Atividade para casa #4

Soluções de Sistemas lineares
Métodos para matrizes especiais

May 18, 2025

1 Contextualização

Até esse ponto vimos já alguns métodos poderosos para resolver sistemas lineares. Nossa construção de raciocínio tem se baseado na busca por uma melhoria constante no algoritmo de solução visando a obtenção de soluções confiáveis com economia gradual de recursos computacionais. Vimos por exemplo que o método de eliminação Gaussiana resolve o problema do alto custo computacional associado ao cálculo de determinantes de matrizes cheias, demandado por exemplo pela regra de Crammer, representando assim uma melhoria. Vimos também que o método de Gauss-Jordan pode ser uma melhoria em relação à eliminação Gaussiana a partir do momento em que a diagonalização do sistema original nos permite pular a etapa de substituição. Já a aplicação da decomposição L.U representa também uma economia de recursos computacionais ao não precisar modificar o vetor $\{b\}$ quando comparada com a eliminação Gaussiana, apesar de incluir uma etapa adicional de substituição na estimativa do vetor $\{d\}$. De toda sorte, você agora compreende não só como esses algoritmos são construídos, mas é também capaz de fazer uma análise do número de operações em ponto flutuante associada a cada algoritmo.

A ideia aqui é avançar no campo da proposição de novos métodos de solução de sistemas lineares, porém incorporando reflexões adicionais sobre funcionalidades de cada método. Por exemplo, no caso da decomposição L.U, podemos utilizá-la não só para resolver um sistema linear, como também para invertir uma matriz quadrada. Isso é útil em contextos nos quais deseja-se simplesmente inverter uma matriz e não necessariamente resolver um sistema linear. Um exemplo prático de situação física no qual isso é necessário ocorre por exemplo na descrição do comportamento físico de suspensões líquido-sólido (figura 1).

Quando estamos descrevendo o movimento de pequenas partículas sólidas imersas num líquido base, estamos no fundo falando de um problema de N corpos. Pense nisso! Quando uma partícula imersa em um fluido se desloca ela induz um escoamento que acaba afetando partículas vizinhas. Essas partículas afetadas acabam induzindo outros escoamentos que por sua vez pertubam as demais e por aí vai. Podemos entender que essas partículas encontram-se presas a uma rede de interações de longo alcance nos quais afetam-se mutuamente. Essas interações são chamadas no campo da microhidrodinâmica de interações hidrodinâmicas. Existe um campo maduro do conhecimento chamado de microhidrodinâmica que estuda esse mecanismo físico. Como a velocidade de cada partícula é descrita pela segunda lei de Newton, podemos escrever para cada partícula que

$$m_p \frac{d\boldsymbol{v}}{dt} = \sum \boldsymbol{f}_h + \sum \boldsymbol{f}_{nh}, \tag{1}$$

em que m_p, $\boldsymbol{v}$ e t representam respectivamente a massa da partícula, a velocidade da mesma e o tempo. Já $\boldsymbol{f}_h$ e $\boldsymbol{f}_{nh}$ representam as forças hidrodinâmicas e não hidrodinâmicas as quais a partícula encontra-se sujeita respectivamente. No limite em que $m_p \to 0$, ou seja, para partículas muito

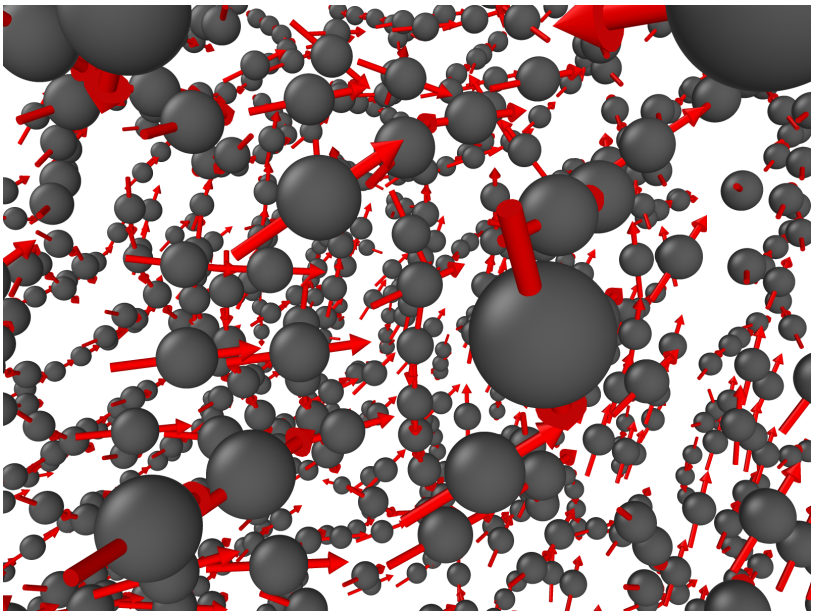

Figure 1: Interior da microestrutura de um fluido magnético: exemplo de uma suspensão líquido-sólido regida por uma física de interações de longo alcance

