

Escreva o seu nome e RA na primeira folha. Numere todas as folhas. Começa cada questão numa nova folha. Pode usar a calculadora. Cada questão (1), (2), (3) tem peso 1/3 na nota da prova.

1. Se deseja determinar a solução da equação $x^2 - \sin(x) - 1 = 0$

- O problema é bem posto? Quantas soluções o problema tem?
- Descreva e implementa dois passos do método da falsa posição para achar a solução maior da equação.
- Descreva um método do ponto fixo diferente daquele de Newton para determinar a solução maior da equação.
- Dizer se (ou quando) o método do ponto fixo proposto no item anterior pode convergir, motivando a sua resposta.

Resolução:

- Para que o problema seja considerado bem posto devemos assegurar três condições: existência, unicidade de solução e que pequenos erros nos dados de entrada impliquem me pequenas variações na solução.

Do gráfico da função x^2 em azul e da função $\sin(x) + 1$ em vermelho notamos que somente no intervalo $[-\pi/2, -\pi/2]$ as duas curvas se intersecam : $x^2 = \sin(x) + 1$ em dois pontos, ou seja **existem duas soluções da equação $x^2 - \sin(x) - 1 = 0$ em $\mathbb{R}$, ver Figura 1. Então o problema é malposto por causa da falta de unicidade da solução.**

Se quisermos analisar como varia a solução perturbando de pouco poderemos analisar quais zeros vai ter a $f + \delta$ (dado de input perturbado). Ou seja resolver o problema $f(x) + \delta = 0$ que equivale a encontrar os x que determinam a interseção de $x^2 = \sin(x) + 1 + \delta$. Do gráfico abaixo usando um $\delta = 0.1$ notamos que a x da interseção varia de pouco e de menos de 0.1 então temos que pequenas variações dos dados levam a pequenas variações da solução, ver Figura 2

Existe uma única solução entre 1 e 2, portanto poderia também dizer que **o problema de encontrar a solução de $f(x) = 0$ em $[1, 2]$ é bem posto.**

- No intervalo $[1, 2]$ temos uma única raiz, que será maior das duas. Porque vimos isto graficamente, mas se pode ver também matematicamente

$$f(1) \approx -0.8415$$

$$f(2) \approx 2.0907$$

E sendo $f'(x) = 2x - \cos(x) > 0$ porque se $x \in [1, 2]$ então $x > \frac{\cos(x)}{2}$. Isso acontece porque o maximo de $\frac{\cos(x)}{2}$ é $\frac{1}{2}$. Logo f' é constante em sinal e $f(1) * f(2) < 0$ então pelo teorema de localização dos zeros existe um único zero em $[1, 2]$.

Logo o zero é único em $[1, 2]$. usando o critério de paragem $|b_k - a_k| < \epsilon$ com $\epsilon = 10^{-5}$

