SOLUÇÃO DE UM SISTEMA LINEAR MAL-CONDICIONADO
ASSOCIADO À EQUAÇÃO DO CALOR

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pelo Sr. TOMÁS HUMBERTO DÍAZ VALENCIA e aprovada pela Comissão Julgadora.

Campinas, 8 de janeiro de 1992

Profa. Dra. Vera Lucia da Rocha Lopes
VERA LÚCIA DA ROCHA LOPES
Orientadora

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de MESTRE em Metemática Aplicada.

SOLUÇÃO DE UM SISTEMA LINEAR MAL-CONDICIONADO ASSOCIADO À EQUAÇÃO DO CALOR

TOMÁS HUMBERTO DÍAZ VALENCIA

IMECC - UNICAMP

1992

AGRADECIMENTOS

À Profa. Dra. Vera Lúcia da Rocha Lopes pelas muitas sugestões e correções feitas no meu trabalho e acima de tudo pela confiança brindada nas horas de consulta.

Ao Prof. Dr. José Vitório Zago por me emprestar a sua área do Vax para as práticas no laboratório e pela assistência na parte numérica da tese.

Ao Prof. Dr. João Frederico da Costa A. Meyer e Dorival I. de Oliveira pelas facilidades brindadas no uso do Laboratório de Matemática Aplicada do IMECC.

À Profa. Dra. Márcia Aparecida Gomes Ruggiero pelas valiosas aulas de Métodos Computacionais em Algebra Linear.

Aos professores do Departamento de Matemática Aplicada, sem exceção, que contribuiram na minha formação profissional.

Ao pessoal da secretaria de Pós-Graduação do IMECC pela constante ajuda na minha condição de estrangeiro.

Á Angles de Fatima T. Espindola secretária do Departamento de Matemática Aplicada pela paciência e vontade na datilografia da tese.

Ao Fermín e minha turma maravilhosa: Luiz, Lilian, Diomar, Bete, Gilli, Eduardo, Gustavo, Marcílio, Marco, Didi e Sandra pelas alegrias e tristezas compartilhadas durante o período de 1989 até 1991.

A todas aquelas pessoas anônimas que fizeram possível minha convivência universitária.

Às Instituições CAPES, UNICAMP E UNICAUCA pelo suporte financeiro.

NO HAY OLVIDO

(Poema)

(...)

Si me preguntáis de dónde vengo,
tengo que conversar con cosas rotas,
con utensilios demasiado amargos,
con grandes bestias a menudo podridas
y con mi acongojado corazón.

No son recuerdos los que se han cruzado
ni es la paloma amarillenta que duerme en el olvido,
sino caras con lágrimas,
dedos en la garganta,
y lo que se desploma de las hojas:
la oscuridad de un día transcurrido,
de un día alimentado con nuestra triste sangre.

(...)

Pero no penetremos más allá de esos dientes,
no mordamos las cáscaras que el silencio acumula,
porque no sé qué contestar:
hay tantos, muertos,
y tantos malecones que el sol rojo partía,
y tantas cabezas que golpean los buques,
y tantas manos que han encerrado besos,
y tantas cosas que quiero olvidar.

Pablo Neruda (1904-1973)

NÃO HÁ ESQUECIMENTO

(Poema)

(...)

Se me perguntar de onde venho,
terei que falar com coisas rompidas,
com utensílios demasiados amargos,
com grandes bestas freqüentemente podres
e com meu aflito coração.

Não são recordações as que se cruzaram
nem é a pomba amarelenta que dorme no esquecimento
senão caras com lágrimas,
dedos na garganta,
e o que cai das folhas:
a escuridão de um dia transcorrido
de um dia alimentado com nosso triste sangue.
(...)
Mas não penetremos além desses dentes,
não mordamos as cascas que o silêncio acumula,
porque não sei que responder:
há tantos mortos,
e tantos diques que o sol vermelho partia
e tantas cabeças que batem os navios,
e tantas mãos que encerraram beijos
e tantas coisas que quero esquecer.

Pablo Neruda (1904-1973)

ÍNDICE

INTRODUÇÃO

O estudo que vamos fazer se propõe a dar solução ao sistema linear que resultou da discretização, por Elementos Finitos, de princípios extremos duais aplicados à equação do calor $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$ com condições de fronteira $u(0, t) = u(1, t) = 0$ e condição inicial $u(x, 0) = u_0(x)$.

Tal sistema é muito mal-condicionado e uma primeira tentativa de solução pelo método dos gradientes conjugados sem pré-condicionamento levou-nos a resultados pessimistas.

O sucesso dos métodos aqui expostos, que dão solução ao sistema, é devido a três sugestões chaves que gostaríamos de salientar:

1) Aproveitamento da esparsidade da matriz;
2) Normalização da matriz e emprego de precisão dupla;
3) Uso das decomposições de Choleski incompletas como pré-condicionadoras no método dos gradientes conjugados.

A primeira foi fornecida por Vera Lopes. Baseados nesta fizemos uma rotina que chamamos de BANDA 7 que faz o produto da matriz do sistema por um vetor aproveitando ao máximo a esparsidade da matriz. Com ela descobrimos que nesta simples operação, básica no método dos gradientes conjugados, íamos acumulando muito erro nos cálculos.

A segunda foi proporcionada por José Vitório Zago ao ter conhecimento dos maus resultados que obtivemos no começo.

A terceira é de novo sugestão de Lopes. Assim enfrentamos o estudo dessas decomposições e fizemos testes que depois de alguns arranjos, aproveitando a esparsidade da matriz, nos conduziram a resultados promissores.

Com respeito ao conteúdo do trabalho podemos dizer que:

O primeiro capítulo faz ênfase nas contribuições de Zago e de Lopes [16] na utilização dos princípios extremos duais para aproximar solução da equação do calor, chegando até o sistema que originou esta tese.

O segundo capítulo faz um estudo geral do método dos gradientes conjugados e seus pré-condicionadores para resolver sistemas lineares esparsos onde a matriz do sistema é simétrica positiva definida. Nesta parte utilizamos pré-condicionadores da forma $C = EE^T$, sendo EE^T a decomposição de Choleski usual de C ou uma decomposição de Choleski incompleta de A. É incluída também a análise da convergência do método e limitantes para o erro em cada iteração nos dois casos de decomposições.

O terceiro capítulo dá os fundamentos básicos de Meijerink [18] e [19] e Kershaw [14] com relação às decomposições incompletas usadas em combinação com o método dos gradientes conjugados para resolver sistemas lineares. No final do capítulo, seguindo estes pesquisadores, resolvemos o sistema mencionado acima.

O quarto capítulo apresenta resultados numéricos e comentários referentes a nossas experiências na solução do sistema linear.

Uma parte final, quinto capítulo, sobre conclusões e linhas de pesquisas futuras como continuação e enriquecimento de nosso problema é incluída.

Finalmente queríamos acrescentar que na escrita deste material houve momentos fugazes, fronteira do efêmero com o eterno, em que tentamos unir *Ciência e Literatura* da maneira como já o fizeram Sigmund Freud, o pai da psicanálise e Bertrand Russel o ilustre filósofo e matemático inglês, ganhador do prêmio Nobel de Literatura. Talvez não o tenhamos conseguido, mas estamos certos que tivemos muito prazer nesta atividade.

CAPÍTULO I

UM SISTEMA LINEAR ASSOCIADO A EQUAÇÃO DO CALOR

1.1 AS ORIGENS DO PROBLEMA

O problema resolvido nesta tese tem suas origens mais remotas na dissertação de Zago [30], onde ele aplica os princípios extremos duais estabelecido por Noble e Sewell em 1972 [21] para aproximar solução da equação do calor:

$$\begin{cases} \dfrac{\partial u}{\partial t} + b(t)\dfrac{\partial u}{\partial x} = \dfrac{\partial}{\partial x}\left[a(x,t)\dfrac{\partial u}{\partial x} \right] + q(x,t), \\ u(0,t) = u(1,t) = 0,\ a(x,t) > 0\ \forall x \in (0,1],\ \forall t \in (0,T], \\ u(x,0) = u_0(x). \end{cases} \tag{1.1}$$

Os ditos princípios, resumidos abaixo, são usados para problemas com uma estrutura Hamiltoniana generalizada:

$$\begin{cases} T^{*}v = \partial X / \partial w, \\ Tw = \partial X / \partial v, \end{cases} \tag{1.2}$$

onde T e T^{*} são operadores lineares fechados adjuntos e $X(v,w)$ é um funcional sobre $H_v \times H_w$, convexo em v e côncavo em w. H_v e H_w são espaços de Hilbert com produtos internos $(,)$ e $\langle,\rangle$, respectivamente.

Definindo os dois novos funcionais J_α e K_β por:

$$J_\alpha(v,w) = \langle w, \partial X/\partial w \rangle - X(v,w),$$

$$K_\beta(v,w) = (v, \partial X/\partial v) - X(v,w), \tag{1.3}$$

podemos resumir os princípios extremos duais, baseados na interpretação de Lopes [16], no seguinte:

TEOREMA 1: Se $X(v,w)$ é convexo em v e côncavo em w, então para qualquer solução $(\hat{v},\hat{w})$ de (2) temos:

a) $J_\alpha(\hat{v},\hat{w})$ é a solução de $\begin{cases} \min J_\alpha(v,w) \\ \text{sujeito a } T^* v = \partial X / \partial w \end{cases}$

b) $K_\beta(\hat{v},\hat{w})$ é a solução de $\begin{cases} \max K_\beta(v,w), \\ Tw = \partial X / \partial v \ , \end{cases}$

Além disso $J_\alpha(\hat{v},\hat{w}) = K_\beta(\hat{v},\hat{w})$.

1.2 OS PRINCÍPIOS EXTREMOS DUAIS E A EQUAÇÃO DO CALOR

Zago [30], visando aproximar solução da equação do calor, leva (1.1) ao sistema (1.2) e, usando os princípios extremos duais de Noble e Sewell, chega à seguinte formulação:

a) *Princípio do mínimo*

$$\begin{cases} \min J_\alpha(w_1,w_2) \\ \text{sujeito à restrição } R_1 \end{cases}$$

onde
$$J_\alpha(w_1,w_2) = 0.5 \int_0^1 \int_0^T a(x,t)\left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2\right]dtdx$$

$$- \int_0^1 \int_0^T q_1(x,t)w_1(x,t)dtdx +$$

$$+ 0.5\int_0^1 \left[w_1^2(x,T) + w_2^2(x,0) - 2v_0(x)w_1(x,T)\right]dx.$$

e
$$R_1: \begin{cases} \dfrac{\partial w_1}{\partial t} + b(t)\dfrac{\partial w_1}{\partial x} = \dfrac{\partial}{\partial x}\left[a(x,t)\dfrac{\partial w_2}{\partial x}\right] + q_2(x,t) \\ w_1(x,0) + w_2(x,0) = u_0(x) \\ w_1(1,t) = w_1(0,t) = 0 \end{cases}$$

b) *Princípio do máximo*

$$\begin{cases} \max K_\beta(w_1, w_2) \\ \text{sujeito à restrição } R_2 \end{cases}$$

onde $K_\beta(w_1, w_2) = 0.5 \int_0^1 \int_0^T a(x,t)\left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2\right]dtdx +$

$$+ \int_0^1 \int_0^T q_2(x,t)w_2(x,t)dtdx$$

$$- 0.5\int_0^1 \left[w_1^2(x,t) + w_2^2(x,0) - 2u_0(x)w_2(x,0)\right]dx.$$

e $$R_2: \begin{cases} \frac{\partial w_2}{\partial t} + b(t)\frac{\partial w_2}{\partial x} = \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial w_1}{\partial x}\right] + q_1(x,t), \\ w_1(x,T) - w_2(x,T) = v_0(x), \\ w_2(1,t) = w_2(0,t) = 0 . \end{cases}$$

sendo $w_1 = 0.5(u+v)$ $\qquad$ $w_2 = 0.5(u-v)$

$q_1 = 0.5(q+r)$ $\qquad$ $q_2 = 0.5(q-r)$

e a equação adjunta de (1.1):

$$\begin{cases} -\frac{\partial v}{\partial t} - b(t)\frac{\partial v}{\partial x} = \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial v}{\partial x}\right] + r(x,t), \\ v(0,t) = v(1,t) = 0 \quad , \ 0 \le t \le T, \\ v(x,T) = v_0(x) \quad , \ a(x,t) > 0 \quad \forall x \in (0,1] \ , \ \forall t \in (0,T]. \end{cases} \tag{1.4}$$

Aliás $v_0(x)$ e $r(x,t)$ são quaisquer funções de quadrado integrável.

Lopes [16] introduz algumas simplificações ao fazer $v_0(x) \equiv 0$, $r(x,t) \equiv 0$ e consegue demonstrar que os princípios extremos duais para a equação (1.1) podem ser reformulados como segue:

ã) *Princípio do mínimo*

$$\begin{cases} \min J_{\alpha}(w_1, w_2) \\ \text{sujeito à restrição } \tilde{R}_1 \end{cases}$$

sendo,

$$J_{\alpha}(w_1, w_2) = 0.5 \int_0^1 \int_0^T a(x,t) \left[\left(\frac{\partial w_1}{\partial x} \right)^2 + \left(\frac{\partial w_2}{\partial x} \right)^2 \right] dt dx$$

$$- 0.5 \int_0^1 \int_0^T q(x,t) w_1(x,t) dt dx$$

$$+ 0.5 \int_0^1 \left[w_1^2(x,T) + w_2^2(x,0) \right] dx$$

e

$$\tilde{R}_1 : \begin{cases} \dfrac{\partial w_1}{\partial t} + b(t) \dfrac{\partial w_1}{\partial x} = \dfrac{\partial}{\partial x} \left[a(x,t) \dfrac{\partial w_2}{\partial x} \right] + \dfrac{1}{2} q(x,t) \\ w_1(1,t) = w_1(0,t) = 0 \end{cases}$$

b̃) *Princípio do máximo*

$$\begin{cases} \max K_{\beta}(w_1, w_2) \\ \text{sujeito à restrição } \tilde{R}_2 \end{cases}$$

sendo,

$$K_{\beta}(w_1, w_2) = - 0.5 \int_0^1 \int_0^T a(x,t) \left[\left(\frac{\partial w_1}{\partial x} \right)^2 + \left(\frac{\partial w_2}{\partial x} \right)^2 \right] dt dx +$$

$$+ 0.5 \int_0^1 \int_0^T q(x,t) w_2(x,t) dt dx$$

$$- 0.5 \int_0^1 \left[w_1^2(x,T) + w_2^2(x,0) - 2u_0(x)w_2(x,0) \right] dx.$$

e

$$\tilde{R}_2: \begin{cases} \dfrac{\partial w_2}{\partial t} + b(t) \dfrac{\partial w_2}{\partial x} = \dfrac{\partial}{\partial x} \left[a(x,t) \dfrac{\partial w_1}{\partial x} \right] + 0.5\ q(x,t), \\ w_2(1,t) = w_2(0,t) = 0 \ . \end{cases}$$

1.3 OS FUNDAMENTOS MATEMÁTICOS DO PROBLEMA

No passo seguinte Lopes [16] define o espaço de Hilbert X onde vai trabalhar. Ela escolhe um apropriado subespaço fechado X em H_x^1 x H_x^1,

$$H_x^1 = \left\{ p(x,t) \in L^2([0,1] \times [0,T]) \,\middle|\, \frac{\partial p}{\partial x} \in L^2([0,1] \times [0,T]) \right\}.$$

Assim X será Hilbert uma vez fique provado que H_x^1 é Hilbert, o que ela faz com detalhe.

Na próxima etapa define uma forma bilinear S sobre X x X, mostrando que ela é contínua e coerciva:

Para $x = (w_1, w_2)$ e $y = (\tilde{w}_1, \tilde{w}_2)$ elementos de X, define

$$S(x,y) = \int_0^1 \int_0^T a(x,t) \left[\frac{\partial w_1}{\partial x} \cdot \frac{\partial \tilde{w}_1}{\partial x} + \frac{\partial w_2}{\partial x} \cdot \frac{\partial \tilde{w}_2}{\partial x} \right] dt dx +$$

$$+ \int_0^1 \left[w_2(x,T)\tilde{w}_2(x,T) + w_2(x,0)\tilde{w}_2(x,0) \right] dx.$$

Logo aplicando o teorema de Riesz (veja, por exemplo, Céa [5], obtém,

$$S(x,y) = < x, Ry > \quad e \quad \int_0^1 u_0(x)w_2(x,0)dx = < b, x >,$$

para certo operador R e o vetor b.

Daqui mostra que achar max $K_\beta(w_1, w_2)$ é equivalente a resolver o sistema Rx = b, que possui solução única. Isto decorre da aplicação do Lema de Lax-Milgram e seu corolário (veja [5]). De maneira semelhante pode-se trabalhar o problema do mínimo de $J_\alpha(w_1, w_2)$.

1.4 APROXIMAÇÃO DA SOLUÇÃO DO PROBLEMA

No capítulo II de sua tese, Lopes [16] encara a questão de aproximar solução para a equação do calor, em espaços de dimensão finita, desenvolvendo os cálculos para o caso particular onde $b(t) \equiv 0$, $q(x,t) \equiv 0$, $a(x,t) \equiv 1$. Disto obtém,

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2}, \\ u(0,t) = u(1,t) = 0 \quad , \quad 0 \le t \le T, \\ u(x,0) = u_0(x) \quad , \quad 0 \le x \le 1 \quad . \end{cases} \tag{1.5}$$

Daí, os princípios extremos duais ficam:

$$\max K_\beta (w_1, w_2)$$

sujeito a

$$\begin{cases} \dfrac{\partial w_2}{\partial t} = \dfrac{\partial^2 w_1}{\partial x^2} \\ w_2(0,t) = w_2(1,t) = 0, \end{cases}$$

onde,

$$K_\beta(w_1, w_2) = 0.5 \int_0^1 \int_0^T \left[\left(\frac{\partial w_1}{\partial x} \right)^2 + \left(\frac{\partial w_2}{\partial x} \right)^2 \right] dt dx$$

$$- 0.5 \int_0^1 \left[w_2^2(x,T) + w_2^2(x,0) - 2\, u_0(x) w_2(x,0) \right] dx,$$

e $\min J_\alpha(w_1, w_2)$

sujeito a

$$\begin{cases} \dfrac{\partial w_1}{\partial t} = \dfrac{\partial^2 w_2}{\partial x^2} \\ w_1(0,t) = w_1(1,t) = 0, \end{cases}$$

onde $J_\alpha(w_1, w_2) = 0.5 \int_0^1 \int_0^T \left[\left(\dfrac{\partial w_1}{\partial x} \right)^2 + \left(\dfrac{\partial w_2}{\partial x} \right)^2 \right] dt dx +$

$$+ 0.5 \int_0^1 \left[w_1^2(x,T) + u_0^2(x) - 2\, u_0(x) w_1(x,0) + w_1^2(x,0) \right] dx.$$

Como fez anteriormente, ela trabalha lá com o princípio do máximo, apenas.

A aproximação é feita com funções chapéu em t e B- splines cúbicas em x. As funções chapéu constituem um espaço $\{\phi_i(t)\}$ de funções onde,

$$\phi_i(t) = \begin{cases} \dfrac{t - t_{i-1}}{\Delta t}, & t_{i-1} \le t \le t_i \\ \dfrac{t_{i+1} - t}{\Delta t}, & t_i \le t \le t_{i+1} \\ 0, & t < t_{i-1}, \ t > t_i . \end{cases} \qquad (1.6)$$

Em nosso caso os t_i são pontos de uma malha de [0,T].

