UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA APLICADA

Otimização do Método SOR para Matrizes p-cíclicas Consistentemente Ordenadas

por

José Caleffi

Dissertação submetida como requisito parcial
para a obtenção do grau de
Mestre em Matemática Aplicada

Prof. Dr. **Oclide José Dotto**
Orientador

Porto Alegre, novembro de 2000.

CIP – Catálogo na publicação

Caleffi, José

Otimização do Método SOR para Matrizes p-cíclicas Consistentemente Ordenadas / José Caleffi.-Porto Alegre: PPGMAP da UFRGS, 2000.

87 p.:il.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Matemática Aplicada, Porto Alegre, 2000. Orientador: Dr. Oclide José Dotto.

Dissertação: Algebra Linear, Análise Numérica.

Agradecimentos

Agradeço de um modo muito especial ao meu orientador, prof. Dr. Oclide José Dotto, por ter-me mostrado o verdadeiro sentido da descoberta do conhecimento através de suas colocações, não só no que se refere à Matemática, mas também no que diz respeito à vida.

Esse agradecimento se estende:

- aos colegas Sérgio, Fábio, Roque, Adalberto, Lucile, Mauren, Simone, Elizabete, Rui, e Sirlei pelas horas de estudo em conjunto, e aos demais colegas que também sempre me incentivaram e me ajudaram;

- aos meus colegas da Universidade de Caxias do Sul que sempre me ajudaram, em especial à Vânia, Isolda, Helena e Ricardo;

- aos professores Vilmar, Rudnei, Maria Paula e Maria Cristina que, quando precisei, me atenderam;

- aos meus pais e aos meus irmãos, que sempre me incentivaram nos estudos;

- a minha amiga Cleusa, pela ajuda durante os meus estudos;

- a minha esposa Sylvia e aos meus filhos Fabrício e Maurício, pela compreensão das horas furtadas para meus estudos.

Sumário

Lista de figuras

Lista de tabelas

Lista de abreviaturas

Símbolo ..... Significado

$\rho(\mathbf{M})$ ..... raio espectral da matriz **M**

$\sigma(\mathbf{M})$ ..... espectro da matriz **M**

$\mathbf{M}^t$ ..... transposta da matriz **M**

$\mathbf{M}^{-1}$ ..... inversa da matriz **M**

$\|\bullet\|$ ..... norma (geralmente euclidiana)

D ..... matriz diagonal

L ..... matriz triangular inferior estrita

U ..... matriz triangular superior estrita

$\mathbf{B}_J$ ..... matrizes de escalares de iteração de Jacobi

$\mathbf{B}_{GS}$ ..... matrizes de escalares de iteração de Gauss-Seidel

$\mathbf{B}_{SOR}$ ..... matrizes de escalares de iteração do método SOR

ω ..... parâmetro de relaxação

ω_o ..... parâmetro ótimo de relaxação

B ..... matriz de blocos de Jacobi

$\mathbf{S}_\omega$ ..... matriz de blocos do SOR

$\mathcal{L}_N$ ..... classe de matrizes definida a partir das matrizes-escada

sse ..... se e somente se

Resumo

Estudamos a otimização do método SOR clássico, para a resolução de um sistema linear **Ax = b**, com **A** não-singular, a partir dos resultados de Young [55, 57] e Varga [50, 51] para matrizes de blocos p-cíclicas consistentemente ordenadas. Num primeiro nível, a otimização refere-se à escolha do parâmetro de relaxação do SOR que produz a maior velocidade de convergência, e, num segundo nível, à escolha da p-ciclicidade que apresenta o melhor desempenho com os valores ótimos do parâmetro, e damos ênfase ao caso 2-cíclico. Além disso, descrevemos a otimização do parâmetro em três generalizações:

a) num relaxamento das condições sobre o espectro da matriz de Jacobi associada a **A**;

b) no método SOR para matrizes singulares;

c) num novo método SOR, que substitui a decomposição **A = D – L – U**, onde **D**, **L** e **U** são a diagonal de **A**, a parte triangular inferior estrita de **A** e a parte triangular superior estrita de **A**, pela **A = D – P – Q**, onde **P** pertence a uma classe de matrizes construída a partir das matrizes-escada.

Descrevemos também a aplicação do caso singular às cadeias de Markov, comentamos a computação paralela aplicada ao SOR, e apresentamos diversas simulações relativas à otimização desse método.

Palavras-chaves: sistema linear, SOR, parâmetro, otimização, matriz-es-cada.

Abstract

We study the optimization of the classic SOR method for solving a linear system **Ax** = **b**, where **A** is a nonsingular p-cyclic consistently ordered block matrix, based on the discoveries of Young [55, 57] and Varga [50, 51]. In a first level, the optimization refers to the choice of the SOR relaxation parameter, which produces the greatest convergence speed and, in a second level, to the p-cyclicity that presents the best performance with the optimal parameter values and emphasize the 2-cyclic case. Moreover we describe three SOR generalizations concerning optimization:

a) by weakening the conditions on the spectrum of Jacobi matrix associated with **A**;

b) by considering the SOR method for singular matrices;

c) by approaching a new SOR, that replaces the splitting $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$, where **D**, **L** and **U** are the diagonal of **A**, the strict lower triangular part of **A** and the strict upper triangular part of **A**, respectively, by this one $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$, where **P** is a stair matrix or a matrix even more general than a stair matrix.

We also describe the application of the singular case to Markov chains, discuss parallel computing applied to SOR method, and present several simulations regarding the optimization of that method.

Key words: linear system, SOR, parameter, optimization, stair matrix.

Introdução

Dado um Sistema de Equações Lineares Algébricas (SELAS),

$$\mathbf{Ax} = (\mathbf{D} - \mathbf{L} - \mathbf{U})\mathbf{x} = \mathbf{b}, \tag{1}$$

onde **A** é uma matriz quadrada particionada em blocos, **D** é a matriz diagonal de **A** com os blocos diagonais quadrados não-singulares, e **L** e **U** são matrizes especiais, nosso trabalho gira em torno da otimização do *método da sobre-relaxação sucessiva* (SOR),

$$\mathbf{x}^{(n+1)} = (\mathbf{D} - \omega\mathbf{L})^{-1}\left((1-\omega)\mathbf{D} + \omega\mathbf{U}\right)\mathbf{x}^{(n)} + (\mathbf{D} - \omega\mathbf{L})^{-1}\omega\mathbf{b}, \tag{2}$$

$n = 0, 1, 2, \ldots, 0 \neq \omega \in \mathbb{R}$, cuja finalidade é calcular iterativamente uma aproximação numérica para a solução exata do SELAS (1).

Até o início da última década, fora desenvolvido o SOR somente para o caso em que **A** é não-singular e as matrizes **L** e **U** em (1) triangulares inferior e superior estritas. Mas, a necessidade prática, particularmente ligada às cadeias de Markov, incentivou a pesquisa do SOR em torno de um SELAS (1) singular consistente, originando publicações como [05, 22, 31, 32]. O capítulo 4 se ocupa com SELAS nessa última condição.

O grande volume das pesquisas relativas à otimização do SOR foi feito para o caso não-singular, com **L** e **U** como descritas acima, iniciado por Young [55, 56,57] em 1950 e depois por Varga [49, 50] em 1958. Com a percepção de que os métodos não-estacionários, como o método do gradiente conjugado, que têm o mesmo objetivo do SOR, isto é, aproximar iterativamente a solução de um SELAS grande e esparso, apresentavam melhor desempenho, este último método ficou ofuscado por quase duas décadas. Mas a computação paralela veio despertar a pesquisa em torno dos métodos estacionários, nos quais se enquadra o SOR, tanto que, na década passada, houve inúmeras investigações sobre a otimização do SOR [13, 46, 52, 53]. Esse fato levou Varga a reeditar e revisar, no corrente ano, seu livro [50], que marcou época em 1961. Além disso, a idéia de DeLong e Ortega, [12, 13] de usar o SOR como precondicionador dos métodos do tipo do gradiente conjugado, como o GMRES [42] (não do método do gradiente conjugado puro, que requer a simetria da matriz dos coeficientes) foi por eles testada comparativamente com sucesso, em conjugação com a

computação paralela, aplicada ao caso de uma matriz consistentemente ordenada com o ordenamento preto-vermelho [55].

A otimização do SOR pode ser vista de dois ângulos. Um deles, objeto da maior investigação, consiste em determinar o valor do parâmetro ω que produz a convergência mais rápida da seqüência (2) para a solução do SELAS (1). Uma medida da velocidade dessa convergência é obtida com raio espectral $\rho(\mathbf{S}_\omega)$ da *matriz do SOR*,

$$\mathbf{S}_{\text{SOR}} := (\mathbf{D} - \omega\mathbf{L})^{-1}\left((1-\omega)\mathbf{D} + \omega\mathbf{U}\right).$$

É um fato conhecido, apresentado no capítulo 1, que (2) converge, seja qual for o vetor inicializador $\mathbf{x}^{(0)}$, se e somente se $\rho(\mathbf{S}_\omega) < 1$. Ainda, quanto menor esse raio espectral, mais rápida a convergência. Então o aspecto da otimização do SOR de que estamos falando consiste em determinar, quando existe,

$$\rho(\omega_o) := \min_{\omega} \rho(\mathbf{S}_\omega).$$

Nesse caso dizemos que ω_o é o *parâmetro ótimo* ou *valor ótimo do parâmetro* ω. Os resultados obtidos não contemplam todos os SELAS, mas apenas aqueles, cuja matriz **A** dos coeficiente é p-cíclica consistentemente ordenada. Tais resultados são devidos a Young [55] e a Varga [51], e são apresentados no capítulo 2. As demonstrações dos lemas que aí ocorrem são nossas, seguindo sugestões de Varga. No capítulo 4, destinado ao caso em que **A** é singular, tratamos brevemente de uma generalização dos resultados de Varga, precisamente como preparação para a abordagem do caso singular.

Que é uma matriz p-cíclica consistentemente ordenada? – Suponhamos que, dada uma matriz quadrada de blocos **A**, com blocos diagonais não-singulares, exista uma matriz de permutação **P** tal que (o correspondente *reordenamento simétrico* de **A** seja)

$$\mathbf{PAP}^{\mathrm{t}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_{1p} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{32} & \mathbf{A}_{33} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{pp-1} & \mathbf{A}_{pp} \end{bmatrix}, \text{ com } p \geq 2. \qquad (3)$$

Pondo $\mathbf{D} := \operatorname{diag}(\mathbf{A}_{11}, \mathbf{A}_{22}, \ldots, \mathbf{A}_{pp})$, a *matriz de Jacobi* $\mathbf{B} := \mathbf{I} - \mathbf{D}^{-1}(\mathbf{PAP}^{-1})$ associada toma a forma

$$\mathbf{B} := \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_{1p} \\ \mathbf{B}_{21} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{32} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_{pp-1} & \mathbf{0} \end{bmatrix}. \qquad (4)$$

É fácil verificar que $\mathbf{B}_{1p} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{1p}$ e $\mathbf{B}_{ii-1} = -\mathbf{A}_{ii}^{-1}\mathbf{A}_{ii-1}$, $i = 2, 3, \ldots, p$. Matrizes na forma (4), as designamos *matrizes fracamente cíclicas de índice p* e, nesse caso, denominamos a matriz **A** de p-*cíclica*. Equivalentemente, se, para uma matriz **A**, a matriz de Jacobi associada $\mathbf{B} := \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$ tem um reordenamento simétrico do tipo (4), dizemos que **B** é fracamente cíclica de índice *p*, e **A**, p-cíclica

Seja $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ a decomposição, como acima, de uma matriz quadrada p-cíclica. Dizemos que **A** é *consistentemente ordenada* sse os autovalores da matriz

$$\mathbf{B}(\alpha) := \alpha\mathbf{D}^{-1}\mathbf{L} + \alpha^{-(p-1)}\mathbf{D}^{-1}\mathbf{U},$$

independem de α, para $\alpha \neq 0$. Nesse caso dizemos também que a matriz de iteração de Jacobi $\mathbf{B}(1)$ é consistentemente ordenada.

Matrizes p-cíclicas consistentemente ordenadas ocorrem muito na prática, por exemplo, na discretização de problemas de contorno com equações diferenciais parciais, particularmente, em problemas de Poisson. Citamos as matrizes tridiagonais, como exemplo de matrizes 2-cíclicas. Matrizes 2-cíclicas estão entre as mais importantes, e, por isso, reservamos um tópico especial para tratar delas no capítulo 2.

Uma matriz p-cíclica, com $p \geq 3$, também pode ser q-cíclica com $2 \leq q < p$. Então cabe pesquisar o q ótimo. Em outras palavras: supondo que tenhamos determinado os parâmetros ótimos para os casos $2 \leq q_1 < q_2 < \ldots <$ p–cíclicos, perguntamos qual é a melhor q_i–ciclicidade, isto é, a que conduz mais rapidamente à solução do SELAS. Vem a ser como que uma otimização de segunda ordem. No capitulo 2 reservamos uma secção para tratar disso. Veremos que o SOR 2–cíclico leva vantagem sobre os demais, o que valoriza o trabalho pioneiro de Young.

Para o caso de um SELAS (1) singular consistente, tratado no capítulo 4, como já informamos acima, os resultados gerais se mantêm, desde que substituamos convergência por *semiconvergência* e o papel do raio espectral $\rho(\mathbf{S}_\omega)$ da matriz do SOR, pelo *fator de semiconvergência*

$$\gamma(\mathbf{S}_\omega) := \max\{|\lambda| \mid \lambda \in \sigma(\mathbf{S}_\omega), |\lambda| \neq 1\},$$

onde $\sigma(\mathbf{S}_\omega)$ indica o conjunto dos autovalores da matriz do SOR $\mathbf{S}_\omega$.

O método SOR clássico não é plenamente adequado para a computação paralela. A razão disso é o tipo de decomposição da matriz $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ em (1). Para

adequar melhor o SOR à computação paralela, sem perder o benefício da rapidez de convergência, ultimamente foi estudado [26] um novo SOR, onde as matrizes triangulares **L** e **U** foram substituídas por outra classe de matrizes, definidas a partir de uma classe básica, a das *matrizes-escada*. Dedicamos o capítulo 3 para dissertar sobre essa idéia, em vista da otimização do parâmetro. Veremos que os resultados sobre o SOR clássico se mantêm no novo SOR, com poucas alterações.

Abre-se aqui um campo de pesquisa – a pesquisa de outros métodos SOR, a partir de outras matrizes que desempenhem o papel das matrizes triangulares no SOR clássico, ou das matrizes-escada no novo SOR, com os quais possamos ter um possível ganho também no precondicionamento dos métodos não-estacionários. Aliás, o estudo do precondicionamento desses métodos não-estacionários mediante o novo SOR também não foi feito ainda.

O capítulo mais longo deste trabalho é o primeiro, no qual são revistos os conceitos e resultados básicos em torno dos processos estacionários e, particularmente, a relação fundamental entre os autovalores da matriz do SOR e a matriz de Jacobi, para matrizes p-cíclicas consistentemente ordenadas, base para o estudo da otimização do SOR e outras conseqüências nos capítulos posteriores.

Devido ao hoje estreito relacionamento do método SOR com a computação paralela, no Apêndice *A* fazemos uma breve descrição desse sistema de computação.

O Apêndice *B* traz a implementação conjunta no MATLAB dos algoritmos de Jacobi, Gauss-Seidel e SOR, além do algoritmo para matrizes-escada.

1 - Métodos Iterativos Estacionários

1.1. Introdução

Escrevemos um sistema de N equações lineares algébricas (SELAS) com a notação matricial

$$\mathbf{Ax} = \mathbf{b}, \tag{1.1}$$

onde $\mathbf{A}$ é a *matriz dos coeficientes* de ordem N, $\mathbf{x} \in \mathbb{R}^N$ é o *vetor variável* e $\mathbf{b} \in \mathbb{R}^N$ é o *vetor dos termos independentes*. Neste trabalho, exceto no capítulo 4, suporemos sempre que $\mathbf{A}$ é não-singular. *Resolver* (1.1), ou *achar a solução* de (1.1) é determinar em $\mathbb{R}^N$ o único vetor

$$\mathbf{x}_0 := \mathbf{A}^{-1}\mathbf{b} \in \mathbb{R}^N.$$

A maneira mais direta de aplicar um método iterativo para aproximar $\mathbf{x}_0$ é formular (1.1) como um problema de ponto fixo linear

$$\mathbf{x} = (\mathbf{I} - \mathbf{A})\mathbf{x} + \mathbf{b}.$$

Para esse problema, a seqüência das iterações é definida por

$$\mathbf{x}^{(n+1)} = (\mathbf{I} - \mathbf{A})\mathbf{x}^{(n)} + \mathbf{b}, \quad n \geq 0.$$

Genericamente, a formulação de (1.1) como método de ponto fixo linear é esta

$$\mathbf{x} = \mathbf{Mx} + \mathbf{c}, \qquad \mathbf{x}^{(n+1)} = \mathbf{Mx}^{(n)} + \mathbf{c}. \tag{1.2}$$

Designamos à matriz $\mathbf{M}$ de *matriz de iteração* e à seqüência $(\mathbf{x}^{(n)})$, de *seqüência das iterações* do método. Métodos iterativos desse tipo são chamados *métodos iterativos estacionários*, porque $\mathbf{x}^{(n+1)}$ só depende de $\mathbf{x}^{(n)}$ e não da história das iterações executadas. Contrapõem-se aos métodos estacionários os métodos não-estacionários, como, por exemplo, os métodos de Krylov, entre os quais o mais popular é o método do gradiente conjugado.

Ocupar-nos-emos com os métodos iterativos estacionários clássicos de Jacobi, de Gauss-Seidel e o método SOR, com o objetivo principal de aprofundar este último.

O uso de métodos iterativos, ao invés de métodos diretos (estes são os que teoricamente encontram a solução exata com um número finito de operações), é preferido quando a matriz dos coeficientes é grande e *esparsa*, isto é, constituída de grande quantidade de elementos nulos, pelo fato de que os métodos diretos acarretam perda de esparsidade e, com isso, uma tarefa mais árdua para os computadores.

1.2. Idéia básica da estruturação de (1.2)

De maneira geral, construímos um método iterativo do tipo (1.2) para um sistema (1.1), substituindo a matriz **A** por outra matriz não-singular **B**, mais fácil de manipular que a matriz **A**:

$$\mathbf{Bx} = (\mathbf{B} - \mathbf{A})\mathbf{x} + \mathbf{b}.$$

Daí obtemos

$$\mathbf{x} = \mathbf{B}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{x} + \mathbf{B}^{-1}\mathbf{b},$$

que se enquadra na forma (1.2) com $\mathbf{M} := \mathbf{B}^{-1}(\mathbf{B} - \mathbf{A})$ e $\mathbf{c} := \mathbf{B}^{-1}\mathbf{b}$.

1.3. Condições para convergência de um método iterativo do tipo (1.2)

Dizemos que um método iterativo (1.2) *converge* sse a correspondente seqüência $(\mathbf{x}^{(n)})$ das iterações converge. O interesse no estudo de um método iterativo gira em torno da convergência, rapidez e precisão com que o método aproxima a solução de um SELAS.

Antes de tudo precisamos saber quando há convergência, isto é, determinar em que condições o método converge. Para tal fim, apresentaremos alguns resultados, entre os quais o Teorema 1.3.3, que nos fornece uma condição necessária e suficiente para haver essa convergência. Os outros resultados nos servirão de subsídios para a demonstração do Teorema 1.3.3.

Na seqüência do trabalho, usaremos constantemente os conceitos de raio espectral e espectro de uma matriz. *Raio espectral* de uma matriz **A** é o maior valor $\rho(\mathbf{A})$ do conjunto dos módulos dos autovalores de **A**; e *espectro de* **A**, representado por $\sigma(\mathbf{A})$, é o conjunto dos autovalores de **A**. Normalmente usaremos os símbolos λ e μ para indicar autovalores.

O teorema a seguir é fundamental, pois conduz à definição da convergência ou não de um método iterativo estacionário (1.2).

1.3.1. Teorema. *Seja uma matriz quadrada* **M**. *Temos* $\lim_{n\to\infty} \mathbf{M}^n \to \mathbf{0} \Leftrightarrow \rho(\mathbf{M}) < 1$.

Prova. Nesta demonstração nos apoiamos na Fórmula Canônica de Jordan [27], pela qual **M** é semelhante a uma matriz bloco-diagonal especial: existe uma matriz não-singular **P** tal que

$$\mathbf{P}^{-1}\mathbf{M}\mathbf{P} = \begin{bmatrix} \mathbf{J}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{J}_r \end{bmatrix},$$

onde cada bloco $\mathbf{J}_i$ (*bloco de Jordan*) tem a forma

$$\mathbf{J}_i = \begin{bmatrix} \lambda_i & 1 & & \mathbf{0} \\ & \lambda_i & 1 & \\ & & \ddots & \ddots \\ & & & 1 \\ \mathbf{0} & & & \lambda_i \end{bmatrix}.$$

Aqui os λ_i, para $i = 1, 2, \ldots, r$, são os autovalores reais de **M**. Por indução, facilmente mostramos que

a $n^{ésima}$ potência de $\mathbf{J}_i$ é a matriz

$$\mathbf{J}_i^n = \begin{bmatrix} \lambda_i^n & n\lambda_i^{n-1} & \binom{n}{2}\lambda_i^{n-2} & \cdots & \binom{n}{k-1}\lambda_i^{n-k+1} \\ 0 & \lambda_i^n & n\lambda_i^{n-1} & \cdots & \binom{n}{k-2}\lambda^{n-k+2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_i^n & n\lambda_i^{n-1} \\ 0 & 0 & 0 & 0 & \lambda_i^n \end{bmatrix},$$

para $n \geq k-1$, onde k é a ordem de $\mathbf{J}_i$ e $\binom{n}{p}$ indica o número binomial n sobre p. Também vale:

$$\mathbf{M}^n = \mathbf{P}\begin{bmatrix} \mathbf{J}_1^n & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{J}_r^n \end{bmatrix}\mathbf{P}^{-1}.$$

Como a multiplicação matricial é contínua e a convergência de uma seqüência de matrizes se dá componente a componente, então $\mathbf{M}^n \to \mathbf{0}$ sse $\mathbf{J}_i^n \to 0$ para todo i, e $\mathbf{J}_i^n \to \mathbf{0}$ sse as componentes de $\mathbf{J}_i^n$ têm todas limite zero quando $n \to \infty$.

($\Rightarrow$) Se $\lim_{n\to\infty} \mathbf{M}^n = \mathbf{0}$, pelo exposto, todas as componentes de $\mathbf{J}_i^n$ tendem a zero quando $n \to \infty$. Em particular $\lambda_i^n \to 0$ para todo i, e isso somente é possível se $|\lambda_i| < 1$ para todo i. Logo $\rho(\mathbf{M}) < 1$.

($\Leftarrow$) Se $\rho(\mathbf{M}) < 1$, então $|\lambda_i| < 1$, para todo i. Nesse caso, as componentes de $\mathbf{J}_i^n$ tendem a zero; portanto, $\mathbf{J}_i^n \to \mathbf{0}$ para todo i e, por isso, $\lim_{n\to\infty} \mathbf{M}^n = \mathbf{0}$.

Para ver que, de fato, as componentes de $\mathbf{J}_i^n$ tendem a zero quando $|\lambda_i| < 1$, requer verificar que

$$\lim_{n\to\infty}\binom{n}{p}\lambda_i^n = 0, \text{ para } 0 \leq p \leq k-1.$$

Se $p = 0$, isso é quase imediato [33]. Em outro caso,

$$\lim_{n\to\infty}\binom{n}{p}\lambda_i^n = \lim_{n\to\infty} n^p\lambda_i^n = \lim_{x\to\infty}\frac{x^p}{\lambda_i^{-x}} = \lim_{x\to\infty} c\lambda_i^x = 0,$$

onde c é a constante proveniente da aplicação da regra de L'Hôpital p vezes . □

Com base no Teorema 1.3.1, fica fácil estabelecer uma condição necessária e suficiente para que haja convergência de um método (1.2). Mas antes apresentamos um resultado que nos servirá de ponte.

1.3.2. Teorema. *Um método iterativo* (1.2) *converge, com uma escolha inicial arbitrária* $\mathbf{x}^{(0)}$ *se e somente se* $\mathbf{M}^n \to \mathbf{0}$.

Prova. Para facilitar a linguagem, expressamos o fato $\mathbf{M}^n \to \mathbf{0}$ dizendo que *a matriz* $\mathbf{M}$ *converge.* Partindo dos vetores

$$\mathbf{x} = \mathbf{M}\mathbf{x} + \mathbf{c} \quad \text{e} \quad \mathbf{x}^{(n+1)} = \mathbf{M}\mathbf{x}^{(n)} + \mathbf{c},$$

subtraímos o primeiro do segundo, para obter

$$\mathbf{x}^{(n+1)} - \mathbf{x} = \mathbf{M}\left(\mathbf{x}^{(n)} - \mathbf{x}\right). \tag{1.3}$$

Já que essa igualdade vale para qualquer n, podemos escrever

$$\mathbf{x}^{(n)} - \mathbf{x} = \mathbf{M}\left(\mathbf{x}^{(n-1)} - \mathbf{x}\right). \tag{1.4}$$

Usando a (1.4) em (1.3), obtemos

$$\mathbf{x}^{(n+1)} - \mathbf{x} = \mathbf{M}^2\left(\mathbf{x}^{(n-1)} - \mathbf{x}\right).$$

Com mais n passos retroativos desse tipo, resulta, por indução, que, para todo n,

$$\mathbf{x}^{(n+1)} - \mathbf{x} = \mathbf{M}^{n+1}\left(\mathbf{x}^{(0)} - \mathbf{x}\right),$$

equivalentemente,

$$\mathbf{x}^{(n)} - \mathbf{x} = \mathbf{M}^{n}\left(\mathbf{x}^{(0)} - \mathbf{x}\right), \text{ para todo } n \geq 0. \tag{1.5}$$

Isto nos mostra que $\mathbf{x}^{(n)}$ converge para o ponto fixo $\mathbf{x}$ (solução do SELAS $\mathbf{Ax} = \mathbf{b}$), seja qual for o ponto inicializador $\mathbf{x}^{(0)}$, se e somente se $\mathbf{M}$ é convergente. □

1.3.3. Teorema. *Um método iterativo* (1.2) *converge, com uma escolha inicial arbitrária* $\mathbf{x}^{(0)}$, *se e somente se* $\rho(\mathbf{M}) < 1$.

Prova. Decorre imediatamente dos Teoremas 1.3.1 e 1.3.2. □

Temos o corolário útil seguinte:

1.3.4. Corolário. *Uma condição suficiente para um método* (1.2) *convergir para o ponto fixo, com uma escolha inicial arbitrária* $\mathbf{x}^{(0)}$, *é que* $\|\mathbf{M}\| < 1$, *para alguma norma matricial* $\|\cdot\|$.

Prova. Para toda matriz $\mathbf{M}$ e toda norma matricial $\|\cdot\|$, temos

$$\rho(\mathbf{M}) \leq \|\mathbf{M}\|.$$

De fato, seja $(\lambda, \mathbf{x})$ um *autopar* de $\mathbf{M}$, isto é, $\mathbf{Mx} = \lambda\mathbf{x}$, com $\mathbf{x} \neq \mathbf{0}$; temos $\|\mathbf{Mx}\| = \|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$ e $\|\mathbf{Mx}\| \leq \|\mathbf{M}\|\|\mathbf{x}\|$; como $\mathbf{x} \neq \mathbf{0}$, vem $|\lambda| \leq \|\mathbf{M}\|$. Agora o corolário segue do Teorema 1.3.3. □

Nota. A estratégia da demonstração do Teorema 1.3.3 é devida a Varga [51]. Axelsson [01] e Demmel [14] apresentam uma demonstração não-construtiva, sem o recurso da forma canônica de Jordan. □

A utilidade do Corolário 1.3.4 vem de que algumas normas, como as normas infinito e da soma, são fáceis de calcular.

O Corolário 1.3.4 também decorre do Teorema do Ponto Fixo de Banach [30]. De fato, na condição desse corolário, a transformação afim $T: \mathbf{x} \mapsto \mathbf{Mx} + \mathbf{b}$ é uma contração, isto é, $\|T(x) - T(y)\| \leq \|\mathbf{M}\| \cdot \|x - y\|$ com $\|\mathbf{M}\| < 1$, e, então, por esse teorema, a seqüência das iterações converge para o ponto fixo.

Associamos a um método iterativo (1.2) a *seqüência dos (vetores) erros*, $(\mathbf{e}^{(n)})$, cujo termo geral é definido por,

$$\mathbf{e}^{(n)} := \mathbf{x}^{(n)} - \mathbf{x},$$

onde $\mathbf{x}$ é o ponto fixo. Então, utilizando (1.5), podemos escrever

$$\mathbf{e}^{(n)} = \mathbf{M}^n \mathbf{e}^{(0)}, \quad n \geq 0. \tag{1.6}$$

1.4. Algumas matrizes especiais – Grafos

Aqui caracterizaremos algumas matrizes quadradas $\mathbf{A} = [a_{ij}]$ de ordem N, importantes para os objetivos do trabalho, e apresentaremos alguns teoremas em torno delas.

Uma matriz $\mathbf{A}$ tal que

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{ij}|, \quad i = 1, 2, \ldots, N,$$

a chamamos de matriz *diagonal-dominante*. Se for verificada a desigualdade estrita para todo i, a chamaremos de matriz *diagonal-dominante estrita*.

Matrizes extremamente importantes são simétricas ou hermitianas que satisfazem

$$\mathbf{x}^t \mathbf{A} \mathbf{x} > 0, \text{ para todo } \mathbf{x} \in \mathbb{R}^N \text{ não-nulo}^1.$$

Chamamos a essas matrizes de *matrizes simétricas* (*hermitianas*, para o caso complexo; nesse caso interpretamos $\mathbf{A}^t$ como a transposta conjugada de $\mathbf{A}$) *positivas definidas* (spd). Uma matriz spd sempre define uma produto interno em $\mathbb{R}^N$ ($\mathbb{C}^N$) e, mais, todo produto interno em $\mathbb{R}^N$ é definido por uma matriz spd, fixada uma base. Alguns outros fatos relevantes a respeito de uma matriz quadrada $\mathbf{A}$ e o conceito de matriz spd [20] são:

- $\mathbf{A}$ é positiva definida sse os seus autovalores são positivos;
- se $\mathbf{A}$ é positiva definida, então seus elementos diagonais são todos positivos;
- se $\mathbf{A}$ é positiva definida, sua inversa $\mathbf{A}^{-1}$ também o é;
- se $\mathbf{A}$ é simétrica e diagonal-dominante estrita, com elementos diagonais positivos, é positiva definida, fato que decorre imediatamente do Teorema de Gerschgorin [08].

Queremos dar ênfase à definição de matriz redutível, para o que precisamos de matriz de permutação: *matriz de permutação* é toda matriz quadrada tal que o único elemento não-nulo em cada linha e em cada coluna é 1; equivalentemente, $\mathbf{P}$ é uma matriz de permutação sse é obtida da matriz identidade permutando linhas e/ou colunas. O efeito de uma matriz de permutação $\mathbf{P}$ sobre uma matriz quadrada $\mathbf{A}$ é este: se $\mathbf{P}$ é obtida da matriz identidade $\mathbf{I}$ permutando somente linhas (*matriz de permutação-linha*), então $\mathbf{PA}$ é a matriz obtida de $\mathbf{A}$ efetuando nesta a mesma transformação que a feita em $\mathbf{I}$ para obter $\mathbf{P}$, e $\mathbf{AP}^t$ é a matriz obtida de $\mathbf{A}$ aplicando nesta a mesma alteração de ordem das colunas que a feita nas linhas de $\mathbf{I}$ para obter $\mathbf{P}$. Assim, cada uma das matrizes $\mathbf{PAP}^t$ e $\mathbf{P}^t\mathbf{AP}$ é simétrica sse $\mathbf{A}$

[1] Neste trabalho os vetores serão sempre vetores-coluna.

é simétrica. Dizemos que $\mathbf{PAP}^t$ e $\mathbf{P}^t\mathbf{AP}$ são *reordenamentos simétricos* de **A**. É imediato que toda matriz de permutação é ortogonal, isto é, $\mathbf{PP}^t = \mathbf{I}$.

1.4.1. Definição. Dizemos que uma matriz quadrada **A**, com ordem $N \geq 2$, é *redutível* sse existe uma matriz de permutação **P** tal que

$$\mathbf{PAP}^t = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}, \tag{1.7}$$

onde $\mathbf{A}_{11}$ é uma matriz quadrada de ordem R, e $\mathbf{A}_{22}$, uma matriz quadrada de ordem $N - R$, com $1 \leq R < N$. Caso tal matriz de permutação não exista, dizemos que **A** é *não-redutível* ou *irredutível*.

Uma aplicação imediata do conceito de matriz redutível é a seguinte. Suponhamos que uma matriz **A** seja redutível e já reduzida à forma (1.7). Então o SELAS $\mathbf{Ax} = \mathbf{b}$, ou

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

pode ser escrito

$$\begin{aligned} \mathbf{A}_{11}\mathbf{x}_1 + \mathbf{A}_{12}\mathbf{x}_2 &= \mathbf{b}_1 \\ \mathbf{A}_{22}\mathbf{x}_2 &= \mathbf{b}_2, \end{aligned}$$

forma que nos traz considerável economia na solução do SELAS.

Na maioria das vezes é difícil ver se uma matriz é ou não redutível. O recurso dos grafos nos ajuda a decidir. Dada uma matriz quadrada $\mathbf{A} = [a_{ij}]$, de ordem N, associamos a ela N pontos distintos $P_1, P_2, \ldots, P_N$, ditos *nodos* ou *vértices*, que podemos posicionar de maneira arbitrária no plano. Para cada $a_{ij} \neq 0$, traçamos uma seta (retilínea ou curvilínea) de P_i a P_j, dita *aresta*, que é identificada ao par ordenado (i, j). A representação de uma aresta (i, i), quando $a_{ii} \neq 0$, é feita por um laço, que muitas vezes omitimos. A figura resultante ou, mais propriamente, o par ordenado $G := G(\mathbf{A}) := (V, E)$, em que V é o conjunto dos vértices e E, o conjunto das arestas, é dito *o grafo orientado* de **A**.