$$[a_0, b_0] = [1, 2] \text{ com } f(a_0)f(b_0) < 0, |b_0 - a_0| < \epsilon$$

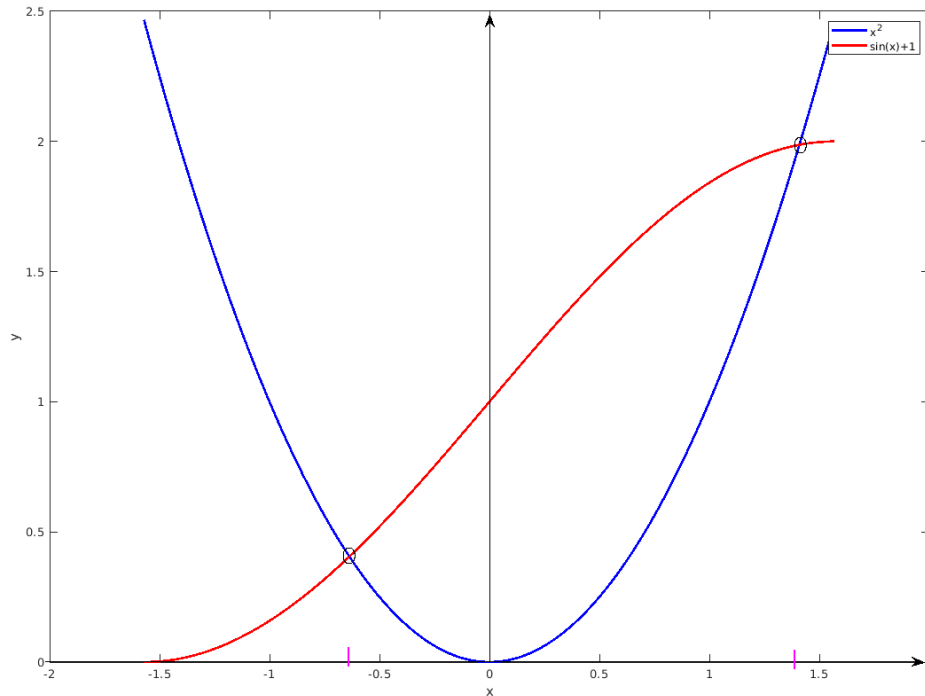


Figura 1:

$$x_0 = \frac{a_0 f(b_0) - b_0 f(a_0)}{f(b_0) - f(a_0)} = 1.287$$

$$f(x_0) = -0.3037 < 0$$

$$\Rightarrow z \in (x_0, b_0), \quad a_1 = x_0 = 1.287$$

$$x_1 = \frac{a_1 f(b_1) - b_1 f(a_1)}{f(b_1) - f(a_1)} = 1.3774$$

- No método do ponto fixo reescrevemos equação $f(x) = 0$ na forma $x = g(x)$ com $g(x)$ dita função do ponto fixo. Note que assim o ponto fixo de g será o zero de f .

Determinamos no nosso caso a $g(x)$ função do ponto fixo para que o método iterativo do ponto fixo) $x_{k+1} = g(x_k)$ possa convergir a raiz maior de $f(x)$:

$$\square \quad f(x) = 0 \Rightarrow x^2 = \sin(x) + 1 \text{ iff } x = \pm \sqrt{\sin(x) + 1}$$

Note que sempre $\sin(x) + 1 \geq 0$ então $\sqrt{\sin(x) + 1}$ é bem definida.

- Sendo que queremos achar a raiz maior que é positiva temos de escolher como função do ponto fixo a

$$g(x) = +\sqrt{\sin(x) + 1}$$

que é sempre positiva e que via gerar portanto $x_{k+1} = g(x_k) > 0$ positiva.

