MAP 2320 - Métodos Numéricos em EDP's
Equações Elípticas

Marco Alexandre Claudino*
IME/USP

1 A equação de Laplace no quadrado unitário

Vamos iniciar o estudo dos métodos de resolução numérica das equações elípticas por meio da chamada **Equação de Poisson**:

$$-\Delta u = u_{xx} + u_{yy} = f(x,y) \quad (x,y) \in \Omega \tag{1}$$

onde $f : \Omega \mapsto \mathbb{R}$ é uma função contínua e para simplificar a apresentação, consideraremos $\Omega = (0,1) \times (0,1)$. Para que o problema possua uma única solução é necessário que sejam fornecidas condições de contorno para os trechos $[t,0]$, $[1,t]$, $[t,1]$ e $[0,t]$ com $t \in [0,1]$ (isto é, as fronteiras do quadrado). Temos três tipos de condições possíveis:

1) **Condições de Dirichlet:** São fornecidos valores para a função na fronteira.

2) **Condições de Neumann:** São fornecidas as derivadas normais em cada ponto da fronteira.

3) **Condições de Robin:** São fornecidas ponderações entre as derivadas normais e valores da função na fronteira.

Para facilitar o estudo, trabalharemos apenas com condições de Dirichlet no contorno. Logo, será dada também a condição

$$u(x,y) = g(x,y) \quad (x,y) \in \partial\Omega \tag{2}$$

Observe que a solução $u(x,y)$ deve ser tal que $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$, ou seja, de classe $\mathcal{C}^2$ no interior de Ω e contínua até o bordo.

2 Discretização do problema

Dado N um número natural, considere a discretização do quadrado unitário em espaçamentos uniformes dados por $h = \frac{1}{N}$. Denotamos por $u_{ij} \approx u(x_i, y_j)$ a aproximação da função u e no ponto $x_i = ih$, $y_j = ih$ para $i,j = 1,2,...,N-1$. Para os índices igual a 0 ou igual a N a condição de contorno do problema é fornecida.

Para obter a discretização da segunda derivada em x e em y vamos utilizar a combinação das Séries de Taylor avançada e retrógrada:

$$u(x_{i+1},y_j) = u(x_i,y_j) + hu_x(x_i,y_j) + \frac{h^2}{2}u_{xx}(x_i,y_j) + \frac{h^3}{6}u_{xxx}(x_i,y_j) + \frac{h^4}{24}u_{xxxx}(x_i,y_j) + \cdots \tag{3}$$

$$u(x_{i-1},y_j) = u(x_i,y_j) - hu_x(x_i,y_j) + \frac{h^2}{2}u_{xx}(x_i,y_j) - \frac{h^3}{6}u_{xxx}(x_i,y_j) + \frac{h^4}{24}u_{xxxx}(x_i,y_j) + \cdots \tag{4}$$

*claudino@ime.usp.br

Somando as equações (3) e (4) temos que

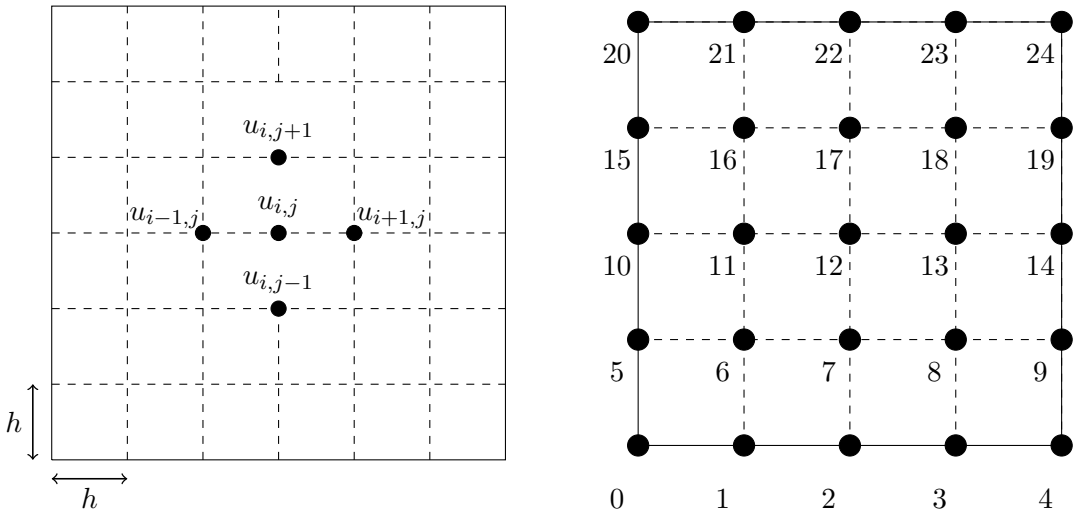
$$u_{xx} = \frac{u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)}{h^2} + \tau_x(x_i, y_j) \tag{5}$$

onde $\tau_x(x_i, y_j) = \mathcal{O}(h^2)$ é o erro de discretização local cometido por esta aproximação no ponto (x_i, y_j). Analogamente, obtemos que a segunda derivada em relação a y é dada por:

$$u_{yy} = \frac{u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})}{h^2} + \tau_y((x_i, y_j) \tag{6}$$

