NOVOS RESULTADOS SOBRE FÓRMULAS SECANTES E APLICAÇÕES

Tese de Doutorado

Autor: Mário César Zambaldi
DMA–IMECC–UNICAMP

Orientador: Prof. Dr. José Mario Martínez
DMA-IMECC-UNICAMP

NOVOS RESULTADOS SOBRE FÓRMULAS SECANTES E APLICAÇÕES

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pelo Sr. Mário César Zambaldi e aprovada pela Comissão Julgadora.

Campinas, 04 de junho de 1993

Prof. Dr. José Mario Martínez
Orientador

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

AGRADECIMENTOS

a Mario Martínez pela judiciosa orientação e confiança na elaboração deste trabalho.

aos professores e funcionários do Depto. de Matemática Aplicada pela colaboração e pela convivência durante este período.

aos companheiros de pós-graduação pelo apoio e incentivo, especialmente a Sandra, Daniel, Fermin e Clarice.

a Ilza, Vanessa, Vinícius e Leonardo pela imensa compreensão.

a FAPESP, Fundação de Amparo a Pesquisa do Estado de São Paulo, pelo criterioso acompanhamento do trabalho e pelo suporte financeiro.

Conteúdo

1 Introdução **2**

2 Métodos Secantes para Sistemas Não Lineares com Termos Não Diferenciáveis **7**

2.1 INTRODUÇÃO 7
2.2 O MÉTODO BÁSICO 8
2.3 CONVERGÊNCIA LOCAL 11
2.4 CONVERGÊNCIA COM TAXA IDEAL 19
2.5 CONCLUSÕES 29

3 Atualização de uma coluna do Jacobiano Inverso **30**

3.1 INTRODUÇÃO 30
3.2 O MÉTODO ICUM 31
3.3 RESULTADOS DE CONVERGÊNCIA 33
3.4 IMPLEMENTAÇÃO COMPUTACIONAL 39
3.5 EXPERIMENTOS NUMÉRICOS 41
3.6 CONCLUSÕES 43

4 Métodos de Newton Inexatos com Precondicionadores Secantes **44**

4.1 INTRODUÇÃO 44
4.2 PRECONDICIONADORES SECANTES 47
4.3 ALGORITMOS 52
4.4 O GMRES 58
4.5 OS PROBLEMAS 62
4.6 EXPERIMENTOS NUMÉRICOS 64

4.7 CONCLUSÕES 71

Capítulo 1

Introdução

Os Sistemas de Equações não Lineares aparecem em muitos problemas da vida real. São os casos, por exemplo, de problemas de Mecânica dos fluidos, fluxo de cargas em redes de transmissão de energia elétrica e simulações numéricas de reservatórios. Um trabalho recente de Moré [47] relaciona uma coleção de problemas práticos, oriundos de diversas áreas e aplicações, cuja formulação conduz a um sistema de equações não lineares.

O Problema

O problema típico em que estamos interessados é o seguinte:

Dada uma função, $F : \mathbb{R}^n \to \mathbb{R}^n$, $F = (f_1, \ldots, f_n)^T$, buscamos obter a solução de

$$F(x) = 0. \tag{1.1}$$

Denotaremos a matriz de derivadas parciais de F, a matriz Jacobiana, ou simplesmente o Jacobiano, por $J(x)$. Portanto,

$$J(x) \equiv F'(x) \equiv \begin{pmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_n(x)^T \end{pmatrix} \equiv \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \cdots & \frac{\partial f_n}{\partial x_n}(x) \end{pmatrix}$$

Frequentemente, em problemas reais, n é relativamente grande e a matriz Jacobiana apresenta uma estrutura esparsa. Isto significa que a maioria dos elementos de $J(x)$ são nulos. Muitas vezes podemos tirar vantagem do "padrão de esparsidade" de $J(x)$ para obter eficientes algoritmos de resolução para (1.1).

Vários métodos têm sido propostos para a resolução numérica de sistemas não lineares. A maioria e os mais conhecidos deles são métodos "locais". Um método local é um esquema iterativo que converge se a aproximação inicial está "suficientemente próxima" de uma solução particular. Felizmente, em muitos problemas práticos, o domínio de convergência dos métodos locais é suficientemente grande. A conotação deste trabalho é o tratamento de métodos locais.

Um aspecto essencial desses métodos de resolução é a taxa de convergência, que nos diz algo sobre a velocidade assintótica do processo de convergência à solução.

O método de resolução mais conhecido, e que serve de base para obtenção de outros métodos eficientes, é o método de Newton. Dada uma estimativa inicial x_0 da solução de (1.1), este método considera, a cada iteração, a aproximação

$$F(x) \approx L_k(x) \equiv F(x_k) + J(x_k)(x - x_k) ,$$

e obtém x_{k+1} como a solução do sistema linear $L_k(x) = 0$. Esta solução existe e é única se $J(x_k)$ é não singular. Portanto, uma iteração de Newton é descrita por

$$J(x_k)s_k = -F(x_k) \tag{1.2}$$

$$x_{k+1} = x_k + s_k \tag{1.3}$$

A cada iteração do Método de Newton, devemos calcular o Jacobiano, $J(x_k)$ e resolver o sistema linear (1.2). Usando técnicas modernas de diferenciação automática [31], podemos calcular $F(x)$ e $J(x)$ de maneira econômica. Se no lugar de Jacobiano verdadeiro em (1.3) usamos uma aproximação por diferenças de $J(x_k)$, que geralmente envolve um alto custo computacional, obtemos o método de Newton com diferenças finitas, cujas propriedades de convergência são muito semelhantes às do Método de Newton.

O maior problema do método de Newton está na resolução do sistema linear (1.2). Se n é pequeno, o sistema linear pode ser resolvido usando fatoração LU, com pivotamento parcial ou a fatoração QR [25]. Se n é grande, esta tarefa envolve um alto custo computacional. Em muitas situações, entretanto, onde $J(x_k)$ é esparsa, podemos recorrer

a técnicas especiais considerando a estrutura da matriz. Neste caso, os fatores no processo de decomposição são gerados de maneira que haja o mínimo enchimento (fill-in), no contexto de esparsidade, sem comprometer a estabilidade numérica. O principal resultado de convergência relativo ao método de Newton é dado pelo seguinte teorema.

Teorema 1.1

Suponhamos $F : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^n, \Omega$ um conjunto aberto e convexo, $F \in C^1(\Omega), F(x_*) = 0, J(x_*)$ não singular, e que existam $L, p > 0$ tais que para todo $x \in \Omega$,

$$||J(x) - J(x_*)|| \leq L||x - x_*||^p \,. \tag{1.4}$$

Então existe $\varepsilon > 0$ tal que se $||x_0 - x_*|| \leq \varepsilon$, a seqüência $\{x_k\}$ gerada por (1.3) - (1.4) está bem definida, converge para x_* e satisfaz

$$||x_{k+1} - x_*|| \leq c||x_k - x_*||^{p+1} \,. \tag{1.5}$$

Prova.

Veja Ortega e Rheinboldt [48]. ■

A propriedade (1.5) (convergência quadrática se $p = 1$) depende da condição de Hölder (1.4). Sem essa condição, a convergência é apenas superlinear.

Os métodos quase-newtonianos, alternativamente, buscam, através de aproximação para o Jacobiano, obter algo próximo da propriedade atrativa de convergência do método de Newton e, por outro lado, superá-lo em sua maior deficiência, resolvendo o sistema linear (1.2) da maneira mais adequada possível. São caracterizados pelo processo iterativo,

$$x_{k+1} = x_k - B_k^{-1} F(x_k) \,. \tag{1.6}$$

A classe de métodos quase-newtonianos mais bem sucedida é a dos métodos secantes. Nestes métodos, escolhem-se as matrizes de aproximação para o Jacobiano, de modo a satisfazer a Equação Secante:

$$B_{k+1} s_k = F(x_{k+1}) - F(x_k) \,. \tag{1.7}$$

Esta equação significa que o "modelo linear" $L_{k+1}(x) = F(x_{k+1}) + B_{k+1}(x - x_{k+1})$ interpola F em x_k e x_{k+1}. Nas últimas décadas, muito se desenvolveu sobre teoria e prática dos métodos secantes. Do ponto de vista teórico, buscam-se métodos superlinearmente convergentes, onde a seqüência de pontos é gerada de modo a satisfazer:

$$\lim_{k\to\infty} \frac{||x_{k+1} - x_*||}{||x_k - x_*||} = 0 .$$

Recentemente, teorias gerais foram desenvolvidas para os métodos secantes. A mais ampla delas é a teoria LSCU (Least Change Secant Update) de Martínez [40], que envolve uma diversidade de métodos conhecidos.

Uma outra maneira de atacar o nosso problema, é investir na resolução dos sistemas lineares (1.2) através do uso de métodos iterativos lineares. Isto caracteriza os métodos de Newton Inexatos. Esta técnica também tem a filosofia de economizar na resolução de (1.2), conservando boas propriedades de convergência.

Martínez [43], conjugou as idéias expostas acima, onde procura acelerar os métodos de Newton-Inexatos através do uso de fórmulas secantes precondicionadoras.

O objetivo deste trabalho é fornecer novos resultados sobre fórmulas secantes e aplicações. Neste sentido, desenvolvemos os seguintes temas:

(*i*) Extensão da teoria geral LCSU de Martínez para sistemas não lineares com termos não diferenciáveis.

(*ii*) Introdução de uma nova fórmula secante, dando origem a um novo método, que apresenta desempenho muito satisfatório para uma importante classe de problemas.

(*iii*) Implementação prática dos precondicionadores secantes.

Os ítens (*i*), (*ii*) e (*iii*), são desenvolvidos nos capítulos 2, 3 e 4 respectivamente.

Apesar da unidade temática desenvolvida neste trabalho, procuramos dar um caráter de independência entre os capítulos. Neste sentido, o leitor que estiver interessado em um dos temas, não precisará, a rigor, de uma leitura sistemática dos outros, uma vez que estabelecemos conexões necessárias por referências rápidas.

Capítulo 2

Métodos Secantes para Sistemas Não Lineares com Termos Não Diferenciáveis

2.1 INTRODUÇÃO

As fórmulas secantes para aproximar Jacobianos, no contexto de resolução de sistemas não lineares, não são recentes (Ortega e Rheinboldt [48] Cap.8). A aproximação do Jacobiano usando os $n+1$ últimos pontos interpolantes de um método iterativo é denominada Fórmula Secante Seqüencial (Barnes[3],Wolfe[58], Gragg-Stewart[29], Martínez [36]). Esta fórmula apresenta problemas de estabilidade e falta de manutenção de estrutura, pelo qual não é muito usada atualmente. A partir de 1959 (Davidon [12]), popularizaram-se as fórmulas secantes baseadas apenas na última "equação secante" : $B_{k+1}(x_{k+1} - x_k) = F(x_{k+1}) - F(x_k)$. Broyden, Dennis e Moré [9], Dennis e Moré [17] e Dennis e Walker [19] desenvolveram a teoria LCSU(Least Change Secant Update) que envolve as mais conhecidas fórmulas secantes. Martínez [40] desenvolveu a teoria LCSU mais amplamente. Tudo isto foi desenvolvido considerando sistemas não lineares diferenciáveis.

Neste capítulo será desenvolvida a extensão da teoria geral de Martínez [40], a problemas parcialmente diferenciáveis.

Consideremos o problema de resolver

$$F(x) = 0 \tag{2.1}$$

onde $F : \Omega \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$, Ω aberto, $F = F_1 + F_2$ e $F_1 \in C^1(\Omega)$. Denotaremos $J(x) = F_1'(x)$ para todo $x \in \Omega$.

Estamos interessados em métodos de resolução de (2.1). Dado $x_k \in \Omega$, a aproximação corrente para a solução de (2.1), a aproximação seguinte x_{k+1} será dada por

$$x_{k+1} = x_k - B_k^{-1} F(x_k) \,, \tag{2.2}$$

onde B_k é uma matriz não singular convenientemente escolhida. O caso onde $B_k = J(x_k)$ foi estudado por Zabrejko e Nguen [62], Chen e Yamamoto [11], Yamamoto [57] e Yamamoto e Chen [60]. Mais recentemente, Chen [10] analisou o caso onde as matrizes B_k são geradas usando a fórmula de Broyden [9,17,18].

