PUC

Série: Monografias em Ciência da Computação

Nº 1/82

MÉTODOS ITERATIVOS PARA SISTEMAS LINEARES E SUA APLICAÇÃO NA SOLUCÃO DE EQUAÇÕES DIFERENCIAIS PARCIAIS

por

Peter Albrecht

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

RUA MARQUÊS DE SÃO VICENTE, 225 – CEP-22453

RIO DE JANEIRO – BRASIL

PUC/RJ - DEPARTAMENTO DE INFORMÁTICA

Série: Monografias em Ciência da Computação

Nº 1/82

Editor: Marco A. Casanova

Março 1982

MÉTODOS ITERATIVOS PARA SISTEMAS LINEARES E SUA APLICAÇÃO NA SOLUÇÃO DE EQUAÇÕES DIFERENCIAIS PARCIAIS *

por

Peter Albrecht

* Este trabalho foi patrocinado em parte pela FINEP, e teve como colaboradores os professores Marcelo Klein, Therezinha C. F. Chaves e Carlos Tomei

ABSTRACT:

This text gives a reasonably complete introduction to classical iteration methods for linear algebraic systems and their application to certain partial differential equation problems. The aim was to collect the relevant material in easily accessible form and, in the last chapter, to present new results with possibilities for further research.

RESUMO:

Este texto contêm uma introdução razoàvelmente completa dos métodos iterativos clássicos para solução de sistemas algébricos lineares e de sua aplicação em problemas oriundos do tratamento numérico de certas classes de equações diferenciais parciais. O objetivo foi o de compilar material relevante numa forma acessível e de apresentar, no último capítulo, resultados recentes com possibilidades de pesquisas futuras.

PREFÁCIO

Os métodos iterativos foram um tópico de destaque na década de 60, mas esse interesse diminuiu na medida em que novos programas para tratamento de grandes sistemas esparsos por métodos diretos foram desenvolvidos facilitados pelo maior poder computacional das novas gerações de computadores.

A razão pela preferência dos métodos diretos não reside fundamento em uma pretensa maior eficiência - de fato os métodos iterativos, adequadamente aplicados, frequentemente são mais rápidos - mas sim na maior simplicidade de aplicação, não exigindo nenhuma participação do usuário (por exemplo, considerações de convergência, determinação de parâmetros de aceleração etc.). Parece, entretanto, que novos desenvolvimentos vão levar a um renascimento dos métodos iterativos produzindo um novo interesse por eles.

Este texto fornece uma introdução razoavelmente completa dos clássicos métodos iterativos para a solução de sistemas algébricos lineares e a sua aplicação em problemas que surgem na discretização de equações diferenciais parciais. Ele se baseia em cursos que o autor deu nos últimos anos no Departamento de Informática da PUC - RJ e na Universidade de Dortmund em Alemanha.

Um dos objetivos foi o de preparar um texto autosuficiente exigindo um mínimo possível de pré-requesitos; por isso; o capítulo I começa com resultados elementares sobre matrizes incluindo também certas noções sobre grafos orientados que serão usadas mais tarde.

O capítulo II está fortemente baseado nos livros de Varga [10] e Young [12] adicionando-se uns poucos resultados próprios e fazendo-se certas modificações, ora com objetivos didáticos, ora para levar em consideração desenvolvimentos mais recentes. Por exemplo, para evitar as confusões entre matrizes cíclicas, p - cíclicas e fracamente cíclicas foram dispensados os primeiros dois conceitos definindo-se somente matrizes fracamente cíclicas (def. 5.4). Assim foi ne-

cessário desvincular a definição das matrizes consistentemente ordenadas do conceito das matrizes p - cíclicas o que foi conseguido através de uma definição feita por Verner e Bernal [11].

No capítulo III procurou-se dar, através de exemplos, uma idéia das aplicações em equações diferenciais parciais e problemas de contôrno. Este capítulo se baseia também no Varga [10] e no livro de Forsythe e Wasow [4], um texto clássico e, apesar de ser a fonte mais velha usada, ainda hoje indispensável.

Os resultados do capítulo IV foram obtidos em intensa colaboração com o Prof. Marcelo Klein [7] em 1976; eles ainda oferecem um amplo campo de pesquisa uma vez que podem ser usados nos novíssimos "métodos multigrid" e também porquê são fortemente relacionados a certas modificações explícitas, A - estáveis do método implícito de Euler para sistemas lineares de equações diferenciais ordinárias.

Agradeço ao Prof. Carlos Tomei pela ajuda e valiosas sugestões na elaboração dos primeiros capítulos na Língua Portuguesa e à Profa. Therezinha Chaves pela ajuda no terceiro capítulo. O capítulo IV foi escrito junto com o Prof. Marcelo Klein para publicação na revista "Matemática Aplicada e Computacional" da SBMAC.

Rio de Janeiro, 8 de março de 1982

Peter Albrecht

S U M Á R I O

CAPITULO I

NOÇÕES FUNDAMENTAIS SOBRE MATRIZES

Conceitos fundamentais e resultados da teoria de matrizes são apresentadas nesse capítulo.

Notação:

$\mathbb{C}(m,n)$: Espaço das matrizes mxn a valores complexos

$\mathbb{R}(m,n)$: Espaço das matrizes mxn a valores reais

$I \in \mathbb{R}(n,n)$: Matriz identidade; $\mathcal{O}$: matriz nula.

A^T : Matriz transposta de $A = (a_{ij})$; $A^T := (a_{ij})$

$\bar{A}$: Matriz conjugada de $A = (a_{ij})$; $\bar{A} := (\bar{a}_{ij})$

A^* : Matriz adjunta de $A = (a_{ij})$; $A^* = \bar{A}^T$; $(A^*)^* = A$

A^{-1} : Matriz inversa de $A \in \mathbb{C}(n,n)$; $A^{-1}A = AA^{-1} = I$.

$|A|$: Matriz dos módulos de $A = (a_{ij})$; $|A| = (|a_{ij}|)$.

§1 Autovalores

Definição 1.1

Seja $A \in \mathbb{C}(m,n)$, $\lambda \in \mathbb{C}$, $x \in \mathbb{C}^n$ tal que

$$Ax = \lambda x, \quad x \neq \mathcal{O}$$

Então λ é um autovalor e x seu autovetor associado.
Os autovalores de $A \in \mathbb{C}(n,n)$ são raízes do polinômio característico:

$$\det(\lambda I - A) = \lambda^n + c_1\lambda^{n-1} + \ldots + c_{n-1}\lambda + C_n \tag{1.1}$$

com $c_1 = \sum_{i=1}^{n} a_{ii}$ e $c_n = (-1)^n \det A$.

O traço de $A \in \mathbb{C}(n,n)$ é $trA := \sum_{i=1}^{n} a_{ij}$.

Então, se λ_j, $j=1(1)n$, são os autovalores de A, vale

$$\lambda_j + \lambda_2 + \ldots + \lambda_n = -c_1 = trA \tag{1.2}$$

$$\text{e} \quad \lambda_1 \cdot \lambda_2 \cdot \ldots \lambda_n = (-1)^n c_n = \det A.$$

Matrizes diferentes podem ter o mesmo polinômio característico. Em particular, a matriz de Frobenius

$$F = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & -c_n \\ 1 & 0 & 0 & \dots & 0 & -c_{n-1} \\ 0 & 1 & 0 & \dots & 0 & -c_{n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -c_1 \end{pmatrix}$$

tem o polinômio característico (1.1).

Teorema 1.1

Sejam $A,T \in \mathbb{C}(n,n)$, T inversível. Então A e $T^{-1}AT$ têm o mesmo polinômio característico, e, consequentemente, os mesmos autovalores.

Prova: $\det(T^{-1}AT-\lambda I) = \det(T^{-1}(A-\lambda I)T)$

$= \det T^{-1}.\det(A-\lambda I).\det T$

$= \det(A-\lambda I)$ □

A transformação $T^{-1}AT$ é chamada transformação por similaridade e $T^{-1}AT$ e A são ditas similares.

Definição 1.2:

$A \in \mathbb{C}(n,n)$ é hermitiana se $A=A^* = \bar{A}^T$. Uma matriz hermitiana é positiva definida se para todo $x \in \mathbb{C}^n$, $x \neq \mathcal{O}$, $\bar{x}^T Ax > 0$.

Teorema 1.2

Os autovalores de uma matriz hermitiana A são reais e a forma quadrática $Q(x) = \bar{x}^T Ax$, $x \in C^n$ é real.

Prova: $Q(x) = \bar{x}^T Ax = \bar{x}^T \bar{A}^T x$, porque A é hermitiana,

$= x^T \overline{Ax}$, porque para todo x,y,A, $x^T Ay = y A^T x$,

$= \overline{\bar{x}^T Ax} = \overline{Q(x)}$

Logo $Q(x)$ é real. Se λ é autovalor, com autovetor associado y, então $Q(y) = \lambda|y|^2$. Logo λ é real. □

A prova mostra também que os autovalores de uma matriz positiva definida são positivos. Mas ainda:

Teorema 1.3

Uma matriz hermitiana é positiva definda se e somente se todos os autovalores são positivos.

Teorema 1.4

Para toda $A\in\mathbb{C}(n,n)$, AA^* e A^*A têm os mesmos autovalores $\eta_j, j = 1(1)n$; mais ainda, $\eta_j\in\mathbb{R}$ e $\eta_j\geq 0$.

Prova: Se $\eta=0$ é autovalor de AA^*, também é de A^*A. Seja $\eta\neq 0$ autovalor de AA^* e $x\neq 0$ seu autovetor associado. Multiplicando à esquerda por A^* e fazendo $A^*x=y$ obtemos: $A^*Ay = \eta y$.

Como $\sum_{j=1}^{n} |y_j|^2 = y^*y = x^*AA^*x = \eta x^*x = \eta \sum_{j=1}^{n} |x_j|^2$, η é autovalor real de A^*A e $\eta>0$. □

Teorema 1.5

Se λ é autovalor de A, λ^m é autovalor de A^m, $m\in\mathbb{N}$.

O resultado vale também para $-m\in\mathbb{N}$, se A é inversível.

Prova: Por indução, a partir de $Ax = \lambda x$ e $A^{-1}x = \lambda^{-1}x$. □

Teorema 1.6

Toda matriz $A\in\mathbb{C}(n,n)$ pode ser levada através de uma transformação por similaridade à sua forma de Jordan:

$$N := T^{-1}AT = \begin{pmatrix} B_1 & & & \sigma \\ & \ddots & & \\ & & B_2 & \\ \sigma & & & \ddots \\ & & & & B_p \end{pmatrix} \qquad (1.3)$$

com blocos de Jordan $B_\ell \in C(m_\ell, m_\ell)$, $\sum_{\ell=1}^{p} m_\ell = n$, da forma

$$B_\ell := \begin{pmatrix} \lambda_\ell & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_\ell & 1 & \dots & 0 & 0 \\ . & . & . & \dots & . & . \\ 0 & 0 & 0 & \dots & \lambda_\ell & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda_\ell \end{pmatrix} \qquad (1.4)$$

onde λ_ℓ, $\ell = 1(1)p$, são os autovalores de A (não necessàriamente distintos).

Uma prova pode ser encontrada em |5|.

Definição 1.3

Os polinômios característicos dos blocos de Jordan B_ℓ se chamam divisores elementares de A

O divisor elementar associado a um autovalor λ é linear, se e somente se o bloco de Jordan associado tem dimensão 1. Se todos os divisores elementares forem lineares, a forma de Jordan da matriz é diagonal.

Exemplo: Seja N a forma de Jordan de uma matriz com autovalores λ_1 λ_2 e λ_3 :

$$N = \begin{pmatrix} \lambda_1 & 0 & & & & & & \\ 0 & \lambda_1 & & & & & & \\ & & \lambda_2 & 1 & & & & \\ & & 0 & \lambda_2 & 0 & & & \\ & & & & \lambda_2 & 0 & & \\ & & & & & \lambda_3 & 0 & 0 \\ & & & & & 0 & \lambda_3 & 1 \\ & & & & & 0 & 0 & \lambda_3 \end{pmatrix}$$

Associado a λ_1, existem apenas divisores elementares lineares, que têm a forma $(\lambda-\lambda_1)$. A λ_2 correspondem os divisores elementares $(\lambda-\lambda_2)^2$ e $(\lambda-\lambda_2)$; a λ_3, corresponde $(\lambda-\lambda_3)^3$

Teorema 1.7

Seja N como em (1.3). Para todo $r \in \mathbb{N}$ tal que $r \geq \max_{1\leq\ell\leq p} (m_\ell - 1)$, vale

$$N^r = \begin{pmatrix} B_1^r & & & \\ & B_2^r & & 0 \\ & & \ddots & \\ 0 & & & B_p^r \end{pmatrix}, \text{ com } B_\ell^r = \begin{pmatrix} \lambda_\ell^r & \binom{r}{1}\lambda_\ell^{r-1} & \lambda_\ell^{r-1} & \cdots & \binom{r}{m_\ell-1}\lambda_\ell^{r-(m_\ell-1)} \\ & \lambda_\ell^r & \binom{r}{1}\lambda_\ell^{r-1} & \cdots & \binom{r}{m_\ell-2}\lambda_\ell^{r-(m_\ell-2)} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_\ell^r & \binom{r}{1}\lambda_\ell^{r-1} \\ & & & & \lambda_\ell^r \end{pmatrix} \tag{1.5}$$

Prova: Seja $D \in \mathbb{R}(m,m)$ com $d_{i,i+1} = 1$, $i=1(1)m-1$ e $d_{ik} = 0$, se $k \neq i+1$.

Seja $F:=D^r$. Se $0\le r\le m-1$, os elementos de F são $f_{i,i+r}=1$, $i=1(1)m-r$, e $f_{ik}=0$, se $k\neq i+r$. Se $r\ge m$, $F=\mathcal{O}$.

Daí, para $B_\ell=\lambda_\ell+D$, com $r\ge m$,

$$B_\ell^r = (\lambda_\ell I+D)^r = \sum_{j=0}^{m-1}\binom{r}{j}\lambda^{r-j}D^j$$

De (1.5) segue que

a) $\lim\limits_{r\to\infty} N^r=\mathcal{O}$ se e somente se $|\lambda_j|<1$, $j=1(1)n$.

b) Se $N^r:=(P_{ij}(r))$, $\sup\limits_{r\in\mathbb{N}}|P_{ij}(r)|\le$ constante se e somente se aos autovalores de módulo 1 estão associados divisores elementares lineares e $|\lambda_j|\le 1$, $j=1(1)n$.

Como $A^r=T^{-1}N^rT$, segue o próximo teorema, importante em aplicações futuras.

Teorema 1.8

Sejam λ_j, $j=1(1)n$, os autovalores de $A\in\mathbb{C}(n,n)$ e $a_{ij}(r)$ os elementos de A^r. Então

a) $\lim\limits_{r\to\infty} A^r=\mathcal{O}$ $\iff$ $|\lambda_j|<1$, $j=1(1)n$

b) $\sup\limits_{r\in\mathbb{N}}|a_{ij}(r)|\le$ constante[1] $\iff$ $|\lambda_j|\le 1$, $j=1(1)n$ e associados aos autovalores de módulo 1 só existem divisores elementares lineares.

(1) Com a norma de matrizes apresentada no parágrafo seguinte isso equivale à relação $\|A^r\|\le \text{Const}$, $\forall r\in\mathbb{N}$.

§2 Normas de matrizes

Definição 2.1 (normas em $\mathbb{C}(n,n)$)

$\|A\|$ é uma norma de $A\in\mathbb{C}(n,n)$ se

a) $\|A\| \geq 0$, para todo $A\in\mathbb{C}(n,n)$.

b) $\|A\| = 0 \iff A=\mathcal{O}$ (2.1)

c) $\|A+B\| \leq \|A\| + \|B\|$, para todos $A,B\in\mathbb{C}(n,n)$ (2.2)

d) $\|cA\| = |c| \cdot \|A\|$, para todo $A\in\mathbb{C}(n,n)$, $c\in\mathbb{C}$. (2.3)

As matrizes de $\mathbb{C}(n,n)$ podem ser interpretadas como operadores lineares de $\mathbb{C}^n \to \mathbb{C}^n$. Por isso é conveniente poder tratar $\|A\|$ como norma de operadores. Defineremos então:

Definição 2.2

Uma norma de matrizes em $\mathbb{C}(n,n)$ é consistente com uma norma de vetores em $\mathbb{C}^n$ se para todo $A\in\mathbb{C}(n,n)$, $x\in\mathbb{C}^n$

$$\|Ax\| \leq \|A\| \, \|x\| \; ; \tag{2.4}$$

Uma norma de vetores é dita subordinada a uma norma consistente de matrizes se para dada $A\in\mathbb{C}(n,n)$, existe um $y\in\mathbb{C}^n$, $y\neq\mathcal{O}$ tal que

$$\|Ay\| = \|A\| \, \|y\| \; . \tag{2.5}$$

Interpretando A como operador no espaço normado $\mathbb{C}^n$, consistência equivale a dizer que A é limitado.

A relação (2.5) mostra que $\|A\|$ é uma norma de operador, isto é, $\|A\| := \sup_{x\neq\mathcal{O}} \frac{\|Ax\|}{\|x\|}$. De fato, o ínfimo é alcançado para toda norma vetorial, isto é, a norma matricial definda por $\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$ é subordinada à norma vetorial usada.

Para normas matriciais com normas vetoriais subordinadas, temos que

$$\|I\| = 1 \qquad e \qquad \|A^m\| \leq \|A\|^m \,, \quad m \in \mathbb{N} \tag{2.7}$$

e de $\|ABy\| = \|AB\| \, \|y\| \leq \|A\| \, \|B\| \, \|y\|$ segue que para $A, B \in \mathbb{C}(n,n)$

$$e) \quad \|AB\| \leq \|A\| \cdot \|B\| \tag{2.8}$$

Frequentemente as propriedades (a)-(e) são dadas como definição de norma de matrizes.

As normas matriciais mais frequentemente usadas são:

1 - <u>A norma linha</u>

$$\|A\|_\infty := \max_i \sum_{j=1}^{n} |a_{ij}| \tag{2.9}$$

Como exercício, pode ser mostrado que $\|A\|_\infty$ é subordinada à norma de vetores $\|x\|_\infty := \max_j |x_j|$ (2.10)

2 - <u>A norma coluna</u>

$$\|A\|_1 := \max_j \sum_{i=1}^{n} |a_{ij}|, \tag{2.11}$$

subordinada à norma vetorial

$$\|x\|_1 := \sum_{j=1}^{n} |x_j| \tag{2.12}$$

3 - <u>A norma espectral</u>

$$\|A\|_S := \max_i | \sqrt{\mu_i} |,$$

onde μ_1 são os autovalores de A^*A, (2.13)

é subrodinada a norma euclidiana

$$\|x\|_E = \sqrt{\sum_{i=1}^{n} |x_i|^2} \qquad (2.14)$$

4) - A norma da soma dos quadradados

$$\|A\|_E = \left| \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2} \right| = \sqrt{tr(AA^*)}, \qquad (2.15)$$

é consistente mas não subordinada com a norma euclidiana. Apesar disso,

$$\|AB\|_E \leq \|A\|_E \|B\|_E .$$

Teorema 2.1

$$\|A\|_S = \max_{x \neq 0} \frac{\|Ax\|_E}{\|x\|_E} \qquad (2.16)$$

Prova: Seja μ_1 o maior autovalor de A^*A. Para todo $x \neq 0$,

$$\|A\|_S^2 - \frac{\|Ax\|_E^2}{\|x\|_E^2} = \mu_1 \frac{x^*A^*Ax}{x^*x} = \frac{x^*(\mu_1 I - A^*A)x}{x^*x} \geq 0.$$

Porque, pelo teorema 1.4, os autovalores de $(\mu_1 I - A^*A)$ são maiores ou iguais a zero, igualdade ocorrendo quando x for o autovetor de A^*A correspondente a μ_1. □

A equação (2.16) pode ser tomada como definição do raio espectral.

Teorema 2.2

Seja $A \in \mathbb{C}(n,n)$. Então $\|A\|_S \leq \|A\|_E \leq \sqrt{n}\ \|A\|_S$ (2.17)

Prova: De (2.15) vem que $\|A\|_E^2 = trAA^*$.

$$\|A\|_E^2 = trAA^* = \sum_{j=1}^{n} \mu_j \qquad \text{(v.(1.2))}$$

onde μ_j são os autovalores de AA^*. Como estes não são negativos e fazendo $\mu_1 \geq \mu_j$, $j=2(1)n$, $\mu_1 \leq \sum_{j=1}^{n} \mu_j \leq n\mu_1$. □

Na próxima seção, normas de matrizes serão usadas na obtenção de estimativas de erros de sistemas lineares.

§ 3 Considerações sobre erros

Considere o seguinte exemplo de T.S. Wilson[1)] :

$$\begin{aligned} 10x + 7y + 8z + 7w &= 32 \\ 7x + 5y + 6z + 5w &= 23 \\ 8x + 6y + 10z + 9w &= 33 \\ 7x + 5y + 9z + 10w &= 31 \end{aligned} \tag{3.1}$$

Com as aproximações

$$x = 9.2 \quad ; \quad y = -12.6 \quad ; \quad z = 4.5 \quad ; \quad w = -1.1 \tag{3.2}$$

obtemos

$$b_1 = 32.1 \quad ; \quad b_2 = 22.9 \quad ; \quad b_3 = 33.1 \quad ; \quad b_4 = 30.9 \quad ,$$

o que faz pensar que (3.2) é uma boa aproximação da solução de (3.1).

Com as soluções

$$x = 1.82 \quad ; \quad y = -0.36 \quad ; \quad z = 4.5 \quad ; \quad w = 0.79 \tag{3.3}$$

obtemos

$$b_1 = 32.01 \quad ; \quad b_2 = 22.99 \quad ; \quad b_3 = 33.01 \quad ; \quad b_4 = 30.99 .$$

Evidentemente, (3.3) é uma aproximação melhor que (3.2), mas ainda está longe da solução exata:

$$x = y = z = w = 1.$$

Consequentemente, o erro $\Delta b := A\bar{x} - b$ não é um critério útil para avaliar a qualidade da aproximação $\bar{x}$, o que leva a perguntar quando fenômenos do tipo indicado podem ocorrer.

Temos que detA=1. Substituindo, por exemplo, o elemento

1) S. Todd,J.: Survey of Numerical Analysis, McGraw-Hill,1962 p.242

a_{11} = 10 de A por 9.96, então detA=0. O fato de detA ser pequeno é uma das causas da sensibilidade com a qual a solução x reage com pequenas alterações do vetor b. Entretanto, isto não explica totalmente a situação.