As funções B- splines constituem um outro espaço de funções $\{S_3(z)\}$ onde,

$$S_3(z) = \begin{cases} \dfrac{1}{6}(z^3 + 6z^2 + 12z + 8) & ; \ -2 \le z \le -1, \\ \dfrac{1}{6}(-3z^3 - 6z^2 + 4) & ; \ -1 \le z \le 0, \\ \dfrac{1}{6}(3z^3 - 6z^2 + 4) & ; \ 0 \le z \le 1, \\ \dfrac{1}{6}(-z^3 + 6z^2 - 12z + 8); & 1 \le z \le 2. \end{cases}$$

Já que nosso intervalo para x é [0,1], devemos aplicar as correspondentes fórmulas de transformação. O leitor interessado nas propriedades destes espaços de funções, dados acima, pode consultar Prenter, P.M. [23].

Lopes e Zago tomam as aproximações w_1 e w_2 nos referidos espaços da seguinte forma:

$$w_1(x,t) = \sum_{i=1}^{N} \sum_{j=1}^{M} C_{ij}\phi'_j(t)\Psi_i(x), \quad \Psi_i(0) = 0 \ \forall i$$

tendo-se $\Psi_i(x) = S_3\left(\frac{x}{\Delta x} - i\right)$, após feita a transformação dos intervalos.

Das relações $\frac{\partial w_2}{\partial t} = \frac{\partial^2 w_1}{\partial x^2}$ e $w_2(0,t) = w_2(1,t) = 0$ derivamos,

$$w_2(x,t) = \sum_{i=1}^{N} \sum_{j=1}^{M} C_{ij}\varphi_j(t)\Psi''_i(x), \quad \Psi''_i(0) = \Psi''_i(1) = 0, \ \forall i. \qquad (1.9)$$

Agora teremos $K_\beta = K_\beta(C_{11},\ldots,C_{1M},C_{21},\ldots,C_{2M},\ldots,C_{N1},\ldots,C_{NM})$, ou escrito mais breve, $K_\beta = K_\beta(C_{ij})$. Portanto,

$$K_\beta(C_{ij}) = -0.5 \int_0^1\int_0^T \left[\sum_{i=1}^{N} \sum_{j=1}^{M} C_{ij}\phi'_j(t)\Psi_i(x) \right]^2 dtdx$$

$$- 0.5 \int_0^1\int_0^T \left[\sum_{i=1}^{N} \sum_{j=1}^{M} C_{ij}\phi_j(t)\Psi'''_i(x) \right]^2 dtdx$$

$$- 0.5 \int_0^1 \left[\sum_{i=1}^{N} \sum_{j=1}^{M} C_{ij}\phi_j(T)\psi''_i(x) \right]^2 dx$$

$$- 0.5 \int_0^1 \left[\sum_{i=1}^{N} \sum_{j=1}^{M} C_{ij}\phi_j(0)\psi''_i(x) \right]^2 dx$$

$$+ \int_0^1 u_0(x) \left[\sum_{i=1}^{N} \sum_{j=1}^{M} C_{ij}\phi_j(0)\psi''_i(x) \right]^2 dx. \qquad (1.10)$$

Para achar max $K_\beta(C_{ij})$ fazemos $\nabla K_\beta(C_{ij}) = 0$; esta condição,

depois de alguns cálculos e arranjo de termos, pode ser escrita na forma de um sistema linear Ax = b,

$$A = \begin{bmatrix} AA & BB & 0 & 0 & 0 & \cdots & 0 \\ BB & A1 & BB & 0 & 0 & \cdots & 0 \\ 0 & BB & A1 & BB & 0 & \cdots & 0 \\ 0 & 0 & BB & A1 & BB & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & \cdots & BB & A1 & BB \\ 0 & 0 & 0 & \cdots & 0 & BB & AA \end{bmatrix} \qquad (1.11)$$

$$x = \begin{bmatrix} C_1 \\ C_2^1 \\ \vdots \\ C_M \end{bmatrix}_{N.Mx1}, \text{ com } C_j = \begin{bmatrix} C_{1j} \\ C_{2j}^1 \\ \vdots \\ C_{Nj} \end{bmatrix}_{Nx1} \text{ e } \quad b = \begin{bmatrix} b_1 \\ b_2^1 \\ \vdots \\ b_M \end{bmatrix}_{NMx1}$$

$$b_1 = \begin{bmatrix} \int u_0 \Psi_1'' \\ \int u_0 \Psi_2'' \\ \vdots \\ \int u_0 \Psi_N'' \end{bmatrix}_{Nx1} \quad \text{e} \quad b_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{Nx1}, \quad j = 2, \ldots, M$$

sendo, AA, BB, A1 matrizes NxN,

$$AA = (aa_{mi}) \ , \ BB = (bb_{mi}) \ , \ A1 = (a1_{mi}) \ ,$$

$$aa_{mi} = \frac{1}{\Delta t}\int_0^1 \Psi_i'(x)\Psi_m'(x)dx + \frac{\Delta t}{3}\int_0^1 \Psi_i'''(x)\Psi_m'''(x)dx + \int_0^1 \Psi_i''(x)\Psi_m''(x)dx,$$

$$bb_{mi} = -\frac{1}{\Delta t}\int_0^1 \Psi_i'(x)\Psi_m'(x)dx + \frac{\Delta t}{6}\int_0^1 \Psi_i'''(x)\Psi_m'''(x)dx,$$

$$a1_{mi} = \frac{2}{\Delta t}\int_0^1 \Psi_i'(x)\Psi_m'(x)dx + \frac{4\Delta t}{6}\int_0^1 \Psi_i'''(x)\Psi_m'''(x)dx.$$

No capítulo I Lopes [16] demonstra o seguinte teorema que nos permitirá escrever a aproximação da solução de (1.1)

TEOREMA.2: Se $w_1(x,t)$ e $w_2(x,t)$ são funções suficientemente regulares que satisfazem

$$\begin{cases} \dfrac{\partial w_1}{\partial t} + b(t)\dfrac{\partial w_1}{\partial x} = \dfrac{\partial}{\partial x}\left[a(x,t)\dfrac{\partial w_2}{\partial x}\right] + \dfrac{1}{2}\, q(x,t) \\ w_1(1,t) = w_1(0,t) = 0 \\ w_2(1,t) = w_2(0,t) = 0 \quad , \end{cases}$$

e (w_1, w_2) minimiza o funcional J_α de ã) , então

$$u(x,t) = w_1(x,t) + w_2(x,t)$$

é solução da equação (1.1).

Com isso, para um nó genérico (x_ℓ, t_k) , a proximação da solução é dada por,

$$u(x_\ell, t_k) \approx w_2(x_\ell, t_k) + w_1(x_\ell, t_k) =$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} C_{ij}\phi_j(t_k)\Psi_i''(x_\ell) + \sum_{i=1}^{N}\sum_{j=1}^{M} C_{ij}\phi_j'(t_k)\Psi_i(x_\ell).$$

Fazendo alguns cálculos algébricos e aplicando

$$\phi_\alpha'(t_k) = \lim_{t \to t_k^-} \phi_\alpha'(t) \ , \ \alpha = k-1,\ k,\ k+1 \ ,$$

fica afinal:

$$u(x_\ell, t_k) \approx \left[\frac{2}{3\Delta t} - \frac{2}{\Delta x^2}\right] C_{\ell k} + \left[\frac{1}{6\Delta t} + \frac{1}{\Delta x^2}\right]\left[C_{\ell-1,k} + C_{\ell-1,k}\right] -$$

(1.12)

$$- \frac{1}{6\Delta t}\left[C_{\ell-1.k-1} + 4C_{\ell,k-1} + C_{\ell+1,k-1}\right], \text{ ou}$$

se aplicarmos $\phi'_\alpha(t_k) = \lim_{t \to t_k^+} \phi'_\alpha(t)$, 2 = k-1, k, k+1 ficará,

$$u(x_\ell, t_k) \approx -\left[\frac{2}{3\Delta t} + \frac{2}{\Delta x^2}\right] C_{\ell k} + \left[\frac{1}{\Delta x^2} - \frac{1}{6\Delta t}\right]\left[C_{\ell-1,k} + C_{\ell+1,k}\right] +$$

(1.13)

$$+ \frac{1}{6\Delta t}\left[C_{\ell-1,k+1} + 4C_{\ell,k+1} + C_{\ell+1,k+1}\right].$$

Lopes [16] encerra o capítulo II com o teorema da convergência do método de aproximação:

Seja U_k a solução de max $K_\beta(w_1^k, w_2^k)$ no espaço NM-dimensional

> X_k e U a solução de max $K_\beta(w_1, w_2)$ no espaço X. Ela demonstra que $U_k \longrightarrow U$, mostrando que $D = \overset{\infty}{\underset{k=1}{\cup}} X_k$ é denso em X pois no teorema 4 do capítulo I, tinha provado que: Se $\{X_k\}$ é uma sequência de sub-espaço de X tais que $D = \overset{\infty}{\underset{k=1}{\cup}} X_k$ é denso em X, então $U_k \longrightarrow U$, D = span $\left[\left[\left(\varphi'_j(t)\Psi_i(x),\ \varphi_j(t)\Psi'_j{}'(x)\right)\right]\right]$.

1.5 MAU-CONDICIONAMENTO DO PROBLEMA

Nossa matriz A obtida na seção anterior resulta ser esparsa, tridiagonal por blocos, simétrica positiva definida e cada bloco é banda 7. No entanto ela apresenta sérios problemas numéricos no cálculo. Por exemplo, ao fazer o produto de A por um vetor, ordem

de A sessenta, achamos que o erro máximo no produto está entre 10 e 20, usando precisão simples. Isto nos remete a seguinte questão:
Suponha que vamos resolver o sistema linear Ax = b, onde A é uma matriz não singular, como em nosso caso; qual é o efeito sobre a solução x, se A e b sofrem uma pequena perturbação?
Se δx, δA, δb são as perturbações em x, A e b, respectivamente e $A(x+\delta x) = b + \delta b$ (supondo δA pequeno), pode-se provar que, salvo uma quantidade desprezível,

$$\frac{\|\delta x\|}{\|x\|} \leq k(A) \left[\frac{\delta A}{\|A\|} + \frac{\delta b}{\|b\|} \right], \quad k(A) = \|A\|\|A^{-1}\|. \tag{1.14}$$

k(A) é chamado o número de condição de A. Esta expressão está nos dizendo que para δA e δb pequenos o erro relativo, no cálculo da solução, depende fortemente de k(A). Se k(A) tem valores moderados (em cujo caso dizemos que A é bem condicionada), pequenas perturbações em A e/ou b produzem uma pequena perturbação na solução. Se pelo contrário, k(A) é relativamente grande (agora dizemos que A é mal-condicionada), pequenas perturbações em A e/ou b, podem produzir perturbações *catastróficas* na solução. Na prática, com o computador, a forma que o número de condição afeta os dados depende também da precisão da máquina.

O leitor interessado em aprofundar mais estes assuntos pode consultar FORSYTHE [8] e [7], NOBLE [20] e GOLUB [9]. Neste último é apresentado um algoritmo para estimar o número de condição, dado por CLINE, MOLER, STEWART E WILKINSON (1979).

Para a matriz A da seção 1.4 calculamos o número de condição para vários tamanhos, usando norm dois tal que k(A) é o quociente entre o maior e o menor autovalor de A. Daí obtivemos para:

A	21x21	k(A) é da ordem de 10^4,	
A	45x45	k(A) é da ordem de 10^6,	
A	93x93	k(A) é da ordem de 10^8.	(1.15)

Todos estes fatores servem para explicar o mal - condicionamento de nosso problema.

Nos dois capítulos seguinte apresentaremos a teoria e algumas técnicas, que nos permitirão encarar o mal-condicionamento de A.

CAPÍTULO II

TEORIA GERAL DO MÉTODO DOS GRADIENTES CONJUGADOS E SEUS PRÉ-CONDICIONADORES

2.1 INTRODUÇÃO

Desde sua descoberta independentemente por Hestenes e Stiefel em 1952, o Método dos Gradientes Conjugados tem apresentado *eficiência e simplicidade* em inúmeros problemas que levam à resolução de um sistema linear, cativando a atenção de muitos pesquisadores em conexão com assuntos numéricos. Estas suas duas virtudes principais foram assinaladas com ênfase por seus descobridores na introdução de [10].

O método dos gradientes conjugados é a grosso modo, um método iterativo para resolver sistemas lineares com matrizes simétricas positivas definidas e está intimamente ligado ao método iterativo de Máxima Descida. Os métodos iterativos são muito úteis na resolução de sistemas lineares esparsos de grande porte, principalmente porque eles são fáceis de programar, pode-se armazenar só os elementos não nulos da matriz, podem ser usados para refinar soluções obtidas por métodos diretos, fornecem um bom chute inicial para a solução de certos problemas, não precisamos calcular mais casas decimais que as requeridas. Por isso faremos nas duas seções seguintes um breve resumo ou curto estudo destes métodos.

Nas seções 2.3, 2.4 e 2.5 seguiremos na sua essência [3], [13], [17] e na 2.6 usaremos [18].

2.2 MÉTODOS ITERATIVOS EM GERAL

Uma opção para resolver tanto problemas lineares quanto não lineares é aplicar algum método iterativo. Um método iterativo, em geral, partindo de um chute inicial x^0 para a solução vai gerando uma

sequência de aproximações x^1, x^2,..., que em princípio devem convergir para a solução.

De maneira abstrata um método iterativo para resolver um sistema linear

$$Ax = b,$$

pode se definir como funções ϕ_0, ϕ_1,...,ϕ_{k+1},... onde

$$x^0 = \phi_0(A,b),$$

$$x^1 = \phi_1(x^0;A,b),$$

$$\vdots$$

$$x^{k+1} = \phi_{k+1}(x^0,x^1,\ldots,x^k;A,b) \ , \ k = 0,1,\ldots$$

veja [29].

Muito conhecidos são os chamados métodos iterativos estacionários lineares básicos: Método de Jacobi, Método JOR ("Jacobi Overrelaxation Method) associado ao Método de Jacobi, Método de Gauss-Seidel, Método SOR (Successive overrelaxation), Método de Richardson e Método de Richardson Generalizado.

Todos eles podem ser colocados na forma geral,

$$x^{k+1} = Gx^k + f, \ k = 0,1,\ldots \qquad (2.1)$$

para alguma matriz G nxn e algum vetor f.

Em conexão com um método iterativo há duas questões fundamentais a resolver: por um lado o assunto da convergência da sequência $\{x^0,x^1,\ldots,x^k,\ldots\}$ para qualquer chute inicial x^0; por outro lado a rapidez ou taxa de convergência da mesma.

Um resultado muito importante é que o método iterativo definido por (2.1) converge se e somente se $\rho(G) < 1$, onde $\rho(G)$ é o raio espectral de G definido por

$$\rho(G) = \max_{1 \leq i \leq n} |\lambda_i|, \text{ sendo } \lambda_i \text{ autovalor de G.} \qquad (2.2)$$

Para um aprofundamento do aqui exposto recomendamos em ordem de profundidade dos temas tratados [11], [2], [22], [27] e [29].

2.3 MÉTODO DE MÁXIMA DESCIDA

Sejam $S \subset \mathbb{R}^n$, A uma matriz nxn simétrica e $f: S \longrightarrow \mathbb{R}$ o funcional dado por $f(x) = \frac{1}{2} x^T Ax - b^T x + c$, $x \in \mathbb{R}^n$, onde $b \in \mathbb{R}^n$ fixo e $c \in \mathbb{R}$. Diz-se neste caso que f é um funcional quadrático.

Visto que gradiente de f, no caso que tenha derivadas parciais de primeira ordem contínuas, é Ax - b, concluímos que encontrar um ponto estacionário para f (isto é achar $\hat{x}$ tal que $\nabla f(\hat{x}) = 0$) é equivalente a resolver o sistema Ax - b = 0. O gradiente de um funcional quadrático f será simbolizado por g ($g(x) = \nabla f(x) = Ax - b$).

Um conceito ligado a um funcional quadrático, em nosso caso, que tem a ver com a interpretação geométrica do comportamento dos métodos numéricos para achar o mínimo do funcional, é a *noção de superfície de nível*.

Para o funcional dado acima o conjunto,

$$L_k = \{ x \in S;\ f(x) = k,\ k \in f(S)\} \qquad (2.3)$$

é uma *superfície de nível* para cada k.

Resultados do cálculo vetorial mostram que se $x \in L_k$, então g(x) é ortogonal a L_k em x e g(x) aponta na direção que f cresce mais rapidamente. Assim temos que o gradiente próximo de um mínimo de f aponta para fora.

Agora, seja A uma matriz simétrica positiva definida e $S = \mathbb{R}^n$.

Queremos dar uma idéia do comportamento que apresentam as *superfície de nível* de f numa vizinhança de um ponto estacionário $\hat{x}$ de f.

Para isto reescrevemos f como,

$$f(x) = \frac{1}{2}(x - \hat{x})^T A(x - \hat{x}) + \hat{c}, \tag{2.4}$$

onde $\hat{c} = -\ 1/2\ b^T\hat{x} + c$.

Mas o fato de A ser simétrica, positiva definida implica que existem um conjunto $\{\lambda_i\}_{i=1}^n$ de autovalores de A positivos e uma base ortonormal de autovetores associados $\{v_i\}_{i=1}^n$ em $\mathbb{R}^n$. Assim a matriz $V = [v_1,\ldots,v_n]$ é ortogonal e se $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_n)$ e $z = V^T(x - \hat{x})$ então, $AV = V\Lambda$ e

$$\begin{aligned} f(x) &= f(\hat{x} + Vz) \\ &= \tilde{f}(z) = \frac{1}{2}(Vz)^T A(Vz) + \hat{c} \\ &= \frac{1}{2} z^T(V^TAV)z + \hat{c} \\ &= \frac{1}{2} z^T\Lambda z + \hat{c}, \\ &= \frac{1}{2}\sum_{i=1}^{n}\lambda_i z_i^2 + \hat{c},\ z = (z_1,\ldots,z_n). \end{aligned}$$

Logo, se $f(x) = k > \hat{c}$ tem-se

$$\sum_{i=1}^{n}\lambda_i z_i^2 = \hat{k}, \tag{2.5}$$

onde $\hat{k} = 2(k - \hat{c}) > 0$.

A expressão (2.5) diz que as curvas de nível de um funcional quadrático associado a uma matriz simétrica positiva definida são elipsoides.

Para efeitos de nossa análise definamos agora:

$$D_k = \inf_{y \in S_k}\left\{\sup_{x \in L_k}\|x - y\| \,/\, \inf_{x \in L_k}\|x - y\|\right\}, \tag{2.6}$$

sendo S_k o conjunto de pontos interiores a L_k. Tomaremos (2.6) como

a medida de distorção de L_k. De (2.6) temos que $D_k \geq 1$ e $D_k = 1$ se L_k é uma esfera. Supondo $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ encontramos para L_k dado em (2.6):

$$D_k = \frac{\lambda_n}{\lambda_1} = k(A). \tag{2.7}$$

Já podemos dizer, segundo (2.7) e o significado de número de condição de uma matriz, que os métodos numéricos para achar o mínimo de um funcional se comportam melhor quando as superfícies de nível numa vizinhança do domínio forem esferas ou próximas a esferas, e são mal comportados se elas mostrarem muita distorção. Isto é, se elas diferem muito de esferas.

Com os conceitos que precedem passamos a descrever o método de máxima descida:
Para minimizar o funcional quadrático associado a uma matriz simétrica definida positiva usamos iterações do tipo

$$x^{k+1} = x^k + \tau_k d^k, \quad k = 0,1,\ldots, \tag{2.8}$$

onde τ_k é um parâmetro e d^k é um vetor ou direção de busca.