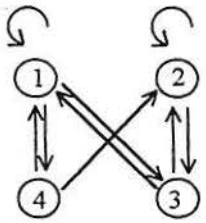
Notemos que um grafo orientado não depende dos valores dos elementos não-nulos da matriz. Logo, ao discutir o grafo de uma matriz **A**, basta considerar a, assim chamada, *matriz booleana* associada a **A**, que tem somente elementos 1 ou 0 (1 na posição (i, j) sse $a_{ij} \neq 0$). Além disso, é logo visto que, para um reordenamento simétrico $\mathbf{B} := \mathbf{PAP}^t$ de **A**, os grafos de **A** e **B** diferem apenas pela ordem dos nodos, e diferir pela ordem dos nodos é irrelevante.

Um grafo orientado é dito *fortemente conexo* sse, para cada par ordenado (P_i, P_j) de nodos, com $i \neq j$, existe um *caminho orientado* $\{(i_0, i_1), (i_1, i_2), \ldots, (i_{r-1}, i_r)\}$ com $i_0 = i$ e $i_r = j$.

1.4.2. Exemplo. A matriz

$$\mathbf{A} := \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

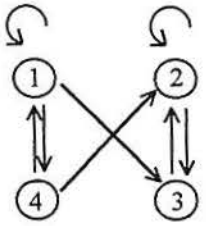
tem o seguinte grafo orientado, que é fortemente conexo:



1.4.3. Exemplo. A matriz

$$\mathbf{A} := \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

tem o grafo orientado



Esse grafo não é fortemente conexo, porque não há caminho orientado de P_3 para P_1, por exemplo.

1.4.4. Teorema. *Uma matriz é irredutível se e somente se o seu grafo é fortemente conexo.*

Prova. ($\Rightarrow$) Seja **P** uma matriz de permutação tal que valha (1.7). Claramente não há caminho orientado que inicie em nodos correspondentes a $\mathbf{A}_{22}$ e termine em nodos correspondentes a $\mathbf{A}_{11}$. Logo $G(\mathbf{PAP}^t)$ não é fortemente conexo.

($\Leftarrow$) Seja $S(\mathbf{A})$ um subgrafo de **A**, constituído apenas de uma componente fortemente conexa de $G(\mathbf{A})$. Reenumeremos as linhas (e colunas) de **A** de maneira que todos os nodos em $S(\mathbf{A})$ sejam os primeiros. Se **P** é a matriz de permutação que opera essa reenumeração, então $\mathbf{PAP}^t$ será da forma (1.7). □

Notemos que segue do Teorema 1.4.4 que a matriz do Exemplo 1.4.3, embora aparentemente redutível, é, de fato, irredutível.

1.4.5. Exemplo. Um exemplo importante de matriz irredutível é uma matriz tridiagonal, cuja matriz booleana associada é

$$\mathbf{A} := \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ & & & \vdots & \\ 0 & 0 & \cdots & 1 & 1 & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 \end{bmatrix}.$$

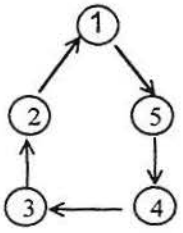
Seu grafo é este

O conceito de grafo orientado de uma matriz pode ser generalizado. Seja $\mathbf{A} := [\mathbf{A}_{ij}]$ uma matriz de blocos: construímos o *grafo orientado generalizado* G_g de $\mathbf{A}$ com o mesmo procedimento que para o grafo usual de $\mathbf{A}$; uma aresta $(i, j) \in G_g$ sse o bloco $\mathbf{A}_{ij} \neq \mathbf{0}$. Assim o grafo orientado generalizado de uma matriz coincide com o grafo orientado, quando os blocos são escalares.

1.4.6. Exemplo. O grafo generalizado da matriz

$$\begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} & \mathbf{0} \end{bmatrix},$$

é este



Após termos associado os conceitos de matriz redutível a seu grafo, vamos definir *grafo cíclico de índice p*, que é muito útil para lidar com SELAS oriundos da discretização de problemas de contorno que envolvem certo tipo de equações diferenciais parciais.

Comprimento de um caminho orientado é o número de arestas que ele contém. Para exemplificar, no Exemplo 1.4.2, os caminhos orientados {(1,1)}, {(1,3), (3,2)} e {(1,3), (3,2), (2,1)} têm, respectivamente, comprimentos 1, 2 e 3. Além disso, o primeiro e o último são também *caminhos fechados*, porque começam e terminam no mesmo nodo. Notemos que {(1,3), (3,2), (2,1), (1,3), (3,2), (2,1)}, {(1,3), (3,2), (2,1), (1,3), (3,2), (2,1), (1,3), (3,2), (2, 1)}, etc. são também caminhos fechados de comprimentos 6, 9, etc., do grafo nesse exemplo.

1.4.7. Definição Seja $G(\mathbf{A})$ um grafo fortemente conexo e p o máximo divisor comum de todos os comprimentos dos caminhos orientados fechados de $G(\mathbf{A})$. Então $G(\mathbf{A})$ é dito um *grafo cíclico de índice p* sse $p > 1$, e um *grafo primitivo* sse $p=1$.

Uma conseqüência dessa definição é que, se uma matriz tem algum elemento não-nulo na diagonal, seu grafo é primitivo. As matrizes dos Exemplos 1.4.2 e 1.4.3 têm ambas grafos primitivos e

o grafo de toda matriz não-nula com diagonal nula é cíclico de algum índice.

Definição 1.4.8. Uma matriz quadrada $\mathbf{A} = [\mathbf{A}_{ij}]$ de blocos é chamada *fracamente cíclica de índice p*, $p > 1$, sse existe uma matriz de permutação **P** tal que

$$\mathbf{PAP}^t = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{1p} \\ \mathbf{A}_{21} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{32} & & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_{pp-1} & \mathbf{0} \end{bmatrix},$$

onde as matrizes nulas na diagonal são quadradas. Essa forma de matriz é dita a *forma normal* de uma matriz fracamente cíclica de índice *p*.

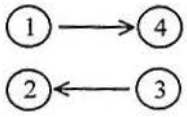
A importância das matrizes fracamente cíclicas é grande para os métodos iterativos, como veremos. Por isso, o teorema seguinte, destinado a identificar essa matrizes, é de grande valia.

1.4.9.Teorema. *Se o grafo* G(**A**) = (V, E) *de uma matriz* **A** *se torna cíclico de índice* p *e fortemente conexo pela identificação de nodos* i, j ∈ V tais que (i, j) ∉ E, *ou acréscimo de arestas* (i, j) *a* E, *então* **A** *é fracamente cíclica de índice* p.

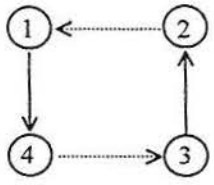
Antes de demonstrar esse teorema, exemplifiquemos seu significado e uso.

1.4.10. Exemplo. Seja a matriz **A** e seu grafo $G(\mathbf{A}) = (V, E)$,

$$\mathbf{A} := \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$



1) Evidentemente essa matriz é fracamente cíclica de índice 4. Também concluímos isso pela aplicação do Teorema 1.4.9, acrescentando a E as arestas (2, 1) e (4, 3):



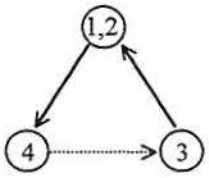
Aqui **P** é a identidade e o particionamento é o que produz todos os blocos com ordem 1×1.

2) Identificando os nodos 1 e 2 e os nodos 3 e 4, obtemos o um grafo cíclico (gereralizado) de índice 2 de **A**:

Logo, pelo Teorema 1.4.9, **A** é também fracamente cíclica de índice 2. A correspondente partição de **A** é esta (a matriz de permutação **P** usada é a identidade):

$$\mathbf{A}_4 = \mathbf{A} = \left[\begin{array}{cc:cc} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \hdashline 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

3) Identificando os nodos 1 e 2 e adicionando a aresta (4, 3), obtemos um grafo cíclico de índice 3:



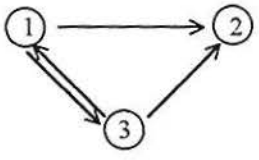
Então a matriz **A** é igualmente fracamente cíclica de índice 3. Para verificar isso com a matriz **A**, fazemos a partição

$$\mathbf{P}_3 = \mathbf{PAP}^t = \left[\begin{array}{c:c:c:c} 0 & 0 & 0 & 1 \\ \hdashline 0 & 0 & 0 & 0 \\ \hdashline 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]. \square$$

1.4.11. Exemplo. A matriz

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix},$$

tem para grafo



que *não* pode ser tornado fortemente conexo e cíclico. □

Prova do Teorema 1.4.9. Claramente, sempre podemos tornar $G(\mathbf{A})$ fortemente conexo pela adição de arestas e identificação de nodos não ligados por arestas. Suponhamos que isso tenha sido feito, tendo o grafo ficado cíclico de índice p, e seja $\{V_1, V_2, \ldots, V_M\}$ o conjunto dos nodos de $G(\mathbf{A})$. Consideremos todos os caminhos orientados fechados que iniciam em V_1 e associemos, começando em V_1, ordenadamente, números 1, 2, 3, . . . aos nodos, pelos quais esses caminhos passam. A cada nodo

ficam associados infinitos números, porque por ele passam infinitos caminhos. Indicamos com C_j, $j = 1, 2, \ldots p$, o conjunto dos nodos aos quais está associado o mesmo número (mod p).

Demonstremos que os conjuntos C_j são disjuntos. Suponhamos que exista $V_k \in C_r \cap C_s$, onde $s \neq r$. Porque $V_k \in C_r$, existe um caminho de V_1 a V_k de comprimento $r - 1$ (mod p) e outro caminho de V_k a V_1 de comprimento $p - r + 1$ (mod p). Porque $V_k \in C_s$, também existe um caminho de V_1 a V_k de comprimento $s - 1$ (mod p). O caminho fechado que inicia em V_1, passa por V_k e termina em V_1 tem comprimento $(s - 1) + (p - r + 1)$ (mod p) $= s - r$ (mod p), portanto não-nulo (mod p), o que contraria a hipótese de $G(\mathbf{A})$ ser cíclico de índice p. Logo os conjuntos C_j são disjuntos.

Agora, dos nodos dos conjuntos C_j, $j = 1, 2, \ldots, p$ saem setas apenas para nodos dos conjuntos $C_{j+1 \pmod p}$, pois se sai uma seta de $V_r \in C_j$ para $V_s \in C_k$, então o caminho orientado fechado que começa em V_1, passa por V_r e por V_s e termina em V_1, tem comprimento $(j - 1) + 1 + (p - k + 1)$ (mod p), que é diferente de zero (mod p), se k (mod p) $\neq j + 1$ (mod p).

Segue que a matriz de permutação **P**, obtida da matriz identidade pela seguinte permuta de linhas,

$$(1, 2, \cdots, N-1, N) \mapsto (I_p, I_{p-1}, \cdots, I_1),$$

onde I_j, $j = 1, 2, \ldots, p$, é a seqüência dos índices dos nodos em V_j numa ordem qualquer, transforma **A** ($\mathbf{A} \mapsto \mathbf{PAP}^t$) na forma normal das matrizes fracamente cíclicas. □

A relação de ordem num conjunto de matrizes reais de dimensões fixas que adotaremos é a ordem usual componente a componente: dadas $\mathbf{A} := [a_{ij}]$ e $\mathbf{B} := [b_{ij}]$, escrevemos $\mathbf{A} \leq \mathbf{B}$ sse $a_{ij} \leq b_{ij}$ para todos os i, j e escrevemos $\mathbf{A} < \mathbf{B}$ sse $a_{ij} < b_{ij}$ para todos os i, j; dizemos que **A** é *não-negativa* (*positiva*) sse $a_{ij} \geq 0$ ($a_{ij} > 0$).

1.5. Decomposição de uma matriz

Consideremos um SELAS como em (1.1). Um método iterativo estacionário se apóia, para resolver esse SELAS, sobre decomposições da matriz **A** dos coeficientes. Genericamente, expressamos uma *decomposição de* **A** na forma

$$\mathbf{A} = \mathbf{C} - \mathbf{R}, \tag{1.8}$$

onde **C** é não-singular. Com a decomposição (1.8), podemos escrever o SELAS (1.1) como $\mathbf{Cx} = \mathbf{Rx} + \mathbf{b}$, de onde surge o *método iterativo básico*

$$\mathbf{Cx}^{(n+1)} = \mathbf{Rx}^{(n)} + \mathbf{b}, \text{ com } n = 0, 1, 2, \ldots,$$

ou, na forma (1.2),

$$\mathbf{x}^{(n+1)} = \mathbf{C}^{-1}\mathbf{Rx}^{(n)} + \mathbf{C}^{-1}\mathbf{b}, \text{ com } n = 0, 1, 2, \ldots \tag{1.9}$$

Claramente, podemos decompor a matriz **A** de muitos modos, nem todos úteis. O critério de utilidade, para decompor a matriz **A**, é tornar os cálculos que visam a encontrar iterativamente a solução do sistema (1.1) o menos custoso possível. Por exemplo, boas escolhas para a matriz **C** são os tipos diagonais ou triangulares (não-singulares).

O estudo da decomposição envolve as chamadas matrizes monótonas: uma matriz quadrada real **A** é *monótona* sse, para todo vetor compatível **x**, $\mathbf{Ax} \geq \mathbf{0}$ implica $\mathbf{x} \geq \mathbf{0}$. O teorema a seguir fornece outra maneira de caracterizar matrizes monótonas.

1.5.1. Teorema. *Uma matriz quadrada real* $\mathbf{A}$ *é monótona* $\Leftrightarrow$ $\mathbf{A}$ *é não singular e* $\mathbf{A}^{-1} \geq \mathbf{0}$.

Prova. ($\Rightarrow$) Fixemos $\mathbf{x} \in R^N$ arbitrariamente, sendo N a ordem de $\mathbf{A}$. Se $\mathbf{Ax} = \mathbf{0}$, então $\mathbf{x} \geq \mathbf{0}$, porque estamos supondo que $\mathbf{A}$ seja monótona; mas $\mathbf{Ax} = \mathbf{0}$ implica $\mathbf{A}(-\mathbf{x}) = \mathbf{0}$; então $-\mathbf{x} \geq \mathbf{0}$, porque $\mathbf{A}$ é monótona; logo $\mathbf{x} \leq \mathbf{0}$. Em resumo, $\mathbf{Ax} = \mathbf{0}$, implica $\mathbf{x} = \mathbf{0}$. Isso prova que $\mathbf{A}$ é não-singular.

Tomemos o $i^{\text{ésimo}}$ vetor $\mathbf{e}_i$ da base canônica de R^N. Temos $\mathbf{AA}^{-1}\mathbf{e}_i = \mathbf{e}_i \geq \mathbf{0}$. Daí concluímos que $\mathbf{A}^{-1}\mathbf{e}_i \geq \mathbf{0}$ para $i = 1, 2, \ldots, N$. Mas $\mathbf{A}^{-1}\mathbf{e}_i$ é a $i^{\text{ésima}}$ coluna de $\mathbf{A}^{-1}$ e, portanto, $\mathbf{A}^{-1} \geq \mathbf{0}$.

($\Leftarrow$) Trivial. □

Uma decomposição $\mathbf{A} = \mathbf{C} - \mathbf{R}$ é dita

- *regular* sse $\mathbf{C}$ é monótona e $\mathbf{R} \geq \mathbf{0}$;
- *regular fraca* sse $\mathbf{C}$ é monótona e $\mathbf{C}^{-1}\mathbf{R} \geq \mathbf{0}$;
- *não-negativa* sse $\mathbf{C}^{-1}$ é não-singular e $\mathbf{C}^{-1}\mathbf{R} \geq \mathbf{0}$;
- *convergente* sse $\mathbf{C}$ é não-singular e $\rho(\mathbf{C}^{-1}\mathbf{R}) < 1$, isto é, o método (1.9) é convergente.

A definição de decomposição regular foi introduzida por [50] em 1962, de regular fraca, por [35] em 1970, e de não-negativa, por [47] em 1991.

1.5.2. Exemplo. Seja a matriz $\mathbf{A} = \begin{bmatrix} 0.7 & -0.7 \\ -0.3 & 0.4 \end{bmatrix}$, que é monótona, pois $\det(A) = 0.07 \neq 0$ e $\mathbf{A}^{-1} = \begin{bmatrix} 40/7 & 10 \\ 30/7 & 10 \end{bmatrix} > \mathbf{0}$. Efetuemos a decomposição

$$\mathbf{A} = \mathbf{C} - \mathbf{R} = \begin{bmatrix} 1 & -1 \\ 0 & 0.8 \end{bmatrix} - \begin{bmatrix} 0.3 & -0.3 \\ 0.3 & 0.4 \end{bmatrix}.$$

A matriz $\mathbf{C}$ é não-singular, pois $\det(C) = 4/5$ e

$$\mathbf{C}^{-1}\mathbf{R} = \begin{bmatrix} 40/7 & 10 \\ 30/7 & 10 \end{bmatrix},$$

cujos autovalores são 7/8 e 3/10; logo $\rho(\mathbf{C}^{-1}\mathbf{R}) = 7/8 < 1$, e nossa decomposição é convergente. □

O resultado geral seguinte foi enunciado sem demonstração por Beauwes [04], demonstrado por Axelsson em 1996 [01] e estende um teorema clássico, que data de 1961, devido a Varga [51].

Teorema. Seja $\mathbf{A} = \mathbf{C} - \mathbf{R}$ uma decomposição não-negativa da matriz $\mathbf{A}$. Coloquemos $\mathbf{B} := \mathbf{C}^{-1}\mathbf{R}$. São equivalentes as seguintes afirmações:

- *essa decomposição é convergente, isto é,* $\rho(\mathbf{B}) < 1$;
- $\mathbf{I} - \mathbf{B}$ *é monótona*;
- $\mathbf{A}$ *é não-singular e* $\mathbf{G} := \mathbf{A}^{-1}\mathbf{R} \geq \mathbf{0}$;
- $\mathbf{A}$ *é não-singular e* $\rho(\mathbf{A}) = \dfrac{\rho(\mathbf{G})}{\mathbf{1} + \rho(\mathbf{G})}$.

Tiramos duas conseqüências relevantes do Teorema 1.5.3.

1.5.4. Corolário. *Se* $\mathbf{A} = \mathbf{C} - \mathbf{R}$ *é uma decomposição regular fraca, então essa decomposição é convergente* $\Leftrightarrow$ $\mathbf{A}$ *é monótona.*

Prova. ($\Rightarrow$) Pondo $\mathbf{B} := \mathbf{C}^{-1}\mathbf{R}$, como $\rho(\mathbf{B}) < 1$, pelo Teorema 1.5.3, $\mathbf{I} - \mathbf{B}$ é monótona e daí segue que $\mathbf{A}^{-1} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{C}^{-1} \geq \mathbf{0}$.

($\Leftarrow$) A hipótese implica $(\mathbf{I} - \mathbf{B})\mathbf{A}^{-1} = \mathbf{C}^{-1}$. Então, para $n = 0, 1, 2, \ldots$,

$$\begin{aligned}(\mathbf{I} + \mathbf{B} + \cdots + \mathbf{B}^n)\mathbf{C}^{-1} &= (\mathbf{I} + \mathbf{B} + \cdots + \mathbf{B}^n)(\mathbf{I} - \mathbf{B})\mathbf{A}^{-1} \\ &= (\mathbf{I} - \mathbf{B}^{n+1})\mathbf{A}^{-1}.\end{aligned}$$

Com base nesse resultado e no fato de que $\mathbf{B}, \mathbf{A}^{-1} \geq \mathbf{0}$, temos, para todo vetor positivo $\mathbf{v}$ e todo n,

$$\begin{aligned}(\mathbf{I} + \mathbf{B} + \cdots + \mathbf{B}^n)\mathbf{C}^{-1}\mathbf{v} &= (\mathbf{I} - \mathbf{B}^{n+1})\mathbf{A}^{-1}\mathbf{v} \\ &= \mathbf{A}^{-1}\mathbf{v} - \mathbf{B}^{n+1}\mathbf{A}^{-1}\mathbf{v} \\ &\leq \mathbf{A}^{-1}\mathbf{v}.\end{aligned}$$

Isso mostra que a série $\mathbf{I} + \mathbf{B} + \cdots + \mathbf{B}^n + \cdots$ de matrizes converge porque $\mathbf{C}^{-1}\mathbf{v} \geq \mathbf{0}$. Logo $\mathbf{B}^n \to \mathbf{0}$ e, portanto, pelo Teorema 1.3.1, $\rho(\mathbf{B}) < 1$. □

1.5.5. Corolário. *Se* $\mathbf{A} = \mathbf{C} - \mathbf{R}$ *é uma decomposição regular fraca e* $\mathbf{A}$ *é monótona, então essa decomposição é convergente, isto é, a seqüência* $(\mathbf{x}^{(n)})$ *das iterações* $\mathbf{x}^{(n+1)} = \mathbf{C}^{-1}\mathbf{R}\mathbf{x}^{(n)} + \mathbf{C}^{-1}\mathbf{b}$, $n = 0, 1, 2, \ldots$, *converge para a solução do* SELAS $\mathbf{Ax} = \mathbf{b}$ *para todo vetor inicial* $\mathbf{x}^{(0)}$.

Prova. Segue imediatamente do Teorema 1.5.3 e do Corolário 1.3.4. □

1.6. Convergência de um método iterativo

Voltemos brevemente para o estudo da convergência em geral. Consideremos o método iterativo geral (1.2), isto é, considerado o SELAS na forma $\mathbf{x} = \mathbf{Mx} + \mathbf{c}$, escrevemos a seqüência das iterações

$$\mathbf{x}^{(n+1)} = \mathbf{Mx}^{(n)} + \mathbf{c}, \quad n = 0, 1, 2, \ldots,$$

onde a matriz de iteração $\mathbf{M}$ tem ordem N. Se $\mathbf{I} - \mathbf{M}$ é não-singular, sendo $\mathbf{I}$ a matriz identidade de mesma ordem de $\mathbf{M}$, temos uma única solução para o SELAS

$$(\mathbf{I} - \mathbf{M})\mathbf{x} = \mathbf{c}$$

e, se o vetor-erro é definido para as iterações sucessivas por

$$\mathbf{e}^{(n)} := \mathbf{x}^{(n)} - \mathbf{x},$$

então

$$\mathbf{e}^{(n)} = \mathbf{Me}^{(n-1)} = \ldots = \mathbf{M}^n\mathbf{e}^{(0)},$$

e, daí,

$$\left\|\mathbf{e}^{(n)}\right\| \leq \left\|\mathbf{M}^n\right\| \cdot \left\|\mathbf{e}^{(0)}\right\|.$$

Caso $\mathbf{e}^{(0)}$ não seja o vetor nulo,

$$\frac{\left\|\mathbf{e}^{(n)}\right\|}{\left\|\mathbf{e}^{(0)}\right\|} \leq \left\|\mathbf{M}^n\right\|.$$

Portanto $\left\|\mathbf{M}^n\right\|$ nos fornece uma cota superior otimizada da razão à esquerda dessa desigualdade, para n iterações, e nos servirá como base de comparação de diferentes métodos iterativos.

Definições 1.6.1. Seja **M** uma matriz quadrada . Se, para algum n inteiro positivo, $\left\|\mathbf{M}^n\right\| < 1$, definimos a *razão média de convergência por iteração para n iterações* da matriz **M** como

$$R(\mathbf{M}^n) := -\ln\left(\left\|\mathbf{M}^n\right\|^{1/n}\right) = -\frac{\ln\left\|\mathbf{M}^n\right\|}{n}.$$

Para duas matrizes quadradas **M** e **N**, se $R(\mathbf{M}^n) < R(\mathbf{N}^n)$, então dizemos que **N** é *iterativamente mais rápida para n iterações que* **M**.

Para entendermos o significado computacional da razão média de convergência $R(\mathbf{M})$, introduzimos o seguinte valor:

$$\varphi := \left(\frac{\left\|\mathbf{e}^{(n)}\right\|}{\left\|\mathbf{e}^{(0)}\right\|}\right)^{1/n},$$

que é o *fator de redução médio por iteração*, para n iterações. Se $\|\mathbf{M}^n\| < 1$, então, pela definição, vem

$$\varphi \leq \|\mathbf{M}^n\|^{1/n} = e^{-R(\mathbf{M}^n)}.$$

Isso significa que $R(\mathbf{M}^n)$ é a taxa de decaimento exponencial para uma cota superior otimizada da redução média φ do erro por iteração, nesse processo iterativo de n passos. Pondo

$$N_n := R(\mathbf{M}^n)^{-1},$$

e usando a desigualdade anterior, vem

$$\varphi^{N_n} \leq \frac{1}{e}.$$

Isso mostra que N_n é uma medida do número de iterações necessárias para reduzir a norma do vetor-erro inicial de um fator e.

Até o presente momento não levamos em consideração o raio espectral das matrizes de iteração, para comparar a convergência dos métodos iterativos. Consideremos duas matrizes **M** e **N**. Vale, para todo $m = 1, 2, \ldots$,

$$\left\|\mathbf{M}^m\right\| = \rho(\mathbf{M})^m,$$

e, portanto, se

$$\rho(\mathbf{M}) < \rho(\mathbf{N}) < 1,$$

então

$$\left\|\mathbf{M}^m\right\| < \left\|\mathbf{N}^m\right\| < 1, \quad \text{para todo } m = 1, 2, \ldots$$

Infelizmente, embora tenhamos $\left\|\mathbf{M}^n\right\| \to 0$, para qualquer matriz convergente **M**, não impor-

tando qual seja a norma matricial, a norma euclidiana $\|\mathbf{M}^n\|$ pode tornar-se muito errática. Em termos das Definições 1.6.1, é possível uma matriz $\mathbf{M}$ ser iterativamente mais rápida que uma matriz $\mathbf{N}$ para m iterações, mas iterativamente mais lenta para $n \neq m$ iterações, como mostra o exemplo seguinte.

1.6.2. Exemplo. Tomemos as matrizes

$$\mathbf{M} := \begin{bmatrix} 0.98 & 4 \\ 0 & 0.98 \end{bmatrix}, \quad \mathbf{N} := \begin{bmatrix} 0.98 & 0 \\ 0 & 0.99 \end{bmatrix}.$$

Usando o MATLAB, obtemos a Tab.1.1 com diversos valores de m e normas euclidianas. Para valores pequenos de m, temos $\|\mathbf{M}^m\| > \|\mathbf{N}^m\|$, mas, para valores grandes, observamos que $\|\mathbf{M}^m\| < \|\mathbf{N}^m\|$, monotonamente, com diferenças relativas cada vez maiores, à medida que $m \to \infty$. De outro modo: parece inicialmente que $\|\mathbf{N}^m\| \to \mathbf{0}$ mais rapidamente que $\|\mathbf{M}^m\| \to \mathbf{0}$, mas verificamos o oposto em definitivo. □

m	$\|\mathbf{M}^m\|$	$\|\mathbf{N}^m\|$	m	$\|\mathbf{M}^m\|$	$\|\mathbf{N}^m\|$	m	$\|\mathbf{M}^m\|$	$\|\mathbf{N}^m\|$
1	4.2272	0.9900	90	59.6247	0.4047	900	$4.6617\text{x}10^{-5}$	$1.1794\text{x}10^{-4}$
2	7.9559	0.9801	100	54.1308	0 .3660	950	$1.7920\text{x}10^{-5}$	$7.1357\text{x}10^{-5}$
3	11.6012	0.9703	150	29.5691	0.2215	1000	$6.8693\text{x}10^{-6}$	$4.3171\text{x}10^{-5}$
4	15.1154	0.9606	200	14.3575	0.1340	1100	$1.0021\text{x}10^{-7}$	$1.5802\text{x}10^{-5}$
5	18.4915	0.9510	250	6.5357	0.0811	1200	$1.4498\text{x}10^{-8}$	$5.7811\text{x}10^{-6}$
6	21.7302	0.9415	300	2.8561	0.0490	1300	$2.0829\text{x}10^{-8}$	$2.1172\text{x}10^{-6}$
7	24.8339	0.9321	350	1.2135	0.0297	1400	$2.9749\text{x}10^{-9}$	$7.7495\text{x}10^{-7}$
8	27.8060	0.9227	400	0.5050	0.0180	1500	$4.2271\text{x}10^{-10}$	$2.8366\text{x}10^{-7}$
9	30.6501	0.9133	450	0.2069	0.0109	2000	$2.3121\text{x}10^{-14}$	$1.8638\text{x}10^{-9}$
10	33.3699	0.9044	500	0.0837	0.0066	3000	$5.8369\text{x}10^{-23}$	$8.0461\text{x}10^{-14}$
20	54.5068	0.8179	550	0.0335	0.0040	4000	$1.3098\text{x}10^{-31}$	$3.4756\text{x}10^{-18}$
30	66.7985	0.7397	600	0.0133	0.0024	5000	$2.7554\text{x}10^{-40}$	$1.4996\text{x}10^{-22}$
40	72.7701	0.6690	650	0.0053	0.0015	6000	$5.5647\text{x}10^{-49}$	$6.4739\text{x}10^{-27}$
50	74.3221	0.6050	700	0.0021	$8.8031\text{x}10^{-4}$	7000	$1.0926\text{x}10^{-57}$	$2.7949\text{x}10^{-31}$
60	72.8714	0.5472	750	$8.0436\text{x}10^{-4}$	$5.3259\text{x}10^{-4}$	8000	$2.1015\text{x}10^{-66}$	$1.2066\text{x}10^{-35}$
70	69.4644	0.4948	800	$3.1245\text{x}10^{-4}$	$3.2222\text{x}10^{-4}$	9000	$3.9789\text{x}10^{-75}$	$5.2090\text{x}10^{-40}$
80	64.8655	0.4475	850	$1.2090\text{x}10^{-4}$	$1.9495\text{x}10^{-4}$	10000	$7.4403\text{x}10^{-84}$	$2.2488\text{x}10^{-44}$

Tab.1.1 – Comparação de velocidades de convergência intermediárias e definitivas

Concluímos que, para obter uma medida que nos dê com precisão a rapidez da convergência, devemos considerar $m \to \infty$, o que nos leva à seguinte definição.

Definição 1.6.3. Chamamos de *razão de convergência assintótica* [51], ou simplesmente *razão de convergência* [55] de uma matriz quadrada convergente $\mathbf{M}$, ou do método iterativo (1.2) que ela define, ao número positivo

$$R_\infty(\mathbf{M}) := \lim_{n\to\infty} R(\mathbf{M}^n).$$

Varga [51] demonstra que o limite nessa definição existe e vale $-\ln\rho(\mathbf{M})$. Demos ênfase a esse fato:

$$R_\infty(\mathbf{M}) = -\ln\rho(\mathbf{M}).$$

1.6.4. Exemplo. Para as matrizes do Exemplo 1.6.2,

$$\rho(\mathbf{M}) = 0.98 \quad \text{e} \quad \rho(\mathbf{N}) = 0.99.$$

Então, dentro da precisão de máquina,

$$R_\infty(\mathbf{M}) = 2.020270731751947 \times 10^{-2} \quad \text{e} \quad R_\infty(\mathbf{N}) = 1.005033585350145 \times 10^{-2},$$

o que mostra que a matriz **M** converge, em definitivo, duas vezes mais rapidamente que **N**. □

Como, para toda matriz quadrada **M**,

$$\rho(\mathbf{M})^m \le \left\|\mathbf{M}^m\right\|, \text{ para todo } m = 1, 2, \ldots,$$

temos, para toda matriz (complexa) quadrada e todo número natural m tal que $\left\|\mathbf{M}^m\right\| < 1$,

$$R(\mathbf{M}^m) \le R_\infty(\mathbf{M}).$$

Se tivéssemos usado a base decimal de logaritmos para definir razão de convergência, como fazem alguns autores [01, 14], $R_\infty(\mathbf{M})$ seria o ganho em número de casas decimais corretas por iteração na solução do SELAS.

1.6.5. Critérios de parada. Antes de entrar nos métodos específicos, escrevamos algumas linhas sobre critérios de parada de um método iterativo. Quando usamos métodos iterativos, a solução exata na maioria dos casos não é atingida. Então precisaremos de algum critério para findar o processo, escolhendo como solução aproximada o último termo da seqüência de iteração ($\mathbf{x}^{(n)}$) calculado. Quando ocorre a convergência, a seqüência de iteração é de Cauchy, e, portanto, $\mathbf{x}^{(n+1)}$ está mais próximo de $\mathbf{x}^{(n)}$ que este termo de $\mathbf{x}^{(n-1)}$, e isso equivale a dizer que $\mathbf{x}^{(n+1)}$ é uma aproximação melhor da solução exata **x** do que $\mathbf{x}^{(n)}$. Então, um critério de parada natural é dado em termos das distâncias sucessivas relativas dos termos dessa seqüência: parar as iterações quando

$$\frac{\left\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\right\|}{\left\|\mathbf{x}^{(n)}\right\|} \le \varepsilon, \tag{1.10}$$

ou parar quando o número das iterações excede um número preestabelecido. Aqui $\varepsilon > 0$ é preestabelecido de acordo com a precisão desejada.

Outro critério de parada pode ser dado em termos dos resíduos: parar as iterações quando

$$\left\|\mathbf{A}\mathbf{x}^{(n+1)} - \mathbf{b}\right\| \le \varepsilon\left(\left\|\mathbf{A}\right\| \cdot \left\|\mathbf{x}^{(n+1)}\right\| + \left\|\mathbf{b}\right\|\right), \tag{1.11}$$

ou quando o número de iterações excede um predeterminado número. A tolerância ε é escolhida de maneira que seja $\mu < \varepsilon < 1$, com μ igual à precisão de máquina.