pequenas, de tal sorte que possam ser consideradas não-massivas, ou livres de efeitos inerciais, temos simplesmente um equilíbrio instantâneo entre forças hidrodinâmicas e forças não hidrodinâmicas. Para o caso limite em que temos apenas uma única partícula isolada, sedimentando sob ação da gravidade em um fluido viscoso e sujeita a um escoamento em baixos Reynolds no limite $Re \to 0$, temos simplesmente que

$$-6\pi\eta a\boldsymbol{v} = \sum \boldsymbol{f}_{nh} \to \boldsymbol{v} = -\left(\frac{1}{6\pi\eta}\right)\boldsymbol{I}\cdot\sum \boldsymbol{f}_{nh}, \tag{2}$$

ou em outras palavras

$$\boldsymbol{v} = \boldsymbol{M}\cdot\sum \boldsymbol{f}_{nh}, \quad \text{com} \quad \boldsymbol{M} = -\left(\frac{1}{6\pi\eta}\right)\boldsymbol{I}. \tag{3}$$

Aqui, dizemos que $\boldsymbol{M}$ é a matriz mobilidade do problema. Nesse caso, para partículas não massivas temos uma relação direta entre a velocidade das partículas e as forças não hidrodinâmicas que atuam entre elas. Para uma única partícula isolada, essa matriz é uma matriz diagonal, que leva em conta o único efeito hidrodinâmico possível para uma partícula isolada em baixo Reynolds que é o próprio arrasto de Stokes. Esse racional pode ser estendido para um problema de N corpos em termos de uma formulação mobilidade geral, dada por

$$\begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}^{11} & \boldsymbol{M}^{12} & \cdots & \boldsymbol{M}^{1N} \\ \boldsymbol{M}^{21} & \boldsymbol{M}^{22} & \cdots & \boldsymbol{M}^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{M}^{N1} & \boldsymbol{M}^{N2} & \cdots & \boldsymbol{M}^{NN} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ \vdots \\ \boldsymbol{f}_N \end{pmatrix}, \tag{4}$$

em que agora cada termo $\boldsymbol{u}_i$ do vetor no lado esquerdo da equação (4) é um vetor, representando as componentes x, y, z da velocidade de cada partícula i. Da mesma forma, cada termo $\boldsymbol{M}^{ij}$ no interior da grande matriz do lado direito da equação é uma matriz 3×3 que conectará a posição da partícula i com a partícula j em questão. Finalmente, cada vetor $\boldsymbol{f}_i$ do lado direito da equação (4) representa o somatório das forças não hidrodinâmicas que atuam em cada partícula i do sistema. A equação (4) é a síntese do que conhecemos como *formulação mobilidade* no contexto de interações hidrodinâmicas em baixos Reynolds. Desenvolvimentos analíticos no campo da microhidrodinâmica nos permitem

conhecer a grande matriz mobilidade do sistema com base na configuração instantânea do sistema. Utilizando essa abordagem, toda a questão da determinação das velocidades das partículas em uma suspensão passa a ser uma questão de multiplicação de matrizes. É claro que existem armadilhas no caminho. A matriz mobilidade é uma matriz cheia, preenchida com caras funções transcendentais e para gerar comportamentos estatisticamente convergentes das propriedades de transporte dessas suspensões deve ser computada em espaços periódicos utilizando uma técnica sofisticada de somas nestes tipos de espaço, conhecida como somas de Ewald. De toda sorte, apesar do alto custo computacional não deixa de ser um grande problema de multiplicação matricial.

Um gargalo dessa formulação é que ela só vale para partículas não massivas. Caso estivéssemos interessados em resolver um problema que fuja do limite assintótico $m_p \to 0$, precisaríamos conhecer as forças hidrodinâmicas atuantes em cima de cada partícula. Um caminho para isso seria inverter a grande matriz mobilidade e ao invés de partir das forças para calcular as velocidades, partiríamos das velocidades instantâneas para calcular as forças que iriam deslocar o nosso sistema para uma nova configuração. A inversa da matriz mobilidade é chamada de matriz resistência. O custo computacional de se resolver um problema de resistência é ainda maior do que o custo associado à solução de um problema de mobilidade devido à necessidade de inversão de uma grande matriz. Mas em muitos contextos essa é a saída para a simulação desse tipo de problema. A pergunta que aparece agora é: como construir algoritmos que façam o processo de inversão matricial?