Substituindo (5) e (6) na equação (1) temos que

$$\frac{-u(x_{i-1}, y_j) - u(x_i, y_{j-1}) + 4u(x_i, y_j) - u(x_{i+1}, y_j) - u(x_i, y_{j+1})}{h^2} = f(x_i, y_j) + \underbrace{\tau_x(x_i, y_j) + \tau_y(x_i, y_j)}_{\mathcal{T}_h = \mathcal{O}(h^2)} \tag{7}$$



(a) Malha do problema

(b) Numeração Lexicográfica

Figura 1: Descrição da Malha do problema.

Ignorando o erro de discretização local, temos que as aproximações nos pontos interiores satisfazem a seguinte relação

$$\frac{-u_{i-1,j} - u_{i,j-1} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1}}{h^2} = f(x_i, y_i)$$

Podemos ainda reescrever esta equação como um sistema linear contendo $(N-1)$ equações e $(N-1)$ incógnitas, dado por

$$-u_{i-1,j} - u_{i,j-1} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1} = h^2 f(x_i, y_i) \tag{8}$$

de forma que, utilizando a enumeração lexicográfica (veja na Figura 2(b)), temos que as aproximações são obtidas através da resolução do sistema linear

$$\begin{bmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{bmatrix} \begin{pmatrix} u_6 \\ u_7 \\ u_8 \\ u_{11} \\ u_{12} \\ u_{13} \\ u_{16} \\ u_{17} \\ u_{18} \end{pmatrix} = \begin{pmatrix} h^2 f_6 + u_1 + u_5 \\ h^2 f_7 + u_2 \\ h^2 f_8 + u_3 + u_9 \\ h^2 f_{11} + u_{10} \\ h^2 f_{12} \\ h^2 f_{13} + u_{14} \\ h^2 f_{16} + u_{15} + u_{21} \\ h^2 f_{17} + u_{22} \\ h^2 f_{18} + u_{19} + u_{23} \end{pmatrix} \tag{9}$$

Observe que cada elemento da diagonal relaciona-se a, no máximo, outros quatro elementos. Naturalmente, as questões a serem abordadas a partir de agora são:

- Este sistema sempre possui solução?
- As aproximações obtidas convergem para a solução do problema quando $h \to 0$?
- Qual a maneira mais eficiente de resolver o sistema associado?

3 Existência de Solução do Sistema Linear

3.1 Demonstração da existência via álgebra linear

Considere as seguintes definições:

Definição 3.1. *Uma matriz $A = [a_{i,j}]_{n\times n}$ é dita **irredutível** se não existir uma matriz de permutação P tal que*

$$PAP^T = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$$

onde $k < n$, $B = [b_{i,j}]_{k\times k}$, $C = [c_{i,j}]_{k\times(n-k)}$ e $D = [d_{i,j}]_{(n-k)\times(n-k)}$. Ou seja, a matriz não pode ser decomposta de forma que um conjunto de variáveis fique independente das outras.

Definição 3.2. *Seja n um inteiro. Uma matriz $A = [a_{i,j}]_{n\times n}$ é dita **diagonal dominante** se*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{i,j}|, \quad i = 1, 2, \cdots, n$$

*e é dita **fracamente diagonal dominante** se*

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{i,j}|, \quad i = 1, 2, \cdots, n$$

e vale a desigualdade estrita para alguma linha.

Observe no exemplo (9) que os pontos próximos da fronteira possuem a desigualdade estrita enquanto que o ponto interior u_{12} possui a entrada da diagonal igual a soma dos outros elementos da linha 12. Além disso, é possível mostrar que a matriz do sistema é irredutível, de forma que o resultado a seguir garante a existência e unicidade de solução para os sistemas em estudo:

Teorema 3.3 (Existência de Solução do sistema)**.** *Toda matriz diagonal dominante é inversivel e toda matriz fracamente diagonal dominante e irredutível é inversível.*

Demonstração. Seja A uma matriz diagonal dominante e suponha, por absurdo, que $det(A) = 0$. Então existe um vetor $x \neq \vec{0}$ tal que $Ax = \vec{0}$. Seja i o índice tal que

$$|x_i| = M = \max_j |x_j|$$

Temos que a i-ésima linha do sistema é dada por

$$\sum_{j=1}^{n} a_{i,j}x_j = 0 \Rightarrow a_{ii}x_i = -\sum_{j\neq i} a_{i,j}x_j \Rightarrow x_i = -\sum_{j\neq i} \frac{a_{i,j}}{a_{i,i}}x_j$$