Consideremos o erro Δx provocado pelo erro Δb em b:

$$A(x+\Delta x) = b+\Delta b.$$

$$A\Delta x = \Delta b$$

$$\|\Delta x\| \leq \|A^{-1}\| \quad \|\Delta b\| \text{ , com normas subordinadas.}$$

De $\|b\| = \|Ax\| \leq \|A\| \|x\|$, segue $\|x\| \geq \|b\| \|A\|^{-1}$ e, para o erro relativo de x obtemos a estimativa

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|}{\|b\|} \frac{\|\Delta b\|}{\|A\|^{-1}}$$

Fica demonstrado assim o

Teorema 3.1

Seja A inversível e Δb o erro de b, então vale para o erro Δx da solução de Ax=b:

$$\frac{\|\Delta x\|}{\|x\|} \leq K(A) \frac{\|\Delta b\|}{\|b\|}, \text{ com } K(A) := \|A\| \|A^{-1}\|.$$

(3.4)

Definição 3.1

O número

$$K(A) := \|A\| \cdot \|A^{-1}\| \quad (A \text{ inversível}) \tag{3.5}$$

é o <u>número de condição</u> de A.

$K_S(A) := \|A\|_S \cdot \|A^{-1}\|_S$ é o número de condição espectral. O problema Ax=b é mal condicionado se K(A) for grande. No exemplo (3.1), $K_S(A) = 2984$.

Além de (3.5) podem ser definidos vários outros números de condição, por ex., o de TURING:

$$N(A) := \frac{1}{n} \|A\|_E \cdot \|A^{-1}\|_E \tag{3.6}$$

o de TODD:

$$P(A) := \frac{\max|\lambda_i|}{\min|\lambda_i|}, \text{ onde } \lambda_i \text{ são os autovalores de } A. \tag{3.7}$$

Para matrizes hermitianas, vale $P(A) = K_S(A)$.

Teorema 3.2

$$K_S(A) \leq K_S(A^*A) \tag{3.8}$$

Prova: Sejam μ_{max} e μ_{min} maior e menor autovalor de A^*A. Então

$$\|A\|_S = \sqrt{\mu_{max}}, \quad \|A^{-1}\|_S = \sqrt{\mu_{min}^{-1}},$$

$$\|A^*A\|_S = \mu_{max} \quad \text{e} \quad \|(A^*A)^{-1}\|_S = \mu_{min}^{-1}.$$

Daí,

$$K_S(A) = \sqrt{\frac{\mu_{max}}{\mu_{min}}} \leq \frac{\mu_{max}}{\mu_{min}} = K_S(A^*A) \qquad \square$$

Logo o problema $A^*Ax=b$ é geralmente pior condicionado do que Ax=b. Não é recomendável então multiplicar Ax=b por A^* para simetrizar !

Como em (3.4), podemos obter um limite para o erro relativo de x no caso de A sofrer uma pertubação ΔA.

Teorema 3.3

Seja ΔA o erro da matriz inversível A. Suponhamos que $\|A^{-1}\| \cdot \|\Delta A\| < 1$, onde a norma de matrizes indicada é norma de operadores.

Então, para o erro Δx da solução de $Ax=b$ vale

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{K(A)}{1-K(A)\frac{\|\Delta A\|}{\|A\|}} \frac{\|\Delta A\|}{\|A\|} \qquad (3.9)$$

Prova: Se $G = (I+A^{-1}\Delta A)^{-1}$. Como $\|A^{-1}\| \|\Delta A\| < 1$, G existe e pode-se mostrar (v. teorema 4.4) que

$$\|G\| \leq \frac{1}{1 - \|A^{-1}\| \cdot \|\Delta A\|} \qquad (3.10)$$

Como $GA^{-1} = (A+\Delta A)^{-1}$, $(A+\Delta A)^{-1}$ também existe.

De $(A+\Delta A)(x+\Delta x) = b$ vem que

$$(A+\Delta A)\Delta x = -\Delta A x$$

$$\Delta x = -(A+\Delta A)^{-1}\Delta A x = -[A(I+A^{-1}\Delta A)]^{-1}\Delta A x$$

$$= -GA^{-1}\Delta A x.$$

Logo $\|\Delta x\| \leq \|G\| \cdot \|A^{-1}\| \cdot \|\Delta A\| \cdot \|x\|$

usando (3.10),

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|\Delta A\|}{1 - \|A^{-1}\| \cdot \|\Delta A\|} = \frac{\|A^{-1}\| \cdot \|A\| \frac{\|\Delta A\|}{\|A\|}}{1 - \|A^{-1}\| \|A\| \frac{\|\Delta A\|}{\|A\|}}$$

§ 4 Limitações para os autovalores

Definição 4.1

Sejam λ_j, $j=1(1)n$, os autovalores de $A\in\mathbb{C}(n,n)$; então $\rho(A) = \max_{1\le j\le n} |\lambda_j|$ é chamado raio espectral de A.

Teorema 4.1

Para toda matriz A e todo $\varepsilon>0$ existe uma norma de matrizes subordinada tal que

$$\rho(A) \le \|A\| \le \rho(A) + \varepsilon. \tag{4.2}$$

Esta norma depende de A e de ε.

Teorema 4.2

Se $A\in\mathbb{C}(n,n)$ é inversível e a norma de matrizes é consistente, então

$$\frac{1}{\|A^{-1}\|} \le |\lambda_j| \le \|A\| \quad , \quad j = 1(1)n. \tag{4.3}$$

Prova: Seja x autovetor associado a λ_j. Então $Ax = \lambda_j x$

$$|\lambda_j| \cdot \|x\| = \|\lambda_j x\| = \|Ax\| \le \|A\| \, \|x\|,$$

Logo $|\lambda_j| \le \|A\|$. De $A^{-1}x = \lambda_j^{-1}x$ vem que $|\lambda_j^{-1}| \le \|A^{-1}\|$, isto é,

$$\|A^{-1}\|^{-1} \le |\lambda_j| .$$

□

Corolário 4.2.1

$$\rho(A) \leq \max_i \sum_{j=1}^{n} |a_{ij}| \tag{4.4}$$

$$\rho(A) \leq \max_j \sum_{i=1}^{n} |a_{ij}| \tag{4.5}$$

$$\rho(A) \leq \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} \tag{4.6}$$

$$\frac{1}{\|A^{-1}\|_S} = \sqrt{\mu_{min}(A^*A)} \leq |\lambda_j(A)| \leq \sqrt{\mu_{max}(A^*A)} = \|A\|_S \tag{4.6}$$

onde $\mu_{min}(A^*A)$ é o autovalor de menor módulo de A^*A.

(4.4) e (4.5) aplicados na matriz de Frobenius fornecem avaliações para as raizes do polinômio (1.1) :

$$|\lambda_j| \leq \max \{1+|c_1|, \ 1+|c_2|, \ \ldots, \ 1+|c_{n-1}|, \ |c_n|\}$$

$$|\lambda_j| \leq \max \{1, \sum_{j=1}^{n} |c_j|\} \tag{4.8}$$

Aplicando (4.4) e (4.5) em $D^{+1}AD$ com

$$D := \begin{pmatrix} d_1 & & & \\ & d_2 & & 0 \\ & & \ddots & \\ 0 & & & d_n \end{pmatrix} \quad ; \quad d_j > 0, \ j=1(1)n$$

resulta no

Corolário 4.2.2

Para números positivos arbitrários d_j, j=1(1)n, vale

$$\rho(A) \leq \min (m_1, m_2),$$

onde
$$m_1 := \max_i \{\frac{1}{d_i} \sum_{j=1}^{n} d_j |a_{ij}|\} = \| D^{-1}AD \|_\infty$$

$$m_2 := \max_j \{d_j \sum_{i=1}^{n} d_i^{-1} |a_{ij}|\} = \| D^{-1}AD \|_1$$

Com esse corolário podem ser obtidas limitações mais precisas para o raio espectral.

Exemplo

$$A = \begin{pmatrix} 3 & 0 & 1 \\ 3 & 2 & 3 \\ 2 & 0 & 2 \end{pmatrix} \quad ; \quad \lambda_1 = 1 \; ; \quad \lambda_2 = 2; \quad \lambda_3 = 4$$

(4.4) e (4.5) dão $\rho(A) \leq 8$. Com $d_1 = d_3 = 1$, $d_2=3$, obtem-se $m_1=4$, $m_2=6$, Logo $\rho(A) \leq 4$.

Um critério simples para limitação de autovalores é dado pelo teorema de Gerschgorin:

Teorema 4.3

Seja $A = (a_{ij}) \in \mathbb{C}(n,n)$ e $r_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$, i=1(1)n.

Então todos os autovalores de A estão dentro dos discos

$$|\lambda - a_{ii}| \leq r_i, \; i=1(n) \qquad (4.10)$$

no plano complexo.

Prova: Seja x o autovetor associado a λ com $\max_{1\le j\le n} |x_j| = 1$.

De $Ax=\lambda x$ segue que

$$(\lambda - a_{ii})x_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j \quad , \quad i=1(1)n.$$

Logo λ é tal que

$$|\lambda - a_{ii}| \le \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}||x_j| \le \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| = r_i$$

Observação: Como A^T e A têm os mesmos autovalores, r_i pode também ser tomado como $\sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|$.

Exemplo: Os autovalores de $A = \begin{pmatrix} 3 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 4 \end{pmatrix}$ estão na união dos discos $|\lambda-3| \le 1$; $|\lambda-2| \le 1$; $|\lambda-4| \le 1$.

O corolário seguinte, apresentado sem demonstração[1] aplica-se para matrizes irredutíveis definidas na pg.22.

Corolário 4.3

Seja A irredutível. Se um autovalor λ está na fronteira da união dos discos, então todos os círculos

$|\lambda - a_{ii}| = r_i$, $i=1(1)n$, passam por λ.

O teorema seguinte será usado mais tarde:

---

[1] veja Varga[10], pag. 20

Teorema 4.4

Seja $A \in \mathbb{C}(n,n)$ e $\rho(A) < 1$. Então $(I-A)^{-1}$ existe e vale

$$(I-A)^{-1} = I+A+A^2+ \ldots$$

Prova: Se λ é autovalor de A, $(1-\lambda)$ é autovalor de $(I-A)$. Como $|\lambda| < 1$, $(I-A)$ é inversível.

De $I = (I-A)(I+A+\ldots+A^k) + A^{k+1}$, segue que

$$(I-A)^{-1} = I+A+\ldots+A^k + (I-A)^{-1}A^{k+1}.$$

Assim, para todo $k \in \mathbb{N}$,

$$\| (I-A)^{-1} - I - A - \ldots - A^k \| \leq \| (I-A)^{-1} \| \cdot \| A^{k+1} \| .$$

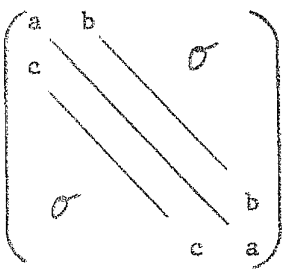
Pelo teorema 1.8, $\lim_{k \to \infty} \| A^k \| = 0$ □

Finalmente será útil o seguinte teorema:

Teorema 4.5

Sejam a, b, c reais e $bc > 0$;
então a matriz tri-diagonal
$T \in \mathbb{R}(n,n)$ tem os autovalores
$\lambda_j = a + 2\sqrt{bc}\cos\left(\frac{j\pi}{n+1}\right)$, $j = 1(1)n$.

$T :=$



§ 5 Matrizes especiais

5.1 Grafos orientados

A uma matriz $A \in \mathbb{C}(n,n)$ associamos um grafo orientado da seguinte maneira:

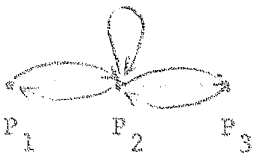
Fixados n vértices P_i, $i=1(1)n$, ligamos P_i a P_j por um caminho orientado se $a_{ij} \neq 0$

Definição 5.1

O grafo construído acima é o grafo de A, indicado por G(A). G(A) é fortemente conexo se para quaisquer dois vértices P_i, P_j, $i \neq j$, existe um caminho orientado de P_i a P_j. O comprimento de um caminho orientado é o número de seus segmentos.

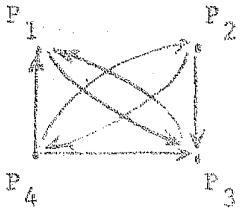
Exemplos:

$$A_1 = \begin{pmatrix} 0 & 2 & 0 \\ -1 & 3 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$ tem o grafo



$G(A_1)$ é fortemente conexo.

$$A_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$ tem grafo



$G(A_2)$ não é fortemente conexo, pois não há caminho orientado indo de P_3 a P_2. O menor caminho de P_2 a P_1 tem comprimento 2.

Definição 5.2

Seja G(A) fortemente conexo e seja p o maior divisor comum dos comprimentos de todos os caminhos fechados e orientados de G(A). Então, G(A) é dito cíclico de índice p, se p>1, e primitivo, se p=1.

Na próxima seção, veremos que propriedades da matriz podem ser deduzidas de seu grafo.

5.2 Matrizes redutíveis e fracamente cíclicas

Linhas ou colunas de uma matriz podem ser rearranjadas através de multiplicação por uma matriz de permutação. Matrizes de permutação têm exatamente um elemento valendo 1 em cada linha e coluna e os outros zeros, e serão denotadas por P.

Exemplo: Multiplicando uma matriz $A \in \mathcal{C}(4,4)$ à esquerda pela matriz de permutação

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

obtém-se $B := PA$, com as mesmas linhas de A, permutadas como $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}$. Análogamente $C := AP^T$ tem as mesmas colunas de A, permutas como $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}$

A transformação PAP^T apresenta permutações simultâ-

neas das linhas e colunas de A. Como $PP^T = I$ temos que

$$P^T = P^{-1} \tag{5.1}$$

Logo PAP^T é uma transformação por similaridade.

É fácil ver que PAP^T deixa G(A) invariante a menos de permutação na numeração dos vértices. Logo as propriedades de A que podem ser deduzidas de G(A) são invariantes por PAP^T, em particular as propriedades definidas a seguir.

Definição 5.3

Uma matriz $A \in \mathbb{C}(n,n)$, $n \geq 2$, é redutível se existe uma matriz de permutação P tal que

$$PAP^T = \begin{pmatrix} A_{11} & \mathcal{O} \\ A_{21} & A_{22} \end{pmatrix}, \text{ com } A_{11} \in \mathbb{C}(r,r),\ A_{21} \in \mathbb{C}(n-r,n-r),$$

$$A_{22} \in \mathbb{C}(n-r,n-r),\ 1 \leq r < n. \tag{5.2}$$

Caso contrário, A é irredutível.

Se A é redutível, o sistema $Ax=b$ pode ser reduzido a dois sistemas de ordem menor:

$$A_{11}\ x_1 = b_1 \quad ; \quad A_{22}x_2 + A_{21}x_1 = b_2 \tag{5.3}$$

Pode ser visto em G(A) se A é redutível ou não:

Teorema 5.1:

$A \in \mathbb{C}(n,n)$ é irredutível se e somente se G(A) é fortemente conexo.

Exemplo: A matriz A_2 da pag. 23 é redutível, como mostra $G(A_2)$. A permutação $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}$ agindo nas linhas e colunas de A_2 via PA_2P^T com

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{fornece } B := PA_2P^T = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

$G(B)$ é obtido de $G(A)$ renomeado os vértices como em $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}$

Definição 5.4

Uma matriz $A_B \in (n,n)$[1] é fracamente cíclica de índice p, $p>1$, se através de uma transformação PA_BP^T é levada à forma

$$PA_BP^T = \begin{pmatrix} \sigma & \sigma & \cdots & \sigma & A_{1p} \\ A_{21} & \sigma & \cdots & \sigma & \sigma \\ \sigma & A_{32} & \cdots & \sigma & \sigma \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma & \sigma & \cdots & A_{p,p-1} & \sigma \end{pmatrix} \qquad (5.4)$$

onde P é uma matriz de permutação e as matrizes nulas da diagonal principal são quadradas, (5.4) é a forma normal de uma matriz fracamente cíclica

(1) O índice B indica uma partição da matriz em blocos, e será sempre usado para enfatizar a partição.

Observação: Dependendo da partição em blocos, uma matriz pode ser fracamente cíclica de diversos índices. A matriz A_1, por exemplo, é fracamente cíclica de índice 2,3 ou 4:

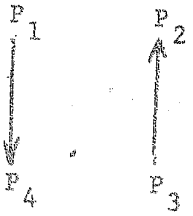
$$A_1 = \left(\begin{array}{c:c:c:c} 0 & 0 & 0 & 1 \\ \hdashline 0 & 0 & 0 & 1 \\ \hdashline 0 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 \end{array}\right) \quad ; \quad \text{(partição para índice 3).}$$

Temos o teorema(*)

Teorema 5.2:

$A \in \mathbb{C}(n,n)$ é fracamente cíclica de índice p se e somente se G(A) é cíclico de índice p ou torna-se cíclico de índice p identificando vértices ou adicionando caminhos.

Exemplos

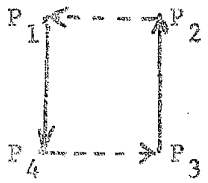
1. A matriz $A_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ tem grafo 

a) identificando P_1 com P_2 e P_3 com P_4, obtém-se um grafo cíclico de índice 2.

(*) provado em Albrecht: "Análise Numérica", pg. 92

b) identificando P_1 com P_2 e adicionando um segmento orientado de P_4 a P_3, obtém-se um grafo de índice 3.

c) adicionando segmentos de P_2 a P_1 e P_4 a P_3, um grafo de índice 4:

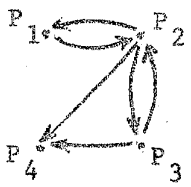


A_1, então, dependendo da partição dos blocos, é fracamente cíclica de índices 2,3 ou 4.

2. A matriz (redutível)

$$A_2 = \begin{pmatrix} 0 & x & 0 & 0 \\ x & 0 & x & x \\ 0 & x & 0 & x \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

com grafo



não pode ser colocada em forma cíclica, porque toda identificação de vértices ou adição de caminhos sempre resulta em grafos primitivos.

5.3 - Redução cíclica de um sistema

Matrizes cíclicas são de grande utilidade nos méto

dos iterativos para solução de sistemas lineares. O exemplo seguinte usa matrizes cíclicas para reduzir a ordem de um sistema linear.

Considere um sistema linear da forma

$$\begin{pmatrix} I_{11} & -A_{12} \\ -A_{21} & I_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad \begin{matrix} I_{11}\in\mathbb{R}(r,r),\ I_{22}\in\mathbb{R}(n-r,n-r) \\ X_1\in\mathbb{R}^r,\ X_2\in\mathbb{R}^{n-r}, \end{matrix} \tag{5.5}$$

que pode ser escrito na forma

$$(I - A_B)\cdot X = b, \quad \text{com} \quad A_B = \begin{pmatrix} 0 & A_{12} \\ A_{21} & 0 \end{pmatrix} \tag{5.6}$$

A_B é fracamente cíclica de índice 2.

Multiplicação à esquerda por $(I+A_B)$ dá

$$\begin{pmatrix} I_{11}-A_{12}\cdot A_{21} & 0 \\ 0 & I_{22}-A_{21}\cdot A_{12} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} b_1+A_{12}b_2 \\ b_2+A_{21}b_1 \end{pmatrix}, \tag{5.7}$$

que é redutível:

$$\begin{aligned} (I_{11} - A_{12}\cdot A_{21})X_1 &= b_1 + A_{12}b_2 \\ (I_{22} - A_{21}\cdot A_{12})X_2 &= b_2 + A_{21}b_1 \end{aligned} \tag{5.8}$$

Se $X_1\in\mathbb{R}^n$ é uma solução de (5.8), $X_2\in\mathbb{R}^{n-r}$ pode ser obtido de (5.5) por uma simples multiplicação:

$$X_2 = A_{21}X_1 + b_2.$$

Logo a solução de um sistema de ordem n foi reduzida à solução de um de ordem r. Esse método chama-se redução cíclica do sistema (5.6), e baseia-se no fato de que A_B^p é redutível quando A_B é fracamente cíclica de índice p.

5.4 Matrizes não negativas

Definição 5.5

$A = (a_{ij}) \in \mathbb{R}(n,n)$ é não negativa se para todo i,j: $a_{ij} \geq 0$. Usa-se, neste caso, a notação $A \geq \mathcal{O}$. $A \geq B$ se $A-B \geq \mathcal{O}$.

Observação: Pode acontecer $A \geq B$ sem que $A > B$ nem $A = B$ (por exemplo, $I \geq \mathcal{O}$ mas $I \not> \mathcal{O}$ nem $I = \mathcal{O}$).

A teoria das matrizes não negativas foi desenvolvida por Perron (1907) e Frobenius (1912); os resultados mais importantes estão reunidos no teorema a seguir, cuja prova pode ser encontrada na literatura:

Teorema 5.3

Seja $A=(a_{ij}) \in \mathbb{R}(n,n)$ não negativa e irredutível. Então

a) $\lambda_1 = \rho(A) > 0$ é autovalor simples de A.

b) Todos os autovalores λ_j de A com $|\lambda_j| = \rho(A)$, $j=1(1)r$ são diferentes e soluções de $\lambda^r - \lambda_1^r = 0$.

c) Associação a λ_1, existe um autovetor x_1, com componentes positivas: $x > \mathcal{O}$.

d) Não existe autovetor não-negativo linearmente independente de x_1.

e) $\rho(A)$ aumenta, quando qualquer a_{ij} aumenta.

O teorema seguinte oferece limitações para o raio espectral de matrizes não negativas irredutíveis.

Teorema 5.4

Seja $A=(a_{ij})\in\mathbb{R}(n,n)$ não-negativa, irredutível. Então ou

$$\sum_{j=1}^{n} a_{ij}=\rho(A), \quad \text{para todo } i=1(1)n \tag{5.9}$$

ou

$$\min_{1\le i\le n} \sum_{j=1}^{n} a_{ij} \le \rho(A) \le \max_{1\le i\le n} \sum_{j=1}^{n} a_{ij}. \tag{5.10}$$

Prova:

a) Seja $\sum_{j=1}^{n} a_{ij}=a$, para todo $i=1(1)n$. Então, por (4,4),

$\rho(a)\le a$.

