Gabarito da primeira prova de Cálculo Numérico
Primeiro trimestre de 2012
prof. Rodrigo Fresneda

23 de abril de 2012

1. (2 pontos) A representação de ponto flutuante de precisão dupla (64bits) no padrão IEE754 é caracterizada por 1 bit de sinal S, 11 bits para o expoente E, e 52 bits para a fração

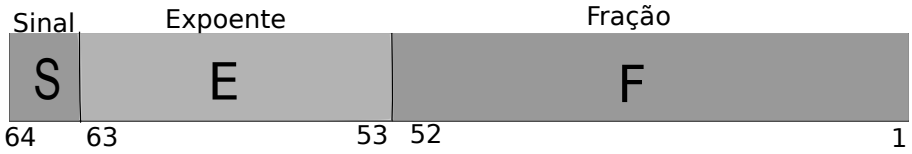

Figura 0.1: Representação de ponto flutuante em IEE754-64bits

Nessa representação, números normalizados são dados pela fórmula

$$x = (-1)^S * 2^{E-1023} * 1.F$$

em que $0 < E < 2^{11} - 1$ e $0 \leqslant F < 1$. Em relação a esse padrão, responda:

(a) Qual o maior número normalizável (base 10)?

$$\begin{aligned} |x| &= 2^{2046-1023} * 1.11111... = 2^{1023} * \left(2 - 2^{-52}\right) \\ &= 1.79 \times 10^{308} \end{aligned}$$

(b) Qual o menor número normalizável (base 10)?

$$\begin{aligned} |x| &= 2^{1-1023} * 1.000... = 2^{-1022} \\ &= 2.22 \times 10^{-308} \end{aligned}$$

(c) Quanto vale eps?

$$\beta = 2^{-52} = 2.22 \times 10^{-16}$$

(d) Represente 0.1 (determine os inteiros S,E e F na base 2).
Na base 2, 1/10 tem a forma $\frac{1}{10} = \left(0.0001\overline{10011}\right)_2$. Na forma normalizada, temos

$$\frac{1}{10} = (-1)^0 \cdot 2^{-4} \cdot (1.10011001100110011001100110011001100110011001100110011010)_2$$

Assim, $S = 0$, $E = 1023 - 4 = 1019 = (01111111011)_2$ e

$$F = 10011001100110011001100110011001100110011001100110011010$$

2. (3.0 pontos) Considere a equação $xe^x - 1 = 0$.

(a) Encontre um intervalo que contenha uma raiz positiva dessa equação. Para tanto, esboçe um gráfico e use o teorema do anulamento.
Podemos encontrar o(s) zero(s) da $f(x)$, $x \neq 0$, esboçando o gráfico das funções $g(x) = e^x$ e $h(x) = \frac{1}{x}$:

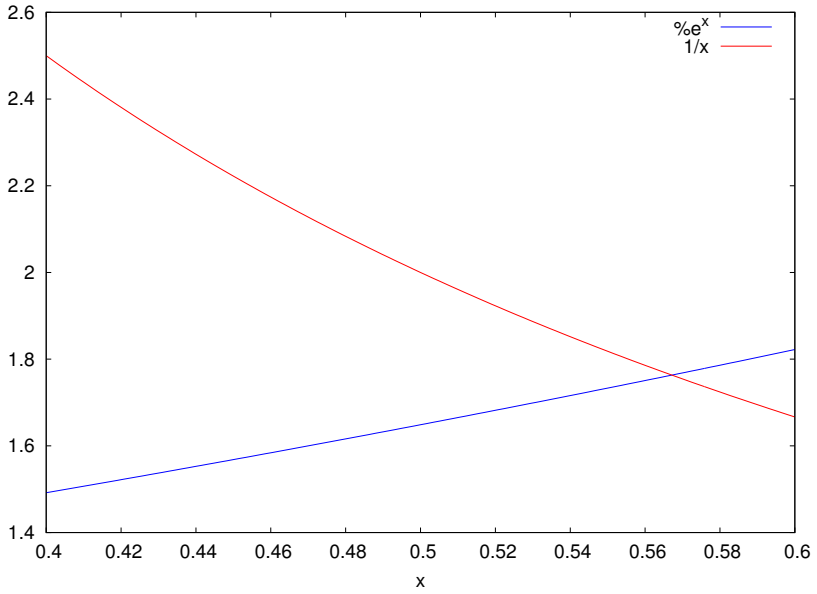

Figura 0.2: gráfico de e^x e $\frac{1}{x}$ no intervalo $[0.4, 0.6]$

Ou podemos esboçar o gráfico de $f(x) = e^x - \frac{1}{x}$ utilizando ferramentos do Cálculo. Para determinar seus extremos locais, resolvemos $f'(x) = (x+1)e^x = 0$, o que dá $x = -1$. Para $x < -1$, tem-se $f'(x) < 0$ ($f(x)$ é decrescente), e para $x > -1$ tem-se $f'(x) > 0$ ($f(x)$ é crescente). Logo, a função $f(x)$ tem um mínimo local em $x = -1$, $f(-1) = -\frac{1}{e} - 1$. Para analisar a concavidade da $f(x)$, tomamos a segunda derivada, $f''(x) = (x+2)e^x$. Temos que $x = -2$ é um ponto de inflexão ($f'''(-2) > 0$). Para $x < -2$, $f'' < 0$ ($f(x)$ tem concavidade para baixo), e para $x > -2$ $f'' > 0$ ($f(x)$ tem concavidade