Tomando o módulo dos dois lados, temos que

$$|x_i| = \Big|\sum_{j\neq i} \frac{a_{i,j}}{a_{i,i}}x_j\Big| \leq \sum_{j\neq i} \Big|\frac{a_{i,j}}{a_{i,i}}\Big||x_j| \leq \underbrace{\sum_{j\neq i} \Big|\frac{a_{i,j}}{a_{i,i}}\Big|}_{<1} M < M \quad \text{(contradição!)}$$

Logo $\det(A) \neq 0$. No caso em que a matriz é fracamente diagonal dominante, note que

$$\sum_{j\neq i} \frac{|a_{i,j}|}{|a_{i,i}|}|x_j| = M$$

ocorre se, e somente se

$$|x_j| = M \quad \text{quando } a_{i,j} \neq 0$$

Como a matriz é irredutível, podemos reordenar os índices onde $|x_j| = M$ de forma a colocá-los em sequência crescente. Como existe pelo menos uma linha onde a desigualdade é estrita, nesta linha caimos novamente no caso anterior. ■

3.2 Demonstração da existência via princípio do máximo discreto

Para que o sistema (8) possua uma única solução é necessário e suficiente que o problema homogêneo associado admita apenas o vetor nulo como solução. Isto implica dizer que o problema

$$\begin{cases} \Delta_h w_h(x,y) = 0, & (x,y) \in \Omega \\ w_h(x,y) = 0, & (x,y) \in \Gamma \end{cases} \tag{10}$$

possui como única solução $w(x,y) \equiv 0$.

Definição 3.4 (Funções harmônicas)**.** *Seja Ω um aberto contido no $\mathbb{R}^n$ e $w : \Omega \to \mathbb{R}$ uma função de classe $\mathcal{C}^2(\Omega)$ tal que $\Delta w = 0$. Então, dizemos que w é uma* ***função harmônica****.*

Sendo w uma função harmônica e w_h sua restrição à malha Ω_h, observe que

$$\Delta_h w_h(x_i,y_j) = 0 \quad (x_i,y_j) \in \Omega \Leftrightarrow w(x_i,y_j) = \frac{1}{4}\left(w(x_{i-1},y_j) + w(x_{i+1},y_j) + w(x_i,y_{j-1}) + w(x_i,y_{j+1})\right)$$

ou seja, $w(x_i,y_j)$ é a média aritmética dos quatro vizinhos. Os resultados a seguir garantem que o máximo (e o mínimo) de uma função harmônica são atingidos na fronteira.

Proposição 3.5 (Princípio do Máximo Discreto)**.** *Seja $u : \overline{\Omega}_h \to \mathbb{R}$ uma função de classe $\mathcal{C}^2(\Omega)$. Se $\Delta_h u \geq 0$ em Ω_h, então:*

$$\max_{(x,y)\in\Omega_h} u(x,y) = \max_{(x,y)\in\Gamma_h} u(x,y) \tag{11}$$

Demonstração. Suponha, por absurdo, que u assuma o valor máximo em um ponto $(x_i, y_j) \in \Omega_h$. Sendo $u_{i,j} = u(x_i, y_j)$ temos que

$$\begin{aligned}\Delta_h u(x_i, y_j) &= \frac{1}{4}\left(u_{i+1,j} + u_{i-1,j} - 4u_{i,j} + u_{i,j+1} + u_{i,j-1}\right) \\ &= \frac{1}{4}\left(\underbrace{(u_{i+i,j} - u_{i,j})}_{\leq 0} + \underbrace{(u_{i-1,j} - u_{i,j})}_{\leq 0} + \underbrace{(u_{i,j+1} - u_{i,j})}_{\leq 0} + \underbrace{(u_{i,j-1} - u_{i,j})}_{\leq 0}\right) \\ &\leq 0\end{aligned}$$

Por hipótese, temos que $\Delta_h u \geq 0$. Logo $\Delta_h u(x_i, y_i) = 0$ o que implica que u é constante em $\overline{\Omega}_h$. ■

Trocando u por $-u$ no resultado anterior, se $\Delta_h v \leq 0$ em Ω_h então

$$\min_{(x,y)\in\Omega_h} u(x,y) = \min_{(x,y)\in\Gamma_h} u(x,y) \tag{12}$$

Teorema 3.6. *Seja $u : \overline{\Omega}_h \to \mathbb{R}$ uma função de classe $\mathcal{C}^2(\Omega)$. Se $\Delta_h u = 0$ em Ω_h, então:*

$$\min_{(x,y)\in\Gamma_h} u(x,y) \leq u(\bar{x},\bar{y}) \leq \max_{(x,y)\in\Gamma_h} u(x,y), \quad \forall (\bar{x},\bar{y}) \in \Omega_h \tag{13}$$

Demonstração. O resultado é imediato a partir das igualdades (11) e (12). ■

Utilizando este resultado aplicado ao problema (10), temos que

$$w(x,y) = 0$$

para todo ponto $(x, y) \in \Omega_h$ e, portanto, o sistema possui solução única para todo valor de h.

