Solução de Sistemas Algébricos Lineares

Notas de aula - Métodos Iterativos

Prof. Yuri Dumaresq Sobral

Departamento de Matemática
Universidade de Brasília

2025

- Já conhecemos um pouco mais sobre os processos iterativos matriciais e sua convergência, podemos construir métodos iterativos para resolver sistemas de equações lineares.
- Para tal, vamos nos concentrar em problemas do tipo

$$A\mathbf{x} = \mathbf{b},$$

em que a matriz dos coeficientes A é inversível, ou seja, existe A^{-1} (matriz inversa), e o sistema possui uma única solução.

- Vamos tentar, portanto, transformar este sistema em um processo iterativo do tipo

$$\mathbf{x}_{n+1} = T\mathbf{x}_n + \mathbf{c},$$

em que T é construída a partir da matriz dos coeficientes A e o vetor $\mathbf{c}$ a partir de A e de $\mathbf{b}$. Queremos que a solução do sistema seja o ponto fixo $\mathbf{x}^*$ deste processo iterativo, isto é

$$\mathbf{x}^* = T\mathbf{x}^* + \mathbf{c}.$$

- A idéia, portanto, é construir os processos iterativos de tal forma que $\mathbf{x}^*$ seja um ponto fixo assintoticamente estável.
- Para isto, precisamos analisar como o erro do processo iterativo evolui ao longo das iterações:

$$\mathbf{e}_{n+1} = \mathbf{x}_{n+1} - \mathbf{x}^* = (T\mathbf{x}_n + \mathbf{c}) - (T\mathbf{x}^* + \mathbf{c}) = T(\mathbf{x}_n - \mathbf{x}^*) = T\mathbf{e}_n \Leftrightarrow$$

$$\mathbf{e}_{n+1} = T\mathbf{e}_n.$$

- Portanto, o erro do processo itertativo satisfaz um processo iterativo matricial linear, e já sabemos que sua solução geral é

$$\mathbf{e}_n = T^n\mathbf{e}_0.$$

- Então, teremos $\mathbf{e}_n \to \mathbf{0}$ com $n \to \infty$ se e somente se a matriz T for convergente, isto é, $T^n \to \mathbb{O}$ com $n \to \infty$. Para isto, precisamos que seu raio espectral $\rho(T) < 1$.

- Desta forma, devemos ter cuidado com a maneira de construir os processos iterativos!
- Mas como construir estes métodos a partir do sistema original? Vamos ver alguns exemplos.
- Vamos começar com um exemplo de um sistema típico com três equações e três incógnitas:

$$\begin{cases} a_{11}x + a_{12}y + a_{13}z = b_1 \\ a_{21}x + a_{22}y + a_{23}z = b_2 \\ a_{31}x + a_{32}y + a_{33}z = b_3 \end{cases}$$

- Uma maneira ingênua de gerar um processo iterativo é isolando uma variável em cada uma das equações. Por exemplo, se $a_{ii} \neq 0$:

$$\begin{cases} a_{11}x = b_1 - a_{12}y - a_{13}z \\ a_{22}y = b_2 - a_{21}x - a_{23}z \\ a_{33}z = b_3 - a_{31}x - a_{32}y \end{cases} \Leftrightarrow \begin{cases} x = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}y - \frac{a_{13}}{a_{11}}z \\ y = \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x - \frac{a_{23}}{a_{22}}z \\ z = \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}}x - \frac{a_{32}}{a_{33}}y \end{cases}$$

$$\begin{cases} x_{n+1} = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} y_n - \frac{a_{13}}{a_{11}} z_n \\ y_{n+1} = \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x_n - \frac{a_{23}}{a_{22}} z_n \\ z_{n+1} = \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}} x_n - \frac{a_{32}}{a_{33}} y_n \end{cases}$$

- Este sistema pode ser escrito em forma matricial da seguinte forma:

$$\underbrace{\begin{pmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{pmatrix}}_{\mathbf{x}_{n+1}} = \underbrace{\begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} \\ -\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 \end{pmatrix}}_{T} \underbrace{\begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix}}_{\mathbf{x}_n} + \underbrace{\begin{pmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \frac{b_3}{a_{33}} \end{pmatrix}}_{\mathbf{c}}.$$

- Notem que esta é exatamente a forma que temos em mente para o processo iterativo. PORÉM... Já fica clara uma grande limitação destes métodos: é muito difícil (ou impossível) determinar a priori se T será ou não uma matriz convergente.