1.7. Métodos iterativos de Jacobi, Gauss-Seidel e SOR com matrizes de escalares

Nesta seção abordaremos brevemente os métodos clássicos de Jacobi, Gauss-Seidel e SOR, para resolver SELAS em que a matrizes dos coeficientes é considerada como matriz de escalares, isto é, não particionada em blocos.

Recordamos que estamos supondo que a matriz **A** (real ou complexa) dos coeficientes de um SELAS (1.1) seja quadrada e não-singular, de ordem N. Supomos ainda não-nulos os elementos diagonais de **A**.

A decomposição da matriz **A** será do tipo

$$\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U},$$

onde **D** é a matriz diagonal de **A**, **L** é a parte triangular inferior estrita de **A** e **U** é a parte triangular superior estrita de **A**.

1.7.1. Método iterativo de Jacobi. O método de Jacobi escreve o SELAS (1.1) na forma de um problema linear de ponto fixo,

$$\mathbf{Dx} = (\mathbf{L}+\mathbf{U})\mathbf{x} + \mathbf{b},$$

ou

$$\mathbf{x} = \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b}\,. \tag{1.12}$$

A matriz

$$\mathbf{B}_J := \mathbf{D}^{-1}(\mathbf{D}+\mathbf{U})$$

é chamada de *matriz de iteração de Jacobi* e o vetor

$$\mathbf{b}_J := \mathbf{D}^{-1}\mathbf{b}$$

é chamado de *vetor de Jacobi*. Dessa maneira podemos reescrever (1.12), junto com a *seqüência das iterações de Jacobi*, do seguinte modo

$$\mathbf{x} = \mathbf{B}_J\mathbf{x} + \mathbf{b}_J,\ \ \mathbf{x}^{(n+1)} = \mathbf{B}_J\mathbf{x}^{(n)} + \mathbf{b}_J, \tag{1.13}$$

por onde vemos que o método de Jacobi é do tipo (1.2). Ainda, podemos escrever o processo de Jacobi na forma escalar ($\mathbf{A} = [a_{ij}]$, $\mathbf{x} = [x_i]$, $\mathbf{b} = [b_i]$),

$$x_i^{(n+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}x_j^{(n)} \right),\quad i = 1, 2, \cdots, N. \tag{1.14}$$

Pelo Teorema 1.3.3, o método de Jacobi (1.13) converge, com uma escolha inicial arbitrária $\mathbf{x}^{(0)}$, sse o raio espectral da matriz de Jacobi $\rho(\mathbf{B}_J) < 1$.

O algoritmo do método de Jacobi é, de fato, simples, como também o é sua implementação no MATLAB. Na implementação no MATLAB (Apêndice *B*) usamos um critério de parada do tipo (1.11) simplificado e a norma-euclidiana. Os dados de entrada são a matriz **A** dos coeficientes, o vetor coluna **b** dos termos independentes, um vetor $\mathbf{x}^{(0)}$ inicializador do processo iterativo, a tolerância *tol*, e o número máximo *nmax* de iterações.

1.7.1.1 Algoritmo para o método de Jacobi

Entrada: **A**, **b**, $\mathbf{x}^{(0)}$, tol, *n*max

Para $n = 0, 1, 2, 3, \ldots$, iterar até que o critério de parada fique satisfeito;

$$x_i^{(n+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}x_j^{(n)} \right), \quad i = 1, 2, \ldots, N;$$

Se o critério de parada está satisfeito, então $\mathbf{x} = \mathbf{x}^{(n+1)}$.

Ao invés da forma escalar (1.14) das iterações, poderíamos usar no algoritmo a forma matricial (1.13). Bastaria introduzir um dispositivo que calcule a matriz de Jacobi e o vetor de Jacobi.

Notemos que a ordem, segundo a qual as equações do SELAS são processadas no método de Jacobi é irrelevante, uma vez que são processadas independentemente. Por isso, o método de Jacobi é conhecido também como o *método dos deslocamentos simultâneos*, pois as atualizações podem ser feitas simultaneamente.

1.7.2. Método de Gauss-Seidel. O método de Gauss-Seidel é semelhante ao de Jacobi, exceto na maneira da atualização das soluções aproximadas em cada passo iterativo: o método de Jacobi usa todas as componentes de $(\mathbf{x}^{(n)})$ (e somente estas) para calcular as componentes de $(\mathbf{x}^{(n+1)})$; o de Gauss-Seidel, precisa de menor armazenamento, pois, para calcular $x_i^{(k+1)}$, usa $x_j^{(k+1)}, j = 1, 2, \ldots, i\text{-}1$, já disponíveis quando do cálculo de $\mathbf{x}^{(n+1)}$:

$$x_i^{(n+1)} = \frac{1}{a_{ii}}\left(b_i - \underbrace{\sum_{j=1}^{i-1} a_{ij}x_j^{(n+1)}}_{\text{atualização de x}} + \underbrace{\sum_{j=i+1}^{N} a_{ij}x_j^{(n)}}_{\text{último x}} \right), \; i = 1, 2, \ldots, N; \; n = 0, 1, 2, \ldots \tag{1.15}$$

A formulação matricial do método de Gauss-Seidel é feita assim (com as notações acima):

$$\mathbf{x}^{(n+1)} = (\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}\mathbf{x}^{(n)} + (\mathbf{L} + \mathbf{D})^{-1}\mathbf{b}, \; n = 0, 1, 2, \ldots$$

Notemos que a matriz $\mathbf{L} + \mathbf{D}$ é a parte triangular inferior da matriz dos coeficientes. As matrizes

$$\mathbf{B}_{GS} := (\mathbf{L} + \mathbf{D})^{-1}\mathbf{U} \;\text{ e }\; \mathbf{b}_{GS} := (\mathbf{L} + \mathbf{D})^{-1}\mathbf{b}$$

são ditas a *matriz de Gauss-Seidel* e o *vetor de Jacobi*, respectivamente. Então a formulação do problema linear $\mathbf{Ax} = \mathbf{b}$, junto com o método de Gauss-Seidel pode ser expressa assim

$$\mathbf{x} = \mathbf{B}_{GS}\mathbf{x} + \mathbf{b}_{GS}, \;\; \mathbf{x}^{(n+1)} = \mathbf{B}_{GS}\mathbf{x}^{(n)} + \mathbf{b}_{GS}, \; n = 0, 1, 2, \ldots$$

Pelo Teorema 1.3.3, o método de Gauss-Seidel (1.15) converge, com uma escolha inicial arbitrária $\mathbf{x}^{(0)}$, sse o raio espectral da matriz de Gauss-Seidel $\rho(\mathbf{B}_{GS}) < 1$.

Pela formulação do método de Gauss-Seidel, é de esperar que a ordem das equações que constituem o SELAS influenciem a convergência. De fato, por exemplo, o método de Gauss-Seidel,

aplicado ao SELAS

$$\begin{cases} 2x_1 + x_2 = 4 \\ x_1 + 2x_2 = 5 \end{cases}$$

converge com essa ordem das equações, mas diverge com a outra ordem. Podemos ver isso, calculando os raios espectrais das matrizes de Gauss-Seidel: para a ordem acima, o raio espectral é ¼ e para a outra ordem, é 4. Para indicar que as iterações no método de Gauss-Seidel dependem da ordem, é às vezes chamado de método dos deslocamentos sucessivos.

Como no método de Jacobi, podemos facilmente construir o algoritmo de Gauss-Seidel.

1.7.2.1 Algoritmo para o método de Gauss-Seidel

Entrada: **A**, **b**, $\mathbf{x}^{(0)}$, tol, *n*max

Para $n = 0, 1, 2, 3, \ldots$, iterar até que o critério de parada fique satisfeito;

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

Se o critério de parada está satisfeito, então $\mathbf{x} = \mathbf{x}^{(n+1)}$.

1.7.3. Método das sobre-relaxações sucessivas. O método de Gauss-Seidel desaponta pela sua lentidão na convergência, quando $\rho(\mathbf{B}_{GS})$ (menor que 1) é próximo de 1. Por isso, introduzimos um parâmetro ω no método Gauss-Seidel, que, se bem escolhido, acelera significativamente a convergência, para certos tipos de matrizes que ocorrem freqüentemente. Esse parâmetro é chamado *parâmetro de relaxação* e o método, resultante da modificação do método de Gauss-Seidel, de *método das sobre-relaxações sucessivas*, ou, abreviadamente, *método* SOR (*successive overrelaxation*).

Multiplicando $\mathbf{Ax} = \mathbf{b}$ por ω, obtemos $\omega\mathbf{Ax} = \omega\mathbf{b}$, e, seguindo o esquema de Gauss-Seidel, resulta a seqüência das iterações do SOR, em termos das coordenadas,

$$x_i^{(n+1)} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(n)} \right) + (1-\omega) x_i^{(n)}, \quad i = 1, 2, \ldots, N; \; n = 0, 1, 2, \ldots \qquad (1.15)$$

Com notação matricial, o processo pode ser expresso na forma

$$\mathbf{x}^{(n+1)} = (\mathbf{D} - \omega\mathbf{L})^{-1}\left((1-\omega)\mathbf{D} + \omega\mathbf{U}\right)\mathbf{x}^{(n)} + \omega(\mathbf{D} - \omega\mathbf{L})^{-1}\mathbf{b}, \; n = 0, 1, 2, \ldots$$

As matrizes

$$\mathbf{B}_{sor} := (\mathbf{D} - \omega\mathbf{L})^{-1}\left((1-\omega)\mathbf{D} + \omega\mathbf{U}\right) \quad \text{e} \quad \mathbf{b}_{sor} := \omega(\mathbf{D} - \omega\mathbf{L})^{-1} \qquad (1.16)$$

são chamadas *matriz do* SOR e *vetor do* SOR, respectivamente. O parâmetro de ponderação ω às vezes é referido com *parâmetro do* SOR. Então podemos escrever o SELAS junto com as iterações do método SOR assim

$$\mathbf{x} = \mathbf{B}_{sor}\mathbf{x} + \mathbf{b}_{sor}, \quad \mathbf{x}^{(n+1)} = \mathbf{B}_{sor}\mathbf{x}^{(n)} + \mathbf{b}_{sor}.$$

Observamos que a escolha $\omega = 1$ particulariza o SOR no método de Gauss-Seidel. Às vezes,

quando $\omega < 1$, falamos em *método de sub-relaxação*, e reservamos a designação de *método de sobre-relaxação* para o caso $1 < \omega$. Neste trabalho unificamos as designações para *método* SOR.

Outra observação interessante: (1.15) mostra que, para cada componente do vetor que está sendo calculado, a iteração atual é um tipo de média ponderada (os pesos são positivos somente quando $\omega < 1$) entre a atualização já computada de Gauss-Seidel e o valor achado na iteração predecessora. A idéia simples do SOR com a ponderação é tentar aproximar mais a parte da solução aproximada computada de Gauss-Seidel da solução exata do SELAS.

Aqui também usamos o Teorema 1.3.3 para concluir que o método SOR converge, com uma escolha inicial arbitrária $\mathbf{x}^{(0)}$, sse o raio espectral da matriz do SOR $\rho(\mathbf{B}_{sor}) < 1$.

O esquema do algoritmo para o método SOR é semelhante aos dos métodos anteriores.

1.7.3.1. Algoritmo para o método SOR[2]

Entrada: **A**, **b**, $\mathbf{x}^{(0)}$, tol, *n*max;

Para $n = 0, 1, 2, 3, \ldots$, iterar até que o critério de parada fique satisfeito;

$$x_i^{(n+1)} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(n)} \right) + (1-\omega) x_i^{(n)}, \qquad i = 1, 2, \ldots, N;$$

Se o critério de parada está satisfeito, então $\mathbf{x} = \mathbf{x}^{(n+1)}$.

1.7.3.2. Escolha de ω para a convergência do método SOR. Estamos particularmente interessados na melhor escolha do parâmetro ω. Este tema vai-nos ocupar preponderantemente neste trabalho. Mas já adiantamos que, em geral, não é possível determinar previamente o valor ótimo de ω em relação à razão de convergência do SOR. Mesmo quando é possível, o custo do cálculo pode ser proibitivo. Por isso, freqüentemente é usada alguma estimativa heurística.

Encontramos também sofisticadas implementações do algoritmo do SOR. Por exemplo, o ITPACK [58] emprega um esquema de estimação adaptável do parâmetro, para tentar ir em direção ao valor apropriado de ω, estimando a razão segundo a qual o processo está convergindo.

Agora provaremos um resultado fácil, mas importante, de William Kahan [28].

Teorema 1.7.3.2.1. *O método* SOR *não converge, seja qual for a aproximação inicial, se* $\omega \notin (0; 2)$.

Prova. Consideremos a matriz do SOR em (1.16). A matriz $(\mathbf{D} - \omega\mathbf{L})^{-1}$ é uma matriz triangular inferior, cujos elementos diagonais são $1/a_{ii}$, $i = 1, 2, \ldots, N$, e $(1-\omega)\mathbf{D} - \omega\mathbf{U}$ é uma matriz triangular superior, cujos elementos diagonais são $(1 - \omega)a_{ii}$, $i = 1, 2, \ldots, N$. Logo

$$\det(\mathbf{B}_{sor}) = (1-\omega)^N .$$

Como o determinante de uma matriz é o produto de seus autovalores, denotando com $\lambda_1, \lambda_2, \ldots, \lambda_N$ os autovalores de $\mathbf{B}_{sor}$,

$$\rho(\mathbf{B}_{sor})^N \geq |\lambda_1||\lambda_2|\cdots|\lambda_n| = |1-\omega|^N .$$

Portanto

---

[2] Os algoritmos 1.7.1.1, 1.7.2.1 e 1.7.3.1 estão implementados em MATLAB no Apêndice *B*, com o nome `jasor`.

$$\rho(\mathbf{B}_{sor}) \geq |1-\omega|.$$

Como o processo não converge se $\rho(\mathbf{B}_{sor}) \geq 1$, então *não* converge se $|1 - \omega| \geq 1$, ou, equivalentemente, se

$$\omega \leq 0 \text{ ou } 2 \leq \omega. \square$$

O seguinte teorema, conhecido como Teorema de Ostrowski-Reich [36], mostra que a condição necessária do Teorema 1.7.3.2.1 é também suficiente quando a matriz **A** é spd.

Teorema 1.7.3.2.2. *Se* **A** *é uma matriz spd e* $0 < \omega < 2$, *então o método SOR converge, seja qual for a escolha inicial* $\mathbf{x}^{(0)}$.

Prova. Seja $\mathbf{e}^{(0)}$ um vetor-erro inicial arbitrário não-nulo. Então a seqüência $(\mathbf{e}^{(n)})$ dos vetores-erro é dada por

$$\mathbf{e}^{(n+1)} := \mathbf{B}_{sor}\mathbf{e}^{(n)}, \quad n = 0, 1, 2, \ldots \tag{1.17}$$

Como **A** é simétrica, podemos escrever a primeira igualdade em (1.16) como

$$\mathbf{B}_{sor} = (\mathbf{D} - \omega\mathbf{L})^{-1}\left((1-\omega)\mathbf{D} + \omega\mathbf{L}^t\right).$$

Então a (1.17) é equivalente à igualdade

$$(\mathbf{D} - \omega\mathbf{L})\mathbf{e}^{(n+1)} = \left((1-\omega)\mathbf{D} + \omega\mathbf{L}^t\right)\mathbf{e}^{(n)}, \quad n = 0, 1, 2, \ldots \tag{1.18}$$

Ponhamos também

$$\mathbf{d}^{(n)} := \mathbf{e}^{(n)} - \mathbf{e}^{(n+1)}, \; n = 0, 1, 2, \ldots$$

Obtemos a partir disso e da (1.18), lembrando que estamos usando a decomposição $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{L}^t$,

$$(\mathbf{D} - \omega\mathbf{L})\mathbf{d}^{(n)} = \omega\mathbf{A}\mathbf{e}^{(n)}, \quad n = 0, 1, 2, \ldots \tag{1.19}$$

Também

$$\omega\mathbf{A}\mathbf{e}^{(n+1)} = (1-\omega)\mathbf{D}\mathbf{d}^{(n)} + \omega\mathbf{L}^t\mathbf{d}^{(n)}, \; n = 0, 1, 2, \ldots$$

Agora multiplicamos à esquerda as duas últimas equações por $\mathbf{e}_n^t$ e $\mathbf{e}_{n+1}^t$, respectivamente, usamos o fato de **A** ser simétrica e combinamos as igualdades resultantes, após algumas manipulações algébricas, na única equação

$$(2-\omega)\left(\mathbf{d}^{(n)}\right)^t\mathbf{D}\mathbf{d}^{(n)} = \omega\left(\left(\mathbf{e}^{(n)}\right)^t\mathbf{A}\mathbf{e}^{(n)} - \left(\mathbf{e}^{(n+1)}\right)^t\mathbf{A}\mathbf{e}^{(n+1)}\right), \quad n = 0, 1, 2, \ldots \tag{1.20}$$

Agora particularizemos $\mathbf{e}^{(0)}$ para um autovetor de $\mathbf{B}_{sor}$ associado a um autovalor λ. Então $\mathbf{e}^{(1)} = \mathbf{B}_{sor}\mathbf{e}^{(0)} = \lambda\mathbf{e}^{(0)}$ e $\mathbf{d}^{(0)} = (1 - \lambda)\mathbf{e}^{(0)}$. Usando isso em (1.20), vem

$$\left(\frac{2-\omega}{\omega}\right)|1-\lambda|^2\left(\mathbf{e}^{(n)}\right)^t\mathbf{D}\mathbf{e}^{(n)} = \left(1-|\lambda|^2\right)\left(\mathbf{e}^{(0)}\right)^t\mathbf{A}\mathbf{e}^{(0)}. \tag{1.21}$$

Notemos que necessariamente $\lambda \neq 1$, caso contrário seria $\mathbf{d}^{(0)} = \mathbf{0}$, e daí e de (1.19) resultaria $\mathbf{A}\mathbf{e}^{(0)} = \mathbf{0}$,

o que contraria a hipótese de **A** ser spd. Como $0 < \omega < 2$, o primeiro membro de (1.21) é positivo (**D** é spd), e, sendo $(\mathbf{e}^{(0)})^t\mathbf{A}\mathbf{e}^{(0)} > 0$, obtemos finalmente $|\lambda| < 1$, pelo que o método SOR converge. □

Observamos que ambos os Teoremas 1.7.3.2.1 e 1.7.3.2.2 valem também para matrizes complexas, contanto que, para o segundo teorema, a matriz **A** seja hermitiana (a conjugada transposta de **A** é igual a **A**) positiva definida.

1.8. Matrizes p-cíclicas

O desenvolvimento do método SOR foi impulsionado pela necessidade de resolver SELAS oriundos da discretização do problema de Dirichlet num retângulo [18] e de uma grande classe de equações diferenciais parciais elípticas em regiões gerais [55]. Mas os autores citados aumentaram muito a classe de matrizes, para as quais o método SOR se aplica e é vantajoso. Tais matrizes se caracterizam por certas propriedades, que vamos estudar nesta secção, o que atenderá ao objetivo de nosso trabalho de entender o comportamento do parâmetro ótimo.

Consideremos um SELAS $\mathbf{A}\mathbf{x} = \mathbf{b}$, onde a matriz $\mathbf{A} = [a_{ij}]$ tem ordem $N \geq 2$, e a particionemos em submatrizes (*blocos*) $\mathbf{A}_{ij}$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} & \cdots & \mathbf{A}_{1n} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & \cdots & \mathbf{A}_{2n} \\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} & \cdots & \mathbf{A}_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{A}_{n1} & \mathbf{A}_{n2} & \mathbf{A}_{n3} & \cdots & \mathbf{A}_{nn} \end{bmatrix}, \tag{1.22}$$

sendo os blocos diagonais matrizes quadradas. Podemos ter $\mathbf{A}_{ij} = a_{ij}$, caso em que chamaremos **A** de matriz de escalares. Separamos a submatriz

$$\mathbf{D} := \begin{bmatrix} \mathbf{A}_{11} & & & & \\ & \mathbf{A}_{22} & & & \\ & & \mathbf{A}_{33} & & \\ & & & \ddots & \\ & & & & \mathbf{A}_{nn} \end{bmatrix} \tag{1.23}$$

dos blocos diagonais de **A**, que vamos supor todos não-singulares, o que acarreta que **D** é também não-singular. A matriz quadrada **B** de ordem N, definida por

$$\mathbf{B} := -\mathbf{D}^{-1}\mathbf{A} + \mathbf{I} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}), \tag{1.24}$$

é a *matriz de Jacobi de blocos*, correspondente ao particionamento de **A** em (1.22). O nosso interesse reside nas seguintes matrizes de blocos:

$$\mathbf{A}_1 := \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_{1p} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{32} & \mathbf{A}_{33} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{pp-1} & \mathbf{A}_{pp} \end{bmatrix}, \text{ com } p \geq 2, \tag{1.25}$$

e

$$\mathbf{A}_2 := \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & & & \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \mathbf{A}_{n-1,n} \\ & & & \mathbf{A}_{nn-1} & \mathbf{A}_{nn} \end{bmatrix}. \tag{1.26}$$

Um matriz do tipo $\mathbf{A}_2$ será referida com *matriz tridiagonal de blocos*. As matrizes $\mathbf{A}_1$ e $\mathbf{A}_2$ originam, respectivamente, as matrizes de Jacobi de blocos,

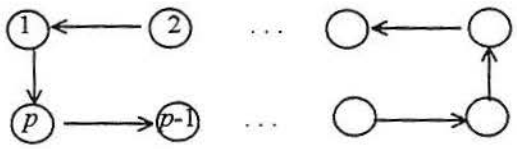
$$\mathbf{B}_1 := \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_{1p} \\ \mathbf{B}_{21} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{32} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_{pp-1} & \mathbf{0} \end{bmatrix}, \tag{1.27}$$

$$\mathbf{B}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{12} & & & \\ \mathbf{B}_{21} & \mathbf{0} & \mathbf{B}_{23} & & \\ & \mathbf{B}_{32} & \ddots & \ddots & \\ & & \ddots & \ddots & \mathbf{B}_{n-1,n} \\ & & & \mathbf{B}_{nn-1} & \mathbf{0} \end{bmatrix}. \tag{1.28}$$

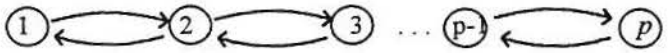
As matrizes $\mathbf{B}_1$ e $\mathbf{B}_2$ têm propriedades interessantes. A matriz $\mathbf{B}_1$, que está na forma da matriz na Definição 1.4.8, é uma matriz fracamente cíclica de índice p, e, operando convenientes permutações dos blocos na matriz $\mathbf{B}_2$, mostramos que ela é fracamente cíclica de índice 2. Estas constatações nos conduzem à definição que segue.

Definição 1.8.1. Se a matriz de Jacobi de blocos **B** em (1.24), relativa ao particionamento da matriz **A** em (1.22) é fracamente cíclica de índice $p \geq 2$, então dizemos que **A** é uma *matriz p-cíclica* relativa a esse particionamento de **A**.

Através dos grafos orientados generalizados podemos determinar quando uma dada matriz de blocos **A** é p-cíclica, ou equivalentemente quando a matriz de Jacobi de blocos **B**, relativa à matriz **A**, é fracamente cíclica de índice p. Vamos ilustrar com os grafos das matrizes $\mathbf{B}_1$ e $\mathbf{B}_2$. Para matrizes do tipo $\mathbf{B}_1$ o grafo é



Vemos que os comprimentos dos caminhos orientados fechados são todos múltiplos de p e, portanto, o grafo é cíclico de índice p. Para matrizes $\mathbf{B}_2$, o grafo é este



que, evidentemente, é cíclico de índice 2. O exposto nos ajuda a compreender a descrição de matrizes p-cíclicas em termos geométricos.

Teorema 1.8.2. *Suponhamos que o grafo orientado generalizado da matriz de Jacobi de blocos* **B** em (1.24) *seja fortemente conexo. Então, se esse grafo é cíclico de índice p, a matriz* **A** (1.22) *é p-cíclica.*

Prova. Se o grafo orientado generalizado G de **B** é fortemente conexo, então G é cíclico de índice p somente se a matriz particionada **B** é fracamente cíclica de índice p. □

Observamos que estão incluídas na classe de matrizes 2-cíclicas as matrizes com *a propriedade A*, segundo a definição de Young [55], pois estas são as matrizes **A** para as quais as submatrizes $\mathbf{A}_{ii}$ em (1.22) têm ordem 1.

1.9. Matrizes consistentemente ordenadas

A seguir definiremos matrizes consistentemente ordenadas. Estas compõem uma classe bastante grande e importante de matrizes, para as quais é possível determinar o valor ótimo do parâmetro de relaxação. Por exemplo, qualquer matriz que esteja na forma bloco-tridiagonal é consistentemente ordenada. Em verdade, até a presente data, só para matrizes consistentemente ordenadas sabemos como achar teoricamente o parâmetro ótimo do SOR.

1.9.1. Definição. Seja $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ a decomposição de uma matriz quadrada p-cíclica como em (1.22), onde **D** é a diagonal de **A**, e **L** e **U**, são as partes triangulares inferior e superior estritas de **A**, respectivamente. Dizemos que **A** é *consistentemente ordenada* sse os autovalores da matriz

$$\mathbf{B}(\alpha) := \alpha\mathbf{D}^{-1}\mathbf{L} + \alpha^{-(p-1)}\mathbf{D}^{-1}\mathbf{U},$$

independem de α, para $\alpha \neq 0$. Nesse caso dizemos também que a matriz de iteração de Jacobi $\mathbf{B}(1)$ é consistentemente ordenada. Caso contrário, dizemos que as matrizes **A** e **B** são inconsistentemente ordenadas.

Notemos que $\mathbf{B}(1)$ é a mesma matriz **B** que em (1.24).

1.9.2. Exemplo.Tomemos a seguinte matriz 2-cíclica,

$$\mathbf{A} := \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}.$$

A decomposição $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ é

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

$$\mathbf{B}(1) = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}) = -\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}\left(\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & -1/4 & 0 \\ -1/4 & 0 & -1/4 \\ 0 & -1/4 & 0 \end{bmatrix}.$$

Os autovalores de $\mathbf{B}(1)$ são 0 e $\pm\sqrt{2}/4$.

Por outro lado, tomando p =2 ($\mathbf{A}$ é 2-cíclica),

$$\mathbf{B}(\alpha) = \alpha\mathbf{L} + \alpha^{-(p-1)}\mathbf{U} = -\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}\left(\alpha\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + \alpha^{-1}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}\right)$$

$$= -\begin{bmatrix} 0 & 1/4\alpha & 0 \\ \alpha/4 & 0 & 1/4\alpha \\ 0 & \alpha/4 & 0 \end{bmatrix}.$$

Calculando os autovalores de $\mathbf{B}(\alpha)$, por exemplo, com o recurso da computação simbólica do MATLAB, obtemos os mesmos três autovalores acima. □

1.9.3. Exemplo. A matriz p-cíclica $\mathbf{A}_1$ em (1.25) é consistentemente ordenada. De fato, consideremos a correspondente matriz de Jacobi $\mathbf{B}_1$ em (1.27). Temos

$$\mathbf{B}_1(\alpha) := \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \alpha^{-(p-1)}\mathbf{B}_{1p} \\ \alpha\mathbf{B}_{21} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{B}_{32} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \alpha\mathbf{B}_{pp-1} & \mathbf{0} \end{bmatrix}.$$

E é fácil ver por multiplicação direta que $\mathbf{B}_1^p(\alpha) = \mathbf{B}_1^p$, para todo $\alpha \neq 0$. Então os autovalores de $\mathbf{B}_1^p(\alpha)$ não dependem de α. Logo a matriz p-cíclica $\mathbf{A}_1$ é consistentemente ordenada. □

1.9.4. Exemplo. Vamos mostrar que a matriz 2-cíclica $\mathbf{A}_2$ em (1.26) é consistentemente ordenada. Sejam $-\mathbf{L}_2$ e $-\mathbf{U}_2$ as partes triangulares inferior e superior de $\mathbf{B}_2$. Basta demonstrar que as matrizes

$\alpha\mathbf{L}_2 + \alpha^{-1}\mathbf{U}_2$ e $\mathbf{L}_2 + \mathbf{U}_2$ são semelhantes. Consideremos a matriz não-singular

$$\mathbf{P} := \begin{bmatrix} \mathbf{I}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{I}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \alpha^2\mathbf{I}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \alpha^{n-1}\mathbf{I}_n \end{bmatrix},$$

onde as matrizes identidades $\mathbf{I}_i$ têm as mesmas ordens que $\mathbf{A}_{ij}$. É fácil verificar que

$$\mathbf{P}^{-1}(\alpha\mathbf{L}_2 + \alpha^{-1}\mathbf{U}_2)\mathbf{P} = \mathbf{L}_2 + \mathbf{U}_2.$$

Por exemplo, se $n = 3$,

$$\mathbf{P}^{-1}(\alpha\mathbf{L}_2 + \alpha^{-1}\mathbf{U}_2)\mathbf{P} = -\begin{bmatrix} \mathbf{I}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha^{-1}\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \alpha^{-2}\mathbf{I}_3 \end{bmatrix}\begin{bmatrix} \mathbf{0} & \alpha^{-1}\mathbf{B}_{12} & \mathbf{0} \\ \alpha\mathbf{B}_{21} & \mathbf{0} & \alpha^{-1}\mathbf{B}_{23} \\ 0 & \alpha\mathbf{B}_{32} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{I}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \alpha^2\mathbf{I}_3 \end{bmatrix} =$$

$$= -\begin{bmatrix} \mathbf{0} & \mathbf{B}_{12} & \mathbf{0} \\ \mathbf{B}_{21} & \mathbf{0} & \mathbf{B}_{23} \\ 0 & \mathbf{B}_{32} & \mathbf{0} \end{bmatrix} = \mathbf{L}_2 + \mathbf{U}_2. \ \square$$

O Exemplo 1.9.4 é importante porque mostra que toda matriz tridiagonal de blocos, cujos blocos diagonais são não-singulares, é uma matriz 2-cíclica consistentemente ordenada. Mas nem todas as matrizes p-cíclicas são consistentemente ordenadas.

1.9.5. Exemplo. Por cálculos diretos, verificamos que a matriz seguinte é 2-cíclica relativamente ao particionamento tal que as submatrizes diagonais sejam de ordem 1,

$$\begin{bmatrix} 1 & -1/4 & 0 & -1/4 \\ -1/4 & 1 & -1/4 & 0 \\ 0 & -1/4 & 1 & -1/4 \\ -1/4 & 0 & -1/4 & 1 \end{bmatrix}.$$

Mas não é consistentemente ordenada, embora seja spd. □

1.9.6. Exemplo. Neste exemplo exibimos uma matriz spd que, particionada de maneira que as submatrizes na diagonal sejam de ordem 1, não é p-cíclica de ordem $p \geq 2$, mas, se particionada de maneira que as submatrizes diagonais sejam de ordem 2, então **A** é 2-cíclica consistentemente ordenada:

$$\begin{bmatrix} 20 & -4 & -4 & -1 & 0 & 0 \\ -4 & 20 & -1 & -4 & 0 & 0 \\ -4 & -1 & 20 & -4 & -4 & -1 \\ -1 & -4 & -4 & 20 & -1 & -4 \\ 0 & 0 & -4 & -1 & 20 & -4 \\ 0 & 0 & -1 & -4 & -4 & 20 \end{bmatrix}.$$

Observações. 1) O Exemplo 1.9.5 mostra que existem matrizes p-cíclicas, que não são consistentemente ordenadas, mas é fácil ver que, para toda matriz **A**, existe uma matriz de permutação **P**, que permuta os blocos de **A** de maneira que o reordenamento simétrico $\mathbf{PAP}^t$ seja uma matriz p-cíclica e consistentemente ordenada. De fato, pela Definição 1.4.8, existe uma matriz de permutação **P** que permuta os blocos da matriz de Jacobi **B** de modo que $\mathbf{PAP}^t$ seja fracamente cíclica de índice p e da forma (1.27), que, pelo Exemplo 1.9.3, é consistentemente ordenada. No entanto, a matriz **P** que produz ordenamentos consistentes não é única [51].

2) Varga [49] tem outras definições equivalentes de matriz consistentemente ordenada e de matrizes p-cíclicas.

1.10. Os métodos de Jacobi e SOR para matrizes de blocos

1.10.1. Os métodos. Consideremos uma matriz particionada

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} & \cdots & \mathbf{A}_{1k} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & \cdots & \mathbf{A}_{2k} \\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} & \cdots & \mathbf{A}_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{A}_{k1} & \mathbf{A}_{k2} & \mathbf{A}_{k3} & \cdots & \mathbf{A}_{kk} \end{bmatrix},$$

onde as matrizes $\mathbf{A}_{ii}$ são quadradas e não-singulares, e a correspondente matriz diagonal de blocos $\mathbf{D} = [\mathbf{A}_{ii}]$. Tomemos a decomposição $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$, onde **L** e **U** são as partes triangulares inferior e superior estritas de **A**, respectivamente. Ponhamos

$$\overline{\mathbf{L}} := \mathbf{D}^{-1}\mathbf{L} \quad \text{e} \quad \overline{\mathbf{U}} := \mathbf{D}^{-1}\mathbf{U}.$$

Então podemos escrever os métodos de Jacobi e SOR, respectivos, para matrizes de blocos, relativos a um SELAS $\mathbf{Ax} = \mathbf{b}$:

$$\mathbf{x}^{(n+1)} = \left(\overline{\mathbf{L}} + \overline{\mathbf{U}}\right)\mathbf{x}^{(n)} + \mathbf{D}^{-1}\mathbf{b}, \; n = 0, 1, 2, \ldots \tag{1.29}$$

$$\left(\mathbf{I} - \omega\overline{\mathbf{L}}\right)\mathbf{x}^{(n+1)} = \left(\omega\overline{\mathbf{U}} + (1-\omega)\mathbf{I}\right)\mathbf{x}^{(n)} + \omega\mathbf{D}^{-1}\mathbf{b}, \; n = 0, 1, 2, \ldots \tag{1.30}$$

Em cada iteração n é preciso resolver k SELAS, $\sum_{j=1}^{k} \mathbf{A}_{ij}\mathbf{X}_j = \mathbf{b}_j$, $i = 1, 2, \ldots, k$, por métodos

diretos, pois a inversão de **D** não se reduz aqui a uma simples divisão, como no caso em que os elementos de **D** são escalares.