para cima). Para determinar os extremos globais, calculamos $\lim f(x)_{x\to\infty} = \infty$ e $\lim_{x\to-\infty} f(x) = -1$

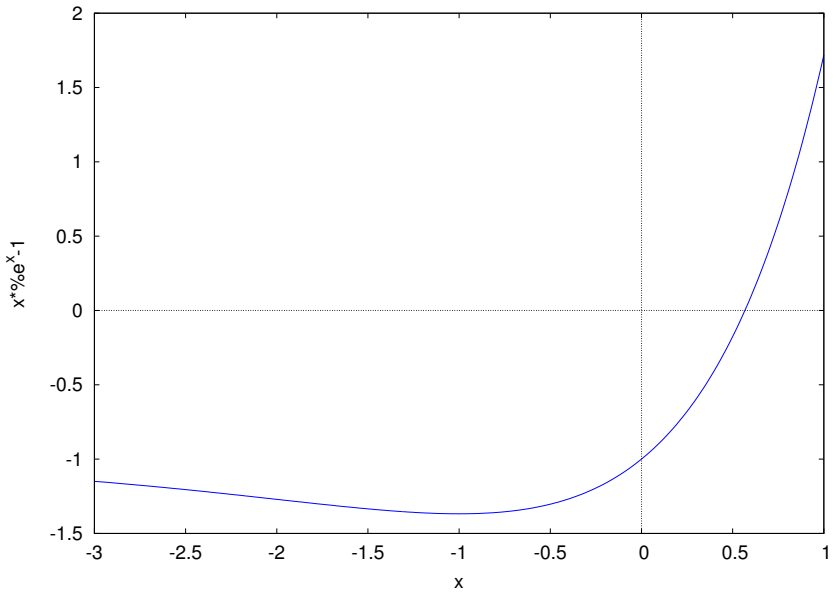

Figura 0.3: Gráfico da função $xe^x - 1$ no intervalo $[-1, 1]$

Aplicando o teorema do anulamento no intervalo $[0.4, 0.6]$ temos,

$$f(0.4)\, f(0.6) = -0.0376 < 0\,.$$

Logo, há uma raiz de $f(x)$ no intervalo $[0.4, 0.6]$.

(b) Determine uma função de iteração apropriada ao problema. Justifique sua escolha com base no teorema de convergência do método iterativo linear.
Consideremos a função de iteração $\varphi(x) = e^{-x}$ ($xe^x - 1 = 0 \Leftrightarrow x = e^{-x}$). φ e φ'são contínuas, e para termos $|\varphi'(x)| < 1$, x deve estar no intervalo

$$\left|e^{-x}\right| = e^{-x} < 1 \Rightarrow x > 0\,.$$

Como a raiz é positiva, podemos escolher qualquer intervalo I centrado na raiz e contido em $(0, \infty)$ e $x_0 \in I$, que o método iterativo linear com função de iteração φ convergirá.

(c) Utilize a função do item anterior para encontrar a raiz localizada no item (a) com erro inferior a 0.01.

i	$x_i = \varphi(x_{i-1})$	$\varepsilon = \lvert x_i - x_{i-1}\rvert / \lvert x_i\rvert$
0	$x_0 = 0.5$	
1	$x_1 = 0.6065$	0.1756
2	$x_2 = 0.5453$	0.1122
3	$x_3 = 0.5797$	0.05934
4	$x_4 = 0.5601$	0.03499
5	$x_5 = 0.5712$	0.01944
6	$x_6 = 0.5648$	0.01133
7	$x_7 = 0.5685$	0.006508

(d) Escreva uma função de iteração para esse problema tal que a convergência seja mais rápida.
Podemos utilizar a função de iteração do método de Newton:

$$\varphi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{xe^x - 1}{(x+1)e^x}$$

3. (2 pontos) Considere o sistema linear $Ax = b$, onde:

$$A = \begin{pmatrix} 1 & \alpha & 3 \\ \alpha & 1 & 4 \\ 5 & 2 & 1 \end{pmatrix},$$

(a) A matriz A é decomponível no produto LU? Justifique.
Para que A seja decomponível no produto LU, é necessário que os primeiros menores principais sejam não-nulos:

$$\det A_1 = 1\,, \det A_2 = 1 - \alpha^2,\ \det A_3 = -\alpha^2 + 26\alpha - 22$$

Assim, devemos ter $\alpha \neq \pm 1$ e $\alpha \neq \frac{26 \pm \sqrt{588}}{2}$

(b) O sistema pode ser resolvido por Cholesky? Justifique.
Não, pois não é simétrica para nenhum α.