Queremos escolher d^k de maneira que f diminua e achar τ_k de sorte que,

$$f(x^k + \tau_k d^k) = \min_{\tau \geq 0} f(x^k + \tau d^k). \tag{2.9}$$

TEOREMA 2.3.1

Seja $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ o funcional quadrático $f(x) = \frac{1}{2} x^t Ax - b^t x + c$ associado a uma matriz A simétrica positiva definida. Para achar o mínimo de f usando o método de máxima descida descrito pelas equações (2.8) e (2.9) tomamos:

$$d^k = - g(x^k), \tag{2.10}$$

$$\tau_k = - d^{kT} g^k / d^{kT} A d^k, \tag{2.11}$$

sendo g o gradiente de f e $g^k = g(x^k)$, $k = 0,1,\ldots$.

PROVA:

Pelo teorema de Taylor do cálculo vetorial tem-se

$$f(x + h) = f(x) + g^T(x)h + 0 (\|h\|). \tag{2.12}$$

Escolhendo em (2.12) $x = x^k$, $h = \tau_k d^k$, vem

$$f(x^{k+1}) = f(x^k) + \tau_k g^T(x^k) d^k + 0(\tau_k).$$

Para f diminuir devemos ter $f(x^{k+1}) < f(x^k)$. Isto acontece se:

$$g^T(x^k) d^k < 0, \tag{2.13}$$

e τ_k é convenientemente pequeno para ser $\tau_k g^T(x^k) d^k + 0(\tau_k) < 0$. Mas $d^k = - g(x^k)$ satifaz (2.13).

Isto prova que f diminue na direção de menos o gradiente e mais ainda: entre todas as direções de busca em x^k essa é a direção segundo a qual f dimunue mais rapidamente. De fato, se minimizarmos as derivadas direcionais em x^k dadas por $g^T(x^k)y$, onde y com $\|y\| = 1$ é a direção de busca, de

$$|g^T(x^k)y| \leq \|g(x^k)\| \|y\| = \|g(x^k)\|,$$

achamos que o mínimo é atingido para $y = - g(x^k)/\|g(x^k)\|$.

Para provar (2.11), em (2.9) fazemos $\frac{d}{d\tau}[f(x^k+\tau d^k)] = 0$ e aplicando uma das regras da cadeia do cálculo vetorial escrevemos,

$$0 = \frac{d}{d\tau}[f(x^k + \tau d^k)] = g^T(x^k + \tau d^k)d^k$$

$$= [A(x^k + \tau d^k) - b]^T d^k$$

$$= (Ax^k - b)^T d^k + \tau {d^k}^T A d^k.$$

Então fazendo $\tau = \tau_k$ resulta (2.11), sempre que $d^k \neq 0$.

Se prova que $d^k = 0$ só acontecerá quando a solução já tiver sido alcançada (veja [3]). A mesma referência pode ser consultada para o assunto relacionado com a análise da convergência do método de máxima descida.

2.4 MÉTODO DOS GRADIENTES CONJUGADOS

Nos propomos de novo achar o mínimo do funcional quadrático

$$f(x) = \frac{1}{2} x^T Ax - b^T x + c, \quad x \in \mathbb{R}^n, \tag{2.14}$$

onde A é uma matriz simétrica positiva definida, usando iterações do tipo

$$x^{k+1} = x^k + \tau_k d^k \ , \quad k = 0, 1, \ldots \tag{2.15}$$

selecionando τ_k como em (2.11) e

$$d^{k+1} = - g^{k+1} + \beta_k d^k \ , \quad k = 0, 1, \ldots \tag{2.16}$$

onde $d^0 = - g^0$ e $\beta_0, \beta_1, \ldots$ deverão ser tais que as direções d^k, d^{k+1} sejam ortogonais em relação a um certo produto interno que induz uma

norma em $\mathbb{R}^n$ e introduzimos a seguir:

Sejam x, y $\in \mathbb{R}^n$ e A uma matriz nxn simétrica positiva definida, então define-se,

$$< x, y >_A = x^T A y, \tag{2.17}$$

$$\|x\|_A = < x, x >_A^{1/2}. \tag{2.18}$$

LEMA 2.4.1

As expressões (2.17) e (2.18) definem respectivamente um produto interno e uma norma em $\mathbb{R}^n$.

PROVA:

É suficiente verificar estas propriedades:

$$< x, x >_A \geq 0, \text{ e } < x, x >_A = 0 \text{ se e somente se } x = 0, \tag{2.19}$$

$$< x + y, z >_A = < x, z >_A + < y, z >_A, \tag{2.20}$$

$$< \lambda x, y >_A = \lambda < x, y >_A, \tag{2.21}$$

$$< x, y >_A = < y, x >_A, \tag{2.22}$$

$$\|x\|_A \geq 0 \text{ e } \|x\|_A = 0 \text{ se e somente se } x = 0, \tag{2.23}$$

$$\|x + y\|_A \leq \|x\|_A + \|y\|_A, \tag{2.24}$$

$$\|\lambda x\|_A = |\lambda| \|x\|_A, \tag{2.25}$$

para todo $x, y, z \in \mathbb{R}^n$ e $\lambda \in \mathbb{R}$.

Provaremos (2.24) as outras derivam-se diretamente das definições (2.17) e (2.18). Para qualquer parâmetro real t,

$$0 \leq \|x + ty\|_A^2 = t^2 < y, y >_A + 2t < x, y >_A + < x, x >_A.$$

Disto decorre que a função de variável real t

$$g(t) = < y, y >_A t^2 + 2t < x, y >_A + < x, x >_A,$$

é uma função quadrática em t e $g(t) \geq 0$, para todo t.

Logo, $4 < x, y >_A^2 - 4 < x, x >_A < y, y >_A \leq 0$, derivando daí (2.24).

Visando reformular o algoritmo dos gradientes conjugados observemos que a condição de d^k e d^{k+1} serem ortogonais com respeito ao produto $< \cdot >_A$ é equivalente, segundo (2.16), a escrever

$$\beta_k = < g^{k+1}, d^k >_A / < d^k, d^k >_A.$$

Com isso, se quisermos encontrar o mínimo do funcional (2.14) ou equivalentemente resolver o sistema linear $Ax = b$, pelo método dos gradientes conjugados, fazemos:

$$x^0 \in \mathbb{R}^n \text{ arbitrário e } d^0 = - g^0 = Ax^0 - b, \tag{2.26}$$

$$\tau_k = - g^{kT} d^k / < d^k, d^k >_A, \tag{2.27}$$

$$x^{k+1} = x^k + \tau_k d^k, \tag{2.28}$$

$$d^{k+1} = - g^{k+1} + \beta_k d^k, \tag{2.29}$$

$$\beta_k = < g^{k+1}, d^k >_A / < d^k, d^k >_A, \tag{2.30}$$

onde $g^k = g(x^k) = Ax^k - b$, $k = 0,1,\ldots$.

No resto desta seção iremos salientando algumas propriedades importantes do método descrito de (2.26) a (2.30) que vão nos permitir fazer uma análise da convergência e estimativas de erro. Uma das motivações de ter introduzido o produto $< , >_A$ e a norma $\|\cdot\|_A$ é precisamente simplificar esta análise.

LEMA 2.4.2

Para $m = 0,1,\ldots$

$$\text{span } \{d^0,\ldots,d^m\} = \text{span } \{g^0,\ldots,g^m\} = \text{span } \{g^0,Ag^0,\ldots,A^mg^0\},$$

sendo $\text{span}\{v^1,\ldots,v^m\}$ o espaço gerado por $v^1,\ldots,v^m$.

PROVA:

Apliquemos indução sobre m: para $m = 0$ o lema é verdadeiro, pois $d^0 = - g^0$ por (2.26).

Suponhamos que seja verdadeiro para $m = k$; isto quer dizer que

$$\text{span } \{d^0,\ldots,d^k\} = \text{span } \{g^0,\ldots,g^k\} = \text{span } \{g^0,Ag^0,\ldots,A^kg^0\}.$$

Primeiro vejamos que $\text{span } \{g^0,\ldots,g^{k+1}\} \subset \text{span } \{g^0,Ag^0,\ldots,A^{k+1}g^0\}$; pela hipótese de indução só falta ver que $g^{k+1} \in \text{span}\{g^0,Ag^0,\ldots,A^{k+1}g^0\}$; mas de (2.28) multiplicando por A deduzimos,

$$g^{k+1} = g^k + \tau_k \, Ad^k, \tag{2.31}$$

$Ad^k \in \text{span } \{g^0,Ag^0,\ldots,A^{k+1}g^0\}$ pois $d^k \in \text{span } \{g^0,Ag^0,\ldots,A^kg^0\}$, também $g^k \in \text{span } \{g^0,Ag^0,\ldots,A^{k+1}g^0\}$ pela hipótese de indução. Então $\text{span } \{g^0,\ldots,g^{k+1}\} \subset \text{span } \{g^0,Ag^0,\ldots,A^{k+1}g^0\}$.

Para deduzir a outra inclusão basta provar que: $A^{k+1}g^0 \in \text{span}\{g^0,\ldots,g^{k+1}\}$. Pela hipótese de indução $A^kg^0 \in \text{span}\{d^0,\ldots,d^k\}$, assim $A^{k+1}g^0 \in \text{span}\{Ad^0,\ldots,Ad^k\}$; mas por (2.31) cada d^j, $0 \le j \le k$, pode se exprimir em termos de g^{j+1} e g^j, portanto $A^{k+1}g^0 \in \text{span}\{g^0,\ldots,g^{k+1}\}$. Temos demonstrado que $\text{span}\{g^0,\ldots,g^{k+1}\} = \text{span}\{g^0,Ag^0,\ldots,A^{k+1}g^0\}$.

Por outro lado, aplicando (2.29) derivamos diretamente

$$\text{span}\{g^0,\ldots,g^{k+1}\} = \text{span}\{d^0,\ldots,d^{k+1}\},$$

chegando com isto à prova do lema.

LEMA 2.4.3

$$g^{i\,T}g^j = 0 \quad , \quad \text{se } i \neq j \tag{2.32}$$

$$< d^i,\ d^j >_A = 0 \quad , \quad \text{se } i \neq j \tag{2.33}$$

PROVA:

Aplicamos indução sobre i e j.

Suponhamos $0 \le i,\ j \le 1$; então $< d^0,\ d^1 >_A = 0$ pois, pelo já visto, $<d^k,\ d^{k+1} >_A = 0$. Agora,

$$g^{1\,T}d^0 = g^T(x^1)d^0 = g^T(x^0 + \tau_0 d^0)d^0 = \frac{d}{d\tau}\left[f(x^1 + \tau d^0)\right]_{\tau=\tau_0} = 0$$

porque τ_0 é calculado sendo ótimo. Aplicando agora (2.26) obtemos $g^{1\,T}g^0 = 0$.

Supondo o lema verdadeiro para $0 \le i,\ j \le k$, devemos

demonstrar que

$$g^{k+1^T} g^j = 0, \quad j = 0,1,\ldots,k \text{ e } < d^{k+1}, \; d^j >_A = 0, \quad j = 0,1,\ldots,k \; .$$

De span $\{d^0,\ldots,d^j\}$ = span $\{g^0,\ldots,g^j\}$, veja lema anterior, tiramos $d^j = \sum_{i=0}^{j} c^i g^i$, para algumas constantes c^i, então

$$g^{k^T} d^j = \sum_{i=0}^{j} c^i g^{i^T} g^k = 0, \quad j = 0,1,\ldots,k-1$$

De (2.31) deduzimos,

$$g^{k+1^T} d^j = g^{k^T} d^j + \tau_k < d^k, \; d^j >_A = 0, \quad j = 0,1,\ldots,k-1,$$

e

$$g^{k+1^T} d^k = g(x^k + \tau_k d^k)^T d^k = \frac{d}{d\tau} f(x^k + \tau d^k)_{\tau=\tau_k} = 0,$$

pois, τ_k escolhido sendo ótimo. Já podemos escrever

$$g^{k+1^T} d^j = 0, \quad j = 0,1,\ldots,k.$$

Com isto, aplicando o lema anterior para m = j,

$$g^{k+1^T} g^j = g^{k+1^T} \left(\sum_{r=0}^{j} \beta^r d^r \right),$$

para algumas constantes β^r,

$$= \sum_{r=0}^{j} \beta^r g^{k+1^T} d^r = 0,$$

ou seja, $g^{k+1^T} g^j = 0, \quad 0,1,\ldots k.$

Para provar a segunda parte, usando (2.29) vem,

$$< d^{k+1},\ d^j >_A = < -\ g^{k+1} + \beta_k d^k,\ d^j >_A,$$

$$= -\ < g^{k+1},\ d^j >_A + \beta_k < d^k,\ d^j >_A.$$

Mas $< d^k,\ d^j >_A = 0,\ j < k$, pela hipótese de indução e

$$< g^{k+1},\ d^j >_A = g^{k+1^T} A d^j.$$

De (2.31) (tiramos) que

$$Ad^j \in \text{span}\ \{g^0, \ldots, g^{j+1}\}$$

e portanto,

$$Ad^j = \sum_{r=0}^{j+1} \gamma^r g^r,$$

para algumas constantes γ^r.

Logo,

$$< g^{k+1},\ d^j >_A = \sum_{r=0}^{j+1} \gamma^r g^{k+1^T} g^r = 0,\ j = 0, 1, \ldots, k-1$$

Disto deduzimos que,

$$< d^{k+1},\ d^j >_A = 0,\ j = 0, 1, \ldots, k-1,$$

mas,

$$< d^{k+1},\ d^k >_A = 0,$$

com o qual,

$$< d^{k+1},\ d^j >_A = 0,\ j = 0, 1, \ldots, k\ .$$

Com o visto até aqui, pode-se demonstrar que o método dos gradientes conjugados, na ausência de erros de arredondamento, converge no máximo em n passos, sendo n a ordem da matriz A; porém se n for grande o resultado não tem interesse do ponto de vista computacional. Isto nos levará a introduzir o conceito de pré-condicionamento que, tem como objetivo acelerar a convergência e diminuir os efeitos do mau-condicionamento se a matriz do sistema tiver número de condição grande.

TEOREMA 2.4.1

No método dos gradientes conjugados, para algum $m \leq n$,

$$Ax^m = b$$

onde n é a ordem da matriz A.

PROVA:

De acordo com (2.32) $\{g^0, g^1, \ldots, g^n\}$ é um conjunto ortogonal de vetores em $\mathbb{R}^n$. Mas existe pelo menos algum $g^m = 0$, $0 \leq m \leq n$, pois a dimensão de $\mathbb{R}^n$ é n e $\{g^0, g^1, \ldots, g^n\}$ tem n + 1 elementos.

Logo $g^m = Ax^m - b = 0$. Isto quer dizer que x^m é a solução de $Ax = b$.

LEMA 2.4.4

No método dos gradientes conjugados $x^k - x^0$ é a projeção ortogonal do erro inicial $x - x^0$, com respeito ao produto $\langle \, , \, \rangle_A$, sobre o espaço

$$W_k = \text{span } \{d^0, \ldots, d^{k+1}\}$$

PROVA:

Devemos demonstrar que,

$$< x^k - x^0,\ d^j >_A = < x - x^0,\ d^j >_A,\ j = 0,1,\ldots,\ k - 1. \qquad (2.34)$$

Usando (2.15), após várias substituições, escrevemos:

$$x^k = x^0 + \sum_{j=0}^{k-1} \tau_j d^j,\ k = 0,1,2,\ldots \qquad (2.35)$$

Daí,

$$< x^k,\ d^k >_A = < x^0,\ d^k >_A + \sum_{j=0}^{k-1} \tau_j < d^j,\ d^k >_A,$$

donde,

$$< x^k,\ d^k >_A = < x^0,\ d^k >_A,\ k = 0,1,2,\ldots\ . \qquad (2.36)$$

De (2.35),

$$< x^k - x^0,\ d^i >_A = \sum_{j=0}^{k-1} \tau_j <d^j,\ d^i >_A = \tau_i < d^i,\ d^i >_A,\ \text{se } i = 0,1,\ldots,k-1.$$

Daqui vem,

$$\tau_i = \frac{< x^k - x^0,\ d^i >_A}{< d^i,\ d^i >_A},\ i = 0,1,\ldots,k-1. \qquad (2.37)$$

Por outro lado de (2.11)

$$\tau_k = \frac{- g^{k^T} d^k}{<d^k,\ d^k>_A},\ k = 0,1,\ldots$$

mas,

$$- g^{k^T} d^k = - (Ax^k - b)^T d^k = - (Ax^k - Ax)^T d^k = - (x^k - x)^T A d^k =$$

$$= < x - x^k,\ d^k >_A = < x,\ d^k >_A - < x^k,\ d^k >_A$$

$$= < x,\ d^k >_A - < x^0,\ d^k >_A \text{ por} \qquad (2.36)$$

Então,

$$\tau_k = \frac{< x - x^0,\ d^k >_A}{< d^k,\ d^k >_A}, \quad k = 0, 1, \ldots \qquad (2.38)$$

Em particular de (2.37) e (2.38) para $j = 0, 1, \ldots, k-1$ temos,

$$< x^k - x^0,\ d^j >_A = < x - x^0,\ d^j >_A .$$

(2.34) é equivalente a $< x - x^k,\ d^j >_A = 0$, $j = 0, 1, \ldots, k-1$ o que quer dizer que o vetor erro $e^k = x - x^k$ em cada iteração é ortogonal ao subespaço gerado por $\{d^0, d^1, \ldots, d^{k-1}\}$.

Com os seguintes dois teoremas enfrentamos a questão da estimativa de erro para o método dos gradientes conjugados:

TEOREMA 2.4.2

Para o método dos gradientes conjugados,

$$\|x - x^k\|_A \leq \max_j |p_k(\lambda_j)| \|x - x^0\|_A, \quad \forall p_k \in \hat{P}_k, \qquad (2.39)$$

onde $\hat{P}_k$ é o conjunto de polinômios $p_k(z) = \sum_{j=0}^{k} \beta_j z^j$ de grau menor ou igual a k com $\beta_o = 1$.

PROVA:

Pelo lema 2.4.4 teremos,

$$\|x - x^k\|_A = \|(x - x^0) - (x^k - x^0)\|_A \leq \|(x - x^0) - \omega\|_A,$$

para todo $\omega \in W_k$.

Mas,

$$W_k = \text{span } \{d^0, \ldots, d^{k-1}\} = \text{span } \{g^0, Ag^0, \ldots, A^{k-1}g^0\},$$

e temos também,

$$g^0 = Ax^0 - b = Ax^0 - Ax = - A(x, x^0).$$

Com isto,

$$W_k = \text{span } \{A(x - x^0), \ldots, A^k(x - x^0)\}$$

e $\omega \in W_k$ pode se exprimir como,

$$\omega = \sum_{j=1}^{k} \alpha_j A^j(x - x^0),$$

portanto,

$$\|x - x^k\|_A \leq \left\|(x - x^0) - \sum_{j=1}^{k} \alpha_j A^j(x - x^0)\right\|_A$$

$$= \left\|(I - \sum_{j=1}^{k} \alpha_j A^j)\ (x - x^0)\right\|_A$$

$$= \|p_k(A)\ (x - x^0)\|_A$$

$$\leq \|p_k(A)\|_A \|x - x^0\|_A.$$

Pela equivalência das normas em $\mathbb{R}^n$, aplicando o fato de que

para uma matriz B simétrica de autovalores $\lambda_1,\ldots,\lambda_n$, $\|B\|_2 = \max_j|\lambda_j|$ e as propriedades dos autovalores para soma e potência de matrizes, obteremos finalmente (2.39).