Por outro lado o vetor com componentes $x_i=1$, $i=1(1)n$ é autovetor, com autovalor associado igual a a. Pela definição de $\rho(A)$, $\rho(A) \ge a$, Logo $\rho(A)=a$.

b) Se as somas $\sum_{j=1}^{n} a_{ij}$ não são todas iguais, considere $B\ge\mathcal{O}$, $B\neq\mathcal{O}$ obtida quando os elementos de A [illegible] forem aumentados de modo a valer

$$\sum_{j=1}^{n} b_{ij} = b = \max_{1\le i\le n} \sum_{j=1}^{n} a_{ij}, \text{ para todo } i=1(1)n.$$

Por a), $\rho(B)=b$ e pelo teorema 5.3 (e), $\rho(A)<b$

Reduzindo adequadamente os elementos de A, obtêm-se uma matriz irredutível C, $C \geq \mathcal{O}$, $C \neq \mathcal{O}$ com

$$\sum_{j=1}^{n} c_{ij} = c = \min_{1 \leq i \leq n} \sum_{j=1}^{n} a_{ij}, \text{ para todo } i=1(1)n \text{ e } \rho(A) > C. \quad \square$$

Observação: Como A é A^T tem os mesmos autovalores, as somas das linhas pode ser trocada pelas somas das colunas em (5.10).

Teorema 5.5

Seja $A \in \mathbb{R}(n,n)$ e $B \in \mathbb{C}(n,n)$ com $|B| \leq A$. Então

$$\rho(B) \leq \rho(A) \tag{5.11}$$

Se $A = |B|$, em particular,

$$\rho(B) \leq \rho(|B|) . \tag{5.12}$$

Prova: Seja $\varepsilon > 0$ arbitrário e $A_1 := (\rho(A)+\varepsilon)^{-1} A$ e

$B_1 := (\rho(A)+\varepsilon)^{-1} B$.

Então, como $\rho(cA) = c\rho(A)$, c positivo,

$$\rho(A_1) < 1 \quad \text{e} \quad |B_1|^r \leq A_1^r, \text{ para } r \in \mathbb{N}.$$

Pelo teorema 1.8, $\lim_{r \to \infty} A_1^r = \mathcal{O}$, logo $\lim_{r \to \infty} B_1^r = \mathcal{O}$ e $\rho(B_1) < 1$.

Como $\rho(B) = (\rho(A)+\varepsilon)\rho(B_1) < \rho(A)+\varepsilon$, para todo ε, $\rho(B) \leq \rho(A)$. $\square$

Como toda matriz redutível não negativa pode ser aproximada arbritàriamente por matrizes irredutíveis não negativas vale a seguinte generalização do teorema 5.3:

Teorema 5.6:

Seja $A=(a_{ij})\in\mathbb{R}(n,n)$ não negativa. Então

a) $\lambda_1=\rho(A)\geq 0$ é autovalor de A,

b) associado a λ_1 existe um autovetor x_1 com componentes não negativas: $x_1\geq \mathcal{O}$,

c) $\rho(A)$ não diminue se os a_{ij} aumentam.

O resultado seguinte será útil no capítulo II :

Teorema 5.7

Sejam $B,C\in\mathbb{R}(n,n)$, B inversível, $B^{-1}\geq\mathcal{O}$, $C\geq\mathcal{O}$, $S=B-C$.

Então existe S^{-1}, $S^{-1}\geq\mathcal{O}$ e

$$\rho(B^{-1}C) = \frac{\rho(B^{-1}C)}{1+\rho(S^{-1}C)} < 1 \qquad (5.13)$$

Prova:

Provamos inicialmente: Se x é auvetor de $G := S^{-1}$ e λ é autovalor associado, então x também é autovetor de $B^{-1}C$ e $\mu=\lambda(1+\lambda)^{-1}$ é o autovalor associado. Se, reciprocamente, μ é autovalor de $B^{-1}C$, então $\lambda=\mu(1-\mu)^{-1}$ é autovalor de G.

Temos que

$$B^{-1}C=(S+C)^{-1}C=[S(I+S^{-1}C)]^{-1}C=(I+G)^{-1}.G \qquad (5.14)$$

Se $\lambda\neq 1$ é autovalor de G, isto é, $Gx=\lambda x$, então vale

$(I+G)x=(\lambda+1)x$ e, por (5.14), $B^{-1}Cx=(I+G)^{-1}Gx = \lambda(1+\lambda)^{-1}x$.

Logo, x é autovetor de $B^{-1}C$ e $\mu:=(1+\lambda)^{-1}$ é o autovalor associa

do.

Por outro lado, se x é autovetor de $B^{-1}C$ e μ o autovalor associado, então $B^{-1}Cx = (I+G)^{-1}Gx=\mu x$, isto é, $Gx = \mu(I+G)x$, para $\mu\neq 1$ segue: $Gx = \mu(1-\mu)^{-1}x$.

Logo, x é auvetor de G e $\lambda=\mu(1-\mu)^{-1}$ é o autovalor associado.

Desde que $G\geq 0$, $\rho(G)=\lambda_o\geq 0$ é autovalor de G.

Sendo $\frac{\lambda}{1+\lambda}$ monótono crescente para $\lambda\geq 0$ obtemos

$$\mu_o = \rho(B^{-1}C) = \frac{\lambda_o}{1+\lambda_o} < 1.$$

Capítulo II

Métodos iterativos para sistemas lineares

§ 6 - Introdução

O problema $f(x) = 0$, $f: \mathbb{R} \to \mathbb{R}$, pode ser tratado iterativamente, colocando-o na forma $x = \Phi(x)$ e definindo $x^{(k+1)} = \Phi(x^{(k)})$ a partir de um valor inicial x_0.

Analogamente, pela decomposição $A = M-N$, $A,M,N \in \mathbb{C}(n,n)$, A,M inversíveis, o sistema $Ax=b$, se torna

$$Mx = Nx+b \tag{6.1}$$

o que leva ao processo iterativo $Mx^{(k+1)} = Nx^{(k)}+b$, $k = 0,1\ldots$

(6.2)

com um valor inicial x_0.

Existem infinitas decomposições de A, e para cada uma está associada um processo iterativo como descrito acima, cuja convergência depende do raio espectral de $M^{-1}N$.

Definição 6.1 : $H := M^{-1}N$ é a matriz de iteração.

Métodos iterativos são úteis para sistemas grandes, que ocorrem, em particular, na solução por discretização de equações diferenciais parciais. Considere o seguinte exemplo simples:

Seja a equação parcial elíptica

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{6.3}$$

Procura-se a solução no quadrado $Q = \{(x,y) | x \in (0,1)$, $y \in (0,1)\}$, com valores dados na fronteira de Q.

Discretizando (6.3) nos pontos $P_{ij}=(x_i,y_j)$, $x_j=hi$ $y_j=hj$, $h = \frac{1}{n+1}$, com diferenças finitas, obtêm-se o seguinte sistema linear com n^2 incógnitas u_{ij} nos pontos interiores P_{ij}, i,j=1(1)n:

$$\left(u_{i+1,j}-2u_{i,j}+u_{i-1,j}\right) h^{-2} + \left(u_{i,j+1}-2u_{i,j}+u_{i,j-1}\right)h^{-2}=0$$

isto é ,

$$u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1}-4u_{ij}=0 \qquad (6.4)$$

O lado direito do sistema é montado levando-se em conta as condições de fronteiras.

Para n=3,

$$\left(\begin{array}{ccc|ccc|ccc}
4 & -1 & 0 & -1 & 0 & 0 & & & \\
-1 & 4 & -1 & 0 & -1 & 0 & & \sigma & \\
0 & -1 & 4 & 0 & 0 & 1 & & & \\
\hline
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\
0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\
\hline
 & & & -1 & 0 & 0 & 4 & -1 & 0 \\
 & \sigma & & 0 & -1 & 0 & -1 & 4 & -1 \\
 & & & 0 & 0 & -1 & 0 & -1 & 4
\end{array}\right)
\left(\begin{array}{c} u_{11} \\ u_{12} \\ u_{13} \\ u_{21} \\ u_{22} \\ u_{23} \\ u_{31} \\ u_{32} \\ u_{33} \end{array}\right) =
\left(\begin{array}{c} (u_{01}+u_{10}) \\ u_{20} \\ (u_{30}+u_{41}) \\ u_{02} \\ 0 \\ u_{42} \\ (u_{03}+u_{14}) \\ u_{24} \\ (u_{34}+u_{43}) \end{array}\right) \qquad (6.5)$$

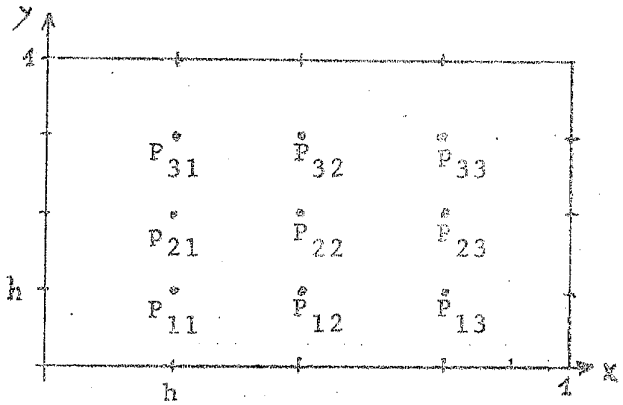


Fig. 6.1

Para discretizações mais finas, isto é, n maiores, o número de entradas da matriz cresce rapidamente; são 160.000 para n=20, dos quais menos de 2000 são não-nulos.
Por isso, na prática, A não é armazenada no computador. Em vez disso, associamos aos pontos interiores P_{ij} uma locação na memória e nela armazenamos a k-ésima aproximação $u_{ij}^{(k)}$. Por (6.4) pode ser calculado um novo valor $u_{ij}^{(k+1)}$ como média aritmética das k-ésimas aproximações nos pontos vizinhos, conforme

$$u_{ij}^{(k+1)} = \frac{1}{4} \left(u_{i+1,j}^{k)} + u_{i-1,j}^{(k)} + u_{i,j+1}^{(k)} + u_{i,j-1}^{(k)} \right).$$

Este procedimento corresponde à resolução de (6.5) para os elementos diagonais e subsequente iteração. Nela podemos utilizar para a determinação de $u_{ij}^{(k+1)}$:

a) ou apenas os valores da k-ésima iteração ou seja, $u_{ij}^{(k)}$

b) ou os valores já obtidos da (k+1)-ésima iteração para obter os demais valores dela. A teoria mostrará que o método b) é mais eficiente nesse caso.(o tratamento completo deste problema encontra-se no §12).

§ 7 Três Métodos Clássicos

Seja $A = D - E - F$, com $\det D \neq 0$ e

$$D = \begin{pmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn} \end{pmatrix}, \quad -E = \begin{pmatrix} 0 & & & 0 \\ -a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{pmatrix},$$

$$F = \begin{pmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ & 0 & \ddots & \vdots \\ & & \ddots & -a_{n-1,n} \\ 0 & & & 0 \end{pmatrix}$$

$$L := D^{-1}E, \qquad U := D^{-1}F \tag{7.1}$$

As decomposições

$$A = D - (E+F) \tag{7.2}$$

$$A = (D - E) - F \tag{7.3}$$

$$A = \frac{1}{\omega}(D - \omega E) - \frac{1}{\omega}(\omega F + (1-\omega)D) \tag{7.4}$$

fornecem os métodos de Jacobi, Gauss-Seidel e sobre-relaxação, descritos a seguir.

7.1 - A iteração de Jacobi

A decomposição (7.2) fornece o método iterativo de Jacobi:

$$x^{(k+1)} = Jx^{(k)} + D^{-1}b \; ; \; J := L+U, \tag{7.5}$$

o que dá para cada coordenada de x,

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

$$i = 1(1)n, \qquad (7.5b)$$

isto é, o sistema é resolvido pelos elementos diagonais sendo iterado depois.

Critérios para convergência são obtidos nos teoremas 9.1, 9.2 e 10.3.

Definição 7.1

A matriz J = L+U é a matriz de Jacobi de A.

Exemplo

$$\begin{aligned} 3\ x_1 + x_2 + x_3 &= 5 \\ x_1 + 2x_2 &= 3 \\ x_1 + \frac{1}{2}\ x_2 + 2x_3 &= 6, \end{aligned} \qquad (7.6)$$

a iteração de Jacobi é dada por

$$\begin{aligned} x^{(k+1)} &= \frac{-1}{3} x_2^{(k)} - \frac{1}{3} x_3^{(k)} + \frac{5}{3} \\ x_2^{(k+1)} &= \frac{-1}{2} x_1^{(k)} + \frac{3}{2} \\ x_3^{(k+1)} &= \frac{-1}{2} x_1^{(k)} - \frac{1}{4} x_2^{(k)} + 3 \end{aligned}$$

e a matriz de Jacobi é

$$J = \begin{pmatrix} 0 & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & -\frac{1}{4} & 0 \end{pmatrix} \quad ; \quad \rho(J) = \frac{1}{4}\left(1 + \sqrt{\frac{7}{3}}\right) \approx 0.63 \qquad (7.7)$$

7.2 A iteração de Gauss-Seidel

A decomposição (7.3) fornece o método iterativo de Gauss-Seidel:

$$x^{(k+1)} = Lx^{(k+1)} + Ux^{(k)} + D^{-1}b \qquad (7.8a)$$

$$x^{(k+1)} = (I-L)^{-1}\,Ux^{(k)} + (D-E)^{-1}b \qquad (7.8b)$$

(as inversas existem porque $\det(I-L) = 1$ e $\det D \neq 0$).

Para cada coordenada de x, então, temos

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}}, \quad i=1(1)n \qquad (7.8c)$$

Como se vê, o método de Gauss-Seidel difere do de Jacobi porque componentes $x_\ell^{(k+1)}$ já calculadas são usadas para o cálculo dos elementos seguintes, $x_m^{(k+1)}$, $m > \ell$. Por isso, no método de Gauss-Seidel, não é necessário armazenar simultâneamente o $x^{(k)}$ e o $x^{(k+1)}$. Em vez disso, o $x^{(k)}$ vai sendo substituido pelos elementos de $x^{(k+1)}$.
Em certos casos, o método de Gauss-Seidel converge mais rapidamente que o de Jacobi (ver teoremas 10.1 e 11.1)

Definição 7.2

A matriz $R_1 = (I-L)^{-1}U$ é a matriz de Gauss-Seidel de A.

Exemplo: Para o sistema (7.6), temos a iteração de Gauss-Seidel

$$\begin{aligned} x_1^{(k+1)} &= -\frac{1}{3}x_2^{(k)} - \frac{1}{3}x_3^{(k)} + \frac{5}{3} \\ x_2^{(k+1)} &= -\frac{1}{2}x_1^{(k+1)} + \frac{3}{2} \\ x_3^{(k+1)} &= -\frac{1}{2}x_1^{(k+1)} - \frac{1}{4}x_2^{(k+1)} + 3 \end{aligned} \qquad (7.9)$$

E por substituições sucessivas dos valores de $x_1^{(k+1)}$ e $x_2^{(k+1)}$ obtemos a matriz de Gauss-Seidel

$$R_1 = \begin{pmatrix} 0 & -1/3 & -1/3 \\ 0 & 1/6 & 1/6 \\ 0 & 1/8 & 1/8 \end{pmatrix} \quad ; \quad \rho(R_1) = \frac{7}{24} \simeq 0.29 \qquad (7.10)$$

Observação: Para a aplicação do método, a matriz R_1 não é calculada! Calculamos R_1, aqui, somente para estudar a convergência a ser explicado no §9.

7.3 A Sobre-relaxação

O problema $f(x) = 0$ pode ser atacado por iteração na forma $x = x + \omega f(x)$ com um parâmetro ω arbitrário. Analogamente, o sistema $b - Ax=0$ pode ser reduzido à forma $x=x + \omega D^{-1}(b-Ax)$, com $\omega \in \mathbb{R}$ qualquer, é resolvido por iteração (se converge) usando, como no método de Gauss-Seidel, os valores já determinados. Este procedimento corresponde à decomposição (7.4) de A, resultando no método de sobre-relaxação:

$$x^{(k+1)} = x^{(k)} + \omega D^{-1}(b-Dx^{(k)} + Ex^{(k+1)} + Fx^{(k)}) \qquad (7.11)$$

$$= (1-\omega)x^{(k)} + \omega(Lx^{(k+1)} + Ux^{(k)} + D^{-1}b) \qquad (7.12a)$$

O método de sobre-relaxação, então, pode ser interpretado como uma média ponderada entre $x^{(k)}$ e a aproximação $x^{(k+1)}$ pelo método de Gauss-Seidel.

Com $\det(I-\omega L)^{-1} = 1$ e $R_\omega=(I-\omega L)^{-1}(\omega U + (1-\omega)I)$, segue de (7.12a) que

$$x^{(k+1)} = R_\omega x^{(k)} + \omega(D-\omega E)^{-1}b \qquad (7.12b)$$

ou que para cada componente vale

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega\left(-\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ij}}x_j^{(k+1)} - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^{(k)} + \frac{b_i}{a_{ii}}\right) \qquad (7.12c)$$

Será mostrado no teorema 9.3 que $0<\omega<2$ é condição necessária para convergência. Em muitos casos, $\omega>1$, isto é, a 'correção' em (7.11) é maior que o erro $D^{-1}(b-Ax)$; daí o nome Sobre-relaxação.

Critérios de convergência para o método são encontrados nos teoremas 9.1, 9.2, 9.3, 10.2, 10.4, 11.4., 11.6., 11.7.

Definição 7.3

A matriz $R_\omega = (I-\omega L)^{-1}(\omega U+(1-\omega)I)$ é a matriz de sobre-relaxação de A.

Teorema 7.1

Seja $L \in \mathbb{C}(n,n)$ triangular inferior.

Então $(I-\omega L)^{-1} = I+\omega L+(\omega L)^2 + \ldots + (\omega L)^{n-1}$ (7.13)

e $\det(I-\omega L)^{-1} = 1$ (7.14)

Prova: $I-(\omega L + \ldots + (\omega L)^{n-1}) = I-\omega^n L^n$, e $L^n = \mathcal{O}$ □

Exemplo: A aplicação da sobre-relaxação em (7.6) fornece:

$$\begin{aligned}
x_1^{(k+1)} &= (1-\omega)\, x_1^{(k)} + \omega\, (-\tfrac{1}{3} x_2^{(k)} - \tfrac{1}{3} x_3^{(k)} + \tfrac{5}{3}) \\
x_2^{(k+1)} &= (1-\omega)\, x_2^{(k)} + \omega\, (-\tfrac{1}{2} x_1^{(k+1)} + \tfrac{3}{2}) \\
x_3^{(k+1)} &= (1-\omega)\, x_3^{(k)} + \omega\, (-\tfrac{1}{2} x_1^{(k+1)} - \tfrac{1}{4} x_2^{(k+1)} + 3)
\end{aligned} \qquad (7.15)$$

Compare-se (7.15) com a iteração de Gauss-Seidel, (7.9).

§ 8 Iteração de blocos

Seja $A \in \mathbb{C}(n,n)$ particionada em blocos

$$A_B = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{pmatrix}$$

com A_{ii} quadradas. Frequentemente, uma partição é sugerida pela forma de A. No exemplo (6.5), ela é consequência da enumeração dos pontos P_{ij}: Aos pontos P_{ij}, i=1(1) N, j fixo, (isto é, a cada linha de (6.1)), estão associados N blocos A_{ij}. Nesse contextos, os métodos-blocos são chamados métodos - linhas (line -methods).

Aplicando os métodos de iteração apresentados na seção anterior ao sistema:

$$\sum_{j=1}^{N} A_{ij}X_j = b_i;\ b_i, x_i \in \mathbb{R}^{m_i};\ i=1(1)N,$$

obtemos novos métodos da forma (6.2). Com as matrizes-blocos

$$D_B = \begin{pmatrix} A_{11} & & 0 \\ & A_{12} & \\ & & \ddots \\ 0 & & A_{NN} \end{pmatrix}, \quad E_B = \begin{pmatrix} 0 & & & 0 \\ & \ddots & & \\ -A_{21} & & \ddots & \\ \vdots & & & \\ -A_{N1} & \cdots & -A_{N,N-1} & 0 \end{pmatrix};$$

$$F_B = \begin{pmatrix} 0 & -A_{12} & \cdots & -A_{1N} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \\ 0 & & & 0 \end{pmatrix} \qquad (8.1)$$

as matrizes $L_B := D^{-1}E_B$ e $U_B := D_B^{-1}F_B$ são, respectivamente, triangular-bloco inferior e triangular-bloco superior. Como no §7, as partições (7.2) a (7.4) levam aos métodos bloco Jacobi, de Gauss-Seidel e de sobre-relaxação, tratados a seguir.

Agora, em cada iteração, devem ser resolvidos N sistemas lineares (com métodos não iterativos, por exemplo), pois a inversão de D_B não é mais feita por uma simples divisão pelos elementos diagonais, como no caso dos métodos do §7 .

Em geral, é aconselhável calcular e armazenar as inversas A_{ii}^{-1} antes da iteração. Frequentemente (por ex., em (6.5)) as matrizes A_{ii} consistem de uma diagonal principal, outra superior e outra inferior (matrizes tridiagonais), cujo tratamento é particularmente simples.

Vantagens da iteração-bloco:

1. Pode ser aplicada em casos em que o sistema é grande demais para ser armazenado.
2. Em certos casos, a iteração-bloco converge mais rápido que a iteração simples (ver teorema 10.5 e os exemplos (8.3) e (8.6)).

8.1 - O Método bloco de Jacobi

A decomposição $M=D_B$ e $N=E_B+F_B$, fornece o método bloco de Jacobi:

$$x^{(k+1)} = J_B\, x^{(k)} + D_B^{-1} b \quad ; \quad J_B := L_B+U_B \qquad (8.2)$$

Definição 8.1

$J_B = L_B + U_B$ é a matriz bloco de Jacobi de A.

Exemplos

1. Seja

$$A_B := \left(\begin{array}{cc|cc|c} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ \hline 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ \hline 0 & 0 & 0 & -1 & 1 \end{array}\right) \quad ; \quad b = \begin{pmatrix} 3 \\ -5 \\ 9 \\ -6 \\ 2 \end{pmatrix} \qquad (8.3)$$

Com a participação dada, $D_B x^{(k+1)} = (E_B + F_B)\, x^{(k)} + b$ é

$$\begin{array}{rcl} 2x_1^{(k+1)} - x_2^{(k+1)} & = & 3 \\ -x_1^{(k+1)} + 2x_2^{(k+1)} & = & x_3^{(k)} - 5 \\ \hline 2x_3^{(k+1)} - x_4^{(k+1)} & = & x_2^{(k)} + 9 \\ -x_3^{(k+1)} + 2x_4^{(k+1)} & = & x_5^{(k)} - 6 \\ \hline x_5^{(k+1)} & = & x_4^{(k)} + 2 \end{array} \qquad (8.4)$$

$$D_B^{-1} = \frac{1}{3}\left(\begin{array}{cc|cc|c} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ \hline 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ \hline 0 & 0 & 0 & 0 & 3 \end{array}\right) \quad \text{segue:} \quad J_B = \left(\begin{array}{cc|cc|c} 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 2/3 & 0 & 0 \\ \hline 0 & 2/3 & 0 & 0 & 1/3 \\ 0 & 1/3 & 0 & 0 & 2/3 \\ \hline 0 & 0 & 0 & 1 & 0 \end{array}\right)$$

1) Veja, por exemplo, Varga [/9/], pg. 195.