1.10.2. Exemplo. Exemplifiquemos o método de Jacobi para matrizes de blocos. Seja o SELAS

$$\left[\begin{array}{ccc:cc} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ \hdashline 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ \hdashline 0 & 0 & 0 & -1 & 1 \end{array}\right]\left[\begin{array}{c} x_1 \\ x_2 \\ \hdashline x_3 \\ x_4 \\ \hdashline x_5 \end{array}\right] = \left[\begin{array}{c} 3 \\ -5 \\ \hdashline 9 \\ -6 \\ \hdashline 2 \end{array}\right].$$

Temos a iteração genérica $\mathbf{x}^{(n+1)} = \left(\overline{\mathbf{L}} + \overline{\mathbf{U}}\right)\mathbf{x}^{(n)} + \mathbf{D}^{-1}\mathbf{b}$, sendo

$$\mathbf{D}^{-1} = \frac{1}{3}\left[\begin{array}{cc:cc:c} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 3 \end{array}\right] \quad \text{e} \quad \overline{\mathbf{L}} + \overline{\mathbf{U}} = \left[\begin{array}{cc:cc:c} 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 2/3 & 0 & 0 \\ \hdashline 0 & 2/3 & 0 & 0 & 1/3 \\ 0 & 1/3 & 0 & 0 & 2/3 \\ \hdashline 0 & 0 & 0 & 1 & 1 \end{array}\right].$$

Podemos expressar o processo de forma escalar:

$$\begin{cases} \begin{cases} 2x_1^{(n+1)} - x_2^{(n+1)} = 3 \\ -x_1^{(n+1)} + 2x_2^{(n+1)} = x_3^{(n)} - 5 \end{cases} \\ \begin{cases} 2x_3^{(n+1)} - x_4^{(n+1)} = x_2^{(n)} + 9 \\ -x_3^{(n+1)} + 2x_4^{(n+1)} = x_5^{(n)} - 6 \end{cases} \\ x_5^{(n+1)} = x_4^{(n)} + 2. \end{cases}$$

Resolvendo esses três sistemas, obtemos

$$\begin{cases} \begin{cases} x_1^{(n+1)} = \dfrac{1}{3}x_2^{(n)} + \dfrac{1}{3} \\ x_2^{(n+1)} = \dfrac{2}{3}x_3^{(n)} - \dfrac{7}{3} \end{cases} \\ \begin{cases} x_3^{(n+1)} = \dfrac{2}{3}x_2^{(n)} + \dfrac{1}{3}x_5^{(n)} + 4 \\ x_4^{(n+1)} = \dfrac{1}{3}x_2^{(n)} + \dfrac{2}{3}x_5^{(n)} - 1 \end{cases} \\ x_5^{(n+1)} = x_4^{(n)} + 2. \ \square \end{cases}$$

1.10.3. O SOR e a computação paralela. Com o surgimento da computação paralela, o método SOR para matrizes de blocos adquiriu força renovada, tornando-se concorrente, por exemplo, com o método do gradiente conjugado.

Por outro lado, o método de Jacobi, mesmo sendo considerado ideal para a paralelização, pois, como mostra o Exemplo 1.10.2, nele a atualização de um grupo de coordenadas em cada passo

pode ser feita independentemente das de outros grupos, apresenta, tanto para os esquemas em série como para os esquemas em paralelo, uma convergência muito lenta.

Os dois tipos de conexões de computadores, usados atualmente na computação paralela, estão tratados de maneira breve no Apêndice *A*. Aqui comentamos o ordenamento *preto-vermelho*, também conhecido por ordenamento *preto-branco*, no qual atribuímos duas cores aos nodos da grade, distribuídos como num tabuleiro de xadrez (caso bidimensional).

Dado um problema de contorno, por exemplo, com equação de Poisson ou Laplace, o ordenamento preto-vermelho, descrito por diversos autores [01, 14, 41], consiste em colorir os nodos da grade de tal maneira que um nodo preto fique circundado somente por nodos vermelhos, e um vermelho fique circundado somente por nodos pretos. A situação mais simples é a de um problema unidimensional, por exemplo, do calor que se propaga através de uma barra metálica, Fig. 3.1, com temperaturas inicial e final $T_i > T_f$, que estabelecem as condições de contorno.

Fig.1.1 – Modelo preto-vermelho unidimensional

Vamo-nos ater a um caso mais rico, o caso bidimensional. Seja o caso típico da equação de Poisson $\nabla^2 u = f(x,y)$ no retângulo $(0\ ;\ 1)\times(0\ ;\ 1)$ com a condição de contorno: $u(x, y) = 0$ na fronteira. Discretizamos o problema no sentido de aproximar u em pontos (x_i, y_j) com $x_i = ih$ e $y_j = jh$, e $h = 1/(N+1)$. A localização desses pontos no retângulo $[0\ ;\ 1]\times[0\ ;\ 1]$ forma uma grade e esses pontos são designados *nodos* da grade. A cada nodo interior do retângulo corresponderá uma incógnita e, portanto, uma equação do SELAS,

$$u_{i-1j} + u_{i+1j} + u_{ij-1} + u_{ij+1} - 4u_{ij} = h^2 f_{ij},\ \ i,j = 1, 2, \ldots, N^2,$$

onde $u_{ij} := u(x_i, y_j)$ e $f_{ij} := f(x_i, y_j)$.

A cada ordenamento dos nodos corresponde um ordenamento das equações no SELAS. Lembremos que a convergência do SOR depende estreitamente da ordem das equações. Por exemplo, usando a *ordenamento natural*, isto é, pelas colunas, de cima para baixo, da esquerda para a direita, para $N = 3$, obtemos um SELAS da forma

$$\mathbf{Au} = \begin{bmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{bmatrix} \bullet \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \\ b_8 \\ b_9 \end{bmatrix}.$$

O ordenamento preto-vermelho está visualizado na Fig.1.2. Aí contamos primeiro os nodos pretos, por

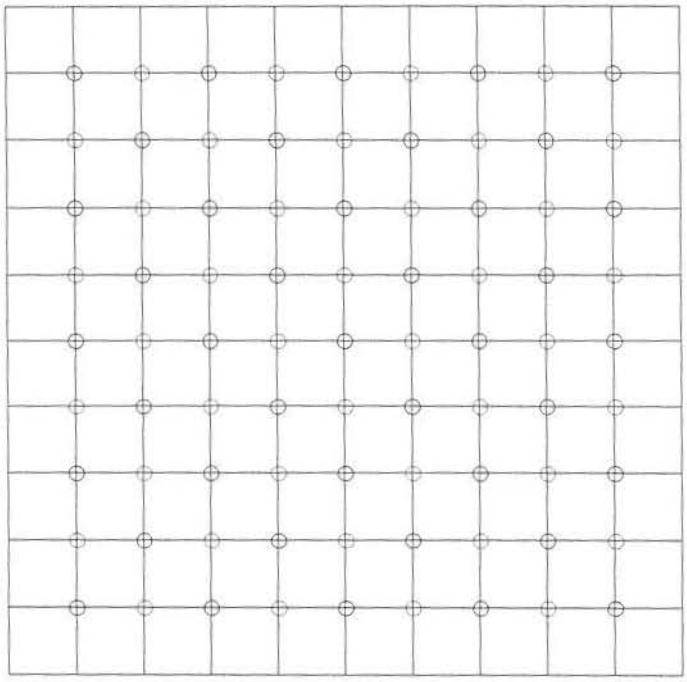


Fig.1.2 - Ordenamento preto-vermelho

coluna. Isso corresponde a efetuar o seguinte reordenamento simétrico na matriz **A**:

$$\mathbf{PAP}^{-1} = \begin{bmatrix} 4 & & & & & & -1 & -1 & & \\ & 4 & & & & & -1 & & -1 & \\ & & 4 & & & & -1 & -1 & -1 & -1 \\ & & & 4 & & & & -1 & & -1 \\ & & & & 4 & & & & -1 & -1 \\ -1 & -1 & -1 & & & 4 & & & & \\ -1 & & -1 & -1 & & & 4 & & & \\ & -1 & -1 & & -1 & & & 4 & & \\ & & -1 & -1 & -1 & & & & 4 & \end{bmatrix}.$$

No ordenamento preto-vermelho, uma vez que os nodos pretos são adjacentes somente a nodos vermelhos, nas iterações do SOR, se atualizarmos primeiro todos os nodos pretos, serão usados apenas dados anteriores relativos a nodos vermelhos, e, então, quando atualizamos os nodos vermelhos, já que são adjacentes apenas a nodos pretos, serão usados somente os novos dados relativos aos nodos pretos. Abaixo exemplificamos a atualização dos nodos pretos, após a inicialização:

$$\begin{cases} u_1^{(1)} = u_1^{(0)} + \dfrac{\omega}{4}\left(b_1 - 4u_1^{(0)} - u_6^{(0)} - u_7^{(0)}\right) \\ u_2^{(1)} = u_2^{(0)} + \dfrac{\omega}{4}\left(b_7 - 4u_2^{(0)} - u_6^{(0)} - u_8^{(0)}\right) \\ u_3^{(1)} = u_3^{(0)} + \dfrac{\omega}{4}\left(b_5 - 4u_3^{(0)} - u_6^{(0)} - u_7^{(0)} - u_8^{(0)} - u_9^{(0)}\right) \\ u_4^{(1)} = u_4^{(0)} + \dfrac{\omega}{4}\left(b_3 - 4u_4^{(0)} - u_7^{(0)} - u_9^{(0)}\right) \\ u_5^{(1)} = u_5^{(0)} + \dfrac{\omega}{4}\left(b_9 - 4u_5^{(0)} - u_8^{(0)} - u_9^{(0)}\right) \end{cases}$$

Em seguida atualizamos os nodos vermelhos:

$$\begin{cases} u_6^{(1)} = u_6^{(0)} + \dfrac{\omega}{4}\left(b_4 - 4u_6^{(0)} - u_1^{(1)} - u_2^{(1)} - u_3^{(1)}\right) \\ u_7^{(1)} = u_7^{(0)} + \dfrac{\omega}{4}\left(b_2 - 4u_7^{(0)} - u_1^{(1)} - u_3^{(1)} - u_7^{(1)}\right) \\ u_8^{(1)} = u_8^{(0)} + \dfrac{\omega}{4}\left(b_8 - 4u_8^{(0)} - u_2^{(1)} - u_3^{(1)} - u_5^{(1)}\right) \\ u_9^{(1)} = u_9^{(0)} + \dfrac{\omega}{4}\left(b_6 - 4u_9^{(0)} - u_3^{(1)} - u_4^{(1)} - u_5^{(1)}\right). \end{cases}$$

O esquema preto-vermelho pode ser resumido no seguinte algoritmo, onde a iteração n +1 atualiza os resultados da iteração n.

1.10.3.1. Algoritmo para o método preto-vermelho.

Para todos os nodos pretos da grade,

$$u_i^{(n+1)} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}u_j^{(n+1)} - \sum_{j=i+1}^{N} a_{ij}u_j^{(n)}\right) + (1-\omega)u_i^{(n)}, \qquad i = 1, 2, \ldots, N;$$

fim para

Para todos os nodos vermelhos da grade

$$u_i^{(n+1)} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}u_j^{(n+1)} - \sum_{j=i+1}^{N} a_{ij}u_j^{(n)}\right) + (1-\omega)u_i^{(n)}, \qquad i = 1, 2, \ldots, N;$$

fim para

Com relação à computação paralela, distribuímos os cálculos relativos aos nodos pretos em diversos processadores e, após o término desses cálculos, passamos para os nodos vermelhos já com as atualizações dos nodos pretos feitas, o que ilustramos na Fig.1.3.

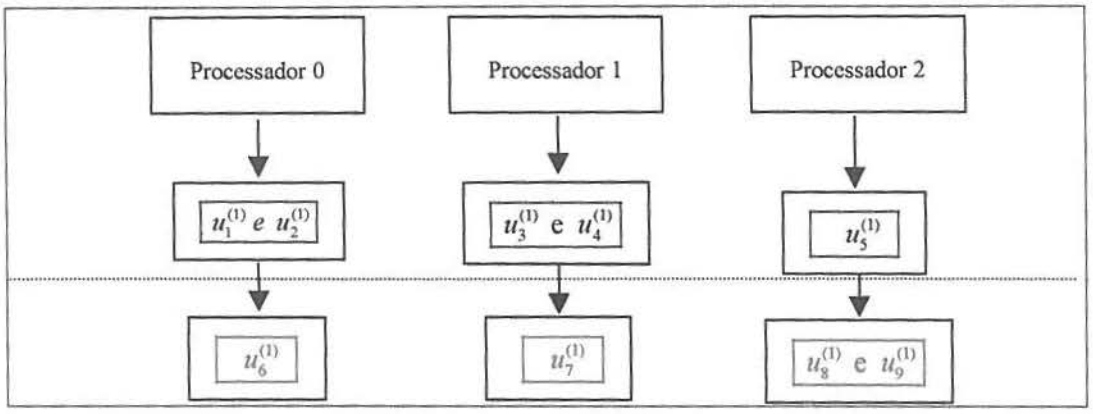


Fig.1.3 – Esquema preto-vermelho na computação paralela

Para equações diferenciais parciais que envolvam discretizações em dimensões maiores que 2, utilizamos mais do que duas cores como está descrito em [08]. Em [19] é feito um estudo da implementação do método SOR multicolorido em um supercomputador vetorial.

Além do modelo colorido para a execução do SOR na computação paralela, usamos o conceito de *decomposição de domínio*. A decomposição de domínio divide a região das malhas em subdomínios, como mostrado na Fig.1.4.

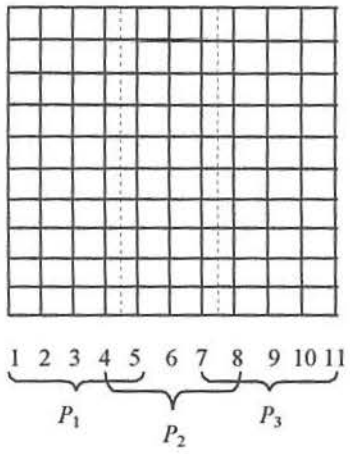


Fig.1.4 – Exemplo de decomposição de domínio em três subdomínios.

O objetivo da subdivisão é utilizar, para cada subdomínio, um processador, repartindo a tarefa a ser executada. Por exemplo, com o objetivo de modelar matematicamente o grau de poluição do Rio Guaíba, em [15] foi construído um modelo, baseado em Equações Shallow Water (SWEs: equações que governam o escoamento bidimensional) para os agentes poluentes despejados nesse rio, e usados quatro processadores para os quatro subdomínios em que foi dividido o domínio, Fig.1.5.

Quando determinamos os subdomínios, devemos cuidar do balanceamento, ou seja, tentar compartilhar de maneira balanceada o melhor possível a tarefa, para que não haja sobrecarga de cálculos para algum processador. Na Fig.1.4 a divisão é simples, mas no caso do Rio Guaíba o cuidado foi maior, devido a que o contorno do rio é irregular. A Fig.1.4 esquematiza uma implementação real do método SOR em [54], onde foi realizada uma comparação entre quatro supercomputadores diferentes, todos do tipo MIMD (cf. Apêndice *A*), obtendo em todos eles um bom desempenho. Aqui foi usada

a linguagem PFORTRAN e, no caso do Guaíba, a linguagem *C*, e, para ambos os casos, foi usada a biblioteca MPI para troca de mensagens (cf. Apêndice *A*).

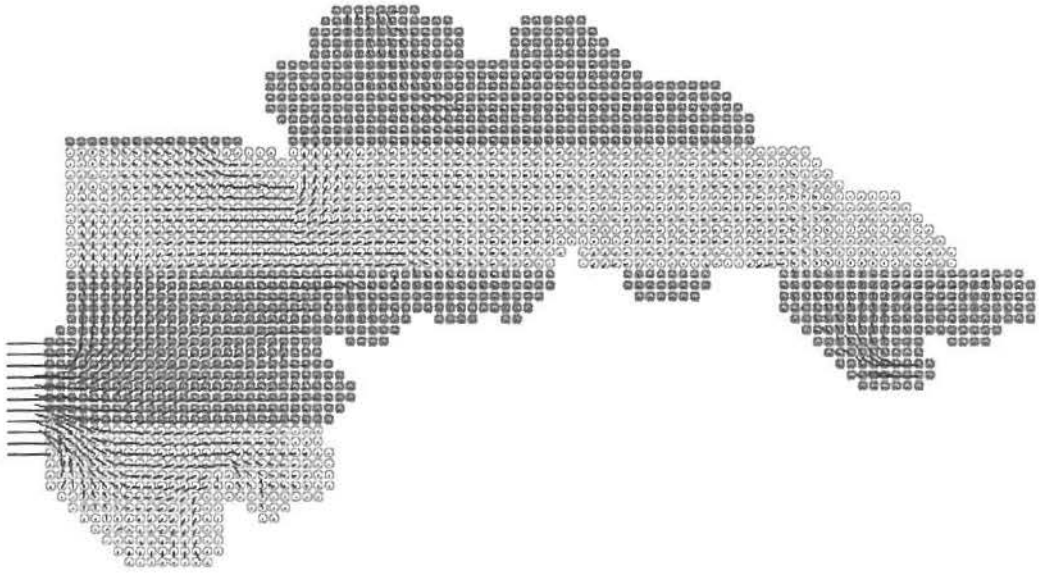


Fig.1.5 – Divisão do Rio Guaíba em quatro subdomínios

1.11. Relação entre os autovalores da matriz de Jacobi e do SOR

O resultado central para obter o estabelecimento da expressão do parâmetro ótimo do SOR é a relação (1.34) abaixo entre os autovalores μ da matriz de blocos de Jacobi,

$$\mathbf{B} := \overline{\mathbf{L}} + \overline{\mathbf{U}}, \tag{1.31}$$

e os autovalores λ da matriz de blocos do SOR,

$$\mathbf{S}_\omega := \left(\mathbf{I} - \omega\overline{\mathbf{L}}\right)^{-1}\left(\omega\overline{\mathbf{U}} + (1-\omega)\mathbf{I}\right). \tag{1.32}$$

1.11.1. Lema. Se $\mathbf{B} := \overline{\mathbf{L}} + \overline{\mathbf{U}}$ é uma matriz quadrada de ordem N consistentemente ordenada e fracamente cíclica de índice p, então, para quaisquer constante complexas α, β, γ,

$$\det\left(\gamma\mathbf{I} - \alpha\overline{\mathbf{L}} - \beta\overline{\mathbf{U}}\right) = \det\left(\gamma\mathbf{I} - \left(\alpha^{p-1}\beta\right)^{1/p}\left(\overline{\mathbf{L}} + \overline{\mathbf{U}}\right)\right). \tag{1.33}$$

Prova. Chamando de σ_i e τ_i os autovalores respectivos das matrizes $\alpha\overline{\mathbf{L}} + \beta\overline{\mathbf{U}}$ e $\left(\alpha^{p-1}\beta\right)^{1/p}\left(\overline{\mathbf{L}} + \overline{\mathbf{U}}\right)$, o primeiro e segundo membros de (1.33) valem, respectivamente (polinômios característicos),

$$(\gamma-\sigma_1)(\gamma-\sigma_2)\ldots(\gamma-\sigma_N) \quad \text{e} \quad (\gamma-\tau_1)(\gamma-\tau_2)\ldots(\gamma-\tau_N).$$

Basta mostrar que o conjunto dos sigmas é igual ao conjunto dos taus. Se α ou β são nulos, então todos os sigmas e os taus são nulos (notemos que, nesse caso, $\alpha\bar{\mathbf{L}}+\beta\bar{\mathbf{U}}$ é triangular estrita). Suponhamos, então, que ambos os α e β sejam não-nulos e ponhamos $\nu := (\alpha/\beta)^{1/p}$. Vem

$$\alpha\bar{\mathbf{L}}+\beta\bar{\mathbf{U}} = \left(\alpha^{p-1}\beta\right)^{1/p}\left(\nu\bar{\mathbf{L}}+\nu^{-(p-1)}\bar{\mathbf{U}}\right).$$

Segue pelas hipóteses e pela Definição 1.9.1 que os autovalores de $\nu\bar{\mathbf{L}}+\nu^{-(p-1)}\bar{\mathbf{U}}$ independem de $\nu\neq 0$. Portanto os autovalores de $\alpha\bar{\mathbf{L}}+\beta\bar{\mathbf{U}}$ são iguais aos de $\left(\alpha^{p-1}\beta\right)^{1/p}\left(\bar{\mathbf{L}}+\bar{\mathbf{U}}\right)$, porque estes últimos são os mesmos que os de $\left(\alpha^{p-1}\beta\right)^{1/p}\left(\nu\bar{\mathbf{L}}+\nu^{-(p-1)}\bar{\mathbf{U}}\right)$. □

Precisamos ainda de um resultado, conhecido como Teorema de Romanovsky [39], que enunciaremos sem demonstrar, para estabelecer o resultado central do capítulo, o Teorema 1.11.3.

1.11.2. Teorema. *Se* **A** *é uma matriz de ordem* N×N *fracamente cíclica de índice* p, *então o polinômio característico de* **A** *é*

$$t^n\prod_{i=1}^{r}\left(t^p-\sigma_i^p\right),$$

onde $n + rp = N$ *e os* σ_i *são autovalores não-nulos de* **A.**

1.11.3. Teorema. *Seja a matriz* **A** *particionada como em* (1.22), *que supomos p-cíclica e consistentemente ordenada, com blocos diagonais quadrados não-singulares* $\mathbf{A}_{ii}$, $i = 1, 2, \ldots, p$, *e seja* $\omega \neq 0$.

(a) *Se* $\lambda \neq 0$ *é um autovalor da matriz do* SOR $\mathbf{S}_\omega$ *em* (1.32) e se μ *satisfaz*

$$\left(\lambda+\omega-1\right)^p = \lambda^{p-1}\omega^p\mu^p, \tag{1.34}$$

então μ *é um autovalor da matriz de Jacobi* **B** *em* (1.31). *Reciprocamente,*

(b) *se* μ *é um autovalor da matriz* **B** *e se* λ *satisfaz* (1.34), *então* λ *é um autovalor de* $\mathbf{S}_\omega$.

Prova. Os autovalores de $\mathbf{S}_\omega$ são as raízes de

$$\det(\lambda\,\mathbf{I}-\mathbf{S}_\omega)=0.$$

Como $\det\left(\mathbf{I}-\omega\bar{\mathbf{L}}\right)=1$,

$$\begin{aligned}\det(\lambda\mathbf{I}-\mathbf{S}_\omega) &= \det\left(\mathbf{I}-\omega\bar{\mathbf{L}}\right)\det\left(\lambda\mathbf{I}-\mathbf{S}_\omega\right)=\det\left[\left(\mathbf{I}-\omega\bar{\mathbf{L}}\right)\left(\lambda\mathbf{I}-\mathbf{S}_\omega\right)\right]\\ &= \det\left[\left(\lambda+\omega-1\right)\mathbf{I}-\lambda\omega\bar{\mathbf{L}}-\omega\bar{\mathbf{U}}\right]=:\varphi(\lambda).\end{aligned}$$

Pelo Lema 1.11.1,

$$\varphi(\lambda)=\det\left[\left(\lambda+\omega-1\right)\mathbf{I}-\lambda^{(p-1)/p}\omega\mathbf{B}\right]. \tag{1.35}$$

Como, por hipótese, **A** é p-cíclica, a matriz **B**, e portanto também a matriz $\lambda^{(p-1)/p}\omega\mathbf{B}$, é fracamente cíclica de índice p. Pelo Teorema de Romanovsky 1.11.2,

$$\varphi(\lambda) = (\lambda+\omega-1)^n \prod_{i=1}^{r}\left[(\lambda+\omega-1)^p - \lambda^{p-1}\omega^p\mu_i^p\right], \tag{1.36}$$

onde os μ_i são não-nulos se $r \geq 1$.

(b) Seja μ um autovalor de **B** e suponhamos que λ satisfaz (1.34). Então um dos fatores de (1.36) se anula, o que significa que λ é um autovalor de $\mathbf{S}_\omega$.

(a) Seja λ um autovalor não-nulo de $\mathbf{S}_\omega$. Então, pelo menos um fator de (1.36) se anula.

Se $\mu \neq 0$ satisfaz (1.34), então $(\lambda+\omega-1) \neq 0$. Logo

$$(\lambda+\omega-1)^p = \lambda^{p-1}\omega^p\mu_i^p,$$

para algum $\mu_i \neq 0$, com $i \in \{1, 2, \ldots, r\}$. Olhando para essa equação e a (1.34), escrevemos, após igualar os segundos membros,

$$\lambda^{p-1}\omega^p\left(\mu^p - \mu_i^p\right) = 0.$$

Logo $\mu^p = \mu_i^p$. Portanto

$$\mu = \mu e^{2\pi r i/p}, \text{ para algum } r \in \{0, 1, \ldots, p\text{-}1\}.$$

Lembrando que **B** é fracamente cíclica, pelo Teorema de Romanovsky, μ é também um autovalor de **B**.

Agora, se $\mu = 0$ verifica (1.34), e se $\lambda \neq 0$ é um autovalor de $\mathbf{S}_\omega$, resulta $\varphi(\lambda) = 0$, e, de (1.34), que $\lambda + \omega - 1 = 0$; conseqüentemente, obtemos de (1.35) que $\det(\mathbf{B}) = 0$, ou $\det(\mu\mathbf{I} - \mathbf{B}) = \det(\mathbf{B}) = 0$, o que mostra que $\mu = 0$ é um autovalor de **B**. □

Apliquemos o Teorema 1.11.3 ao caso Gauss-Seidel: $\omega = 1$. Sejam **A** e **B** como nesse teorema. Se μ é um autovalor de **B**, então $\lambda = \mu^p$ é um autovalor da matriz de iteração de Gauss-Seidel $\mathbf{S}_1$; reciprocamente, se λ é um autovalor não-nulo de $\mathbf{S}_1$ e $\mu^p = \lambda$, então μ é um autovalor de **B**. Resulta daí que, para matrizes p-cíclicas consistentemente ordenadas, o método iterativo de Jacobi para matrizes de blocos converge sse o método iterativo de Gauss-Seidel para matrizes de Blocos converge, e, nesse caso,

$$\rho(\mathbf{S}_1) = (\rho(\mathbf{B}))^p < 1, \tag{1.37}$$

e, tomando logaritmos naturais,

$$R_\infty(\mathbf{S}_1) = pR_\infty(\mathbf{B}), \tag{1.38}$$

o que mostra que o método de Gauss-Seidel converge p vezes mais rápido que o de Jacobi.

1.11.4. Exemplo. Apliquemos o método de Jacobi e Gauss-Seidel ao SELAS $\mathbf{Ax} = \mathbf{b}$, com a matriz **A** [3] abaixo, de ordem 900×900, consistentemente ordenada e 2-cíclicas (cf. secções 1.8 e 1.9); ela se origina da discretização do problema de contorno de Poisson a duas dimensões:

[3] A Fig.3.1, pág. 65, visualiza a estrutura dessa matriz.

$$\begin{cases} \Delta u = f(x,y), & (x,y) \in \Omega := (0;1)\times(0;1), \\ u / \partial\Omega = g(x,y). \end{cases}$$

$$\mathbf{A} := \left[\begin{array}{cccccccccccccccccccc}
4 & -1 & & & & & & & & -1 & & & & & & & & & & \\
-1 & 4 & -1 & & & & & & & & -1 & & & & & & & & & \\
 & -1 & 4 & -1 & & & & & & & & -1 & & & & & & & & \\
 & & -1 & 4 & & & & & & & & & -1 & & & & & & & \\
 & & & & 4 & -1 & & & & & & & & -1 & & & & & & \\
 & & & & -1 & 4 & -1 & & & & & & & & -1 & & & & & \\
 & & & & & -1 & 4 & -1 & & & & & & & & -1 & & & & \\
 & & & & & & -1 & 4 & -1 & & & & & & & & -1 & & & \\
 & & & & & & & \ddots & \ddots & \ddots & & & & & & & & \ddots & & \\
-1 & & & & & & & & \ddots & \ddots & \ddots & & & & & & & & -1 & \\
 & -1 & & & & & & & & \ddots & \ddots & \ddots & & & & & & & & -1 \\
 & & \ddots & & & & & & & & -1 & 4 & -1 & & & & & & & \\
 & & & -1 & & & & & & & & -1 & 4 & -1 & & & & & & \\
 & & & & -1 & & & & & & & & -1 & 4 & -1 & & & & & \\
 & & & & & -1 & & & & & & & & -1 & 4 & -1 & & & & \\
 & & & & & & -1 & & & & & & & & -1 & 4 & & & & \\
 & & & & & & & -1 & & & & & & & & & 4 & -1 & & \\
 & & & & & & & & -1 & & & & & & & & -1 & 4 & -1 & \\
 & & & & & & & & & -1 & & & & & & & & -1 & 4 & -1 \\
 & & & & & & & & & & -1 & & & & & & & & -1 & 4
\end{array}\right] \quad \text{e} \quad \mathbf{b} := \left[\begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array}\right]$$

A malha da discretização é de 5 pontos e tem 30 nodos e, conseqüentemente, a matriz **A** tem largura de banda igual a $2\times30+1 = 61$. A equação geradora do SELAS tem a forma

$$u_{i-1j} + u_{i+1j} + u_{ij-1} + u_{ij+1} - 4u_{ij} = h^2 f_{ij}, \quad i,j = 1, 2, \ldots, 900,$$

onde $h = 1/30$. Estamos supondo que as funções f e g sejam tais que o segundo membro de nosso sistema seja $\mathbf{b} = [1, 1, \ldots, 1]^t$.

Resolvemos o SELAS com o MATLAB, com o uso do programa `jasor.m`. O computador usado foi um Pentium III de 450 MHz e 128 MB de memória RAM. Os resultados obtidos estão na Tab.1.2,

	Número de iterações	Erro	Tempo
Jacobi	1661	9.9955×10^{-7}	437.98
Gauss-Seidel	899	9.9590×10^{-7}	245.19

Tab.1.2 – Comparação das velocidades dos métodos de Jacobi e Gauss-Seidel

Constatamos, como era esperado, que o método de Gauss-Seidel converge aproximadamente duas vezes mais rápido que o método de Jacobi.

2 - Parâmetro Ótimo do SOR

2.1. Introdução

A ferramenta montada no capítulo 1, especialmente a igualdade (1.34), que relaciona os autovalores da matriz $\mathbf{S}_\omega$ do SOR e os autovalores da matriz **B** de Jacobi, relativamente a um SELAS $\mathbf{Ax} = \mathbf{b}$, particionado em blocos, para o caso em que **A** é p-cíclica consistentemente ordenada, vai permitir [49] determinar teoricamente o *parâmetro ótimo* ω_o, que é o que satisfaz

$$\rho(\mathbf{S}_{\omega_o}) = \min_{\omega \in \mathbb{R}} \rho(\mathbf{S}_\omega),$$

em outras palavras, é o valor do parâmetro ω do SOR de modo que o raio espectral $\rho(\mathbf{S}_\omega) < 1$ de $\mathbf{S}_\omega$ seja o menor possível, e acarrete, em conseqüência, a convergência mais rápida possível da matriz $\mathbf{S}_\omega$.

A relação (1.34), base para os próximos resultados sobre o parâmetro ótimo, foi estabelecida para o caso de **A** ser não-singular. Mas, no capítulo 4, consideraremos também, com vistas ao parâmetro ótimo, a situação de um SELAS singular, que tenha solução (clássica).

Além das condições acima sobre o SELAS, suporemos que os autovalores μ de **B** satisfaçam

$$0 \le \mu^p \le \rho(\mathbf{B}^p) < 1,$$

o que inclui a condição de **B** ser convergente. Com essas hipóteses, percorreremos o caminho do estabelecimento de que o valor ótimo ω_o de ω é a solução única da equação em ω,

$$\rho(\mathbf{B})^p \omega^p = p^p (p-1)^{1-p} (\omega - 1), \tag{2.1}$$

no intervalo

$$\left(1 \,;\, \frac{p}{p-1}\right),$$

e de que o correspondente raio espectral da matriz do SOR é

$$\rho(\mathbf{S}_{\omega_o}) = (p-1)(\omega_o - 1). \tag{2.2}$$

2.1.1.Exemplo. Consideremos a matriz **A** 2-cíclica consistentemente ordenada do Exemplo 1.9.2, e a correspondente matriz de Jacobi **B**,

$$\mathbf{A} := \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & -1/4 & 0 \\ -1/4 & 0 & -1/4 \\ 0 & -1/4 & 0 \end{bmatrix}.$$

O raio espectral de $\mathbf{B}$ é $\sqrt{2}/4$, que, levado em (2.1), fornece para ω os valores $16 \pm 4\sqrt{14}$. Vemos que $\omega = 16+4\sqrt{14} > 2 \notin (1;2)$; mas o valor $\omega_o = 16-4\sqrt{14} \in (1;2)$, como afirmado acima, é o valor do parâmetro ótimo. Com esse valor construímos a matriz do SOR ótimo

$$\mathbf{S}_{\omega_o} = \begin{bmatrix} -15+4\sqrt{14} & 4-\sqrt{14} & 0 \\ -116+31\sqrt{14} & 15-4\sqrt{14} & 4-\sqrt{14} \\ -898+240\sqrt{14} & 116-31\sqrt{14} & 15-4\sqrt{14} \end{bmatrix}.$$

Ainda, com ω_o podemos obter através de (2.2) o raio espectral da matriz $\mathbf{S}_{\omega_o}$,

$$\rho\left(\mathbf{S}_{\omega_o}\right) = 15-4\sqrt{14}.$$

Com o MATLAB confirmamos esses resultados, pois obtemos $15-4\sqrt{14} \approx 0.0334$ e, para autovalores aproximados de $\mathbf{S}_{\omega_o}$,

$$\begin{array}{r} 0.0334 + 0.0000\mathrm{i} \\ 0.0334 - 0.0000\mathrm{i} \\ -0.0334. \end{array}$$

2.2. Derivação do parâmetro ótimo

De (1.34), pondo $\bar{\mu} := \rho(\mathbf{B})$ e $\lambda =: z^p$ e, extraindo a raiz de índice p, resulta a equação polinomial em z,

$$z^p - \omega\bar{\mu}z^{p-1} + (\omega - 1) = 0.$$

Definimos, então, para cada valor do parâmetro ω, o polinômio na variável z,

$$f_\omega(z) := z^p - \omega\bar{\mu}z^{p-1} + (\omega - 1). \tag{2.3}$$

Para preparar a dedução do resultado principal, apresentamos quatro lemas, enunciados em [49] com alguma indicação da demonstração, nos quais suporemos $\bar{\mu} > 0$, uma vez que o caso $\bar{\mu} = 0$ é trivial.