Segundo o teorema anterior para estimar o fator pelo qual se reduz o erro inicial $\|x - x^0\|_A$ após k passos é suficiente construir um polinômio p_k de grau menor ou igual a k tal que $p_k(0) = 1$ e p_k sendo o menor possível no intervalo $[\lambda_1, \lambda_n]$ contendo os autovalores de A, $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

Os polinômios que satisfazem estas condições são os polinômios de Chebyshev. Uma estimativa está dada pelo seguinte teorema:

TEOREMA 2.4.3

Para o método dos gradientes conjugados temos a seguinte estimativa de erro,

$$\|x - x^k\|_A \leq T_k[(\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)]^{-1}\|x - x^0\|_A, \qquad (2.40)$$

além disso se $p(\epsilon)$ é o menor inteiro k tal que

$$\|x - x^k\|_A \leq \epsilon\, \|x - x^0\|_A, \quad \forall x^0 \in \mathbb{R}^n,$$

então,

$$p(\epsilon) \leq 1/2\ \sqrt{k(A)}\ \ell n(2/\epsilon) + 1, \qquad (2.41)$$

onde T_k e o polinômio de Chebyshev de grau k e os autovalores de A estão ordenados como $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

PROVA:

A melhor estimativa para (2.39) se obtém tomando

$$\min_{p_k \in \Pi_k^1} \max_{\lambda} |p_k(\lambda)|,$$

onde Π_k^1 é o conjunto de polinômios p_k de grau k tal que $p_k(0) = 1$. Mas pela teoria de aproximações,

$$\max_{\lambda_1 \leq \lambda \leq \lambda_n} |\tilde{P}_k(\lambda)| = \min_{p_k \in \Pi_k^1} \max |p_k(\lambda)|,$$

sendo,

$$\tilde{P}_k(\lambda) = \frac{T_k[(\lambda_n + \lambda_1 - 2\lambda)/(\lambda_n - \lambda_1)]}{T_k[(\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)]},$$

onde T_k é o polinômio de Chebyshev de grau k e

$$\max_{\lambda_1 \leq \lambda \leq \lambda_n} |\tilde{P}_k(\lambda)| = 1/T_k[(\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)].$$

Para demonstrar (2.41), seja $p(\epsilon)$ o menor inteiro k que satisfaz

$$1/T_k[(\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)] < \epsilon.$$

Isto quer dizer que $p(\epsilon) - 1$ não satisfaz. Ou seja,

$$1/T_{p(\epsilon)-1}[(\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)] \geq \epsilon,$$

ou

$$1/T_{p(\epsilon)-1}\left[\frac{k(A) + 1}{k(A) - 1}\right] \geq \epsilon, \quad k(A) = \frac{\lambda_n}{\lambda_1}.$$

Segundo a identidade,

$$T_k(x) = \frac{1}{2}\left[(x + \sqrt{x^2-1}\,)^k + (x - \sqrt{x^2-1}\,)^k\right],\ x \in \mathbb{R}$$

derivamos a relação,

$$T_k\left[\frac{\alpha + 1}{\alpha - 1}\right] = \frac{1}{2}\left[\left[\frac{\sqrt{\alpha} + 1}{\sqrt{\alpha} - 1}\right]^k + \left[\frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1}\right]^k\right] > \frac{1}{2}\left[\frac{\sqrt{\alpha} + 1}{\sqrt{\alpha} - 1}\right]^k,$$

disto decorre,

$$T_{p(\epsilon)-1}\left[\frac{k(A) + 1}{k(A) - 1}\right] > \frac{1}{2}\left[\frac{\sqrt{k(A)} + 1}{\sqrt{k(A)} - 1}\right]^{p(\epsilon)-1},$$

com o qual

$$\frac{1}{2}\left[\frac{\sqrt{k(A)} + 1}{\sqrt{k(A)} - 1}\right]^{p(\epsilon)-1} < T_{p(\epsilon)-1}\left[\frac{k(A) + 1}{k(A) - 1}\right] \leq 1/\epsilon\ ,$$

portanto,

$$\left[p(\epsilon) - 1\right] \ell n \left[\frac{\sqrt{k(A)} + 1}{\sqrt{k(A)} - 1}\right] \leq \ell n(2/\epsilon).$$

Usando o fato de que,

$$\ell n[(\sqrt{\alpha} + 1)/(\sqrt{\alpha} - 1)] > 2/\sqrt{\alpha}\ ,\ \alpha > 1,$$

obtemos (2.41).

Desta análise (2.40) poderia se escrever como,

$$\|x - x^k\|_A \leq 2\left[\frac{\sqrt{k(A)} - 1}{\sqrt{k(A)} + 1}\right]^k \|x - x^0\|_A, \qquad (2.42)$$

donde podemos concluir que se A for mal-condicionada a convergência do método é muito lenta.

Dependendo da distribuição dos autovalores de A no intervalo $[\lambda_1, \lambda_n]$, o limitante dado em (2.42) pode ser reduzido. Veja, por exemplo, [12] e [26].

O fato de, no método dos gradientes conjugados a taxa de convergência ser dependente do número de condição da matriz em questão, nos leva a estudar técnicas para acelerá-la. Uma maneira de fazer isto é introduzir no método uma nova matriz C que seja uma boa aproximação de A de sorte que $C^{-1}A$ esteja próxima da matriz identidade ou pelo menos que o número de condição de $C^{-1}A$ diminua consideravelmente. Estes conceitos serão explorados ou na próxima seção.

2.5 PRÉ-CONDICIONADORES POR DECOMPOSIÇÃO COMPLETA

Se formos resolver o sistema linear Ax = b, A matriz simétrica positiva definida, pelo método dos gradientes conjugados, estamos aqui interessados em introduzir no método uma matriz C simétrica definida positiva, de maneira que resolver Ax = b seja equivalente a resolver um outro sistema $\tilde{A}y = \tilde{b}$ envolvendo a matriz C, onde se tenham estas vantagens: $k(\tilde{A})$ deve ser muito menor que k(A) e resolver o sistema Ch = g seja mais eficiente que resolver Ax = b, usando para C uma fatoração conveniente como $C = EE^T$, a conhecida decomposição de Choleski; ou $C = LDL^T$, sendo L uma matriz triangular inferior e D uma matriz diagonal.

Equivalentemente, queremos reduzir o problema de minimizar o funcional

$$f(x) = \frac{1}{2} x^T Ax - b^T x + c$$

ao de minimizar um outro funcional

$$\tilde{f}(y) = \frac{1}{2} y^T \tilde{A} y - \tilde{b}^T y + \tilde{c}.$$

Seja, então, C uma matriz simétrica positiva definida e C =

EE^T a decomposição de Choleski usual, que chamaremos *decomposição de Choleski completa* de C para diferenciá-la de outro tipo de decomposição a ser discutida na próxima seção e consideremos o funcional associado à matriz A em estudo

$$f(x) = \frac{1}{2} x^T Ax - b^T x + c.$$

Definamos agora o novo funcional $\tilde{f}(y)$ empregando a transformação

$$y = E^T x \tag{2.43}$$

por

$$f(x) = f(E^{-T}y) = \tilde{f}(y) = \frac{1}{2} (E^{-T}y)^T A(E^{-T}y) - b^T(E^{-T}y) + c$$

$$= \frac{1}{2} y^T(E^{-1}AE^{-T})y - E^{-1}b^T y + c,$$

ou,

$$\tilde{f}(y) = \frac{1}{2} y^T\tilde{A}y - \tilde{b}^T y + \tilde{c}, \tag{2.44}$$

onde,

$$\tilde{A} = E^{-1}AE^{-T}\ ,\ \tilde{b} = E^{-1}b \text{ e } \tilde{c} = c. \tag{2.45}$$

LEMA 2.5.1

A matriz $\tilde{A}$ associada ao funcional quadrático de (2.44) é simétrica positiva definida e tanto $\tilde{A}$ quanto $C^{-1}A$ possuem os mesmos autovalores.

PROVA:

De (2.45)

$$\tilde{A}^T = (E^{-1}AE^{-T})^T = (E^{-T})^T A^T (E^{-1})^T = E^{-1}AE^{-T} = \tilde{A},$$

logo $\tilde{A}$ é simétrica.

De (2.43) e (2.45),

$$x^TAx = (E^{-T}y)^TA(E^{-T}y),$$

$$= y^T(E^{-1}AE^{-T})y$$

$$= y^T\tilde{A}y > 0 \ \forall y \neq 0,$$

pois $x^TAx > 0 \ \forall x \neq 0$ por A ser simétrica positiva definida, assim temos que $\tilde{A}$ é positiva definida.

Para demonstrar que $\tilde{A}$ e $C^{-1}A$ possuem os mesmos autovalores, basta ver que elas são semelhantes:

$$E^{-T}\tilde{A}E^T = E^{-T}(E^{-1}AE^{-T})\ E^T,$$

$$= E^{-T}E^{-1}A = (EE^T)^{-1}A = C^{-1}A$$

a igualdade

$$E^{-T}\tilde{A}E^T = C^{-1}A, \qquad (2.46)$$

prova que $\tilde{A}$ e $C^{-1}A$ são matrizes semelhantes.

COROLÁRIO 2.5.1

Ao resolver o sistema linear $\tilde{A}y = \tilde{b}$ pelo método dos gradientes conjugados geramos a sequência

$$y^0,\ y^1,\ldots,y^k,\ldots\ .$$

Seja

$$x^k = E^{-T}y^k\ ,\ k = 0,1,\ldots \qquad e \qquad \hat{y},\ \hat{x}$$

as soluções dos sistemas

$$\tilde{A}y = \tilde{b} \quad e \quad Ax = b,$$

respectivamente. Então,

1) $\hat{y} = E^T\hat{x}$,

2) $\lim_{k \to \infty} y^k = \hat{y}$ e $\lim_{k \to \infty} x^k = \hat{x}$,

3) A taxa de convergência de $\{x^k\}$ está determinada por $k(\tilde{A})$.

PROVA:

Para provar (1) é suficiente ver que $E^T\hat{x}$ é solução de $\tilde{A}y = \tilde{b}$. Vejamos,

$$\tilde{A}(E^T\hat{x}) = (E^{-1}AE^{-T})\ (E^T\hat{x}) = E^{-1}(A\hat{x}) = E^{-1}b = \tilde{b},$$

ou seja, $E^T\hat{x}$ é solução de $\tilde{A}y = \tilde{b}$, portanto $\hat{y} = E^T\hat{x}$.

Que $\{y^k\}$ converge para $\hat{y}$ é aplicação direta do teorema 2.4.1.

Por outro lado, a continuidade do operador E^{-T} nos leva a,

$$\lim_{k \to \infty} x^k = \lim_{k \to \infty} (E^{-T}y^k) = E^{-T}(\lim_{k \to \infty} y^k) = E^{-T}\hat{y} = \hat{x} \text{ por (1).}$$

Para demonstrar (3) apliquemos a relação (2.4.2) ao sistema $\tilde{A}y = \tilde{b}$,

$$\|\hat{y} - y^k\|_{\tilde{A}} \leq 2\left[\frac{\sqrt{k(\tilde{A})} - 1}{\sqrt{k(\tilde{A})} + 1}\right]^k \|\hat{y} - y^0\|_{\tilde{A}} , \qquad (2.47)$$

mas,

$$\|y^k - \hat{y}\|^2_{\tilde{A}} = \|E^T(x^k - \hat{x})\|^2_{\tilde{A}} ,$$

$$= [E^T(x^k - \hat{x})]^T\tilde{A}\,[E^T(x^k - \hat{x})],$$

$$= (x^k - \hat{x})^T\,[E(E^{-1}AE^{-T})\,E^T(x^k - \hat{x})],$$

$$= (x^k - \hat{x})^T A(x^k - \hat{x}),$$

$$= \|x^k - \hat{x}\|^2_A .$$

Da mesma maneira,

$$\|\hat{y} - y^0\|_{\tilde{A}} = \|\hat{x} - x^0\|_A$$

e substituindo em (2.47) derivamos (3).

Com o objetivo de simplificar a referência damos as seguintes definições:

DEFINIÇÃO 2.5.1

A matriz $C = EE^T$ introduzida nesta seção é chamada *matriz pré-condicionadora*. Às vezes diremos simplesmente que $C = EE^T$ é um *pré-condicionador* para sistema $Ax = b$.

A matriz $\tilde{A} = E^{-1}AE^T$ é chamada *matriz pré-condicionada*.

O método descrito pelo algoritmo exposto de (2.26) a (2.30) é chamado de o método dos gradientes conjugados sem pré-condicionar.

Tendo em vista a apresentação de um algoritmo eficiente

computacionalmente tanto do método dos gradientes conjugados sem pré-condicionar quanto do método dos gradientes conjugados pré-condicionado, que definiremos depois, damos o seguinte teorema.

TEOREMA 2.5.1

O algoritmo dos gradientes conjugados sem pré-condicionar é equivalente ao seguinte:

Selecionamos $x^0 \in \mathbb{R}^n$ arbitrário, calculamos

$$g^0 = Ax^0 - b, \quad d^0 = - g^0 \tag{2.48}$$

fazendo logo,

$$\tau_k = g^{k^T} g^k / d^{k^T} A d^k, \tag{2.49}$$

$$x^{k+1} = x^k + \tau_k d^k, \tag{2.50}$$

$$g^{k+1} = g^k + \tau_k A d^k, \tag{2.51}$$

$$\beta_k = g^{k+1^T} g^{k+1} / g^{k^T} g^k, \tag{2.52}$$

$$d^{k+1} = - g^{k+1} + \beta_k d^k. \tag{2.53}$$

PROVA:

Já tendo demonstrado (2.31), é suficiente provar as equivalências das expressões (2.27) e (2.49), (2.30) e (2.52).

Para a primeira parte: de (2.29) escrevemos

$$d^k = - g^k + \beta_{k-1} d^{k-1}, \quad d^{k-1} = - g^{k-1} + \beta_{k-2} d^{k-2}, \ldots$$

assim fica, substituindo recorrentemente na primeira igualdade,

$$d^k = - g^k - \sum_{j=1}^{k} \gamma_j^k g^{k-j}, \tag{2.54}$$

onde

$$\gamma_j^k = \beta_{k-1}\beta_{k-2}\cdots\beta_{k-j}.$$

Daí obtemos, depois de aplicar (2.32):

$$d^{k^T} g^k = - g^{k^T} g^k, \tag{2.55}$$

o que prova a equivalência entre (2.27) e (2.49).

Para demonstrar a outra parte: de (2.31) resulta,

$$Ad^k = \tau_k^{-1} \ (g^{k+1} - g^k),$$

portanto,

$$g^{k+1^T} Ad^k = \tau_k^{-1} \ g^{k+1^T} g^{k+1} \tag{2.56}$$

e

$$d^{k^T} Ad^k = \left[-g^k - \sum_{j=1}^{k} \gamma_j^k g^{k-j}\right]^T \tau_k^{-1} \ (g^{k+1} - g^k)$$

ou

$$d^{k^T} Ad^k = \tau_k^{-1} \ g^{k^T} g^k \tag{2.57}$$

Aplicando (2.56) e (2.57) em (2.30) derivamos (2.60), completando deste modo a prova do teorema (2.5.1).

Se aplicarmos o algoritmo do teorema (2.5.1) à solução do sistema $\tilde{A}y = \tilde{b}$, com $\tilde{A}$ e $\tilde{b}$ dados por (2.45) vem:

Selecionamos $y^0 \in \mathbb{R}^n$ arbitrário, calculamos

$$\tilde{g}^0 = \tilde{A}y^0 - \tilde{b}, \quad \tilde{d}^0 = -\tilde{g}^0 \tag{2.58}$$

e fazemos,

$$\tilde{\tau}_k = \tilde{g}^{k^T}\tilde{g}^k \,/\, \tilde{d}^k \tilde{A} \tilde{d}^k, \tag{2.59}$$

$$y^{k+1} = y^k + \tilde{\tau}_k \tilde{d}^k, \tag{2.60}$$

$$\tilde{g}^{k+1} = \tilde{g}^k + \tilde{\tau}_k \tilde{A} \tilde{d}^k, \tag{2.61}$$

$$\tilde{\beta}_k = \tilde{g}^{k+1^T}\tilde{g}^{k+1} / \tilde{g}^{k^T}\tilde{g}^k, \tag{2.62}$$

$$\tilde{d}^{k+1} = -\tilde{g}^{k+1} + \tilde{\beta}_k \tilde{d}^k; \qquad k = 0, 1, \ldots \tag{2.63}$$

onde

$$\tilde{g}^k = \tilde{A}y^k - \tilde{b}, \quad y^k = E^T x^k.$$

Além disso, $\tilde{g}^k$ e $\tilde{d}^k$ satisfazem:

LEMA 2.5.2

No método dos gradientes conjugados aplicado a $\tilde{A}y = \tilde{b}$ tem-se:

$$\tilde{g}^k = E^{-1} g^k \tag{2.64}$$

$$\tilde{d}^k = E^T d^k \tag{2.65}$$

PROVA:

$$\tilde{g}^k = \tilde{A}y^k - \tilde{b} = E^{-1}AE^{-T}y^k - E^{-1}b$$

$$= E^{-1}(AE^{-T}y^k - b)$$

$$= E^{-1}(Ax^k - b) = E^{-1}g^k.$$

Para demonstrar (2.65); de (2.28) derivamos

$$E^T x^{k+1} = E^T x^k + \tau_k E^T d^k$$

ou

$$y^{k+1} = y^k + \tau_k E^T d^k,$$

mas por (2.60)

$$y^{k+1} = y^k + \tilde{\tau}_k \tilde{d}^k.$$

Das duas últimas igualdades se tomarmos d^k e $\tilde{d}^k$ do mesmo comprimento obteremos (2.65).

Em resumo, até aqui temos dito que resolver o sistema linear $Ax = b$ é equivalente a resolver o sistema linear $\tilde{A}y = \tilde{b}$ com a vantagem de que $k(\tilde{A})$ pode ser muito menor que $k(A)$. A tranferência de um sistema ao outro foi feita introduzindo o pré-condicionador $C = EE^T$.

Agora queremos dar um algoritmo, de reconhecida eficiência computacional, equivalente ao descrito de (2.58) a (2.63) para resolver o sistema $\tilde{A}y = \tilde{b}$, usando a matriz C e as variáveis antigas, isto é aquelas do método dos gradientes conjugados para resolver $Ax = b$. Isto evitará fazer o passo cada vez que formos resolver o dito sistema aplicando um pré-condicionador. Este algoritmo corresponde ao método que chamaremos no que segue *Método dos Gradientes Conjugados Pré-condicionado.*

TEOREMA 2.5.2

O algoritmo dado de (2.66) a (2.71) é equivalente a este outro algoritmo:

Seja $x^0 \in \mathbb{R}^n$ arbitrário,

$$g^0 = Ax^0 - b \ , \ h^0 = C^{-1}g^0, \ d^0 = - h^0 \tag{2.66}$$

e façamos:

$$\tau_k = g^{k^T} h^k \ / \ d^{k^T} Ad^k, \tag{2.67}$$

$$x^{k+1} = x^k + \tau_k d^k, \tag{2.68}$$

$$g^{k+1} = g^k + \tau_k \ Ad^k, \tag{2.69}$$

$$h^{k+1} = C^{-1} g^{k+1}, \tag{2.70}$$

$$\beta_k = g^{k+1^T} h^{k+1} \ / \ g^{k^T} h^k, \tag{2.71}$$

$$d^{k+1} = - h^{k+1} + \beta_k d^k, \tag{2.72}$$

onde

$$g^k = g(x^k) = Ax^k - b \quad \text{e} \quad k = 0, 1, \ldots$$

A operação (2.70) deverá se fazer resolvendo o sistema $Ch^{k+1} = g^{k+1}$.

PROVA:

$$\tilde{g}^{k^T} \tilde{g}^k = (E^{-1}g^k)^T (E^{-1}g^k)$$

$$= (E^{-T}E^{-1}g^k)^T g^k$$

$$= (C^{-1}g^k)^T g^k = h^{k^T} g^k.$$

$$\tilde{d}^{k^T}\tilde{A}\,\tilde{d}^k = \tilde{d}^{k^T}(E^{-1}AE^{-T})\tilde{d}^k$$

$$= (E^{-T}\tilde{d}^k)^T A(E^{-T}\tilde{d}^k)$$

$$= d^{k^T}Ad^k,$$

pelo lema (2.5.2).