2.2 O MÉTODO BÁSICO

Descreveremos o algoritmo geral para resolução de (2.1) seguindo as linhas de [40].

Seja X um espaço vetorial de dimensão finita. Para todo $x, z \in \Omega$, consideremos $\|\cdot\|_{xz}$ uma norma em X, associada a algum produto escalar $\langle \, , \rangle_{xz}$. O operador projeção no conjunto $\mathcal{C} \subset X$ relativo a $\|\cdot\|_{xz}$ será denotado por $P_{\mathcal{C},xz}$.

Para todo $x, z \in \Omega$ consideremos $V(x,z) \subset X$ uma variedade linear. Seja $\mathcal{D} \subset X$ um conjunto aberto e $\varphi : \Omega \times \mathcal{D} \longrightarrow \mathbb{R}^{n \times n}$ uma função contínua.

Para um ponto arbitrário $x_0 \in \Omega$, $E_0 \in \mathcal{D}$, $B_0 = \varphi(x_0, E_0)$, consideramos a seqüência gerada por

$$x_{k+1} = x_k - B_k^{-1} F(x_k) \,, \tag{2.3}$$

onde

$$B_{k+1} \in \{\varphi(x_k, E_k),\, \varphi(x_{k+1}, E_k),\, \varphi(x_k, E_{k+1}),\, \varphi(x_{k+1}, E_{k+1})\} \,, \tag{2.4}$$

$$E_{k+1} \in \{E_k,\ P_k(E_k)\} \tag{2.5}$$

e

$$P_k \equiv P_{V(x_k,x_{k+1}),x_kx_{k+1}} \tag{2.6}$$

para todo $k = 0, 1, 2, \ldots$.

Geralmente, consideramos somente o mais interessante dos casos, onde

$$B_{k+1} = \varphi(x_{k+1}, E_{k+1}) \tag{2.7}$$

e

$$E_{k+1} = P_k(E_k) \tag{2.8}$$

para todo $k = 0, 1, 2, \ldots$.

Exemplo 2.1 O Método de Broyden

Na extensão do método de Broyden para o caso não diferenciável, considerado por Chen [10], temos $X = I\!R^{n\times n}$, $\|\cdot\|_{xz} = \|\cdot\|_F$ (a norma de Frobenius) para todo $x, z \in \Omega$,

$$V(x,z) = \{B \in X \mid B(z-x) = F_1(z) - F_1(x)\} \tag{2.9}$$

e

$$\varphi(x, B) \equiv B\ . \tag{2.10}$$

Portanto,

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T}{s_k^T s_k} \tag{2.11}$$

para todo $k = 0, 1, 2, \ldots$, onde $s_k = x_{k+1} - x_k$ e $y_k = F_1(x_{k+1}) - F_1(x_k)$. ([18],cap. 8).

Exemplo 2.2. O Método BFGS

Suponhamos que para todo $x, z \in \Omega$, a matriz $\tilde{J}(x,z)$ é simétrica e positiva definida, onde

$$\tilde{J}(x,z) = \int_0^1 J(x + t(z-x))dt\,. \tag{2.12}$$

Definimos $X = I\!R^{n\times n}$, e para todo $x, z \in \Omega$,

$$||E||_{xz} = ||L(x,z)^T E L(x,z)||_F\,,$$

onde $L(x,z)L(x,z)^T$ é a fatoração de Cholesky de $\tilde{J}(x,z)$,

$$\varphi(x,E) \equiv E^{-1}\ ,$$

$$V(x,z) = S \cap \{E \in X \mid E[F_1(z) - F_1(x)] = z - x\}\,,$$

e S e o subespaço das matrizes simétricas de $I\!R^{n\times n}$. Uma manipulação clássica dos cálculos mostra que [18]

$$\begin{aligned} E_{k+1} = B_{k+1}^{-1} &= E_k + \frac{(s_k - E_k y_k)s_k^T + s_k(s_k - E_k y_k)^T}{s_k^T y_k} \\ &- \frac{(s_k - E_k y_k)^T y_k s_k s_k^T}{(s_k^T y_k)^2}\ , \end{aligned}$$

onde s_k e y_k são definidos como no exemplo 2.1.

Exemplo 2.3. Métodos de Atualização Direta da Fatoração

Suponha que para todo $x, z \in \Omega$, $\tilde{J}(x,z) = \mathcal{A}(x,z)^{-1}\mathcal{R}(x,z)$, onde $\tilde{J}$ é definido por (2.12) e $(\mathcal{A}(x,z), \mathcal{R}(x,z)) \in S$, uma variedade linear em $X \equiv I\!R^{n\times n} \times I\!R^{n\times n}$. Para

todo $x, z \in \Omega$, $(A, R) \in X$, definimos $||(A,R)||^2_{xz} = ||A||^2_F + ||R||^2_F$. Ainda, para todo $x, z \in \Omega$,

$$V(x,z) = \{(A,R) \in S \mid R(z-x) - A[F_1(z) - F_1(x)] = 0\}$$

e

$$\varphi(x,(A,R)) \equiv A^{-1}R.$$

Este tipo de método é muito conveniente quando o Jacobiano da parte diferenciável de F admite um espaço de fatoração definido pela variedade S (Veja [40,41]).

2.3 CONVERGÊNCIA LOCAL

Nesta seção, provaremos que o método definido por (2.3)-(2.6) está bem definido e converge para a solução de (2.1), desde que o ponto inicial esteja numa vizinhança da solução e o parâmetro inicial E_0 esteja suficientemente próximo de um parâmetro ideal E_*. Precisamos de quatro hipóteses para demonstrar tal propriedade.

A primeira hipótese refere-se à função F.

Hipótese H1

Suponhamos que existam $x_* \in \Omega$ tal que $F(x_*) = 0$, $J(x_*)$ não singular e $L, p > 0$ tais que

$$|J(x) - J(x_*)| \leq L|x - x_*|^p \tag{2.13}$$

para todo $x \in \Omega$, onde $|\cdot|$ denota uma norma arbitrária em $\mathbb{R}^n$. Isto implica que

$$|F_1(z) - F_1(x) - J(x_*)(z-x)| \leq L|z-x|\sigma(x,z)^p \tag{2.14}$$

para todo $x, z \in \Omega$, onde $\sigma(x,z) = \max\{|x - x_*|, |z - x_*|\}$.

Suponhamos também que exista $\alpha > 0$ tal que

$$|F_2(x) - F_2(x_*)| \leq \alpha|x - x_*| \tag{2.15}$$

para todo $x \in \Omega$.

A segunda hipótese refere-se à função φ.

Hipótese H2

Existe $E_* \in D$ tal que $\varphi(x_*, E_*)$ seja não singular e

$$|I - \varphi(x_*, E_*)^{-1} J(x_*)| + \alpha|\varphi(x_*, E_*)^{-1}| = r_* < 1. \tag{2.16}$$

A terceira hipótese é fundamental dentro da teoria que estamos desenvolvendo. Ela afirma que as variedades $V(x, z)$ estão suficientemente próximas de E_*.

Hipótese H3

Seja $\|\cdot\|$ uma norma fixa em X, associada ao produto escalar $\langle , \rangle$, e $c_1 \geq 0$ uma constante. Para todo $x, z \in \Omega$ existe $E = E(x, z) \in V(x, z)$ tal que

$$\|E - E_*\| \leq c_1 \sigma(x, z)^p. \tag{2.17}$$

A última hipótese refere-se a relação entre diferentes normas em X.

Hipótese H4

Existe $q > 0$, $c_2 \geq 0$ tal que, para todo $x, z \in \Omega$, $E \in X$,

$$\|E\|_{xz} \leq [1 + c_2 \sigma(x, z)^q] \|E\| \tag{2.18}$$

e

$$||E|| \leq [1 + c_2\sigma(x,z)^q]||E||_{xz}\,. \tag{2.19}$$

O significado desta hipótese é que as normas $||E||_{xz}$ tornam-se muito próximas de $||E||$, quando x e z estão muito próximos de x_*.

Lema 2.1

Seja $r_1 \in (r_*, 1)$. Se F satisfaz a Hipótese H1 e φ, x_*, E_* satisfazem a Hipótese H2, então existe $\varepsilon_1 = \varepsilon_1(r_1) > 0$ tal que

$$|x - \varphi(x_*, E_*)^{-1}F(x) - x_*| \leq r_1|x - x_*| \tag{2.20}$$

sempre que $|x - x_*| \leq \varepsilon_1$.

Prova. Seja $\varepsilon_1 > 0$ tal que

$$r_1 \geq r_* + |\varphi(x_*, E_*)^{-1}|L\varepsilon_1^p\,. \tag{2.21}$$

Então, por H1, H2 e (2.21),

$$|x - \varphi(x_*, E_*)^{-1}F(x) - x_*|$$

$$\leq |x - \varphi(x_*, E_*)^{-1}J(x_*)(x - x_*) - x_*| + |\varphi(x_*, E_*)^{-1}||F(x) - J(x_*)(x - x_*)|$$

$$= |[I - \varphi(x_*, E_*)^{-1}J(x_*)](x - x_*)| + |\varphi(x_*, E_*)^{-1}|\,|F_1(x) + F_2(x) - J(x_*)(x - x_*)|$$

$$\leq |I - \varphi(x_*, E_*)^{-1}J(x_*)|\,|x - x_*| + |\varphi(x_*, E_*)^{-1}|\,|F_2(x) - F_2(x_*)|$$

$$+|\varphi(x_*, E_*)^{-1}|\,|F_1(x) - F_1(x_*) - J(x_*)(x - x_*)|$$

$$\leq [||I - \varphi(x_*, E_*)^{-1} J(x_*)| + \alpha|\varphi(x_*, E_*)^{-1}| + |\varphi(x_*, E_*)^{-1}|\ L|x - x_*|^p]|x - x_*|$$

$$\leq (r_* + |\varphi(x_*, E_*)^{-1}|L\varepsilon_1^p)|x - x_*| \leq r_1|x - x_*|\,.$$

Logo, a prova está completa. ■

O Lema 2.1 mostra que, sob as Hipóteses H1 and H2, a iteração "ideal" $x_{k+1} = x_k - \varphi(x_*, E_*)^{-1}F(x_k)$ é localmente convergente. O restante desta seção consiste em provar que a iteração principal (2.3)-(2.6) compartilha desta propriedade.

O próximo teorema mostra que se x e E estão suficientemente próximos de x_* e E_*, respectivamente, a aplicação $x \to x - \varphi(x, E)^{-1}F(x)$ aproxima o ponto x de x_* com uma taxa próxima de r_*.

Teorema 2.2

Seja $r \in (r_*, 1)$. Suponha que F, φ, x_*, E_* satisfaçam H1 e H2. Então, existem $\varepsilon_2 = \varepsilon_2(r)$, $\delta_2 = \delta_2(r) > 0$ tais que, para todo x, z, E satisfazendo $||x - x_*|| \leq \varepsilon_2$, $||z - x_*|| \leq \varepsilon_2$, $||E - E_*|| \leq \delta_2$, temos que, $||E||, |\varphi(z, E)|$ e $|\varphi(z, E)^{-1}|$ são uniformemente limitados, e

$$|x - \varphi(z, E)^{-1}F(x) - x_*| \leq r|x - x_*|\,. \tag{2.22}$$

Prova. Sejam $\varepsilon_3, \delta_3 > 0$ tais que $\varphi(z, E)^{-1}$ existe, sempre que $|z - x_*| \leq \varepsilon_3$, $||E - E_*|| \leq \delta_3$. Portanto, por compacidade, $||E||$, $|\varphi(z, E)|$ e $|\varphi(z, E)^{-1}|$ são uniformemente limitados para $|z - x_*| \leq \varepsilon_3$ e $||E - E_*|| \leq \delta_3$.