2 Decomposição L.U e inversão matricial

Uma boa notícia no campo da construção de algoritmos de inversão matricial é que a própria decomposição L.U pode ser usada para tal finalidade. A sequência de passos responsável por realizar o processo de inversão matricial a partir da decomposição L.U é descrita a seguir.

1. Calculamos as matrizes $[L]$ e $[U]$;
2. Concebemos um vetor $\{d_1\}$ associado ao cálculo
$$[L]\{d_1\} = \begin{Bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{Bmatrix}, \tag{5}$$
3. Calculamos $\{d_1\}$ por substituição progressiva;
4. Façamos $[U]\{x_1\} = \{d_1\}$ e calculamos $\{x_1\}$ por substituição progressiva;
5. $\{x_1\}$ passa a ser a primeira coluna de $[A]^{-1}$;
6. Façamos agora
$$[L]\{d_2\} = \begin{Bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{Bmatrix}, \tag{6}$$
7. Calculamos $\{d_2\}$ por substituição progressiva;
8. Façamos $[U]\{x_2\} = \{d_2\}$ e calculamos $\{x_2\}$ por substituição progressiva;
9. $\{x_2\}$ passa a ser a segunda coluna de $[A]^{-1}$;
10. Repetimos os passos anteriores até estimarmos a última coluna de $[A]^{-1}$ e nossa matriz está montada;

3 Métodos para matrizes especiais

Nessa seção vamos olhar para alguns métodos dedicados à solução de sistemas lineares envolvendo matrizes especiais. Iremos considerar aqui dois tipos de matrizes especiais: as matrizes de banda e as matrizes simétricas. Mas antes de adentrarmos nos meandros dos métodos utilizados para tal finalidade é interessante fornecer um pouco de contexto para justificar esse estudo.

Esses tipos de matrizes aparecem naturalmente em problemas de engenharia. Quando pensamos num problema físico descrito por equações diferenciais, o cenário unidimensional do problema quando discretizado para fins de simulação sempre recairá numa matriz de banda, mais do que isso, para o caso 1D, no qual cada nó da nossa rede nodal discretizada se comunica apenas com os vizinhos da frente e de traz acaba culminando em uma matriz dos coeficientes para o sistema discretizado com largura de banda igual à 3. Esse tipo de sistema é conhecido como sistema tridiagonal e está sujeito à aplicação de um algoritmo próprio para isso, conhecido como algoritmo de Thomas para sistemas tridiagonais. O algoritmo de Thomas nada mais é do que uma variante da técnica de decomposição L.U aplicada a esse tipo particular de sistema. É interessante notar que a existência de algoritmos mais específicos voltados à solução de sistemas lineares é sempre pautada por exemplos práticos que acabam recaindo neste tipo de sistema. A vantagem por exemplo de trabalharmos com esses tipos de sistemas é a redução de custo computacional obtida por meio de algoritmos aplicados a sistemas especiais que demandam menos espaço de armazenamento na memória do computador e menos laços aninhados (*nested loops*). Isso ficará claro com a exposição dos métodos voltados à solução de sistemas envolvendo matrizes especiais nessa seção.

Um outro tipo de sistema especial que iremos ver aqui nessa seção diz respeito à sistemas envolvendo matrizes simétricas. Nesse ponto cabe um pouco mais de contexto para nos motivar antes de prosseguirmos. É sempre importante nos perguntarmos: em que tipos de sistemas surgem matrizes espontaneamente simétricas? Na verdade, essa é por exemplo uma discussão muito profunda no campo de algumas ciências físicas baseada na mecânica dos meios contínuos, como é o caso da Mecânica dos Fluidos. Sabemos por exemplo que o tensor de tensões que descreve o estado dos carregamentos mecânicos num ponto infinitesimal de um fluido Newtoniano em escomento é por natureza simétrico. Esse tensor é geralmente estruturado e organizado na forma de uma matriz 3×3 na forma