Isto prova a equivalência de (2.59) e (2.67). De (2.60),

$$E^T x^{k+1} = E^T x^k + \tilde{\tau}_k E^T d^k,$$

ou

$$x^{k+1} = x^k + \tilde{\tau}_k d^k.$$

Fazemos $\tilde{\tau}_k = \tau_k$ e fica provada a equivalência entre (2.60) e (2.68).

De (2.61),

$$E^{-1}g^{k+1} = E^{-1}g^k + \tilde{\tau}_k(E^{-1}AE^{-T})\ (E^T d^k)$$

ou

$$g^{k+1} = g^k + \tilde{\tau}_k\, Ad^k,$$

mas temos que $\tilde{\tau}_k = \tau_k$; assim resulta a equivalência entre (2.61) e (2.69).

A equivalência entre (2.62) e (2.71) segue do que, já foi demonstrado no início, com $\beta_k = \tilde{\beta}_k$.

De (2.63) deduzimos,

$$E^T d^{k+1} = -\,E^{-1}g^{k+1} + \tilde{\beta}_k E^T d^k,$$

ou

$$d^{k+1} = -\,(E^{-T}E^{-1})g^{k+1} + \beta_k d^k,$$

com o qual

$$d^{k+1} = - h^{k+1} + \beta_k d^k.$$

A equivalência entre (2.58) e (2.66) é direta.

2.6 PRÉ-CONDICIONADORES POR DECOMPOSIÇÃO INCOMPLETA

Para certo tipo de matrizes A não singulares correspondentes a uma família de sistemas lineares Ax = b, que definiremos nesta seção, é possível fazer uma única decomposição LU de uma certa matriz K próxima a A, de modo que L e U sejam obtidas da decomposição $\tilde{L}\,\tilde{U}$ de A, selecionando com antecedência alguns lugares (i,j) fora da diagonal de A onde os elementos correspondentes em L e U serão nulos. A vantagem da nova decomposição, que se chama Decomposição Incompleta de A, é que escrevendo A = K - R há um método iterativo para resolver Ax = b associado a essa expressão de rápida convergência. Se além disso A for simétrica positiva definida teremos a decomposição de Choleski de K ou decomposição de Choleski Incompleta de A, que poderá ser usada como pré-condicionador no método dos gradientes conjugados pré-condicionado para resolver Ax = b.

Aqui apresentaremos o desenvolvimento feito por Meijerink e Van der Vorst [18] com respeito a este assunto. Antes introduziremos algumas definições e resultados do livro de Varga [27], fechando a introdução com um lema de Ky Fan [15] adaptado à nossas circunstâncias.

2.6.1 DEFINIÇÃO

1) Sejam $A = (a_{ij})$ e $B = (b_{ij})$ matrizes reais mxn. Dizemos que $A \geq B$ se $a_{ij} \geq b_{ij}$ para todo i, j $\quad 1 \leq i \leq m$, $1 \leq j \leq n$.

   Em particular, se B = 0 teremos a definição para $A \geq 0$.

2) Uma matriz real $A = (a_{ij})$ nxn não singular com $a_{ij} \leq 0$ para todo $i \neq j$ e $A^{-1} \geq 0$ é chamada uma M-matriz.

3) Sejam A, K, R matrizes nxn, A não singular. Se

$$A = K - R,$$

e K é não singular dizemos que $A = K - R$ é um splitting de A. Se além disso, $K^{-1} \geq 0$ e $R \geq 0$ dizemos que $A = K-R$ é um Splitting regular de A.

É fácil ver que os elementos da diagonal de uma M-matriz A nxn são positivos. De fato, se

$$A = (a_{ij}) \text{ e } A^{-1} = (r_{ij})$$

então fazendo o produto da linha i pela coluna i em $A^{-1}A = I$ temos,

$$\sum_{k=1}^{n} r_{ik} a_{ki} = 1$$

ou

$$r_{ii} a_{ii} - \sum_{\substack{k=1 \\ k \neq i}}^{n} r_{ik} |a_{ki}| = 1.$$

Como $r_{ii} \geq 0$, desta última igualdade se deduz que $a_{ii} > 0$, $1 \leq i \leq n$.

Agora se tivermos o sistema linear $Ax = b$ e $A = K - R$ for um splitting de A então existe um método iterativo para resolver o sistema anterior associado ao Splitting $A = K - R$. Seja $\hat{x}$ a solução do sistema; assim,

$$K\hat{x} = R\hat{x} + A\hat{x} = R\hat{x} + b,$$

e então podemos escrever o seguinte método iterativo:

$$Kx^{k+1} = Rx^{k} + b \; ; \; k = 0, 1, \ldots \qquad (2.73)$$

TEOREMA 2.6.1

Seja $A = (a_{ij})$ uma nxn M-matriz e $B = (b_{ij})$ matriz nxn se

$$a_{ij} \leq b_{ij} \leq 0 \quad \text{para} \quad i \neq j \quad \text{e} \quad 0 < a_{ii} \leq b_{ii}.$$

Então B é também uma M- matriz.

TEOREMA 2.6.2

Se $A = K - R$ é um splitting regular de A e $A^{-1} \geq 0$, então o método iterativo (2.73) converge para qualquer chute inicial x^0.

TEOREMA 2.6.3

Toda M-matriz simétrica é uma matriz positiva definida.

LEMA 2.6.1 (Lema de Ky Fan)

Seja $A = (a_{ij})$ uma M-matriz de ordem n e $C = (c_{ij})$ uma matriz de ordem n - 1 definida por,

$$c_{ij} = \frac{a_{ij}a_{nn} - a_{in}a_{nj}}{a_{nn}} \; ; \; i, j = 1,2,\ldots,n-1,$$

Então, C é também uma M-matriz.

A seguir o teorema da Decomposição de LU incompleta.

TEOREMA 2.6.4

Seja $A = (a_{ij})$ uma nxn M-matriz e

$$P_n = \{(i, j) \;/\; i \neq j, \quad 1 \leq i \leq n, \quad 1 \leq j \leq n\}.$$

Então, para para cada $P \subset P_n$ existe uma matriz triangular inferior $L = (\ell_{ij})$ com diagonal unitária, uma matriz triangular superior $U = (u_{ij})$ e uma matriz $R = (r_{ij})$ com

$$\ell_{ij} = 0 \quad \text{se} \quad (i, j) \in P,$$

$$u_{ij} = 0 \quad \text{se} \quad (i, j) \in P,$$

$$r_{ij} = 0 \quad \text{se} \quad (i, j) \notin P,$$

tais que o splitting $A = LU - R$ é regular. Além disso, os fatores L e U são únicos.

PROVA:

Faremos a demonstração construindo L e U. Para isto definamos

$$A^k = (a^k_{ij}) \;, \quad \tilde{A}^k = (\tilde{a}^k_{ij}), \quad L^k = (\ell^k_{ij}) \quad \text{e} \quad R^k = (r^k_{ij}),$$

onde

$$A^0 = A, \tag{2.74}$$

$$\tilde{A}^k = A^{k-1} + R^k, \tag{2.75}$$

$$A^k = L^k \tilde{A}^k \;; \quad k = 1,2,\ldots,n-1 \text{ e} \tag{2.76}$$

$$r^k_{kj} = - a^{k-1}_{kj} \;, \quad \text{se } (k, j) \in P, \tag{2.77}$$

$$r_{ik}^{k} = - a_{ik}^{k-1} \quad , \quad \text{se } (i,\ k) \in P \text{ e } r_{ij}^{k} = 0 \text{ nos demais casos,} \qquad (2.78)$$

L^k é a matriz identidade, exceto na coluna k dada pelo vetor

$$\left[0,0,\ldots,1, \ - \ \frac{\tilde{a}_{k+1,k}^{k}}{\tilde{a}_{kk}^{k}} \ , \ - \ \frac{\tilde{a}_{k+2,k}^{k}}{\tilde{a}_{kk}^{k}} \ ,\ldots, \ - \ \frac{\tilde{a}_{nk}^{k}}{\tilde{a}_{kk}^{k}} \right]^{T} . \qquad (2.79)$$

A prova tem duas partes fundamentais: de um lado a prova de que A^k, $\tilde{A}^k$ são M-matrizes e L^k, $R^k \geq 0$, $k = 1,\ldots,n-1$; de outro lado a construção de L, U e R. No final vamos definir $U = A^{n-1}$, $L = (L^{n-1}L^{n-2}\ldots L^{1})^{-1}$ e $R = R^1 + \ldots + R^{n-1}$.

Para k = 1,

$$\left.\begin{array}{l} r_{1j}^{1} = - a_{1j}^{0} = - a_{1j} \geq 0, \ (1,\ j) \in P \\ r_{i1}^{1} = - a_{i1}^{0} = - a_{i1} \geq 0, \ (1,\ j) \in P \end{array}\right\} \text{pois } a_{ij} \leq 0, \ (i,\ j) \in P,$$

daqui temos que $R^1 \geq 0$.

Provemos que $\tilde{A}^1$ é uma M-matriz,

$$\tilde{A}^1 = A^0 + R^1 = A + R^1$$

isto é,

$$\tilde{a}_{ij}^{1} = a_{ij} + r_{ij}^{1} \geq a_{ij}, \text{ pois } r_{ij}^{1} \geq 0; \text{ e}$$

$$\tilde{a}_{ij}^{1} = a_{ij} + r_{ij}^{1} = a_{ij} + \left\{ \begin{array}{c} - \ a_{ij} \\ \text{ou} \\ 0 \end{array} \right\} \leq 0 \ ;$$

as duas últimas expressões levam a,

$$a_{ij} \leq \tilde{a}_{ij}^{1} \leq 0.$$

Por outro lado,

$$\tilde{a}^1_{11} = a_{11} + r^1_{11} = a_{11},$$

então, pelo teorema (2.6.1), $\tilde{A}^1$ é M-matriz.

Provemos que $L^1 \geq 0$. L^1 é quase a identidade, a menos da coluna 1 que é o vetor

$$\left[1, -\frac{\tilde{a}^1_{21}}{\tilde{a}^1_{11}}, \ldots, -\frac{\tilde{a}^1_{n1}}{\tilde{a}^1_{11}}\right]^T$$

onde

$$\tilde{a}^1_{11} > 0 \quad e \quad \tilde{a}^1_{11} \leq 0$$

pois $\tilde{A}^1$ é M-matriz. Portanto $L^1 \geq 0$.

Para ver que A^1 é M-matriz, consideremos a expressão

$$A^k = L^k \tilde{A}^k = (I + O^k)\tilde{A}^k = \tilde{A}^k + O^k \tilde{A}^k,$$

onde O^k é quase a matriz nula, exceto na coluna k dada pelo vetor

$$\left[\underbrace{0,0,\ldots 0}_{k}, \underbrace{-\frac{\tilde{a}^k_{k+1,k}}{\tilde{a}^k_{kk}}, -\frac{\tilde{a}^k_{k+2,k}}{\tilde{a}^k_{kk}}, \ldots, -\frac{\tilde{a}^k_{nk}}{\tilde{a}^k_{kk}}}_{n-k}\right]^T,$$

Agora, as entradas (i, j) dos produtos $O^k \tilde{A}^k$ estão dadas por

$$\sum_{p=1}^{n} o^k_{ip} \tilde{a}^k_{pj} \quad ; \quad i = k+1, \ldots, n \qquad j = 1,2,\ldots,n,$$

$$= o^k_{ik} \tilde{a}^k_{kj} \quad , \quad i = k+1, \ldots, n \qquad j = 1,2,\ldots,n,$$

escrevendo $i = k + m$ como $m = 1,2,\ldots, n-k$ vem,

$$o^k_{ik} \tilde{a}^k_{kj} = o^k_{k+m,k} \tilde{a}^k_{kj} \quad , \quad j = 1,2,\ldots, n,$$

logo,

$$a^{k}_{k+m,j} = \tilde{a}^{k}_{k+m,j} + o^{k}_{k+m,k}\,\tilde{a}^{k}_{kj} \quad , \quad j = 1,2,\ldots, n$$

ou

$$a^{k}_{k+m,j} = \tilde{a}^{k}_{k+m,j} - \frac{\tilde{a}^{k}_{k+m,k}}{\tilde{a}^{k}_{kk}} \cdot \tilde{a}^{k}_{kj} \quad , \quad j = 1,2,\ldots, n$$

substituindo $k = 1$ e aplicando o lema (2.6.1) obtemos que A^1 é M-matriz.

De forma semelhante pode-se demonstrar que A^k, $\tilde{A}^k$ são M-matrizes e L^k, $R^k \geq 0$, $k = 1,\ldots, n-1$.

Para construir L, U, R observemos que,

$$L^k R^m = R^m \quad \text{se} \quad k < m, \text{ pois}$$

$$L^k R^m = (I + O^k) R^m = R^m + O^k R^m \text{ e}$$

$$(O^k R^m)_{ij} = \sum_{p=1}^{n} o^{k}_{ip} r^{m}_{pj} = o^{k}_{ik} r^{k}_{kj} \ ,$$

mas $r^{m}_{kj} = 0$ se $k < m$ de acordo com a definição de R.

Finalmente,

$$\begin{aligned}
A^{n-1} &= L^{n-1}\tilde{A}^{n-1} = L^{n-1}(A^{n-2} + R^{n-1}) \\
&= L^{n-1}A^{n-2} + L^{n-1}R^{n-1} \\
&= L^{n-1}(L^{n-2}\tilde{A}^{n-2}) + L^{n-1}R^{n-1} \\
&= L^{n-1}L^{n-2}\tilde{A}^{n-2} + L^{n-1}R^{n-1} \\
&= \ldots \\
&= L^{n-1}L^{n-2}\ldots L^{1}A^{0} + L^{n-1}L^{n-2}\ldots L^{1}R^{1} + L^{n-1}L^{n-2}\ldots L^{2}R^{2}\ldots + L^{n-1}R^{n-1} \\
&= L^{n-1}L^{n-2}\ldots L^{1}(A^{0} + R^{1} + \ldots + R^{n-1})
\end{aligned}$$

Se definirmos

$$U = A^{n-1} \ , \ L = (L^{n-1}L^{n-2}\ldots L^{1})^{-1} \ , \ R = R^{1} + \ldots + R^{n-1},$$

temos que,

$$A = LU - R, \text{ pois } A^{0} = A.$$

Demonstraremos que $A = LU - R$ é um splitting regular de A.

Confiramos que $(LU)^{-1} \geq 0$ e $R \geq 0$; $(LU)^{-1} = U^{-1}L^{-1}$, mas $U^{-1} = (A^{n-1})^{-1} \geq 0$, pois A^{n-1} é M-matriz; $L^{-1} = L^{n-1}\ldots L^{1}$, onde cada $L^{k} \geq 0$, disto tudo, $(LU)^{-1} \geq 0$.

A unicidade deriva da relação que segue e do fato que L possui diagonal unitária:

$$A = \begin{cases} LU & \text{se } (i, j) \notin P, \\ -R & \text{se } (i, j) \in P \ . \end{cases}$$

Se além de A ser M-matriz, A é simétrica, então pelo teorema (2.6.3) A é positiva definida e podemos demonstrar o seguinte corolário do teorema anterior:

COROLÁRIO 2.6.1

Se A é uma M-matriz simétrica, então para cada $P \subset P_{n}$, com a propriedade que $(i, j) \in P$ implica $(j, i) \in P$, existe uma única matriz triangular inferior L e uma matriz R não negativa ($R \geq 0$) com $\ell_{ij} = 0$ se $(i, j) \in P$ e $r_{ij} = 0$ se $(i, j) \notin P$ tal que $A = LL^{T} - R$ é um splitting regular.

Do teorema (2.6.2) extraímos o seguinte:

TEOREMA 2.6.5

Se A, L, U e R estão definidas como no teorema (2.6.4) então

o método iterativo

$$LU\, x^{k+1} = R\, x^{k} + b, \quad k = 0,1,\ldots, \tag{2.80}$$

converge à solução de $Ax = b$ para qualquer chute inicial x^0.

A estabilidade da decomposição LU incompleta está garantida pelo seguinte teorema:

TEOREMA 2.6.6

Seja $A = (a_{ij})$ uma nxn M-matriz e $B = (b_{ij})$ com $a_{ij} \leq b_{ij} \leq 0$ se $i \neq j$ e $0 < a_{ii} \leq b_{ii}$. Sejam A^1 e B^1 as matrizes que surgem de A e B por eliminação da primeira coluna usando a primeira linha. Então,

$$a^1_{ij} \leq b^1_{ij} \leq 0\ , \quad 0 < a^1_{ii} \leq b^1_{ii}\ ,$$

e B^1 é uma M-matriz.

PROVA:

$$a^{(1)}_{ij} = a_{ij} - \frac{a_{i1}}{a_{11}}\, a_{1j}$$

De $a_{ij} \leq b_{ij}$, derivamos $a_{i1} \leq b_{i1}$ e $a_{1j} \leq b_{1j}$, com o que

$$a_{i1}\, a_{1j} \geq b_{i1} b_{1j}\ , \text{ pois } a_{ij},\ b_{ij} \leq 0 \text{ se } i \neq j.$$

Por outro lado, $0 < a_{11} \leq b_{11}$ implica $1/a_{11} \geq 1/b_{11}$, assim

$$\frac{a_{11} a_{1j}}{a_{11}} \geq \frac{b_{11} b_{1j}}{b_{11}}\ ,$$

ou

$$-\frac{a_{i1}a_{1j}}{a_{11}} \leq -\frac{b_{i1}b_{1j}}{b_{11}} \quad ;$$

isto e o fato que $a_{ij} \leq b_{ij}$ levam a

$$a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \leq b_{ij} - \frac{b_{i1}b_{1j}}{b_{11}} \quad ,$$

ou seja,

$$a_{ij}^{(1)} \leq b_{ij}^{(1)} \quad .$$

Para provar que B^1 é M-matriz, observemos que

$$b_{ij}^{(1)} = b_{ij} - \frac{b_{i1}b_{1j}}{b_{11}} = \frac{b_{ij}b_{11} - b_{i1}b_{1j}}{b_{11}} \leq 0,$$

pois B é M-matriz, pelas hipóteses do teorema (Veja teorema (2.6.1)).

A^1 é M-matriz pelo lema (2.6.1). Aliás $a_{11}^1 > 0$ (propriedade já vista das M-matrizes). De novo pelo teorema (2.6.1) B^1 é M-matriz.

No método iterativo que surge ao fazer decomposição LU incompleta de uma matriz A correspondente ao sistema linear Ax = b, temos uma estimativa de erro dada no teorema que segue:

TEOREMA 2.6.7

O método iterativo (2.80) tem a seguinte estimativa de erro:

$$\|x^k - \hat{x}\|_A \leq \left[\max\left\{|1 - \lambda_{min}| \ , \ |1 - \lambda_{max}|\right\}\right]^k \|x^0 - \hat{x}\|_A \qquad (2.81)$$

onde $\hat{x}$ é a solução de Ax = b, e λ_{min}, λ_{max} são o menor e o maior autovalor de $(LU)^{-1}A$, respectivamente.

PROVA:

$$\text{De } LUx^{k+1} = Rx^k + b \text{ temos,}$$

$$LUx^{k+1} = Rx^k + A\hat{x} \text{ então,}$$

$$x^{k+1} = (LU)^{-1}A\hat{x} + (LU)^{-1}Rx^k$$

$$x^1 = (LU)^{-1}A\hat{x} + (LU)^{-1}Rx^0$$

$$= (LU)^{-1}A\hat{x} + (LU)^{-1}[LU - A]x^0$$

$$= (LU)^{-1}A(\hat{x} - x^0) + x^0. \text{ Assim,}$$

$$x^1 - \hat{x} = [\ I - (LU)^{-1}A\ ]\ (\hat{x} - x^0).$$

Indutivamente se prova que, em geral,

$$x^k - \hat{x} = [\ I - (LU)^{-1}A\ ]^k\ (\hat{x} - x^0),$$

de onde se conclui que

$$\|x^k - \hat{x}\|_A \leq \|\ [\ I - (LU)^{-1}A\ ]^k\ \|_2 \|\hat{x} - x^0\|_A.$$

Mas, se λ é um autovalor de B, $(1 - \lambda)$ é autovalor de I-B, $\|B\|_2 = \max_j |\lambda_j|$ e $(1-\lambda)^k$ é um autovalor de $(I-B)^k$. Com isto tudo concluímos o teorema.