Seja $r_1 \in (r_*, r)$, $0 < \varepsilon_2 \leq \min\{\varepsilon_1(r_1), \varepsilon_3\}$, $0 < \delta_2 \leq \delta_3$ tais que

$$r_1 + |\varphi(z, E)^{-1} - \varphi(x_*, E_*)^{-1}|(|J(x_*)| + L\varepsilon_2^p + \alpha) \leq r \tag{2.23}$$

sempre que $|z - x_*| \leq \varepsilon_2$, $||E - E_*|| \leq \delta_2$. Se $|x - x_*| \leq \varepsilon_2$, temos, por (2.14) e (2.15), que

$$|F(x)| \leq |F_1(x) - F_1(x_*)| + |F_2(x) - F_2(x_*)|$$

$$\leq (|J(x_*)| + L|x - x_*|^p + \alpha)|x - x_*|$$

$$\leq (|J(x_*)| + L\varepsilon_2^p + \alpha)|x - x_*| \,. \tag{2.24}$$

Portanto, por (2.20), (2.24) e (2.23),

$$|x - \varphi(z, E)^{-1}F(x) - x_*|$$

$$\leq |x - \varphi(x_*, E_*)^{-1}F(x) - x_*| + |\varphi(z, E)^{-1} - \varphi(x_*, E_*)^{-1}|\,|F(x)|$$

$$\leq [r_1 + |\varphi(z, E)^{-1} - \varphi(x_*, E_*)^{-1}|(|J(x_*)| + L\varepsilon_2^p + \alpha)]|x - x_*|$$

$$\leq r|x - x_*| \,. \tag{2.25}$$

Isto completa a demonstração ■

Agora, estabeleceremos propriedades de deterioração limitada que são necessárias para provar convergência local de (2.3)-(2.6). Os princípios de deterioração limitada foram introduzidos por Dennis[13] e popularizados no trabalho de Broyden, Dennis e Moré [9]. A deterioração limitada, em nosso contexto, significa que a distância entre $P_{xz}(E)$ e E_*, não pode ser muito maior que a distância entre E e E_*, isto é, a possível deterioração nas aproximações E em relação a E_*, ocorre de maneira controlada.

Lema 2.3

Sejam F, φ, V, E_* satisfazendo as Hipóteses H1 a H4. Suponhamos que Ω' é um subconjunto limitado de Ω. Então existem constantes positivas c_3, c_4 tal que para todo $x, z \in \Omega'$, $E \in X$,

$$||P_{xz}(E) - E_*|| \leq [1 + c_4\sigma(x,z)^q]||E - E_*|| + c_3\sigma(x,z)^p. \tag{2.26}$$

Prova. A prova é muito similar à do lema 3.1 de [40], e a incluimos aqui por completude. Por (2.19), temos

$$||P_{xz}(E) - E_*|| \leq [1 + c_1\sigma(x,z)^q]||P_{xz}(E) - E_*||_{x,z}. \tag{2.27}$$

Seja $\tilde{E}$ a projeção ortogonal de E_* em $V(x,z)$, com relação à norma $||\cdot||$. Portanto, por (2.27),

$$||P_{xz}(E) - E_*|| \leq [1 + c_1\sigma(x,z)^q][||P_{xz}(E) - \tilde{E}||_{x,z} + ||\tilde{E} - E_*||_{x,z}]. \tag{2.28}$$

Mas P_{xz} é a projeção em V, e $\tilde{E} \in V$. Logo,

$$||P_{xz}(E) - \tilde{E}||_{x,z} \leq ||E - \tilde{E}||_{x,z} \leq ||E - E_*||_{x,z} + ||\tilde{E} - E_*||_{x,z}. \tag{2.29}$$

portanto, por (2.18), (2.28), e (2.29),

$$\begin{aligned} ||P_{xz}(E) - E_*|| &\leq [1 + c_1\sigma(x,z)^q][||E - E_*||_{x,z} + 2||\tilde{E} - E_*||_{x,z}] \\ &\leq [1 + c_1\sigma(x,z)^q]^2[||E - E_*|| + 2||\tilde{E} - E_*||]. \end{aligned}$$

Agora, pela hipótese H3, $||\tilde{E} - E_*|| \leq c_2\sigma(x,z)^p$. Portanto,

$$\begin{aligned} ||P_{xz}(E) - E_*|| &\leq [1 + c_1\sigma(x,z)^q]^2[||E - E_*|| + 2c_2\sigma(x,z)^p] \\ &= [1 + 2c_1\sigma(x,z)^q + c_1^2\sigma(x,z)^{2q}]||E - E_*|| \\ &+ 2[1 + c_1\sigma(x,z)^q]^2 c_2\sigma(x,z)^p. \end{aligned}$$

Então, fazendo

$$d_1 = \sup\{|x - x^*| \,|x \in \Omega\}, \tag{2.30}$$

temos

$$\begin{aligned} ||P_{xz}(E) - E_*|| &\leq [1 + (2c_1 + c_1^2 d_1^q)\sigma(x,z)^q]||E - E_*|| \\ &+ 2[1 + c_1 d_1^q]^2 c_2 \sigma(x,z)^p. \end{aligned}$$

Logo, (2.26) se segue com $c_3 = 2[1 + c_1 d_1^q]^2 c_2, c_4 = 2c_1 + c_1^2 d_1^q$. ■

Corolário 2.4

Suponha as hipóteses do Lema 2.3. Seja $s = \min\{p, q\}$. Se $\mathcal{N}$ é um subconjunto limitado de X, então existe $c_5 > 0$ tal que

$$||P_{xz}(E) - E_*|| \leq ||E - E_*|| + c_5|x - x_*|^s, \tag{2.31}$$

sempre que $x, z \in \Omega'$, $E \in \mathcal{N}$ and $|z - x_*| \leq |x - x_*|$.

Prova. Definimos d_1 como em (2.30) e $d_2 = \sup\{||E - E_*||, E \in \mathcal{N}\}$. Então, por (2.26),

$$\begin{aligned} ||P_{xz}(E) - E_*|| &\leq [1 + c_4|x - x^*|^q]||E - E_*|| + c_3|x - x^*|^p \\ &\leq ||E - E_*|| + c_4 d_2|x - x^*|^q + c_3|x - x^*|^p. \end{aligned} \tag{2.32}$$

Portanto, (2.31) segue diretamente de (2.32). ■

Para finalizar esta seção, provaremos o teorema de convergência local para (2.3)-(2.6), baseando-nos nos dois prévios princípios de deterioração limitada. A demonstração consiste em mostrar que a seqüência de parâmetros E_k gerada pelo algoritmo, pertence à bola de raio δ_2 definida no teorema 2.2, se E_0 está suficientemente próximo de E_*.

Teorema 2.5

Sejam F, φ, V, E_* satifazendo as Hipóteses H1 a H4. Suponha que $\{x_k\}$ seja definida por (2.3)-(2.6) e seja $r \in (r_*, 1)$. Então existem $\varepsilon = \varepsilon(r)$, $\delta = \delta(r)$ tais que, se $|x_0 - x_*| \leq \varepsilon$, $||E_0 - E_*|| \leq \delta$, a seqüência $\{x_k\}$ está bem definida, converge para x_* e

$$|x_{k+1} - x_*| \leq r|x_k - x_*| \tag{2.33}$$

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

Prova. A prova é muito semelhante à prova do teorema 3.2 de [40].

Seja $\varepsilon_2 = \varepsilon_2(r)$, $\delta_2 = \delta_2(r)$ como dada pelo Teorema 2.2. Seja $\varepsilon, \delta > 0$ tal que $\varepsilon \leq \varepsilon_2(r)$, $\delta \leq \delta_2(r)$ e

$$\delta + c_5\varepsilon^s/(1 - r^s) < \delta_2 . \tag{2.34}$$

Provaremos por indução em k que, para todo $k = 0, 1, 2, \ldots$,

(i) x_{k+1} está bem definida,

(ii) $|x_{k+1} - x_*| \leq r|x_k - x_*|$,

(iii) $|x_{k+1} - x_*| \leq r^{k+1}\varepsilon$,

(iv) $||E_{k+1} - E_*|| \leq \delta + c_5\varepsilon^s \sum_{j=0}^{k} r^{sj}$.

Como $\varepsilon \leq \varepsilon_2(r)$ and $\delta \leq \delta_2(r)$, a tese segue trivialmente do Teorema 2.2 e do Corolário 2.4 para $k = 0$.

Suponhamos agora as hipóteses de indução para $k - 1$. Então,

$$||E_k - E_*|| \leq \delta + c_5\varepsilon^s \sum_{j=0}^{k-1} r^{sj} \leq \delta + \frac{c_5\varepsilon^s}{1 - r^s} < \delta_2 .$$

Portanto, por (2.4), (2.5) e pelo teorema 2.2, x_{k+1} está bem definida e (i)-(iii) se verificam. Finalmente, (iv) se segue das hipóteses indutivas e da propriedade de deterioração limitada (2.31). ■

2.4 CONVERGÊNCIA COM TAXA IDEAL

Nesta seção consideraremos o método definido por (2.3), (2.7) e (2.8). Suponhamos que H1-H4 se verificam e que a seqüência $\{x_k\}$ converge para x_* com taxa linear r. Naturalmente, não há perda de generalidade em considerar que (2.33) se verifica para $k \geq 0$ em vez de para k suficientemente grande, pois, em último caso, podemos simplesmente reenumerar as iterações. Existem métodos para os quais isto é o único resultado que pode ser provado, por exemplo, o método de Newton Modificado, onde

$$x_{k+1} = x_k - J(x_0)^{-1} F(x_k),$$

mas existem métodos para os quais, assumindo a convergência linear com taxa $r > r_*$, pode-se provar que essa convergência se acelera naturalmente para atingir a taxa r_* (superlinear no caso diferenciável quando $r_* = 0$).

Nesta seção, estudaremos condições suficientes sob as quais isto ocorre. Ressaltamos ainda que, embora tenhamos provado anteriormente que uma condição suficiente para convergência com taxa linear r, é a existência de vizinhanças apropriadas, isto não é uma hipótese agora. Ou seja, a hipótese em questão é que x_k converge para x_* com taxa pelo menos $r > r_*$, seja por qualquer motivo.

O primeiro teorema desta seção estabelece uma condição suficiente do tipo Dennis-Moré-Walker, para convergência com taxa ideal.

Teorema 2.6

Suponhamos que as Hipóteses H1 e H2 são satisfeitas e que a seqüência gerada por (2.3) e (2.6)-(2.8) está bem definida, converge para x_*, e que existe $r \geq 0$ tal que

$$|x_{k+1} - x_*| \leq r|x_k - x_*| \tag{2.35}$$

para todo $k = 0, 1, 2, \ldots$. Suponhamos que

$$\lim_{k\to\infty} \frac{|[\varphi(x_k, E_k) - \varphi(x_*, E_*)](x_{k+1} - x_k)|}{|x_{k+1} - x_k|} = 0\,. \tag{2.36}$$

Então,

$$\overline{\lim_{k\to\infty}} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq r_*\,. \tag{2.37}$$

Prova. Façamos $B_k = \varphi(x_k, E_k)$, $B_* = \varphi(x_*, E_*)$. Por (2.36), temos

$$\lim_{k\to\infty} \frac{|(I - B_*^{-1}B_k)(x_{k+1} - x_k)|}{|x_{k+1} - x_k|} = 0\,. \tag{2.38}$$

Agora,

$$\begin{aligned} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} &= \frac{|x_k - B_k^{-1}F(x_k) - x_*|}{|x_k - x_*|} \\ &\leq \frac{|x_k - x_* - B_*^{-1}F(x_k)|}{|x_k - x_*|} + \frac{|[B_*^{-1} - B_k^{-1}]F(x_k)|}{|x_k - x_*|}\,. \end{aligned} \tag{2.39}$$

Pelo Lema 2.1,

$$\overline{\lim_{k\to\infty}} \frac{|x_k - x_* - B_*^{-1}F(x_k)|}{|x_k - x_*|} \leq r_*\,, \tag{2.40}$$

Logo, resta provar que o segundo termo do lado direito de (2.39) tende a 0 quando $k \to \infty$.

Agora,

$$\frac{|[B_*^{-1} - B_k^{-1}]F(x_k)|}{|x_k - x_*|} = \frac{|(I - B_*^{-1}B_k)(x_{k+1} - x_k)|}{|x_k - x_*|}$$

$$= \frac{|(I - B_*^{-1}B_k)(x_{k+1} - x_k)|}{|x_{k+1} - x_k|} \frac{|x_{k+1} - x_k|}{|x_k - x_*|}. \tag{2.41}$$

Mas, por (2.35), $|x_{k+1} - x_k| \leq |x_{k+1} - x_*| + |x_k - x_*| \leq (1 + r)|x_k - x_*|$. Portanto, por (2.38) e (2.41),

$$\lim_{k \to \infty} \frac{|[B_*^{-1} - B_k^{-1}]F(x_k)|}{|x_k - x_*|} = 0. \tag{2.42}$$

Logo, (2.37) segue de (2.39), (2.40) e (2.42). ■

O teorema 2.6 não oferece, por si só, uma condição prática para a convergência com taxa ideal. Com efeito, ele se refere a ação de B_k^{-1} sobre $F(x_k)$, ou ainda, de B_k sobre $(x_{k+1} - x_k)$. Mas x_{k+1} é obtido a partir de B_k "sem liberdade", embora tenhamos liberdade para escolher B_{k+1}, conhecidos x_k e x_{k+1}. Felizmente, estabeleceremos um resultado que permitirá transformar (2.36) numa condição natural relacionada à equação secante, que se refere a ação de B_{k+1} sobre $(x_{k+1} - x_k)$. É para isso que se encaminham os próximos resultados.