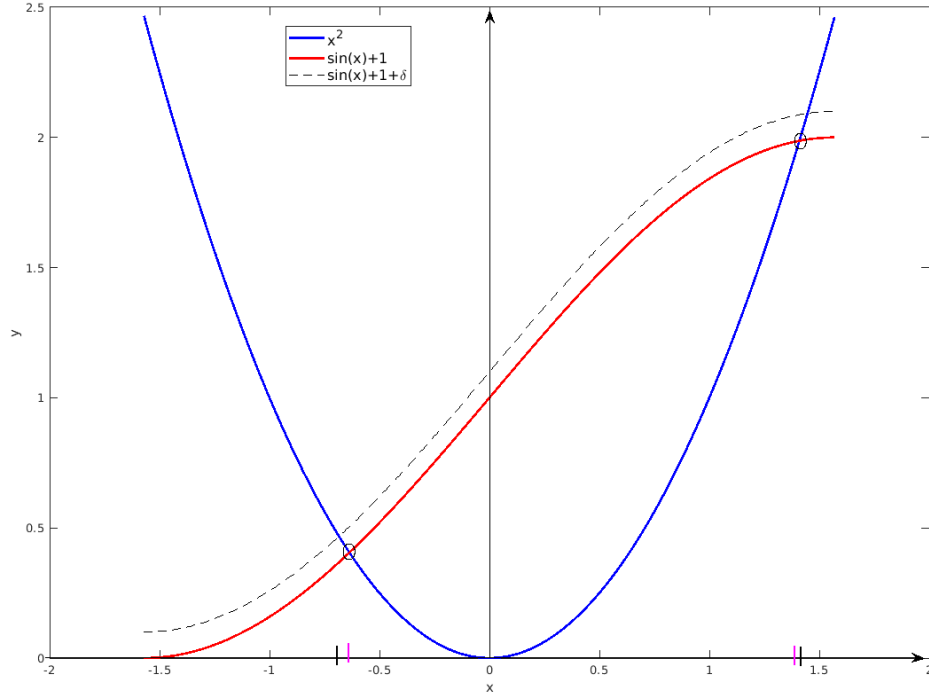


Figura 2: Solução com f perturbada. Interseção de x^2 com $\sin(x) + 1$ perturbada. Se obtem a mesma abscissa x da interseção se perturbasse somente x^2 .

- Analizamos a convergencia possivel deste método do ponto fixo $x_{k+1} = g(x_k) = \sqrt{\sin(x) + 1}$. A região (intervalo) I onde poder pegar x_0 tem de ser tal que g é derivavel e com derivada continua (ok: este acontece sempre , porque $\sin(x) + 1 > 0$). E temos de verificar quando $|g'(x)| < 1$. Note que $|g'(x)| = \left| \frac{\cos(x)}{2\sqrt{\sin(x)+1}} \right|$ portanto

$$|g'(x)| < 1 \iff 2\sqrt{\sin(x) + 1} > \cos(x) \iff \sin(x) + 1 > \frac{\cos(x)^2}{4} \iff \sin(x) > \frac{\cos(x)^2}{4} - 1$$

Note que temos sempre (para cada x): $\frac{\cos(x)^2}{4} - 1 < 0$ porque $0 \leq \frac{\cos(x)^2}{4} \leq \frac{1}{4}$. Então a condição de convergencia $|g'(x)| < 1$ vale para qualquer x . Então podemos pegar um qualquer intervalo $I \subset \mathbb{R}_+$ para poder determinar a raiz positiva de f partindo de $x_0 \in I$ e aplicando o método do ponto fixo $x_{k+1} = \sqrt{\sin(x) + 1}$.

2. Considere o seguinte sistema linear retangular

$$\begin{cases} 3x - 10^2y + 3z + w = -10 \\ 10^5x - 2y + z - 2w = 3 \\ x + y + z + w = 0 \end{cases}$$

- O sistema é consistente? Se sim quantas soluções tem?
- Fixe o parâmetro $w = 1$, prove que existirão únicos (x, y, z) que satisfazem o sistema com $w = 1$.
- Escreva o sistema que se obtém do item anterior. Encontre a solução (x, y, z) deste sistema usando o método de eliminação de Gauss com pivotamento parcial na aritmética FP(10,4,-99,99,A).
Pelo menos faça o primeiro passo do método, e explique o que faltaria para encontrar a solução.
- Porque é esperado que o método com pivotamento parcial fornece resultados melhores que o método de eliminação clássico para resolver o sistema do item anterior?

Resolução

- Para que o sistema dado $\bar{A}x = \bar{b}$ seja consistente ele deve admitir ao menos uma solução, e este acontece se $\text{posto}(\bar{A}|\bar{b}) = \text{posto}(\bar{A})$.
Neste caso temos um sistema 3×4 com mais incógnitas ($n = 4$) que equações ($m = 3$), a matriz

$$\bar{A} = \begin{bmatrix} 3 & -10^2 & 3 & 1 \\ 10^5 & -2 & 1 & -2 \\ 1 & 1 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{3 \times 4}$$

tem $\text{posto} = 3 = m$. Isso porque as primeiras três colunas são linearmente independentes : note que

$$\det \begin{bmatrix} 3 & -10^2 & 3 \\ 10^5 & -2 & 1 \\ 1 & 1 & 1 \end{bmatrix} > 0. \quad (1)$$

ouseja $\text{posto}(\bar{A}) = 3$, porque 3 é o numero máximo de colunas l.i. de $\bar{A}$.