Para findar este capítulo temos de dizer que se A for M-matriz simétrica e portanto positiva definida (Veja teorema (2.6.3)) então para resolver o sistema Ax = b pelo método dos gradientes conjugados, poderá se empregar o algoritmo descrito de (2.66) e (2.72) na seção 2.5 trocando C por LL^T.

CAPÍTULO III

ALGUNS PRÉ-CONDICIONADORES POR DECOMPOSIÇÃO INCOMPLETA NO MÉTODO DOS GRADIENTES CONJUGADOS

O universo das decomposições incompletas é muito rico e cheio de resultados surpreendentes. Neste capítulo queremos estudar uma variedade delas que, em combinação com o método dos gradientes conjugados, tem sido bem sucedidas e foram estudas primeiro por Meijerink [18] e depois por Kershaw [14], que fez uma extensão dos conceitos daquele colocando hipóteses mais fracas.

Seguindo o roteiro de Meijerink e Kershaw daremos no final do capítulo solução ao sistema linear associado à equação do calor surgida no capítulo I.

3.1 UMA CLASSE DE DECOMPOSIÇÕES INCOMPLETAS

Em seu primeiro trabalho Meijerink [18], logo após ter desenvolvido sua teoria sobre decomposição incompleta, dá duas aplicações dela. Nestas aplicações usa o sistema linear que resulta da aproximação por discretização com cinco pontos da equação diferencial:

$$-\frac{\partial}{\partial x}\left(P(x,y)\frac{\partial}{\partial x}u(x,y)\right)-\frac{\partial}{\partial y}\left(Q(x,y)\frac{\partial}{\partial y}u(x,y)\right)+\sigma(x,y)u(x,y)=f(x,y),$$

com $P(x,y)$, $Q(x,y) > 0$, $\sigma(x,y) \geq 0$ e $(x,y) \in R$, sendo R uma região quadrada, e com condições de fronteira apropriadas. A matriz $A = (a_{ij})$ do sistema é simétrica positiva definida e diagonalmente dominante de ordem n e portanto M-matriz, segundo os Capítulos III e VI de Varga [27]. A figura 3.1 exemplifica a estrutura da matriz A.

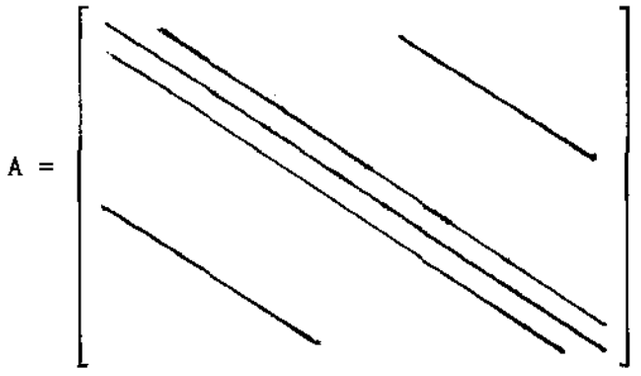


FIGURA 3.1

Na primeira aplicação o conjunto P do teorema 2.6.4 escolhido por Meijerink [18] para se fazer a decomposição incompleta de A é sinalizado por P^* sendo,

$$P^* = \{(i,j)/\ |i-j| \neq 0,1,m\},$$

onde m é a semi-banda de A. Agora se $A = GG^T$ é a decomposição de Choleski Completa de A então a decomposição incompleta LL^T de A é obtida de maneira que L^T pegue de G^T só aquelas três diagonais que conservam a estrutura de A, parte superior, o resto é zero (veja figura 3.2, linhas sem pontilhar). Neste caso o método que usa o algoritmo do teorema 2.5.2, Capítulo II com $C = LL^T$ é referido como ICCG(0): *"Incomplete Choleski & Conjugate Gradients, with 0 extra diagonal"*.

Na segunda aplicação o conjunto P do teorema 2.6.4 para se fazer a decomposição incompleta de A é sinalizado por P^3,

$$P^3 = \{(i,\ j)\ /\ |i - j| \neq 0,1,2,m-2,m-1,m\},$$

isto é, além dos lugares originais de A, parte superior, agora permite preencher em L^T as duas diagonais de G^T mais próximas à última diagonal superior e mais uma diagonal, aquela cujos elementos satisfazem $j - i = 2$; o resto é zero. No total deixa preencher 3 diagonais extras (Veja fig. 3.2). Esta variante é referida como ICCG(3).

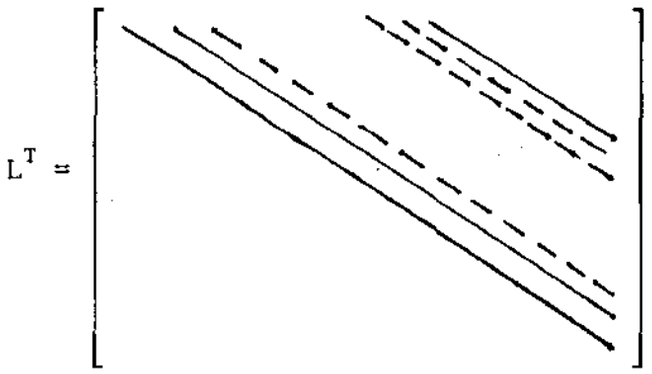


FIGURA 3.2

Dentre as causas pelas quais o método funciona há duas que possuem muito peso. A primeira é que a decomposição de Choleski completa de A é estável. Isto quer dizer , neste caso, que podemos substituir na decomposição completa alguns elementos relativamente pequenos por zero e a decomposição que daí deriva terá quase o mesmo comportamento que a original.

A segunda causa está baseada na observação feita por Meijerink no sentido que as entradas não nulas decresciam rapidamente na direção acima da diagonal principal e abaixo da última diagonal, na decomposição completa. Algumas matrizes têm esta propriedade que pode ser aproveitada escolhendo uma decomposição incompleta apropriada, tendo assim que LL^T é uma boa aproximação de A e portanto $(LL^T)^{-1}A$ está próxima da matriz identidade. Dito de outra maneira, a maior parte dos autovalores de $(LL^T)^{-1}A$ devem estar próximos a um com o que garantimos rápida convergência no ICCG (Incomplete Choleski and Conjugate Gradients).

Num segundo trabalho Meijerink [19], resolvendo outros problemas práticos, introduz diferentes tipos de decomposições incompletas da matriz resultante, deixando preencher em L^T algumas diagonais convenientes obtendo diferentes ICCG.

Como já visto, na primeira parte do seu trabalho, Meijerink consegue construir decomposições incompletas de uma matriz A sendo esta

M-matriz ou M-matriz simétrica, condição forte na demonstração da existência e unicidade. No entanto Kershaw [14] coloca uma hipótese mais fraca considerando que A seja só simétrica positiva definida, mas o preço disso é que a validade que ele obtém do método é só experimental.

Essencialmente, o método de Kershaw é um ICCG(0). Seu método pode-se resumir como segue:

Suponha que vamos resolver o sistema linear $Ax = b$, sendo $A = (a_{ij})$ uma matriz nxn simétrica positiva definida, pelo método dos gradientes conjugados, empregando como pré-condicionador, uma decomposição incompleta de A da forma $C = LL^T$ ou $C = LDL^T$, onde L é uma matriz triangular inferior e D é uma matriz diagonal. Se tivéssemos $C = A = LDL^T$ teríamos uma decomposição completa de A e os coeficientes de L e D seriam determinados recorrentemente, coluna por coluna:

$$\ell_{j1} = a_{j1} \quad , \quad j = 1,2,\ldots,n$$

$$\ell_{ji} = a_{ji} - \sum_{k=1}^{i-1} \ell_{jk}\ell_{ik}d_{kk} \; ; \; i \geq 2, \; j = i,i+1,\ldots,n$$

$$d_{ii} = (\ell_{ii})^{-1}. \qquad (3.1)$$

Definimos como padrão de esparsidade o conjunto P,

$$P = \left\{(i,\ j) \ / \ a_{ij} = 0 \ ; \ i,j = 1,\ldots,n\right\} ,$$

de sorte que faremos $\ell_{ij} = 0$ se $(i,\ j) \in P$. Quer dizer que L é forçada a ter o mesmo padrão de esparsidade que A. Por outro lado, como a condição de ser positiva definida não garante que fazendo tal escolha tenha que ter sempre $\ell_{ii} > 0$, então o algoritmo pode se interromper em alguma parte. Isto ocorrerá quando tivermos no processo $\ell_{ii} = 0$ ou $\ell_{ii} < 0$. É bom lembrar que na decomposição completa isto não acontecerá.

Para corrigir este defeito, se acontecer $\ell_{ii} \leq 0$ substituímos este valor por algum número positivo e prosseguimos com o algoritmo (3.1). Escrevendo

$$A = LDL^T + E,$$

onde E é uma matriz de erro cujas entradas não nulas estão no conjunto P pode-se ver que a decisão tomada quando $\ell_{ii} \leq 0$ faz com que o i-ésimo elemento da diagonal da matriz de erro seja não nulo. As outras entradas de E ainda estarão no conjunto P.

Deve ser claro que se $\ell_{ii} \leq 0$ acontecer poucas vezes o algoritmo aumentará as possibilidades de funcionar muito bem. Isto dependerá fortemente de LDL^T ser uma boa aproximação de A.

Ao aplicar o método dos gradientes conjugados, uma vez obtida a decomposição incompleta LL^T ou LDL^T, colocamos estas no lugar de C no algoritmo do teorema (2.5.2), capítulo II.

3.2 SOLUÇÃO DO PROBLEMA INICIAL

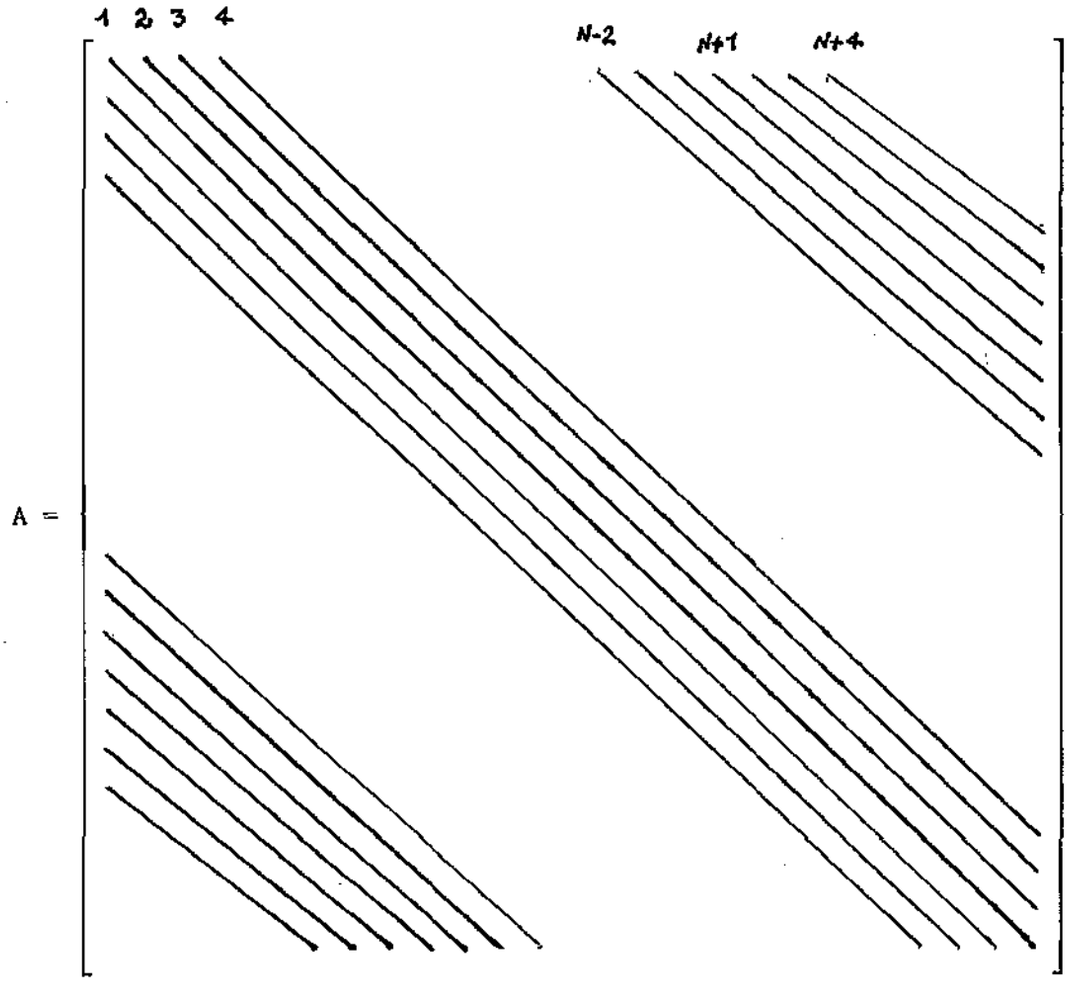
Os métodos que aqui usaremos para resolver o sistema Ax = b do capítulo I (Veja expressão (1.11)) tem a ver de um lado com os conceitos de Meijerink propostos em seus dois trabalhos [18] e [19] e de outro lado com a variante introduzida por Kershaw em [14]. Isto porque a matriz A em questão é simétrica positiva definida, mas não é M-matriz. A validade destes métodos está baseada principalmente nas experiências numéricas feitas e na estabilidade da decomposição de Choleski Completa: $A = GG^T$.

Além do mais lembremos que A é também tridiagonal por blocos, tendo só três deles diferentes: AA, BB, A1, cada um banda 7. Desta maneira A possui só 11 diagonais superiores não nulas espaçadas segundo o tamanho dos blocos. Isto mostra que A é muito esparsa (veja fig. 3.3).

Em nome da simplicidade, neste capítulo só descrevemos os métodos para a solução do sistema, deixando a implementação computacional, resultados numéricos e comentários para o próximo capítulo. No entanto queremos destacar antes de continuar, que, dado o mau-condicio-

namento da matriz e os erros de arredondamento, o método dos gradientes conjugados sem pré-condicionar converge para a solução só após um número de iterações muito alto. Por exemplo, para a matriz A de (1.11) de ordem 255, o método fez 2685 iterações com o vetor chute inicial sendo o vetor nulo e usando como critério de parada a norma do gradiente menor que 10^{-6}. Para o mesmo sistema, com as mesmas condições, empregando um dos tipos ICCG conseguimos convergência em apenas 22 iterações.

Vamos agora fazer a descrição dos tipos de ICCG que foram utilizados na solução do problema em estudo.



N é a ordem de cada bloco de A.

N ≥ 7.

FIGURA 3.3

ICCG(0):

É o mesmo empregado por Meijerink com a variante dada por Kershaw adaptada à matriz da fig. 3.3. Isto quer dizer que L^T terá 11 diagonais não nulas espaçadas como na fig.3.3; manteremos também zero nas entradas (i, j) das diagonais não nulas onde $a_{ij} = 0$. Com isso a outra parte que deveria aparecer na decomposição é ignorada.

ICCG(3):

Neste caso conservamos em L^T a estrutura de A e deixamos preencher mais 3 diagonais. Elas são as diagonais número 5, 6 e 7. Veja fig. 3.4 em linhas pontilhadas:

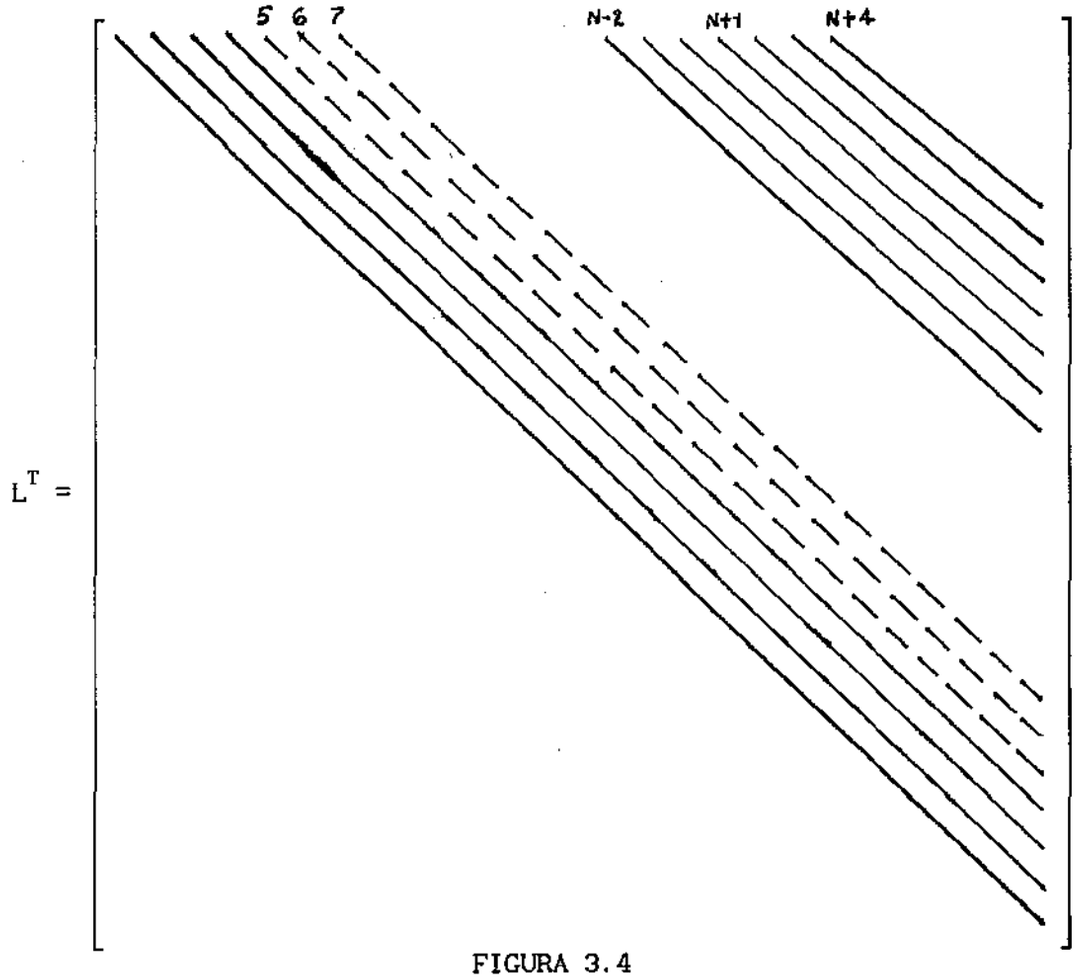


FIGURA 3.4

ICCG(2):

Além dos lugares originais de A, parte triangular superior, deixamos preencher em L^T mais duas diagonais de G^T: N-4 e N-3, sendo N a ordem de cada bloco. O resto é zero. L^T teria então 13 diagonais não nulas. Neste caso só trabalhamos com $N \geq 9$.

ICCG(-3):

Em L^T desta vez só aparecem as quatro primeiras diagonais e as quatro últimas de G^T. É dizer, com respeito à estrutura original de A, parte triangular superior, ter jogado fora os lugares N-2, N-1, N, onde N é a ordem de cada bloco de A. Assim L^T terá no total 8 diagonais não nulas.