Nos teoremas que se seguem, estabeleceremos o comportamento da seqüência $\{E_k\}$ sob as hipóteses de convergência linear.

Teorema 2.7

Suponha que as hipóteses H1-H4 se verificam e que existe $r \in (0, 1)$ tal que

$$|x_{k+1} - x_*| \leq r|x_k - x_*| \tag{2.43}$$

para todo $k = 0, 1, 2, \ldots, .$ Então, $||E_k||$ é uniformemente limitado e

$$\lim_{k\to\infty} ||E_{k+1} - E_k|| = 0\,. \tag{2.44}$$

Prova. A prova é muito similar à do teorema 3.3 de [40], e a incluimos aqui por completude.

Estabeleceremos, primeiro, uma desigualdade relacionado os E_k's, que nos será útil. Por (2.43) e o Lema 2.3 temos que

$$||E_{k+1} - E_*|| \leq (1 + c_4|x_k - x_*|^q)||E_k - E_*|| + c_3|x_k - x_*|^p \tag{2.45}$$

para todo $k = 0, 1, 2, \ldots$. Então, pelo lema 3.3 de [40], $||E_k - E_*||$ e $||E_k||$ são uniformemente limitados. Portanto, existe $c_6 > 0$ tal que

$$||E_{k+1} - E_*|| \leq ||E_k - E_*|| + c_6|x_k - x_*|^s \tag{2.46}$$

para todo $k = 0, 1, 2, \ldots$, (2.43) e (2.46) implicam que

$$||E_{k+j} - E_*|| \leq ||E_k - E_*|| + c_7|x_k - x_*|^s \tag{2.47}$$

para todo $k, j = 0, 1, 2, \ldots$, onde $c_7 = c_6/(1 - r^s)$. Logo, pela limitação uniforme de $||E_k - E_*||$ e $|x_k - x_*|$, obtemos que existe $c_8 > 0$ tal que

$$||E_{k+j} - E_*||^2 \leq ||E_k - E_*||^2 + c_8|x_k - x_*|^s\,. \tag{2.48}$$

Agora, suponhamos que (2.44) seja falso. Então, existe um conjunto infinito de indices K_1 tal que

$$||E_{k+1} - E_k|| \geq \gamma > 0$$

para todo $k \in K_1$. Portanto, por (2.19),

$$[1 + c_1\sigma(x^k, x_{k+1}^k)^q]||E_{k+1} - E_k||_k \geq \gamma$$

para $k \in K_1$. Logo, para k suficientemente grande e $k \in K_1$,

$$||E_{k+1} - E_k||_k \geq \gamma/2.$$

Portanto,

$$||E_{k+1} - E_k||_k^2 \geq \gamma^2/4 \tag{2.49}$$

para k pertencente a um conjunto infinito de indices K_2.

Sejam $\tilde{E}$ e $\tilde{E}_k$ projeções de E_* em $V(x_k, x_{k+1})$ relativas às normas $||.||$ and $||.||_k$, respectivamente.

Por (2.17), e pelo teorema 2.2, temos

$$||\tilde{E} - E_*|| \leq c_2|x_k - x_*|^p.$$

Portanto, por (2.18),

$$\begin{aligned} ||\tilde{E} - E_*||_k &\leq (1 + c_1|x_k - x_*|^q)c_2|x_k - x_*|^p \\ &\leq (1 + c_1\varepsilon^q)c_2|x_k - x_*|^p = c_8|x_k - x_*|^p \end{aligned}$$

com $c_8 = (1 + c_1\varepsilon^q)c_2$.

Portanto, pela definição de $\tilde{E}_k$,

$$||\tilde{E}_k - E_*||_k \leq c_8|x_k - x_*|^p, \tag{2.50}$$

e, consequentemente,

$$||\tilde{E}_k - E_*||_k^2 \leq c_8^2|x_k - x_*|^{2p}. \tag{2.51}$$

Seja $k \in K_2$. Por (2.51) e pela desigualdade triangular temos

$$\begin{aligned} ||E_{k+1} - E_*||_k^2 &= ||E_{k+1} - \tilde{E}_k||_k^2 + ||\tilde{E}_k - E_*||_k^2 \\ &\leq ||E_{k+1} - \tilde{E}_k||_k^2 + c_8^2|x_k - x_*|^{2p}, \\ &= ||E_k - \tilde{E}_k||_k^2 - ||E_{k+1} - E_k||_k^2 + c_8^2|x_k - x_*|^{2p}. \end{aligned}$$

Portanto, por (2.49), (2.50) e (2.18),

$$\begin{aligned}
||E_{k+1} - E_*||_k^2 &= ||E_k - \tilde{E}_k||_k^2 - \frac{\gamma^2}{4} + c_8^2|x_k - x_*|^{2p} \\
&\leq (||E_k - E_*||_k + ||E_* - \tilde{E}_k||_k)^2 - \frac{\gamma^2}{4} + c_8^2|x_k - x_*|^{2p} \\
&\leq (||E_k - E_*||_k + c_8|x_k - x_*|^p)^2 - \frac{\gamma^2}{4} + c_8^2|x_k - x_*|^{2p} \\
&= ||E_k - E_*||_k^2 + 2c_8||E_k - E_*||_k|x_k - x_*|^p + 2c_8^2|x_k - x_*|^{2p} - \frac{\gamma^2}{4} \\
&\leq ||E_k - E_*||_k^2 + 2c_8(1 + c_1|x_k - x_*|^q)||E_k - E_*|| \, |x_k - x_*|^p \\
&+ 2c_8^2|x_k - x_*|^{2p} - \frac{\gamma^2}{4} \\
&\leq ||E_k - E_*||_k^2 + c_9|x_k - x_*|^p - \frac{\gamma^2}{4}
\end{aligned}$$

com $c_9 = 2c_8(1 + c_1\varepsilon^q)\delta_1 + 2c_8^2\varepsilon^p$.

Portanto, existe $\overline{k}$ tal que, para $k \in K_2$ e $k \geq \overline{k}$,

$$||E_{k+1} - E_*||_k^2 \leq ||E_k - E_*||_k^2 - \frac{\gamma^2}{8}.$$

Logo, por (2.18), (2.19), e o Teorema 2.2, temos, para $\overline{\overline{k}}$ suficientemente grande,

$$\begin{aligned}
||E_{k+1} - E_*||^2 &\leq (1 + c_1|x_k - x_*|^q)^2||E_{k+1} - E_*||_k^2 \\
&\leq (1 + c_1|x_k - x_*|^q)^2 \left(||E_{k+1} - E_*||_k^2 - \frac{\gamma^2}{8} \right) \\
&\leq (1 + c_1|x_k - x_*|^q)^2 \left[(1 + c_1|x_k - x_*|^q)^2||E_k - E_*||^2 - \frac{\gamma^2}{8} \right] \\
&\leq ||E_k - E_*||^2 - \frac{\gamma^2}{16} \qquad (2.52)
\end{aligned}$$

para $k \in K_2, k \geq \overline{\overline{k}}$.

Consideremos agora, $k_0 \geq \overline{\overline{k}}$ tal que, para todo $k \geq k_0$,

$$c_7|x_k - x_*|^s \leq \frac{\gamma^2}{32}. \tag{2.53}$$

Definimos

$$K_3 = \{k \in K_2 | k \geq k_0\} = \{k_1, k_2, k_3, \ldots\}, \qquad k_j < k_{i+1}, \ \ i = 1, 2, 3, \ldots$$

Então, para todo $j = 1, 2, 3, \ldots$, temos, por (2.52),

$$||E_{k_j+1} - E_*||^2 \leq E_{k_j} - E_*||^2 - \frac{\gamma^2}{16}. \tag{2.54}$$

Agora, por (2.48), (2.53) e (2.54),

$$\begin{aligned} ||E_{k_{j+1}} - E_*||^2 &\leq ||E_{k_j+1} - E_*||^2 + c_7|x_k - x_*|^5 \\ &\leq ||E_{k_j} - E_*||^2 - \frac{\gamma^2}{16} + \frac{\gamma^2}{32} \\ &= ||E_{k_j} - E_*||^2 - \frac{\gamma^2}{32}. \end{aligned} \tag{2.55}$$

Mas (2.55) se verifica para todo $j = 1, 2, 3, \ldots$. Portanto,

$$||E_{k_j} - E_*||^2 \leq ||E_{k_1} - E_*||^2 - (j-1)\frac{\gamma^2}{32}. \tag{2.56}$$

E (2.56) implica que $||E_{k_j} - E_*||^2 < 0$ para j suficientemente grande, o que é uma contradição. ■

Teorema 2.8

Suponhamos que as Hipóteses H1-H4 são satisfeitas e que existe $r \in (0,1)$ tal que (2.43) se verifica para todo $k = 0, 1, 2, \ldots$. Suponhamos que existe um conjunto fechado $\Gamma \subset \mathbb{R}^n \times X$ tal que $(x_k, E_k) \in \Gamma \subset \Omega \times D$ para todo $k = 0, 1, 2, \ldots$. Então

$$\lim_{k \to \infty} |\varphi(x_{k+1}, E_{k+1}) - \varphi(x_k, E_k)| = 0. \tag{2.57}$$

Prova. Por (2.43) temos que

$$x_k \in \{x \in I\!R^n \mid |x - x_*| \leq |x_0 - x_*|\} \equiv B_1 \tag{2.58}$$

para todo $k = 0, 1, 2, \ldots$.

Pelo teorema 2.7, existe $M > 0$ tal que

$$E_k \in \{E \in X \mid ||E|| \leq M\} \equiv B_2 \tag{2.59}$$

para todo $k = 0, 1, 2, \ldots$.

Portanto, por (2.58) e (2.59),

$$(x_k, E_k) \in B_1 \times B_2 \subset I\!R^n \times X \tag{2.60}$$

para todo $k = 0, 1, 2, \ldots$, e, obviamente, $B_1 \times B_2$ é compacto. Portanto, por (2.58) e as hipóteses,

$$(x_k, E_k) \in (B_1 \times B_2) \cap \Gamma \equiv C \tag{2.61}$$

para todo $k = 0, 1, 2, \ldots$, e, como $B_1 \times B_2$ é compacto e Γ é fechado, então C é um conjunto compacto de $I\!R^n \times X$ e $C \subset \Omega \times D$. Agora, da convergência de $\{x_k\}$ se segue que $|x_{k+1} - x_k| \to 0$ e pelo Teorema 2.7, $||E_{k+1} - E_k|| \to 0$. Portanto, (2.57) segue da continuidade uniforme de φ em C. ■

Apesar da importância da propriedade (2.57), ela ainda não é suficiente para provar a "convergência ideal". Por exemplo, o método de Newton modificado satisfaz obviamente esta propriedade, e, no entanto, não desfruta de convergência ideal.

Algoritmos secantes para resolução de sistemas não lineares são caracterizados pela equação secante que, no caso não diferenciável, tem a forma

$$B_{k+1}(x_{k+1} - x_k) = F_1(x_{k+1}) - F_1(x_k) \tag{2.62}$$

para todo $k = 0, 1, 2, \ldots$, ou, usando (2.7),

$$\varphi(x_{k+1}, E_{k+1})(x_{k+1} - x_k) = F_1(x_{k+1}) - F_1(x_k)\,. \tag{2.63}$$

Se as hipóteses H1-H4 se verificam, assim como uma equação do tipo secante, obtemos, usando os resultados prévios, a convergência com taxa ideal. É isto que será estabelecido nos próximos teoremas.