4 Análise de convergência

Seja $e_h(x_i, y_j) = u(x_i, y_j) - u_{i,j}$ o erro cometido na aproximação de um dado ponto de $(x_i, y_j) \in \Omega_h$. Utilizando a linearidade do operador Δ_h temos que o erro satisfaz o problema

$$\begin{cases} \Delta_h e_h = \mathcal{T}_h \text{ em } \Omega_h \\ e_h = 0 \text{ em } \Gamma_h \end{cases} \tag{14}$$

onde $\mathcal{T}_h$ é o erro local de discretização cometido no ponto (x_i, y_j). Sabemos, pela construção da Série de Taylor, que $\mathcal{T}_h = \mathcal{O}(h^2)$, para qualquer $h \to 0$. Para garantirmos que $e_h \to 0$ quando $h \to 0$, considere o seguinte resultado:

Teorema 4.1. *Sejam $u : \overline{\Omega}_h \to \mathbb{R}$ uma função de classe $\mathcal{C}^2(\Omega)$ e $f : \Omega_h \to \mathbb{R}$ tais que*

$$\begin{cases} \Delta_h u = f \text{ em } \Omega_h \\ u = 0 \text{ em } \Gamma_h \end{cases}$$

Então, para todo $(x, y) \in \Omega_h$, temos que

$$|u(x,y)| \leq \frac{1}{8} \max_{(x,y)\in\Omega_h} |f(x,y)| \tag{15}$$

A aplicação desse resultado ao problema (14) equivale dizer que

$$|e_h(x,y)| \leq \frac{1}{8} \max_{(x,y)\in\Omega_h} |\mathcal{T}_h| \Rightarrow |e_h(x,y)| \leq Ch^2, \quad \forall (x,y) \in \Omega_h$$

Logo

$$|e_h(x,y)| \to 0 \text{ quando } h \to 0$$

concluindo assim a convergência do método em estudo.

Vamos a demonstração do Teorema (4.1):

Demonstração. Seja $||f||_\infty = \max_{\Omega_h} |f|$. Então, para todo ponto $(x, y) \in \Omega_h$, temos que

$$-||f||_\infty \leq f(x, y) \leq ||f||_\infty \tag{16}$$

Considere a função $w : \bar{\Omega} \to \mathbb{R}$ definida por

$$w(x, y) = \frac{1}{4}\left[\left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2\right]$$

Observe que o valor máximo assumido por w na fronteira é $\frac{1}{8}$ e w satisfaz

$$\Delta_h w = 1, \quad \forall (x, y) \in \Omega$$

pois $\frac{\partial^p w}{\partial x^p} = 0$ e $\frac{\partial^p w}{\partial y^p} = 0$ para todo $p > 2$ (w é um polinômio de ordem dois). Assim, o erro de discretização associado ao operador discreto é zero, pois ele é proporcional às derivadas de quarta ordem em x e y. Usando a desigualdade (16), temos que

$$-||f||_\infty \leq \Delta_h u \leq ||f||_\infty$$

e, portanto

$$\Delta_h u + ||f||_\infty \geq 0 \text{ em } \Omega_h$$

Como

$$||f||_\infty = ||f||_\infty \underbrace{\Delta_h w}_{=1} = \Delta_h(||f||_\infty w)$$

podemos escrever que

$$\Delta_h u + \Delta_h(||f||_\infty w) \geq 0 \Rightarrow \Delta_h(u + ||f||_\infty w) \geq 0 \text{ em } \Omega_h$$

Usando o princípio do máximo

$$u + ||f||_\infty w \leq \max_{\Gamma_h}(u + ||f||_\infty w) = \max_{\Gamma_h}(||f||_\infty w)$$

Como $w(x, y) \geq 0$, concluimos que

$$u \leq \max_{\Gamma_h}(||f||_\infty w) \Rightarrow u \leq \frac{1}{8}||f||_\infty \tag{17}$$

Por outro lado

$$\Delta_h u - ||f||_\infty \leq 0 \text{ em } \Omega_h \Rightarrow \Delta_h(u - ||f||_\infty w) \leq 0 \text{ em } \Omega_h$$