- Vamos generalizar o método ingênuo que discutimos acima. Consideremos um sistema de M equações e M incógnitas dado por $A\mathbf{x} = \mathbf{b}$.
- Suponha que possamos decompor a matriz dos coeficientes A da seguinte maneira:

$$A = L + D + U,$$

isto é, decompomos a matriz A em suas partes estritamente triangular inferior L, diagonal D e estritamente triangular superior U, e reescrevemos o sistema como:

$$A\mathbf{x} = \mathbf{b} \Leftrightarrow (L + D + U)\mathbf{x} = \mathbf{b} \Leftrightarrow D\mathbf{x} + (L + U)\mathbf{x} = \mathbf{b} \Leftrightarrow$$

$$D\mathbf{x} = -(L + U)\mathbf{x} + \mathbf{b} \Leftrightarrow \mathbf{x}_{n+1} = -D^{-1}(L + U)\mathbf{x}_n + D^{-1}\mathbf{b}.$$

- Portanto, $T_J = -D^{-1}(L + U)$ e $\mathbf{c} = D^{-1}\mathbf{b}$. Este processo iterativo é chamado de Método de Gauss-Jacobi.

- Um processo iterativo um pouco mais sofisticado pode ser intuitivamente determinado a partir do Método de Gauss-Jacobi.
- No processo iterativo do Método de Gauss-Jacobi, cada nova aproximação $n+1$ de uma variável é determinada apenas a partir das antigas aproximações n. Por exemplo:

$$\begin{cases} x_{n+1} = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} y_n - \frac{a_{13}}{a_{11}} z_n \\ y_{n+1} = \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x_n - \frac{a_{23}}{a_{22}} z_n \\ z_{n+1} = \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}} x_n - \frac{a_{32}}{a_{33}} y_n \end{cases}$$

- A priori, já teríamos uma melhor estimativa para a variável x, calculada na primeira equação, que poderia ter sido utilizada nas equações subsequentes.
- De fato, o processo iterativo para o sistema acima poderia ser dado por:

$$\begin{cases} x_{n+1} = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} y_n - \frac{a_{13}}{a_{11}} z_n \\ y_{n+1} = \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x_{n+1} - \frac{a_{23}}{a_{22}} z_n \\ z_{n+1} = \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}} x_{n+1} - \frac{a_{32}}{a_{33}} y_{n+1} \end{cases}$$

- Este processo também pode ser escrito em uma forma matricial. Note que os termos que já são conhecidos na nova iteração $n+1$ são aqueles que correspondem aos coeficientes abaixo da diagonal principal de A (associados à matriz L).
- Considerando a mesma decomposição da matriz A feita anteriormente, temos:

$$A\mathbf{x} = \mathbf{b} \Leftrightarrow (L + D + U)\mathbf{x} = \mathbf{b} \Leftrightarrow (L + D)\mathbf{x} + U\mathbf{x} = \mathbf{b} \Leftrightarrow$$

$$(L+D)\mathbf{x} = -U\mathbf{x}+\mathbf{b} \Leftrightarrow \mathbf{x}_{n+1} = -(L+D)^{-1}U\mathbf{x}_n+(L+D)^{-1}\mathbf{b}.$$

- Portanto, $T_S = -(L + D)^{-1}U$ e $\mathbf{c} = (L + D)^{-1}\mathbf{b}$. Este processo iterativo é chamado de Método de Gauss-Seidel.
- Tal como no Método de Gauss-Jacobi, no Método de Gauss-Seidel não temos como garantir a priori que a matriz T_S será convergente.