Então sabendo pela teoria que $\text{posto}(\bar{A}|\bar{b}) \leq \min\{m, n + 1\} = \min 3, 5 = 3$ então o maior numero de colunas l.i. de $(\bar{A}|\bar{b})$ é necessariamente 3, porque as primeiras três colunas de $(\bar{A}|\bar{b})$ são as três colunas l.i. de $\bar{A}$ então temos por definição (de posto) $\text{posto}(\bar{A}|\bar{b}) = 3$.

Logo **o sistema é consistente** $\text{posto}(\bar{A}|\bar{b}) = \text{posto}(\bar{A}) = 3 \equiv p$ e **tem infinitas soluções em particular são** $\infty^{n-p} = \infty^{4-3} = \infty^1$ **soluções, temos um grau de liberdade no sistema.**

Em particular temos uma solução associada a cada parâmetro (ou valor) w inserido no sistema.

- Seja agora $w = 1$ e vamos provar que existirão únicos (x, y, z) que satisfazem o sistema com $w = 1$.

$$\begin{cases} 3x - 10^2y + 3z = -11 \\ 10^5x - 2y + z = 5 \\ x + y + z = -1 \end{cases}$$

O sistema é do tipo $Ax = b$ com $A = \begin{bmatrix} 3 & -10^2 & 3 \\ 10^5 & -2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ e $b = \begin{bmatrix} -11 \\ 5 \\ -1 \end{bmatrix}$ sendo que

foi provado na (1), $\det A \neq 0$, então a matriz A é não singular e portanto as suas linhas são l.i. então o sistema admite uma única solução (x, y, z) . Note que $\text{posto}(A|b) = \text{posto}(A) = 3$.

- Observamos que $\det A \neq 0$ e isso vale para todas as matrizes de cabeça de A então se pode aplica o metodo de eliminação de Gauss. Aplicaremos em particular o método de eliminação de Gauss com pivotamento na artitmetica finita $FP(10, 4, -99, 99)$.

O números de passos é $n - 1 = 2$ passos.

$$(A^{(0)}|b^{(0)}) = \left[\begin{array}{ccc|c} 3 & -10^2 & 3 & -11 \\ 10^5 & -2 & 1 & 5 \\ 1 & 1 & 1 & -1 \end{array} \right]$$

Notamos que $|a_{21}| = 10^5 > 3 = |a_{11}|$ então pelo pivotamento parcial temos trocar a linha 2 pela 1 e obtemos:

$$(\bar{A}^{(0)}|\bar{b}^{(0)}) = \left[\begin{array}{ccc|c} 10^5 & -2 & 1 & 5 \\ 3 & -10^2 & 3 & -11 \\ 1 & 1 & 1 & -1 \end{array} \right]$$

Aplicamos o passo 1 nesta ultima matriz nas linhas 1 e 2.

$$m_{21}^{(0)} = \frac{a_{21}^{(0)}}{a_{11}^{(0)}} = \frac{3}{10^5} = 3 \times 10^{-5} = 0.3 \times 10^{-4}, \quad m_{31}^{(0)} = \frac{a_{31}^{(0)}}{a_{11}^{(0)}} = \frac{1}{10^5} = 0.1 \times 10^{-4}.$$