Teorema 2.9

Sob as hipóteses do teorema 2.8, suponhamos que

$$\lim_{k\to\infty} \frac{|[\varphi(x_{k+1}, E_{k+1}) - \varphi(x_*, E_*)](x_{k+1} - x_k)|}{|x_{k+1} - x_k|} = 0\,. \tag{2.64}$$

Então

$$\overline{\lim_{k\to\infty}} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq r_*\,.$$

Prova. Pelo teorema 2.8, (2.64) implica (2.36). Portanto, o resultado desejado segue do teorema 2.6. ■

Teorema 2.10

Sob as hipóteses do Teorema 2.8, suponhamos que

$$\varphi(x_*, E_*) = J(x_*) \tag{2.65}$$

e que (2.63) se verifica para todo $k = 0, 1, 2, \ldots$.

Então,

$$\overline{\lim_{k\to\infty}} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq r_*\,.$$

Prova. Por (2.14), (2.63) e (2.65) temos que

$$\frac{|[\varphi(x_{k+1}, E_{k+1}) - \varphi(x_*, E_*)](x_{k+1} - x_k)|}{|x_{k+1} - x_k|}$$

$$= \frac{|F_1(x_{k+1}) - F_1(x_k) - J(x_*)(x_{k+1} - x_k)|}{|x_{k+1} - x_k|} \leq L|x_k - x_*|^p .$$

Portanto, o resultado segue do Teorema 2.9. ∎

Teorema 2.11

Existem $\varepsilon, \delta > 0$ tais que, se $|x_0 - x_*| \leq \varepsilon$, $||E_0 - E_*|| \leq \delta$ e (2.64) (respectivamente, (2.65) e (2.63)) se verifica, então a seqüência $\{x_k\}$ está bem definida, converge para x_*, e $\overline{\lim_{k\to\infty}} \dfrac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq r_*$.

Prova. A convergência é provada do Teorema 2.9. Assim, por (2.47),

$$||E_k - E_*|| \leq ||E_0 - E_*|| + c_7|x_0 - x_*|^s$$

para todo $k = 0, 1, 2, \ldots$. Assim, podemos restringir δ e ε para que todos os E_k's pertençam à bola estritamente contida em D cujo centro é E_*. Portanto, o conjunto fechado Γ mencionado nas hipóteses do Teorema 2.8 existe neste caso. Logo, o resultado desejado segue do teorema 2.9. ∎

Exemplo Numérico

Consideremos uma discretização por diferenças finitas, 16×16, de $\Delta u + \alpha|u| = f(s,t,u)$ no quadrado $[0,1] \times [0,1]$, onde $f(s,t,u) = u^3/(1 + s^2 + t^2)$, e as condições de fronteira dadas por $u(0,t) = u(s,0) = 1, u(s,1) = 2 - e^s, u(1,t) = 2 - e^t$. [52].

A matriz Jacobiana da parte diferenciável deste sistema tem uma estrutura bem familiar, resultante da discretização do Laplaciano com a fórmula padrão de cinco pontos.

Definimos $V(x,z)$ o conjunto de matrizes com tal estrutura e que satisfazem a equação secante $B(z-x) = F_1(z) - F_1(x)$. Testamos um método secante, o primeiro método de Broyden, e o método de Newton para diferentes valores de α. Para $\alpha < 0.1$, não foram detectadas diferenças significativas entre os dois métodos. Isto parece confirmar que ambos têm a mesma taxa de convergência na prática, como predito pela teoria. Por exemplo, para $\alpha = 0.08$, a convergência começando de $(-1,\ldots,-1)$, ocorreu em 4 iterações para Newton e 5 iterações para o método secante. Para $\alpha \in (0.001, 0.08)$ ambos os métodos convergiram em 4 iterações. Para $\alpha = 0.001$ Newton usou somente 3 iterações, talvez devido ao fato de que o peso da parte não diferenciável é tão pequeno que o comportamento quadrático do método de Newton diferenciável ocorre.

Poderíamos considerar a possibilidade de definir $V(x,z)$, usando toda a função F, em vez de somente sua parte diferenciável. De fato, é claro que a convergência superlinear pode ser obtida desta maneira, se a solução for um ponto diferenciável. Do ponto de vista prático, uma boa taxa de convergência pode ser obtida deste modo, se a maioria das componentes de F forem diferenciáveis em x_*. Estamos cogitando sobre a possibilidade de obter, teoricamente, a convergência no caso geral, se a equação secante for definida usando toda função F. De qualquer forma, pesquisas posteriores são necessárias para este objetivo.

2.5 CONCLUSÕES

Em termos gerais, as conseqüências dos resultados das seções 2.3 e 2.4 deste capítulo, são que todos os métodos secantes já introduzidos para problemas diferenciáveis podem ser estendidos para problemas parcialmente diferenciáveis, desde que uma condição do tipo (2.16) seja satisfeita. Esta condição afirma, essencialmente, que a parte diferenciável de F pesa mais do que a parte não diferenciável. Uma conseqüência interessante deste resultado, é que a iteração newtoniana ideal, baseada na recorrência $x_{k+1} = x_k - J(x_*)^{-1}F(x_k)$ tem a mesma taxa de convergência linear que a iteração secante que usamos para aproximá-la. Assim, pelo menos na teoria, não haveria vantagem em calcular as derivadas da parte diferenciável de um sistema não linear se não for possível calcular as derivadas de toda F. Experimentos numéricos adicionais ainda são necessários para verificar se este fato se verifica na prática.

Capítulo 3

Atualização de uma coluna do Jacobiano Inverso

3.1 INTRODUÇÃO

Neste capítulo, introduzimos uma nova fórmula de atualização da matriz inversa do Jacobiano aproximado por iteração. Esta nova fórmula pode ser usada em diferentes contextos. Por exemplo, para gerar diferentes métodos quase-newtonianos para sistemas não lineares parcialmente diferenciáveis, como as estudadas no capítulo 2. Também como geradoras de precondicionadores secantes, como veremos no capítulo 4. Mostraremos as propriedades teóricas desta nova fórmula secante na resolução de SNL diferenciáveis. Assim, consideremos o problema de resolver

$$F(x) = 0\,, \tag{3.1}$$

onde $F : I\!R^n \to I\!R^n$ é uma função não linear diferenciável. Estamos especialmente interessados em casos onde n é grande.

É geralmente aceito que a propriedade teórica que explica o bom comportamento prático de métodos direcionados para a resolução de (3.1), é a convergência superlinear. Entretanto, nas últimas duas décadas, alguns autores introduziram algoritmos que, embora pareçam não ter esta propriedade, ainda assim apresentam bom desempenho prático. Entre eles, podemos citar os métodos introduzidos por Tewarson [54], Tewarson e Zhang

[55], Dennis e Marwil [15] e Martínez [38]. Gomes-Ruggiero e Martínez [27] mostraram que o "Column-Updating method" (CUM) introduzido em Martínez [39] é muito efetivo quando comparado com a implementação do primeiro método de Broyden com memória limitada (Gomes-Ruggiero, Martínez e Moretti [28], Griewank [30], Matthies e Strang [46]) na resolução de sistemas não lineares de grande porte.

Neste capítulo, introduzimos um método relacionado ao CUM. De fato, enquanto em CUM uma coluna da matriz de aproximação do Jacobiano é modificada por iteração para satisfazer a equação secante, no método aqui introduzido, atualizamos uma coluna da matriz de aproximação para o inverso do Jacobiano, de modo que a equação secante seja também satisfeita a cada iteração. Veremos que o novo método, que denominaremos ICUM (Inverse Column-Updating Method), tem as mesmas propriedades de convergência de CUM e seu desempenho prático é muito satisfatório quando comparado a este último, com o método de Newton e com o primeiro método de Broyden.

3.2 O MÉTODO ICUM

Suponha m um inteiro positivo, $x_0 \in I\!R^n$ é uma aproximação inicial para a solução de (3.1) e $H_0 \in I\!R^{n\times n}$.

Dado $x_k \in I\!R^n$, $H_k \in I\!R^{n\times n}$, $F(x_k) \neq 0$, x_{k+1} , é definido por

$$x_{k+1} = x_k - H_k F(x_k). \tag{3.2}$$

Se $k+1 \equiv 0 \pmod{m}$, definimos $H_{k+1} \in I\!R^{n\times n}$, de alguma maneira a ser especificada (veja seção 3.4), e a iteração termina. O restante desta seção corresponde ao caso onde $k+1$ não é um múltiplo de m.

Neste caso, definimos

$$s_k = x_{k+1} - x_k \ , \tag{3.3}$$

$$y_k = F(x_{k+1}) - F(x_k)\,, \tag{3.4}$$

e escolhemos $j_k \in \{1, \ldots, n\}$ tal que

$$|y_k^{j_k}| = ||y_k||_\infty \equiv \max\{|y_k^1|, \ldots, |y_k^n|\}. \tag{3.5}$$

onde as coordenadas de um vetor $z \in I\!R^n$ são denotadas por $z^1, ..., z^n$.

Definimos também $H_{k+1}(\equiv (h_{k+1}^{il}))$ igual a $H_k(\equiv (h_k^{il}))$ exceto provavelmente a coluna j_k. A j_k-ésima coluna é definida por

$$h_{k+1}^{ij_k} = (s_k^i - \sum_{l \neq j_k} h_k^{il} y_k^l)/y_k^{j_k}, \tag{3.6}$$

$i = 1, \ldots, n$.

Observações

Por (3.5) e (3.6) é fácil ver que, sempre que $k + 1 \not\equiv 0 \pmod{m}$ e $y_k \neq 0$, temos que

$$H_{k+1} y_k = s_k. \tag{3.7}$$

Esta equação, que é a forma inversa da equação secante que caracteriza os métodos quase-newtonianos mais bem sucedidos [18], tem papel fundamental nestes métodos, pois é satisfeita por uma média de Jacobianos no segmento $[x_k, x_{k+1}]$, devido à identidade:

$$y_k = [\int_0^1 J(x_k + ts_k)dt]s_k. \tag{3.8}$$

Os resultados de convergência se verificam se, em vez de (3.5), definirmos j_k como qualquer índice que satisfaça

$$|y_k^{j_k}| \geq \alpha ||y_k||_\infty, \tag{3.9}$$

para um valor fixo de $\alpha > 0$. Entretanto, não vemos qualquer razão prática para usar $\alpha \neq 1$. Assim sendo, para simplificar a exposição, usaremos diretamente (3.5).

3.3 RESULTADOS DE CONVERGÊNCIA

Denotemos, por ora, $||.||$ como a norma euclideana de vetores e sua norma subordinada de matrizes.

Suponhamos $F : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n, F \in C^1(\Omega), \Omega$ um conjunto convexo, $x_* \in \Omega, F(x_*) = 0$ e

$$||J(x) - J(x_*)|| \leq L||x - x_*||^p, L, p > 0. \tag{3.10}$$

para todo $x \in \Omega$; (3.10) implica que para todo $u, v \in \Omega$

$$||F(v) - F(u) - J(x_*)(v - u)|| \leq L||v - u||\sigma(u, v)^p \tag{3.11}$$

onde $\sigma(u, v) = \max\{||u - x_*||, ||v - x_*||\}$. (Broyden, Dennis e Moré [9]).

Suponhamos que $J(x_*)$ seja não singular e definimos $M = ||J(x_*)^{-1}||$. Por (3.11), deduzimos que, para todo $u, v \in D$,

$$||v - u - J(x_*)^{-1}[F(v) - F(u)]|| \leq ML||v - u||\sigma(u, v)^p. \tag{3.12}$$

Lema 3.1

Existe $\varepsilon_1 > 0$ tal que $F(v) \neq F(u)$ sempre que $v \neq u$, $||v - x_*|| \leq \varepsilon_1$, $||u - x_*|| \leq \varepsilon_1$.