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix}, \tag{7}$$

em que na simbologia σ_{ij}, o índice i representa o plano de atuação daquele esforço e j a direção do carregamento. Num contexto em que as tensões se distribuem simetricamente em todo o material, temos que $\boldsymbol{\sigma} = \boldsymbol{\sigma}^T$, o que significa que $\sigma_{xy} = \sigma_{yx}$, $\sigma_{xz} = \sigma_{zx}$ e $\sigma_{yz} = \sigma_{zy}$. Dessa forma, como 3 das 9 componentes se repetem, precisamos de apenas 6 informações distintas para definir o estado de tensão em um ponto em um fluido Newtoniano. Essa simetria tensorial nos leva a uma redução no espaço necessário para armazenamento das informações que descrevem o estado do carregamento mecânico de um fluido Newtoniano em movimento. A simetria do tensor de tensões de um fluido Newtoniano é passível de demonstração matemática e surge como consequência da aplicação de um princípio físico fundamental sobre um meio contínuo: o princípio de balanço de momento angular. Nesse sentido, essa simetria é uma consequência física da natureza constitutiva de um fluido Newtoniano. É possível provar por exemplo que quando este fluido possui nanopartículas magnéticas em seu interior (o que chamamos de ferrofluido) a presença de torques magnéticos intrínsecos ao meio (quando na presença de um campo externo) leva como consequência a uma quebra na simetria da distribuição de tensões no material. Dessa forma, as propriedades matemáticas do tensor de tensões de um meio contínuo estão intrinsicamente conectadas com as características físicas do material. De toda sorte, podemos encontrar vários exemplos de situações físicas que ao serem modeladas

matematicamente acabam recaindo em sistemas lineares especiais formados tanto por matrizes de banda quando por matrizes simétricas. Veremos a seguir algumas formas de abordarmos esses tipos de sistema.

3.1 Sistemas envolvendo matrizes de banda

Uma matriz de banda é aquela para a qual todos os termos são nulos com exceção da diagonal e de uma faixa centrada em torno da diagonal, como a que vemos na equação (8)

$$[A] = \begin{pmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{44} \\ 0 & 0 & a_{43} & a_{44} \end{pmatrix}, \tag{8}$$

para a matriz apresentada em (8) temos uma largura de banda igual a 3, o que significa que a maior quantidade de termos não nulos por linha é 3. Na verdade, essa é a maior largura de banda possível para uma matriz de banda 4×4. Entretanto, podemos ter matrizes de ordem bem mais alta com essa mesma largura de banda, como a matriz apresentada em (9)

$$[A] = \begin{pmatrix} f_1 & g_1 & 0 & 0 & \cdots & 0 \\ e_2 & f_2 & g_2 & 0 & \cdots & 0 \\ 0 & e_3 & f_3 & g_3 & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & e_{n-1} & f_{n-1} & g_{n-1} & 0 \\ \cdots & \cdots & \cdots & e_n & f_n & g_n \end{pmatrix}, \tag{9}$$

note que para a matriz tridiagonal $n \times n$ representada em (8) podemos armazenar todos os termos não nulos em termos dos vetores e, f, g, cada qual contendo a seguinte quantidade de termos:

- $e = [2, \cdots, n] \rightarrow n - 1$ termos
- $f = [1, \cdots, n] \rightarrow n$ termos
- $g = [1, \cdots, n] \rightarrow n$ termos

de tal sorte que ao invés de armazenarmos n^2 termos, podemos armazenar apenas $3n-1$ termos. Isso gera uma grande economia de espaço de armazenamento. A aplicação da técnica de decomposição L.U para resolver esse tipo de sistema pode ser implementada em termos de um algortimo especial, denominado *algoritmo de Thomas* que resolve o sistema por meio do pseudocódigo em FORTRAN apresentado em (2). Note que ao invés de vários laços entralaçados temos apenas 3 laços simples em sequência, o que gera uma enorme economia de recursos computacionais para a solução do sistema. Na verdade, o que o algoritmo de Thomas faz é basicamente a aplicação da decomposição L.U, porém já direcionada a matrizes tridiagonais, de tal sorte que ao invés de trabalhar com as n^2 componentes originais da matriz $[A]$, o mesmo trabalha apenas com os subvetores e, f, g necessários para a escrita da matriz $[A]$ no limite em que a mesma é uma matriz de largura de banda igual a 3. Além disso, no pseudocódigo apresentado, o vetor $\{b\}$, associado aos termos de fonte, está representado em termos de um vetor $\{r\}$. A pergunta central aqui é: *qual a diferença entre a decomposição L.U e o método de Thomas?* A resposta é: conceitualmente nenhuma, o que muda é apenas o algoritmo.

```
! Decomposição L.U
do k=2,n
e(k)= e(k)/f(k-1)
f(k) = f(k) - e(k)*g(k-1)
end do

!Substituição progressiva = cálculo de {d}
do k=2,n
r(k)=r(k) - e(k)*r(k-1)
end do

!Substituição regressiva = cálculo de {x}
x(n) = r(n)/f(n)
do k= n-1,1
x(k)=(r(k) - g(k)*x(k+1))/f(k)
end do
```

