000\,01100\ldots \times 2^4 \;= 0,011001100\cdots \times 2^{-23} \times 2^4 \\
- & \underline{0,00000000000000000000000 \qquad\qquad\;\; \times 2^4 \qquad\qquad\qquad\qquad\qquad\qquad}
\end{array}
$$

Ou seja,

$E_{abs} = 0,011001100 \cdots \times 2^{-23} \times 2^4 = \underbrace{0,110011001100\ldots}_{0.8} \times 2^{-1} \times 2^{-23} \times 2^4$

$E_{abs} = 0.8 \times 2^{-1} \times 2^{-23} \times 2^4 \approx 7.629394531 \times 10^{-7}$.

Assim,

$(28.8)_{ap} = (28.8)_{real} - 7.629394531 \times 10^{-7} \approx \mathbf{28.79999924}$.