Vale:

$$\rho(J_B) = \frac{1}{3}\sqrt[4]{18} \cong 0.687 < \rho(J) \cong 0.866 \qquad (8.5)$$

A teoria apresentada mais adiante mostrará que o esquema de iteração bloco de Jacobi, nesse exemplo, converge mais rápido que o método (simples) de Jacobi

2. Uma partição possível do sistema (7.6) é

$$A_B = \left(\begin{array}{cc:c} 3 & 1 & 1 \\ 1 & 2 & 0 \\ \hdashline 1 & 1/2 & 2 \end{array}\right) \qquad (8.6)$$

então

$$D_B^{-1} = \frac{1}{10}\left(\begin{array}{cc:c} 4 & -2 & 0 \\ -2 & 6 & 0 \\ \hdashline 0 & 0 & 5 \end{array}\right), \quad J_B = \frac{1}{20}\left(\begin{array}{cc:c} 0 & 0 & -8 \\ 0 & 0 & 4 \\ \hdashline -10 & -5 & 0 \end{array}\right)$$

$$\rho(J_B) = \sqrt{\frac{3}{20}} \cong 0.387 < \rho(J) = \frac{1}{4}(1+\sqrt{\frac{7}{3}}) \cong 0.632 \qquad (8.7)$$

J_B é fracamente cíclico (de índice 2), apesar de J não o ser.

8.2 O método bloco de Gauss-Seidel e a sobre-relaxação bloco

A decomposição $M=D_B-E_B$ e $N=F_B$, fornece o método bloco de Gauss-Seidel:

$$D_B x^{(k+1)} = E_B x^{(k+1)} + F_B x^{(k+1)} + b, \qquad (8.8a)$$

isto é,

$$x^{(k+1)} = R_{1B}x^{(k)} + (D_B-E_B)^{-1}b;\ R_{1B} := (I-L_B)^{-1}U_B \qquad (8.8b)$$

Definição 8.2

$R_{1B} = (I-L_B)^{-1}U_B$ é a matriz bloco de Gauss-Seidel de A.

Para a partição (8.6) vale:

$$R_{1B} = \frac{1}{20}\begin{pmatrix} 0 & 0 & -8 \\ 0 & 0 & 4 \\ 0 & 0 & 3 \end{pmatrix}, \text{ com } \rho(R_{1B}) = \frac{3}{20} = 0.15 \qquad (8.9)$$

A decomposição $M = \frac{1}{\omega}(D_B - \omega E_B)$ e $N = \frac{1}{\omega}[\omega F_B + (1-\omega)D_B]$ dá a Sobre-relaxação bloco:

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega(L_B x^{(k+1)} + U_B x^{(k)} + D_B^{-1}b), \qquad (8.10)$$

isto é, $x^{(k+1)} = R_{\omega B}\, x^{(k)} + \omega(D_B-\omega E_B)^{-1}b$, (8.11)

Com $R_{\omega B} = (I-\omega L_B)^{-1}(\omega U_B + (1-\omega)I)$

Definição 8.3

$R_{\omega B} = (I-\omega L_B)^{-1}(\omega U_B + (1-\omega)I)$ é a matriz bloco de sobre-relaxação

Observa-se mais uma vez que as matrizes J, R_1, R_ω, J_B, R_{1B}, $R_{\omega B}$ não são necessárias no cálculo dos $x^{(k)}$, como mostrado nas equações (7.5b), (7.8c), (7.12c) e (8.4), apesar de serem de interesse teórico (no estabelecimento de critérios de convergência, por exemplo).

§ 9 CONVERGÊNCIA

9.1 Teoremas de Convergência

Todos os métodos descritos são da forma

$$x^{(k+1)} = Tx^{(k)}; \quad x^{(0)} = a; \quad k = 0,1... \tag{9.1}$$

com $Tx = Hx + c$, $H \in \mathbb{C}(n,n)$, $c \in \mathbb{C}^n$.

Para $x_1, x_2 \in \mathbb{C}^n$ arbitrarios e normas subordinadas vale

$$\|Tx_1 - Tx_2\| = \|Hx_1 - Hx_2\| \leq \|H\| \, \|x_1 - x_2\|$$

Logo T é uma contração em $\mathbb{C}^n$ se $\|H\| < 1$.

O teorema de ponto fixo de BANACH fornece a seguinte condição suficiente para convergência:

Teorema 9.1

A iteração $x^{(k+1)} = Hx^{(k)} + c$ converge, qualquer que seja $x^{(0)} \in \mathbb{C}^n$, para a solução u de $x = Hx + c$ se $P := \|H\| < 1$;

$$\|x^{(k+1)} - u\| \leq \frac{P}{1-P} \|x^{(k+1)} - x^{(k)}\| \tag{9.2}$$

O teorema seguinte fornece condições necessarias e suficientes para convergência:

Teorema 9.2

A iteração (9.1) converge para todo $x^{(0)} \in \mathbb{C}^n$ se e somente se

$$\rho(H) < 1 \tag{9.3}$$

sendo $\rho(H)$ o raio especial de H.

Prova:

Seja u solução de $x=Tx$ e $x^{(k)}$ a k-ésima aproximação de u. Então $u=Hu+c$ e $x^{(k+1)}=Hx^{(k)}+c$.

Para o erro $e^{(k)} = x^{(k)}-u$ vale $e^{(k+1)} = He^{(k)}$ e, por indução,

$$e^{(k)} = H^k e(0).$$

O resultado segue do Teorema 1.8 . ∎

Seja $B: S^{-1}HS$ a forma normal de Jordan de H.

Então $H^k = SB^kS^{-1}$. Por (1.5) os elementos de B^k dependem dos autovalores de H. Pode-se mostrar que, para k suficientemente grande, $\|H_1^k\| < \|H_2^k\|$ se $\rho(H_1) < \rho(H_2) < 1$.

Levando em conta (9.4), definimos

Definição 9.1

Sejam H_1 e H_2 matrizes de iteração; então a iteração $x^{(k+1)}=Hx^{(k)}+c$ com $H=H_1$ converge assintoticamente mais rápido que a iteração com $H=H_2$ se $\rho(H_1) < \rho(H_2) < 1$.

Observação: Se $\rho(H) < 1$, $\|H^k\|$ vai a zero, quando $k \to \infty$, mas não monotonamente! Pode acontecer que $\|H_2\|^m < \|H_1\|^m$, mas $\rho(H_1) < \rho(H_2)$. Então, depois de m iterações, pode acontecer que os resultados obtidos com H_2 sejam melhores que com H_1. Eventualmente ("assintoticamente") H_1 fará a iteração convergir mais rápido.

Exemplo:

Seja $H_1 = \begin{pmatrix} 0.90 & 4 \\ 0 & 0.90 \end{pmatrix}$, $H_2 = \begin{pmatrix} 0.90 & 0 \\ 0 & 0.95 \end{pmatrix}$.

Então $\rho(H_1) = 0.90$, $\rho(H_2) = 0.95$. e

$$H_1^k = \begin{pmatrix} (0.90)^k & 4k(0.90)^{k-1} \\ 0 & (0.90)^k \end{pmatrix} \quad ; \quad H_2^k = \begin{pmatrix} (0.90)^k & 0 \\ 0 & (0.95)^k \end{pmatrix};$$

$$\| H_1^k \|_\infty = (0.90)^{k-1} (4k+0.90) \; ; \; \| H_2^k \|_\infty = (0.95)^k$$

Vale $\rho(H_1) < \rho(H_2)$, mas $\| H_1^k \|_\infty > \| H_2^k \|_\infty$ para $k \leq 115$.

9.2 Aplicaçois

Consideremos os exemplos dos parágrafos 7 e 8. Para o sistema (7.6), $\rho(J) \cong 0.63$ e $\rho(R_1) \cong 0.29$ (por (7.7) e (7.10)). Logo, nesse caso, o método de Gauss-Seidel converge assintoticamente mais rápido que o de Jacobi:

A partição em blocos (8.6) do sistema fornece as matrizes J_B e R_{1B}, com $\rho(J_B) \cong 0.39$ e $\rho(R_{1B}) = 0.15$ (por (8.7) e (8.9)). Nesse caso, a iteração por blocos dá uma convergência assintótica ainda melhor.

Desses exemplos, não podemos concluir que a convergência do método de Jacobi é pior do que o de Gauss-Seidel sempre, como mostra esse exemplo de Collatz:

Exemplos

1 - Para $A_1 = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$, temos $J = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & 1 \\ -2 & -2 & 0 \end{pmatrix}$; $R_1 = \begin{pmatrix} 0 & -2 & 2 \\ 0 & -2 & -3 \\ 0 & 0 & 2 \end{pmatrix}$

Então $\rho(J) = 0$, isto é, a iteração de Jacobi converge

e $\rho(R_1) = 2$, isto é, a iteração de Gauss-Seidel diverge.

2 - Para

$$A_2 = \begin{pmatrix} 1 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 1 & 1 \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix} \quad \text{temos} \quad J = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ -1 & 0 & -1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \quad R_1 = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{3}{2} \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}.$$

Com $\rho(J) = \frac{\sqrt{5}}{2}$ (divergência da iteração de Jacobi) e

$\rho(R_1) = \frac{1}{2}$ (convergência da iteração de Gauss-Seidel).

No § 10 serão dados critérios que permitem determinar, em casos particulares importantes, que método converge assintoticamente mais rápido. Com o teorema 9.2, pode ser mostrado que o $<\omega<2$ é necessário para convergência da Sobre-relaxação:

Teorema 9.3. Seja $R_\omega = (I-\omega L)^{-1}(\omega U+(1-\omega)I)$; L, U como em (7.1).

Então, para todo ω real, $\rho(R_\omega) \geq |\omega-1|$.

Prova: Para os autovalores λ_i (i=1(1)n) de R_ω temos (1.2):

$$\prod_{i=1}^{n} \lambda_i = \det R_\omega = \det (I-\omega L)^{-1} \det (\omega U + (1-\omega)I).$$

Como L e U são triangulares,

$$\prod_{i=1}^{n} |\lambda_i| = |1-\omega|^n \leq \rho^n(R_\omega), \text{ já que } |\lambda_i| \leq \rho(R_\omega).$$

Logo $\rho(R_\omega) \geq |\omega-1|$. ∎

Em alguns casos especiais (veja Teoremas 10.2, 11.4 e 11.7), também não são indicados os $\omega \in (0,1)$.

§10 Teoremas de comparação e condições suficientes para convergência

Os exemplos da pag. 49 tornam claro que não existe um método iterativo que é o melhor para todas as matrizes. Entretanto, para classes especiais de matrizes, os métodos podem ser comparados e podem ser estabelecidas condições suficientes de convergência, de aplicação mais fácil do que a condição (9.3)

Nesse parágrafo, estudamos os casos especiais:

10.1 A matriz de Jacobi (ou a matriz bloco de Jacobi) de A é não negativa.

10.2 A satisfaz o "critério de soma de linhas" ou de colunas.

10.3 A é hermitiana.

10.1 - Matrizes de Jacobi não negativas

Nesse caso vale o critério de comparação de Stein-Rosenberg, confrontando os métodos de Jacobi e Gauss-Seidel:

Teorema 10.1

Seja $R_1=(I-L)^{-1}U$ a matriz de Gauss-Seidel e $J=L+U$ a matriz de Jacobi de A. Seja J irredutível e não-negativa. Então vale uma das três afirmações:

$$\begin{aligned} &(a) \quad \rho(R_1) < \rho(J) < 1 \\ &(b) \quad \rho(R_1) = \rho(J) = 1 \\ &(c) \quad \rho(R_1) > \rho(J) > 1 \end{aligned} \tag{10.1}$$

Prova: Pelo teorema 5.3(a): $\rho(J) > 0$; consequentemente, se $\rho(R_1) = 0$ vale (a).

Seja então $\lambda=\rho(R_1) > 0$ e, R_1 sendo não negativa,

$$R_1 x = \lambda x \geq \sigma$$

Logo $(L+\lambda^{-1}U)\, x = x$.

Temos $x>0$ pois $(L+\lambda^{-1}U)$ é irredutível e não negativa.

Seja $f(c) = \rho(L+c^{-1}U)$, então, para c de 0 até ∞, f decresce estritamente monótono e $f(\lambda)=1$, $f(1)=\rho(J)$, $\lambda = \rho(R_1)$. (10.2)

Logo segue (b); (a) e (c) são consequencia de

$$[1-\rho(J)][1-\rho(R_1)] = -(c-1)[f(c)-f(1)] \geq 0$$

□

O teorema seguinte mostra que o parâmetro de relaxação ω não deve ser tomado menor que 1, se J é não negativa e irredutível:

Teorema 10.2

Seja J matriz de Jacobi e R_ω matriz de sobre-relaxação de A (ou da partição bloco correspondente). Se J é não negativa, irredutível, e $\rho(J)< 1$, então

(a) a sobre-relaxação converge para todo $0<\omega \leq 1$,

(b) $\rho(R_{\omega_2}) < \rho(R_{\omega_1}) < 1$, se $0<\omega_1 < \omega_2 \leq 1$.

Prova Com $B_\omega := \frac{1}{\omega}(I-\omega L)$ e $C_\omega := \frac{1}{\omega}(\omega U+(1-\omega)I)$; $R_\omega = B_\omega^{-1} C_\omega$ e $S := B_\omega - C_\omega = I-J$. Se $0<\omega<1$, $C_\omega \geq 0$ e, pelo teorema 4.4, B_ω e S são inversíveis com $B_\omega^{-1} \geq \sigma$ e $S^{-1} \geq \sigma$.

Pelo teorema 5.7 e $R_\omega = B_\omega^{-1} C_\omega$

$$\rho(R_\omega) = \frac{\rho(S^{-1}C_\omega)}{1+\rho(S^{-1}C_\omega)} < 1, \text{ o que prova (a).}$$

Para provar(b), basta ver que $\rho(S^{-1}C_{\omega_1}) > \rho(S^{-1}C_{\omega_2})$ (10.3)

$S^{-1} = I+J+J^2+\ldots$ é não-negativa e irredutível, porque J é. Para $0<\omega< 1$, também $S^{-1}C_\omega$ é não-negativa e irredutível, pois $C_\omega := U + (\frac{1}{\omega} - 1)\, I = (c_{ij}) \geq 0$, com $c_{ii} \neq 0$.

(10.3) segue do teorema 5.3(e)

□

10.2 - O critério da soma das linhas (ou colunas)

Definição 10.1

$A \in \mathbb{C}(n,n)$ satisfaz o

(a) critério da soma das linhas se para todos i=1(1)n

$$r_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < |a_{ii}| \qquad (10.4a)$$

(b) critério da somda das colunas se para todos j=1(1)n

$$r_j = \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| < |a_{jj}| \qquad (10.4b)$$

(c) critério fraco da soma das linhas se para todos i=1(1)n

$$r_i \leq |a_{ii}| \qquad (10.5a)$$

e $r_i < |a_{ii}|$ para algum $i=i_o$

(d) critério fraco da soma das colunas se para todos j=1(1)n

$$r_j \leq |a_{jj}|, \qquad (10.5b)$$

e $r_j < |a_{jj}|$ para algum $j=j_o$

Teorema 10.3

Se $A \in \mathbb{C}(n,n)$ satisfaz

a) o critério da soma das linhas (ou colunas) ou

b) o critério fraco da soma das linhas (ou colunas) e A é irredutível,

então os métodos de Jacobi e Gauss-Seidel convergem.

Prova: No caso (a), pelo teorema 4.3[1] de Gerschgorin.

No caso (b), pelo teorema 5.4, $\rho(|J|) < 1$

Pelo teorema 5.5, $\rho(J) \leq \rho(|J|) < 1$.

Logo a convergência do método de Jacobi está mostrada.

Como a matriz L é triangular superior, vale para $R_1 = (I-L)^{-1}U$:

$$|R_1| = |(I+L+L^2+\ldots+L^{n-1})\,U| \leq (|I+|L|+\ldots+|L^{n-1}|\,|)\,|U|$$
$$\leq (I-|L|)^{-1}|U| := \hat{R}_1.$$

$|J|$ (resp. $\hat{R}_1$) é a matriz da Jacobi (resp. Gauss-Seidel) da matriz não negativa $|A|$. Pelos teoremas 5.5, 5.6 e 10.1

$$\rho(R_1) \leq \rho(|R_1|) \leq \rho(\hat{R}_1) \leq \rho(|J|) < 1. \qquad \square$$

No caso (a), o êrro da iteração pode ser avaliado por (9.2) com

$$P := \|J\|_\infty = \max_{0<i\leq n} \left|\frac{r_i}{a_{ii}}\right| < 1, \text{ ou } P := \|J\|_1 \qquad (10.6)$$

no método de Jacobi e com $P := \|R_1\|_\infty$ ou $P := \|R_1\|_1$, no método de Gauss-Seidel (veja Collatz |2|, pg. 230), dependendo da escolha de norma de vetores.

---

(1) Para o critério da Soma das Colunas, veja as observações nas páginas 48 e 29.

10.3 Matrizes hermitianas

Nesse caso, vale o seguinte teorema de Ostrowski:

Teorema 10.4

Seja $A \in \mathbb{C}(n,n)$ hermitiana, positiva definida, com $A=B-C-C^*$, onde B é positiva definida e $\det(B-\omega C) \neq 0$, para $0<\omega<2$. Então a iteração

$$x^{(k+1)} = (B-\omega C)^{-1}(\omega C^*+(1-\omega)B)x^{(k)} + \omega(B-\omega C)^{-1}b \qquad (10.7)$$

converge para todo $0<\omega<2$.

Prova[1]: Seja $S_\omega := (B-\omega C)^{-1}(\omega C^*+(1-\omega)B)$. Queremos mostrar que $\rho(S_\omega) < 1$. Como $\det(S_\omega - \lambda I) = \det(B-\omega C)^{-1}\det\left(\omega C^* + (1-\omega)B - \lambda(B-\omega C)\right)$, todos os autovalores de S_ω são raizes de

$$\det\ (\omega C^* + (1-\omega)B - \lambda(B-\omega C)) = 0.$$

Seja $x \neq 0$ autovetor de S_ω e λ autovalor correspondente, então

$$\lambda(B-\omega C)x = (\omega C^* + (1-\omega)B)x.$$

Multiplicando por $\frac{2}{\omega}$ e somando (resp. subtraindo) $0 = A-B+C^*+C$ nos parenteses, vem que

$$\lambda[B(\tfrac{2}{\omega} - 1) + A + (C^*-C)]\,x = [\,B(\tfrac{2}{\omega} - 1)-A + (C^*-C)]x. \qquad (10.8)$$

Como $(C^*-C)^* = -(C^*-C)$, $\bar{x}^T(C^*-C)x$ é imaginário puro para $x \neq 0$. Multiplicando (10.8) à esquerda por $\bar{x}^T$, temos

$$\lambda[(\tfrac{2}{\omega} -1)b + a+ic] = [(\tfrac{2}{\omega} -1)b - a+ic]$$

com $a := \bar{x}^T Ax$; $b := \bar{x}^T Bx$; $ic := \bar{x}^T(C^*-C)x$, C real. Se $\omega \in (0,2)$ e λ é autovalor de S_ω,

$$\lambda = \frac{|(\frac{2}{\omega} -1)\ b+ic-a|}{|(\frac{2}{\omega} -1)\ b+ic+a|} < 1, \text{ já que } (\tfrac{2}{\omega} -1)>0,\ a>0,\ b>0 \qquad \square$$

(1) De J. Albrecht.

Para B=D, C=E, C^*=F, obtemos de (10.7) a iteração por sobre-relaxação. Como $a_{ii} > 0$, segue do teorema 10.4

Corolário 10.4

Se A é hermitiana e positiva definida, o método de sobre-relaxação converge para $\omega \in (0,2)$.

O teorema 9.3 já mostrou que $\omega \in (0,2)$ é uma condição necessária para convergência; o corolário 10.4 mostra que também é suficiente no caso de matrizes A hermitianas, positivas definidas.

De maneira análoga são obtidas condições necessárias e suficientes para convergência da sobre-relaxação bloco de matrizes hermitianas positivas definidas.

O sistema Ax=b pode ser transformado num sistema com matriz hermitiana positiva definida via multiplicação por A^*: $A^*Ax=A^*b$ (veja teorema 1.4). Isto não é interessante fazer, porque o novo sistema é, em geral, pior condicionado do que o inicial (veja teorema 3.2).

Um caso particular de matrizes hermitianas positivas definidas são as matrizes de Stieltjes:

Definição 10.2

$A=(a_{ij}) \in R(n,n)$ é uma matriz de Stieltjes, se A é simetrica positiva definida e $a_{ij} \leq 0$ se $i \neq j$.

Teorema 10.5 (Varga)

Sejam $A=M_1-N_1 = M_2-N_2$ duas decomposições de A com $A^{-1} > \mathcal{O}$, $M_i^{-1} \geq \mathcal{O}$, $N_i \geq \mathcal{O}$ (i=1,2).

Se $N_1 \geq N_2 \geq \sigma$ $(N_2 \neq \sigma,\ N_1 \neq N_2)$ então

$$0 < \rho\ (M_2^{-1} N_2) < \rho\ (M_1^{-1} N_1) < 1.$$

Prova: Temos (veja teorema 5.7) $\rho\ (M^{-1} N) = \dfrac{\rho(A^{-1} N)}{1 + \rho(A^{-1} N)}$

Basta, então, provar que $0 < \rho\ (A^{-1} N_2) < \rho\ (A^{-1} N_1)$ (10.9)

Como $A^{-1} > \sigma$ temos $\sigma \leq A^{-1} N_2 \leq A^{-1} N_1$. (10.10)

Se $A^{-1} N_2$ é irredutível: (10.9) segue do teorema 5.3 (e).

Se $A^{-1} N_2$ é redutível : Então existe uma matriz de permutação P tal que

$$P\ (A^{-1} N_2) P^T = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ \sigma & R_{22} & \cdots & R_{2m} \\ \cdots & \cdots & \cdots & \\ \sigma & \sigma & & R_{mm} \end{pmatrix}$$

com R_{ii} irredutíveis ou matrizes (1x1) nulos. Sendo (10.10) invariante à transformação de similaridade com P, aplica-se o argumento acima às submatrizes R_{ii} (os autovalores de $A^{-1}N_2$ não dependem dos R_{ij} com $i \neq j$).