ICCG(-7):

L^T pega de G^T, parte triangular superior, so as quatro primeiras diagonais e joga fora o resto. É dizer com respeito a estrutura original de A, parte superior, tem jogado fora sete diagonais. O resto é zero. Assim L^T tem quatro diagonais não nulas.

O ICCG(-7) nos surpreendeu porque apesar de sua simplicidade, economia de memória e operações, tem um grau de precisão para um mesmo $\|\nabla f(x)\| \leq \epsilon$, pelo menos igual ao dos outros ICCG anteriormente descritos.

Uma vez feita a decomposição incompleta LL^T de A nos diferentes ICCG, devemos aplicar o algoritmo dado no teorema (2.5.2), capítulo II trocando C por LL^T. Daí obteremos a solução do sistema Ax = b associado à equação do calor.

A comparação dos ICCG aqui estudados do ponto de vista do tamanho do sistema e número de iterações, será exibida no próximo capítulo.

CAPÍTULO IV

IMPLEMENTAÇÃO COMPUTACIONAL E RESULTADOS NUMÉRICOS

4.1 NOTA PRELIMINAR

Já tínhamos dito que nossa matriz A em estudo é simétrica positiva definida e tridiagonal por blocos. Cada bloco tem ordem N e A tem ordem NM = NxM, onde N, M são inteiros $N \geq 7$ e $M \geq 3$.

Na prática computacional dada a grande esparsidade desta matriz tivemos que aproveitar ao máximo este fato. Foi por isso que em vez de armazenar a matriz toda guardamos só os três blocos AA, BB, A1 (Veja expressão 1.11). Isto produz grande economia de memória. Assim, por exemplo, se armazenarmos a matriz cheia de ordem 495 (N = 33, M = 15) empregaremos 495 x 495 = 245.025 posições de memória, enquanto que se guardarmos só os 3 blocos usaremos 33 x 33 x 3 = 3267 posições; teremos poupado desta maneira aproximadamente 98.7% de memória. Mas esta vantajosa mudança levou-nos a fazer várias rotinas complexas adaptadas à esparsidade de A.

Dentre elas podemos citar:

TRIBLOCO: que constrói os três blocos AA, BB, A1 de acordo com as fórmulas dadas em (1.11);

BANDA 7: que faz o produto de A por um vetor, utilizando só os três blocos;

DESLOCAR: que identifica cada elemento a_{ij} de A com um elemento dos blocos AA, BB, A1 de maneira que onde tivermos a_{ij} poderemos substituí-lo por aa_{ij} ou bb_{ij} ou $a1_{ij}$, segundo o caso;

CHOLINHO: que guarda a decomposição de Choleski completa de A numa matriz de tamanho NM x (N+4); com isso, para a matriz de ordem 495 conseguimos poupar 92.5% de memória aproximadamente. Também fizemos respectivas rotinas para cada ICCG nos

quais resolvemos o sistema $(LL^T)h = g$, sendo LL^T uma decomposição incompleta de A, empregando só diagonais não nulas de L^T (e portanto de L).

Todos os programas foram feitos em linguagem FORTRAN e implementados no computador digital da linha VAX 11/785 - Sistema Operacional VMS da UNICAMP. Além disso a matriz A foi normalizada dividindo cada entrada por 2^{20} e, salvo dito o contrário, utilizamos precisão dupla.

Na tabela No.1 o produto v = Au foi calculado por meio da rotina BANDA 7 com precisão simples nos cáculos. A rotina BANDA 7 tinha sido testada antes para várias matrizes bem condicionadas, incluindo tamanhos grandes (ordem de A 495, por exemplo). A seleção de u(i) = i, i = 1,...,NM permitiu-nos, por um lado testar a rotina BANDA 7, que estava baseada no posicionamento dos elementos de A, e por outro medir facilmente o erro no produto. Para as matrizes bem condicionadas que construímos simulando a estrutura de A, obtivemos sempre erro zero mesmo com precisão simples. É bom notar que neste caso k(A) = 0.14825 x 10^7.

TABELA No. 1

PRODUTO DA MATRIZ A POR UM VETOR. ORDEM DE A 60

O VETOR É U, ONDE U(I)=I, I=1,...,60.

PRODUTO DE A PELO VETOR U: V=AU, USANDO PRECISÃO SIMPLES.

0.4368917E+07	-0.3495408E+07	0.8738119E+06	0.3625000E+01
0.3062500E+01	0.5125000E+01	0.5125000E+01	0.5125000E+01
0.5125000E+01	0.4500000E+01	0.7625000E+01	0.8500000E+01
0.3659092E+07	-0.1470206E+08	0.1842639E+08	0.2621440E+08
-0.2097152E+08	0.5242892E+07	0.9968750E+01	0.9156250E+01
0.1509375E+02	0.1475000E+02	0.1618750E+02	0.7375000E+01

0.1031250E+02	0.2018750E+02	0.1612500E+02	0.1083530E+08
-0.4334112E+08	0.5417643E+08	0.5242880E+08	-0.4194303E+08
0.1048578E+08	0.2162500E+02	0.1275000E+02	0.3025000E+02
0.2512500E+02	0.2337500E+02	0.2175000E+02	0.2212500E+02
0.2437500E+02	0.2775000E+02	0.1607819E+08	-0.6431264E+08
0.8039082E+08	0.3516772E+08	-0.2802330E+08	0.6959807E+07
0.1300000E+02	0.1025000E+02	0.2125000E+02	0.1200000E+02
0.2925000E+02	0.1950000E+02	0.2200000E+02	0.1925000E+02
0.2200000E+02	0.9745094E+07	-0.3922996E+08	0.4922520E+08

O VETOR ERRO NO PRODUTO ANTERIOR É:

0.1000000E+01	-0.1250000E+01	0.8125000E+00	0.0000000E+00
-0.4375000E+00	0.1500000E+01	-0.1000000E+01	-0.1000000E+01
0.1000000E+01	0.0000000E+00	-0.1000000E+01	0.0000000E+00
0.0000000E+00	-0,2000000E+01	-0.2000000E+01	-0.2000000E+01
0.2000000E+01	0.7000000E+01	-0.5468750E+01	0.7093750E+01
0.6562500E+00	0.7250000E+01	-0.9375000E+00	0.4625000E+01
0.2687500E+01	-0.5437500E+01	0.1375000E+01	-0.1000000E+01
-0.4000000E+01	-0.4000000E+01	0.0000000E+00	-0.1200000E+02
0.1000000E+01	0.7625000E+01	0.4500000E+01	0.5250000E+01
-0.5375000E+01	-0.3625000E+01	-0.5750000E+01	0.1212500E+02
0.2875000E+01	0.1750000E+01	0.1400000E+02	0.8000000E+01
0.0000000E+00	-0.4000000E+01	-0.4000000E+01	0.6500000E+01
0.4000000E+01	0.5250000E+01	0.6250000E+01	0.5000000E+01
0.2500000E+00	0.0000000E+00	0.3000000E+01	-0.1750000E+01
-0.3000000E+01	0.2000000E+01	0.0000000E+00	-0.4000000E+01

O ERRO MÁXIMO NO PRODUTO É: 0.1400000E+02

Na tabela No.2 exibimos o mesmo produto da tabela No.1 mas desta vez usando precisão dupla. Os resultados falam por si só.

TABELA No. 2

PRODUTO DA MATRIZ A POR UM VETOR. ORDEM DE A 60
O VETOR É U, ONDE U(I) = I, I = 1,...,60

PRODUTO DE A PELO VETOR U: V = AU, USANDO PRECISÃO DUPLA.

0.4368917E+07	-0.3495409E+07	0.8738073E+06	-0.1164153E-09
0.3492460E-09	0.2328306E-09	0.4947651E-09	0.2910383E-10
0.0000000E+00	-0.1164153E-09	-0.3492460E-09	0.1154153E-09
0.3659087E+07	-0.1470207E+08	0.1842639E+08	0.2621440E+08
-0.2097152E+08	0.5242880E+07	-0.9749783E-09	0.8294592E-09
0.5966285E-09	0.1469743E-08	0.5384209E-09	-0.1018634E-09
-0.1309672E-08	-0.6111804E-09	-0.3783498E-09	0.1083529E+08
-0.4334114E+08	0.5417643E+08	0.5242880E+08	-0.4194304E+08
0.1048576E+08	-0.6693881E-09	-0.2066372E-08	0.9604264E-09
-0.7566996E-09	0.2968591E-08	-0.1920853E-08	0.1804437E-08
-0.5820766E-10	0.1804437E-08	0.1607817E+08	-0.6431266E+08
0.8039083E+08	0.3516772E+08	-0.2802331E+08	0.6959793E+07
0.1105946E-08	0.4074536E-09	0.1105946E-08	-0.9895302E-09
0.1571607E-08	0.6402843E-09	0.1804437E-08	0.6402843E-09

O VETOR ERRO NO PRODUTO ANTERIOR É:

0.1164153E-09	-0.1164153E-09	-0.5829766E-10	0.2328306E-09
0.4656613E-09	0.1164153E-09	0.2328306E-09	0.2328306E-09
0.0000000E+00	0.2328306E-09	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	-0.4656613E-09	-0.4656613E-09
-0.1396984E-08	0.1164153E-08	-0.8294592E-09	-0.8876668E-09
-0.6548362E-09	0.6839400E-09	0.2182787E-09	0.4365575E-10
0.7858034E-09	0.1251465E-08	-0.3783498E-09	0.2328306E-09

0.1862645E-08	0.0000000E+00	0.0000000E+00	-0.9313226E-09
0.6984919E-09	-0.1455192E-09	0.4161848E-08	-0.1076842E-08
0.5820766E-10	-0.5238689E-09	-0.1804437E-08	0.2502929E-08
-0.1105946E-08	-0.2386514E-08	0.1862645E-08	0.9313226E-09
0.1862645E-08	0.0000000E+00	0.0000000E+00	0.5820766E-09
-0.4074536E-09	0.7566996E-09	0.9995302E-09	0.7566996E-09
0.5820766E-10	0.5238689E-09	-0.1746230E-09	0.5820766E-10
-0.1746230E-09	0.0000000E+00	0.0000000E+00	0.0000000E+00

O ERRO MÁXIMO NO PRODUTO É: 0.4161848E-08

Tendo já bem claro o fenômeno que estava acontecendo quisemos empregar um dos ICCG para resolver o sistema Ax = b, sendo A a mesma matriz utilizada para as tabelas No.1 e No.2 e colocando b_i como a soma dos elementos da linha i de A para poder medir o erro, pois antecipadamente sabemos que a solução exata é $x_i = 1.0$ para $i = 1,2,\ldots,NM$. Assim pegamos o ICCG(2), NM = 60, e calculamos em precisão simples com critério de parada norma do gradiente menor que 10^{-6} e obtivemos 10 iterações para resolver o sistema e erro máximo na solução do sistema 0.02146065, enquanto o mesmo problema em precisão dupla deu 8 iterações e erro máximo na solução do sistema $0.1805181 \times 10^{-11}$.

4.2 COMPARAÇÃO DOS DIFERENTES ICCG

Visando ser coerentes com o exposto no capítulo III escolhemos, por razões de espaço, a menor matriz de nosso estudo: ela é de ordem 21 (N = 7, M = 3, NM = 21). Após normalizada fizemos a decomposição de Choleski completa de A, $A = GG^T$ programando o algoritmo dado nas fórmulas (3.1) da seção 3.1. As diagonais não nulas de G^T aparecem dispostas em colunas na tabela No.3.

A diagonal da posição j tem NM - (j -1) elementos com $j \geq 1$, no resto da coluna colocamos zero.

A tabela 3 nos diz do decrescimento dos elementos de algumas diagonais. Este fato, que aconteceu em todos os casos de nossas experiências numéricas, facilitou a escolha dos ICCG que usamos.

TABELA No. 3

NESTA LISTAGEM APARERECEM AS DIAGONAIS, NÃO NULAS, DE G^TEM COLUNAS SENDO A = G * G^T A DECOMPOSIÇÃO DE CHOLESKI COMPLETA DA MATRIZ A DE TAMANHO 21. ANTES A FOI NORMALIZADA DIVIDINDO POR 2 A 20.

DIAGONAIS NÃO NULAS NÚMEROS 1, 2, 3, 4,

0.2724461E+00	-0.2706306E+00	0.1146904E+00	-0.1881869E-01
0.1795495E+00	-0.2663366E+00	0.1456648E+00	-0.2855525E-01
0.1462526E+00	-0.2591758E+00	0.1616493E+00	-0.3505633E-01
0.1293618E+00	-0.2535853E+00	0.1713118E+00	-0.3963366E-01
0.1192791E+00	-0.2494179E+00	0.1777045E+00	-0.1343678E-03
0.1126630E+00	-0.2007744E+00	-0.5226397E-04	-0.2778238E-03
0.2769147E-01	0.8040439E-04	-0.2949582E-03	-0.4453134E-03
0.3576980E+00	-0.3573203E+00	0.1528655E+00	-0.2542412E-01
0.2344474E+00	-0.3509758E+00	0.1944783E+00	-0.3878985E-01
0.1899778E+00	-0.3409013E+00	0.2161575E+00	-0.4786981E-01
0.1672171E+00	-0.3329513E+00	0.2294055E+00	-0.5438280E-01
0.1534866E+00	-0.3269276E+00	0.2382785E+00	0.1600640E-05
0.1443721E+00	-0.2593101E+00	0.2713916E-04	-0.1117759E-03
0.3117704E-01	0.1558837E-03	-0.1887810E-03	-0.2639810E-03
0.2526713E+00	-0.2506562E+00	0.1059890E+00	-0.1734018E-01
0.1667446E+00	-0.2467751E+00	0.1345399E+00	-0.2627617E-01
0.1360171E+00	-0.2402361E+00	0.1492157E+00	-0.3221252E-01
0.1204872E+00	-0.2351535E+00	0.1580383E=00	-0.3636209E-01
0.1112623E+00	-0.2313929E+00	0.1638430E+00	0.0000000E+00
0.1052419E+00	-0.1869490E+00	0.0000000E=00	0.0000000E+00
0.2694615E-01	0.0000000E+00	0.0000000E+00	0.0000000E+00

DIAGONAIS NÃO NULAS NÚMEROS 5, 6, 7, 8,

0.0000000E+00	0.0000000E+00	0.0000000E+00	0.1337886E+00
0.0000000E+00	0.0000000E+00	-0.1388239E-02	0.8832969E-01
0.0000000E+00	-0.5877263E-03	-0.1286909E-02	0.7234673E-01
-0.2815278E-03	0.6970721E-03	0.1049533E-02	0.6429713E-01
-0.4233971E-03	-0.6382536E-03	-0.8570624E-03	0.5950290E-01
-0.4191525E-03	-0.5642665E-03	-0.7130869E-03	0.5635724E-01
-0.6029034E-03	-0.7691640E-03	-0.9445967E-03	0.1321322E-01
0.3655032E-06	0.2407760E-06	0.5801336E-07	0.1019022E+00
0.4161344E-06	0.4557714E-06	-0.1912589E-03	0.6677444E-01
0.1483522E-05	-0.8634036E-04	-0.2333903E-03	0.5416612E-01
-0.3292160E-04	-0.1729062E-03	-0.1979886E-03	0.4773280E-01
-0.1365054E-03	-0.1693559E-03	-0.1589416E-03	0.4386064E-01
-0.1456925E-03	-0.1531013E-03	-0.1233569E-03	0.4129472E-01
-0.3177049E-03	-0.3407609E-03	-0.3220043E-03	0.8977172E-02
0.2841262E-06	0.2052213E-06	-0.8240182E-07	0.0000000E+00
0.1884403E-06	0.1871065E-06	0.0000000E+00	0.0000000E+00
0.9314815E-06	0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00

DIAGONAIS NÃO NULAS NÚMEROS 9, 10 e 11

-0.1338120E+00	0.5736199E-01	-0.9557998E-02
-0.1310875E+00	0.7263382E-01	-0.1450319E-01
-0.1273094E+00	0.8044520E-01	-0.1780508E-01
-0.1244459E+00	0.8513649E-01	-0.2012990E-01
-0.1223437E+00	0.8822506E-01	0.0000000E+00
-0.9766430E-01	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00
-0.1019199E+00	0.4369064E-01	-0.7279995E-02

-0.1000183E+00	0.5556367E-01	0.1110714E-01
-0.9709674E-01	0.6174255E-01	-0.1370707E-01
-0.9480119E-01	0.6551539E-01	-0.1557281E-01
-0.9306526E-01	0.6803893E-01	0.0000000E+00
-0.7370022E-01	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00
0.0000000E+00	0.0000000E+00	0.0000000E+00

Na tabela número 4 comparamos os diferentes ICCG utilizados para o sistema linear Ax = b de acordo com o tamanho de A e o número de iterações. Os dados foram obtidos fixando em todos os casos N = 15 e critério de parada norma do gradiente menor que 10^{-6}. O termo b foi construído de maneira que b_1 seja a soma dos elementos da linha 1 de A. A solução em todos os casos se mantém entre 10 e 11 casas decimais exatas, salvo no caso de NM = 495 que o ICCG(0) deu nove.

Em todos os casos deste capítulo cada vez que utilizamos ICCG o vetor de chute inicial é o vetor nulo.

TABELA No. 4

N = 15

ICCG	M	NM	No. de iterações
ICCG(0)	5	75	11
ICCG(2)	5	75	9
ICCG(3)	5	75	10
ICCG(-3)	5	75	13
ICCG(-7)	5	75	18
ICCG(0)	17	255	32
ICCG(2)	17	255	21
ICCG(3)	17	255	26
ICCG(-3)	17	255	43
ICCG(-7)	17	255	27
ICCG(0)	33	495	47
ICCG(2)	33	495	24
ICCG(3)	33	495	45
ICCG(-3)	33	495	121
ICCG(-7)	33	495	46
ICCG(0)	65	975	68
ICCG(2)	65	975	24
ICCG(3)	65	975	56
ICCG(-3)	65	975	487
ICCG(-7)	65	975	91

Fomos além da tabela No. 4 e pegamos a matriz de ordem 2015 (N = 31, M = 65) com as mesmas condições com que construímos esta tabela e obtivemos estes resultados:

O ICCG(0) deu 518 iterações, o ICCG(2) alcançou 153

iterações, o ICCG(3) atingiu 339 iterações, o ICCG(-3) conseguiu 568 iterações e o ICCG(-7) chegou a fazer 97 iterações. Cada um alcançou pelos menos 8 casas decimais exatas, salvo o ICCG(-7) que deu 6. Para o ICCG(-7) atingir a mesma precisão tivemos que lhe fixar o critério de parada em 10^{-8} em vez de 10^{-6} e assim obtivemos 108 iterações.

4.3 ESTUDO ESPECTRAL DE ALGUMAS MATRIZES PRÉ-CONDICIONADAS

Visto que o cálculo de autovalores de uma matriz é um algoritmo muito caro computacionalmente, decidimos, visando fazer um estudo espectral de algumas matrizes pré-condicionadas, pegar o sistema Ax = b com ordem de A igual a 150 e b como nos casos anteriores. Na tabela No. 5 colocamos o número de iterações para resolver o sistema (ITER), menor e maior auto-valores da matriz pré-condiciona da: $\lambda min\ (\tilde{A})$ e $\lambda max\ (\tilde{A})$ e também pusemos o número de condição da matriz pré-condicionada $K(\tilde{A})$.