2.2.1. Lema. *Se ω_o satisfaz* (2.1) *e f_ω é definida por* (2.3), *então*

1. *se $1 < \omega < \omega_o$, f_ω tem exatamente dois zeros reais positivos;*
2. *se $\omega = \omega_o$, f_ω tem um único zero real positivo, e esse tem multiplicidade 2;*
3. *se $\omega_o < \omega < p/(p-1)$, f_ω não tem zeros reais positivos.*

Prova. Consideremos $\omega > 1$. Pondo $\zeta := z(\omega-1)^{-1/p}$, obtemos de (2.3)

$$f_\omega(z) = (\omega-1)\left(\zeta^p - \varepsilon(\omega)\zeta^{p-1} + 1\right) = (\omega-1)g_\omega(\zeta),$$

onde pusemos $g_\omega(\zeta) := \zeta^p - \varepsilon(\omega)\zeta^{p-1} + 1$ e

$$\varepsilon(\omega) := \frac{\omega\overline{\mu}}{(\omega-1)^{1/p}}. \tag{2.4}$$

Notemos que, se $\omega \neq 1$, então f_ω e g_ω têm os mesmos zeros. Pela Regra dos Sinais de Descartes [16], f_ω tem no máximo dois zeros reais positivos. Como a derivada

$$\frac{d\varepsilon}{d\omega} = \frac{\overline{\mu}}{p}(\omega-1)^{-1/p}\left(p - \omega(\omega-1)^{-1}\right) < 0,$$

para $\omega \in \left(1; 1/(p-1)\right)$, a função ε é estritamente decrescente nesse intervalo. Por outro lado, de (2.1) e (2.4), com $\overline{\mu} := \rho(\mathbf{B})$, obtemos

$$\varepsilon(\omega_\circ) = \frac{p}{(p-1)^{(p-1)/p}},$$

e vemos que $(p-1)^{1/p}$ é um zero de $g_{\omega_\circ}$, e também que, usando, por exemplo, a derivada $g'_\omega(\zeta)$, a multiplicidade desse zero é 2. Está pois provada a parte 2.

Pela monotonia de ε, resulta $g_\omega\left((p-1)^{1/p}\right) < 0$ quando $1 < \omega < \omega_\circ$, o que implica a parte 1 do teorema.

Vale $g_{\omega_\circ}(\zeta) \geq 0$ para todo $\zeta \geq 0$ e, então, pela monotonia de ε, temos que, para todo $\zeta \geq 0$, $g_\omega(\zeta) > 0$, se $\omega \in \left(1; 1/(p-1)\right)$, ficando provada também a parte 3. □

2.2.2. Lema. *Se* $1 < \omega < \omega_\circ$, *a função* f_ω, *definida em* (2.3), *tem um zero real maior que* $\left[(\omega_\circ - 1)(p-1)\right]^{1/p}$.

Prova. Para todo ω fixo no intervalo $(1\ ;\ \omega_\circ)$, e

$$\zeta_1 := \left[\frac{(\omega_\circ - 1)(p-1)}{\omega - 1}\right]^{1/p},$$

temos, usando a (2.1) com $\overline{\mu} := \rho(\mathbf{B})$,

$$\begin{aligned} g_\omega(\zeta_1) &= \frac{(\omega_\circ-1)(p-1)}{\omega-1} - \frac{\omega\overline{\mu}}{(\omega-1)^{1/p}}\left[\frac{(\omega_\circ-1)(p-1)}{\omega-1}\right]^{\frac{p-1}{p}} + 1 \\ &= \frac{(\omega_\circ-1)(p-1)}{\omega-1} - \frac{\omega p}{\omega_\circ}(p-1)^{\frac{1-p}{p}}(\omega_\circ-1)^{\frac{1}{p}}(\omega-1)^{-\frac{1}{p}}\left[\frac{(\omega_\circ-1)(p-1)}{\omega-1}\right]^{\frac{p-1}{p}} + 1 \\ &= \frac{\omega_\circ - 1}{\omega - 1}\left[\frac{(p-1)\omega_\circ - p\omega}{\omega_\circ}\right] + 1 < 0. \end{aligned}$$

Justifiquemos a desigualdade: primeiro observemos que a fração fora do colchete é maior que 1; então, se mostrarmos que o colchete é negativo, e maior que 1 em módulo, fica provada a desigualdade; da hipótese

$$1 < \omega < \omega_o < p/(p-1),$$

obtemos,

$$p-1 \; < \; \omega(p-1) \; < \; \omega_o(p-1) \; < \; p \; < p\omega,$$

o que mostra que o numerador do colchete é negativo, para todo p, com módulo

$$p\omega - (p-1)\omega_o > \; p\omega - p\omega + \omega_o = \omega_o.$$

Pelo Lema 2.2.1 e pelo tipo de função polinomial que é g_ω, esta deve ter um zero positivo menor que ζ_1 e outro maior que ζ_1, assumindo o mínimo (negativo) no ponto $\zeta = \varepsilon(\omega)(p\text{-}1)/p$, Fig.2.1. Mas

$$\zeta_1 := \left[\frac{(\omega_o - 1)(p-1)}{\omega - 1}\right]^{1/p} > \left[\omega_o - 1)(p-1)\right]^{1/p}.$$

Como f_ω e g_ω têm os mesmos zeros, quando $\omega \neq 1$, o lema está demonstrado. □

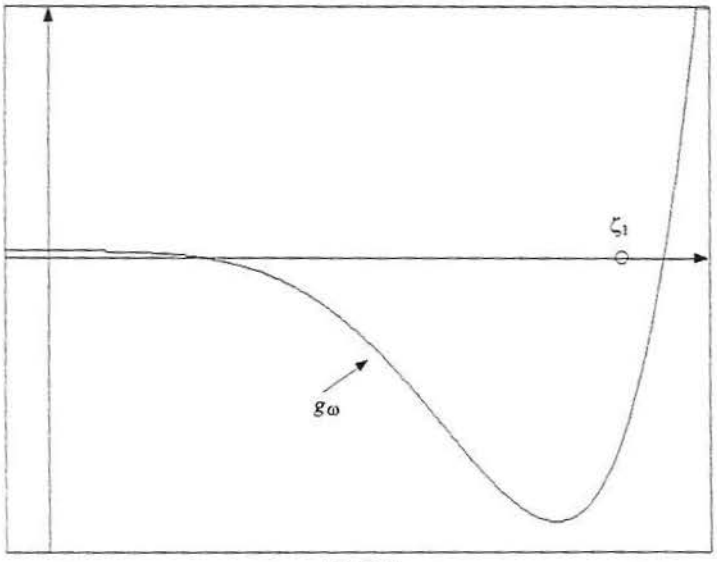


Fig.2.1

2.2.3. Lema. *Para todo $\omega \leq 1$ fixo, a função f_ω tem um zero em módulo maior que* $\left[(\omega_o - 1)(p-1)\right]^{1/p}$.

Prova. Pela definição de f_ω, temos $f_1(z) = z^{p-1}(z - \overline{\mu})$, e, portanto, $\overline{\mu}$ é um zero de f_1. Mostremos que $\overline{\mu}$ é maior que $\left[(\omega_o - 1)(p-1)\right]^{1/p}$. Por (2.1), onde $\overline{\mu} = \rho(\mathbf{B})$, temos

$$\overline{\mu} = \frac{p}{\omega_o}(p-1)^{(1-p)/p}(\omega_o - 1)^{1/p}$$
$$= \frac{p}{(p-1)\omega_o}\left[(p-1)(\omega_o - 1)\right]^{1/p}$$
$$> \left[(p-1)(\omega_o - 1)\right]^{1/p}.$$

A desigualdade vale porque estamos tomando $1 < \omega_o < \frac{p}{p-1}$.

Para $\omega < 1$,

$$f_\omega(\overline{\mu}) = \overline{\mu}^p - \omega\overline{\mu}^p + (\omega - 1)$$
$$= (1-\omega)\left(\overline{\mu}^p - 1\right) < 0,$$

pois estamos tomando $\overline{\mu} < 1$. Como $\lim_{|z|\to+\infty} f_\omega(|z|) = +\infty$, a conclusão se impõe. □

Falta estudar o caso $\omega_o \leq \omega$. Primeiro abordamos a situação da desigualdade estrita. Pela parte 3 do Lema 2.2.1, f_ω não tem zeros reais positivos, se $\omega_o \leq \omega < p/(p-1)$. Consideremos o conjunto

$$R := \{ z \in \mathbb{C} \mid z = Z(\omega), \text{com } \omega \in (\omega_o ; +\infty), \text{é um zero de} f_\omega \},$$

contido no semiplano complexo superior $\mathcal{I}z \geq 0$. A função Z é bivalente e contínua. É possível demonstrar [49] que R está contido no ânulo

$$\left\{ z \in \mathbb{C} \mid \left[(p-1)(\omega_o - 1)\right]^{1/p} < |z| \leq \left[(p-1)(\omega - 1)\right]^{1/p} \right\}.$$

Daí segue o lema,

2.2.4. Lema. *Para* $\omega_o < \omega$, *a função* f_ω *tem um zero* $Z(\omega)$, *que satisfaz*

$$\left[(p-1)(\omega_o - 1)\right]^{1/p} < |Z(\omega)| \leq \left[(p-1)(\omega - 1)\right]^{1/p}.$$

Consideremos agora $\omega = \omega_o$. Isolando $\mu = \mu(\omega)$ em (1.34), obtemos, com $\lambda =: z^p$,

$$\mu(z) = \frac{1}{\omega_o}\left[z + \frac{\omega_o - 1}{z^{p-1}} \right]. \tag{2.5}$$

Seja o número positivo $r := [(\omega_o - 1)(p-1)]^{1/p}$. Podemos verificar que a função $z \mapsto \mu(z)$, definida em (2.5), aplica $C_p := \{ z \in \mathbb{C} \mid |z| > r \}$ (conjunto exterior do círculo centrado na origem de raio r) conformemente sobre o conjunto exterior da curva fechada $\theta \mapsto \mu\left(re^{i\theta}\right)$.

Seja o conjunto compacto $S_p(\overline{\mu}) := \mathbb{C} - \mu(C_p)$. A Fig.2.2 ilustra o caso $p = 3$, que apresenta uma simetria tripla (hipociclóide tricúspide). De maneira geral, a aplicação definida em (2.5) tem simetria p$^{\text{upla}}$ no sentido de que, se $z = re^{i\theta}$, então

$$\mu\left(re^{i(\theta + 3k\pi/p)}\right) = e^{2k\pi i/p}\mu\left(re^{i\theta}\right), \tag{2.6}$$

para $k = 0, 1, 2, \ldots, p-1$.

Estamos prontos para apresentar os principais resultados deste capítulo, os Teoremas 2.2.5 e 2.2.6.

2.2.5. Teorema. *Seja um SELAS não-singular de ordem* N, $\mathbf{Ax} = \mathbf{b}$, *onde* $\mathbf{A}$ *é uma matriz p-cíclica consistentemente ordenada e particionada como em* (1.22), *com blocos diagonais quadrados não-singulares* $\mathbf{A}_{ii}$, i = 1, 2, . . . , p, *e as correspondentes matrizes* $\mathbf{B}$ *de Jacobi e* $\mathbf{S}_\omega$ *do SOR. Se o espectro* $\sigma(\mathbf{B}) \subset S_p(\overline{\mu})$, *onde* $\overline{\mu}$ *é o raio espectral de* $\mathbf{B}$ *com* $0 < \overline{\mu} < 1$, *e* $\rho(\mathbf{S}_\omega)$ *é o raio espectral de* $\mathbf{S}_\omega$, *então, denotando com* ω_o *a única solução de* (2.1) *no intervalo* (1 ; p/((p-1)), temos

1. $\rho(\mathbf{S}_{\omega_o}) = (\omega_o - 1)(p-1)$;
2. $\rho(\mathbf{S}_{\omega_o}) < \rho(\mathbf{S}_\omega)$, *para todo* $\omega \neq \omega_o$;
3. *para* $\omega > \omega_o$,
   a) $\rho(\mathbf{S}_{\omega_o}) < \rho(\mathbf{S}_\omega) < (\omega - 1)(p-1)$, *se* $p > 2$,
   b) $\rho(\mathbf{S}_{\omega_o}) < \rho(\mathbf{S}_\omega) < \omega - 1$, *se* $p = 2$.

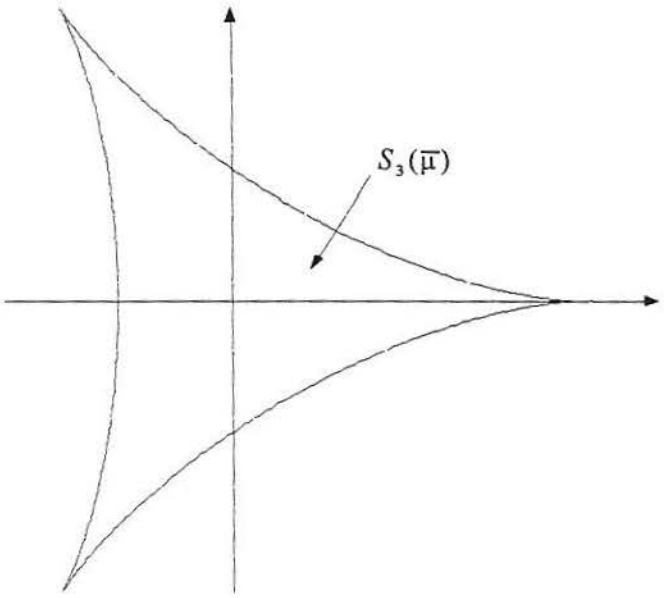


Fig.2.2 - $S_3(\overline{\mu})$ = (interior da curva) $\cup$ curva $\subset \mathbb{C}$

Demonstração. Com base nos Lemas 2.2.1 – 2.2.4, tudo ficará demonstrado, se provarmos que

$$\rho(\mathbf{S}_{\omega_o}) = (\omega_o - 1)(p-1). \tag{2.7}$$

Por hipótese, o espectro $\sigma(\mathbf{B}) \subset S_p(\overline{\mu})$, logo

$$\rho(\mathbf{S}_{\omega_o}) \leq (\omega_o - 1)(p-1). \tag{2.8}$$

Por outro lado, todos os autovalores μ_i de $\mathbf{B}$ satisfazem $|\mu_i| \leq |\overline{\mu}|$, onde a igualdade vale para algum i.

Lembremos que $r := [(\omega_o - 1)(p-1)]^{1/p}$. Logo, por (2.5), $\mu(r) = \overline{\mu}$. Como também, por (2.5),

$$|\mu(z)| \leq \overline{\mu} \text{ sempre que } |z| \leq r,$$

e, por hipótese, o espectro $\sigma(\mathbf{B}) \subset S_p(\overline{\mu})$, a imagem do fecho de C_p, $\mu(\overline{C}_p)$, contém pelo menos um autovalor de **B** e, portanto,

$$\rho(\mathbf{S}_{\omega_o}) \geq (\omega_o - 1)(p-1). \tag{2.9}$$

Com (2.8) e (2.9) obtemos (2.7). □

Para simplificar parte das hipóteses do Teorema 2.2.5 e, mesmo assim manter a validade de suas conclusões, com o intuito de adequar esse teorema à prática, suponhamos que os autovalores de $\mathbf{B}^p$ (potência p^ésima de **B**) sejam reais e não-negativos. O conjunto compacto $S_p(\overline{\mu})$, por (2.6), contém os segmentos

$$\left\{z \mid z = te^{2k\pi i/p}, t \in [0;\overline{\mu}]\right\}, k = 0, 1, \ldots, p-1. \tag{2.10}$$

O fato de que os autovalores de $\mathbf{B}^p$ são reais não-negativos implica, em presença do Teorema 1.10.3 de Romanovsky, válido porque **B** é fracamente cíclica de índice p, que os autovalores de **B** estão sobre os p segmentos (2.10), e, portanto, em $S_p(\overline{\mu})$. Acabamos de demonstrar o teorema seguinte.

2.2.6. Teorema. *Com as notações do Teorema* 2.2.5, *se os autovalores de* $\mathbf{B}^p$ *são reais e não-negativos, e* $0 < \overline{\mu} < 1$, *as conclusões do Teorema* 2.2.5 *são válidas.*

A demonstração do Teorema 2.2.5 abrangeu também a demonstração de que, se o espectro $\sigma(\mathbf{B}) \subset S_p(\overline{\mu}_1)$, $0 < \overline{\mu}_1 < 1$, então a matriz do SOR $\mathbf{S}_{\omega_1}$ é convergente, sendo ω_1 a solução da equação (2.1) com $\overline{\mu}_1$ em lugar de $\rho(\mathbf{B})$.

Comentamos acima que não compensa calcular o parâmetro ótimo a partir da matriz do SOR, devido ao elevado custo computacional. Um caminho para estimar esse parâmetro é estimar o raio espectral $\rho(\mathbf{B})$ da matriz de Jacobi e usar a equação (2.1) para uma correspondente estimativa do parâmetro ótimo. Cumpre então chamar a atenção sobre o que Young [55] observou, para o caso $p = 2$ – e isso se estende para o caso geral – que ω_o varia no mesmo sentido que $\rho(\mathbf{B})$. Portanto, uma estimativa do primeiro por falta ou por excesso produz uma estimativa por falta ou por excesso, respectivamente, para o segundo.

2.3. A geometria do caso $p = 2$

Indubitavelmente, o caso $p = 2$ é o mais importante do SOR p-cíclico, e, historicamente, o primeiro que foi estudado e compreendido [56]. Por causa disso, achamos que vale a pena dar-lhe atenção especial. Primeiro particularizaremos as fórmulas a partir das gerais obtidas e, depois, faremos uma abordagem geométrica independente, que permita uma compreensão mais plena e mais intuitiva.

Particularizando a equação (2.1) para $p = 2$, pondo $\overline{\mu} := \rho(\mathbf{B})$, resulta a equação em ω,

$$\bar{\mu}^2\omega^2 = 4(\omega - 1), \tag{2.11}$$

cuja solução única no intervalo (1 ; 2) produz o valor ótimo $\omega = \omega_o$ do parâmetro do SOR para matrizes 2-cíclicas consistentemente ordenadas. Resolvendo essa equação, obtemos

$$\omega = \frac{2 \pm 2\sqrt{1-\bar{\mu}^2}}{\bar{\mu}^2}. \tag{2.12}$$

A raiz correspondente ao sinal mais, uma vez que deve verificar-se $0 < \bar{\mu} < 1$, está fora do intervalo (1 ; 2) (cf. Exemplo 2.1.1). Então, satisfeitas todas as condições do Teorema 2.2.5 com $p = 2$, o valor ótimo do parâmetro do SOR é

$$\omega_o = \frac{2-2\sqrt{1-\bar{\mu}^2}}{\bar{\mu}^2} = \frac{2}{1+\sqrt{1-\bar{\mu}^2}}. \tag{2.13}$$

O Teorema 2.2.5, parte 1, também nós dá, para $p = 2$, que o raio espectral da matriz do SOR ótimo é

$$\rho\left(\mathbf{S}_{\omega_o}\right) = \omega_o - 1 = \frac{1+\sqrt{1-\bar{\mu}^2}}{1-\sqrt{1-\bar{\mu}^2}}.$$

Agora procedemos à confirmação geométrica de (2.13). Com $p = 2$, a equação (1.34) se escreve

$$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2, \tag{2.14}$$

que reescrevemos na forma

$$\frac{\lambda}{\omega} + 1 - \frac{1}{\omega} = \pm\lambda^{1/2}\mu. \tag{2.15}$$

Definimos funções f_ω e g por

$$f_\omega(\lambda) := \frac{\lambda}{\omega} + 1 - \frac{1}{\omega}, \quad \omega \in (0; 2) \tag{2.16}$$

e

$$g(\lambda) := \lambda^{1/2}\mu, \quad 0 \le \mu \le \bar{\mu} < 1. \tag{2.17}$$

Para todo ω, o gráfico de f_ω é uma reta (em verde na Fig.2.3), que passa pelo ponto (1, 1) com declividade $1/\omega$. O gráfico de $\lambda \mapsto \pm g(\lambda)$ é uma curva como a da Fig.2.3, que vamos indicar com $C(\mu)$. Para cada valor de ω, as soluções de (2.15) são as intersecções da reta f_ω e a curva $C(\mu)$.

A declividade da reta f_ω decresce monotonamente quando $\omega > 0$ descreve o intervalo (0 ; 2) na ordem natural, isto é, a reta gira no sentido horário em torno do ponto (1, 1), e o efeito sobre a abscissa do ponto de intersecção $(\lambda, g_\omega(\lambda))$ (este é o de maior abscissa entre os dois pontos de intersecção e esta abscissa corresponde ao raio espectral da matriz do SOR $\mathbf{S}_\omega$) é decrescer até um mínimo, que ocorre quando os dois pontos de intersecção coincidem e a reta f_ω se torna tangente à curva $C(\mu)$, Fig.2.4. Nessa figura salientamos o caso ótimo. Destacamos também a situação de Gauss-Seidel: a reta f_ω passa pela origem).

Para achar o valor $\bar{\omega}$ de ω, que corresponde ao ponto de tangência, usamos a equação (2.15)

com o sinal mais e determinamos o valor de ω de maneira que essa equação em λ tenha uma única solução. Na verdade a consideraremos uma equação em $x := \sqrt{\lambda}$:

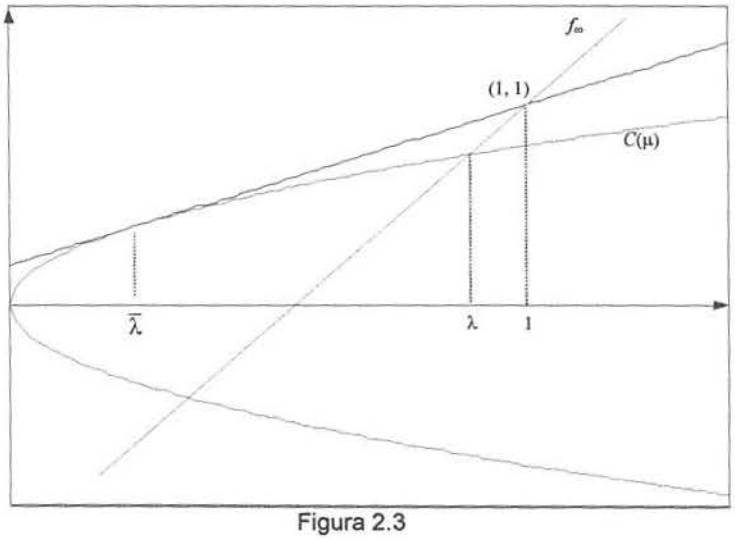


Figura 2.3

$$x^2 - \eta\omega x + (\omega - 1) = 0. \tag{2.18}$$

Para que essa equação tenha uma única solução é necessário e suficiente que o discriminante seja

$$\mu^2\omega^2 - 4\omega + 4 = 0, \tag{2.19}$$

ou

$$\omega = 2\left[\frac{1 \pm \sqrt{1-\mu^2}}{\mu^2}\right].$$

Devemos adotar o sinal menos, caso contrário teríamos $\omega > 2$ (cf. Teorema 1.7.3.2.1). Ponhamos

$$\overline{\omega} := 2\left[\frac{1-\sqrt{1-\mu^2}}{\mu^2}\right]. \tag{2.20}$$

Calculamos a abscissa no ponto de tangência assim: determinamos a raiz única da equação (2.18), com o uso de (2.19), obtendo $x = \overline{\omega}\mu/2 = \sqrt{\lambda}$; agora levamos o valor de μ em termos de $\overline{\omega}$ e λ em (2.14), onde ω deve ser substituído por $\overline{\omega}$, e obtemos a abscissa procurada

$$\overline{\lambda} := \overline{\omega} - 1.$$

Para $\omega > \overline{\omega}$, o primeiro membro de (2.19) é negativo, o que significa geometricamente que a reta f_ω na Fig.2.3 não intercepta $C(\mu)$ – cf. Fig.2.4. Nesse caso, a equação (2.18) tem duas raízes complexas conjugadas

$$\frac{\mu\omega \pm \sqrt{-\mu^2\omega^2 + 4\omega - 4}\, i}{2},$$

cujo quadrado do módulo é $\omega - 1$; e este é o módulo das raízes (complexas) do polinômio (2.14), quando $\omega > \overline{\omega}$. E, importante, vemos que esse módulo cresce quando ω cresce. Conseqüentemente, para um autovalor μ fixo de **B**, o valor de ω que minimiza a raiz de maior módulo da equação em λ, (2.14), é $\overline{\omega}$, calculado por (2.20).

A curva $C(\overline{\mu})$ é a envolvente de todas as curvas $C(\mu)$, $\mu \in [0;\overline{\mu}]$, e concluímos que

$$\min_{\omega \in \Omega} \rho(\mathbf{S}_\omega) = \rho(\mathbf{S}_{\omega_o}) = \omega_o - 1,$$

onde ω_o é dado em (2.13).

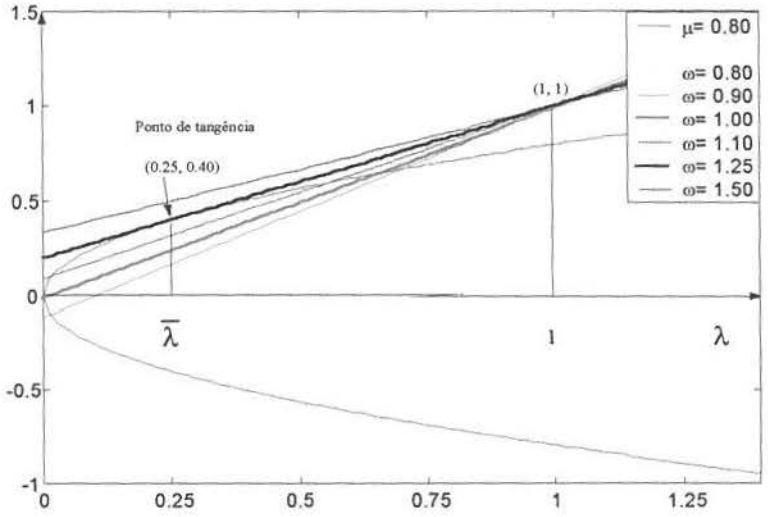


Figura 2.4

Na Figura 2.5, ilustramos o caso para um valor $\lambda > 1$, que leva o método SOR à divergir.

2.4. Comparação do SOR ótimo com Gauss-Seidel e Jacobi

Pelo Teorema 2.2.5, parte 1,

$$R_\infty(\mathbf{S}_{\omega_o}) = -\ln[(\omega_o - 1)(p-1)],$$

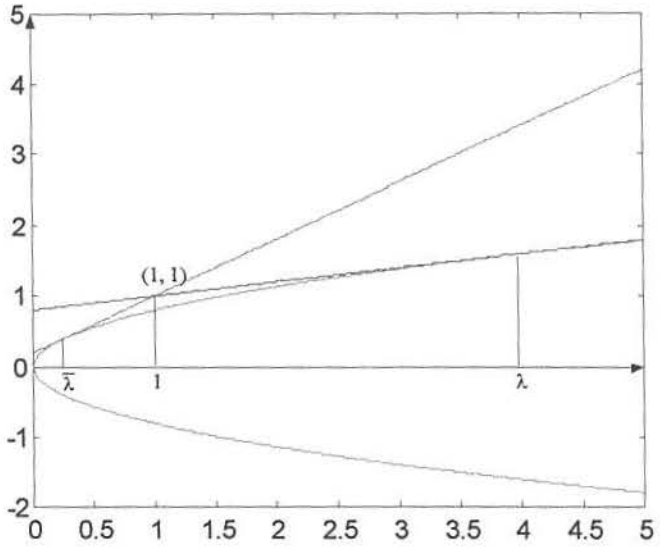


Figura 2.5

e, por (1.37),

$$R_\infty(\mathbf{S}_1) = -p \ln \rho(\mathbf{B}).$$

Usando a (2.1) com $\rho(\mathbf{B}) = \bar{\mu}$ e $\omega = \omega_o$, e aplicando duas vezes a Regra de l'Hôpital, obtemos o limite,

$$\lim_{\bar{\mu}\to 1^-} \frac{R_\infty(\mathbf{S}_{\omega_o})}{\left[R_\infty(\mathbf{S}_1)\right]^{1/2}} = \left(\frac{2p}{p-1}\right)^{1/2}. \tag{2.21}$$

Daí, também, com o apoio na igualdade (1.38), obtemos

$$\lim_{\bar{\mu}\to 1^-} \frac{R_\infty(\mathbf{S}_{\omega_o})}{\left[R_\infty(\mathbf{B})\right]^{1/2}} = \left(\frac{2p^2}{p-1}\right)^{1/2}. \tag{2.22}$$

Os resultados (2.21) e (2.22) mostram que o SOR ótimo, aplicado a matrizes p-cíclicas consistentemente ordenadas, produz considerável aceleração na convergência, em relação aos métodos de Gauss-Seidel e Jacobi, quando estes são lentos, isto é, os raios espectrais das respectivas matrizes de iteração são (menores que) próximos de 1.

2.4.1. Exemplo. Suponhamos que $\bar{\mu} = \rho(\mathbf{B}) = 0{,}9999$ e $p = 2$. Então

$$\omega_o = \frac{2}{1+\sqrt{1-(0{,}9999)^2}} \approx 1.9721,$$

$$R_\infty\left(\mathbf{S}_{\omega_o}\right) = -\ln\left(\omega_o - 1\right) \approx 0{,}0283,$$

$$R_\infty\left(S_1\right) = -p\ln\overline{\mu} \approx 0{,}0002,$$

$$R_\infty(\mathbf{B}) = \frac{R_\infty\left(S_1\right)}{p} \approx 0{,}0001.$$

Aqui a fração

$$\frac{R_\infty\left(S_{\omega_o}\right)}{R_\infty\left(S_1\right)} \approx \frac{0{,}0283}{0{,}0002} = 141{,}5$$

das razões assintóticas de convergência indica, a grosso modo, que, para obter exatidões comparáveis, precisa mais de 100 iterações do método do Gauss-Seidel para uma iteração do SOR. □

2.5. Variação da razão de convergência com p

No capítulo 1 mostramos que certas matrizes podem ser dispostas de modo que sejam p-cíclicas para diversos valores de p. Então vale a pena verificar como varia a razão de convergência do SOR com p. Para isso, olhemos para a relação (2.21) e ponhamos

$$\varphi(p) := \sqrt{\frac{2p}{p-1}}, \quad p \geq 2.$$

Temos

$$\varphi'(p) = \frac{-1}{\sqrt{2p(p-1)^3}} < 0, \text{ para todo } p \geq 2.$$

Então φ é estritamente decrescente com o máximo em $p = 2$. Essa conclusão é interessante porque, como já comentamos, o caso $p = 2$ é freqüente na discretização de equações diferenciais que modelam problemas de Física.

2.6. Considerações práticas sobre o SOR

Na busca de soluções numéricas de problemas reais por métodos iterativos estacionários, não é normalmente viável determinar previamente o raio espectral das matrizes de iteração de blocos de Jacobi e de Gauss-Seidel. O recurso é, então, estimar o valor ótimo do parâmetro do SOR a partir de estimativas do raio espectral $\rho(\mathbf{B})$ da matriz de Jacobi, usando a (2.1). Por essa relação facilmente podemos ver que a toda perturbação de $\rho(\mathbf{B})$ corresponde uma perturbação de ω_o de mesmo sinal. Contudo Young [57] demonstrou que a repercussão sobre ω_o de uma pequena perturbação de $\rho(\mathbf{B})$ é menor se esta última for positiva. A razão disso fica bastante clara para $p = 2$, visualizando a situação geometricamente.

Para fazer essa visualização, recorremos ao seguinte resultado, que vem de resolver a equa-

ção (1.34) dentro das condições do Teorema 2.2.6, com $p = 2$,

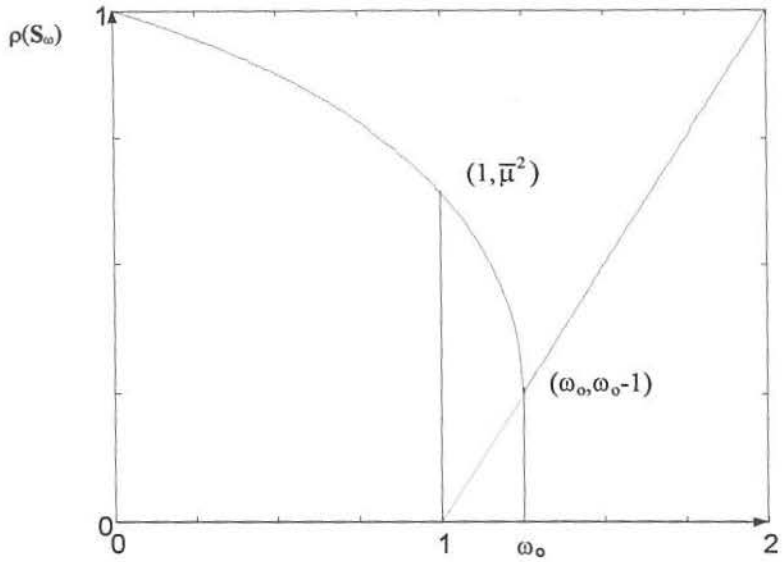


Fig.2.6 –Comportamento de $\rho(S_\omega)$ versus ω

$$\rho(S_\omega) = \begin{cases} \left(\dfrac{\omega\bar{\mu} + \sqrt{(\omega^2\bar{\mu}^2 - 4(\omega-1)}}{2} \right)^2, & \text{se } 0 < \omega \le \omega_o \\ \omega - 1, & \text{se } \omega_o \le \omega < 2. \end{cases} \tag{2.23}$$

Com esse resultado, construímos a Fig.2.6, onde a linha vermelha mostra o comportamento de $\rho(S_\omega)$ versus ω. Em particular, revela o fato de que a derivada à esquerda de ρ no ponto ω_o é infinita, motivo pelo qual ρ é muito sensível a perturbações negativas de ω_o.