Prova

Por (3.12) temos que, para todo $u, v \in D$,

$$||v - u|| - ||J(x_*)^{-1}[F(v) - F(u)]|| \leq ML||v - u||\sigma(u, v)^p.$$

Portanto,

$$||F(v) - F(u)|| \geq ||v - u||(\frac{1}{M} - L\sigma(u, v)^p) \tag{3.13}$$

Seja $\varepsilon_1 > 0$ tal que

$$\varepsilon_1^p < \frac{1}{2ML}. \tag{3.14}$$

Por (3.14), se $||u - x_*|| \leq \varepsilon_1, ||v - x_*|| \leq \varepsilon_1$, temos que:

$$L\sigma(u,v)^p \leq L\varepsilon_1^p < \frac{1}{2M}.$$

Logo

$$\frac{1}{M} - L\sigma(u,v)^p > \frac{1}{M} - \frac{1}{2M} = \frac{1}{2M}. \tag{3.15}$$

Assim, por (3.13) e (3.15),

$$||F(v) - F(u)|| \geq \frac{1}{2M}||v - u||. \tag{3.16}$$

De (3.16) o resultado desejado é obtido diretamente. ■

O resultado de convergência local é estabelecido no seguinte teorema:

Teorema 3.2

Consideremos as seqüências $\{x_k\}$ e $\{H_k\}$ geradas pelo método ICUM e suponhamos que $F(x_k) \neq 0$ para todo $k = 0, 1, 2, \ldots$. Seja $r \in (0,1)$. Existem $\varepsilon = \varepsilon(r), \delta = \delta(r)$ tais que, se $||x_0 - x_*|| \leq \varepsilon$ e $||H_k - J(x_*)^{-1}|| \leq \delta$, sempre que $k \equiv 0 \pmod{m}$, então as seqüências $\{x_k\}$ e $\{H_k\}$ estão bem definidas, $\{x_k\}$ converge para x_* e

$$||x_{k+1} - x_*|| \leq r||x_k - x_*|| \tag{3.17}$$

para todo $k = 0, 1, 2, \ldots$.

Prova

Sejam $c_1 = 2nM^2L, c_2 = n^{3/2}$. Dados $\varepsilon, \delta > 0$, definimos $b_i(\varepsilon, \delta), i = 0, 1, \ldots, m-1$ por

$$b_0(\varepsilon, \delta) = \delta, \tag{3.18}$$

$$b_i(\varepsilon,\delta) = c_2 b_{i-1}(\varepsilon,\delta) + c_1\varepsilon^p, \tag{3.19}$$

$i = 1, 2, \ldots, m-1$.

Claramente, temos, para quaisquer $\varepsilon, \delta > 0$

$$0 < b_0(\varepsilon,\delta) < b_1(\varepsilon,\delta) < \cdots < b_{m-1}(\varepsilon,\delta) \tag{3.20}$$

e

$$\lim_{\varepsilon,\delta\to 0} b_i(\varepsilon,\delta) = 0 \tag{3.21}$$

para $i = 0, 1, \ldots, m-1$.

Por (3.20), (3.21) podemos escolher $\varepsilon = \varepsilon(r) > 0, \delta = \delta(r) > 0$ tais que $\varepsilon \leq \varepsilon_1$ e

$$b_i(\varepsilon,\delta) + L\varepsilon^p < r/M_1 \tag{3.22}$$

para $i = 0, 1, \ldots, m-1$, onde $M_1 = \max\{||J(x_*)||, 2M\}$.

Suponhamos que $||x_0 - x_*|| \leq \varepsilon$ e $||H_k - J(x_*)^{-1}|| \leq \delta$ sempre que $k \equiv 0 \pmod m$. Provaremos por indução em k que se $k \equiv q \pmod m$ então H_k é não singular,

$$||x_{k+1} - x_*|| \leq r||x_k - x_*||, \tag{3.23}$$

$$||H_k - J(x_*)^{-1}|| \leq b_q(\varepsilon,\delta) \tag{3.24}$$

e

$$||H_k|| \leq 2M \tag{3.25}$$

para todo $q = 0, 1, \ldots, m-1$.

Para $k = 0$, por hipótese,

$$||H_0 - J(x_*)^{-1}|| \leq \delta = b_0(\varepsilon,\delta). \tag{3.26}$$

Portanto, por (3.22) e (3.26),

$$\begin{aligned} ||H_0|| &\leq ||J(x_*)^{-1}|| + ||H_0 - J(x_*)^{-1}|| \\ &\leq ||J(x_*)^{-1}|| + \delta \\ &\leq ||J(x_*)^{-1}|| + 1/||J(x_*)|| \\ &\leq 2||J(x_*)^{-1}|| = 2M. \end{aligned}$$

Além disso, por (3.11),

$$\begin{aligned} ||x_1 - x_*|| &= ||x_0 - x_* - H_0F(x_0)|| \\ &= ||x_0 - x_* - H_0[F(x_0) - J(x_0) - J(x_*)(x_0 - x_*)]|| \\ &+ ||H_0J(x_*)(x_0 - x_*)|| \\ &\leq ||[I - H_0J(x_*)](x_0 - x_*)|| + 2ML||x_0 - x_*||^{p+1} \\ &\leq (||J(x_*)||\,||J(x_*)^{-1} - H_0|| + 2ML||x_0 - x_*||^p)||x_0 - x_*|| \\ &\leq M_1(\delta + L\varepsilon^p)||x_0 - x_*|| \\ &= M_1(b_0(\varepsilon,\delta) + L\varepsilon^p)||x_0 - x_*|| \leq r||x_0 - x_*||. \end{aligned}$$

Consideramos agora $k > 0, k \equiv q \pmod{m}$. Se $q = 0$, à prova de (3.23) - (3.25) é análoga a prova para $k = 0$. Suponha $q > 0$. Provemos primeiro que H_k está bem definida e que (3.24) se verifica.

Por hipótese, temos que $F(x_{k-1}) \neq 0$ e H_{k-1} é não singular, portanto $s_{k-1} \neq 0$ e $x_k \neq x_{k-1}$. Assim, pelo Lema 3.1, $y_{k-1} \equiv F(x_k) - F(x_{k-1}) \neq 0$. Isto prova que o denominador de (3.6) não se anula. Portanto, H_k está bem definida.

Seja j_{k-1} tal que

$$|y_{k-1}^{j_{k-1}}| = \max\{|y_{k-1}^1|, \ldots, |y_{k-1}^n|\},$$

pelos argumentos prévios, sabemos que $y_{k-1}^{j_{k-1}} \neq 0$.

Agora, por (3.6), temos, para $i = 1, \ldots, n$,

$$\begin{aligned} h_k^{ij_{k-1}} &= (s_{k-1}^i - \sum_{l \neq j} h_{k-1}^{il} y_{k-1}^l)/y_{k-1}^{j_{k-1}} \\ &= (s_{k-1}^i - \sum_{l \neq j_{k-1}} h_*^{il} y_{k-1}^l + \sum_{l \neq j_{k-1}} h_*^{il} y_{k-1}^l - \sum_{l \neq j_{k-1}} h_{k-1}^{il} y_{k-1}^l)/y_{k-1}^{j_{k-1}} \end{aligned}$$

onde $J(x_*)^{-1} \equiv (h_*^{il})$.

Portanto, por (3.12),

$$\begin{aligned} |h_k^{ij_{k-1}} - h_*^{ij_{k-1}}| &\leq |s_{k-1}^i - \sum_{l=1}^n h_*^{il} y_{k-1}^l|/|y_{k-1}^{j_{k-1}}| \\ &+ \sum_{l \neq j_{k-1}} |h_*^{il} - h_{k-1}^{il}||y_{k-1}^l|/|y_{k-1}^{j_{k-1}}| \\ &\leq ||s_{k-1} - J(x_*)^{-1} y_{k-1}||/|y_{k-1}^{j_{k-1}}| + \sum_{l=1}^n |h_*^{il} - h_{k-1}^{il}| \\ &\leq \sqrt{n}||s_{k-1} - J(x_*)^{-1} y_{k-1}||/||y_{k-1}|| + n||H_{k-1} - J(x_*)^{-1}|| \\ &\leq \sqrt{n} M L ||s_{k-1}|| \varepsilon^p / ||y_{k-1}|| + n||H_{k-1} - J(x_*)^{-1}||. \end{aligned} \tag{3.27}$$

Agora, por (3.16),

$$||y_{k-1}|| \geq \frac{1}{2M}||s_{k-1}||. \tag{3.28}$$

Assim sendo, por (3.27) e (3.28),

$$|h_k^{ij_{k-1}} - h_*^{ij_{k-1}}| \leq 2\sqrt{n}M^2L\varepsilon^p + n||H_{k-1} - J(x_*)^{-1}||. \tag{3.29}$$

Logo, por (3.19),

$$\begin{aligned} ||H_k - J(x_*)^{-1}|| &\leq 2nM^2L\varepsilon^p + n^{3/2}||H_{k-1} - J(x_*)^{-1}|| \\ &\leq 2nM^2L\varepsilon^p + n^{3/2}b_{q-1}(\varepsilon,\delta) \\ &= c_2b_{q-1}(\varepsilon,\delta) + c_1\varepsilon^p = b_q(\varepsilon,\delta). \end{aligned} \tag{3.30}$$

Logo, (3.24) está provado.

Assim, por (3.22),

$$||H_k - J(x_*)^{-1}|| \leq r/M_1 \leq 1/2M.$$

Portanto, pelo de Lema de Banach [25], H_k é não singular.

Portanto,

$$\begin{aligned} ||H_k|| &\leq ||J(x_*)^{-1}|| + ||H_k - J(x_*)^{-1}|| \\ &\leq ||J(x_*)^{-1}|| + r/M_1 \\ &\leq ||J(x_*)^{-1}|| + 1/||J(x_*)|| \leq 2||J(x_*)^{-1}|| = 2M \end{aligned} \tag{3.31}$$

Logo, (3.25) está provado.

Finalmente, por (3.11), (3.31), (3.22),

$$||x_{k+1} - x_*|| = ||x_k - x_* - H_k F(x_k)||$$

$$= ||x_k - x_* - H_k[F(x_k) - J(x_*)(x_k - x_*)] - H_k J(x_*)(x_k - x_*)||$$

$$\leq ||[I - H_k J(x_*)](x_k - x_*)|| + 2ML||x_k - x_*||^{p+1}$$

$$\leq (||J(x_*)||\ ||J(x_*)^{-1} - H_k|| + 2ML||x_k - x_*||^p)||x_k - x_*||$$

$$\leq M_1(b_q(\varepsilon, \delta) + L\varepsilon^p)||x_k - x_*|| \leq r||x_k - x_*||.$$

Logo, (3.23) também está provado. Isto completa a demonstração do teorema. ■

3.4 IMPLEMENTAÇÃO COMPUTACIONAL

Nesta seção descreveremos uma implementação computacional de ICUM direcionada para problemas de grande porte.

Definimos H_k, s_k, y_k e j_k como na seção 2 deste capítulo. É fácil ver que, quando $k+1 \not\equiv 0 \pmod{m}$, a fórmula (3.6) produz

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)e_{j_k}^t}{y_k^t e_{j_k}}, \tag{3.32}$$

onde $\{e_1, \ldots, e_n\}$ é a base canônica de $\mathbb{R}^n$.

Portanto, definindo

$$u_k = s_k - H_k y_k, \tag{3.33}$$

temos que, se $\beta \equiv 0 \pmod{m}$, $\nu \in \{1, \ldots, m-1\}$,

$$H_{\beta+\nu} = H_\beta + \sum_{l=0}^{\nu-1} u_{\beta+l} e_{j_{\beta+l}}^T / y_{\beta+l}^{j_{\beta+l}}. \tag{3.34}$$

A fórmula (3.34) mostra que uma iteração de ICUM pode ser implementada usando no máximo $O(nm)$ posições de memória, mais o necessário para armazenar H_k para $k \equiv (\text{mod } m)$.

Nos experimentos da seção 3.5, escolhemos

$$H_k = [P_\tau(J(x_k))]^{-1}, \text{ se } k \equiv 0(\text{mod } m), \tag{3.35}$$

onde P_τ é a projeção no espaço das matrizes tridiagonais. Como H_k é a inversa da matriz tridiagonal, o produto $H_k v$ pode ser computado usando $O(n)$ flops. De fato, H_k não é armazenada, mas sim a fatoração LU de H_k^{-1} [25].

Na implementação computacional de ICUM usamos, no lugar de (3.2), a fórmula mais efetiva

$$s_k = -\lambda_k H_k F(x_k), \tag{3.36}$$

$$x_{k+1} = x_k + s_k.$$

onde $\lambda_k \in (0, 1]$ é usado para prevenir passos excessivamente grandes. Calculamos λ_k por

$$\lambda_k = 1 \text{ se } ||H_k F(x_k)|| \leq \text{ BIG} \tag{3.37}$$

$$\text{BIG } / ||H_k F(x_k)|| \text{ caso contrário,}$$

onde BIG é um número positivo grande. Usamos BIG $= \min\{10^6, 10^6||x_k||\}$. Obviamente, esta modificação não altera os resultados de convergência, pois, numa vizinhança da solução, $||H_k F(x_k)||$ é pequeno.