Fazendo $M_1 = D - E$ e $M_2 = D_B - E_B$ e levando em consideração que $A^{-1} > 0$ para matrizes de Stieltjes irredutíveis (veja [Ø], pag.85) segue do teorema 10.5. o

Corolário 10.5

Seja A uma matriz de Stieltjes irredutível. Se existe uma partição em blocos de A tal que pelo menos uma das matrizes A_{ii} tem elementos não-nulos fora da diagonal principal, então o método bloco converge assintóticamente mais rápido do que o método simples.

10.4 Exemplo

Consideremos o sistema $Ax = b$, com A dada por (8.3):

$$A = \left(\begin{array}{cc|cc|c} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ \hline 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ \hline 0 & 0 & 0 & -1 & 2 \end{array}\right) \quad \text{então } J = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 \end{pmatrix} \tag{10.11}$$

A satisfaz o critério fraco da soma das linhas, logo, pelo teorema 10.3, convergem as iterações de Gauss-Seidel e Jacobi.

J é não negativa e irredutível, logo pelo teorema 10.1 o método de Gauss-Seidel converge (assintoticamente) mais rápido que o de Jacobi.

A é hermitiana, logo(pelo teorema de Gerschgorin (pg.) e o Corolário 4.3) positiva definida. A sobre-relaxação converge para $\omega \in (0,2)$, pelo corolário 10.4.

Pelo teorema 10.2, a convergência assintótica é mais rápida para $\omega \geq 1$. (por resultados dos próximos parágrafos, pode-se mostrar que $\omega_o \simeq 1.333$ é o valor ótimo para ω e, nesse caso $\rho(R_{\omega_o}) \simeq 0.333$).

A é uma matriz de Stieltjes e irredutível. Logo pelo corolário 10.5, a iteração Gauss-Seidel por blocos converge (assintoticamente) mais rápido do que a iteração simples de Gauss-Seidel.

§ 11 Sistemas Consistentemente Ordenados

Para matrizes de Jacobi não-negativas, o teorema de Stein-Rosenberg (10.1) garante a convergência dos métodos de Jacobi e Gauss-Seidel.

Nesse parágrafo, comparamos, com hipóteses diferentes, os métodos de Jacobi e a sobre-relaxação, demonstrando uma relação entre autovalores das suas matrizes de iteração. Assim, pode-se tornar mais preciso o teorema de Stein-Rosenberg para certas matrizes e, em casos particulares, pode-se calcular o parâmetro ótimo $\omega = \omega_o$. Além disso, estuda-se o efeito de reordenações do sistema na convergência da sobre-relaxação.

11.1 Uma relação entre os autovalores de J e R_ω

Nessa seção, consideremos a pergunta: Sob que hipótese existe uma relação funcional entre os autovalores de $J = L+U$ e $R_\omega = (I-\omega L)^{-1} \cdot (\omega U + (1-\omega) I)$?

Definição 11.1

A matriz de Jacobi $J = L+U$ (ou matriz bloco de Jacobi) é consistentemente (r,q)-ordenada ($r,q \in \mathbb{N}$ primos entre si) quando os autovalores de $\alpha^r L + \alpha^{-q} U$ são independentes de $\alpha \neq 0$. O sistema $Ax=b$ é consistentemente ordenado se a matriz de Jacobi (ou matriz bloco de Jacobi) de A for consistentemente ordenada.[A]

A pergunta acima é respondida pelo teorema seguinte:

Teorema 11.1

Seja $J = L+U$ matriz de Jacobi (ou matriz bloco de Jacobi) e R_ω matriz de sobre-relaxação (ou matriz bloco de sobre-relaxação) de A, e seja J consistentemente (r,q) - ordenada.

(a) Se $\lambda \neq 0$ é autovalor de R_ω e se $\mu \in \mathbb{C}$ é dada por

$$(\lambda + \omega - 1)^p = \omega^p \mu^p \lambda^q \quad ; \; p := q+r; \; \omega \in (0,2) \qquad (11.1)$$

então μ é autovalor de J.

(b) Se μ é autovalor de J, satisfazendo (11.1) para algum λ, então λ autovalor de R_ω.

Além disso, $\mu \cdot \exp(\frac{k}{p} \cdot 2\pi i)$, $k = 1(1)\ p-1$, também são autovalores de J.

A) Observe-se que um sistema pode ser consistentemente ordenado com relação ao método bloco de Jacobi e não em relação ao método simples

Prova

(a) Seja $\lambda \neq 0$ o autovalor de $R_\omega \in \mathcal{C}(n,n)$. Então $\det (R_\omega - \lambda I) = 0$. Por (7.14)

$$\det (\omega \lambda L + \omega U - (\omega+\lambda - 1) I) = 0 \tag{11.2}$$

Dividindo por $(\omega \lambda^{q/p})^n$, obtemos

$$\det (\alpha^r L + \alpha^{-q} U - \mu I) = 0, \tag{11.3}$$

com $\alpha := \lambda^{1/p}$, $\mu := \dfrac{\omega + \lambda - 1}{\omega \lambda^{q/p}}$. (11.4)

Sendo J consistentemente (r,q) - ordenada, então, por (11.3), os $\mu \exp(\frac{k}{p} 2\pi i)$, $k = 0(1)p-1$, são autovalores de $L+U$, para todo λ (onde o fator $\exp(\frac{k}{p} 2\pi i)$ é consequência das raízes $\lambda^{1/p}$) isto é, vale (11.1).

(b) Se μ é autovalor de $J = L+U$, de $\det (L+U-\lambda I)=0$ e do fato de J ser consistentemente ordenada, temos que

$$\det (\alpha^r L + \alpha^{-q} U - \mu I) = 0$$

para $\alpha \neq 0$; mais ainda, se $\alpha = \lambda^{1/p}$ $(\lambda \neq 0)$,

$$\det (\lambda^{r/p} L + \lambda^{-q/p} U - \mu I) = 0$$

isto é,

$$\det (\omega\lambda L + \omega U + \omega\lambda^{q/p} \mu I) = 0. \tag{11.5}$$

Se $\lambda^{1/p} = 1$, $\mu \exp(\frac{k}{p} 2\pi i)$, $k = 0(1)\ p-1$, é autovalor.

Como λ satisfaz (11.1), $\mu\omega\lambda^{q/p} = \omega + \lambda - 1$, logo, de (11.5), vem que

$$\det(\omega\lambda L + \omega U - (\omega+\lambda-1)I) = 0,$$

o que mostra que λ é autovalor de R_ω (cf.(11.2)). □

No caso especial $\omega=1$, de (11.1) obtemos o seguinte

Corolário 11.1

Se o sistema Ax=b é consistentemente (r,q)-ordenado, então as iterações de Jacobi e Gauss-Seidel (e as iterações bloco) convergem ou divergem simultaneamente.

Se convergem, $\rho(R_1) = (\rho(J))^{(1+\frac{q}{r})} < 1$, (11.6)

isto é, a iteração de Gauss-Seidel converge mais rápido.

Temos, assim, um resultado mais preciso para convergência assintóticamente dos dois métodos do que o fornecido pelo teorema de Stein-Rosenberg, sob hipótese bem diferentes, entretanto.

Exemplo: A transformação por similaridade $S(\alpha)JS^{-1}(\alpha)$ não altera os autovalores de J. Logo estes independem de α.

No caso da matriz (8.3) temos $S(\alpha)JS^{-1}(\alpha)=\alpha L + \alpha^{-1}U$,

com

$$S(\alpha) = \begin{pmatrix} 1 & & & & \\ & \alpha & & 0 & \\ & & \alpha^2 & & \\ & 0 & & \alpha^3 & \\ & & & & \alpha^4 \end{pmatrix}, \alpha \neq 0$$

Logo, os autovalores de $\alpha L+\alpha+\alpha^{-1}U$ são independentes de α e J é consistentemente (1,1)-ordenada.

De $\rho(J)=0,866$ (cf.(8.5)) vem pelo corolário 11.1: $\rho(R_1)=(0.866)^2 \simeq 0.750$ e para a iteração bloco, vem de $\rho(J_B)=\frac{1}{3}\sqrt[4]{18}=0,687$: $\rho(R_{1B}) = \frac{1}{3}\sqrt{2} \simeq 0.471$

A próxima seção mostra como decidir se um sistema é consistentenmente ordenado, numa situação particular.

11.2 Uma classe de matrizes consistentemente (r,q)-ordenadas

Em geral, é difícil decidir se um sistema é consistentemente ordenado, ou pode vir a sê-lo, por reordenações. Um conjunto de matrizes (frequentes na prática) com essa propriedade são as matrizes banda, definida a seguir. Um conjunto de matrizes que por reordenação simultânea de linhas e colunas se tornam consistentemente ordenadas são, como será mostrado, as matrizes fracamente cíclicas (cf. seção 5.2). Obviamente, J é consistentemente (r,q)-ordenada se existem matrizes inversíveis $S(\alpha)$ tais que

$$S(\alpha)JS^{-1}(\alpha) = \alpha^r L + \alpha^{-q}U;\ \alpha \neq 0;\ r,q \in \mathbb{N},\ \text{primos entre si.} \quad (11.7)$$

No caso especial

$$S(\alpha) = \begin{pmatrix} I_1 & & & \mathcal{O} \\ & \alpha I_2 & & \\ & & \alpha^2 I_3 & \\ \mathcal{O} & & & \ddots \alpha^{N-1} I_N \end{pmatrix} ; \; I_j \text{: matrizes identidade}, \; j = 1(1)N \qquad (11.8)$$

temos que

$$S(\alpha)JS^{-1}(\alpha) = \begin{pmatrix} \mathcal{O} & \alpha^{-1}U_{12} & \alpha^{-2}U_{13} \cdots\cdots & \alpha^{-N+1}U_{1N} \\ \alpha L_{21} & \mathcal{O} & \alpha^{-1}U_{23} \cdots\cdots & \alpha^{-N+2}U_{2N} \\ \alpha^2 L_{31} & \alpha L_{32} & \mathcal{O} \cdots\cdots & \alpha^{-N+3}U_{3N} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \alpha^{N-1}L_{N1} & \alpha^{N-2}L_{N2} & \alpha^{N-3}L_{N3} \cdots\cdots & \mathcal{O} \end{pmatrix} \qquad (11.9)$$

e a matriz é consistentemente (r,q)-ordenada se todas as submatrizes são nulas, com exceção das matrizes da r-ésima diagonal inferior (i.é., as matrizes $L_{r+j,j}$, j=1(1)N-r) e da q-ésima diagonal superior (i.é., $U_{j,q+j}$, j= 1(1)N-q).

As matrizes descritas acima são chamadas de <u>matrizes (r,q)-bandas</u>. Segue o seguinte teorema:

<u>Teorema 11.2</u>

Matrizes (r,q)-bandas são consistentemente (r,q)-ordenadas.

<u>Exemplos:</u>

1 - A matriz de Jacobi (10.11) é consistentemente (1,1)-ordenada , assim como a matriz bloco de Jacobi J_B (veja pag. 43).

2.- A matriz de Jacobi

$$\left(\begin{array}{c|ccc|c} 0 & 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 1 & 0 \end{array}\right) \qquad (11.10)$$

é consistentemente (1,1)-ordenada, como mostra a partição em blocos. (note que J não precisa ser uma matriz de iteração bloco).

3.- A matriz bloco de Jacobi do sistema (6.5) é consistentemente (1,1)-ordenada, mas sua matriz de Jacobi não.

OBSERVAÇÃO: Note que a matriz J pode se tornar uma matriz banda com uma determinada partição em blocos. Isto não implica que métodos bloco tenham que ser aplicado!

Matrizes fracamente cíclicas de índice p na forma normal (5.4) são consistentemente (1,p-1)-ordenadas, como (11.9) mostra; além disso, todas as matrizes (r,q)-bandas são fracamente cíclicas de índice(r+q), como mostram seu grafo e o teorema 5.2. Logo vale o seguinte teorema:

Teorema 11.3

Se J é fracamente cíclica de índice p (em particular, se J é (r,q)-banda, r+q=p) então existe uma matriz de permutação P tal que PJP^T é consistentemente (1,p-1) ordenada.

Assim, pelo teorema 5.2, podemos decidir se uma matriz pode ser reordenada numa forma consistente.

Exemplos:

1. A matriz

$$J_1 = \begin{pmatrix} 0 & 1 & 0 & -4 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 5 \\ -2 & 0 & 1 & 0 \end{pmatrix} \tag{11.11}$$

não é consistentemente ordenada. Ela é, entretanto, fracamente cíclica de índice 2 e (pelo teorema 11.2), pode ser tornada consistentemente (1,1)-ordenada. No caso, P é dada por

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{e} \quad PJP^T = \begin{pmatrix} 0 & 0 & 1 & -4 \\ 0 & 0 & 1 & 5 \\ 1 & 2 & 0 & 0 \\ -2 & 1 & 0 & 0 \end{pmatrix}$$

2. A matriz

$$J_2 = \frac{1}{4} \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{11.12}$$

é uma matriz (2,3)-banda, logo consistentemente (2,3)-ordenada. Sob a ação da permutação $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 5 & 2 & 4 \end{pmatrix}$, J_2 torna-se consistentemente (1,4)-ordenada.

3. Seja

$$A = \left(\begin{array}{cc|cc|cc} 20 & -4 & -4 & -1 & 0 & 0 \\ -4 & 20 & -1 & -4 & 0 & 0 \\ \hline -4 & -1 & 20 & -4 & -4 & -1 \\ -1 & -4 & -4 & 20 & -1 & -4 \\ \hline 0 & 0 & -4 & -1 & 20 & -4 \\ 0 & 0 & -1 & -4 & -4 & 20 \end{array}\right) \qquad (11.13)$$

Sua matriz de Jacobi J não é fracamente cíclica. Com a partição indicada, entretanto, J_B é uma matriz (1,1)-banda, logo consistentemente (1,1)-ordenada.

4. A matriz de Jacobi J do sistema (6.5) é fracamente cíclica de índice 2. A permutação $P=\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 3 & 5 & 7 & 9 & 2 & 4 & 6 & 8 \end{pmatrix}$ faz com que J se torne matriz (1,1)-banda;

$$\hat{J} := PJP^T = \frac{1}{4}\begin{pmatrix} & & & & & 1 & 1 & 0 & 0 \\ & & & & & 1 & 0 & 1 & 0 \\ & & \sigma & & & 1 & 1 & 1 & 1 \\ & & & & & 0 & 1 & 0 & 1 \\ & & & & & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & & & & \\ 1 & 0 & 1 & 1 & 0 & & & & \\ & & & & & & & \sigma & \\ 0 & 1 & 1 & 0 & 1 & & & & \\ 0 & 0 & 1 & 1 & 1 & & & & \end{pmatrix}$$

Entretanto, não é facil vêr, com base da teoria apresentada, que J já é consistentemente (1,1)-ordenada.

A matriz bloco J_B associada a (6.5) é uma matriz (1,1)-banda e, consequentemente, (1,1)-ordenada.

5. Seja (Varga)

$$A = \left(\begin{array}{cc|cc|cc} 8 & -1 & 0 & -1 & -1 & -1 \\ -1 & 8 & 0 & -1 & -1 & -1 \\ \hline -1 & 1 & 8 & -1 & 8 & -8 \\ 8 & -8 & -1 & 8 & -1 & 1 \\ \hline -1 & -1 & -1 & 0 & 8 & -1 \\ -1 & -1 & -1 & 0 & -1 & 8 \end{array}\right)$$

A matriz bloco de Jacobi (associada à partição em blocos indicada) tem a forma

$$J_B = \begin{pmatrix} 0 & B_{12} & B_{13} \\ B_{21} & 0 & B_{23} \\ B_{31} & B_{32} & 0 \end{pmatrix}$$

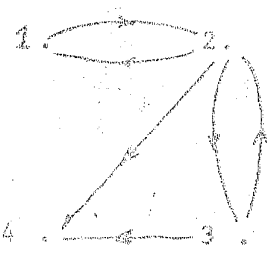
que, numa primeira inspeção, não parece fracamente cíclica. O cálculo revela, entretanto, que J_B é fracamente cíclica de índice 2:

$$J_B = \frac{1}{7}\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & -7 & -7 \\ -7 & -7 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad (11.14)$$

Concluimos, então, que uma matriz bloco pode ser fracamente cíclica se o seu "grafo bloco" (o grafo que resulte interpretando os blocos como elementos) for primitivo. Entretanto, caso o "grafo bloco" fosse cíclico, então a matriz seria fracamente cíclica.

6. O grafo G da matriz

$$J = \frac{1}{4}\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$



não é cíclico e nem pode ser tornado (veja exemplo 2, pg. 27). Pelo teorema 5.2, J não é fracamente cíclico. Entretanto, J é consistentemente (1,1)-ordenada, pois
$\det(\alpha L + \alpha^{-1}U - \mu I) = \mu^4 - \frac{1}{8}\mu^2$ não depende de α.

Este exemplo mostra que não são apenas as matrizes fracamente cíclicas que podem ser tornadas consistentemente ordenadas. A seguir, um outro exemplo:

7. Seja (Broyden[1])

$$J = \begin{pmatrix} 0 & -b & b & -b \\ -b & 0 & 0 & -b \\ b & 0 & 0 & -b \\ -b & -b & -b & 0 \end{pmatrix}, \quad S(\alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1+\alpha}{2} & \frac{1-\alpha}{2} & 0 \\ 0 & \frac{1-\alpha}{2} & \frac{1+\alpha}{2} & 0 \\ 0 & 0 & 0 & \alpha \end{pmatrix} \qquad \alpha \neq 0.$$

Então $S(\alpha)JS^{-1}(\alpha) = \alpha L + \alpha^{-1}U$. J é consistentemente (1,1)-ordenada, mas não é fracamente cíclica.

---

1) C.G.Broyden: "Some geralizations of the theory of successive over-relaxation" Num.Math. 6, 1964, p.269-284

II.3 Influência das reordenações na convergência dos métodos

Seja $P \in \mathbb{R}^{(n,n)}$ matriz de permutação. Então a reordenação PAP^T não altera o raio espectral da matriz de Jacobi. Se A tem a matriz de Jacobi $J_1 = L+U = -D^{-1}(A+I)$, PAP^T tem a matriz de Jacobi

$$J_2 = -PD^{-1}IP^TPAP^T + I = PJ_1P^T \quad (\text{já que } P^TP=I).$$

Pelo teorema 1.1.: $\rho(J_1) = \rho(J_2)$.

Entretanto, reordenações 1) alteram o raio espectral da matriz de sobre-relaxação. Por isso, é possível melhorar a convergência da sobre-relaxação por reordenações.

Exemplo:

$$\text{Se } A = \begin{pmatrix} 2 & -1 & 0 \\ 0 & 2 & -1 \\ -1 & 0 & 2 \end{pmatrix}, \text{ então } J = \begin{pmatrix} 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix}; R_1 = \begin{pmatrix} 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \\ 0 & 1/4 & 0 \end{pmatrix}$$

com $\rho(J) = \frac{1}{2}$, $\rho(R_1) = \sqrt[3]{\frac{1}{8}}$.

J é uma matriz (2,1) banda e fracamente cíclica de índice 3. Pelo teorema II.2, pode ser reordenada para uma matriz consistentemente (1.2)-ordenada.

---

1) No texto, 'reordenação' quer sempre dizer reordenação simultânea de linhas e colunas de A, via PAP^T. Isso corresponde a uma permutação das equações, do sistema, junto com a permutação equivalente das incógnitas, x_k.

Com $P=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}$, obtemos um outro sistema

$\hat{A}y = \hat{b}$, com $y = Px$, $\hat{b}=Pb$

$$\hat{A}=PAP^T=\begin{pmatrix}2&0&-1\\-1&2&0\\0&-1&2\end{pmatrix} \quad ,\hat{J} = \begin{pmatrix}0&0&\frac{1}{2}\\\frac{1}{2}&0&0\\0&\frac{1}{2}&0\end{pmatrix} \quad \hat{R}_1=\begin{pmatrix}0&0&\frac{1}{2}\\0&0&\frac{1}{4}\\0&0&\frac{1}{8}\end{pmatrix}$$

Nesse caso, $\rho(\hat{J}) = \frac{1}{2}$, $\rho(\hat{R}) = \frac{1}{8}$

$\hat{J}$ é consistentemente (1,2)-ordenada. Este exemplo confirma o corolário 11.1 e mostra que reordenações podem ter efeitos favoráveis à convergência assintótica do método de Gauss-Seidel. Se existem várias ordenações consistentes de um sistema (como, por exemplo, no caso das matrizes (r-q)-banda, com r+q>2), podemos decidir, usando o corolário 11.1, qual delas é a mais favorável para a convergência assintótica do método de Gauss-Seidel.

Resultados para o caso mais geral do método de sobre-relaxação são dados pelo seguinte teorema, que também permite calcular o melhor parâmetro ω_o para a sobre-relaxação. Afirmações sobre a convergência da sobre-relaxação, sob hipótese ainda mais fracas, podem ser obtidas por teorema 11.1.

Teorema 11.4 [9]

Suponha que a matriz de Jacobi J (ou de Jacobi em blocos) seja consistentemente (r,q)-ordenada, $\rho(J)< 1$ e os autovalores de J^p ($p:=r+q$) sejam reais e não negativos. Então

a) se $r\neq 1$, $\min_\omega \rho(R_\omega) = \rho(R_1) = (\rho(J))^{(1+\frac{q}{r})}$

b) se $r= 1$, $\min_\omega \rho(R_\omega) = \rho(R_{\omega_o}) = (\omega_o-1)_q < 1$,

onde ω_o é a raiz (única) de

$$(\rho(J)\omega_o)^p = p^p\, q^{-q}(\omega_o -1)^q \; ; \; p := q+r \qquad (11.15)$$

no intervalo $[1, (q+1)/q]$.

A prova se baseia em uma análise de (11.1). Vamos nos restringir ao caso importante r=q=1(ao qual pertencem, entre outros, as matrizes de Jacobi simétricas, fracamente cíclicas de índice 2).

Seja J consistentemente (1,1)-ordenada e suponhamos que os autovalores de J^2 sejam não negativos; então, pelo teorema 1.5, J só tem autovalores reais e, pelo teorema 11.1(b), se μ é autovalor, $-\mu$ também é; logo $\mu = \rho(J)$ é autovalor de J. De (11.1), vem que

$$\left|\lambda(\mu,\omega)\right| = \left|\frac{1}{2}\,\omega^2\mu^2 - (\omega-1) \pm \frac{1}{2}\,\omega\mu\sqrt{\omega^2\mu^2 - 4(\omega-1)}\right| \qquad (11.16)$$

é autovalor de R_ω.

Se $\omega^2\mu^2 \le 4(\omega-1)$: $|\lambda|^2 = (\omega-1)^2$, isto é, $|\lambda| = |\omega-1|$ para todos os autovalores λ de R_ω .