2.6.1. Exemplo. Neste exemplo comparamos as velocidades de convergência do SOR, na busca de solução numérica de um SELAS particular, relativas a diversos valores de ω. Todos os procedimentos foram executados com o MATLAB, versão 5.3, num computador Pentium III, com 128 MB de memória e 450 MHz de velocidade-relógio.

A matriz dos coeficientes foi a matriz 2-cíclica consistentemente ordenada (cf. Exemplo 1.9.4) de ordem 3000×3000, gerada com `tridiag(3000)`:

$$\mathbf{A} := \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

A razão da escolha dessa matriz **A**, que se origina da discretização com diferenças centrais e nodos eqüiespaçados, da equação unidimensional de Poisson, para nossa ilustração, é dupla:

1) essa matriz é spd, pois seus autovalores são [14]

$$\lambda_i = 2 - 2\cos\left(\frac{\pi i}{N+1}\right) > 0,\ i = 1, 2, \ldots, N,$$

onde N é a ordem de **A** (no nosso caso $N = 3000$), e, portanto, o método SOR convergirá para todo valor $\omega \in (0\ ;\ 2)$, conforme o Teorema de Ostrowski-Reich (Teorema 1.7.3.2.2);

2) como o raio espectral (obtido com a função `eigs` do MATLAB) da correspondente matriz de Gauss-Seidel é muito próximo de 1, a convergência do método de Gauss-Seidel é muito lenta, como já mostramos no Exemplo 2.4.1, e a velocidade da convergência do SOR é muito sensível em relação à variação de ω.

O vetor **b** do segundo membro foi um vetor gerado de forma aleatória com componentes 0 e 1, obtido com `b=randint(3000,1,2)`.

O vetor inicializador $\mathbf{x}_0$ das iterações foi o vetor nulo.

O parâmetro de relaxação ótimo foi obtido com a fórmula (2.13), após o cálculo do raio espectral da matriz **B** de Jacobi com `max(eig(full(B)))`, e é

$$\omega_o = 1.997908492672649.$$

O comando, usado sucessivamente, para os diversos valores do parâmetro de relaxação w (ω), foi este

```
[x,er,iter,c]=jasor(A,b,x,w,1.5e+6,1e-6);
```

ω	**Iterações**	ω	**Iterações**	ω	**Iterações**	ω	**Iterações**
0.100	954307	1.40000	495866	ω_o	**4957**	1.999960	175352
0.200	928471	1.50000	442514	1.99900	7141	1.999970	232501
0.300	902580	1.60000	383823	1.99950	14313	1.999975	280373
0.400	875346	1.70000	317985	1.99960	18007	1.999980	349457
0.500	846343	1.80000	241779	1.99980	36089	1.999983	409607
0.600	815507	1.90000	148032	1.99985	46625	1.999985	466467
0.700	782877	1.91000	137080	1.99988	58457	1.999987	535642
0.800	748465	1.92000	125709	1.99990	70358	1.999988	580628
0.900	712216	1.93000	113858	1.99991	79106	1.999989	634574
1.000	**674000**	1.94000	101451	1.99992	88245	1.999990	697612
1.100	633609	1.95000	88382	1.99993	100349	1.999991	775619
1.200	590750	1.96000	74502	1.99994	118156	1.999992	871600
1.300	545019	1.97000	59584	1.999950	139502	1.999993	994785

Tab. 2.1 – Número de iterações do SOR versus ω

Os resultados foram organizados na Tab.2.1 e visualizados na Fig.2.7 e na Fig.2.8. Aí vemos que o número de iterações está entre um mínimo de 4 957 (correspondente ao parâmetro ótimo) e um máximo de 994 785. Os tempos de execução (não mostrados na tabela) variaram entre 51.57 seg e 2 h 51 min 40 seg. É interessante comparar as Fig.2.6 e Fig.2.7. A Fig.2.8 amplia a escala da porção da Fig.2.7 relativa à variação do parâmetro no intervalo $[\omega_o\ ;\ 1.999993]$, onde vemos mais claramente o comportamento do número de iterações do SOR com a variação do parâmetro à direita do valor ótimo.

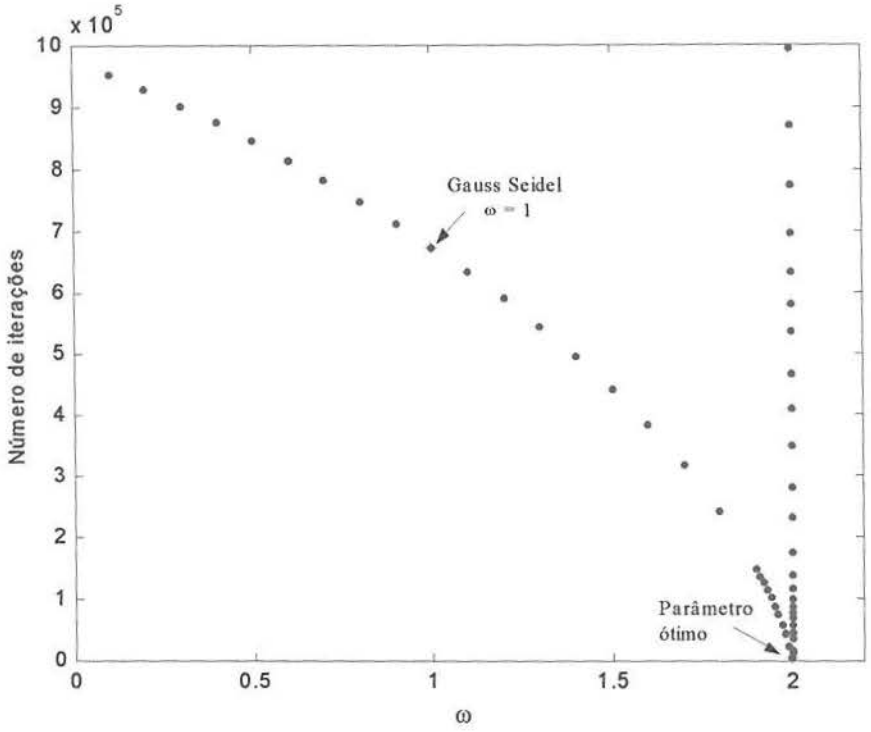


Fig.2.7 –Visualização da Tab.2.1

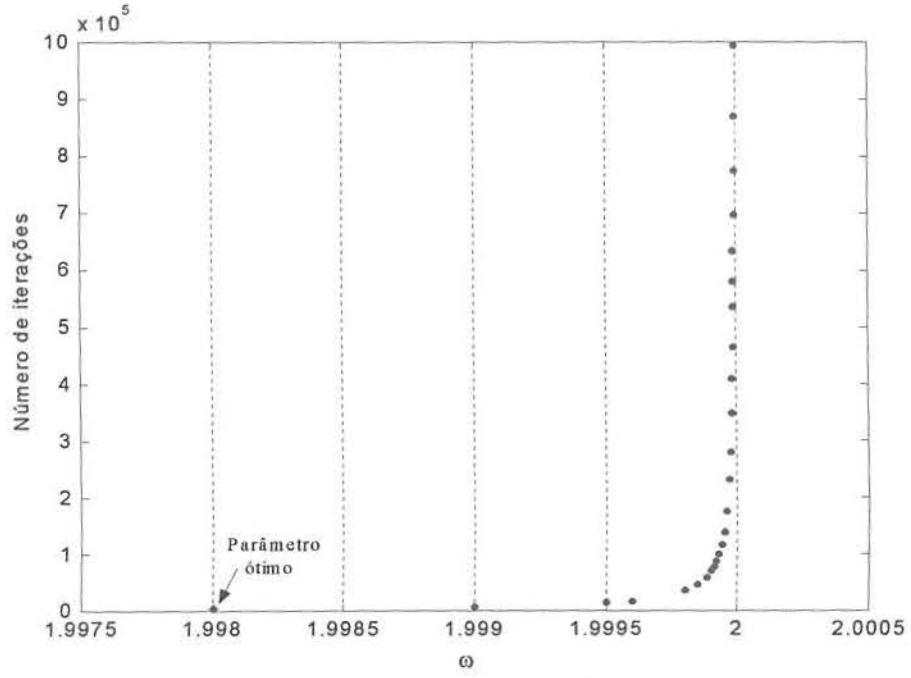


Fig. 2.8 – Retângulo [1.9975 ; 2.0005] × [0 ; 10×10^5] com zum, contido na Fig.2.7

Na Tab.2.1 e Fig.2.7 salientamos, além do caso do parâmetro ótimo, o caso Gauss-Seidel. Constatamos que o SOR com o valor ótimo do parâmetro é em torno de 136 vezes mais rápido que o método de Gauss-Seidel. Entretanto, notamos o que também é teoricamente previsto, que, se $\omega < 1$, ou se ω é muito próximo à esquerda de 2, o SOR é mais lento que o Gauss-Seidel; para $\omega = 0.1$, por exemplo, este último é quase uma vez e meia mais rápido que o SOR. □

2.7 Otimalidade relativa a *p* do SOR p-cíclico

A matriz $\mathbf{A}_1$ (aqui vamos indicá-la com $\mathbf{A}_p$) p-cíclica consistentemente ordenada em (1.25) pode ser particionada de várias maneiras numa matriz q-cíclica, com $2 \le q < p$ (cf. Exemplo 1.4.10). Por exemplo, a matriz 4-cíclica

$$\mathbf{A}_4 := \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{0} & \mathbf{A}_{14} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{32} & \mathbf{A}_{33} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{43} & \mathbf{A}_{44} \end{bmatrix}$$

pode ser particionada nas formas 2-cíclica e 3-cíclica, respectivamente,

$$\mathbf{A}_2 = \left[\begin{array}{c|ccc} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{0} & \mathbf{A}_{14} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{32} & \mathbf{A}_{33} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{43} & \mathbf{A}_{44} \end{array}\right] \quad \text{e} \quad \mathbf{A}_3 = \left[\begin{array}{ccc|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{0} & \mathbf{A}_{14} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{32} & \mathbf{A}_{33} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{A}_{43} & \mathbf{A}_{44} \end{array}\right].$$

Em torno disso, em 1990, Pierce, Hadjidimos e Plemmons [23] demonstraram um resultado, que impressionou muito, e, dois anos mais tarde, Eiermann, Niethammer e Ruttan [17] deram ao mesmo resultado outra demonstração mais elegante. Enunciemos esse resultado.

2.7.1. Teorema. *Seja* $\mathbf{A}_p$ *a matriz p-cíclica consistentemente ordenada* (1.25). *Particionemos* $\mathbf{A}_p$ *em uma matriz q-cíclica* $\mathbf{A}_q$ *de blocos, com* $2 \le q < p$. *Sejam* $\mathbf{B}_p$ *e* $\mathbf{B}_q$ *as matrizes de iteração de Jacobi fracamente cíclicas associadas, de índices respectivos* p *e* q. *Suponhamos que o espectro de* $\mathbf{B}_p^p$ *e o raio espectral de* $\mathbf{B}_p$ *satisfaçam*

- $\sigma\left(\mathbf{B}_p^p\right) \subset [0; \beta^p]$ e $0 < \beta := \rho\left(\mathbf{B}_p\right) < 1$,

ou

- $\sigma\left(\mathbf{B}_p^p\right) \subset (-\alpha^p; 0]$ e $0 < \alpha := \rho\left(\mathbf{B}_p\right) < \dfrac{p}{p-2}$.

Então o método SOR q-cíclico ótimo é assintoticamente mais rápido que o $(q+1)$*-cíclico ótimo, isto é,*

$$\rho\left(\mathbf{S}_{\omega_q}\right) < \rho\left(\mathbf{S}_{\omega_{q+1}}\right),$$

onde ω_q *denota o valor ótimo do parâmetro do método SOR q-cíclico.*

O que diz o Teorema 2.7.1 é que a função $q \mapsto \rho\left(\mathbf{S}_{\omega_q}\right)$ é estritamente crescente no intervalo $[2\ ;\ p]$. Uma conseqüência desse teorema é que, nas condições de seu enunciado sobre o espectro de $\mathbf{B}_p^p$ e o raio espectral de $\mathbf{B}_p$, com $p \geq 3$, temos

$$\rho\left(\mathbf{S}_{\omega_2}\right) < \rho\left(\mathbf{S}_{\omega_p}\right),$$

o que significa que o SOR 2-cíclico é o ótimo em termos da convergência assintótica, o que valoriza os resultados de Young [55] e faz jus ao fato de que os SELAS 2-cíclicos ocorrem com freqüência.

Poderíamos objetar que a transformação de $\mathbf{A}_p$ em $\mathbf{A}_q$, $2 \leq q < p$, possivelmente acarreta mais flops por iteração para o SOR aplicado a $\mathbf{A}_q$. Ocorre, entretanto, que o trabalho é, de fato, o mesmo para todas as formas cíclicas de uma matriz fixa $\mathbf{A}$. Para ver isso, basta comparar o custo do SOR q-cíclico com o custo relativo ao Gauss-Seidel, pois, se representarmos a $(n+1)^{\text{ésima}}$ iteração de Gauss-Seidel por

$$\mathbf{x}^{(n)} = \mathbf{x}^{(n-1)} + \mathbf{y}^{(n+1)},$$

onde $\mathbf{y}^{(n+1)}$ depende de $\mathbf{x}^{(n)}$, então o SOR toma a forma

$$\mathbf{x}^{(n)} = \mathbf{x}^{(n-1)} + \omega \mathbf{y}^{(n+1)}.$$

Mas os custos para uma iteração nos casos p-cíclico e q-cíclico de blocos, $2 \leq q < p$, para $\omega = 1$, são matematicamente equivalentes, porque, aqui, (1.15), escrito matricialmente, se torna

$$(\mathbf{D} - \mathbf{L})\mathbf{x}^{n+1} = \mathbf{U}\mathbf{x}^n + \mathbf{b},$$

e $\mathbf{U}$ envolve apenas o bloco $\mathbf{A}_{1p}$ de $\mathbf{A}_p$ em (1.25) (neste referido número, está $\mathbf{A}_1$ em vez de $\mathbf{A}_p$) independentemente do particionameneto de $\mathbf{A}_p$ em blocos para o caso q-cíclico, $q \geq 2$.

No entanto, se as condições do Teorema 2.7.1 não forem preenchidas, pode não ocorrer a otimalidade do SOR 2-cíclico ótimo entre os q-cíclicos ótimos. De fato, em 1994, Eiermann et al. [17] mostraram que, se a matriz de Jacobi $\mathbf{B}_p$ associada a um SELAS é fracamente cíclica de índice $p > 2$ e

$$-\alpha^p, \beta^p \in \sigma\left(\mathbf{B}_p^p\right) \subset [-\alpha^p; \beta^p], \text{ com } \alpha := \frac{(p-2)\beta}{p} \text{ e } 0 < \boldsymbol{\beta} < \mathbf{1},$$

então

$$\rho\left(\mathbf{S}_{\omega_p}\right) < \rho\left(\mathbf{S}_{\omega_2}\right),$$

em outros termos, o método SOR p-cíclico ótimo é assintoticamente mais rápido que o 2-cíclico ótimo.

3 – Novo SOR

3.1. Introdução

O método SOR teve sua consagração com a tese de Young [56] em 1950 e depois com Varga [51]. Posteriormente o método ficou ofuscado pelos métodos não-estacionários, mas despertou com muita força no início da última década [12, 13, 19, 26, 31, 40, 41, 42, 52, 53, 54], parcialmente devido à computação paralela. A idéia de usar o SOR como precondicionador dos métodos da família do gradiente conjugado, devida a DeLong e Ortega [12, 14] também impulsionou a investigação. Atualmente continua a ser campo ativo de pesquisa. Neste capítulo queremos abordar um aspecto dessa pesquisa [26], especificamente, uma modificação e generalização do método SOR, aplicável a uma classe de matrizes mais ampla que a classe das p-cíclicas consistentemente ordenadas.

Introduziremos as matrizes-escada, que, analogamente às matrizes triangulares, servem de base para novo tipo útil de decomposição de matrizes para métodos iterativos estacionários. O método resultante tem a vantagem do método de Jacobi (adequado à computação paralela) e do SOR (mais rapidamente convergente).

3.2. Matrizes-escada

3.2.1. Definição. Uma matriz tridiagonal $\mathbf{A} = [a_{ij}]$ com $i, j = 1, 2, \ldots, N$, onde a_{ij} pode ser número ou matriz de ordem $N_i \times N_j$, é dita uma *matriz-escada* sse uma das duas condições seguintes é satisfeita:

1. $a_{12} = 0$, e $a_{ii-1} = a_{ii+1} = 0$ para $i = 3, 5, \ldots, N-1$, se N é par,
   e $a_{12} = 0$, $a_{NN-1} = 0$, e $a_{ii-1} = a_{ii+1} = 0$ para $i = 3, 5, \ldots, N-2$, se N é ímpar;
2. $a_{NN-1} = 0$, e $a_{ii-1} = a_{ii+1} = 0$ para $i = 2, 4, \ldots, N-2$ se N é par,
   e $a_{ii-1} = a_{ii+1} = 0$ para $i = 2, 4, \ldots, N-1$ se N é ímpar.

Usaremos as denominações matriz-escada do *tipo* 1 ou do *tipo* 2, de acordo com a condição satisfeita na Definição 3.2.1.

3.2.1.1. Exemplo. Matrizes-escada dos tipos 1 e 2 de ordem 6 têm a seguinte estrutura respectiva:

$$\begin{bmatrix} x & & & & & \\ x & x & x & & & \\ & & x & & & \\ & & x & x & x & \\ & & & & x & \\ & & & & x & x \end{bmatrix}, \quad \begin{bmatrix} x & x & & & & \\ & x & & & & \\ & x & x & x & & \\ & & & x & & \\ & & & x & x & x \\ & & & & & x \end{bmatrix}. \square$$

É claro que uma matriz-escada $\mathbf{A} = [a_{ij}]$ é não-singular sse seus blocos diagonais são não-singulares e, nesse caso, se indicamos com **D** a matriz bloco-diagonal de **A**, verificamos, meramente fazendo os cálculos, que,

$$\mathbf{A}^{-1} = \mathbf{D}^{-1}(2\mathbf{D} - \mathbf{A})\mathbf{D}^{-1}.$$

O algoritmo seguinte objetiva aproveitar ao máximo a estrutura de uma matriz-escada $\mathbf{A} = [a_{ij}]$, para resolver um SELAS $\mathbf{Ax} = \mathbf{b}$.

3.2.2. Algoritmo para matriz-escada (no MATLAB: `escada.m`).

Dados de entrada: **A**, **b**.
Saída: solução do SELAS $\mathbf{Ax} = \mathbf{b}$.
se A é do tipo 1,
para $i = 1{:}2{:}\,2\left\lfloor\frac{N-1}{2}\right\rfloor+1$,
$b_i = a_{ii}^{-1} b_i$;
fim para
para $i = 2{:}2{:}\,2\left\lfloor\frac{N}{2}\right\rfloor$,
$b_i = a_{ii}^{-1}\left(b_i - a_{ii-1}b_{i-1} - a_{ii+1}b_{i+1}\right)$;
fim para
fim se
se A é do tipo 2,
para $i = 2{:}2{:}\,2\left\lfloor\frac{N}{2}\right\rfloor$,
$b_i = a_{ii}^{-1} b_i$;
fim para
para $i = 1{:}2{:}\,2\left\lfloor\frac{N-1}{2}\right\rfloor+1$,
$b_i = a_{ii}^{-1}\left(b_i - a_{ii-1}b_{i-1} - a_{ii+1}b_{i+1}\right)$;
fim para
fim se
onde $b_i = 0$ se $i < 1$ ou $i > N$.

Se **A** é uma matriz de escalares, o algoritmo requer apenas $3N$ operações aritméticas, N adições, N multiplicações e N divisões.

Se **A** é uma matriz de blocos, são requeridas N multiplicações matriz-vetor, N adições vetoriais e a resolução de N subsistemas lineares, sendo, portanto, o algoritmo, adequado para computação paralela (cf. Apêndice *A*). Por exemplo, se **A** é uma matriz-escada do tipo 1, primeiro para i ímpar, os cálculos de $a_{ii}^{-1}b_i$ podem ser feitos por diferentes processadores simultaneamente; e, em seguida, para i par, os valores $a_{ii}^{-1}\left(b_i - a_{ii-1}b_{i-1} - a_{ii+1}b_{i+1}\right)$ são calculados paralelamente. Ainda maior paralelismo é alcançado, se **A** é de escalares complexos.

Testamos o algoritmo `escada` no MATLAB com uma matriz-escada não-singular de ordem 10 000, na forma esparsa, num computador de 750 MHz de velocidade e 256 MB de memória, e levou apenas 0.4400 seg para fornecer a solução.

3.2.3. Ampliação da classe das matrizes-escada. Daqui em diante usaremos a notação $\mathbf{A} = [\mathbf{A}_{ij}]$,

quando quisermos salientar a decomposição de $\mathbf{A}$ em blocos $\mathbf{A}_{ij}$.

Com o objetivo de obter uma boa decomposição matricial para um método iterativo (1.9), adequada à paralelização, ampliamos a classe das matrizes-escada para uma classe $\mathcal{L}_N^K$, mediante a definição indutiva:

- $\mathcal{L}_N^1 := \{\mathbf{A} \mid \mathbf{A}$ é uma matriz-escada $N \times N$ de escalares $\}$;
- $\mathcal{L}_N^K := \{\mathbf{A} \mid \mathbf{A} = [\mathbf{A}_{ij}]$ é uma matriz-escada $N \times N$ de blocos e cada bloco diagonal $\mathbf{A}_{ii}$ é uma matriz de ordem $N_i \times N_i$, com $\mathbf{A}_{ii} \in \mathcal{L}_{N_i}^r$ se $r < K\}$.

Segue diretamente da definição que

$$\mathcal{L}_n^1 \subset \mathcal{L}_n^2 \subset \cdots \subset \mathcal{L}_n^n \subset \cdots$$

e, mediante indução sobre N, que

$$\mathcal{L}_N^K = \mathcal{L}_N^N, \quad \text{para } K \geq N.$$

Introduzimos a definição: $\mathcal{L}_N := \mathcal{L}_N^N$. Resulta que todas as matrizes triangulares de ordem N pertencem à classe $\mathcal{L}_N$, assim como pertencem a essa classe todas as matrizes quadradas de ordem N com as estruturas

$$\begin{bmatrix} x & & & & & & \\ x & x & x & x & x & x & \cdots \\ x & & x & & & & \\ x & & x & x & x & x & \cdots \\ x & & x & & x & & \\ x & & x & & x & x & \cdots \\ x & & x & & x & & \\ \vdots & & \vdots & & \vdots & & \end{bmatrix} \quad \text{e} \quad \begin{bmatrix} x & x & x & x & x & x & \cdots \\ & x & & & & & \\ & x & x & x & x & x & \cdots \\ & x & & x & & & \\ & x & & x & x & x & \cdots \\ & x & & x & & x & \\ & \vdots & & \vdots & & \vdots & \end{bmatrix},$$

que chamaremos *matrizes-zebra*.

O determinante das matrizes triangulares de blocos e, evidentemente, das matrizes-zebra é o produto dos blocos diagonais. Essa propriedade se estende para toda matriz da classe $\mathcal{L}_N$. Mais, a conjugada transposta de uma matriz complexa da classe $\mathcal{L}_N$ pertence a $\mathcal{L}_N$ e, se é não-singular, sua inversa também está nessa classe [26].

Se $\mathbf{A} = [\mathbf{A}_{ij}]$ é uma matriz-escada em $\mathcal{L}_N$, podemos aplicar o Algoritmo 3.2.1 repetidamente para resolver um SELAS $\mathbf{Ax} = \mathbf{b}$. Particionamos $\mathbf{b} = [\mathbf{b}_1 \ldots \mathbf{b}_m]^t$ de acordo com a partição de $\mathbf{A}$, sendo m o número de blocos diagonais de $\mathbf{A}$. Se N_i é a ordem do bloco $\mathbf{A}_{ii}$, cada produto $\mathbf{A}_{ik}\mathbf{b}_k$, com $k = i - 1$, $i + 1$, necessita no máximo N_iN_k multiplicações e $(N_i - 1)N_k$ adições, e o cálculo de $\mathbf{b}_i - \mathbf{A}_{ii-1}\mathbf{b}_{i-1} - \mathbf{A}_{ii+1}\mathbf{b}_{i+1}$ necessita de no máximo $2(N_iN_{i-1} + N_iN_{i+1})$ operações aritméticas. Então, se $L(N)$ é o número de operações aritméticas para resolver um SELAS $\mathbf{Ax} = \mathbf{b}$ de ordem N,

$$L(N) \leq \sum_{i=1}^{m} L(N_i) + 2\sum_{i=1}^{m-1} N_iN_{i+1}.$$

Usando agora a indução, obtemos $L(N) \leq N^2$. Portanto, o custo para resolver um SELAS $\mathbf{Ax} = \mathbf{b}$, com $\mathbf{A}$ em $\mathcal{L}_N$ não é maior que o exigido para resolver um SELAS triangular. No entanto, muitos SELAS-escada

esparsos são resolvidos facilmente em paralelo, executando repetidamente o Algoritmo 3.2.1.

3.3. Novo processo SOR

Acima vimos que o método iterativo (1.9) é facilmente paralelizado, se $\mathbf{C} \in \mathcal{L}_N$. Nesta secção generalizamos o método SOR baseando-o na decomposição $\mathbf{A} = \mathbf{C} - \mathbf{R}$, onde $\mathbf{C} \in \mathcal{L}_N$.

Seja **A** uma matriz não-singular, com os elementos diagonais não-nulos. Tomamos a decomposição $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$, onde **D** é a diagonal de **A**, as diagonais de **P** e **Q** são nulas, e $\mathbf{P} \in \mathcal{L}_N$. Dado um SELAS $\mathbf{Ax} = \mathbf{b}$, definimos o novo processo SOR por

$$\mathbf{x}^{(n+1)} = (\mathbf{D} - \omega\mathbf{P})^{-1}\left((1-\omega)\mathbf{D} + \omega\mathbf{Q}\right)\mathbf{x}^{(n)} + \omega(\mathbf{D} - \omega\mathbf{P})^{-1}\mathbf{b}, \quad n = 0, 1, 2, \ldots \tag{3.1}$$

Aqui também, as matrizes

$$\mathbf{S}_\omega := (\mathbf{D} - \omega\mathbf{P})^{-1}\left((1-\omega)\mathbf{D} + \omega\mathbf{Q}\right) \quad \text{e} \quad \mathbf{b}_\omega := \omega(\mathbf{D} - \omega\mathbf{P})^{-1} \tag{3.2}$$

são ditas a matriz do SOR e o vetor do SOR, respectivamente. Vemos que, formalmente, o novo SOR é análogo ao clássico, com **P** em lugar de **L** e **Q** em lugar de **U**, mas a diferença efetiva é substancial.

O Teorema de Kahan (Teorema 1.7.3.2.1) mantém-se com a mesma argumentação para a demonstração.

3.3.1. Parâmetro ótimo. No novo SOR os resultados e as demonstrações no contexto da teoria da determinação do parâmetro ótimo são essencialmente os mesmos que os do SOR clássico. Inicialmente estendemos o conceito de matriz consistentemente ordenada p-cíclica.

3.3.1.1. Definição. Seja uma matriz $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$, onde **D** é a diagonal de **A**. Se existe um número natural $p > 1$ tal que

$$\det\left(\beta\mathbf{D} - \alpha\mathbf{P} - \alpha^{-(p-1)}\mathbf{Q}\right) = \det\left(\beta\mathbf{D} - \mathbf{P} - \mathbf{Q}\right),$$

para quaisquer constantes $\alpha \neq 0$ e β, dizemos que **A** é uma matriz *p-consistentemente ordenada em relação a* (**P**, **Q**).

A Definição 1.9.1 de *matriz* **A** *consistentemente ordenada p-cíclica* é equivalente à Definição 3.3.1.1 de *matriz p-consistentemente ordenada em relação a* (**L**, **U**). Aqui **L** e **U** são as partes triangulares estritamente inferior e superior de **A**, respectivamente.

O seguinte teorema é do tipo do Teorema de Romanovsky (Teorema 1.11.2) e mostra como é o polinômio característico da matriz de Jacobi **B** correspondente a uma matriz **A** p-consistentemente ordenada em relação a (**P**, **Q**).

3.3.1.2. Teorema. *Se* **A** *é uma matriz de ordem* N×N, *com diagonal não-singular, p-consistentemente ordenada em relação a* (**P**, **Q**), *então*

$$\det(t\mathbf{I} - \mathbf{B}) = t^n \prod_{i=1}^{r}\left(t^p - \mu_i^p\right),$$

onde $n + rp = N$ *e os* μ_i *são autovalores não-nulos de* **B.**

O teorema a seguir é o análogo ao Teorema 1.11.3 e fornece a relação entre os autovalores da matriz **B** de Jacobi e os da matriz $\mathbf{S}_\omega$ do novo SOR.

3.3.1.3. Teorema. *Seja uma matriz* $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$ *p-consistentemente ordenada em relação a* (**P**, **Q**), *de ordem* N×N, *onde* **D** *é a matriz bloco-diagonal dos blocos diagonais, quadrados e não-singulares, de* **A** *e* $\mathbf{P} \in \mathcal{L}_N$, *e seja* $\omega \neq 0$.

(a) *Se* $\lambda \neq 0$ *é um autovalor da matriz* $\mathbf{S}_\omega$ *em* (3.2) e se μ *satisfaz*

$$(\lambda + \omega - 1)^p = \lambda^{p-1}\omega^p\mu^p, \tag{3.3}$$

então μ *é um autovalor da matriz de Jacobi* **B** correspondente a **A**. *Reciprocamente*,

(b) *se* μ *é um autovalor da matriz* **B** *e se* λ *satisfaz* (3.3), *então* λ *é um autovalor de* $\mathbf{S}_\omega$.

Com o fato $\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{P} \in \mathcal{L}_N$, fácil de verificar, e os resultados apresentados acima, a demonstração do Teorema 3.3.1.3 segue o mesmo esquema que a do Teorema 1.11.3.

Se $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$ satisfaz as condições do Teorema 3.3.1.3, o Teorema 2.2.5 se mantém ipsis litteris para o novo SOR. Para $p = 2$, a fórmula (2.13) continua a valer para o novo SOR, assim como o resultado (2.23) de Young [55].

A seguir descreveremos vários exemplos para mostrar que o novo SOR, não apenas é facilmente paralelizado, mas também, para ampla classe de matrizes, produz convergência tão rápida quanto o SOR clássico.

No que segue usaremos "escada(a_{ii-1}, a_{ii}, a_{ii+1})" para simbolizar uma matriz escada com elementos a_{ii} na diagonal, e a_{ii-1} e a_{ii+1}, na subdiagonal e na superdiagonal, respectivamente, contíguas à diagonal principal. Também usaremos "escada1(a_{ii-1}, a_{ii}, a_{ii+1})" e "escada2(a_{ii-1}, a_{ii}, a_{ii+1})" para especificar matrizes-escada dos tipos 1 e 2, respectivamente.

3.3.1.4. Exemplo. Seja uma matriz tridiagonal $\mathbf{A} = [a_{ij}]$. Então **A** é 2-consistentemente ordenada em relação a (**L**, **U**), onde, como sempre, **L** é a parte triangular estritamente inferior e **U**, a parte triangular estritamente superior de **A**. Consideremos a decomposição $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$, onde

$$\mathbf{P} = -\text{escada1}(a_{ii-1}, 0, a_{ii+1}),\ \mathbf{Q} = -\text{escada2}(a_{ii-1}, 0, a_{ii+1}).$$

Então $\det\left(\beta\mathbf{D} - \alpha\mathbf{P} - \alpha^{-1)}\mathbf{Q}\right)$ é independente de $\alpha \neq 0$, porque

$$\tilde{\mathbf{D}}^{-1}\left(\beta\mathbf{D} - \alpha\mathbf{P} - \alpha^{-1}\mathbf{Q}\right)\tilde{\mathbf{D}} = \beta\mathbf{D} - \mathbf{P} - \mathbf{Q},$$

onde

$$\tilde{\mathbf{D}} := \text{diag}(\alpha_i\mathbf{I}) \text{ e } \alpha_i := \begin{cases} 1, & \text{se } i \text{ é impar} \\ \alpha, & \text{se } i \text{ é par.} \end{cases}$$

Então, o método SOR clássico (1.15) e o novo (3.1) têm a mesma razão de convergência assintótica ótima, pois, para ambos, o parâmetro ótimo é dado por (2.13), mas o segundo é implementado muito mais facilmente na computação paralela. □

3.3.1.5. Exemplo. Neste exemplo mostramos que as condições que possibilitam a determinação do parâmetro ótimo do novo SOR se mantêm para uma grande classe de matrizes, para as quais as correspondentes condições para o SOR clássico falham. Consideremos uma matriz de ordem $2m \times 2m$ com a

estrutura

$$\mathbf{A} := \text{tridiag}(a_{ii-1}, a_{ii}, a_{ii+1}) + \begin{bmatrix} 0 & 0 & \cdots & 0 & a_{1,2m} \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ a_{2m,1} & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Está mostrado em [51, 56] (cf. Exemplo 1.9.5) que **A** não é p-consistentemente ordenada em relação a (**L**, **U**). Então a teoria do SOR clássico não é aplicável para matrizes com a estrutura de **A**.