Figure 2: Pseudocódigo em FORTRAN para implementação do algoritmo de Thomas

3.2 Sistemas envolvendo matrizes simétricas

No contexto de sistemas lineares envolvendo matrizes simétricas, uma das técnicas de solução consiste na decomposição ou fatoração de Cholesky. Essa técnica se deve ao cartógrafo francês André-Louis Cholesky e consiste na proposição de relações simples de recorrência para decompor uma matriz $[A]$ simétrica no produto entre duas matrizes triangulares na forma $[A] = [L][L]^T$. Para apresentar a ideia da decomposição de Cholesky, consideremos um caso 3×3. Para esse cenário, definindo $[L]$ como

$$[L] = \begin{pmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{pmatrix}, \tag{10}$$

temos que

$$[L][L]^T = \begin{pmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{pmatrix} \cdot \begin{pmatrix} \ell_{11} & \ell_{21} & \ell_{31} \\ 0 & \ell_{22} & \ell_{32} \\ 0 & 0 & \ell_{33} \end{pmatrix} = \begin{pmatrix} \ell_{11}^2 & \ell_{11}\ell_{21} & \ell_{11}\ell_{31} \\ \ell_{11}\ell_{21} & \ell_{21}^2 + \ell_{22}^2 & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} \\ \ell_{11}\ell_{31} & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} & \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 \end{pmatrix}. \tag{11}$$

É fácil perceber que o produto $[L][L]^T$ nesse caso resulta em uma matriz simétrica, de tal sorte que se $[A] = [A]^T$, podemos então decompor $[A]$ em $[A] = [L][L]^T$. Além disso, para esse cenário 3×3 podemos escrever as seguintes relações entre os termos de $[A]$ e os termos da matriz $[L]$:

$$\begin{gathered} a_{11} = \ell_{11}^2, \quad a_{12} = a_{21} = \ell_{11}\ell_{21}, \quad a_{13} = a_{31} = \ell_{11}\ell_{31} \\ a_{22} = \ell_{21}^2\ell_{22}^2, \quad a_{23} = a_{32} = \ell_{21}\ell_{31} + \ell_{22}\ell_{32}, \quad a_{33} = \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2. \end{gathered} \tag{12}$$

As relações (12) podem ser invertidas, reorganizadas e generalizadas para obtenção das componentes de $[L]$ em sistemas de ordem n em termos das seguintes relações de recorrência:

$$\begin{gathered} \ell_{ki} = \frac{a_{ki} - \sum_{j=1}^{i-1} \ell_{ij}\ell_{kj}}{\ell_{ii}} \longrightarrow \text{para } i = 1, 2, \cdots, k-1 \\ \ell_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} \ell_{kj}^2} \end{gathered} \tag{13}$$

As relações (13) representam o que conhecemos como decomposição de Cholesky. Após o processo de decomposição é necessário resolver o sistema linear. Isso é feito assumindo que se $[A] = [L][L]^T$, então o sistema linear é dado por $[L][L]^T\{x\} = \{b\}$. Chamando $[L]^T\{x\} = \{y\}$, podemos reescrever esse sistema em termos de duas equações:

$$[L]\{y\} = \{b\} \quad \text{e} \quad [L]^T\{x\} = \{y\}. \tag{14}$$

Como as duas equações em (14) envolvem apenas matrizes triangulares podemos utilizar substituições progressivas e regressivas para calcular $\{y\}$ e $\{x\}$ respectivamente.

4 O método de Gauss-Seidel

Todos os métodos vistos até aqui são métodos que organizam a solução do sistema sem a necessidade de processos iterativos. Em outras palavras, os métodos que possuem sua origem na ideia da eliminação Gaussiana, não demandam um chute inicial da solução que vai sendo gradativamente refinado por meio de um processo iterativo. Nessa sessão vamos introduzir um esquema iterativo muito utilizado no campo da solução de sistemas lineares, conhecido como método de Gauss-Seidel. Para compreendermos a lógica do método de Gauss-Seidel, comecemos com um sistema 3×3 do tipo:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \tag{15}$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \tag{16}$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3. \tag{17}$$

Isolando x_1, x_2, x_3 da primeira, segunda e terceira equação que compõe o sistema, temos

$$x_1 = \left(b_1 - a_{12}x_2 - a_{13}x_3\right)/a_{11} \tag{18}$$
$$x_2 = \left(b_2 - a_{21}x_1 - a_{23}x_3\right)/a_{22} \tag{19}$$
$$x_3 = \left(b_3 - a_{31}x_1 - a_{32}x_2\right)/a_{33}. \tag{20}$$