Se $\omega^2\mu^2 > 4(\omega-1)$: $\rho(R_\omega) = |\lambda(\mu,\omega)|$, para $\mu=\rho(J)$.

$\rho(R_\omega)$ é monotonamente decrescente para $\omega\in(0,\omega_o)$ (Veja figura 11.1).

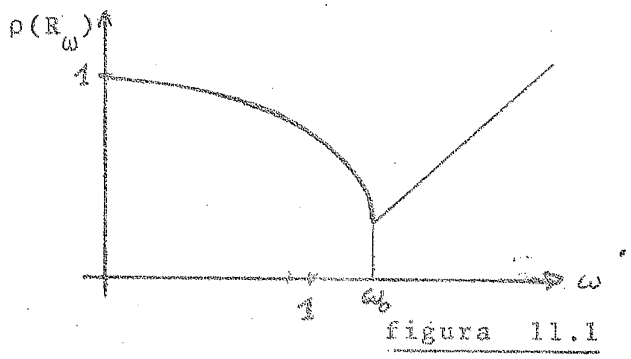


figura 11.1

Além disso, $\rho(R_\omega)$ é mínimo em $\omega = \omega_o$, com

$$\omega_o^2\mu^2 = 4(\omega_o-1) \text{ , isto é, } \omega_o = \frac{2}{1+\sqrt{1-\mu^2}} \qquad (11.17)$$

o que prova(11.15) se r=q=1. □

Como $\mu = \rho(J)$, em geral, não é conhecido exatamente, na prática ω_o só pode ser estimado. A figura 11.1 mostra que é mais favorável tornar ω_o grande do que pequeno.

Exemplo: O sistema

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ 2x_2 - x_3 &= 0 \\ -x_1 + 2x_3 &= 1 \end{aligned} \qquad (11.18)$$

tem a matriz de Jacobi

$$J = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 \end{pmatrix} \qquad (11.19)$$

J é uma matriz (2,1)-banda, logo o sistema (11.18) é consistentemente (2,1)-ordenado. Pelo teorema 5.4, $\rho(J) = \frac{1}{2}$ e pelo corolário 11.1 , $\rho(R_1) = \sqrt{\frac{1}{8}} = 0.354$.

Como J é fracamente cíclica de índice 3, o sistema (11.18), pelo teorema 11.2, pode ser reordenado para tornar-se consistentemente (1,2) ordenado. Basta tomar $P = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$, obtendo

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ -x_1 + 2x_3 &= 1 \\ -x_3 + 2x_2 &= 0 \end{aligned} \qquad (11.20)$$

com a matriz de Jacobi J consistentemente (1,2) ordenada

$$\hat{J} = \begin{pmatrix} 0 & 0 & 1/2 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix} \quad ; \quad \rho(\hat{J}) = \frac{1}{2} \quad . \tag{11.21}$$

Pelo corolário 11.1, vale para a matriz de Gauss-Seidel $\hat{R}_1$:

$$\rho(\hat{R}_1) = \frac{1}{8} = 0.125$$

A iteração de Gauss-Seidel converge assintoticamente muito mais rápido no caso do sistema reordenado (11.20), do que no caso do sistema original (11.18) .

Pelo teorema 11.1(b), os autovalores de $\hat{J}$ são

$$\mu_k = \frac{1}{2} \exp\left(\frac{k}{3} 2\pi i\right) \quad , \quad k = 0,1,2.$$

J^3, então, só tem autovalores positivos e de (11.15) obtem-se o ω_o ótimo como raiz real de

$$\left(\frac{1}{2}\omega_o\right)^3 = \frac{27}{4}(\omega_o - 1)$$

Daí, $\omega_o \cong 1.0196$ e o teorema 11.3(b) fornece $\rho(R_{\omega_o}) = 0.0392$.

11.4. Resultados adicionais

Nessa seção, serão apresentados, sem demonstração, resultados adicionais [1] que, num certo sentido, representam uma generalização da teoria da seção anterior para uma classe maior de matrizes.

Seja S a classe das matrizes J com as propriedades:

1. $J \geq \mathcal{O}$ com elementos diagonais nulos,
2. J é irredutível e $0 < \rho(J) < 1$,
3. J é simétrica, $J = L + L^T$,

e seja $R_\omega = (I-\omega L)^{-1}\,[\omega L^T + (1-\omega)I]$ a matriz da sobre-relaxação associada a J.

Teorema 11.5

Seja $J \in S$; então $\rho^2(J) \leq \rho(R_1) < \frac{\rho(J)}{2-\rho(J)}$ (11.22)

com igualdade somente se J for fracamente cíclica de índice 2 e consistentemente (1,1)-ordenada.

Para matrizes $J \in S$; isto é um resultado mais forte do que o de teorema 10.1 de Stein-Rosenberg.

Mesmo se $J \in S$ não for consistentemente ordenada, $\rho(R_1)$ difere do melhor valor $\rho^2(J)$ por menos que $\frac{1}{2}(5\sqrt{5}-11) \approx 0{,}090$.

---

[1] de Varga [10], capítulo 4.4

Teorema 11.6

Seja $J \in S$ e $\omega_o = \dfrac{2}{1 + \sqrt{1-\rho^2(J)}}$ (veja (11.17));

então $$\omega_o - 1 \leq \rho(R_{\omega_o}) < \sqrt{\omega_o - 1} < \rho(J)$$

com $\rho(R_{\omega_o}) = \omega_o - 1$ somente se J for fracamente cíclica de índice 2 e consistentemente (1,1)-ordenada.

Este resultado mostra que o parâmetro ω_o - mesmo <u>não</u> sendo o melhor para todos os $J \in S$ - é uma boa escolha para $J \in S$.

Teorema 11.7

Seja $J \in S$ e ω_o como no teorema anterior; então para todo $\omega: (\omega_o - 1) \leq \min_{\omega} \rho(R_\omega)$ com igualdade somente se J for fracamente cíclica de índice 2 e consistentemente (1,1)-ordenada.

Se $J \in S$, então todas as reordenações PJP^T de J com matrizes de permutação P são também elementos de S. Por isso, o teorema 11.6 implica que entre todas as reordenações de $J \in S$ a ordenação consistente - quando existe - resulta na melhor convergência assintótica.

Capítulo III

APLICAÇÕES EM EQUAÇÕES DIFERENCIAIS PARCIAIS

§ 12 A discretização da equação diferencial parcial (6.3)

12.1 Convergência da iteração

Nessa seção, serão aplicados os resultados obtidos no problema apresentado no § 6("problema de Dirichlet").

A discretização de (6.3) fornece para n=3 o sistema (6.6), cuja matriz de Jacobi e matriz bloco de Jacobi são, resp.,

$$J=\frac{1}{4}\left(\begin{array}{ccc|ccc|ccc}
0&1&0&1&0&0&0&0&0\\
1&0&1&0&1&0&0&0&0\\
0&1&0&1&0&1&0&0&0\\
\hline
1&0&1&0&1&0&1&0&0\\
0&1&0&1&0&1&0&1&0\\
0&0&1&0&1&0&1&0&1\\
\hline
0&0&0&1&0&1&0&1&0\\
0&0&0&0&1&0&1&0&1\\
0&0&0&0&0&1&0&1&0
\end{array}\right) \qquad \text{e}$$

$$J_B=\frac{1}{56}\left(\begin{array}{ccc|ccc|ccc}
0&0&0&15&4&1&0&0&0\\
0&0&0&4&16&4&0&0&0\\
0&0&0&1&4&5&0&0&0\\
\hline
15&4&1&0&0&0&15&4&1\\
4&16&4&0&0&0&4&16&4\\
1&4&5&0&0&0&1&4&5\\
\hline
0&0&0&15&4&1&0&0&0\\
0&0&0&4&16&4&0&0&0\\
0&0&0&1&4&5&0&0&0
\end{array}\right) \qquad (12.1)$$

J e J_B são não-negativas e irredutíveis. Pelo teorema 5.4

$$\frac{1}{2} < \rho(J) < 1 \qquad \text{e} \qquad \frac{20}{56} < \rho(J_B) < \frac{48}{56} \tag{12.2}$$

As duas iterações convergem, pelo teorema 9.2.(e, pelo teorema 10.5., a iteração bloco converge assintóticamente mais rápida).
As iterações de Gauss-Seidel (e Gauss-Seidel bloco) convergem assintóticamente mais rápido que as respectivas iterações de Jacobi, pelo teorema de Stein-Rosenberg.

A matriz A do sistema (pg. 34, (6.6)) é simétrica, logo só tem autovalores reais λ_j, $j = 1(1)9$, pelo teorema 1.2. Pelo teorema de Gerschgorin, $\lambda_j \geq 0$. Se existisse um autovalor λ nulo, então, pelo corolário 4.3, todos os circulos passariam por $\lambda=0$; logo $\lambda_j > 0$.

A, então, é positiva definida, pelo teorema 1.3, e, como $a_{ij} \leq 0$ para $i \neq j$, é uma matriz de Stieltjes. Pelo corolário 10.5, a iteração bloco de Gauss-Seidel converge assintóticamente mais rápido que a iteração simples. J_B é uma matriz (1,1)-banda, logo consistentemente (1,1)ordenada.
J_B^2 só tem autovalores reais não-negativos, porque J_B é simétrica (teoremas 1.2 e 1.5). Pelo teorema 11.4, $\rho(R_{1B}) = (\rho(J_B))^2$.

$$\rho(R_{\omega_o B}) = (\omega_o - 1), \text{ com } \omega_o = \frac{2}{1 + \sqrt{1-(\rho(J_B))^2}} \tag{12.3}$$

De (12.2),

$$\frac{2}{1 + \sqrt{1-(\frac{20}{56})^2}} \simeq 1.034 < \omega_o < \frac{2}{1 + \sqrt{1-(\frac{48}{56})^2}} \simeq 1.320$$

Pelo teorema 11.4, a sobre-relaxação bloco converge para $1.034 < \omega_o < 1.320$ assintóticamente mais rápido que a iteração bloco de Gauss-Seidel.

O valor exato de $\rho(J_B)$ é $\rho(J_B) = \frac{\cos \frac{\pi}{4}}{2-\cos \frac{\pi}{4}} \simeq 0.5469$.
Por (12.3): $\omega_o \simeq 1.176$; daí $\rho(R_{\omega_o B}) \simeq 0,176$ (pelo teorema 11.4).

A sobre-relaxação bloco com $\omega_o \approx 1.176$ converge assintóticamente mais rápido que todos os métodos descritos. A aplicação deste método, somente exige uma única inversão, a do primeiro bloco do diagonal. O êrro após k iterações por (9.4) depende de H^k, H: matriz de iteração. Para ter uma idéia das diferentes eficiências dos métodos foram calculados $\|J^k\|_1$, $\|J_B^k\|_1$, $\|R_1^k\|_1$, $\|R_{1B}^k\|_1$ e $\|R_{\omega_o B}^k\|_1$, para k = 10.

$$\|J^{10}\|_1 \approx 0{,}47.10^{-1} \quad ; \qquad \|R_1^{10}\|_1 \approx 0{,}30.10^{-2}$$

$$\|J_B^{10}\|_1 \approx 0{,}29.10^{-2} \quad ; \qquad \|R_{1B}^{10}\|_1 \approx 0{,}24.10^{-4}$$

$$\|R_{\omega_o B}^{10}\|_1 \approx 0{,}33.10^{-6}$$

que comprova a superioridade da sobre-relaxação com o parâmetro ótimo ω_o.

12.2 Convergência da discretização de (6.3)

Seja $u_{ij}=u(x_i,y_j)$ a solução exata de (6.3) no ponto $P_{ij}=(x_i,y_j)$ do interior do quadrado Q; $x_i=ih$, $y_j=jh$, $h=\frac{1}{n+1}$

Para u suficientemente diferenciável obtemos por desenvolvimento de Taylor:

$$u_{i+1,j}=\left(u+h\frac{\partial u}{\partial x}+\frac{1}{2}h^2\frac{\partial^2 u}{\partial x^2}+\frac{1}{6}h^3\frac{\partial^3 u}{\partial x^3}+\frac{1}{24}\frac{\partial^4 u}{\partial x^4}+\ldots\right)_{ij}$$

$$u_{i-1,j}=\left(u-h\frac{\partial u}{\partial x}+\frac{1}{2}h^2\frac{\partial^2 u}{\partial x^2}-\frac{1}{6}h^3\frac{\partial^3 u}{\partial x^3}+\frac{1}{24}\frac{\partial^4 u}{\partial x^4}+\ldots\right)_{ij}$$

Assim

$$u_{i+1,j}-2u_{i,j}+u_{i-1,j}=\left(h^2\frac{\partial^2 u}{\partial x^2}+\frac{1}{12}h^4\frac{\partial^4 u}{\partial x^4}+\ldots\right)_{ij}$$

Da mesma maneira:

$$u_{i,j+1}-2u_{i,j}+u_{i,j-1}=\left(h^2\frac{\partial^2 u}{\partial y^2}+\frac{1}{12}h^4\frac{\partial^4 u}{\partial y^4}+\ldots\right)_{ij}$$

Então para i=1(1)n, j=1(1)n:

$$h^{-2}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{ij}) = [(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) +$$

$$+ \frac{1}{12} h^2 (\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}) + \ldots]_{ij}$$

Segue então de $u_{xx} + u_{yy}=0$ o sistema

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{ij} = h^2 t_{i,j}(h) \qquad (12.4)$$

$$\text{com } t_{i,j}(h) = -\frac{1}{12} h^2 [u_{xxxx}(x_i+\xi_i h, y_i) + u_{yyyy}(x_i, y_j+\eta_j h)], \quad |\xi_i|<1, |\eta_j|<1 \qquad (12.5)$$

$$\text{ou seja} \qquad Au = b + h^2 t_h \qquad (12.6)$$

com $u=(u_{1,1},\ldots,u_{n,n})$, $t_h=(t_{1,1}(h),\ldots,t_{n,n}(h))$
onde A e b têm a forma indicada em (6.5).

Se as derivadas de u são contínuas até a ordem 4, então existe um número M tal que $|t_{i,j}(h)| \leq Mh^2$. (12.7)

O método de discretização consiste na substituição do sistema (12.6) por $Au_h=b$; $u_h=(U_{11},\ldots,U_{nn})$ (12.8)

esperando que, para h suficientemente pequeno, o vetor u_h obtido de (12.8) aproxima a solução exata u de (6.3) nos pontos P_{ij}.

Definição 12.1.

O vetor u_h obtido por algum método de discretização com passo h converge para a solução u do problema (6.3) se $\lim_{h\to 0} \|u-u_h\|_\infty = 0$; (12.9)

O método tem ordem de convergência p, se $\|u-u_h\|_\infty = O(h^p)$.

É obvio que a convergência do método iterativo aplicado no cálculo de u_h não implica na convergência de u_h para u.
Por isso, ainda é necessário provar (12.9).

De (12.6) e (12.8) obtemos

$$A(u-u_h) = h^2 t_h$$

Daí, $$\|u-u_h\|_\infty = h^2 \|A^{-1} t_h\|_\infty$$

De (12.7) e $A^{-1} \geq 0$: $\leq h^2 M \|A^{-1} e\|_\infty$ com $e := h^2(1,\ldots,1)^T$.

Pode-se demonstrar que, para todos n, $\|A^{-1}e\| \leq$ const.

Então, para $h \to 0$, u_h converge com ordem 2 para u.

§13 Um problema de contôrno para uma equação diferencial ordinária

Consideremos o problema

$$-y'' + q(x)y = f(x) \quad ; \quad y(a) = \alpha; \; y(b) = \beta \tag{13.1}$$

com $q, f \in C[a,b]$ e $q(x) \geq 0$ em $[a,b]$.

Este problema tem solução única em $[a,b]$.

Seja $h = \frac{b-a}{n+1}$, $x_i = a+hi$, $y_i := y(x_i)$ $(i=0,1,\ldots,n)$.

Substituindo $y''(x_i)$ por $\Delta^2 y_i = h^{-2}(y_{i+1} - 2y_i + y_{i-1})$ cometemos o erro $t_i(h) := y''(x_i) - \Delta^2 y_i$. Se $y \in C^4[a,b]$, obtemos por desenvolvimento de Taylor

$$y_{i\pm1} = y_i \pm hy_i' + \frac{h^2}{2!}y_i'' \pm \frac{h^3}{3!}y_i''' + \frac{h^4}{4!}y^{(4)}(x_i \pm \xi_i^{\pm} h); \quad 0 < \xi_i^{\pm} < 1$$

Assim, $$\Delta^2 y_i = y_i'' + \frac{h^2}{24}\left[y^{(4)}(x_i + \xi_i^+ h) + y^{(4)}(x_i - \xi_i^- h)\right]$$

e da continuidade de $y^{(4)}$, segue para o erro de discretização:

$$t_i(h) := y''(x_i) - \Delta^2 y_i = -\frac{h^2}{12} y^{(4)}(x_i + \xi_i h), \quad |\xi_i| < 1 \tag{13.2}$$

Segue, então, de (13.1) o sistema

$$y_0 = \alpha$$

$$h^{-2}(2y_i - y_{i+1} - y_{i-1}) + q(x_i)y_i = f(x_i) + t_i(y_i), \quad i = 1(1)n$$

$$y_{n+1} = \beta.$$

ou seja $Ay = k + t_h$ (13.3)

$$A := \frac{1}{h^2}\begin{pmatrix} 2+q_1h^2 & -1 & & & \mathcal{O} \\ -1 & 2+q_2h^2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ \mathcal{O} & & & -1 & 2+q_nh^2 \end{pmatrix}$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad ; \quad t_h = \begin{pmatrix} t_1(h) \\ t_2(h) \\ \vdots \\ t_n(h) \end{pmatrix} \quad ; \quad k = \begin{pmatrix} f_1+h^{-2}\alpha \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n+h^{-2}\beta \end{pmatrix}$$

Lema 13.1

Se $q(x) \geq 0$, então A é uma matriz de Stieltjes e irredutível, logo $A^{-1} > 0$.

Prova: O grafo de A é fortemente conexo, logo A é irredutível (teorema 5.1). A é simétrica, então só tem autovalores λ_j reais; λ_j é consequência do teorema 4.3 (de Gerschgorin) e do corolário 4.3. ;A, então, é positiva definida e, assim, matriz de Stieltjes.

Lema 13.2

Vale $A^{-1} \leq A_o^{-1}$ com $A_o = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$.

Prova: $h^2A = D(I-J)$, D : diagonal de A ; $J \geq \mathcal{O}$

$h^2A_o = D_o(I-J_o)$, $D_o = 2I$; $J_o \geq \mathcal{O}$

$\rho(J) < 1$, $\rho(J_o) < 1$ (do teorema 5.4), então as seguintes inversas existem:

$$(I-J_o)^{-1} = I + J_o + J_o^2 + \ldots \geq \mathcal{O}$$

$$(I-J)^{-1} = I + J + J^2 + \ldots \geq \mathcal{O}$$

Segue de $0 \leq J \leq J_o$ e $\mathcal{O} \leq D^{-1} \leq D_o^{-1}$:

$$\mathcal{O} \leq (h^2A)^{-1} = (I-J)^{-1}D^{-1} \leq (I-J_o)^{-1}D_o^{-1} = (h^2A_o)^{-1}. \quad \blacksquare$$

Agora, o método de discretização consiste na substituição de problema (13.1) por

$$A\,y_h = k \quad , \quad y_h = (Y_1, Y_2, \ldots, Y_n)^T \tag{13.4}$$

e no cálculo dos Y_i como aproximações da solução $y(x_i)$ de (13.1).

Devemos provar (compare com (12.9.) que y_h converge para y se $h \to 0$.

Teorema 13.1

Seja $y \in C^4[a,b]$, $|y^{(4)}(x)| \le M$ em $[a,b]$ e $y_h \in \mathbb{R}^n$ solução de (13.4).

Se $q(x) \ge 0$ em $[a,b]$, então vale para $i=1(1)n$: $|y(x_i)-Y_i| \le \frac{Mh^2}{24}(x_i-a)(b-x_i)$

(13.5)

Prova: Segue de (13.3) e (13.1):

$$A(y-y_h) = t_h$$

$$\|y-y_h\|_\infty = \|A^{-1}t_h\|_\infty$$

De $|t_i(h)| \le \frac{Mh^2}{12}$ (veja (13.2)), $A^{-1} > 0$ e do tema 13.2. obtemos

$$\|y-y_h\|_\infty \le \frac{Mh^2}{12}\|A^{-1}e\|_\infty \le \frac{Mh^2}{12}\|A_0^{-1}e\|_\infty \quad ; \quad e := (1,1,\ldots,1)^T .$$

Sendo $w=(w_1,\ldots,w_n)^T$, $w_i = \frac{1}{2}(n+1-i)h^2 = \frac{1}{2}(x_i-a)(b-x_i)$

solução de $A_0 w = e$, segue (13.5).

A estrutura simples do problema (13.1) facilitou a demonstração de convergência que, em casos mais gerais, pode ser bem mais complexa. O procedimento geral, entretanto, sempre é o mesmo:

Substituimos a equação diferencial por equações de diferenças (no caso acima (13.1) por (13.4)) cometendo um "erro de discretização" $t_h = O(h^p)$ $(h \to 0)$.

A ordem p de t_h (que pode ser calculada por desenvolvimento de Taylor) é um critério, usado para estimar-se a precisão dessa substituição; p é chamada de ordem de consistência (com o problema original) do método.

A solução das equações de diferenças é interpretada como aproximação à solução da equação diferencial nos pontos x_i.

Se o método converge para $h \to 0$, a ordem de convergência, frequentemente , é igual à ordem de consistência.