Usando o princípio do mínimo

$$u - ||f||_\infty w \geq \min_{\Gamma_h}(-||f||_\infty w) \Rightarrow u \geq -\max_{\Gamma_h}(||f||_\infty w)$$

$$u \geq -\frac{1}{8}||f||_\infty \tag{18}$$

Logo, de (17) e (18) temos que

$$-\frac{1}{8}||f||_\infty \leq u \leq \frac{1}{8}||f||_\infty$$

e, portanto

$$|u| \leq \frac{1}{8}||f||_\infty = \frac{1}{8}\max_{(x,y)\in\Omega_h} |f(x, y)|$$

■

5 Métodos Iterativos para a resolução de sistemas lineares

Antes de iniciarmos o estudo de métodos iterativos para a resolução de sistemas lineares, uma pergunta interessante a ser feita é: **Por que utilizar métodos iteraivos para resolver sistemas lineares?**

Sabemos que as aproximações do problema (1) sujeito à condição de contorno (2) serão obtidas através da resolução de um sistema linear cuja matriz possui $(N-1)^2$ elementos, onde apenas 5 deles são não nulos em cada uma das linhas. Utilizando o método de **Eliminação de Gauss seriam necessárias em torno de** $(N-1)^6$ **operações** para construir as aproximações do problema em estudo.

Para exemplificar o tempo de processamento que seria gasto, considere uma malha com $N = 100$ (ou seja, $h = 0.01$) e que todas as operações realizadas demorem o mesmo tempo de execução de uma multiplicação. Desta forma, utilizando um computador capaz de realizar 10^{10} multiplicações por segundo seriam necessários aproximadamente 95 segundos. Porém, aumentando o valor para $N = 200$ ($h = 0.005$) o tempo necessário aumenta para aproximadamente 1 hora e 43 minutos e para e $N = 500$ ($h = 0.002$) o tempo aumenta para mais de 17 dias. Com isto, podemos concluir que o método de Eliminação de Gauss não é uma boa estratégia para a resolução deste problema.

Considere $A = [a_{i,j}]_{n\times n}$ uma matriz real tal que $det(A) \neq 0$ e o sistema linear $Ax = b$. Nossa estratégia será escrever $A = M - N$, de modo que seja possível escrever um método iterativo da forma

$$Mx - Nx = b \Rightarrow Mx^{k+1} = b + Nx^k$$

Considere também a hipótese de que $det(M) \neq 0$. Assim, caso tal processo seja convergente, então o limite do processo será a solução do sistema $Ax = b$ pois, se

$$\lim_{k\to\infty} x^k = \bar{x} \Rightarrow M\bar{x} = b + N\bar{x} = (M-N)\bar{x} = b \Rightarrow A\bar{x} = b$$

Definição 5.1. *Seja A uma matriz diagonalizável. Dizemos que o maior autovalor (em módulo) de uma matriz A é o raio espectral da matriz, denotado por $\rho(A)$.*

Proposição 5.2 (Condição necessária para a convergência dos métodos iterativos)**.** *Seja $A = (M-N)$, onde A e M são inversíveis. Um método iterativo será convergente se, e somente se, $\rho(M^{-1}N) < 1$.*

Demonstração. Definindo $e^k = \bar{x} - x^k$ temos que

$$Me^{k+1} = Ne^k \Rightarrow e^{k+1} = M^{-1}Ne^k \Rightarrow e^k = (M^{-1}N)^k e^0$$

Logo,

$$e^k \to 0 \Leftrightarrow (M^{-1}N)^k \to 0$$

Mas isto ocorre se, e somente se, $\rho(M^{-1}N)^k < 1$. ■

Escrevendo $A = L + D + U$, onde L é a parte triangular inferior (*lower*), D a diagonal da matriz e U a parte triangular superior (*upper*), podemos escrever os métodos de **Jacobi** e **Gauss-Seidel**

- Método de Jacobi: M = D, N = L+U.

  O método será convergente se, e somente se

  $$\rho(D^{-1}(L+U)) < 1$$

- Método de Gauss-Seidel: M = L+D, N = U

  O método será convergente se, e somente se

  $$\rho((D+L)^{-1}U) < 1$$

Gostaríamos agora de saber se estes métodos podem ser aplicados para a resolução do sistema (8). O teorema a seguir fornece algumas condições nas quais os métodos de Jacobi e Gauss-Seidel são convergentes.

Teorema 5.3. *Se A é uma matriz irredutível e fracamente diagonal dominante, então os métodos de Jacobi e Gauss-Seidel são convergentes.*

Demonstração.

- **Método de Jacobi:** Seja λ um autovetor da matriz $D^{-1}(L+U)$. Temos que

  $$det(D^{-1}(L+U)-\lambda I)=0 \Rightarrow det(D^{-1}(L+U-\lambda D))=0 \Rightarrow det(D^{-1})det((L+U-\lambda D))=0$$

  Como A é fracamente diagonal dominante e irredutível, temos que $det(D)\neq 0$ e, portanto, $det(D^{-1})\neq 0$. Logo

  $$det((L+U-\lambda D))=0$$

  Caso $|\lambda|\geq 1$ temos que $(L+U-\lambda D)$ é irredutível e fracamente diagonal dominante. Pelo Teorema (3.3), temos que a matriz é inversível e, portanto, $det((L+U-\lambda D))\neq 0$. Logo, $|\lambda|<1$ para todos os autovalores de $D^{-1}(L+U)$ de modo que $\rho(D^{-1}(L+U))<1$. Pela Proposição (5.2), temos que o método será convergente.