Esse sistema pode ser transformado em um sistema iterativo por meio de duas formas diferentes. A primeira delas é ilustrada a seguir:

$$x_1^{j+1} = \left(b_1 - a_{12}x_2^j - a_{13}x_3^j\right)/a_{11} \tag{21}$$
$$x_2^{j+1} = \left(b_2 - a_{21}x_1^{j+1} - a_{23}x_3^j\right)/a_{22} \tag{22}$$
$$x_3^{j+1} = \left(b_3 - a_{31}x_1^{j+1} - a_{32}x_2^{j+1}\right)/a_{33}, \tag{23}$$

em que $j = 0, \cdots, N_{iter}$ sendo N_{iter} o número de iterações. Note que nas equações (21-23) a partir do momento em que obtemos uma melhor aproximação para um dos termos de $\{x\}$ imediatamente inserimos o novo valor na estimativa do próximo termo da solução que buscamos. Essa é a essência do método de Gauss-Seidel. Uma outra opção para a montagem do sistema iterativo é apresentada a seguir:

$$x_1^{j+1} = \left(b_1 - a_{12}x_2^j - a_{13}x_3^j\right)/a_{11} \tag{24}$$
$$x_2^{j+1} = \left(b_2 - a_{21}x_1^j - a_{23}x_3^j\right)/a_{22} \tag{25}$$
$$x_3^{j+1} = \left(b_3 - a_{31}x_1^j - a_{32}x_2^j\right)/a_{33}. \tag{26}$$

As equações (24-26) partem de um conjunto de pontos iniciais e usam esse conjunto para evoluir em bloco a solução para o próximo ponto. Esse esquema é conhecido como iteração de Jacobi e constitui uma variante do método de Gauss-Seidel. Como a proposta do método de Gauss-Seidel já considera o último valor atualizado de uma determinada componente do vetor solução que buscamos na estimativa da próxima, essa é a escolha dominante no campo de esquemas numéricos iterativos.

4.1 A convergência do método de Gauss-Seidel

Uma questão importante em qualquer esquema iterativo voltado à solução de problemas matemáticos utilizando métodos numéricos diz respeito à convergência do método. Nesse sentido, é importante entendermos como certas características da matriz dos coeficientes podem afetar a convergência do método de Gauss-Seidel. E para entendermos isso, vamos voltar a um problema mais simples: o de determinação de zeros de funções.

Um esquema análogo ao método de Gauss-Seidel no campo da determinação de raízes de equações é o método de iteração de ponto fixo. Vamos apresentar aqui esse método e discutirmos questões relativas à convergência do método de iteração de ponto fixo para estimativa de raízes de equações de uma única variável para em seguida traçarmos os devidos paralelos com a análise de convergência do método de Gauss-Seidel. Se temos uma função $f(x)$ cuja raiz queremos obter, uma estratégia ingênua e direta de obtermos a raiz é tentar manipular algebricamente a função $f(x)$ para isolarmos x em termos de uma outra função $g(x)$, de tal sorte que $x = g(x)$. Para entendermos melhor a ideia, vamos a um exemplo. Considere

$$f(x) = x + \frac{x^2 - 3}{\sqrt{x}} = 0 \longrightarrow x = -\frac{x^2 - 3}{\sqrt{x}} = g(x). \tag{27}$$

Podemos inclusive usar artimanhas criativas para isolar a variável x de uma função transcendental. Suponha $f(x) = \sin(x) = 0$, caso queiramos obter a raiz dessa função podemos somar x dos dois lados para obter $x = x + \sin(x) = g(x)$. Baseado nessa ideia, o método da iteração do ponto-fixo consiste em transformar a relação $x = g(x)$ numa relação iterativa, como $x_{i+1} = g(x_i)$. Dessa forma, partimos de um valor x_0 inicial, inserimos esse valor em $g(x_0)$ para estimar o próximo valor $x_1 = g(x_0)$ e prosseguimos assim até uma possível convergência para a raiz que estamos buscando.

Suponha agora que a solução verdadeira do nosso problema é $x_r = g(x_r)$, dessa forma podemos escrever

$$x_r - x_{i+1} = g(x_r) - g(x_i), \tag{28}$$

agora, se $g(x)$ e $g'(x)$ são ambas funções contínuas no intervalo $a \leq x \leq b$, então existe um valor intermediário $x = \zeta$ nesse intervalo $[a, b]$ para o qual

$$g'(\zeta) = \frac{g(b) - g(a)}{b - a}. \tag{29}$$

A equação (29) é a representação de um teorema matemático chamado de teorema do valor médio para derivadas. Se fizermos $a = x_i$ e $b = x_r$, podemos obter da equação (29):

$$g(x_r) - g(x_i) = (x_r - x_i)g'(\zeta), \tag{30}$$

mas se $x_{i+1} = g(x_i)$, então

$$x_r - x_{i+1} = (x_r - x_i)g'(\zeta), \tag{31}$$

se definirmos o erro verdadeiro para i-ésima iteração como a diferença entre o valor verdadeiro de x e o valor de x avaliado na iteração i, de tal sorte que $\varepsilon_{a,i} = x_r - x_i$, podemos finalmente escrever

$$\varepsilon_{a,i+1} = \varepsilon_{a,i}g'(\zeta), \tag{32}$$

o que significa afirmar que se $|g'(x)| < 1$ os erros diminuem a cada iteração e o método é convergente, enquanto para $|g'(x)| > 1$ os erros daumentam a cada iteração e o método diverge. Nesse sentido, para situações convergentes os erros vão diminuindo proporcionalmente a cada iteração, o que classifica o método da iteração de ponto-fixo como um método linearmente convergente.