Contudo, definamos uma matriz-zebra $\mathbf{P} = [p_{ij}]$ de ordem $2m \times 2m$ por

$$p_{ij} := \begin{cases} -a_{ij}, \text{ para } j = i-1, i+1, i = 2, 4, \cdots, 2m-2, \\ -a_{ij}, \text{ para } j = 1, 2m-1, i = 2m, \\ 0, \text{ em outro caso,} \end{cases}$$

e decomponhamos $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$. Então $\det\left(\beta\mathbf{D} - \alpha\mathbf{P} - \alpha^{-1}\mathbf{Q}\right)$ é independente de $\alpha \neq 0$, porque

$$\tilde{\mathbf{D}}^{-1}\left(\beta\mathbf{D} - \alpha\mathbf{P} - \alpha^{-1}\mathbf{Q}\right)\tilde{\mathbf{D}} = \beta\mathbf{D} - \mathbf{P} - \mathbf{Q},$$

onde $\tilde{\mathbf{D}} := \text{diag}(\mathbf{I}_1, \alpha\mathbf{I}_2, \mathbf{I}_3, \alpha\mathbf{I}_4, \cdots, \mathbf{I}_{2m-1}, \alpha\mathbf{I}_{2m})$. Aqui as $\mathbf{I}_j$ são matrizes identidades com ordens convenientes. Em conseqüência, **A** é 2-consistentemente ordenada em relação a (**P**, **Q**). □

3.3.1.6. Exemplo. As vantagens do paralelismo em cada iteração, e a convergência rápida, do novo SOR surgem na aplicação do método a matrizes provenientes da discretização de equações diferenciais parciais em espaços de dimensão elevada. Definamos indutivamente uma classe de matrizes por

- $\mathcal{T}_1 := \{\mathbf{A} \mid \mathbf{A} = \text{tridiag}(a_{ii-1}, a_{ii}, a_{ii+1})\}$;
- $\mathcal{T}_k := \{\mathbf{A} \mid \mathbf{A} = \text{tridiag}(\mathbf{A}_{ii-1}, \mathbf{A}_{ii}, \mathbf{A}_{ii+1})$, $\mathbf{A}_{ii-1}$, $\mathbf{A}_{ii+1}$ são matrizes diagonais e $\mathbf{A}_{ii} \in \mathcal{T}_{k-1}\}$.

Muitas matrizes, que se originam na discretização de equações diferenciais parciais, pertencem a essa classe de matrizes. Por exemplo, as matrizes de diferenças provenientes das equações parciais elípticas autoadjuntas a k variáveis estão em $\mathcal{T}_k$. Semelhantemente definimos

- $\mathcal{T}_1^e := \{\mathbf{A} \mid \mathbf{A} = \text{escada}(a_{ii-1}, a_{ii}, a_{ii+1})\}$;
- $\mathcal{T}_k^e := \{\mathbf{A} \mid \mathbf{A} = \text{escada}(\mathbf{A}_{ii-1}, \mathbf{A}_{ii}, \mathbf{A}_{ii+1})$, $\mathbf{A}_{ii-1}$, $\mathbf{A}_{ii+1}$ são matrizes diagonais e $\mathbf{A}_{ii} \in \mathcal{T}_{k-1}^e\}$.

Seja $\mathbf{A} = \text{tridiag}(a_{ii-1}, a_{ii}, a_{ii+1}) \in \mathcal{T}_k$ e $\mathbf{D} := \text{diag}(\mathbf{D}_i)$ a diagonal de **A**, sendo $\mathbf{D}_i$ a diagonal de $\mathbf{A}_{ii}$ Queremos mostrar que existe uma decomposição $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$ com $\mathbf{P}, \mathbf{Q} \in \mathcal{T}_k^e$ e uma matriz diagonal $\hat{\mathbf{D}}$ tal que

$$\hat{\mathbf{D}}^{-1}\left(\beta\mathbf{D} - \alpha\mathbf{P} - \alpha^{-1}\mathbf{Q}\right)\hat{\mathbf{D}} = \beta\mathbf{D} - \mathbf{P} - \mathbf{Q}, \tag{3.4}$$

para quaisquer $\alpha \neq 0$ e β.

Isso é verdade para $k = 1$, como mostramos no Exemplo 3.3.1.4. Suponhamos que (3.4) se

mantenha para toda matriz em $\mathcal{T}_k$. Em particular, para cada bloco diagonal $\mathbf{A}_{ii}$, existem $\mathbf{D}_i$, $\mathbf{P}_i \in \mathcal{T}_k^e$ e uma matriz diagonal $\widehat{\mathbf{D}}_i$ tal que $\mathbf{A}_i = \mathbf{D}_i - \mathbf{P}_i - \mathbf{Q}_i$ e

$$\widehat{\mathbf{D}}_i^{-1}\left(\beta\mathbf{D}_i - \alpha\mathbf{P}_i - \alpha^{-1}\mathbf{Q}_i\right)\widehat{\mathbf{D}}_i = \beta\mathbf{D}_i - \mathbf{P}_i - \mathbf{Q}_i, \tag{3.5}$$

para quaisquer $\alpha \neq 0$ e β. Seja

$$\mathbf{P} := \text{escada1}(-\mathbf{A}_{ii-1}, \mathbf{P}_i, -\mathbf{A}_{ii+1}),\ \mathbf{Q} := \text{escada2}(-\mathbf{A}_{ii-1}, \mathbf{Q}_i, -\mathbf{A}_{ii+1}).$$

Segue imediatamente que $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$. Ponhamos $\widehat{\mathbf{D}}$:=diag$\left(\alpha_i\widehat{\mathbf{D}}_i\right)$, onde α_i é definido no Exemplo 3.3.1.4. Agora com (3.5) chegamos a (3.4).

A igualdade (3.4) mostra que $\mathbf{A}$ é 2-consistentemente ordenada em relação a $(\mathbf{P}, \mathbf{Q})$. Então os resultados teóricos sobre a determinação do parâmetro ótimo são aplicáveis às matrizes em $\mathcal{T}_k$. Outra vantagem da aplicação da decomposição $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$ no processo (3.1) é o completo paralelismo em cada iteração. □

3.3.1.7. Exemplo. Finalizamos o capítulo com uma comparação numérica do desempenho do SOR clássico com o do novo SOR. Nosso SELAS se origina na discretização de 5 pontos, mediante diferenças centrais, com passo uniforme h, do problema bidimensional de contorno de Poisson

$$\begin{cases} \nabla^2 u = 1, & (x,y) \in \Omega := (0;1)\times(0;1), \\ u/\partial\Omega = 0. \end{cases}$$

Os nodos da grade são enumerados conforme a ordem lexicográfica. Resulta um SELAS $\mathbf{Ax} = \mathbf{b}$, onde $\mathbf{A}$ é uma matriz tridiagonal de blocos

$$\mathbf{A} = \text{tridiag}(\mathbf{A}_{ii-1}, \mathbf{A}_{ii}, \mathbf{A}_{ii+1}),$$

com $\mathbf{A}_{ii-1} = \mathbf{A}_{ii+1} = -\mathbf{I}$ e , $\mathbf{A}_{ii}$= tridiag(−1, 4, −1) [4] Essa matriz, exceto quanto ao tamanho, é a do Exemplo 1.11.4 e sua estrutura está mostrada na Fig.3.1. O número de incógnitas será $(h^{-1} - 1)^2$.

Resolvemos o SELAS pelo método do SOR clássico (SORC) e do SOR novo (SORN). A matriz $\mathbf{A}$ dos coeficientes é 2-consistentemente ordenada em relação a $(\mathbf{L}, \mathbf{U})$, (cf. secções 1.8 e 1.9). Como $\mathbf{A} \in \mathcal{T}_2$, para o SORN, a decomposição $\mathbf{A} = \mathbf{D} - \mathbf{P} - \mathbf{Q}$ é obtida como explicado no Exemplo 3.3.1.6, e resulta que $\mathbf{A}$ é também 2-consistentemente ordenada em relação a $(\mathbf{P}, \mathbf{Q})$.

Tomamos $\mathbf{x}^{(0)} := [1, 1, \ldots, 1]^t$ para vetor inicializador. É sabido [01, 14] que o raio espectral da matriz $\mathbf{B}$ de Jacobi relativa a $\mathbf{A}$ é

$$\rho(\mathbf{B}) = \cos(\pi h),$$

e, portanto, usando a (2.13), obtemos o parâmetro ótimo

$$\omega_o(1/h) = \frac{2}{1+\text{sen}(\pi h)}.$$

[4] Esta matriz é gerada no MATLAB 5.3 com o comando `poisson(N)` em `tesmat`, onde N^2 é o número de nodos.

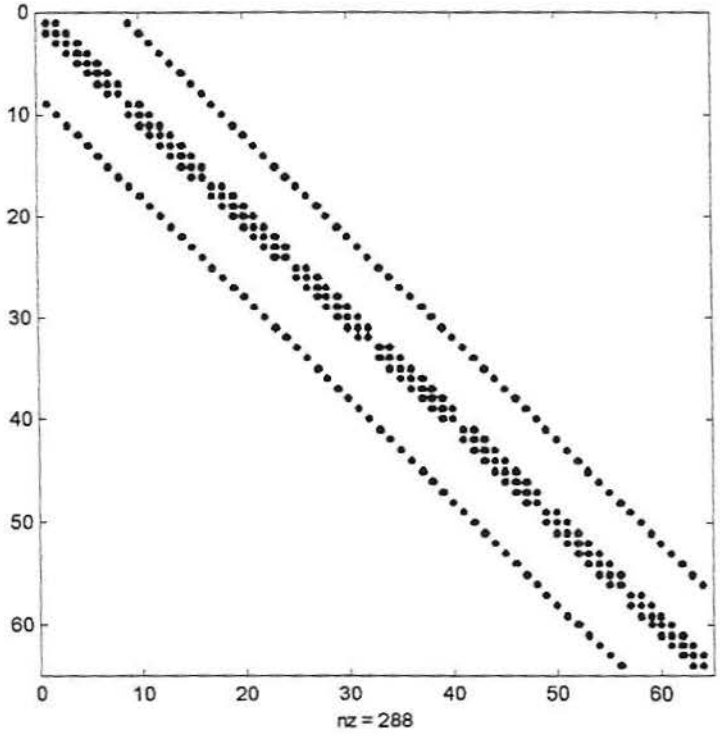


Fig.3.1 – Matriz de Poisson para problema bidimensional

Para critério de parada foi usado $\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(i)}\|_2 / \|\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}\|_2 < 10^{-5}$.

A Tab.3.1 mostra o número de iterações do SORC e do SORN para diversos valores de ω, com destaque para os casos de valor ótimo.

$1/h$	$\omega_o(8)$		$\omega_o(16)$		$\omega_o(32)$		$\omega_o(64)$		$\omega_o(128)$		$\omega_o(256)$	
	SORC	SORN	SORC	SORN	SORC	SORN	SORC	SORN	SORC	SORN	SORC	SORN
8	**19**	**17**	33	31	64	61	127	124	251	245	496	484
16	80	84	**36**	**34**	64	62	128	124	254	248	499	489
32	294	320	154	174	**69**	**69**	128	124	256	246	507	487
64	1034	1149	554	653	291	358	**132**	**129**	256	248	512	490
128	3553	4022	1908	2323	1016	1325	495	679	**259**	**258**	512	492
256	11875	13742	6371	8049	3397	4676	1675	2506	841	1315	**515**	**515**

Tab.3.1 – Comparação dos desempenhos do SOR e do SORN

Observamos na Tab.3.1 que, para o valor ótimo do parâmetro, o número de iterações de ambos os métodos são aproximadamente iguais, embora, para outros valores do parâmetro ocorram algumas diferenças.

Possivelmente, com as matrizes apresentadas neste capítulo, podemos, além do novo SOR, obter outros novos métodos iterativos, na linha de métodos iterativos existentes, assim como nos é sugerido investigar a aplicação das matrizes-escada e suas generalizações na construção de precondicionadores para métodos da família do gradiente conjugado, o que ainda não foi feito até onde vai nosso conhecimento.

4 – O Método SOR para Matrizes Singulares

4.1. Introdução

Nos capítulos anteriores, o método SOR foi estudado para SELAS não-singulares, mas conhecemos resultados relativos a esse método também para SELAS singulares consistentes, particularmente no que toca à otimização do parâmetro de relaxação [05, 22, 31, 32]. E isso é interessante, porque, para muitas aplicações, por exemplo, em cadeias de Markov, a matriz dos coeficientes é singular. O primeiro a examinar o caso singular foi Hadjidimos [21]. Sob as hipóteses de que

- a matriz de Jacobi associada ao SELAS singular consistente seja fracamente cíclica de índice $p \geq 2$;
- o espectro $\sigma\left(\mathbf{B}_p^p\right) \subset [0\,;\infty)$ e o raio espectral $\rho(\mathbf{B}_p) = 1$;
- $\mathbf{B}_p$ tenha 1 como autovalor simples, ou múltiplo, mas neste último caso, associado a blocos de Jordan de ordem 1 por 1,

Hadjidimos mostrou, entre outros resultados, que o valor ótimo ω_p do parâmetro ω é a única raiz de (2.1), no mesmo intervalo que para o caso não-singular, mas com $\rho(\mathbf{B}_p)$ substituído por

$$\gamma\left(\mathbf{B}_p\right) := \max\left\{|\lambda| \;\middle|\; \lambda \in \sigma\left(\mathbf{B}_p\right), |\lambda| \neq 1\right\}.$$

4.2. Uma generalização do resultado de Varga

Em [17] e [31] encontramos uma substancial generalização dos resultados de Varga, contidos nos Teoremas 2.2.5 e 2.2.6, para o caso não-singular. Enunciamos essa generalização no teorema a seguir, para servir de apoio para a análise do caso singular.

4.2.1. Teorema. *Seja um SELAS não-singular* $\mathbf{Ax} = \mathbf{b}$, *onde a matriz* $\mathbf{A}$ *é consistentemente ordenada e da forma da matriz em* (1.25), *e a matriz de iteração de Jacobi associada, da forma da matriz em* (1.27). *Sejam* ω_p e ρ_p *o parâmetro de relaxação ótimo e o correspondente raio espectral do SOR p-cíclico, respectivamente, e suponhamos que*

$$\sigma\left(\mathbf{B}_p^p\right) \subset \left[-\alpha^p;\beta^p\right], \quad -\alpha^p, \beta^p \in \sigma\left(\mathbf{B}_p^p\right), \quad 0 \leq \alpha < \frac{p}{p-2}, \quad 0 \leq \beta < 1. \tag{4.1}$$

Então ω_p e ρ_p *são determinados pelas equações*

$$\left(\frac{\alpha_p+\beta_p}{2}\omega\right)^p-\frac{\alpha_p+\beta_p}{\beta_p-\alpha_p}(\omega-1)=0 \tag{4.2}$$

e

$$\rho_p=\left(\frac{\alpha_p+\beta_p}{2}\omega_p\right)(\omega_p-1)=\left(\frac{\alpha_p+\beta_p}{2}\omega_p\right)^p, \tag{4.3}$$

onde ω_p *é a única raiz positiva de* (4.2) *no intervalo aberto*

$$\left(\min\left\{1,1+\frac{\alpha_p-\beta_p}{\beta_p+\alpha_p}\right\};\ \max\left\{1,1+\frac{\alpha_p-\beta_p}{\beta_p+\alpha_p}\right\}\right) \tag{4.4}$$

e onde

$$\left.\begin{array}{lll} i)\ \alpha_p=\dfrac{p-2}{p}\beta_p,\beta_p=\beta & \Leftrightarrow & \dfrac{\alpha}{\beta}\le\dfrac{p-2}{p} \\ ii)\ \alpha_p=\alpha, \quad \beta_p=\beta & \Leftrightarrow & \dfrac{p-2}{p}\le\dfrac{\alpha}{\beta}\le\dfrac{p}{p-2} \\ iii)\ \alpha_p=\alpha, \quad \beta_p=\dfrac{p-2}{p}\alpha_p & \Leftrightarrow & \dfrac{p}{p-2}\le\dfrac{\alpha}{\beta} \end{array}\right\} \tag{4.5}$$

Nesse enunciado, os casos limites $\alpha = \beta$ (ambos iguais a zero ou diferentes de zero) conduzem a $\omega_p = 1$ e $\rho_p = \alpha^p = \beta^p$; enquanto que, para $\alpha \neq 0$ e $\beta = 0$, supomos $\alpha/\beta = \infty$, como também supomos $p/(p\text{-}2) = \infty$ para $p = 2$.

4.3. Matriz dos coeficientes singular

Sob certas condições, discutidas nesta secção, o método SOR aplicado a um SELAS $\mathbf{Ax} = \mathbf{b}$, se estende, em sua maior parte, para o caso em que $\mathbf{A}$ é singular, contanto que esse SELAS seja consistente.

Há muitos problemas que ocorrem na prática, como o problema de Neumann, o problema de Poisson numa esfera com condições de contorno periódicas, cuja formulação em termos de diferenças finitas conduzem a SELAS singulares; na secção 4.4, veremos que o vetor da distribuição de equilíbrio de uma cadeia de Markov é uma solução de certos SELAS singulares que envolvem a matriz de transição; também o cálculo do vetor de produção do modelo econômico de Leontief requer a solução de um SELAS singular [05]. Em todos os casos, normalmente tais problemas são grandes e esparsos, portanto próprios para serem tratados com métodos iterativos, particularmente o método SOR.

Num processo iterativo para o caso singular, a noção de convergência dá lugar a de semiconvergência. Dizemos que uma matriz quadrada $\mathbf{M}$ é *semiconvergente* sse

$$\lim_{n\to\infty}\mathbf{M}^n \text{ existe.}$$

O seguinte teorema dá condições para haver semiconvergência, e sua demonstração usa a fórmula canônica de Jordan, como ocorreu para a convergência – Teorema 1.3.1.

4.3.1. Teorema. *Uma matriz quadrada* $\mathbf{M}$ *é semiconvergente* $\Leftrightarrow$ $\mathbf{M}$ *possui todas as propriedades*:

1. $\rho(\mathbf{M}) \leq 1$,
2. *se* $\rho(\mathbf{M}) = 1$, *então*
   - *todos os divisores elementares associados ao autovalor* 1 *de* $\mathbf{M}$ *são lineares, equivalentemente,* $\text{posto}(\mathbf{I}-\mathbf{M})^2 = \text{posto}(\mathbf{I}-\mathbf{M})$,
   - *se* λ *é um autovalor de* $\mathbf{M}$ *com módulo* 1, *então* $\lambda = 1$.

Queremos estender a validade do Teorema 1.3.3 para o caso em que $\mathbf{A}$ é singular. Para isso precisamos de mais algum material preparatório.

4.3.2. Lema. *Uma matriz quadrada* $\mathbf{M}$ *é semiconvergente* $\Leftrightarrow$ *existe uma matriz não singular* $\mathbf{P}$ *tal que*

$$\mathbf{M} = \mathbf{P}\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{K} \end{bmatrix}\mathbf{P}^{-1}, \tag{4.6}$$

onde a matriz identidade $\mathbf{I}$ *está ausente quando* 1 *não é autovalor de* $\mathbf{M}$, *e* $\rho(\mathbf{K}) < 1$ *ou* $\mathbf{K}$ *está ausente.*

A demonstração desse lema decorre do Teorema 4.3.1, aplicando a fórmula canônica de Jordan a $\mathbf{M}$.

O *índice* de uma matriz $\mathbf{A}$ é definido como sendo o menor inteiro não-negativo $\text{indice}(\mathbf{A}) =: k$ tal que $\text{posto}(\mathbf{A}^k) = \text{posto}(\mathbf{A}^{k+1})$. Esse conceito não tem nada a ver com o índice de ciclicidade.

Seja $\mathbf{A}$ uma matriz com $\text{indice}(\mathbf{A}) = k$. A *inversa de Drazin* de $\mathbf{A}$ é uma matriz $\mathbf{X}$ tal que

- $\mathbf{XAX} = \mathbf{X}$;
- $\mathbf{AX} = \mathbf{XA}$;
- $\mathbf{A}^n = \mathbf{XA}^{n+1}$, para $n = 0, 1, 2, \cdots$.

Facilmente mostramos que a inversa de Drazin de uma matriz $\mathbf{A}$ é a única matriz $\mathbf{A}^D$ tal que

$$\mathbf{A}^D\mathbf{x} = \begin{cases} \mathbf{y}, \text{ se } \mathbf{Ay} = \mathbf{x}, \text{ com } \mathbf{x} \in \text{col}(\mathbf{A}^k). \\ \mathbf{0}, \text{ se } \mathbf{A}^k\mathbf{x} = \mathbf{0}. \end{cases}$$

A notação $\text{col}(\mathbf{M})$ indica o espaço coluna da matriz $\mathbf{M}$. Dessa caracterização da inversa de Drazin e do Lema 4.3.2 obtemos o lema seguinte.

4.3.3. Lema. *Se* $\mathbf{M}$ *é uma matriz semiconvergente, então*

$$\lim_{n\to\infty} \mathbf{M}^n = \mathbf{I} - (\mathbf{I}-\mathbf{M})(\mathbf{I}-\mathbf{M})^D. \tag{4.7}$$

Temos agora a base para estender a validade do Teorema 1.3.3 para o caso em que $\mathbf{A}$ é (quadrada) singular. Consideremos, como na secção 1.5, uma decomposição

$$\mathbf{A} = \mathbf{C} - \mathbf{R},$$

onde $\mathbf{C}$ é não-singular. Podemos escrever um SELAS $\mathbf{Ax} = \mathbf{b}$ como $\mathbf{Cx} = \mathbf{Rx} + \mathbf{b}$, e o correspondente método iterativo básico como

$$\mathbf{x}^{(n+1)} = \mathbf{M}\mathbf{x}^{(n)} + \mathbf{c}, \text{ com } n = 0, 1, 2, \ldots \tag{4.8}$$

onde $\mathbf{M} := \mathbf{C}^{-1}\mathbf{R}$ e $\mathbf{c} := \mathbf{C}^{-1}\mathbf{b}$.

4.3.4. Teorema. *O método iterativo* (4.8) *converge para alguma solução de* $\mathbf{Ax} = \mathbf{b}$, *seja qual for o vetor inicializador* $\mathbf{x}_0 \Leftrightarrow \mathbf{M}$ *é semiconvergente. Nesse caso*

$$\lim_{n\to\infty} \mathbf{x}^{(n)} = (\mathbf{I}-\mathbf{M})^{D}\mathbf{c} + \left(\mathbf{I} - (\mathbf{I}-\mathbf{M})(\mathbf{I}-\mathbf{M})^{D}\right)\mathbf{x}^{(0)}. \tag{4.9}$$

Prova. Seja $\mathbf{x}$ uma solução do SELAS (ponto fixo). Então $\mathbf{x} = \mathbf{Mx} + \mathbf{c}$. Escrevendo a seqüência dos vetores-erro (1.6)

$$\mathbf{x}^{(n)} - \mathbf{x} = \mathbf{M}^{n}(\mathbf{x}^{(0)} - \mathrm{x}), \; n \geq 0,$$

vemos que $\left(\mathbf{x}^{(n)}\right)$ converge para alguma solução de $\mathbf{Ax} = \mathbf{b}$ se e somente se $\left(\mathbf{x}^{(n)} - \mathbf{x}\right)$ converge para algum vetor do núcleo de $\mathbf{A}$. Mas esse é o caso se e somente se

$$\lim_{n\to\infty} \mathbf{M}^{n}\left(\mathbf{x}^{(0)} - \mathbf{x}\right) \text{existe.} \tag{4.10}$$

Decorre do Lema 4.3.2 que (4.10) se mantém para todo $\mathbf{x}^{(0)}$ se e somente se $\mathbf{M}$ é semiconvergente. Além disso, se $\mathbf{M}$ é semiconvergente, a identidade (4.9) segue imediatamente do Lema 4.3.3. □

Introduzimos a definição de *fator de semiconvergência* de uma matriz $\mathbf{M}$, ou do método estacionário usado para resolver um SELAS $\mathbf{x} = \mathbf{Mx} + \mathbf{b}$, quando $\mathbf{M}$ é semiconvergente, por

$$\gamma(\mathbf{M}) := \max\left\{|\lambda| \;\middle|\; \lambda \in \sigma(\mathbf{M}), |\lambda| \neq 1\right\}.$$

Então $\gamma(\mathbf{M}) = \rho(\mathbf{M})$ quando $\rho(\mathbf{M}) < 1$. Em outro caso, $\gamma(\mathbf{M})$ é o maior dos módulos dos autovalores de $\mathbf{M}$, fora o raio espectral de $\mathbf{M}$. Pelo Teorema 4.3.1, se $\mathbf{M}$ tem um autovalor com módulo 1, então $\gamma(\mathbf{M}) < 1$. Resulta também que $\gamma(\mathbf{M}) = \rho(\mathbf{K})$, sendo $\mathbf{K}$ como em (4.6).

Somos levados a concluir que, se $\mathbf{M}$ é semiconvergente, a razão assintótica de convergência para o método (4.8) é dada por

$$R_{\infty}(\mathbf{M}) = -\ln \gamma(\mathbf{M}).$$

O lema de Hadjidimos [21] a seguir apóia o teorema principal abaixo desse capítulo. Como sempre, as matrizes $\mathbf{S}_{\omega}$ e $\mathbf{B}$ são as matrizes do SOR e de Jacobi de blocos do SELAS $\mathbf{Ax} = \mathbf{b}$.

4.3.5. Lema *Se uma matriz de Jacobi de blocos* (1.27) *satisfaz a hipótese de que* indice$(\mathbf{I} - \mathbf{B}) = 1$, *então para todo* $\omega \in (0\,;\,2) - \{p/(p-1)\}$ *temos que*

$$\mathit{indice}(\mathbf{I} - \mathbf{S}_{\omega}) = \mathit{indice}(\mathbf{I} - \mathbf{B}) = 1.$$

Para o caso geral, que inclui o caso singular, supomos que

$$\sigma(\mathbf{B}_p^p) \subset \left[-\alpha^p, \beta^p\right] \cup \{1\}, \quad -\alpha^p, \beta^p \in \sigma(\mathbf{B}_p^p), \; 0 \leq \alpha < \frac{p}{p-2}, \; 0 \leq \beta < 1.$$

Com essa hipótese e admitindo a condição de que indice($I - \mathbf{B}_p$) = 1, o resultado central – teorema seguinte – a menos do raio espectral $\rho_p := \rho(\omega_p)$, substituído pelo fator de semiconvergência $\gamma_p := \gamma(\omega_p)$, tem enunciado idêntico ao do Teorema 4.2.1.

4.3.6. Teorema. *Seja um SELAS, singular ou não,* $\mathbf{Ax} = \mathbf{b}$, *onde a matriz* $\mathbf{A}$ *é consistentemente ordenada e da forma da matriz em* (1.25), *e a matriz de iteração de Jacobi associada, da forma da matriz em* (1.27). *Sejam* ω_p e $\gamma_p := \gamma(\mathbf{S}_{\omega_p})$ *o parâmetro de relaxação ótimo e o correspondente fator de semiconvergência do SOR p-ciclico, respectivamente, e suponhamos que*

$$\sigma(\mathbf{B}_p^p) \subset [-\alpha^p; \beta^p] \cup \{1\}, \quad -\alpha^p, \beta^p \in \sigma(\mathbf{B}_p^p), \quad 0 \le \alpha < \frac{p}{p-2}, \quad 0 \le \beta < 1, \tag{4.11}$$

e que indice($\mathbf{I} - \mathbf{B}_p$) = 1. *Então* ω_p e γ_p *são determinados pelas equações* (4.2) e (4.3), com γ_p em lugar ρ_p.

Demonstração. Primeiro observemos que, fixados $p \ge 2$ e $y \ge p - 1$, a função $g_y : (0 ; 1] \to \mathbb{R}$, dada por

$$g_y(x) = \frac{(x+y)^p}{x}, \tag{4.12}$$

é estritamente decrescente e, que, portanto, assume o mínimo global em $x = 1$.

Simplificamos as notações pondo $\alpha := \alpha^p$ e $\beta := \beta^p$ e dividamos a demonstração em dois casos: $\alpha \ne \beta$ e $\alpha = \beta$.

a) Suponhamos $\alpha \ne \beta$. Se $\alpha < \beta$, então, conforme (4.4), $\omega_p \in \left(1; 1 + \frac{\beta - \alpha}{\beta + \alpha}\right)$, o que acarreta por (4.5) que

$$0 < \omega_p - 1 < \frac{\beta - \alpha}{\beta + \alpha} \le \frac{1}{p-1}. \tag{4.13}$$

Se $\alpha > \beta$, resulta $\omega_p \in \left(1 + \frac{\beta - \alpha}{\beta + \alpha}; 1\right)$, e daí,

$$0 > \omega_p - 1 > \frac{\beta - \alpha}{\beta + \alpha} \ge -\frac{1}{p-1}. \tag{4.14}$$

De (4.13) e (4.14) obtemos que

$$\omega_p \in \left(\frac{p-2}{p-1}; \frac{p}{p-1}\right) - \{1\},$$

e, por isso, pelo Lema 4.3.5, indice$(\mathbf{I} - \mathbf{S}_{\omega_p}) = 1$. Também, em ambos os casos, em vista de (4.13) e (4.14),

$$0 < |\omega_p - 1| < \frac{|\beta - \alpha|}{\beta + \alpha} \le \frac{1}{p-1}.$$

Analogamente ao caso não-singular, consideremos agora o polinômio

$$f(\lambda,\omega) := (\lambda+\omega-1)^p - \omega^p\lambda^{p-1}. \tag{4.15}$$

Os zeros de f, com $\lambda \neq 1$, são os valores de μ, dados por (1.34). Ponhamos

$$y := \omega - 1,\ y_p := \omega_p - 1,\ z := \frac{\beta+\alpha}{\beta-\alpha} \tag{4.16}$$

e dividamos (4.15) por $\lambda - 1$, para obter

$$\begin{aligned} g(\lambda,y) &:= \frac{f(\lambda,y+1)}{\lambda-1} \\ &= \lambda^{p-1} - \left[\binom{p}{2}y^2 + \cdots + \binom{p}{p}y^p\right]\lambda^{p-2} - \cdots - \left[\binom{p}{p-1}y^{p-1} + \binom{p}{p}y^p\right]\lambda - \binom{p}{p}y^p. \end{aligned} \tag{4.17}$$

O teorema seguirá se provarmos que a equação $g(\lambda, y_p) = 0$ tem raízes de módulo estritamente menor que ρ_p, sendo

$$\rho_p = \frac{\beta+\alpha}{\beta-\alpha}(\omega_p - 1) = zy_p = |z||y_p|. \tag{4.18}$$

O valor $\lambda = 0$ não é zero de (4.17), pois, uma vez que é $\alpha \neq \beta$, temos $\omega_p \neq 1$ e, daí, $g(0, y_p) = -(\omega_p - 1)^p \neq 0$. Para mostrar que as raízes λ de $g(\lambda, y_p) = 0$ têm módulo menor que ρ_p, ponhamos

$$\nu := \frac{\lambda}{zy_p},\quad a := |z|,\quad b := |y_p| \tag{4.19}$$

e mostremos que $|\nu| < 1$. Para tal, consideremos este outro polinômio em ν,

$$\left.\begin{aligned} h(\nu) &:= \frac{g(\lambda,y_p)}{(zy_p)^{p-1}} = \frac{g(zy_p\nu, y_p)}{(zy_p)^{p-1}} \\ &= \nu^{p-1} - \frac{1}{zy_p}\left[\binom{p}{2}y_p^2 + \cdots + \binom{p}{p}y_p^p\right]\nu^{p-2} - \cdots - \frac{1}{(zy_p)^{p-1}}\binom{p}{p}y_p^p \end{aligned}\right\} \tag{4.20}$$

Como $g(0, y_p) \neq 0$, temos $h(0) \neq 0$. Então isolamos ν^{p-1} em $h(\nu) = 0$ e dividimos a igualdade resultante por ν^{p-2}, para obter

$$\nu = \frac{1}{zy_p}\left[\binom{p}{2}y_p^2 + \cdots + \binom{p}{p}y_p^p\right] + \frac{1}{(zy_p)^2}\left[\binom{p}{3}y_p^3 + \cdots + \binom{p}{p}y_p^p\right]\frac{1}{\nu} + \cdots + \frac{1}{(zy_p)^{p-1}}\binom{p}{p}y_p^p\frac{1}{\nu^{p-2}}.$$

Suponhamos que exista um valor de ν, com $|\nu| \geq 1$, que verifique essa última equação. Usando nela (4.19), vem

$$|\nu| \le \frac{1}{ab}\left[\binom{p}{2}b^2 + \cdots + \binom{p}{p}b^p\right] + \frac{1}{(ab)^2}\left[\binom{p}{3}b^3 + \cdots + \binom{p}{p}b^p\right] + \cdots + \frac{1}{(ab)^{p-1}}\binom{p}{p}b^p$$

$$= \frac{ab}{1-ab}\left[\left(\frac{1}{a^2} - \frac{b}{a}\right)\binom{p}{2} + \left(\frac{1}{a^3} - \frac{b^2}{a}\right)\binom{p}{3} + \cdots + \left(\frac{1}{a^p} - \frac{b^{p-1}}{a}\right)\binom{p}{p}\right]$$

$$= \frac{ab}{1-ab}\left\{\left[\left(1+\frac{1}{a}\right)^p - 1 - \frac{p}{a}\right] - \frac{1}{ab}\left[(1+b)^p - 1 - pb\right]\right\},$$

donde

$$|\nu| \le 1 + \frac{b}{1-ab}\left[\frac{(1+a)^p}{a^{p-1}} - \frac{(1+a)^p}{b}\right]. \tag{4.21}$$

Com o apoio de (4.18) e (4.19), a (4.21) implica

$$|\nu| \le 1 + \frac{\rho_p}{(1-\rho_p)a^p}\left[(1-a)^p - \frac{(1-b)^p}{\rho_p}\right]. \tag{4.22}$$

Como $a = |z| \ge p - 1$ independe de $\rho_p \in (0\ ;\ 1)$, então, em vista da observação sobre a função (4.12),

$$\frac{(\rho_p + a)^p}{\rho_p} > (1+a)^p.$$

Portanto, a diferença entre colchetes em (4.22) é negativa, implicando que $|\nu| < 1$, em contradição com a hipótese de ser $|\nu| \ge 1$. Resulta que nenhum zero da função $\lambda \mapsto g(\lambda, y_p)$ em (4.17) pode ser, em módulo, maior ou igual a ρ_p, o que conclui a demonstração do teorema para o caso $\alpha \ne \beta$.

b) $\alpha = \beta > 0$. Esse caso é trivial, pois, aqui, $\omega_p = 1$, resultando

$$g(\lambda, 1) = \frac{f(\lambda, 1)}{\lambda - 1} = \lambda^{p-1},$$

com todas as raízes iguais a zero, e, por isso, estritamente menores que $\rho_p = \alpha^p = \beta^p > 0$. □

Hadjidimos e Plemmons [24] mostraram exaustivamente que, no método SOR estendido, no importante caso 2-cíclico consistentemente ordenado, pequenas perturbações do valor ótimo de ω afeta a velocidade de semiconvergência muito menos que no método usual do SOR, confirmando-se formalmente a validade das observações numéricas, feitas anteriormente por Kontovasilis et al. [31].

