A_1 - Análise Numérica - 2021

1. (**4 pontos**) Seja S o conjunto formado pelos números reais que tem uma representação exata no computador (considere precisão dupla, i.e 64 bits) e cuja representação em ponto flutuante tem exatamente um bit "1".

   (a) Determine o maior e o menor número real positivo contido em S.

   (b) Determine um número real x tal que $x \notin S$, mas $fl(x) \in S$.

a) Temos que esses números têm o formato

(i) $1\ \overbrace{0 — 0}^{11}\ \overbrace{0 — 0}^{52}$; ou

(ii) $0\ 0 - 1 - 0\ 0 — 0$; ou

(iii) $0\ 0 — 0\ 0 - 1 - 0$

Em (i), temos o número -0

Em (ii), temos os números da forma

$$2^{2^{i-1}-1023},$$

em que $i = 1, \ldots, 11$ é a posição do bit 1.

Em (iii), em que o expoente é nulo, temos os números subnormais. Nesse caso temos o formato

$$2^{-1022} \cdot (0 + 2^{-i}) = 2^{-(1022+i)}$$

em que $i = 1, \ldots, 52$ é a posição na mantissa (da esquerda para a direita) do bit 1.

Concluímos que:

$$S = \{2^{2^{i-1}-1023} \mid i = 1, \ldots, 11\} \cup \{2^{-(i+1022)} \mid i = 1, \ldots, 52\} \cup \{-0\}$$

O maior em S é $2^{2^{11-1}-1023} = 2$

O menor em S é $2^{-(52+1022)} = 2^{-1074}$

b) Considere o número

$$0 \; \underbrace{10 - 0}_{\text{expoente}} \; \underbrace{0 - 0}_{\text{mantissa}} \; \underbrace{01}_{\text{resto}}$$

que tem representação decimal

$$(-1)^0 \, 2 \, (1 + 2^{-54}) = 2 + 2^{-53}$$

Temos que $x \notin S$ e de fato nem pode ser repentado por outro número de máquina. Entretanto, na operação de arredondamento da mantissa 01 é cortado e temos o número $fl(x) = 2 \in S$.

2. (**6 pontos**) Considere o sistema de equações lineares de $n \times n$, cuja j-ésima equação ($j = 1, \ldots, n-1$) é

$$\sum_{k=1}^{j} \frac{1}{2^{j-k}} x_k + \sum_{k=j+1}^{n} \frac{1}{2^k} x_k = \frac{1}{j}$$

e cuja n-ésima equação é

$$\sum_{k=1}^{n} \frac{1}{2^{n-k}} x_k = \frac{1}{n}$$

(a) O método de Jacobi aplicado a esse sistema é convergente $\forall n \in \mathbb{N}$? Justifique bem sua resposta.

(b) O método de Gauss-Seidel aplicado a esse sistema é convergente $\forall n \in \mathbb{N}$? Justifique bem sua resposta.

Seja $Ax = b$ nosso sistema.

$$\text{lin}_i(A) = (\overbrace{2^{1-i}, 2^{2-i}, \ldots, 1}^{i}, \overbrace{2^{-(i+1)}, \ldots, 2^{-n}}^{n-i}) \quad (i = 1, \ldots, n-1)$$

$$\text{lin}_n(A) = (2^{1-n}, 2^{2-n}, \ldots, 1)$$

Considere a linha n.

$$\sum_{k=1}^{n-1} \frac{1}{2^{n-k}} = \sum_{j=1}^{n-1} 2^{-j} = \frac{\cancel{\frac{1}{2}}\left(1 - \left(\frac{1}{2}\right)^{n-1}\right)}{\cancel{1 - 1/2}} = 1 - \frac{1}{2^{n-1}} < 1,$$

$\forall n \in \mathbb{N}$, pois $2^{1-n} > 0$.

Isso implica que a matriz é irredutivelmente diagonalmente dominante. Além disso, para $j = 1, \ldots, n-1$,

$$\sum_{k=1}^{j-1} 2^{k-j} + \sum_{k=j+1}^{n} 2^{-k} = \sum_{i=1}^{j-1} 2^{-i} + \sum_{i=j+1}^{n} 2^{-i}$$
$$= \sum_{i=1}^{n} 2^{-i} - 2^{-j}$$
$$= \frac{\cancel{\tfrac{1}{2}}\left(1-\left(\tfrac{1}{2}\right)^n\right)}{\cancel{1-1/2}} - 2^{-j}$$
$$= 1 - 2^{-n} - 2^{-j} < 1,$$

o que implica A ser estritamente diagonalmente dominante. Isso mostra que $\forall n \in \mathbb{N}$, ambos os métodos são convergentes.