- Além disto, apesar de parecer ser melhor, não há garantias de que o Método de Gauss-Seidel será de fato melhor, isto é, que ele convergirá quando o Método de Gauss-Jacobi divergir, e que será mais rápido que o Método de Gauss-Jacobi.
- Como vimos antes, não podemos afirmar a priori, se as matrizes T_J e T_S serão convergentes. Para isto, precisamos determinar seu raio espectral, o que é muito caro.
- Um resultado que pode facilitar um pouco a análise é o seguinte: é possível mostrar, para qualquer matriz A, que

$$\rho(A) \quad \leq \quad \max_{i=1,\ldots,M} \quad \sum_{j=1}^{M} |a_{ij}| = S_{\ell}^{Max}.$$

- Isto é, o raio espectral é limitado pelo maior valor da soma dos valores absolutos de todos os elementos de uma linha, S_{ℓ}^{Max}.

- Um caso particular bem conhecido, e bastante frequente em problemas reais, é o caso de sistemas lineares cuja matriz de coeficientes A é estritamente diagonalmente dominante.
- Nestas matrizes, o valor absoluto do elemento da diagonal, $|a_{ii}|$, é estritamente maior que a soma dos valores absolutos de todos os outros elementos da linha, S_i^p.
- Exemplo:

$$A = \begin{pmatrix} 10 & 0 & -1 & 2 & 4 \\ -2 & -18 & 2 & 1 & -5 \\ 0 & 2 & 6 & 1 & -1 \\ -1 & -2 & -1 & -7 & 1 \\ 0 & 2 & 1 & 0 & 14 \end{pmatrix} \begin{array}{l} S_1^p = |0| + |-1| + |2| + |4| = 7 < |10| = |a_{11}| \\ S_2^p = |-2| + |2| + |1| + |-5| = 10 < |-18| = |a_{22}| \\ S_3^p = |0| + |2| + |1| + |-1| = 4 < |6| = |a_{33}| \\ S_4^p = |-1| + |-2| + |-1| + |1| = 5 < |-7| = |a_{44}| \\ S_5^p = |0| + |2| + |1| + |0| = 3 < |14| = |a_{55}| \end{array}$$

- Neste caso, por causa da divisão pelos elementos da diagonal nos Métodos de Gauss-Jacobi e de Gauss-Seidel, é fácil perceber que $S_\ell^{Max} < 1$ (estritamente menor que 1) e, portanto, tanto $\rho(T_J) < 1$, como $\rho(T_S) < 1$.

- Note que tanto no Método de Gauss-Jacobi, como no Método de Gauss-Seidel, a matriz T tem elementos nulos na diagonal. Isto quer dizer que valores das iterações anteriores da própria variável não entram no cálculo das próximas iterações.
- Podemos, então, pensar similarmente ao que foi feito no Método de Gauss-Seidel e tentar melhorar a convergência dos métodos.
- Para isto, vamos considerar, como anteriormente, a decomposição da matriz A da seguinte forma:

$$A\mathbf{x} = \mathbf{b} \Leftrightarrow (L + D + U)\mathbf{x} = \mathbf{b},$$

e vamos assumir que o método de Gauss-Seidel seja convergente para este sistema.