4.4. Aplicação nas cadeias de Markov

Os resultados desse capítulo sobre o SOR p-cíclico ótimo relativos a SELAS consistentes, possivelmente singulares, têm aplicação em problemas de cadeias de Markov [32] discretas ergódicas [04] com ma-

triz de transição **P** cíclica com período p. Essas cadeias às vezes possuem a propriedade de que o número mínimo de transições, que precisam ser efetivadas, para deixar um estado e retornar a ele, é um múltiplo de algum inteiro $p > 1$. Esses processos são ditos periódicos de período p, ou cíclicos de índice p, ou ainda simplesmente p-cíclicos. Bonhoure et al. [07] mostraram que as cadeias de Markov, que se originam em problemas de filas, freqüentemente têm essa propriedade.

No caso discreto, o SELAS a ser resolvido é

$$\mathbf{Px} = \mathbf{x}, \text{ com } \|\mathbf{x}\|_1 = x_1 + x_2 + \ldots + x_n = 1,$$

ou, equivalentemente,

$$(\mathbf{I} - \mathbf{P})\mathbf{x} = \mathbf{0}, \quad \|\mathbf{x}\|_1 = 1,$$

onde x_i é a probabilidade de o sistema estar no estado i, quando atinge o equilíbrio estatístico. Em particular, vemos que o vetor de estado **x** é um autovetor de **P** associado ao autovalor 1. Mostra-se que esse autovetor com norma-1 igual a 1 é único.

É imediato que, pondo $\mathbf{A} := \mathbf{I} - \mathbf{P}$, e notando que **P** é uma matriz estocástica (a soma dos elementos em cada coluna é 1) da forma (1.27), o correspondente problema homogêneo tem uma matriz dos coeficientes da forma (1.25), e a matriz de Jacobi associada é $\mathbf{B}_p := \mathbf{P}$. Então todos os resultados desse capítulo aplicam-se às cadeias de Markov p-cíclicas, simplesmente substituindo $\mathbf{B}_p$ por **P**. Em particular, a matriz **A** é uma M-matriz singular, e é irredutível quando a cadeia é ergódica. Logo as condições para semiconvergência aplicam-se a esse tipo de cadeias de Markov.

Para cadeias de Markov de grande porte o uso de métodos iterativos para calcular o vetor de equilíbrio **x** é de primeira importância e foram estudos extensivamente, por exemplo, por Berman e Plemmons [05], Courtois e Semal [45] e O'Leary [34]

Para cadeias de Markov p-cíclicas de tempo contínuo, com gerador infinitesimal **Q**, abordados por [6, 7, 25, 31], ocorre a necessidade de resolver SELAS homogêneos

$$\mathbf{Qx} = \mathbf{0}, \quad \|\mathbf{x}\|_1 = 1.$$

A matriz **Q** tem a forma (1.25), quando convenientemente particionada em blocos. Com $\mathbf{P} := \mathbf{Q}\Delta t + \mathbf{I}$, se Δt é suficientemente pequeno, a última equação pode também ser escrita na forma $\mathbf{Px} = \mathbf{x}$.

Resumindo, cadeias de Markov e modelos de filas conduzem a SELAS singulares estruturados irredutíveis do tipo considerado neste capítulo. Modelos probabilísticos de filas estão desempenhando um papel de importância crescente na compreensão dos fenômenos complexos que surgem em sistemas de computadores, comunicação e transporte [9, 10, 29].

A aplicação do novo SOR a SELAS singulares constitui-se um ponto a ser ainda investigado.

Referências Bibliográficas

[01] Axelsson, O., *Iterative Solution Methods*, Cambridge University Press, N.Y., 1994.

[02] Baker, L., Bradley, J. S., *Parallel Programming*, McGraw-Hill, N. Y., 1996.

[03] Barret, R., Berry, M., Chan, T., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., Van der Vorst, H., *Templates for the Solution of Linear Systems: Building Blocks for Iterative Method*, SIAM, Philadelphia, 1994.

[04] Beauwes, R., *Factorization iterative methods, M-operators and H-operators*, Numer. Math. 31, 335–357, 1979.

[05] Berman, A., Plemmons, R. J., *Nonnegative Matrices in the Mathematical Science*, SIAM, Philadelphia, 1994.

[06] Bonhoure, F., Dallery, Y., Stewart, W., *Algorithms for periodic Markov chains. Linear Algebra, Markov chains and queueing model*, IMA Volumes on Applied Mathematics, Springer Verlag , 48, 1992.

[07] Bonhoure, F., Dallery, Y., Stewart, W., *On the efficient use of periodicy properties for the efficient numerical solution of certain Markov chains*, MASI, Tech. Rept., 91-40, Université de Paris,1991.

[08] Carey, G. F., *Parallel Supercomputing: Methods, Algorithms and Applications*, John Wiley & Sons, 1989.

[09] Chan, R. – *Iterative Methods for queueing networks with irrregular state-spaces, Linear Algebra, Markov Chains and Queueing Models*, IMA Volumes on Applied Mathematics, Springer Verlag, 48:1992.

[10] Courtois, P.J. – *Decomposability: Queueing and Computer Systems Applications*, Academic Press, NY, 1997.

[11] Datta, B., *Numerical Linear Algebra and Applications*, Brooks/Cole Publishing Co., 1995.

[12] DeLong, M. A., Ortega, J. M., *SOR as a Preconditioner* II, 440, 1997.

[13] DeLong, M. A., Ortega, J. M., *SOR as a Preconditioner*, Applied Numerical Mathematics, 18:431–440, 1995.

[14] Demmel, J. W., *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[15] Dorneles, R.V., Rizzi, R. L., Zeferino, C. A., Diverio, T. A., Navaux, P. O. A., Susin, A. A., Bampi, S., *Fluvial Flow of Guaiba River – A Parallel Solution for the Shallow Water Equations Model,* Faculdade de Engenharia da Universidade do Porto, p. 885:896, 2000.

[16] Dotto, O. J., *Regra dos sinais de Descartes,* Chronos, v. 28, n. 2, p. 193–197, Caxias do Sul, 1995.

[17] Eiermann, M., Niethammer, W., Ruttan, A., *Optimal successive ovverrelation iterative methods for p-cyclic matrices,* Numer. Math. 57, 593-606, 1990.

[18] Frankel, S., *Convergence rates of iterative treatments of partial differential equations,* Math. Tables and other Aids to Computation, 4, 65–75, 1950.

[19] Fujino, S., Himeno, R., Kojima, A., Terada, K., *Implemention of the Multicolored SOR Method on a Vector Supercomputer,* IEICE, Transf. Inf. & Syst., vol. E80–D, n. 4, 1997.

[20] Golub, G. H., Van Loan, C. F., *Matrix Computation,* The John Hopkins University Press, Baltimore, 1996.

[21] Hadjidimos, A., *On the optimization of the classical iterative schemes for the solution of complex singular linear systems,* SIAM, J. Alg. Disc. Meth., 6(4):555 – 566, 1985.

[22] Hadjidimos, A., Plemmons, R. J., *Optimal p-cyclic SOR,* http://www.library. usyd. edu. au/ Ejournals/NM/67/4/10670475.html.

[23] Hadjidimos, A., Plemmons, R. J., Pierce, D. J., *Optimality relationships for p-cyclic SOR,* Numer. Math. 56, 635 – 643. 1990.

[24] Hadjidimos, A., Plemmons, R. L., *A general theory of optimal p-cyclic SOR,* CSD–TR–92–076, 1992.

[25] Hadjidimos, A., Plemmons, R. L., *Analysis of p-cyclic iterations for Markov chains Linear Algebra, Markov Chains and Queueing models,* IMA volumes on Applied Mathematics. Springer Verlag , 48, 1992.

[26] Hao, L., *Stair matrices and their generalizations with applications to iterative methods I: a generalization of the successive overrelaxation method,* SIAM J. Numer. Anal., vol. 37 . n.1, p.1–17, 1999.

[27] Hoffman, K., Kunze, R., *Linear Algebra,* Prentice-Hall, Inc., N. J., 1961.

[28] Kahan, W., *Numerical Linear Algebra,* Canadian Math. Bull., 9:757–801, 1966.

[29] Kaufman, L. – *Matrix methods for queueing problems,* SIAM J. Sci. Stat. Comp., 4:525-552, 1983.

[30] Kelley, C. T., *Iterative Methods for Linear and Nonlinear Equations,* SIAM, Philadelphia, 1995.

[31] Kontovasilis, K., Plemmons, R. J., Stewart, W. J., *Block cyclic SOR for Markov chains with p-cyclic infinitesimal generator,* Linear Algebra and its apllications, 154–156:145–223, 1991.

[32] Kulkarni, V. G., *Modeling, Analysis, Design, and Control of Stochastic Systems*, Springer Verlag, 1999.

[33] Lima, E. L., *Análise Real, vol.* 1. IMPA, R. J., 1997.

[34] O'Leary, D., *Iterative Methods for finding the stationary vector of Markov chains.* IMA Volumes in Aplied Mathematics, Springer Verlag, 48:1992.

[35] Ortega, J., Rheinboldt, W., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, N. Y., 1970.

[36] Ostrowski, A., M., *On the linear iteration procedures for symmetric matrices*, Rend. Mat. E Appl., 14, 140–163, 1954.

[37] Pacheco, P. S., *Parallel Programing with MPI*, Morgan Kaufmann Publishers, San Francisco, 1996.

[38] Preissler, L. E., Navaux, P. O., *Supercomputador Cray T94*, Instituto de Informática, UFRGS, 2000.

[39] Romanovsky, V., *Recherches sur les chains de Markoff*, Acta Math., 66, 147–251, 1936.

[40] Saad, Y., *Highly Parallel Preconditioners for General Sparse Matrices, in Recent Advanced Iterative Methods*, Springer Verlag, pp. 165–199, 1994.

[41] Saad Y., *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, 1996.

[42] Saad, Y., Schultz, M., *GMRES:A generalized Minimal Residual Algorithm for Solving Nonsymetric Linear Systems*, SIAM, J. Sci. Stat. Comput., 7(3):856–869, 1986.

[43] Sato, L. M., Midorikawa, E. T., Senger, H., *Introdução à programação paralela e distribuída*, http://www.lsi.usp.br./~liria/jai96.html

[44] Semal, P., *Iterative algorithms for large stochastic matrices*, Linear Algebra and its applications, 154–156:65–103, 1991.

[45] Semal, P., Courtois, P. J., *Block iterative algorithms for stochastic matrices*, Linear Algebra Appl., 76:59 –70:65–103, 1986.

[46] Silva, R. M., *Programação Paralela na Rede UNIX*, http://lexandria.cat.cbpf.br/~sun/NTs /nt0197/nt0197.html

[47] Song, Y., *Comparisons of nonnegative splitting of matrices*, Lin. Alg. Appl., 154–156, pp. 433–455, 1991.

[48] Tanenbaum A. S., *Structured Computer Organization*, Prentice Hall, 1999.

[49] Varga, R.S., *p-Cyclic matrices: a generalization of the Young-Frankel successive overrelaxation scheme*, Trans. Math. Soc., 1958.

[50] Varga, R.S., *Matrix Iterative Analysis*, Prentice-Hall, N. J., 1961.

[51] Varga, R.S., *Matrix Iterative Analysis,* Prentice-Hall, Berlin, 2000.

[52] Wong, K. L., *Iterative Solvers for System of Linear Equations,* http://www-jics.cs.utk.edu/PCUE/MOD9_IT/sld001.htm.

[53] Wong, K. L., *Introduction to Parallel Programming,* http://www-jics.cs.utk.edu/I2PP/I2PP60html/sld001.htm.

[54] Xie, D., *New Parallel SOR Method by Domain Partitioning,* SIAM J. Sci. Comput., 1995.

[55] Young, D. M., *Iterative Solution of Large Linear Systems,* Academic Press, Florida, 1971.

[56] Young, D. M., *Iterative Methods for Solving Partial Difference Equations of Elliptic Type,* Tese de Doutorado, Harvard University, Massachussets, 1950.

[57] Young, D. M., *Iterative Methods for Solving Partial Difference Equations of Elliptic Type,* Trans. Amer. Math. Soc., 76, 1954.

[58] Young, D. M., Grimes, R. G., Respess, J. R., Kincaid, D. R., *ITPACK 2C: A Fortran package for solving large sparse linear systems by adaptative accelerated iterative methods,* ACM Trans. Amer. Math. Soft., 8, 302–322, 1982.

Apêndice A

Breve descrição da computação paralela

O uso de diversos processadores trabalhando simultaneamente em tarefas parciais para realizar uma tarefa total maior, havendo comunicação entre eles, é o que podemos chamar de *processamento paralelo*. Uma definição de paralelismo em termos de programação é dada por Silva [46]:

> *Programação paralela consiste basicamente em dividir um programa em vários módulos, que serão executados em diferentes estações paralelamente, visando à solução do problema.*

Para a resolução de um SELAS, o dividimos, então, em sub-SELAS menores, sendo estes resolvidos, cada um, por processadores diferentes, comandados po um programa. Esta divisão hoje é muitas vezes necessária, devido ao grande tamanho dos SELAS que surgem na prática.

As principais razões do processamento paralelo são devidas a sua capacidade de

- resolver problemas que requerem uma grande quantidade de memória e armazenamento: hoje não conseguimos resolver a maioria dos problemas no modelo em serie com a eficiência necessária, e com a computação paralela podemos distribuir a memória em diversos computadores;
- reduzir o tempo de pesquisa e realizar simulações em tempo real;
- freqüentemente reduzir custos, por exemplo, usando os computadores disponíveis, ao invés de adquirir um novo computador mais potente, pois este, além da despesa inicial necessária, pode oferecer restrições, tais como as relativas à mudança de tecnologia, ao suprimento de componentes, ao tamanho da memória, etc.

Atualmente existem dois tipos principais de processadores em paralelo: o tipo MIMD (Multiple Instruction Multiple Data) e o tipo SIMD (Single Instruction Multiple Data). Faremos uma rápida descrição de ambos [48].

No tipo SIMD, Fig.A.1, os computadores são usados para resolver problemas científicos e de engenharia com uma estrutura regular de dados, tais como vetores e matrizes. Aqui os computadores têm uma única unidade de controle, que executa as instruções, uma de cada vez; mas cada instrução opera múltiplos itens de dados. Esse tipo compreende duas classes principais de processadores, a dos processadores matriciais e a dos processadores vetoriais.

Os de processadores matriciais consistem em um número grande de processadores idênticos, que executam a mesma seqüência de instruções em diferentes conjuntos de dados, Fig.A.2. Essa figura mostra que existe uma única unidade de controle, que envia as instruções para todo o conjunto, sendo que cada elemento usa seus próprios dados de sua própria memória, para realizar as tarefas enviadas pela unidade de controle.

Os processadores vetoriais executam uma seqüência de operações em pares de elementos dos dados. Ao contrário dos processadores matriciais, todas as operações aritméticas são feitas em uma única máquina de somar com estrutura *pipeline*. A montagem de uma estrutura *pipeline* está ilustrado na Fig.A.3. No primeiro estágio, E_1, a unidade 1 busca as informações na memória; no segundo estágio, a unidade 2 decodifica estas instruções; no terceiro estágio, a unidade 3 faz a busca dos operado-

res aritméticos ou lógicos; no quarto estágio, a unidade 4 executa as operações; e, finalmente, no quinto estágio, a unidade 5 disponibiliza o resultado destas instruções, enquanto que os outros estágios estão seguindo as instruções para mais um ciclo.

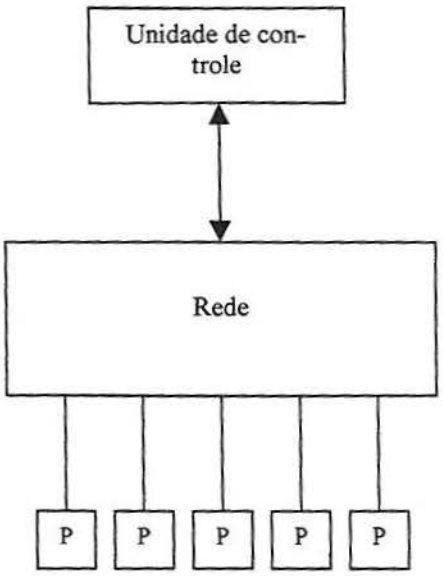


Fig.A.1 – Tipo SIMD

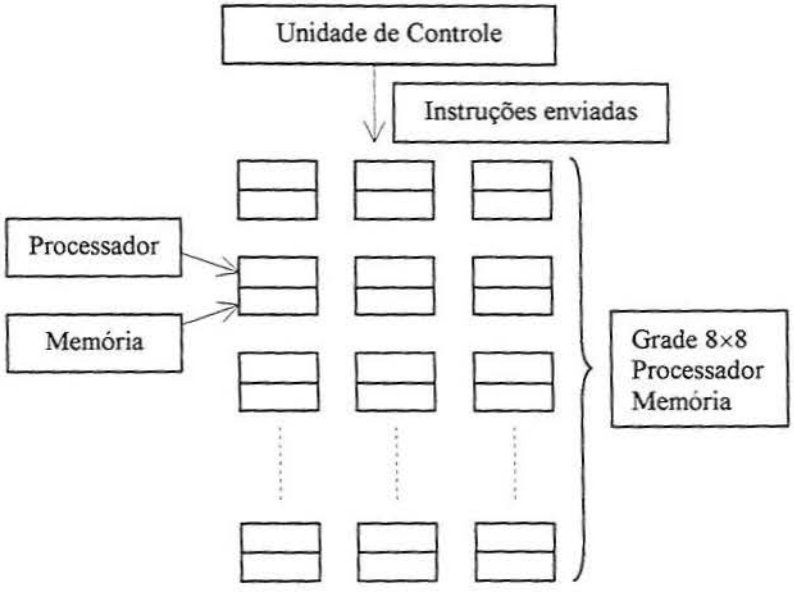


Fig.A.2 – Processador matricial

Na Fig.A.4 mostramos a seqüência da Fig.A.3 em função do tempo. Durante o primeiro ciclo, o primeiro estágio trabalha a instrução 1 trazida da memória; durante o ciclo 2, o segundo estágio

decodifica a instrução 1, enquanto o primeiro estágio traz da memória a instrução 2; durante o ciclo 3, o terceiro estágio busca os operandos para a instrução 1, o segundo estágio decodifica a instrução 2, e o primeiro estágio busca a instrução 3; durante o ciclo 4, o quarto estágio executa instrução 1, enquanto o terceiro estágio traz os operandos para a instrução 2, o segundo estágio decodifica a instrução 3, e o primeiro estágio busca a instrução 4; finalmente, durante o ciclo 5, o quinto estágio escreve o resultado da instrução 1, enquanto os outros estágios trabalham cada um seguindo as próprias instruções.

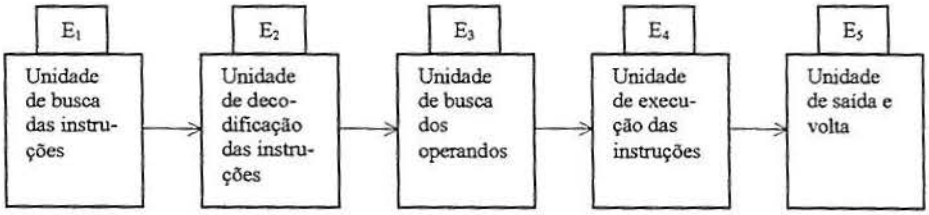


Fig.A.3 – Estrutura *pipeline* de 5 estágios

E_1	1	2	3	4	5	6	7	8	9
E_2		1	2	3	4	5	6	7	8
E_3			1	2	3	4	5	6	7
E_4				1	2	3	4	5	6
E_5					1	2	3	4	5
Ciclos	1	2	3	4	5	6	7	8	9

Fig.A.4 – Seqüência dos estágios em função do tempo com nove fases.

O tipo MIMD é o mais usado. Nesse tipo cada processador roda a sua instrução sobre os seus próprios dados, Fig.A.5, e pode consistir em um sistema de multiprocessadores ou de multicomputadores. Por exemplo, a UFRGS (Universidade Federal do Rio Grande do Sul) utiliza um supercomputador da família Cray T90, o Cray T94, que se enquadra neste tipo [38].

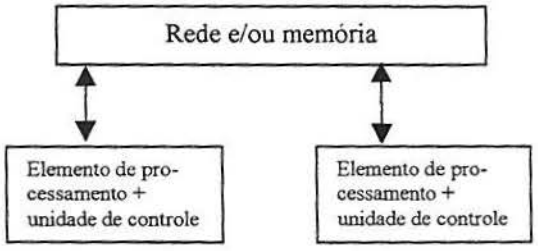


Fig.A.5 – Tipo MIMD

O sistema de multiprocessadores também é chamado de sistema de memória compartilhada. Nesse sistema, todos os processadores trabalham de maneira independente, apesar de terem acesso a toda a memória do sistema, lêem e escrevem de forma assíncrona. A comunicação entre a memória do sistema e os computadores é feita através de barramento. *Barramento* é uma coleção de fios paralelos

para transmitir endereços, dados e sinais de controle, Fig.A.6, ou de *switch* (chaveamento), Fig.A.7. O sistema de barramento é o de uso mais comum, devido a seu menor custo e maior facilidade de implementação, e sua maior limitação está no baixo número de processadores que podem ser conectados à memória, sendo que o máximo conseguido foi de 36 processadores [37]. Caso vários processadores tentem acessar a memória de uma só vez, causarão uma demora no atendimento de todos os pedidos de acesso. Uma forma de reduzir esse retardamento é através de memória cache para uso individual. Contudo o uso de memória cache origina o problema de garantir que os dados contidos na cache local sejam os mesmos daqueles encontrados na memória compartilhada. Para gerenciar isso existem diversos protocolos, de que não trataremos por fugirem ao escopo deste trabalho.

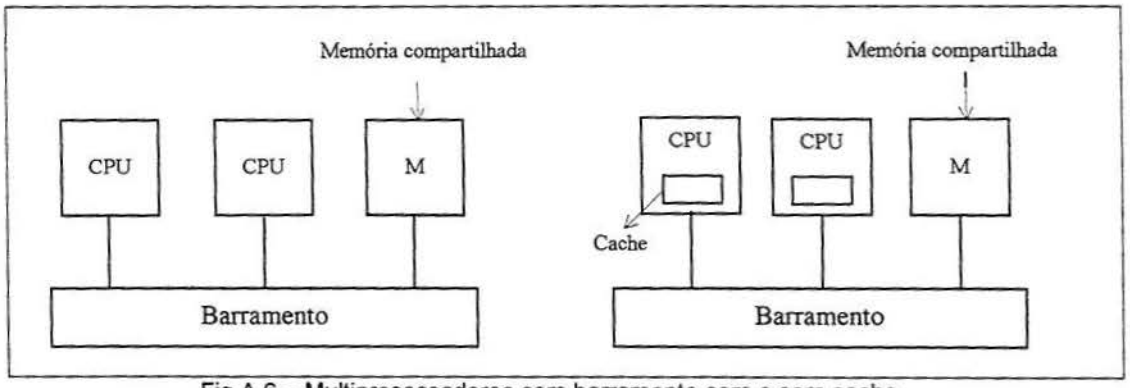


Fig.A.6 – Multiprocessadores com barramento com e sem cache

Para arquiteturas com *switchs* é usada uma interconexão de redes, também conhecida como *crossbar* que pode ser descrita como uma malha retangular, onde existem *switchs* nas intersecções, e as memórias são conectados nas extremidades. A vantagem desse sistema está em que qualquer processador pode acessar ao mesmo tempo qualquer memória não acessada por outro. A desvantagem desse sistema está no alto custo do hardware, já que cada intersecção terá um *switch* associado.

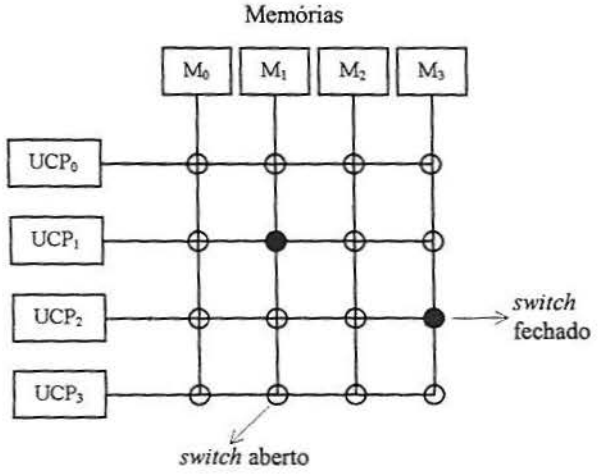


Fig.A.7 – Malha 4×4 com *switchs*

O sistema de multicomputadores, dentro do tipo MIMD, é também chamado de sistema de memória distribuída. Neste modelo cada processador opera de forma independente, e cada um possui sua própria memória, Fig.A.8. Podemos representar um sistema de memória distribuída através de grafos, onde cada vértice representa um par processador/memória, ou um *switch*. Há redes do tipo de ordem linear, que são aquelas em que cada nodo é cercado por outros dois adjacentes, e redes de tipo anel, Fig.A.9, que são uma variante do tipo de ordem linear, onde os dois últimos nodos se juntam para formar o anel. Ainda, numa configuração em duas dimensões, obtemos uma rede tipo grade, Fig.A.10. Além disso, temos distribuições em três ou mais dimensões [41].

Até agora abordamos o processamento paralelo em termos de máquina. Mas, no processamento paralelo é importante ter uma medida de desempenho capaz de informar quanto um algoritmo paralelizado é mais rápido que um algoritmo seqüencial.

Uma medida do ganho de desempenho de um algoritmo, *speedup*, é definida de uma maneira formal por Sato [43]: *speedup*, denotado com S, de um algoritmo paralelo, executado em p processadores, é a razão entre o tempo t_s, levado por um computador executando o algoritmo seqüencial mais rápido e o tempo t_p, levado pelo mesmo computador executando o algoritmo paralelo,

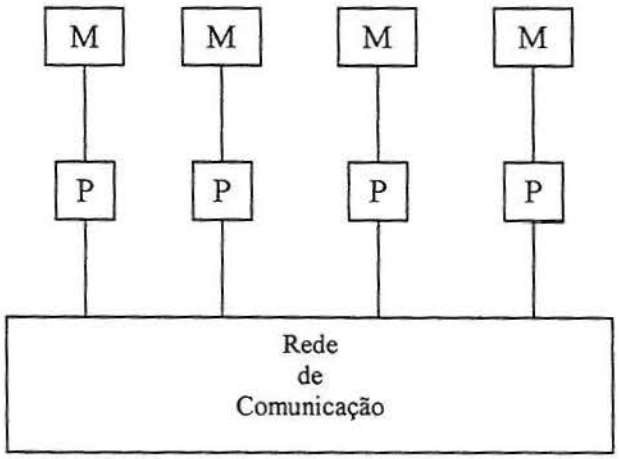


Fig.A.8 - Sistema de memória distribuída

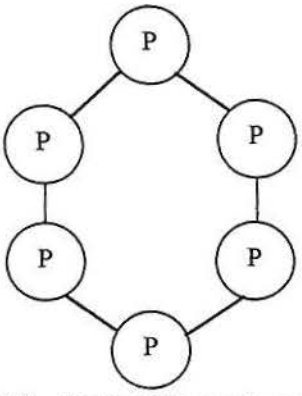


Fig.A.9 – Rede em forma de anel

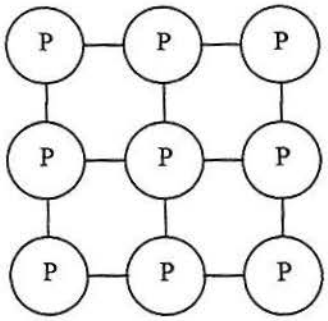


Fig.A.10 – Rede em forma de grade

usando p processadores:

$$S := \frac{t_s}{t_p}.$$

Poderíamos pensar então que, quanto mais processadores para trabalhar em paralelo, maior *speedup* obtemos, mas isto não ocorre, pois cada acréscimo de processadores acarreta uma perda, pela necessidade de aumento de comunicação entre os processadores e de aumento de tarefas de sincronização [02]. A conseqüente redução do *speedup* recebe o nome de *speeddown* .

Papel importante têm também as *bibliotecas* de troca de mensagens, responsáveis pela comunicação dos processos, baseados no envio e recebimento de mensagens. Elas permitem aproveitar a capacidade computacional das estações de trabalho que compõem uma rede. Normalmente essa ferramenta é usada quando os computadores são distribuídos em *clusters* (agregados) numa rede de alta velocidade. Segundo Baker e Bradley [02], o uso de bibliotecas de troca de mensagens é a maneira mais flexível de paralelismo. Algumas bibliotecas que oferecem estes serviços são: PVM, MPI, TCGMSG, PARMACS, P4 e *Express* [43]. Dentre estas, a que mais está em uso para ambientes de memória distribuída, é a MPI (Message Passing Interface), devido a sua portabilidade.

Apêndice B

Implementação dos algoritmos JASOR e ESCADA

Algoritmo JASOR

```
function [x,er,iter,c]  = jasor(A,b,x,w,maxit,tol,op)

% ------------------------------------------------------------------
% JASOR resolve um SELAS não-singular Ax = b iterativamente pelo
% método de Jacobi e SOR.
%
% Entrada: A, matriz dos coeficientes;
%          b, vetor dos termos independentes;
%          x, vetor inicializador;
%          w, parâmetro do SOR no intervalo (0 ; 2);
%          maxit, número máximo de iterações permitido;
%          tol, erro permitido;
%          op,1, para o método de Jacobi; 2, para o método SOR.
%
% Saída:   x, vetor solução;
%          er, norma do vetor erro da solução;
%          iter, número de iterações executadas;
%          c, 0 (se solução não encontrada com maxit iterações
%             dentro da tolerância tol);
%             1 (se solução encontrada dentro da tol
%             preestabelecida
% ------------------------------------------------------------------

c = 1;
iter = 0;
a = norm(b);
if a == 0,
   a = 1;
end

r = b - A*x;
er = norm(r)/a;
if er < tol,
   return,
end

if op == 1,
   M = diag(diag(A));
   N = diag(diag(A)) - A;
elseif op == 2,
```

```
   b = w * b;
   M = w*tril(A, -1) + diag(diag(A));
   N = -w*triu( A,1) + (1.0 - w)*diag(diag(A));
end
for iter = 1:maxit,
   x1 = x;
   x  = M\(N*x + b);
   er = norm(x - x1)/norm(x);
   if er <= tol,
      break,
   end
end

b = b/w;
if ( er > tol ),
   c = 0;
end
```

Algoritmo ESCADA

```
function x=escada(A,b,tipo)

% ------------------------------------------------------------
% ESCADA resolve o SELAS Ax = b, onde A é uma matriz escada do
% primeiro ou segundo tipo.
%
% Entrada: A, matriz escada do primeiro ou segundo tipo;
%          b, vetor (linha ou coluna)dos termos independentes;
%          tipo, identifica o tipo de matriz-escada: 1, para
%             tipo 1, e 2, para tipo 2.
% Saída:   x, solução numérica do SELAS.
%
% Usa-se x = escada(A,b,tipo).
% ------------------------------------------------------------

b = b(:);
n = size(A,1);
x=zeros(n+2,1);
B=[zeros(1,n+2);zeros(n,1) A zeros(n,1);zeros(1,n+2)];
if tipo ~= 1 & tipo ~= 2,
   msgbox('A 3a. entrada, TIPO, deve ser 1 ou
2','Atenção!','error'),
   x = 'Nada feito.';
   return
end

if tipo == 1,
   for i=1:2:2*floor((n-1)/2)+1
      x(i)=inv(A(i,i))*b(i);
```

```
   end
   for i=2:2:2*floor(n/2)
      x(i)=inv(A(i,i))*(b(i)-A(i,i-1)*x(i-1)-B(i+1,i+2)*x(i+1));
   end
end
if tipo == 2,
   for i=2:2:2*floor(n/2)
      x(i)=inv(A(i,i))*b(i);
   end
   for i=1:2:2*floor((n-1)/2)+1
      y=[0;x];
      x(i)=inv(A(i,i))*(b(i)-B(i+1,i)*y(i)-B(i+1,i+2)*x(i+1));
   end
end
x=x(1:n,1);
```

GRÁFICA UFRGS

Impressão: Gráfica UFRGS
Rua Ramiro Barcelos, 2705 - 1º andar
Fone: 316 5088 Fax: 316 5083 - Porto Alegre - RS
E-mail: grafica@vortex.ufrgs.br