- **Método de Gauss-Seidel:** Seja λ um autovetor da matriz $(D+L)^{-1}U$. Temos que

  $$\begin{aligned} det((D+L)^{-1}U-\lambda I)=0 &\Rightarrow det((D+L)^{-1}(U-\lambda(D+L)))=0 \\ &\Rightarrow det((D+L)^{-1})det(U-\lambda(D+L))=0 \end{aligned}$$

  Por hipótese[1] $det((D+L)^{-1}\neq 0$, de modo que

  $$det(U-\lambda(D+L))=0$$

  Caso $|\lambda|\geq 1$ temos que $(U-\lambda(D+L))$ é irredutível e fracamente diagonal dominante. Novamente pelo Teorema (3.3) a matriz $U-\lambda(D+L)$ é inversível e $det(U-\lambda(D+L))\neq 0$. Logo, $|\lambda|<1$ para todos os autovalores de $(D+L)^{-1}U$ de modo que $\rho(-(D+L)^{-1}U)<1$. Pela Proposição (5.2), segue a convergência do método.

■

Agora que sabemos que ambos os métodos são convergentes, seria interessante que pudessemos estimar a velocidade de convergência de cada um desses métodos. Para isto, considere $\bar{x}$ a solução do sistema $Ax=b$, x^k a aproximação obtida na k-ésima iteração do método, $\lambda_1<\lambda_2<\cdots<\lambda_n=\rho(M^{-1}N)$ e $v^1,v^2,\cdots,v^n$ os autovalores e autovetores da matriz $M^{-1}N$. Assim

$$x^k-\bar{x}=(M^{-1}N)^k(x^0-\bar{x})$$

Como os autovetores formam uma base para o $\mathbb{R}^n$, podemos escrever

$$\sum_{i=1}^{n}c_iv^i=x^0-\bar{x}$$

e, para $k\to\infty$

$$(M^{-1}N)^k(x^0-\bar{x})=\sum_{i=1}^{n}\lambda_i^kc_iv^i=\underbrace{\lambda_n^k\sum_{i=1}^{n-1}\left(\frac{\lambda_1}{\lambda_n}\right)^k c_iv^i}_{\to 0}+c_n\lambda_n^kv^n$$

[1] A matriz de iteração deve ser inversível, para que o sistema possa ser resolvido a cada iteração do método.

Desta forma, podemos concluir que

$$||x^{k+1} - x^k|| \approx \rho(M^{-1}N)^k ||x^k - \bar{x}|| \tag{19}$$

e a taxa de convergência das aproximações é controlada pelo valor de $\rho(M^{-1}N)$. Assim, para que um dado método iterativo apresente rápida convergência para a solução do problema devemos ter que $\rho(M^{-1}N)$ deve ser pequeno o suficiente para que poucas iterações sejam suficientes para fornecer boas aproximações.

Para avaliar a convergência do método de Jacobi iremos aplicar o método aos vetores $\psi_{k,l}$, $1 \leq k, l \leq N-1$, cujas componentes são dadas por

$$(\psi_{k,l})_{i,j} = \sin(k\pi x_i)\sin(l\pi y_j) \quad 1 \leq i, j \leq N-1$$

Temos que

$$\begin{aligned}\left[D^{-1}(L+U)\right](\psi_{k,l})_{i,j} &= -\frac{1}{h^2}\left[(\sin(k\pi x_{i+1}) + \sin(k\pi x_{i-1}))\sin(l\pi y_j) + (\sin(l\pi y_{i+1}) + \sin(l\pi y_{i-1}))\sin(k\pi x_i)\right] \\ &= -\frac{1}{h^2}\left[2\cos(k\pi h) + 2\cos(l\pi h)\right](\psi_{k,l})_{i,j}\end{aligned}$$

Portanto

$$\left[D^{-1}(L+U)\right]\psi_{k,l} = \left(\frac{\cos(k\pi h) + \cos(l\pi h)}{2}\right)\psi_{k,l}$$

Desta forma, temos que os vetores $\psi_{k,l}$ formam uma base de autovetores para a matriz $\left[D^{-1}(L+U)\right]$, com os autovalores dados por

$$\mu_{k,l} = \frac{\cos(k\pi h) + \cos(l\pi h)}{2}$$

O maior valor de $\mu_{k,l}$ será dado quando $k = l = 1$ e assim

$$\rho(\left[D^{-1}(L+U)\right]) = \cos(\pi h) \approx 1 - \frac{h^2\pi^2}{2} < 1 \tag{20}$$

o que mostra que **o método de Jacobi possui uma baixa velocidade de convergência**.