- Vamos reescrever, agora, esta decomposição da seguinte maneira:

$$(L + D + U)\mathbf{x} = \mathbf{b} \Leftrightarrow (L + (1+\alpha - \alpha)D + U)\mathbf{x} = \mathbf{b} \Leftrightarrow$$

$$[L + \alpha D + (1 - \alpha)D + U]\mathbf{x} = \mathbf{b}$$

- Vamos pensar em construir um método com base no Método de Gauss-Seidel. Portanto:

$$[L+\alpha D+(1-\alpha)D+U]\mathbf{x} = \mathbf{b} \Leftrightarrow (L+\alpha D)\mathbf{x} = -[(1-\alpha)D+U]\mathbf{x}+\mathbf{b},$$

e, normalmente, dividimos tudo por α e tomamos $\omega = 1/\alpha$, obtendo o seguinte processo iterativo:

$$\mathbf{x}_{n+1} = (\omega L + D)^{-1}[(1-\omega)D - \omega U]\mathbf{x}_n + \omega(\omega L + D)^{-1}\mathbf{b}.$$

- Portanto, a matriz que define este processo iterativo é

$$T_\omega = (\omega L + D)^{-1}[(1-\omega)D - \omega U],$$

e vemos que ela depende de um parâmetro livre ω. Será que podemos escolher adequadamente ω para minimizar (ou pelo menos diminuir) o seu raio espectral $\rho(T_\omega)$?

- Precisamos associar, de alguma maneira, $\rho(T_\omega)$ com ω.
- Vamos calcular o determinante de T_ω:

$$\det(T_\omega) = \det\Big((\omega L + D)^{-1}[(1-\omega)D - \omega U]\Big) =$$

$$= \det\Big((\omega L + D)^{-1}\Big)\det\Big((1-\omega)D - \omega U\Big).$$

- Como L e U são matrizes estritamente triangulares, os elementos de suas diagonais são todos zero. Portanto, os determinantes das matrizes $\omega L + D$ e $(1-\omega)D - \omega U$ são determinados apenas por D e por $1-\omega)D$, respectivamente.
- Usando as propriedades do determinante de uma matriz, temos, então:

$$\det(T_\omega) = \det\Big((\omega L + D)^{-1}\Big)\det((1-\omega)D - \omega U) = \frac{1}{\det(D)}(1-\omega)^M \det(D)$$

- E, portanto, $\det(T_\omega) = (1-\omega)^M$.
- Por outro lado, assumindo que T_ω tem M autovalores e seus M autovetores forem base de $\mathbb{R}^M$ (já vimos esta hipótese antes!), é possível mostrar que

$$\det(T_\omega) = \lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \ldots \cdot \lambda_M,$$

  isto é, o determinante de T_ω é dado pelo produto de seus M autovalores.
- Comparando os dois resultados:

$$|\det(T_\omega)| = |(1-\omega)^M| = |\lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \ldots \cdot \lambda_M| \leq \rho(T_\omega)^M.$$

- Como precisamos que $\rho(T_\omega) < 1$ para que T_ω seja convergente, então:

$$|(1-\omega)^M| \leq \rho(T_\omega)^M < 1 \Leftrightarrow |1-\omega| < 1 \Leftrightarrow$$

$$-1 < 1-\omega < 1 \Leftrightarrow 0 < \omega < 2.$$

- Assim, não podemos escolher qualquer valor para ω!
- O parâmetro ω é chamado de parâmetro de relaxação do método e dizemos que o método é:
  - sub-relaxado, se $0 < \omega < 1$
  - Gauss-Seidel, se $\omega = 1$
  - sobre-relaxado, se $1 < \omega < 2$
- Se $\omega < 0$ ou $\omega > 2$, claramente o método diverge.
- Normalmente, os métodos sobre-relaxados são os mais eficientes!
- Para algumas matrizes especiais, é possível calcular o valor ótimo de ω para que $\rho(T_\omega)$ seja mínimo!
- Exemplo: matrizes simétricas, tridiagonais por blocos, positivas definidas ($\mathbf{x}^T \cdot A \cdot \mathbf{x} > 0 \quad \forall \mathbf{x} \neq \mathbf{0}$) são tais que:

$$\rho(T_J)^2 = \rho(T_S), \quad \omega_{ot} = \frac{2}{1 + \sqrt{1 - \rho(T_S)}} \quad \text{e} \quad \rho(T_\omega) = \omega_{ot} - 1.$$