Observação

Na prática, não se resolve o sistema (13.14) por iteração. Sendo A tri-diagonal da forma

$$A = \begin{pmatrix} b_1 & c_1 & & 0 \\ a_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ 0 & & a_n & b_n \end{pmatrix} \quad , \quad k = \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{pmatrix}$$

aplica-se o seguinte algoritmo recursivo baseado na eliminação de Gauss (veja Forsythe-Wasow [8] pg. 104):

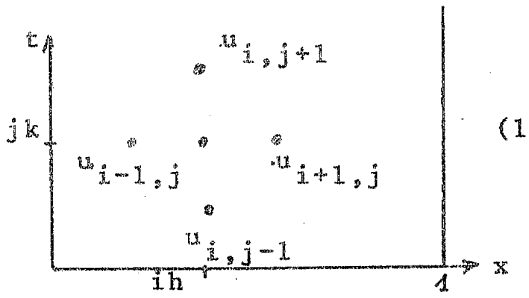
$$\begin{aligned} &v_1 = b_1^{-1} c_1 \quad ; \; v_i = (b_i - a_i v_{i-1})^{-1} c_i && i=2(1)n-1 \\ &w_1 = b_1^{-1} k_1 \quad ; \; w_i = (b_i - a_i v_{i-1})^{-1} (k_i - a_i w_{i-1}) && i=2(1)n \qquad (13.6) \\ &u_n = w_n \quad ; \; u_i = w_i - v_i u_{i+1} && i=(n-1)(1)1. \end{aligned}$$

Este algoritmo necessita menos que 5n multiplicações e/ou divisões e 3n adições, muito menos do que um processo iterativo. Interpretando os elementos a_i, b_i, c_i como sendo matrizes blocos, o algoritmo (13.6) pode ser generalizado para uma matriz tri-diagonal em blocos. A aplicação desta generalização, entretanto, somente é aconselhável em casos particulares como, por exemplo, no caso da matriz (6.6), onde os blocos diagonais são todos iguais e os demais blocos são matrizes identidades. No caso geral, para sistemas grandes, o método exige demais memória no computador. Convém lembrar, neste ponto, que a transferência de dados de uma memória externa poderia consumir mais tempo do que as proprias operações algébricas!

§ 14 - Discretização de uma equação diferencial parcial parabólica.

Consideramos o seguinte problema de transferência de calor:

$$u_t - \sigma u_{xx} = f(x,t,u) \quad ; \ \sigma>0$$
$$u(x,0)=\eta(x) \quad ; \ 0\le x\le 1$$
$$u(0,t)=u(1,t)=0; \ 0\le t\le 1 \qquad (14.1)$$



Fazendo $x_i=ih$, $t_j=jk$, $h=\frac{1}{n_1+1}$, $k=\frac{1}{n_2+1}$,

substituimos u_t resp. u_{xx} por

(a) diferenças progressivas resp. centrais em (x_i,t_j)

(b) diferenças centrais resp. centrais em (x_i,t_j)

(c) diferenças regressivas resp. centrais em (x_i,t_{j+1})

obtendo

$$\text{(a)} \ \frac{1}{k}(U_{i,j+1}-U_{ij}) - \frac{\sigma}{h^2}(U_{i+1,j}-2U_{ij}+U_{i-1,j}) = f_{ij} \quad ;$$

$$\text{(b)} \ \frac{1}{2k}(U_{i,j+1}-U_{i,j-1}) \ \frac{\sigma}{h^2}(U_{i+1,j}-2U_{ij}+U_{i-1,j}) = f_{ij}$$

$$\text{(c)} \ \frac{1}{k}(U_{i,j+1}-U_{ij}) - \frac{\sigma}{h^2}(U_{i+1,j+1}-2U_{i,j+1}+U_{i-1,j+1}) = f_{ij+1}$$

$$f_{ij}=f(x_i,t_j,U_{ij})$$

Observamos que agora temos discretizações com passos differentes nos eixos x e t, caracterizadas pelos parâmetros h e k.

Isto leva aos seguintes métodos:

(a) método explícito

$$U_{i,j+1} = \left(1 - \frac{2\sigma k}{h^2}\right)U_{ij} + \frac{\sigma k}{h^2}(U_{i+1,j} + U_{i-1,j}) + kf_{ij} \qquad (14.2a)$$

$$i=1(1)n_1;\ j=0(1)n_2$$

(b) método de passo 2

$$U_{i,j+1} = U_{i,j-1} + \frac{2\sigma k}{h^2}(U_{i+1,j} - 2U_{ij} + U_{i-1,j}) + 2kf_{ij} \qquad (14.2b)$$

$$i=1(1)n_1;\ j=1(1)n_2$$

(c) método implícito (de Crank-Nicolson)

$$-\frac{\sigma k}{h^2}U_{i+1,j+1} + \left(1 + \frac{2\sigma k}{h^2}\right)U_{i,j+1} - \frac{\sigma k}{h^2}U_{i-1,j+1} = U_{ij} + kf_{ij+1} \qquad (14.2c)$$

$$i=1(1)n_1;\ j=0(1)n_2.$$

Pelo desenvolvimento de Taylor obtemos os seguintes erros de discretização $t_{ij}(h,k)$:

(a) $t_{ij}(h,k) = \frac{k}{2}u_{tt}(x_i,t_j) - \frac{\sigma h^2}{12}u_{xxxx}(x_i,t_j)$ + elementos com ordem maior de que h^2 e k

(b) $t_{ij}(h,k) = \frac{k^2}{6}u_{ttt}(x_i,t_j) - \frac{\sigma h^2}{12}u_{xxxx}(x_i,t_j)$ + elementos com ordem maior do que h^2 e k^2.

(c) $t_{ij}(h,k) = \frac{k}{2}u_{tt}(x_i,t_j) - \sigma k u_{xxt}(x_i,t_j)$ + elementos com ordem maior do que h e k.

Definindo a ordem de consistência $p=\min(q,r)$, sendo q e r, respectivamente as menores potências de h e k em $t_{ij}(h,k)$, obtemos para u suficientemente diferenciável:

O método (14.2a) tem ordem de consistência p=1,

O método (14.2b) tem ordem de consistência p=2,

O método (14.2c) tem ordem de consistência p=1 (e p=2 para $\sigma = \frac{1}{2}$ e $\frac{\partial}{\partial t} f(x,t,u(x,t)) \equiv 0$)

À primeira vista, o método (14.2b) parece ser o melhor; vemos , entretanto, que (14.2b) é um método sem valor prático algum, pois é "instável".

Seja $a := \frac{\sigma k}{h^2}$ e

$$u_o := \begin{pmatrix} \eta(1) \\ \eta(2h) \\ \vdots \\ \eta(n_1 h) \end{pmatrix} \quad ; \quad u_j := \begin{pmatrix} U_{1j} \\ U_{2j} \\ \vdots \\ U_{n_1 j} \end{pmatrix} \quad ; \quad f_j := \begin{pmatrix} f_{1j} \\ f_{2j} \\ \vdots \\ f_{n_1 j} \end{pmatrix}$$

$$A_1 := \begin{pmatrix} (1-2a) & a & & 0 \\ a & (1-2a) & \ddots & \\ & \ddots & \ddots & a \\ 0 & & a & (1-2a) \end{pmatrix} \quad ; \quad C := \begin{pmatrix} -2 & 1 & & 0 \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -2 \end{pmatrix}$$

$$A_3 := \begin{pmatrix} (1+2a) & -a & & 0 \\ -a & (1+2a) & \ddots & \\ & \ddots & \ddots & a \\ 0 & & -a & (1+2a) \end{pmatrix}$$

então os métodos podem ser escritos assim:

(a) $u_{j+1} = A_1 u_j + kf_j \quad , \quad j = 0(1)n_2$ (14.3a)

(b) $u_{j+1} = 2aC\ u_j + u_{j-1} + 2kf_j \quad , \quad j = 1(1)n_2$

ou com $v_j = \begin{pmatrix} u_{j-1} \\ u_j \end{pmatrix}$; $A_2 = \begin{pmatrix} 0 & I \\ I & 2aC \end{pmatrix}$; $g_j = \begin{pmatrix} 0 \\ \vdots \\ f_j \end{pmatrix}$;

$I \in \mathbb{R}(n_1, n_1)$

$v_{j+1} = A_2 v_j + 2kg_j \quad , \quad j = 1(1)n_2$ (14.3b)

(c) $A_3 u_{j+1} = u_j + kf_j \quad , \quad j = 0(1)n_2$

ou seja $u_{j+1} = A_3^{-1} u_j + kA_3^{-1} f_j$. (14.3c)

Temos assim reduzidos os três métodos à forma

$$y_{j+1} = Sy_j + kb_j \quad ; \quad y_o = u_o . \tag{14.4}$$

Seja agora $\bar{y}_j$ solução do método pertubado pelos erros ε_o, então vale para o erro $e_j = \bar{y}_j - y_j$:

$$e_{j+1} = Se_j + k\varepsilon_{j+1} \quad , \quad e_o = \varepsilon_o$$

ou seja $$e_j = Se_o^j + k \sum_{\ell=1}^{j} S^{j-\ell} \varepsilon_\ell \tag{14.5}$$

Obviamente, um método de discretização é aplicável sòmente se o efeito do erro de discretização t_h (ou de outros erros pequenos) na solução, é limitado, qualquer que seja o passo $k \in (o, k_o]$. Tal método é chamado "estável". (14.5) mostra que, no caso dos métodos (14.4), isto sòmente ocorre, se $\|S^j\| \leq$ const. para todos $j \in \mathbb{N}$, ou seja (teorema 1.8(b)), se $\rho(S) \leq 1$.

Definição 14.1

Um método da forma (14.4) é estável para $o<k\leq k_o$, para todo $j\in\mathbb{N}$ se $\|S^j\| \leq$ const. para todo $k\in(o,k_o]$. O método é incondicionalmente estável, se para todo $j\in\mathbb{N}$, $\|S^j\| \leq$ const. para todo $k>o$.

Vejamos, agora, a estabilidade dos métodos (14.2a/b/c)

(a) Os autovalores de A_1 são (teorema 4.5) $\mu_j = (1-2a) + 2a\cos\frac{j\pi}{n_1+1}$, $j=1(1)n_1$, $a>o$; então $\rho(A_1)<1$ para $o<a\leq\frac{1}{2}$; concluimos que o método (14.2a) é estável para $a = \frac{\sigma k}{h^2} \leq \frac{1}{2}$.

Este resultado implica que o passo k no eixo t deve ser consideravelmente menor do que o passo h no eixo x!

(b) Seja $\mu\neq o$ autovalor de A_2, então segue de $\det(A_2-\mu I')=o$ ($I'\in R(2n_1,2n_1)$: matriz identidade) multiplicando (pela esquerda) com $\det\begin{pmatrix} I & \mathcal{O} \\ I & \mu I\end{pmatrix}$ $(-\mu)^{n_1}$

$$\det(2a\mu C + (1-\mu^2)I') = 0$$

$$\det\left(C + \frac{1-\mu^2}{2a\mu}I'\right) = 0$$

Seja λ autovalor de C, então

$$\lambda = \frac{\mu^2-1}{2a\mu} \quad \text{ou seja} \quad \mu = a\lambda + a^2\lambda^2+1$$

Sendo todos os autovalores de C diferentes de zero, existem, para qualquer a, autovalores μ tais que $\mu > 1$.

Assim, o método (14.2/b) é instável para todo h e k.

c) Substituindo a por -a em A_1 obtemos A_3; segue, então, de (a) que todos autovalores de A_3 são maior do que 1.

Consequentemente, $\rho(A_3^{-1})<1$ para todo $a>o$.

Por isso, o método implícita (14.2c) é estável para todo h e k.

Observação

Substituindo em (14.2b) o valor $2U_{ij}$ por $(U_{i,j-1}+U_{i,j+1})$ obtem-se o método de DuFort - Frankel, que é estável para todo h e k.

§ 15 - Problemas elípticos mais gerais

O tratamento do problema de Dirichlet em §12 foi bastante simplificado pela forma simples da equação diferencial (6.3) e, ainda mais, pelo fato que os valores de contôrno foram definidas no limite de um quadrado.
Procuramos, agora, extender o método do §12 a problemas e contôrnos mais gerais, escolhidos entretanto, tais que o sistema linear, obtido pela discretização, ainda tem uma matriz irredutível de Stieltjes.

15.1 - O Problema

Seja $S\in\mathbb{R}^2$ um conjunto aberto, limitado e conexo com o limite ∂S. Consideremos em S a equação

$$-(q(x,y)u_x)_x - (q(x,y)u_y)_y + \sigma(x,y)u(x,y) = f(x,y) \qquad (15.1a)$$

com a condição de contôrno

$$\alpha(x,y)u+\beta(x,y)\frac{\partial u}{\partial n} = \gamma(x,y) \quad , \quad (x,y)\in\partial S \qquad (15.1b)$$

(sendo $\frac{\partial u}{\partial n}$ a derivada de u na direção da normal[1] externa de ∂S).

sejam $q(x,y)$ e $\sigma(x,y)$ contínuos em $\bar{S}:= S\cup\partial S$ e $q(x,y)>o$, $\sigma(x,y)>o$; (15.2)

Sejam α,β,γ secionalmente contínuas em ∂S com

$$\alpha(x,y)\geq o, \ \beta(x,y)\geq o \quad , \quad \alpha+\beta>o \qquad (15.3)$$

Para $\alpha\equiv 1$, $\beta\equiv o$ temos novamente o "problema de Dirichlet", para $\alpha\equiv o$, $\beta\equiv 1$ o "problema de Neumann".

---

1) lembramos que $\frac{\partial}{\partial n}u(x,y) = u_x\dot{y}-u_y\dot{x}$ para ∂S dado por $x=x(t),y=y(t)$ (desde que a direção em que t cresce seja a correspondente à descrição positiva do limite)

15.2 - Os pontos de discretização

Escolhendo pontos P_ℓ em ∂S e ligando-los por retas, substituimos o limite ∂S por um polígono ∂S_h. A rede obtida por paralelos aos eixos, passando pelo P_ℓ, define os pontos (x_i, y_j) com $x_0<x_1<\ldots<x_n, y_0<y_1<\ldots<y_m$ que, se $(x_i y_j)\in \bar{S}$, são chamados de pontos de discretização (veja fig. 15.1) e denominados de P_{ij}.

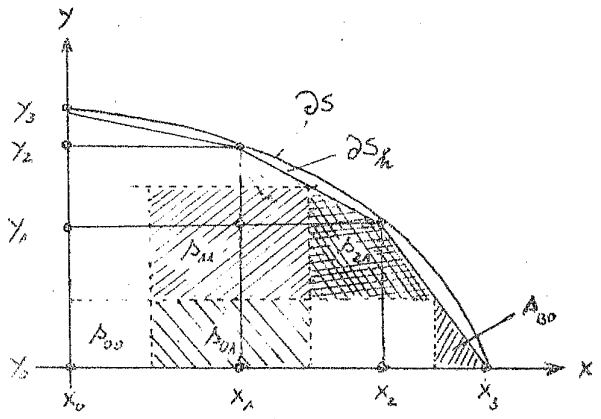


fig. 15.1

Observe-se que os x_i e y_j não necessariamente são equidistantes.

Definimos $h_i := x_{i+1} - x_i$, $k_j := y_{j+1} - y_j$ e associamos a cada P_{ij} a região $s_{ij}\in\bar{S}$, limitado pelas retas $x=x_i-\frac{1}{2}h_{i-1}$, $x=x_i+\frac{1}{2}h_i$, $y=y_j-\frac{1}{2}k_{j-1}$, $y=y_j+\frac{1}{2}k_j$ e, caso que são em ∂S, pelo polígono ∂S_h (veja fig. 15.1).

15.3 - A discretização da equação parcial

Para cada P_{ij}, onde $u_{ij} := u(x_i, y_j)$ é desconhecido, integramos a equação (15.1a) sobre as regiões s_{ij} correspondentes:

$$-\iint_{s_{ij}} \{(qu_x)_x + (qu_y)_y\}\, dx\,dy + \iint_{s_{ij}} \sigma u\, dx\,dy = \iint_{s_{ij}} f\, dx\,dy$$

Do teorema de Green[1] obtemos para funções diferenciaveis quaisquer

1) veja e.g. R. Courant: Cálculo Diferencial e Integral, Vol. II, pg. 366, Editora Globo, Rio de Janeiro, 1955

$g(x,y)$ e $h(x,y)$ em s_{ij}:

$$\iint_{s_{ij}} (g_x - h_y)\, dxdy = \int_{\partial s_{ij}} (hdx+gdy)$$

sendo ∂s_{ij} o limite de s_{ij} e a integral de linha tomada no sentido positivo. Com $g:=qu_x$ e $h:=qu_y$ obtemos

$$-\int_{\partial s_{ij}} (qu_x dy - qu_y dx) + \iint_{s_{ij}} \sigma u dxdy = \iint_{s_{ij}} f dxdy \,. \qquad (15.4)$$

As duas integrais duplas são aproximadas conforme

$$\iint_{s_{ij}} g dxdy \approx g_{ij} a_{ij} \qquad \text{com } g=\sigma u \text{ ou } g=f, \text{ resp.,} \qquad (15.5)$$

sendo a_{ij} a área de s_{ij} $\left(a_{ij} = \frac{1}{4}(h_i+h_{i-1})(k_i+k_{i-1})\right.$ se s_{ij} é um retângulo$\left.\right)$.

Vejamos agora a aproximação da integral de linha I

Caso (a): ∂s_{ij} é retângular (veja fig. 15.2), então

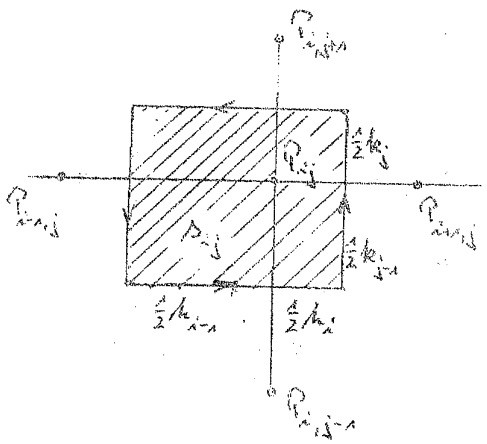


fig. 15.2

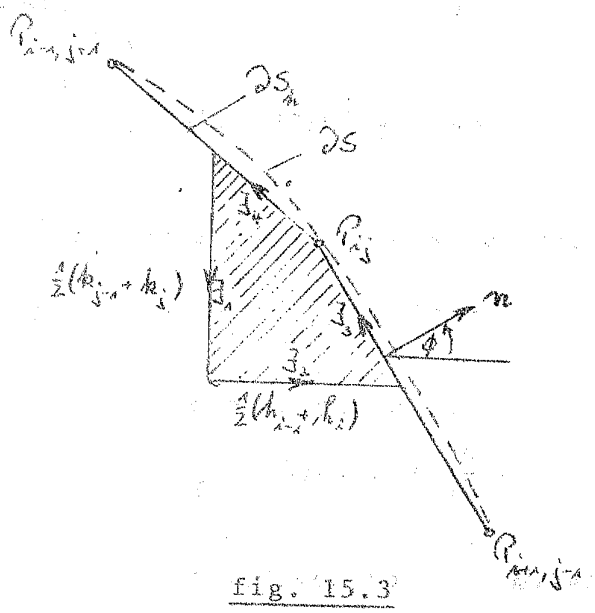


fig. 15.3

os lados tem o comprimento $\frac{1}{2}(h_i+h_{i-1})$ e $\frac{1}{2}(k_i+k_{i-1})$ e a integral de linha sobre os quatros lados é aproximado usando diferenças centrais:

$$I = -\int_{\partial s_{ij}} (qu_x dy - qu_y dx) \simeq \frac{1}{2}(k_j+k_{j-1})\left\{q_{i+\frac{1}{2},j}\left(\frac{u_{ij}-u_{i+1,j}}{h_i}\right) + q_{i-\frac{1}{2},j}\left(\frac{u_{ij}-u_{i-1,j}}{h_{i-1}}\right)\right\} + \frac{1}{2}(h_i+h_{i-1})\left\{q_{i,j+\frac{1}{2}}\left(\frac{u_{ij}-u_{i,j+1}}{k_j}\right) + q_{i,j-\frac{1}{2}}\left(\frac{u_{ij}-u_{i,j-1}}{k_{j-1}}\right)\right\} \tag{15.6}$$

com $q_{i+\frac{1}{2},j} := q(x_i+h_i/2, y_j)$.

Caso (b): Se ∂s_{ij} não é retangular (veja s_{ij} em fig. 15.1 e fig. 15.3) com $u_{ij} \in \partial s_{ij}$ desconhecido (isto é $\beta_{ij} \neq 0$ em (15.1b)), então dividimos a integral de linha I em partes I_j (j=1,2,3,4) associadas aos lados do poligono: $I = I_1+I_2+I_3+I_4$ (veja fig. 15.3).

I_1 e I_2 são aproximadas como no caso (a), ou seja

$$I_1 \simeq \frac{1}{2}(k_j+k_{j-1}) q_{i-\frac{1}{2},j} \left(\frac{u_{ij}-u_{i-1,j}}{h_{i-1}}\right)$$

$$I_2 \simeq \frac{1}{2}(h_i+h_{i-1}) q_{i,j-\frac{1}{2}} \left(\frac{u_{ij}-u_{i,j-1}}{k_{j-1}}\right).$$

Para aproximar I_3 representamos o trecho de ∂S_h associado a I_3 por

$$x(t) = x_{i+\frac{1}{2}} - t\,\mathrm{sen}\phi \quad ; \quad y(t) = y_{i-\frac{1}{2}} + t\cos\phi$$

Sendo ϕ o angulo do normal com o eixo positivo de x(veja fig. 15.3).

Então, $\frac{\partial u}{\partial n} = u_x \cos\phi + u_y \mathrm{sen}\phi \quad , \quad (x,y) \in \partial S_h$.