Embora tenhamos provado que, perto de uma solução isolada não é possível que $F(x_{k+1}) = F(x_k)$, isto pode ocorrer longe de x_*. Mais precisamente, é possível que

$$||F(x_{k+1}) - F(x_k)|| \leq \text{TOL}\ ||F(x_k)||, \tag{3.38}$$

para TOL muito pequeno, conduzindo à instabilidade no cálculo de H_{k+1}. Portanto, em nossa implementação, fazemos $H_{k+1} = H_k$ se (3.38) ocorre para TOL $= 10^{-6}$.

3.5 EXPERIMENTOS NUMÉRICOS

Usamos os testes de Schwandt [52]. Cada teste é gerado pela discretização por diferenças finitas de uma equação de Poisson no quadrado $[0,1] \times [0,1]$. O número de divisões no intervalo é denotado por N. Portanto, o número de variáveis é $n = (N-1)^2$. As equações de Poisson são $\Delta u = f(s,t,u)$ onde

(a_i) $f(s,t,u) = 10^i u^3/(1+s^2+t^2), i = 0,2,4$

$$u(s,t) = \begin{cases} 1. & s = 0, t \in [0,1] \quad ou\ t = 0, s \in [0,1] \\ 2 - e^s & t = 1, s \in [0,1] \\ 2 - e^t & s = 1, t \in [0,1] \end{cases}$$

(b) $f(s,t,u) = u^3$, $u(s,t) = 0$ na fronteira

(c) $f(s,t,u) = e^u$, $u(s,t) = s + 2t$ na fronteira

Usamos os seguintes métodos:

NR. : Newton. Implementação de Gomes-Ruggiero, Martínez e Moretti [28].

BR: Primeiro método de Broyden. Implementação de Gomes-Ruggiero, Martínez e Moretti [28], com $B_k \equiv P_\tau(J(x_k))$ quando $k \equiv 0 \pmod{m}$.

CUM: Column-Updating method. Implementação de Gomes-Ruggiero e Martínez [27] com $B_k \equiv P_\tau(J(x_k))$ quando $k \equiv 0 \pmod{m}$.

ICUM : O método descrito neste capítulo.

Nos problemas $(a_i, i = 0, 2, 4)$ e (c) usamos o critério de parada $||F(x_k)|| \leq 10^{-3}$. No problema (b) paramos o processo quando $||F(x_k)|| \leq 10^{-5}$. O ponto inicial usado foi $x_0 = (-1, \ldots, -1)^T$ em todos os casos. Usamos $m = 30$ quando $N = 32$ e $N = 64$, e $m = 25$ quando $N = 128$.

Os testes foram feitos num VAX11/785 da Universidade Estadual de Campinas - UNICAMP, usando o compilador FORTRAN 77 e o sistema operacional VMS, utilizando precisão simples.

Tabela 1

		MÉTODOS			
PROBLEMAS		NR	BR	CUM	ICUM
a_0	N = 32	(2; 30.6)	(64; 15.6)	(62; 10.6)	(52; 8.5)
	N = 64	*	(268; 1285.3)	(164; 118.3)	(86; 63.7)
	N = 128	*	*	*	(182; 566.1)
a_2	N = 32	(5; 62.3)	(51; 11.8)	(66; 11.3)	(47; 8.1)
	N = 64	*	(101; 106.1)	(176; 125.3)	(80; 56.4)
	N = 128	*	*	(161; 483.4)	(155; 481.2)
a_4	N = 32	(9; 96.7)	(53; 12.4)	(66; 11.1)	(58; 10.0)
	N = 64	*	(78; 77.2)	(85; 59.1)	(75; 52.5)
	N = 128	*	(67; 300.4)	(94; 273.7)	(63; 182.2)
b	N = 32	(2; 31.4)	(101; 23.7)	(75; 11.6)	(64; 9.7)
	N = 64	*	(199; 186.2)	(227; 146.2)	(140; 91.2)
	N = 128	*	*	*	*
c	N = 32	(2; 28.8)	(115; 26.2)	(62; 9.9)	(58; 9.3)
	N = 64	*	(138; 139.7)	(134; 91.1)	(96; 66.4)
	N = 128	*	*	*	(160; 475.5)

Os resultados estão na Tabela 1. Cada experimento é representado pelo par (Kon; Time), onde Kon é o número de iterações e Time é o tempo de CPU em segundos. O símbolo * significa não convergência após 20 minutos em tempo de CPU.

3.6 CONCLUSÕES

Como esperado, a implementação dos métodos secantes onde os recomeços são feitos de acordo com $H_k = P_\tau(J(x_k))^{-1}$ (resp. $B_k = P_\tau(J(x_k))$ no método de Broyden e CUM) foram mais eficientes em termos do tempo computacional do que o método de Newton. O motivo deste fato é que o procedimento da fatoração de $J(x_k)$ usa $O(N^4)$ flops enquanto a fatoração de $P_\tau(J(x_k))$ usa $O(N^2)$ flops. No caso do método de Newton, o número de iterações é independente de N. Portanto, o tempo de CPU usado pelo método de Newton é, geralmente, $O(N^4)$. Observamos, empiricamente, que o número de iterações usados pelos métodos secantes, com recomeços com a parte tridiagonal, é $O(N)$. Isto explica o porquê da superioridade dos métodos secantes sobre o método de Newton, a qual se torna mais acentuada quando N cresce.

Os exemplos de Poisson são representativos de uma vasta e importante classe de sistemas não lineares: aqueles provenientes da discretização de Equações Diferenciais Parciais que ocorrem na Física e Engenharia. Para estes problemas, o método ICUM introduzido neste capítulo, apresenta desempenho superior a todos os outros métodos testados.

Pesquisas futuras são necessárias, tanto do ponto de vista teórico quanto prático. Do lado teórico, talvez seja possível justificar o comportamento da necessidade de "menos que $O(N)$ " iterações em ICUM e os outros métodos com recomeços tridiagonais. Acreditamos ser possível obter resultados de convergência superiores aos apresentados na seção 3.3 deste capítulo.

Do ponto de vista prático, é necessário incorporar estratégias de convergência global, ao mesmo tempo que implementações eficientes empregando paralelismo devem ser desenvolvidas.

Capítulo 4

Métodos de Newton Inexatos com Precondicionadores Secantes

4.1 INTRODUÇÃO

Uma das desvantagens do método de Newton, na resolução de um sistema de equações não lineares, é a necessidade de resolver o sistema linear

$$J(x_k)s_k = -F(x_k) \tag{4.1}$$

a cada iteração, usando algum método direto. Nesse caso, alguma forma de fatorar $J(x_k)$ deve ser usada (sobretudo fatoração LU ou, às vezes, QR [25]). Quando n é grande, técnicas especiais para fatoração, considerando esparsidade do problema, são frequentemente empregadas [63,64]. Entretanto, o padrão de esparsidade da matriz Jacobiana pode ser tão desfavorável, que tais técnicas podem ser ineficientes, conduzindo a excessivo enchimento (fill-in) no procedimento da fatoração. É o caso, por exemplo, da discretização de problemas de contorno tridimensionais. Neste caso, devem ser usados métodos iterativos para obter uma solução aproximada de (4.1). A vantagem dos métodos iterativos, é

que o custo de uma iteração é reduzido quando comparado a de um método direto. Além disso, a memória exigida é aproximadamente a mesma que a necessária para armazenar os dados.

O comportamento de métodos iterativos lineares, quando aplicados a (4.1), foi analisado por Ortega e Rheinboldt em 1970 [48] e Sherman em 1978 [53]. Entretanto, somente em 1982, Dembo, Eisenstat e Steihaug [20] forneceram uma teoria, na qual estabelecem o critério de parada prático para o método iterativo linear, necessário para obter a convergência satisfatória do método não linear. Essa teoria define os Métodos de Newton Inexatos. Neles, obtém-se uma solução aproximada de (4.1), satisfazendo

$$||J(x_k)s_k + F(x_k)|| \leq \theta_k ||F(x_k)|| \tag{4.2}$$

onde $\theta_k \in (0,1)$.

0 principal resultado de convergência é descrito no seguinte teorema:

Teorema 4.1

Suponhamos $F(x_*) = 0$, $J(x_*)$ não singular e contínuo em x_* , e $\theta_k < \theta_{max} < \theta < 1$. Então existe $\varepsilon > 0$ tal que, se $||x_0 - x_*|| < \varepsilon$, a seqüência $\{x_k\}$, obtida usando (4.2) e $x_{k+1} = x_k + s_k$, converge a x_* e satisfaz

$$||x_{k+1} - x_*||_* < \theta ||x_k - x_*||_* \tag{4.3}$$

para todo $k > 0$, onde $||z||_* = ||J(x_*)z||$. Se $\lim_{k\to\infty} \theta_k = 0$, a convergência é superlinear.

Prova

Veja Dembo, Eisenstat e Steihaug[20]. ■

Em geral os métodos iterativos não são eficientes, a menos que implementados com precondicionadores adequados. Por precondicionamento, entendemos que o sistema original (4.1) é substituido por um sistema equivalente, porém mais fácil de resolver:

$$B_k^{-1} J(x_k) s_k = -B_k^{-1} F(x_k) \tag{4.4}$$

onde B_k^{-1}, ou melhor, $B_k^{-1} z$ deve ser fácil de calcular e $B_k \approx J(x_k)$.

Na resolução de (4.4), o primeiro incremento a ser tentado deve ser do tipo:

$$s_k^0 = -\lambda_k B_k^{-1} F(x_k) \, . \tag{4.5}$$

Tal incremento é aceito se satisfaz (4.2). Esta é uma característica comum aos sistemas precondicionados.

Existem algumas técnicas clássicas de precondicionamento para (4.1) :

(a) Precondicionadores baseados em fatorações incompletas (Golub e Van Loan[25], Axelsson [1,2], Martínez[38]) : A idéia básica consiste em obter uma matriz de precondicionamento, digamos M, que seja "próxima", da matriz do sistema original, digamos A, de maneira que a estrutura de esparsidade seja manipulada com facilidade. Um critério simples é que os fatores da matriz M sejam nulos nas posições nas quais os fatores de A o são.

(b) Precondicionadores baseados no problema físico subjacente (Glowinski, Keller e Reinhard[24]). A escolha do precondicionador considera as características funcionais próprias do problema físico. Por exemplo, podemos considerar apenas (ou desprezar) contribuições das derivadas em relação a algumas componentes que tenham muita (ou pouca) influência no problema como um todo.

(*c*) Precondicionadores baseados em outros métodos iterativos (Young [61]). Um determinado método pode ser formulado de maneira a ser acelerado por um outro método iterativo. Por exemplo, o método do Gradiente Conjugado, pode ser precondicionado usando a matriz de um método iterativo estacionário, como Jacobi, onde a matriz de precondicionamento é diagonal. Alternativamente, o Gradiente Conjugado precondicionado por uma matriz M genérica, via fórmula de três termos, é um acelerador de métodos iterativos que tem a forma $Mx_{l+1} = Nx_l + c$.

A seção que se segue descreve o uso dos Precondicionadores Secantes.

4.2 PRECONDICIONADORES SECANTES

Uma técnica de precondicionadores secantes foi desenvolvida inicialmente por Martínez [43]. Definimos, para todo $k \in \mathbb{N}$,

$$x_k^Q = x_k - B_k^{-1}F(x_k), \tag{4.6}$$

$$x_k^N = x_k - J(x_k)^{-1}F(x_k); \tag{4.7}$$

x_k^Q é obtido por algum método quase-newtoniano e x_k^N, obviamente, não é computado no processo.

Dada uma norma $|.|$ em $\mathbb{R}^n$ escolhemos x_{k+1} tal que

$$|x_{k+1} - x_k^N| \leq |x_k^Q - x_k^N| \tag{4.8}$$

Se $|.| = ||.||_2$, a condição (4.8) é obtida usando um número arbitrário de iterações de Gradientes Conjugados (GC) aplicado ao sistema $J(x_k)s_k = -F(x_k)$ (Hestenes e Stiefel

[33] p. 416). Martínez[40] provou que a convergência do esquema (4.6)-(4.8) é idêntica à de um método baseado apenas em (4.6). De fato, as mesmas hipóteses que garantem a convergência local dos métodos quase-newtonianos puros ($x_{k+1} = x^Q$), também garantem a convergência local do processo baseado em (4.8).