- Então, se tivermos uma matriz A tal como acima com $\rho(T_J) = 0.99$ (péssimo!), então teremos que

$$\rho(T_S) = \rho(T_J)^2 = 0.99^2 = 0.9801$$

$$\omega_{ot} = \frac{2}{1 + \sqrt{1 - 0.9801}} = 1.7527 \quad \text{e} \quad \rho(T_\omega) = 0.7527.$$

- Fazendo as contas, vemos que $\rho(T_\omega) \approx \rho(T_S)^{14} \approx \rho(T_J)^{28}$!!!! O método sobre-relaxado é 28 vezes mais rápido que o método de Gauss-Jacobi para estas matrizes!
- Os métodos sobre-relaxados costumam ser chamados de Métodos SOR (successive over-relaxation) e, de fato, podemos reescrevê-los de uma maneira bem fácil e intuitiva!

- Relembrando o método de Gauss-Seidel em sua forma matricial:

$$\mathbf{x}_{n+1} = -(L+D)^{-1}U\mathbf{x}_n+(L+D)^{-1}\mathbf{b} \Leftrightarrow (L+D)\mathbf{x}_{n+1} = -U\mathbf{x}_n+\mathbf{b}.$$

- E, portanto, podemos escrevê-lo como:

$$D\mathbf{x}_{n+1} = -U\mathbf{x}_n+\mathbf{b}-L\mathbf{x}_{n+1} \Leftrightarrow \mathbf{x}_{n+1}^{Sei} = D^{-1}\Big(-U\mathbf{x}_n+\mathbf{b}-L\mathbf{x}_{n+1}\Big).$$

- Por outro lado, voltando ao método SOR:

$$\mathbf{x}_{n+1} = (\omega L + D)^{-1}[(1-\omega)D - \omega U]\mathbf{x}_n + \omega(\omega L + D)^{-1}\mathbf{b} \Leftrightarrow$$

$$(\omega L + D)\mathbf{x}_{n+1} = [(1-\omega)D - \omega U]\mathbf{x}_n + \omega\mathbf{b} \Leftrightarrow$$

$$D\mathbf{x}_{n+1} = (1-\omega)D\mathbf{x}_n - \omega U\mathbf{x}_n + \omega\mathbf{b} - \omega L\mathbf{x}_{n+1} \Leftrightarrow$$

$$\mathbf{x}_{n+1} = (1-\omega)\mathbf{x}_n + \omega D^{-1}\Big(-U\mathbf{x}_n + \mathbf{b} - L\mathbf{x}_{n+1}.\Big) \Leftrightarrow$$

$$\mathbf{x}_{n+1} = (1-\omega)\mathbf{x}_n + \omega\mathbf{x}_{n+1}^{Sei} \qquad (SOR)$$

- Finalmente, temos que decidir quando parar as iterações destes métodos! Existem vários critérios de parada diferentes.
- Todos se baseiam em um determinado valor de tolerância, TOL, que podemos escolher tão pequeno quanto quisermos. Quanto menor for TOL, mais iterações serão necessárias.
- Alguns exemplos de critérios frequentemetne usados são:
  - máxima diferença absoluta das componentes do vetor $\mathbf{x}$ em duas iterações consecutivas:

$$\delta_{n+1} = \max_{1 \leq i \leq M} |x_{i\,n+1} - x_{i\,n}| < TOL$$

  - erro relativo associado a δ_{n+1}:

$$\varepsilon_{n+1} = \frac{\delta_{n+1}}{\max\limits_{1 \leq i \leq M} |x_{i\,n+1}|} < TOL$$

  - norma euclidiana dos resíduos $\mathbf{R}_{n+1}$:

$$\mathbf{R}_{n+1} = \mathbf{b} - A \cdot \mathbf{x}_{n+1}, \qquad |\mathbf{R}_{n+1}| = \sqrt{\sum_{i=1}^{M} R_{i\,n+1}^2} < TOL$$