De maneira análoga, obtemos que para o método de Gauss-Seidel

$$\rho(\left[D^{-1}(L+U)\right]) = \cos^2(\pi h) \approx 1 - h^2\pi^2 < 1 \tag{21}$$

de forma que o método de **Gauss-Seidel possui a velocidade de convergência um pouco maior do que o método de Jacobi**. Em ambos os casos, quando $h \to 0$ a velocidade de convergência da solução vai diminuindo, pois o raio espectral de ambas as matrizes tendem a 1. Ou seja, quanto mais refinada a malha, menor a velocidade de convergência dos métodos de Jacobi e de Gauss-Seidel.

Visando aumentar a velocidade de convergência vamos inserir um parâmetro livre no problema de forma que com a manipulação desse parâmetro seja possível aumentar a velocidade de convergência dos métodos numéricos. Dado o sistema $Ax = b$ e considerando $\omega \in \mathbb{R}$, vamos escrever a k-ésima iteração do método como uma combinação ponderada entre a $(k-1)$-ésima aproximação e a k-ésima aproximação obtida pelo método de Gauss-Seidel. Isto equivale a escrever

$$x_i^k = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j<i} a_{i,j}x_j^k - \sum_{j>i} a_{i,j}x_j^k\right) + (1-\omega)x_i^{k-1}$$

Na forma matricial, temos que

$$(D + \omega L)x^k = ((1-\omega)D - \omega U)x^{k-1} + \omega b$$

onde a matriz de iteração do método será dada por

$$S = (D + \omega L)^{-1}((1-\omega)D - \omega U)$$

e o método será convergente se, e somente se, $\rho(S) < 1$. Este método conhecido como *Sucessive Over-Relaxation* ou **SOR**.

Os resultados a seguir apresentam os resultados de convergência e de velocidade de convergência do método SOR. As demonstrações não serão feitas aqui, pois utilizam alguns conceitos que fogem ao escopo de uma primeira apresentação. Ao leitor interessado, a referência [2] contém as provas dos resultados a seguir:

Teorema 5.4. *Se o método SOR é convergente, então* $0 < \omega < 2$.

Teorema 5.5. *O parâmetro ótimo para o método SOR quando aplicado a matriz do sistema gerado por (8) é dado por:*

$$\omega = \frac{2}{1 + \sin(h\pi)}$$

Com este parâmetro, o raio espectral da matriz de iteração é

$$\rho(S) = \frac{1 - \sin \pi h}{1 + \sin \pi h} \approx 1 - 2\pi h$$

Assim como nos métodos de Jacobi e de Gauss Seidel, observa-se que o raio espectral da matriz tende a 1 quando h tende a 0. Porém, neste método a convergência é linear, enquanto que nos outros métodos a convergência é quadrática.

Vamos analisar como isso influência o erro na prática: Sendo e^k o erro cometido na k-ésima iteração, pela desigualdade (19) temos que o erro na próxima iteração depende diretamente do raio espectral da matriz do método, o qual denotaremos apenas por ρ. Dado um erro e^k vamos a quantidade L de iterações necessárias para o erro ser igual a ε vezes este erro, para algum $\varepsilon < 1$.

Para isto, da desigualdade (19) temos que

$$\begin{cases} e^{k+L} = \varepsilon e^k \\ e^{k+L} \approx \rho^L e^k \end{cases} \Rightarrow \rho^L \approx \varepsilon$$

e, portanto

$$L \approx \frac{\ln \varepsilon}{\ln \rho}$$

Como o raio espectral ρ pode ser escrito da forma $\rho = 1 - \delta$ para algum δ pequeno, podemos utilizar que $\ln(1-\delta) \approx -\delta$. Assim

$$L \approx \frac{\ln \varepsilon}{\ln \rho} = \frac{\ln \varepsilon^{-1}}{-\ln \rho} = \frac{\ln \varepsilon^{-1}}{-\ln(1-\delta)} \approx \frac{\ln \varepsilon^{-1}}{\delta} \Rightarrow L \approx \frac{\ln \varepsilon^{-1}}{\delta}$$

Assim, temos que

- **Método de Jacobi**:
$$L \approx \frac{2}{\pi^2} \ln \varepsilon^{-1} N^2$$
- **Método de Gauss-Seidel**:
$$L \approx \frac{1}{\pi^2} \ln \varepsilon^{-1} N^2$$
- **Método SOR (parâmetro ótimo)**:
$$L \approx \frac{1}{2\pi} \ln \varepsilon^{-1} N$$

Tomando como exemplo $N = 100$ temos que para reduzir o erro em um décimo ($\varepsilon = 0.1$) são necessárias

- **Método de Jacobi**: $L \approx 0.467N^2 = 4670$ iterações.
- **Método de Gauss-Seidel**: $L \approx 0.234N^2 = 2340$ iterações.
- **Método SOR (parâmetro ótimo)**: $L \approx 0.364N = 37$ iterações.