Usamos m_{21} para zerar a linha 2 de $A^{(1)}$ e b^1 , usaremos o alinhamento

$$a_{22}^{(1)} = a_{22}^{(0)} - m_{21}a_{12}^{(0)} = -0.1 \times 10^3 - 0.3 \times 10^{-4}(-0.2 \times 10) = -0.1 \times 10^3 + 0.6 \times 10^{-4} \rightarrow$$

Usando o alinhamento

$$\dots a_{22}^{(1)} = (-0.1 + 0.6 \times 10^{-7})10^3 = (-0.099999994)10^3 \approx -0.110^3.$$

$$\begin{aligned} a_{23}^{(1)} &= a_{23}^{(0)} - m_{21}a_{13}^{(0)} = 0.3 \times 10 - 0.3 \times 10^{-4} = \\ &= (0.3 - 0.310^{-5})10 = (0.3 - 0.000003)10 = 0.299997 \times 10 \approx 0.3 \times 10. \end{aligned}$$

$$b_2^{(1)} = b_2^{(0)} - m_{21}b_1^{(0)} = -0.11 \times 10^2 - 0.3 \times 10^{-4} \cdot (0.5) = (-0.11 - 0.15 \times 10^{-5})10^2 \approx -0.11 \times 10^2$$

Usamos m_{31} para gerar a linha 3 de A^1 e b^1

$$a_{32}^{(1)} = a_{32}^{(0)} - m_{31}a_{12}^{(0)} = 0.1 \times 10 - 0.1 \times 10^{-4}(-0.2 \times 10) = (0.1 - 0.2 \times 10^{-5})10 \approx 0.1 \times 10.$$

$$a_{33}^{(1)} = a_{33}^{(0)} - m_{31}a_{13}^{(0)} = 0.1 \times 10 - 0.1 \times 10^{-4} = (0.1 - 0.1 \times 10^{-5})10 \approx 0.110$$

$$b_3^{(1)} = b_3^{(0)} - m_{31}b_1^{(0)} = -0.1 \times 10 - 0.1 \times 10^{-4}(0.5 \times 10) = (-0.1 - 0.5 \times 10^{-5})10 =$$

-0.1×10 Depois de um passo obtemos a matriz:

$$(A^{(1)}|b^{(1)}) = \left[\begin{array}{ccc|c} 0.1 \times 10^6 & -0.2 \times 10 & 0.1 \times 10 & 0.5 \times 10 \\ 0 & -0.1 \times 10^3 & 0.3 \times 10 & -0.11 \times 10^2 \\ 0 & 0.1 \times 10 & 0.1 \times 10 & -0.1 \times 10 \end{array} \right] = \left[\begin{array}{ccc|c} 10^5 & -2 & 1 & 5 \\ 0 & -10^2 & 3 & -11 \\ 0 & 1 & 1 & -1 \end{array} \right]$$

O método tem de prosseguir verificando o pivotamento : se $|a_{32}^{(1)}| > |a_{22}^{(1)}|$ então efetuamos a troca da terceira com a segunda linha. No nosso caso isso não acontece , então se prossegue eliminando o elemento $a_{32}^{(1)}$ através da operação $linha_3 \text{ nova} = linha_3 - m_{32}linha_2$ na FP.

Assim vamos obter um sistema triangular superior $A^{(3)}x = b^{(3)}$ que resolvemos sempre na FP usando o algoritmo de substituição partindo da ultima equação, obtendo assim antes $x_3 = z$, depois $x_2 = y$ e em fim $x_1 = x$.

- O pivotamento reduz a possibilidade de divisões por números pequenos, que poderia acontecer no calculo dos multiplicadores $m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$ a cada passo. O pivotamento permite de ter um multiplicador limitado , $|m_{ik}| < 1$. Este leva a uma diminuição importante dos erros de arredondamento, gerando um método mais estável que a eliminação de Gauss clássica.