TABELA No. 5

N = 15, M = 10, $K(A) = 0,1033709 \times 10^{7}$

NM = 150

ICCG	ITER.	$\lambda min\ (\tilde{A})$	$\lambda max\ (\tilde{A})$	$K(\tilde{A})$
ICCG(0)	21	0,2476	2,0536	8,2908
ICCG(2)	14	0,4799	1,8028	3,7563
ICCG(3)	17	0,3978	2,2108	5,5556
ICCG(-3)	23	0,1834	6,0122	3,2770
ICCG(-7)	20	0,5163	1,6641	3,2228

Utilizamos sempre, para o cálculo dos auto-valores neste capítulo, uma rotina que calcula todos os auto-valores de uma matriz

real e simétrica por redução de Householder e o algoritmo QL [28]. Essa rotina, F02AAF, se encontra disponível na biblioteca NAG (Numerical Algorithms Groups) instalada no VAX da UNICAMP

A tabela No.5 nos diz ainda que ao passar do sistema Ax = b ao sistema $\tilde{A}y = \tilde{b}$ o número de condição da matriz em questão se reduz consideravelmente. Isto deixa prever que os auto-valores da matriz pré-condicionada se juntam mais.

Na tabela No.6 exibimos os auto-valores da matriz pré-condicionada quando resolvemos o sistema Ax = b com A de ordem 90 e usando ICCG(0). Lembramos que neste caso K(A) = 0.1039039 x 10^7. Enquanto, segundo a tabela No.6, $K(\tilde{A})$ = 3,7435.

TABELA No. 6

AUTO-VALORES DA MATRIZ PRÉ-CONDICIONADA

0.4479627031322158	0.5753976823723465	0.7333834516595589
0.9461145223823008	0.9677188112987500	0.9740936771830067
0.9865899734447878	0.9949143652282402	0.9957190857956069
0.9973024652773482	0.9974666810850144	0.9985370244238822
0.9989620817407342	0.9990721434917120	0.9994278070131190
0.9995642907142579	0.9996611031348182	0.9996920716392482
0.9997895859707906	0.9998132497738346	0.9998407348620523
0.9998504744215039	0.9998560282177615	0.9998647703686030
0.9998816193774478	0.9999166246696718	0.9999250213860755
0.9999294791979163	0.9999348984011355	0.9999439184017629
0.9999540169902654	0.9999629959513538	0.9999643738353435
0.9999667779638563	0.9999984660840078	0.9999986212361197
0.9999996065845255	0.9999997970713921	0.9999999995592053
0.9999999999999987	0.9999999999999995	0.9999999999999995
1.000000000000000	1.000000000000000	1.000000000000000
1.000000000000000	1.000000000000000	1.000000000000000
1.000000000000000	1.000000000000000	1.000000000000001
1.000000000474140	1.000000055064155	1.000000260370386

1.000000798116940	1.000001863190688	1.000008499021923
1.000016200763876	1.000035389548189	1.000038076317347
1.000045004418188	1.000049696930789	1.000056717203054
1.000061200117359	1.000070418219751	1.000097659170203
1.000112253883332	1.000121621037785	1.000145375455194
1.000164160352738	1.000174023451671	1.000203621459794
1.000224633450852	1.000283134731680	1.000328646880011
1.000403834029698	1.000569474262857	1.000592445354540
1.000916491039479	1.000992560642719	1.001549332856819
1.002082249709986	1.002833346900283	1.005430422829039
1.006328500956489	1.018730155094581	1.045187552152834
1.061379924759575	1.375571918377778	1.676986919974718

O NÚMERO DE CONDIÇÃO DA MATRIZ PRÉ-CONDICIONADA É: 0.3743586E +01.

Feita a decomposição de Choleski completa de A nós quisemos saber o que acontecia se, ao resolver o sistema Ax = b pelo método dos gradiente conjugados, usássemos como pré-condicionador a decomposição incompleta $C = LL^T$ tal que L^T tivesse como diagonais não nulas: num primeiro passo a diagonal principal de G^T, num segundo a diagonal principal e a segunda diagonal não nula de G^T, num terceiro a diagonal principal e a segunda e a terceira diagonais não nulas de G^T e assim até chegar à última diagonal não nula de G^T, em cujo caso será $C = LL^T = A$. Foi o interesse por resolver este assunto que nos levou a descobrir o ICCG(-7), no qual L^T só tem como diagonais não nulas as quatro primeiras diagonais não nulas de G^T.

Por simplicidade escolhemos A de ordem 90(N = 15, M = 6). Na tabela No.7 apresentamos os resultados para cada grupo de diagonais (D), colocamos o número de condição da matriz pré-condicionada, $K(\tilde{A})$, e

TABELA No. 7

N = 15 , M = 6

D	ITER.	K(Ã)
3	91	$0,1032x10^{9}$
4	18	3,0100
5	18	3,0100
6	19	3,0100
7	19	3,0100
8	19	3,0101
9	19	3,0099
10	19	3,0102
11	20	3,1156
12	20	3,5716
13	20	4,5886
14	21	6,6587
15	22	10,8360
16	91	$0,3830x10^{12}$
17	91	$0.9497x10^{11}$
18	91	$0.9415x10^{6}$
19	1	1,0

número de iterações (ITER) para resolver o sistema com b como nos casos anteriores. A precisão da solução em cada caso foi de pelo menos oito casas decimais exatas exceto quando usamos três diagonais em que o erro máximo foi de 1.3943; dezesseis diagonais que o erro máximo ficou em 1.7877, dezessete diagonais com erro máximo de 0.944 e dezoito diagonais alcançando erro máximo de 0.000147. O caso para dezenove diagonais era de se esperar pois, neste caso, o pré-condicionador é a decomposição de Choleski completa de A, ou seja, $C = LL^{T} = A$.

CAPÍTULO V

CONCLUSÕES E DESENVOLVIMENTOS FUTUROS

Muitas conclusões deste trabalho foram tiradas à medida em que íamos avançando no material; outras estão implícitas nas tabelas do capítulo IV. No entanto, queremos aqui ressaltar algumas e dar ênfase em outros aspectos que consideramos de importância.

A normalização da matriz A do sistema aumentou o número de iterações e a precisão para resolver o sistema na aplicação dos diferentes ICCG e permitiu estudar com simplicidade as diagonais não nulas de G^T na decomposição de Choleski de A ($A = GG^T$). Assim, segundo a tabela No. 4, capítulo IV, o ICCG (2) fez 24 iterações para resolver o sistema com a matriz normalizada (NM = 495), enquanto noutro teste fez 20 iterações sem normalizar a matriz e deixando fixas as outras condições. Os erros na solução foram 0.1722×10^{-11} e 0.2059×10^{-7} respectivamente.

Aqueles ICCG deram bons resultados, na solução dos sistemas dos exemplos lá expostos, porque os pré-condicionadores escolhidos, baseados numa análise das diagonais não nulas de G^T, eram uma boa aproximação de A. Além disso, para A de tamanhos razoáveis como os da tabela No. 4, ICCG(0), ICCG(2), ICCG(3) e ICCG(-7) convergem rapidamente. O ICCG (-3) deixa prever convergência lenta para tamanhos de A maiores que os da tabela No. 4. Já o teste feito para NM = 2015 permite conjecturar que o ICCG(2) e o ICCG(-7) convergem mais rapidamente para tamanhos de A muito grandes, além dos da tabela No. 4. Fomos mais longe ainda e pegamos A de ordem 4050 (N = 50, M = 90) e, com critério de parada norma do gradiente menor que 10^{-6}, obtivemos para o ICCG(-7) só 148 iterações dando nove casas decimais exatas na solução do sistema.

O teste que levou a construir a tabela No. 7 diz, entre outras coisas, que nem sempre, conservar a estrutura de A ou introduzir mais diagonais na decomposição incompleta de A, é o mais eficiente do

ponto de vista computacional, pois com um estudo tal como esse é possível obter o mesmo grau de precisão e um mínimo de operações. A este respeito, por exemplo, comparamos o ICCG(2) e o ICCG(-7) tomando os dados da tabela No. 4 para NM = 495 e nos propusemos a saber o número total de operações (produtos e divisões) que efetua cada um para resolver o sistema $(LL^T)h = g$, sendo genéricamente LL^T o respectivo pré-condicionador em cada um dos casos. A rotina que no método ICCG(2) resolve $(LL^T)h = g$ é chamada SISTEMA 2 e no ICCG(-7), SISTEMA 7. Delas tiramos estes resultados: o ICCG(2) emprega 26 NM - 18N - 58 operações por iteração e o ICCG(-7), 8NM - 12 operações por iteração. Com isto o ICCG(2) faz aproximadamente 608NM operações no total para resolver o sistema enquanto o ICCG(-7) efetua aproximadamente 369 NM operações. É bom também observar aqui a economia no número de operações das rotinas SISTEMA 2 e SISTEMA 7, já que resultados conhecidos mostram que para resolver o sistema $(LL^T)h = g$, considerando L e L^T como matrizes triangulares chei- as, o número de operações por iteração é da ordem de NM^2, onde NM é a ordem de A e portanto de L e L^T. A causa dessa poupança operacional é que nas rotinas mencionadas pegamos só as diagonais não nulas tanto de L quanto de L^T.

Outra observação pertinente é que, ao tomarmos os ICCG em conjunto, tivemos que fazer a decomposição de Choleski Completa de A e daí extraímos a decomposição incompleta para cada um. No entanto se considerarmos o ICCG(-7) isolado poderemos obter ainda mais economia evitando fazer a decomposição completa empregando fórmulas semelhantes às obtidas por Meijerink [19, p. 138].

Lembremos ainda que um ponto fraco no método de Kershaw [14] é a troca dos ℓ_{ii} por algum número positivo se ℓ_{ii} for ≤ 0. Em nossas experiências numéricas feitas com ordem de A de 21 até 4050 só encontramos dois casos onde este efeito se apresentava e só uma vez por caso: NM = 715 (N = 11, M = 65) e NM = 2015 (N = 31, M = 65). Neste último, por exemplo, só para i = 2015 obtivemos $\ell_{ii} < 0$, ℓ_{ii} = -0,7267492E-04, o qual foi trocado por 1.0 e o resultado foi colocado no capítulo IV (Veja ICCG (0), NM = 2015). Daqui podemos conjecturar que

aplicando o ICCG (0) ao sistema em estudo a matriz erro não sofre muita perturbação. Isto pode ser uma causa dos bons resultados do ICCG(0) no capítulo IV.

Na solução do sistema real, ou seja, usando as expressões (1.11) do capítulo I, com os métodos ICCG(2) e ICCG(-7) para NM = 495 e critério de parada norma gradiente menor que 10^{-8} obtivemos para o ICCG (2) 21 iterações com erro máximo no resíduo de 0.1116×10^{-7} e para o ICCG(-7) 40 iterações e erro máximo no resíduo de 0.3884×10^{-7}. Aqui destacamos que a rotina BANDA 7 que calcula o produto de matriz por vetor tinha sido testada com o vetor u; u(i) = i, i=1,...,NM que aumenta consideravelmente o tamanho das entradas de A (Veja tabela No. 2, capítulo IV). Daí a confiabilidade no cálculo do resíduo. Aliás, essa margem de erro no resíduo se manteve para ordem de A muito maior, por exemplo, NM = 2015.

Em relação aos desenvolvimentos futuros do trabalho da tese temos dois a curto prazo que são: "Vetorização" da matriz do sistema e o uso de pré-condicionadores por blocos, e outros dois que seguirão aos primeiros: a utilização de pré-condicionadores polinomiais e o emprego de algoritmo de Lanczos aprimorado devido a Shao [25] para resolver sistemas de equações lineares mal-condicionados.

Visto que já temos armazenados os blocos AA, BB, A1, entendemos por "Vetorização" de A o armazenamento só das diagonais não nulas de cada bloco (lembramos que cada bloco é BANDA 7). Pela simetria, para cada bloco é suficiente armazenar só 4 diagonais. Temos pensado, para facilitar a implementação computacional, armazenar numa mesma matriz R, N x 12, estes vetores colocando zero nos poucos espaços vazios. Isto levaria a modificar a rotina TRIBLOCO que constrói os três blocos e fazer outra que identifique um elemento de cada bloco com outro correspondente de R. Este trabalho está bem avançado, só falta fazer os testes e as correções do caso. A preocupação nesta parte não é só do ponto de vista estético, que incontestávelmente a tem, mas também econômico. De fato, assim conseguimos considerável economia de memória para aqueles casos que precisarem de malhas mais refinadas.

O pré-condicionamento por blocos é uma motivação que encontramos no artigo de Concus, Golub e Meurant [6]. Eles usam pré-condicionadores por blocos no método dos gradientes conjugados para resolver sistemas lineares tridiagonais por blocos positivos definidos que surgem da discretização de problemas com valores de fronteira de equações parciais elípticas. Um papel importante nestas técnicas desempenha a decomposição de Choleski por blocos e a aproximação da inversa de uma matriz tridiagonal diagonalmente dominante que garante que os elementos da inversa decrescem estritamente conforme se afastam da diagonal principal. Em nosso caso já que não contamos com a propriedade da dominância e que teremos que aproximar a inversa de matrizes de banda maior, tentaremos, por um lado, uma espécie de adaptação do tipo Kershaw [14] para compensar a falta de dominância e por outro lado, usaremos os métodos de Asplund [1] para o cálculo de inversas de matrizes de banda $p \geq 1$ e as idéias de Bevilacqua [4] para reduzir o armazenagem no cálculo de inversas de matrizes banda. Segundo os resultados numéricos obtidos em [4] os pré-condicionadores por blocos levam vantagem sobre os correspondentes pré-condicionadores pontuais em menor número de iterações e trabalho por ponto (Work/n, onde n é a ordem da matriz do sistema). Nesta parte temos o trabalho bibliográfico quase completo.

Um próximo passo em nossas pesquisas é o uso dos pré-condicionadores polinomiais no método dos gradientes conjugados. Eles levam implícito o fascínio dos computadores vetoriais e paralelos. O método que seguiremos será o de Saad [24], usado para resolver sistemas lineares esparsos que surgem de discretização de equações diferenciais parciais. A matriz A do sistema é simétrica positiva definida. As técnicas são baseadas em polinômios de quadrados mínimos no intervalo $[0,b]$, sendo b uma estimativa de Gershgorin do maior autovalor, isto é não precisamos calcular os autovalores de A como noutras versões. Na sua essência, dado um sistema simétrico $Ax = b$, o princípio de pré-condicionamento polinomial consiste em resolver o

sistema (pré-condicionado) p(A)Ax = p(A)b, onde p é algum polinômio normalmente de grau pequeno. A escolha de p é feita de maneira que a matriz p(A)A tenha uma distribuição adequada de autovalores, o que quer dizer que o método dos gradientes conjugados aplicado ao sistema pré-condicionado converge rapidamente. Alías, p é escolhido de forma a minimizar a norma L_2 com uma certa função peso w definida em [0, b] contendo o espectro de A. Neste trabalho estamos ainda em estado rudimentar.

Finalmente, norteados pelo método de Shao [25], queremos aplicar sua versão aprimorada do Algoritmo de Lanczos. Ele observa que a estimativa do resíduo para resolver sistemas lineares pelos algoritmos de Lanczos usuais é confiável para sistemas bem-condicionados mas não é este o caso para sistemas mal-condicionados. Assim, ele propõe um algoritmo que corrige esta imperfeição e mostra bons resultados numéricos obtidos. Nesta parte nosso trabalho está apenas em seu começo.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Asplund, E., Inverses of Matrices $\{a_{ij}\}$ which satisfy $a_{ij} = 0$ for $j > i + p$, Math. Scand., Vol. 7, (1959) 57-60.

[2] Axelsson, O., A class of Iterative Methods for Finite Element Equations, Comput. Methods Appl. Mech. Engrg., Vol. 9, (1976) 123-137.

[3] Axelsson, O. and Barker, V.A., Finite Element Solution of Boundary Value Problems, Academic Press, Inc., 1984.

[4] Bevilacqua, R., Lotti, C. and Romani, F., Storage Compression of Inverses of Band Matrices, Computers Math. Applic., Vol. 20, (1990) 1-11.

[5] Céa, J., Optimisation Théorie et Algorithmes, Dunod, Paris, 1971.

[6] Concus, P., Golub, G.H. and Meurant, Block Preconditioning for the Conjugate Gradient Method, SIAM J. Sci. Stat. Comp., Vol. 6, (1985) 220-252.

[7] Forsythe, G.E. and Moler, C.B., Computer Solution of Linear Algebraic Systems, Prentice-Hall, Englewood Cliffs, N.J., 1967.

[8] Forsythe, G.E., Malcolm, M.A. and Moler, C.B., Computer Methods for Mathematical Computations, Prentice-Hall, Englewood Cliffs, N.J., 1977.

[9] Golub, G.H. and Van Loan, C.F., Matrix Computations, The John Hopkins University Press, Baltimore and London, 1990.

[10] Hestenes, M.R. and Stiefel, E., Methods of Conjugate Gradients for Solving Systems, J.Res. Nat. Bur. Stand., Vol. 49, (1952) 409-436.

[11] Isaacson, E. and Keller, H.B., Analysis of Numerical Methods, John Wiley & Sons, Inc., New York, 1966.

[12] Jennings, A., Influence of the Eigenvalue Spectrum on the Convergence Rate of the Conjugate Gradient Methods, J.Inst. Maths. Applics., Vol. 20, (1977) 61-72.

[13] Johnson, C., Numerical Solution of Partial Differential Equations by the Finite Element Methods, Cambridge University Press, Cambridge, 1987.

[14] Kershaw, D.S., The Incomplete Cholesky-Conjugate Gradient Method for the Iterative Solution of Systems of Linear Equations, J. Comp. Phys., Vol. 26, (1978) 43-65.

[15] Ky Fan (Notre Dame, Indiana), Note on M-matrices, Quart. J. Math. Oxford Ser. 2, Vol. 11 (1960) 43-49.

[16] Lopes, V.L.R., Solução, por Elementos Finitos, de Equações de Difusão Lineares, via Princípios Extremos Duais, Tese de Doutorado, ICM-USP, São Carlos, São Paulo, 1988.

[17] Luenberger, D.G., Linear and Nonlinear Programming Second Edition, Addison-Wesley, 1984.

[18] Meijerink, J.A. and Van der Vorst, H.A., An Iterative Solution Method for Linear Systems of Which the Coefficient Matrix is a Symmetric M-matrix, Math. Comp., Vol. 31., (1977) 148-162.

[19] Meijerink, J.A. and Van der Vorst, H.A., Guidelines for the Usage of Incomplete Decompositions in Solving Sets of Linear Equations as they Occur in Practical Problems, J.Comp. Phys., Vol. 44, (1981) 134-155.155.

[20] Noble, B. and Daniel, J.W., Applied Linear Algebra, Prentice-Hall, Englewood Cliffs, N.J., 1988.

[21] Noble, B. and Sewell, M.J., On dual Extremum Principles in Applied Mathematics, J.Inst. Maths. and its Applics, Vol. 9, (1972) 123-193.

[22] Ortega, J.M., Numerical Analysis, Academic Press, New York, 1972.

[23] Prenter, P.M., Splines and Variational Methods, John Wiley & Sons, New York, 1975.

[24] Saad, Y., Practical use of Polynomial Preconditionings for the Conjugate Gradient Methods, SIAM J.Sci. Stat. Comp., Vol. 6, (1985) 865-881.

[25] Shao, P.L., An Improved Lanczos Algorithm for Solving ill-conditioned Linear Equations, Computers Math. Applic., Vol. 20, (1990) 25-33.

[26] Stewart, G.W., The Convergence of the Method of Conjugate Gradients at Isolated Extreme Points of the Spectrum, Numer. Math., Vol. 24, (1975) 85-93.

[27] Varga, R.S., Matrix Iterative Analysis, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962.

[28] Wilkinson, J.H. and Reinsch, C., Handbook for Automatic Computation, Vol. 2, Linear Algebra, pp. 212-226 and 227-240, Springer Verlag, 1971.

[29] Young, D.M., Iterative Solution of Large Linear Systems, Academic Press, New York, 1971.

[30] Zago, J.V., Approximate Solution of Generalized Hamiltonian Equations with Applications, Ph.D. Thesis, U. of Wisconsin, Madison, 1976.