Assim obtemos com $\ell_1 := \ell_{ij}^{(1)} = \frac{1}{2}\sqrt{h_i^2+k_{j-1}^2}$

$$I_3 = -\int_{\partial s_{ij}} (qu_x dy - qu_y dx) = -\int_0^{\ell_1} (qu_x \cos\phi + qu_y \operatorname{sen}\phi)\, dt =$$

$$= -\int_0^{\ell_1} q \frac{\partial u}{\partial n}\, dt = -\int_0^{\ell_1} q \left(\frac{\gamma(t)-\alpha(t)u(t)}{\beta(t)}\right) dt$$

$$\simeq -q_{ij}\beta_{ij}^{-1}(\gamma_{ij}-\alpha_{ij}u_{ij})\ell_{ij}^{(1)}.$$

I_4 é aproximada na mesma maneira:

$$I_4 \simeq -q_{ij}\beta_{ij}^{-1}(\gamma_{ij} - \alpha_{ij}u_{ij})\ell_{ij}^{(2)} \quad \text{com } \ell_{ij}^{(2)} = \frac{1}{2}\sqrt{h_{i-1}^2+k_j^2}.$$

Temos assim para cada P_{ij} em que u_{ij} é desconhecido, uma equação da forma

$$D_{ij}u_{ij} - L_{ij}u_{i-1,j} - R_{ij}u_{i+1,j} - T_{ij}u_{i,j+1} - B_{ij}u_{i,j-1} =$$

$$= b_{ij} + \frac{1}{4}(h_{i-1}+h_i)(k_{j-1}+k_j)t_{ij}. \qquad (15.7)$$

(t_{ij}: êrros de discretização)

Esta equação interrelaciona (no máximo) 5 incógnitas,

u_{ij}, $u_{i-1,j}$, $u_{i+1,j}$, $u_{i,j-1}$ e $u_{i,j+1}$

(a) Para $P_{ij} \in S$, os seus coeficientes são obtidos de (15.5/6):

$$L_{ij} = q_{i-\frac{1}{2},j}\left(\frac{k_{j-1}+k_j}{2h_{i-1}}\right) \; ; \qquad R_{ij} = q_{i+\frac{1}{2},j}\left(\frac{k_{j-1}+k_j}{2h_i}\right)$$

$$T_{ij} = q_{i,j+\frac{1}{2}}\left(\frac{h_{i-1}+h_i}{2k_j}\right) \; ; \qquad B_{ij} = q_{i,j-\frac{1}{2}}\left(\frac{h_{i-1}+h_i}{2k_{j-1}}\right) \qquad (15.8a)$$

$$D_{ij} = L_{ij}+R_{ij}+T_{ij}+B_{ij} + \frac{1}{4}\sigma_{ij}(h_{i-1}+h_i)(k_{j-1}+k_j)$$

$$b_{ij} = \frac{1}{4} f_{ij}(h_{i-1}+h_i)(k_{j-1}+k_j).$$

(b) Para $P_{ij} \in \partial S_h$ com $\beta_{ij} \neq 0$ vale, sendo a_{ij} a area de s_{ij}:

L_{ij} e B_{ij} como acima; $R_{ij} = T_{ij} = 0$

$$D_{ij} = L_{ij} + B_{ij} + \sigma_{ij}\, a_{ij} + \alpha_{ij}\beta_{ij}^{-1} q_{ij}(\ell_{ij}^{(1)}+\ell_{ij}^{(2)})$$

$$b_{ij} = f_{ij}\, a_{ij} + \gamma_{ij}\beta_{ij}^{-1} q_{ij}(\ell_{ij}^{(1)}+\ell_{ij}^{(2)}) \qquad (15.8b)$$

(15.7) tem a forma $Au=b+t_h$, donde obtemos o sistema $Au_h=b$, (15.9)

para a aproximação u_h de u nos pontos P_{ij}.

15.4 A estrutura da matriz A

0 sistema (15.9) contém, no caso da fig. 15.1, 10 incógnitas[1] : U_{03}, U_{02}, U_{12}, U_{01}, U_{11}, U_{21}, U_{00}, U_{10}, U_{20}, U_{30} que sejam, nessa ordem, os componentes u_i, da solução u_h. Os pontos P_i associados aos u_i assim ordenados são indicados na figura 15.4. (compare com fig. 15.1).

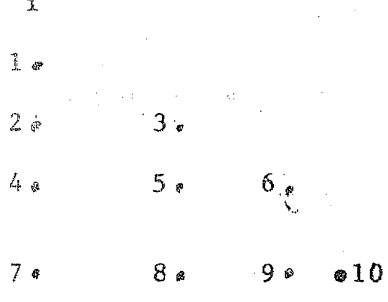


Fig. 15.4

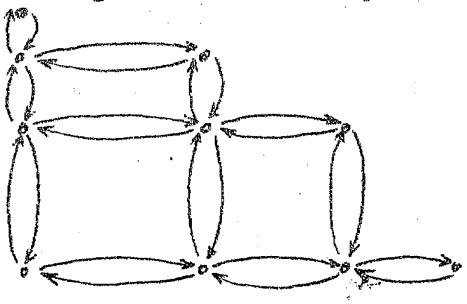


Fig. 15.5

Pelo sistema (15.9) cada ponto P_i é relacionado, no máximo, a 4 outros pontos. Indicamos este relacionamento por caminhos orientados de P_i para estes pontos. Obtemos, assim, o grafo G da figura 15.5., que geométricamente corresponde à figura 15.1. Obviamente, G é o grafo orientado da matriz J de Jacobi de A que, consequentemente, tem a forma indicada no lado.

Vemos que J é irredutível e, com o bloqueamento indicado, uma matriz (1,1)-banda; Assim, J_B é consistentemente (1,1)-ordenada (veja teorema 11.2). Observe-se que o bloqueamento corresponde ao número de pontos nas linhas horizontais em fig. 15.1!

$$J=\left(\begin{array}{c|cc|ccc|cccc}
0 & x & 0 & & & & & & & \\
\hline
x & 0 & x & x & 0 & 0 & & & & \\
0 & x & 0 & 0 & x & 0 & & & & \\
\hline
 & x & 0 & 0 & x & 0 & x & 0 & 0 & 0 \\
 & 0 & x & x & 0 & x & 0 & x & 0 & 0 \\
 & 0 & 0 & 0 & x & 0 & 0 & 0 & x & 0 \\
\hline
 & & & x & 0 & 0 & 0 & x & 0 & 0 \\
 & & & 0 & x & 0 & x & 0 & x & 0 \\
 & & & 0 & 0 & x & 0 & x & 0 & x \\
 & & & 0 & 0 & 0 & 0 & 0 & x & 0
\end{array}\right) \qquad (15.10)$$

1) Assumindo que $\beta \neq 0$ nos eixos

De (15.7) com (15.8 a/b) segue que A é simétrica e de diagonal dominante para $\sigma,q,\alpha,\beta > 0$; assim, A é positiva definida. Os elementos diagonais são positivos e os demais elementos são negativos. Segue o

Teorema 15.1

A é uma matriz de Stieltjes; e sua matriz de Jacobi J é não-negativa, irredutível, com autovalores reais e $\rho(J) < 1$. Na ordenação indicada em fig. 15.4, J_B é consistentemente (1,1)-ordenada.

Isto nos permite aplicar toda a teoria da sobre-relaxação, apresentada no capítulo II, na discretização (15.9) do problema (15.1a/b)

Observe-se que chegávamos ao grafo de J e, assim, à estrutura de A simplesmente pela topologia da fig. 15.1. que levou também, numa maneira bastante fácil, à ordenação consistente e até a um bloqueamento de J que poderia ser usado na aplicação do método bloco de sobre-relaxação (levando à sobre-relaxação "por linhas").

CAPÍTULO IV

ACELERAÇÃO DA CONVERGÊNCIA DA SOBRE-RELAXAÇÃO

A seguir, vamos indicar um algoritmo de relaxação a dois parâmetros, que contém como casos particulares, os três métodos clássicos (Jacobi, Gauss-Seidel e sobrerelaxação) e que, em termos de esforços computacionais, não difere muito da sobrerelaxação.

Tal algoritmo, tem dentre outras, a vantagem de acelerar de uma maneira simples a convergência da sobrerelaxação, fato particularmente relevante em casos onde não é possível estimar teoricamente o melhor parâmetro da sobrerelaxação, por exemplo, quando a matriz J de Jacobi não é (r,q) consistentemente ordenada; também, através deste algoritmo, podemos obter convergência em situações, mesmo onde os três métodos clássicos divergem.

§16. Uma relaxação a dois parâmetros

Considerando o método da sobrerelaxação na forma

$$x^{j+1} = (I - \omega L)^{-1}(\omega U + (1-\omega)I)x^j + c$$

$$= R_\omega x^j + c \qquad (16.1)$$

e fazendo a seguinte média ponderada entre x^j e x^{j+1} obtida por (16.1)

$$x^{j+1} := \frac{\gamma}{\omega}(R_\omega x^j + c) + (1 - \frac{\gamma}{\omega})x^j \quad , \quad \omega \neq 0 \qquad (16.2)$$

obtemos a iteração

$$x^{j+1} = R_\omega(\gamma)x^j + d$$

com

$$R_\omega(\gamma) = (I-\omega L)^{-1}((\gamma-\omega)L+\gamma U + (1-\gamma)I).$$

Esta iteração, contém os três métodos clássicos como casos particulares.

$\omega = \gamma \quad \Longrightarrow R_\omega(\omega) = R_\omega$, Sobrerelaxação

$\omega = \gamma = 1 \quad \Longrightarrow R_1(1) = R_1$, Gauss-Seidel

$\omega = 0;\ \gamma = 1 \Longrightarrow R_o(1) = 0$, Jacobi

Além destes, obtemos um método de Jacobi com um parâmetro γ, definido pelo caso $\omega = 0$, levando à matriz de iteração

$$R_o(\gamma) = J_\gamma = \gamma\ J + (1-\gamma)I$$

Exemplo: Considere o sistema

$$\begin{aligned} x_1 - x_2 &= b_1 \\ -x_1 + 2x_2 + 10x_3 &= b_2 \\ - x_2 + 2x_3 - 2x_4 &= b_3 \\ 2x_3 - x_4 &= b_4, \ b_i \text{ real } i = 1(1)4 \end{aligned}$$

Neste caso,

$$J = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/2 & 0 & -5 & 0 \\ 0 & 1/2 & 0 & 1 \\ 0 & 0 & 2 & 0 \end{pmatrix}$$

é (1,1) consistentemente ordenada e tem como auto valores $\mu_{1,2,3,4} = \pm\sqrt{\pm i}$; logo $\rho(J) = 1$ e pela relação (11.6) com q=r=1, tem-se $\rho(R_1) = 1$ e portanto, os métodos de Jacobi e Gauss-Seidel divergem, bem como a iteração com R_ω.

A iteração com Gauss-Seidel é

$$
\begin{aligned}
x_1^{j+1} &= & x_2^j & & + c_1\\
x_2^{j+1} &= \frac{1}{2}\, x_1^{j+1} & & -5x_3^j & + c_2\\
x_3^{j+1} &= & \frac{1}{2}\, x_2^{j+1} & + x_4^j & + c_3\\
x_4^{j+1} &= & & +2x_3^{j+1} & + c_4
\end{aligned}
$$

com a matriz de iteração $R_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1/2 & -5 & 0 \\ 0 & 1/4 & -5/2 & 1 \\ 0 & 1/2 & -5 & 2 \end{pmatrix}$

Com $\omega = 1$ e $\gamma = \frac{1}{2}$, obtemos de (16.2), o sistema

$$
\begin{aligned}
x_1^{j+1} &= \frac{1}{2}\, x_1^j + \frac{1}{2}\, x_2^j & & + d_1\\
x_2^{j+1} &= & \frac{3}{4}\, x_2^j - \frac{5}{2}\, x_3^j & + d_2\\
x_3^{j+1} &= & \frac{1}{6}\, x_2^j - \frac{3}{4}\, x_3^j + \frac{1}{2}\, x_4^j & + d_3\\
x_4^{j+1} &= & \frac{1}{4}\, x_2^j - \frac{5}{2}\, x_3^j + \frac{3}{2}\, x_4^j & + d_4
\end{aligned}
$$

cuja matriz de iteração é

$$
R_1(1/2) = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 3/4 & -5/2 & 0 \\ 0 & 1/8 & -3/4 & 1/2 \\ 0 & 1/4 & -5/2 & 3/2 \end{pmatrix}
$$

Verifica-se que $\rho(R_1(1/2)) = \frac{1}{2}\sqrt{2} < 1$ e assim, a iteração com o novo método converge.

§17. Estudo comparativo da convergência

Neste parágrafo, serão demonstrados teoremas de comparação entre o novo método e o método de sobrerelaxação, os quais, estão baseados numa relação entre os auto valores das matrizes associadas.

Teorema 17.1

Sejam $\omega \neq 0$, λ auto valor de R_ω e β dado por

$$\beta = \frac{\gamma}{\omega}\lambda + (1 - \frac{\gamma}{\omega}). \qquad (17.1)$$

Então, β é auto valor de $R_\omega(\gamma)$ e reciprocamente, se β é auto valor de $R_\omega(\gamma)$ e λ satisfaz (17.1), então, λ é auto valor de R_ω.

Prova Basta observar que $\frac{\gamma}{\omega} R_\omega + (1 - \frac{\gamma}{\omega})I = R_\omega(\gamma)$

Este resultado, permitirá discutir casos, nos quais o novo método é melhor que a sobrerelaxação.

Teorema 17.2

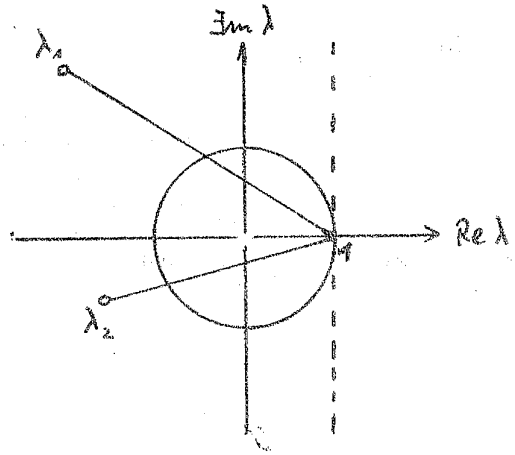
Se a iteração com a matriz R_ω é divergente, isto é, $\rho(R_\omega) = r \geq 1$, então, existe $\gamma_o = \gamma_o(\omega)$ com $|\gamma_o| < \omega$ tal que o método (16.2) converge, se e somente se todos os auto valores γ_j de R_ω satisfazem

a) $\mathrm{Re}\,(\gamma_j) < 1$

ou se todos satisfazem

b) $\mathrm{Re}\,(\gamma_j) > 1$.

A demonstração baseia-se no fato de que a aplicação $\beta: C \Longrightarrow C$ dada por (16.3), tem o valor $\lambda = 1$ como ponto fixo e leva retas passando por $\lambda = 1$ em si mesmas.



Com as hipóteses (a) ou (b), qualquer auto valor λ de R_ω, pode ser levado no interior do círculo unitário, escolhendo-se γ devidamente, sendo positivo no caso (a) e negativo no caso (b).

De maneira análoga, obtemos de (17.1) a seguinte condição suficiente para a convergência do método (16.2).

Teorema 17.3

Se a iteração com a matriz R_ω é convergente, isto é $\rho(R_\omega) = r < 1$ e se todos os auto valores λ_j de R_ω satisfazem

a) $\mathrm{Re}\,(\lambda_j) < r^2$

ou se todos satisfazem

b) $\mathrm{Re}\,(\lambda_j) > r^2$.

Então, existe $\gamma_o = \gamma_o(\omega)$ tal que $\rho(R_\omega(\gamma_o)) < r$ e portanto, o método (16.2) converge assintóticamente mais rápido que a sobrerelaxação.

De um modo geral, não conhecemos o comportamento dos auto valores da matriz de relaxação R_ω e assim, em particular, as hipóteses dos teoremas 17.2 e 17.3 não são facilmente veri-

ficadas e então, como acontece com a iteração com R_ω fica - camos também sem condições de saber se a iteração com $R_\omega(\gamma)$ é ou não convergente.

Mostra-se que no caso particular em que a matriz J de Jacobi é (1,1) consistentemente ordenada, existe uma relação entre os auto valores de $R_\omega(\gamma)$ e os auto valores de J. Assim, em vez de estudar a localização dos auto valores de R_ω (como é feita nos teoremas 17.2 e 17.3), precisamos sòmente de um conhecimento dos auto valores de J.

Desse modo, procederemos de maneira similar à de Young, quando ele estuda em [12] a determinação do melhor parâmetro ω_o da sobrerelaxação R_ω.

§18. O caso das matrizes J consistentemente ordenadas

Para matrizes (r,q) consistentemente ordenadas, vale o seguinte resultado, originalmente devido a Young [12] (caso q=r=1) e generalizado por Verner e Bernal [11]:

Teorema

Seja a matriz J de Jacobi (r,q) consistentemente ordenada

a) Se μ é auto valor de J e λ satisfaz

$$(\lambda + \omega - 1)^p = \omega^p \mu^p \lambda^q, \quad p = q + r \qquad (18.1)$$

então, λ é auto valor de R_ω.

b) Se λ é auto valor de R_ω e μ satisfaz (18.1) então μ é auto valor da matriz J.

Devido à relação (17.1) entre os auto valores β de $R_\omega(\gamma)$ e λ de R_ω, segue-se o seguinte

Teorema 18.1

Seja a matriz J de Jacobi (r,q) consistentemente ordenada.

a) Se μ é auto valor de J e βsatisfaz

$$(\beta + \gamma - 1)^p = \gamma^r \mu^p (\omega \beta + \gamma - \omega)^q, \quad p = q + r \qquad (18.2)$$

então, β é auto valor de $R_\omega(\gamma)$

b) Se β é auto valor de $R_\omega(\gamma)$ e μ satisfaz (18.2), então μ é auto valor de J.

Através da relação (18.2), estabelecem-se critérios de comparação entre a iteração de Jacobi e o novo método.

Teorema 18.2

Se a matriz J de Jacobi for (1,1) consistentemente ordenada e se todos os seus auto valores $\mu = a + i\,b$ satisfazem

$$a^2 - \frac{b^2}{B^2} < 1, \quad B \geq 1 \qquad (18.3)$$

então, para cada ω satisfazendo $1 \geq \omega \geq \frac{2}{1+B}$, existe $\gamma=\gamma(\omega)$, $0 < \gamma < \omega$, tal que a iteração com $R_\omega(\gamma)$ é convergente.

Prova: Foi demonstrado em [7] pag 23 que, com a hipótese (18.3), vale $Re(\lambda_j) < 1$, para todos os auto valores λ_j de R_ω, $1 \geq \omega \geq \frac{2}{1+B}$. Daí, a demonstração segue-se dos teoremas 17.2 e 17.3 .

No próximo teorema, damos condições suficientes para que a convergência assintótica do novo algorítmo seja mais rápida do que a da sobrerelaxação.

Teorema 18.3

Seja J (1,1) consistentemente ordenada e suponhamos que todos os seus auto valores $\mu = a + ib$ satisfazem

$$| a | \geq | b | \tag{18.4}$$

Então, a iteração com $R_\omega(\gamma)$, $\omega \geq 1$, $0 < \gamma < \omega$, converge e mais, se para algum $\omega \leq 1$, a sobrerelaxação com R_ω é convergente, existe $\gamma = \gamma(\omega)$, $0 < \gamma < \omega$ tal que a iteração com $R_\omega(\gamma)$ converge assintóticamente mais rápido que aquela com R_ω.

Prova: Segue-se de (18.1), para r = q = 1, a relação

$$\mu = \frac{1}{\omega}\left(\lambda^{1/2} + \frac{\omega-1}{\lambda^{1/2}}\right),$$

da qual concluímos que $Re(\lambda) \leq 0$ se $\mu \in S$, onde

$$S = \left\{ \mu = a + ib;\ | a | \leq | b | \right\}.$$

Sendo $Re(\lambda) \leq 0$, em particular temos $Re(\lambda) < 1$ e utilizando os teoremas 17.2 e 17.3 completaremos a demonstração.

Quando a matriz J de Jacobi, além de ser (1,1) consistentemente ordenada possuir todos os auto valores reais, a convergência com o novo método a exemplo do que se dá com a sobrerelaxação, só ocorre se tivermos convergência com o método de Jacobi.

Teorema 18.4

Se J é (1,1) consistentemente ordenada e tem todos os seus auto valores reais, então, existe um par (ω,γ) onde $\rho(R_\omega(\gamma)) < 1$ se e sòmente se $\rho(J) < 1$.

Prova: Segue-se de um estudo da equação do 2º grau obtida da relação (18.2), tomando $r = q = 1$.

Se $\rho(J) < 1$, com hipóteses adicionais aos auto valores da matriz J, podemos através do novo método, ter convergência assintòtica mais rápida que a sobrerelaxação, mesmo tomando a iteração com R_{ω_o} onde ω_o é o melhor parâmetro para a sobrerelaxação.

Teorema 18.5

Seja J (1,1) consistentemente ordenada e suponhamos que todos os seus auto valores são reais com

$$0 < \mu_1 := \min_k |\mu_k| \; ; \; \mu_p := \max_k |\mu_k| < 1.$$

Se $$\left(1 - \mu_1^2\right) < \left(1 - \mu_p^2\right)^{1/2} \quad \text{e} \quad \omega_o \leq \omega < \frac{2}{2 - \mu_1^2}$$

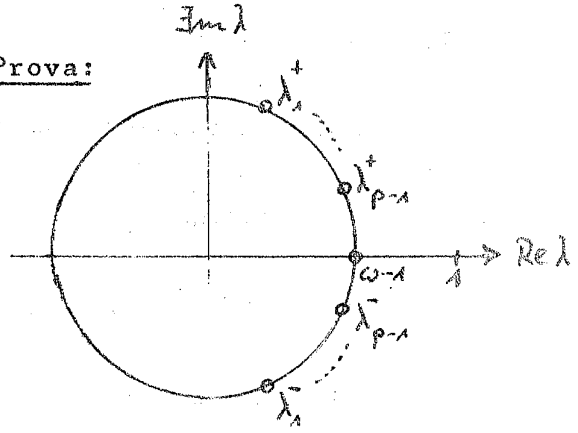
então

a) Para $\tilde{\gamma} = \dfrac{2 - \omega}{1 - \mu_1^2}$ e $\omega < \gamma < \tilde{\gamma}$, temos

$$\rho(R_\omega(\gamma)) < \rho(R_\omega)$$

b) para $\gamma_o = \frac{1}{2}\left(\frac{2-\mu_1^2\,\omega}{1-\mu_1^2}\right)$, (ou seja $\gamma_o = \frac{1}{2}(\omega + \hat{\gamma})$),

temos $\rho(R_\omega(\gamma_o)) < \rho(R_\omega(\gamma))$, γ real, $\gamma \neq \gamma_o$.



Prova: Para tais valores de ω, os auto valores de R_ω são complexos conjugados e se comportam conforme a figura ao lado.

Utilizando então a relação (17.1), poderemos concluir a demonstração.

BIBLIOGRAFIA

/1/ Broyden, C.G.: Some generalizations of the theory of sucæssive overrelaxation, Num. Math. 6,1964, p. 269-284

/2/ Collatz, L.: Functional analysis and numerical mathematics, Academic Press, 1964

/3/ Faddeev e Faddeeva: Computational methods of linear algebra, Freeman 1963

/4/ Forthythe, G.E. e Wasow, W.: Finite difference methods for partial differential equations, Wiley 1960

/5/ Gantmacher, F.R.: The theory of matrices, Vol. I,II Chelsea 1960

/6/ Householder, A.S.: The theory of matrices in numerical analysis, Blaisdell 1965

/7/ Klein, M.P.: Uma relaxação a dois parametros, Tesis de doutoramento IMPA 1976

/8/ Lancaster, P.: Theory of matrices, Academic Press 1969

/9/ Nichols, N.K. e Fox,L.: Generalized consistent ordering and the optimal successive overrelaxation factor Num. Math. 13,1969, p.425-433

/10/ Varga, P.S.: Matrix iterative analysis, Prentice-Hall 1962

/11/ Verner, J.H. e Bernal, M.J.M.: On generalizations of the theory of consistent orderings for successive overrelaxation methods, Num. Math. 2,1968, p.215-222

/12/ Young, D.M.: Iterative solution of large linear systems, Academic Press 1971