No nosso caso especifico este problema de obter multiplicadores grandes aconteceria logo no calculo de m_{21} senão tivesse feito a troca. Em geral o pivotamento parcial ou total tem de ser aplicado quando o sistema apresenta coeficientes muito diferentes de dimensão. Como acontece no nosso caso onde temos uma diferença da ordem de 10^5 graus entre o menor e o maior coeficiente.

3. Considere o sistema linear $Ax = b$

- Determine e escreva o método iterativo de Gauss-Seidel na forma escalar.
- Quais vantagens tem este método sobre o método iterativo de Jacobi?
- Dados

$$A = \begin{bmatrix} 4 & -2 & 1 \\ -1 & 3 & -2 \\ -3 & 2 & 6 \end{bmatrix} ; b = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} ,$$

verifique se o método de Gauss-Seidel pode convergir a solução do sistema $Ax = b$.

- Faça duas iteradas do método de Gauss-Seidel partindo de $x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$.

Arredonda os resultados com 4 dígitos significativos.

Resolução:

- Deduzimos o método de Gauss-Seidel, partindo de um sistema generico quadrado $Ax = b$ de dimensão n . Usamos o splitting da matriz $A = D + L + U$, onde D é a diagonal da matriz, L é parte triangular inferior de A e U é a restante parte triangular superior. Se supomos que $D + L$ é invertível ou seja se todos os a_{ii} foram diferentes de zero podemos determinar como segue o método, gerando a função do ponto fixo:

$$\begin{aligned}
Ax = b &\iff (D+L+U)x = b \iff (D+L)x = b - Ux \iff x = -(D+L)^{-1}Ux + (D+L)^{-1}b \\
&\iff x = (I - (D+L)^{-1}A)x + (D+L)^{-1}b = G_{GS}x + b_{GS}
\end{aligned}$$

Daí obtemos o método iterativo (do ponto fixo) de Gauss -Seidel

$$x^{(k+1)} = G_{GS}x^{(k)} + b_{GS} \quad (1)$$

Note que $G_{GS} = I - (D+L)^{-1}A = -(D+L)^{-1}U$ e $b_{GS} = (D+L)^{-1}b$ Então da forma vetorial (1) obtemos multiplicando por $(D+L)$:

$$(D+L)x^{(k+1)} = -Ux^{(k)} + b \rightarrow Dx^{(k+1)} = b - Lx^{(k+1)} - Ux^{(k)}$$

então agora olhando componente por componente

$$(Dx^{(k+1)})_i = (b - Lx^{(k+1)} - Ux^{(k)})_i \rightarrow a_{ii}x_i = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)}$$

Assim dividindo por a_{ii} deduzimos a forma escalar do método de Gauss Seidel

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right) \quad (2)$$

Logo temos,

$$\begin{aligned}
x_1^{(k+1)} &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^{(k)} \dots - a_{1n}x_n^{(k)}) \\
x_2^{(k+1)} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} \dots - a_{2n}x_n^{(k)}) \\
&\dots \\
x_n^{(k+1)} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} \dots - a_{nn-1}x_{n-1}^{(k+1)})
\end{aligned}$$

Descrevendo passo a passo:

- 1- Obtemos antes $x_1^{(k+1)}$ usando os $x_i^{(k)}$ com $i = 1, \dots, n$
 - 2- Obtemos $x_2^{(k+1)}$ usando os $x_1^{(k+1)}$ com $i = 2, \dots, n$
 - ...
 - i- Obtemos $x_i^{(k+1)}$ usando os $x_j^{(k+1)}$ anteriores com $j = 1, \dots, i-1$ e os $x_j^{(k)}$ com $j = i+1, \dots, n$
 - ...
 - n- Computamos $x_n^{(k+1)}$ usando os $x_j^{(k+1)}$ com $j = 1, \dots, n-1$
- Na maior parte dos casos quando Jacobi converge e também Gauss Seidel temos que Gauss-Seidel tem uma redução mior do erro respeito ao método de Jacobi avançando nas iterações. Este acontece por exemplo quando a matriz A satisfaz o critério forte das linhas. O método de Gauss Seidel utiliza um array a menos no seu algoritmo respeito Jacobi.