Podemos também enxergar esse mecanismo de convergência ou divergência visualmente. Se $x_{i+1} = g(x_i)$, podemos chamar $y_1 = x$ e $y_2 = g(x)$ e traçar as funções y_1 e y_2 em um plano xy para

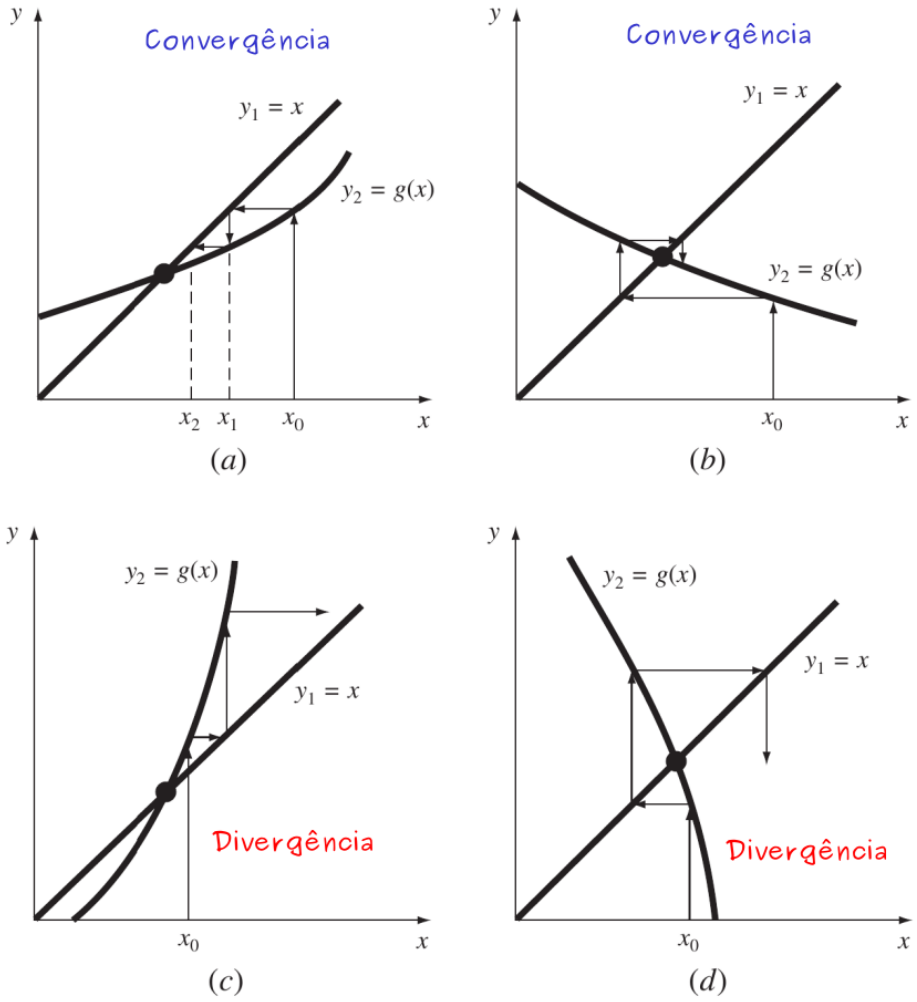

Figure 3: Representação visual da convergência do método de iteração de ponto-fixo

enxergar a evolução da interação entre essas duas funções a partir de um chute inicial x_0 conforme ilustrado na figura (3).