5.1 Implementação computacional

Na resolução numérica do Problema (1) sujeito à condição (2) iremos utilizar no vetor de soluções do sistema dois índices, um representando a linha e outro representando a coluna. A aplicação dos métodos de Jacobi, Gauss-Seidel e SOR seguem a mesma estrutura:

1. Inicialize o vetor $u_{i,j}$ com alguma aproximação inicial. Caso não tenha nenhuma informação sobre o problema, utilize o vetor nulo;
2. Inicialize o lado direito do sistema com as informações da função f;
3. Utilize a função g para atribuir os valores aos contornos do domínio (isto é, para os índices $i, j = 0$ ou N);
4. Atualize os pontos internos de acordo com o método utilizado;
5. Calcule a diferença entre as duas iterações.;
6. Se a diferença for maior do que uma tolerância TOL, retorne ao passo (4). Caso contrário, encerre.

Observe que a tolerância TOL pode ser uma estimativa muito pequena, de modo que o algoritmo pode demorar muito para convergir. Desta forma, é aconselhavel que seja colocado também uma restrição no número máximo de passos a ser executado pelo código.

Algorithm 1: Método de Jacobi.

```
//Inicialize vetor de soluções com alguma condição inicial (ou zero);
//Inicialize o lado direito do sistema;
//Insira as condições de contorno;
para eps < TOL ou k < ITMAX faça
    para j = 1 até N − 1 faça
        para i = 1 até N − 1 faça
            unovo_{i,j} = (1/4)(b_{i,j} + u_{i−1,j} + u_{i+1,j} + u_{i,j−1} + u_{i,j+1});
        fim para
    fim para
    //Calcule a diferença entre as iterações;
    para j = 1 até N − 1 faça
        para i = 1 até N − 1 faça
            u_{i,j} = unovo_{i,j};
        fim para
    fim para
    k = k + 1;
fim para
```

Algorithm 2: Método de Gauss-Seidel.

```
//Inicialize o vetor de soluções com alguma condição inicial (ou zero);
//Inicialize o lado direito do sistema;
//Insira as condições de contorno;
para eps < TOL ou k < ITMAX faça
    para j = 1 até N − 1 faça
        para i = 1 até N − 1 faça
            c = (1/4) (u_{i+1,j} + u_{i−1,j} + u_{i,j+1} + u_{i,j−1} − 4u_{i,j} + h^2 b_{i,j});
            u_{i,j} = u_{i,j} + c;
        fim para
        //Utilize o incremento c para calcular a norma da diferença entre duas iterações;
    fim para
    k = k + 1;
fim para
```

A diferença entre os dois métodos está no fato que o método de Gauss-Seidel pode ser implementado utilizando apenas um vetor pois as novas aproximações já são armazenadas nas entradas correspondentes e utilizadas para construir as aproximações dos outros pontos. Já o método de Jacobi precisa de dois vetores: um para armazenar as novas aproximações e um para armazenar as aproximações anteriores.

Algorithm 3: Método SOR com parâmetro ω.

```
//Inicialize o vetor de soluções com alguma condição inicial (ou zero);
//Inicialize o lado direito do sistema;
//Insira as condições de contorno;
para eps < TOL ou k < ITMAX faça
    para j = 1 até N − 1 faça
        para i = 1 até N − 1 faça
            c = (ω/4) (u_{i+1,j} + u_{i−1,j} + u_{i,j+1} + u_{i,j−1} − 4u_{i,j} + h^2 b_{i,j});
            u_{i,j} = u_{i,j} + c;
        fim para
        //Utilize o incremento c para calcular a norma da diferença entre duas iterações;
    fim para
    k = k + 1;
fim para
```

Em ambas as implementações não foi preciso armazenar as matrizes do sistema linear pois sua estrutura é conhecida e permite que os algoritmos considerem apenas as operações com suas entradas não nulas.

Referências

[1] Stoer,J and Burlisch, R. - *Introduction to Numerical Analysis* - Springer - 3rd Ed. - 2002.

[2] Strikwerda, J.C.- *Finite Difference Schemes and Partial Differential Equations* - SIAM - 2nd ed. - 2004.

[3] LeVeque, R.J. - *Finite Difference Methods for Ordinary and Partial Differential Equations - Steady-State and Time-Dependent problems* - SIAM - 2007

[4] Burden, R.L. and Faires, J.D. - *Numerical Analysis* - 9th ed. - Brooks/Cole - 2010