4. (3.0 pontos) Considere o sistema linear $Ax = b$ onde

$$A = \begin{pmatrix} a & 3 & 1 \\ a & 20 & 1 \\ 1 & a & 6 \end{pmatrix},\ b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

(a) Para que valores de a teremos $\|B\|_\infty < 1$, em que B é a matriz de iteração do método de Jacobi?
A matriz de iteração para $a \neq 0$ é

$$B = -\begin{pmatrix} 0 & \frac{3}{a} & \frac{1}{a} \\ \frac{a}{20} & 0 & \frac{1}{20} \\ \frac{1}{6} & \frac{a}{6} & 0 \end{pmatrix} \text{ se } a \neq 0$$

As somas-linha devem ser inferiores a 1:

$$\left|-\frac{3}{a}\right| + \left|-\frac{1}{a}\right| < 1 \Rightarrow |a| > 4$$
$$\left|-\frac{a}{20}\right| + \left|-\frac{1}{20}\right| < 1 \Rightarrow |a| < 19$$
$$\left|-\frac{1}{6}\right| + \left|-\frac{a}{6}\right| < 1 \Rightarrow |a| < 5$$

Assim, $4 < |a| < 5$. Se $a > 0$, $4 < a < 5$ e se $a < 0$, $-5 < a < -4$.

(b) Para quais valores de a podemos afirmar que $\left\|x^{(k+1)} - \xi\right\|_\infty \leq \frac{1}{2}\left\|x^{(k)} - \xi\right\|_\infty$, em que $x^{(k+1)}$ e $x^{(k)}$ são aproximações para a solução ξ?
Em geral temos

$$\left\|x^{(k+1)} - \xi\right\|_\infty \leq \|B\|_\infty \left\|x^{(k)} - \xi\right\|_\infty$$

Então basta escolher $\|B\|_\infty \leq \frac{1}{2}$. Não há α que satifaça as desigualdades

$$\left|-\frac{3}{a}\right| + \left|-\frac{1}{a}\right| \leq \frac{1}{2} \Rightarrow |a| \geq 8$$
$$\left|-\frac{a}{20}\right| + \left|-\frac{1}{20}\right| \leq \frac{1}{2} \Rightarrow |a| \leq 9$$
$$\left|-\frac{1}{6}\right| + \left|-\frac{a}{6}\right| \leq \frac{1}{2} \Rightarrow |a| \leq 2$$

(c) Resolva o sistema dado pelo método de Jacobi com $a = -1$ e erro relativo inferior a 0.1 na norma do máximo.
O método não converge com $a = -1$.

5. (0.5 ponto) Mostre que se a matriz de coeficientes A é estritamente diagonal dominante, o critério de Sassenfeld é satisfeito, i.e., $\beta_i < 1$, $i = 1, ..., n$, em que

$$\beta_i = \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| \beta_j + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|$$

Demonstração. Se A é estritamente diagonal dominante, então

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}| \Rightarrow \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1\,,\ i = 1, ..., n\,.$$

Temos

$$\beta_1 = \sum_{j=2}^{n} \frac{|a_{1j}|}{|a_{11}|} = \sum_{j \neq 1} \frac{|a_{1j}|}{|a_{11}|} < 1$$

e

$$\beta_2 = \frac{|a_{21}|}{|a_{22}|}\beta_1 + \sum_{j=3}^{n}\frac{|a_{2j}|}{|a_{22}|} < \frac{|a_{21}|}{|a_{22}|} + \sum_{j=3}^{n}\frac{|a_{2j}|}{|a_{22}|} = \sum_{j\neq 2}^{n}\frac{|a_{2j}|}{|a_{22}|} < 1\,.$$

Suponha que $\beta_i < 1$ para $i = 1, ..., m$. Vamos mostrar que $\beta_{m+1} < 1$. Temos

$$\begin{aligned}
\beta_{m+1} &= \sum_{j=1}^{m}\frac{|a_{m+1,j}|}{|a_{m+1,m+1}|}\beta_j + \sum_{j=m+2}^{n}\frac{|a_{m+1,j}|}{|a_{m+1,m+1}|} \\
&< \sum_{j=1}^{m}\frac{|a_{m+1,j}|}{|a_{m+1,m+1}|} + \sum_{j=m+2}^{n}\frac{|a_{m+1,j}|}{|a_{m+1,m+1}|} \\
&= \sum_{j\neq m+1}\frac{|a_{m+1,j}|}{|a_{m+1,m+1}|} < 1
\end{aligned}$$

Logo $\beta_{m+1} < 1$, e provamos $\beta_i < 1$ para $i = 1, ..., n$. □