Agora voltemos à análise de convergência do método de Gauss-Seidel. Para isso consideremos um sistema linear simples de ordem 2 representado genericamente por

$$a_{11}x_1 + a_{12}x_2 = b_1 \rightarrow x_1 = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2 = u \tag{33}$$

$$a_{21}x_1 + a_{22}x_2 = b_2 \rightarrow x_2 = \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1 = v. \tag{34}$$

$$\tag{35}$$

Aqui estamos usando duas funções u e v para representar o nosso sistema 2×2 de tal sorte que a solução do sistema ocorre quando $u = v = 0$. Em outras palavras, essa forma de escrever o sistema nos permite traçar um paralelo entre sistemas lineares e raízes de equações, de tal sorte que o que aprendemos em termos da convergência do método da iteração de ponto-fixo possa ser aplicado à análise da convergência do método de Gauss-Seidel. Enquanto no contexto da iteração de ponto-fixo tínhamos convergência para $|g'(x)| < 1$, para o nosso cenário de um sistema 2×2 esse critério é reescrito como

$$\left|\frac{\partial u}{\partial x_1}\right| + \left|\frac{\partial u}{\partial x_2}\right| < 1 \qquad \text{e} \qquad \left|\frac{\partial v}{\partial x_1}\right| + \left|\frac{\partial v}{\partial x_2}\right| < 1 \tag{36}$$

Para o nosso sitema, temos

$$\frac{\partial u}{\partial x_1} = 0, \quad \frac{\partial u}{\partial x_2} = -\frac{a_{12}}{a_{11}}, \quad \frac{\partial v}{\partial x_1} = -\frac{a_{21}}{a_{22}}, \quad \frac{\partial v}{\partial x_2} = 0. \tag{37}$$

Substituindo (37) em (36) obtemos

$$\left|\frac{a_{12}}{a_{11}}\right| < 1 \qquad \text{e} \qquad \left|\frac{a_{21}}{a_{22}}\right| < 1. \tag{38}$$

A equação (38) nos diz que para um sistema 2×2 o método de Gauss-Seidel será convergente quando os termos da diagonal forem os maiores termos da linha em que se encontram. Essa conclusão pode ser estendida para sistemas gerais de ordem n por meio do seguinte critério geral:

$$|a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}|. \tag{39}$$

Uma forma interessante de controlarmos a convergência do método inclui uma estratégia chamada *relaxação*. Nesse contexto, após o cálculo do valor atual de x_{i+1} pelo método de Gauss-Seidel convencional, aplicamos a seguinte correção

$$x_{i+1} = \lambda x_{i+1} + (1 - \lambda) x_i. \tag{40}$$

Nesse contexto, temos três possibilidades:

- Se $\lambda = 1$, temos o método de Gauss-Seidel sem relaxação original;
- Se $0 < \lambda < 1$, temos o método de Gauss-Seidel com sub-relaxamento. Essa estratégia ajuda alguns sistemas não convergentes a convergirem;
- Se $1 < \lambda < 2$, temos o método de Gauss-Seidel com sobre-relaxamento ou SOR. Essa estratégia acelera a convergência de alguns sistemas de lenta convergência ao dar mais peso para os valores mais recentes de x, assumindo aqui que a solução está caminhando no rumo certo;

5 Para casa

1. Para a matriz $[A]$ abaixo, calcule a inversa de $[A]$ utilizando a técnica decorrente da decomposição L.U e em seguida verifique de $[A][A]^{-1} = [I]$;

$$[A] = \begin{pmatrix} 3 & -0.1 & -0.2 \\ 0.1 & 7 & -0.3 \\ 0.3 & -0.2 & 10 \end{pmatrix} \tag{41}$$

2. Considere o seguinte sistema tridiagonal:

$$\begin{pmatrix} 2.04 & -1 & 0 & 0 \\ -1 & 2.04 & -1 & 0 \\ 0 & -1 & 2.04 & -1 \\ 0 & 0 & -1 & 2.04 \end{pmatrix} \cdot \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 40.8 \\ 0.8 \\ 0.8 \\ 200.8 \end{Bmatrix} \tag{42}$$

   Resolva esse sistema na mão, aplicando a técnica de decomposição L.U como foi apresentada. Em seguida, monte uma tabela e execute na mão os laços apresentados no pseudocódigo vinculado ao algoritmo de Thomas. Compare os procedimentos em termos do número de operações de ponto flutuante e do resultado final obtido a partir de cada caminho.

3. Invente um sistema linear se sua escolha no qual a matriz dos coeficientes seja uma matriz quadrada, simétrica, de ordem 3. Você pode escolher os números que quiser para compor a matriz dos coeficientes e o vetor $\{b\}$, desde que a matriz $[A]$ seja simétrica. Em seguida, aplique as relações de recorrência associadas à decomposição de Cholesky e resolva o sistema linear proposto.

4. Use o método de Gauss-Seidel (a) sem relaxamento e (b) com relaxamento (utilizando $\lambda = 1.2$) para resolver o seguinte sistema de equações com um erro relativo de 5%. Se necessário, reorganize as equações para garantir a convergência.

$$\begin{aligned} 2x_1 - 6x_2 - x_3 &= -38 \\ -3x_1 - x_2 + 7x_3 &= -34 \\ -8x_1 + x_2 - 2x_3 &= -20 \end{aligned}$$