- Pelo critério de convergência de Sassenfeld verificaremos a convergência de Gauss-Seidel

$$\beta_1 = \frac{|a_{12}| + |a_{13}|}{|a_{11}|} = \frac{2 + 1}{4} = \frac{3}{4} = 0.75 < 1$$

$$\beta_2 = \frac{|a_{21}|\beta_1 + |a_{23}|}{|a_{22}|} = \frac{\frac{3}{4} + 2}{3} = \frac{11}{12} \approx 0.9167 < 1$$

$$\beta_3 = \frac{|a_{31}|\beta_1 + |a_{32}|\beta_2}{|a_{33}|} = \frac{49}{72} \approx 0.6806 < 1$$

Logo todos os β_i são menores de 1 então o método converge.

Note também que a matriz A satisfaz o critério fraco das linhas $|a_{11}| = 4 > \sum_{j=2}^3 |a_{1j}| = 3$, $|a_{22}| = 3 = |a_{21}| + |a_{23}| = 3$, $|a_{33}| = 6 > \sum_{j=1}^2 |a_{3j}| = 3$ e tem todos os coeficientes não nulos, e este pela teoria é suficiente para garantir a convergência do método.

•

$$\begin{cases} 4x_1 - 2x_2 + x_3 = 1 \\ -x_1 + 3x_2 - 2x_3 = 1 \\ -3x_1 + 2x_2 + 6x_3 = 0 \end{cases}$$

Usando $x^0 = (0, 0, 0)^t$

$$x_1^{(k+1)} = \frac{1}{a_{11}}(b_1 - a_{12}x_2^{(k)} - \dots - a_{1n}x_n^{(k)}) = \frac{1}{4}(1 - (-2)x_2^k - x_3^k) = \frac{1}{4}(1 + 2x_2^k - x_3^k)$$

$$x_2^{(k+1)} = \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}) = \frac{1}{3}(1 - (-1)x_1^{(k+1)} - (-2)x_3^{(k)}) = \frac{1}{3}(1 + x_1^{(k+1)} + 2x_3^{(k)})$$

$$x_3^{(k+1)} = \frac{1}{6}(0 - (-3)x_1^{(k+1)} - 2x_2^{(k+1)}) = \frac{1}{6}(3x_1^{(k+1)} - 2x_2^{(k+1)})$$

(1a iteração) Para $k = 0$ temos arredondando com quatro dígitos significativos

$$x_1^{(1)} = \frac{1}{4}(1 + 2x_2^{(0)} - x_3^{(0)}) = \frac{1}{4}(1 + 0 + 0) = 0.25$$

$$x_2^{(1)} = \frac{1}{3}(1 + x_1^{(1)} + 2x_3^{(0)}) = \frac{1}{3}(1 + 0.25 + 0) = \frac{5}{12} \approx 0.4167$$

$$x_3^{(1)} = \frac{1}{6}(3x_1^{(1)} - 2x_2^{(1)}) = \frac{1}{6}(3 * 0.25 - 2 * 0.4167) \approx -0.0139$$

(2a iteração) Para $k = 1$

$$x_1^{(2)} = \frac{1}{4}(1 + 2x_2^{(1)} - x_3^{(1)}) = \frac{1}{4}(1 + 2 * 0.4167 + 0.0139) = 0.4618$$

$$x_2^{(2)} = \frac{1}{3}(1 + x_1^{(2)} + 2x_3^{(1)}) = \frac{1}{3}(1 + 0.4618 + 2 * 0.0139) = 0.478$$

$$x_3^{(2)} = \frac{1}{6}(3x_1^{(2)} - 2x_2^{(2)}) = \frac{1}{6}(3 * 0.4618 - 2 * 0.478) = 0.07157$$