(c) Demonstre que para qualquer norma induzida, o método iterativo $x^{(m+1)} = Cx^{(m)} + d$ satisfaz

$$\left\|x^{(m+1)} - x^{(m)}\right\| \leq \|C\| \left\|x^{(m)} - x^{(m-1)}\right\|, \quad \forall m \in \mathbb{N}$$

(d) Determine m, em função de n, de modo que após m iterações do método de Jacobi, se obtenh uma aproximação a solução do sistema com uma precisão de 10^{-5}. Justifique.

c)

$$\|x^{(m+1)} - x^{(m)}\| \overset{\text{passo de iteração}}{=} \|Cx^{(m)} + d - Cx^{(m-1)} - d\|$$
$$\overset{\text{linearidade}}{=} \|C(x^{(m)} - x^{(m-1)})\|$$
$$\leq \|C\|\|x^{(m)} - x^{(m-1)}\|,$$

pois $\|C\| = \sup_{\|x\| \neq 0} \frac{\|Cx\|}{\|x\|} \Rightarrow \|C\| \geq \frac{\|Cx\|}{\|x\|}$, $\forall x \neq 0$ e, portanto,

$\forall x \in \mathbb{R}^n$, $\|C\|\|x\| \geq \|Cx\|$, em que $x = 0$ a desigualdade é trivial e torna-se uma igualdade.

d) Note que

$$\|x^* - x^{(m+1)}\| = \|Cx^* + d - Cx^{(m)} - d\|$$
$$\leq \|C\|\|x^* - x^{(m)}\|$$
$$= \|C\|\|x^* - x^{(m+1)} + x^{(m+1)} - x^{(m)}\|$$
$$\leq \|C\|\|x^* - x^{(m+1)}\| + \|C\|\|x^{(m+1)} - x^{(m)}\|$$
$$\Rightarrow \|x^* - x^{(m+1)}\| \leq \frac{\|C\|}{1 - \|C\|} \|x^{(m+1)} - x^{(m)}\|$$

No nosso caso $\|C\| < 1$, pois o método é convergente, isto é, existe $\|\cdot\|$ induzida com $\|C\| < 1$.

Como $\|x^{(m+1)} - x^{(m)}\| \leq \|C\| \, \|x^{(m)} - x^{(m-1)}\|$, que por indução implica

$$\|x^{(m+1)} - x^{(m)}\| \leq \|C\|^m \|x^{(1)} - x^{(0)}\|.$$

Logo

$$\|x^* - x^{(m+1)}\| \leq \frac{\|C\|^{m+1}}{1 - \|C\|} \|x^{(1)} - x^{(0)}\|.$$

Para $\|x^* - x^{(m)}\| \leq 10^{-5}$, é suficiente que

$$\frac{\|C\|^m}{1 - \|C\|} \leq 10^{-5}$$

$$\Rightarrow m \log \|C\| - \log(1 - \|C\|) \leq -5$$

$$\Rightarrow m \geq \frac{\log(1 - \|C\|) - 5}{\log \|C\|}$$

Falta calcular $\|C\|$, por fim.

No método de Jacobi, $C = -D^{-1}(L + U)$. Eu lembro que $D^{-1} = I$, pois $D_{jj} = 1/a_{jj} = 1$. Logo $C = -(L + U)$. Note que

$$\begin{aligned}
\|C\|_\infty &= \|L + U\|_\infty \\
&= \max_{1 \leq j \leq n} \sum_{k=1}^{n} |c_{jk}| \\
&= \max_{1 \leq j \leq n} \sum_{k=1}^{j-1} \frac{1}{2^{j-k}} + \sum_{k=j+1}^{n} \frac{1}{2^j} \\
&= \max_{1 \leq j \leq n} \left(\sum_{i=1}^{n} \frac{1}{2^i} - \frac{1}{2^j} \right) \\
&= \max \left(1 - \frac{1}{2^n} - \frac{1}{2^j} \right) \\
&= 1 - \frac{1}{2^n} - \min 1/2^j \\
&= 1 - 2^{1-n}
\end{aligned}$$

Assim $\quad m \geqslant \dfrac{\log(1-\|C\|_\infty) - 5}{\log(\|C\|_\infty)}$

$$= \frac{(1-n)\log 2 - 5}{\log(1-2^{1-n})}$$

$$= \frac{(n-1)\log 2 + 5}{\log(1/(1-2^{1-n}))}.$$