A tendência atual para resolver sistemas lineares não simétricos, é usar gradientes conjugados generalizados, ou mais precisamente, métodos de minimização do resíduo em subespaços de Krylov (Saad e Schultz[49], Saad[50], Young[61], Brown[5,6]). Estes métodos (e mais concretamente o GMRES de Saad e Schultz[1986]), quando aplicado a um sistema linear $Ax = b$ trabalham diretamente com a matriz A e não com sua transformação A^TA usada no algoritmo clássico de gradientes conjugados. Por outro lado, são implementados com memória limitada (veja seção 4.4), já que a convergência finita se baseia na acumulação de um vetor adicional na memória por iteração. Porém, o GMRES não possui a propriedade de norma decrescente do erro, como o GC clássico, o que é necessário para a abordagem de Martínez [40]. Por exemplo, consideremos o sistema linear definido por,

$$A = \begin{bmatrix} 0 & 0 & -1 \\ -3 & 0 & 0 \\ 0 & -2 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} -1 \\ -3 \\ -2 \end{bmatrix}$$

com $x_0 = (0,0,0)^T$. Considerando $||e_i|| = ||x_i - x_*||$, temos $||e_1|| \cong 2/3$ e $||e_2|| \cong 9/10$.

Este fato motiva a seguinte discussão:

Sabemos que, de acordo com a teoria de Dembo, Eisenstat e Steihaug [20], obtem-se convergência linear dos métodos de Newton Inexatos se, em cada iteração, o critério de parada do método iterativo linear subjacente for:

$$||J(x_k)s_k + F(x_k)|| \leq \theta_k ||F(x_k)||, \tag{4.9}$$

ressaltando que a convergência superlinear, que é independente da norma, é obtida se $\lim_{k\to\infty} \theta_k = 0$. Isto significa que, para se obter convergência superlinear, precisaríamos

que o critério de parada do método iterativo linear (digamos GMRES) fosse cada vez mais exigente, portanto o número de iterações do método iterativo linear seria cada vez maior, ou seja, o custo de cada iteração principal aumentaria com k, o que não é prático.

Martínez sugeriu contornar este problema através do uso de precondicionadores secantes. Com efeito, (4.1) não é um sistema de equações isolado. É muito provável que $J(x_k) \approx J(x_{k+1})$ especialmente quando k é grande. Este fato motiva a utilização de informação anterior (B_k, $F(x_k)$, $F(x_{k+1})$, x_{k+1} e x_k) quando escolhemos o precondicionador B_{k+1}. A idéia é exigir que o precondicionador satisfaça a equação secante (veja cap. 2). Logo, gostaríamos de introduzir um algoritmo baseado em (4.8) onde a seqüência de precondicionadores gerados são escolhidos de modo a satisfazer a equação

$$B_{k+1}s_k = y_k \tag{4.10}$$

onde $y_k = F(x_{k+1}) - F(x_k)$, para todo $k \in \mathbb{N}$.

Veremos mais adiante algumas fórmulas secantes utilizadas para gerar os precondicionadores.

Descrevemos o seguinte algoritmo básico, que implementa essa idéia.

Algoritmo 4.1

Sejam $x^0 \in \mathbb{R}^n$ arbitrário, $B_0 \in \mathbb{R}^{n\times n}, \theta \in (0,1)$ e $\{\theta_k\}$ uma seqüência que tende a zero. Dados x_k e B_k, executar:

Passo 1 (Passo Secante)

Resolver

$$B_k s_Q = -F(x_k)$$

Passo 2 (Iterativo Linear)

Se s_Q satisfaz $||J(x_k)s_k^Q + F(x_k)|| \leq \theta||F(x_k)||$, definir $s_k = s_k^Q$ e executar o Passo 3.

Caso contrário, usar um método iterativo linear até conseguir

$$||J(x_k)s_k + F(x_k)|| \leq \theta_k||F(x_k)|| \,.$$

Passo 3 (Novo Ponto)

$$x_{k+1} = x_k + s_k$$

Passo 4 (Atualização)

Atualizar B_k^{-1} para B_{k+1}^{-1} usando uma fórmula Secante. □

Martínez [43] provou que, quando se usa uma fórmula secante, enquadrada na teoria do captítulo 2 no passo 4 do algoritmo 4.1, este algoritmo tem convergência superlinear.

Um resultado fundamental da teoria de precondicionadores secantes é o seguinte:

Teorema 4.2

Suponhamos $F : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$, Ω um conjunto aberto e convexo, $F \in C^1(\Omega)$, $J(x_*)$ não singular, $F(x_*) = 0$. Suponhamos também que (3.10) se verifica, $||B_k||$ e $||B_k^{-1}||$ são limitadas e que

$$\lim_{k\to\infty} \frac{||[B_k - J(x_*)](x_{k+1} - x_k)||}{||x_{k+1} - x_k||} = 0. \tag{4.11}$$

Então, existe $\varepsilon > 0$ tal que, se $||x_0 - x_*|| \leq \varepsilon$, a seqüência $\{x_k\}$ gerada pelo algoritmo 4.1 converge superlinearmente a x_*. Além disso, existe $k_0 \in I\!N$ tal que $s_k = s_k^Q$ para todo $k \geq k_0$.

Prova.

Veja Martínez [43]. ■

O Teorema 4.2 afirma que, se usarmos precondicionadores que satisfazem a condição de Dennis-Moré (4.11), a convergência superlinear é obtida sem $\lim_{k\to\infty} \theta_k = 0$. De fato, o incremento inicial s_k^Q do método iterativo linear satisfará o teste do passo 2 do algoritmo 4.1, e portanto será aceito como o incremento corrente s_k, preservando a superlinearidade.

Uma condição suficiente para que um precondicionador secante satisfaça as hipóteses do teorema 4.2, é que

$$\lim_{k\to\infty} ||B_{k+1} - B_k|| = 0\,, \tag{4.12}$$

que é o caso dos métodos exemplificados no cap. 2, entre eles o método de Broyden. Existem, no entanto, métodos que, embora não se enquadrem perfeitamente na teoria do cap. 2, apresentam bom desempenho prático (veja sec. 3.1), como o ICUM descrito no capítulo 3. De fato, não sabemos se (4.12) se verifica para este método.

Os primeiros resultados práticos obtidos quando da implementação do Algoritmo 4.1, serão vistos ainda neste capítulo.

4.3 ALGORITMOS

Como vimos na seção anterior, o precondicionamento para o sistema linear (4.1) será representado pela matriz B_k^{-1}, ou ainda H_k, como no capítulo 3 para ICUM, ambas aproximação para $J(x_k)^{-1}$. No algoritmo do método iterativo linear (veja seção 4.4) não precisaremos explicitar H_k e sim formar o produto $H_k v$, onde $v \in \mathbb{R}^n$.

Gomes-Ruggiero [26] mostra a efetividade do primeiro método de Broyden assim como do método CUM, como já comentado no capítulo 3, onde o mesmo foi feito para o método ICUM. Além destes, vamos considerar o segundo método de Broyden, que denominaremos BROYDEN2 em oposição ao primeiro (veja exemplo 2.1 cap. 2), que iremos referenciar como BROYDEN1.

Assim como o primeiro, o segundo método de Broyden se enquadra na teoria LCSU do capítulo 2, gozando da propriedade de convergência superlinear. Analogamente ao exemplo 2.1, podemos definir a variedade linear em função da forma inversa da equação secante:

$$V(x,z) = \{H \in X \mid H(F(z) - F(x)) = z - x\} \tag{4.13}$$

onde $H \in \mathbb{R}^n$ é uma aproximação para $J(x)^{-1}$.

A fórmula de atualização para o segundo método de Broyden, dada em [18]:

$$H_{k+1} = H_k - \frac{(s_k - H_k y_k)}{y_k^T y_k} y_k^T \tag{4.14}$$

Esta expressão pode ser obtida exigindo que H_{k+1} seja a matriz de $V(s_k, y_k)$ mais próxima de H_k, considerando a norma de Frobenius. Geometricamente, isto significa que H_{k+1} é a projeção ortogonal de H_k em $V(s_k, y_k)$. É isto que caracteriza o seguinte teo-

rema:

Teorema 4.3

Dada uma matriz $H \in \mathbb{R}^n$ e os vetores $s, y \in \mathbb{R}^n$, $y \neq 0$. A matriz $\overline{H}$, dada por

$$\overline{H} = H - \frac{(s - Hy)}{y^T y} y^T$$

é a solução única do problema

$$Min||M - H||_F$$

sujeito a

$$M \in V(s, y).$$

Prova

Veja Dennis e Moré [17] ■

Em verdade, (4.14) é um caso especial da formulação mais geral para a atualização da inversa do Jacobiano:

$$H_{k+1} = H_k - \frac{(s_k - H_k y_k)}{z_k^T y_k} z_k^T \tag{4.15}$$

onde $z_k = y_k$ para o segundo método de Broyden e $z_k = e_{j_k}$, $|y^T e_{j_k}| = ||y_k||_\infty$ para o método ICUM. Aqui, $\{e_1, ..., e_n\}$ é a base canônica de $\mathbb{R}^n$.

(4.15) é uma fórmula de correção de posto um. Isto significa que o padrão de esparsidade de H_k para H_{k+1} não é preservado, logo H_{k+1} pode resultar densa quando H_k for esparsa. Lembrando mais uma vez que é interessante obter $H_{k+1}v$, vamos estabelecer a forma de memória limitada.

A fórmula (4.15) mostra que H_{k+1} pode ser obtido de H_k usando $O(n^2)$ operações no caso denso. Além disso,

$$H_{k+1} = H_k + u_k z_k^T \tag{4.16}$$

com

$$u_k = \frac{(s_k - H_k y_k)}{z_k^T y_k}\,. \tag{4.17}$$

Logo

$$H_{k+1} = H_k + u_0 z_0^T + u_1 z_1^T + ... + u_k z_k^T \tag{4.18}$$

e

$$H_{k+1} v = H_0 v + u_0 z_0^T v + u_1 z_1^T v + ... + u_k z_k^T v. \tag{4.19}$$

A fórmula (4.18) deve ser usada para n grande. Neste caso os vetores $u_0, z_0, ..., u_k, z_k$ são armazenados e o produto $H_{k+1}v$ calculado por (4.19). Desta maneira, o custo da iteração k é $O(kn)$ mais o custo de calcular $H_0 v$.

Se k for grande, o processo deve ser recomeçado periodicamente com $H_l \approx J(x_l)^{-1}$ a cada, digamos, m iterações : $l = 0, m, 2m, ...$ Portanto, a fórmula (4.19) toma a forma:

$$H_{k+1} v = H_l v + u_l z_l^T v + ... + u_k z_k^T v \tag{4.20}$$

O algoritmo 4.2, que se segue, fornece a implementação de (4.20), onde são armazenados no máximo m pares de vetores para uma iteração típica, digamos k. Para simplificar, suponha que estamos interessados em obter $H_k v$.

Algoritmo 4.2 : Obtenção de $H_k v$ por (4.17 - 4.20) (iteração k).

Dado um número inteiro m, e um parâmetro β execute:

Passo 1

Se $k \equiv 0 \; mod(m)$,

obtenha H_l, faça $t = H_l v$, e vá para o passo 5

Caso contrário, faça $t = H_l v$.

Passo 2

Para $j = l, \ldots, k-2$. Faça

$t \leftarrow t + u_j z_j^T v$.

Passo 3

Faça

$$u_{k-1} = \frac{(s_{k-1} - H_{k-1} y_{k-1})}{z_{k-1}^T y_{k-1}}.$$

Passo 4

Faça

$t \leftarrow t + u_{k-1} z_{k-1}^T v$.

Passo 5

Obtenha $\alpha = min\{1, \frac{\beta}{||t||_\infty}\}$,

e faça $s_k^Q = \alpha t$. □

H_l será dada por uma fatoração da matriz Jacobiana truncada (veja sec. 4.5) sempre que $l \equiv 0\ mod(m)$.

A condição para que H_{k+1} seja não singular é $z_k^T y_k \neq 0$, ou ainda, $y_k = F(x_{k+1}) - F(x_k) \neq 0$. Se

$$||F(x_{k+1}) - F(x_k)|| \leq TOL||F(x_k)||,$$

fazemos $H_{k+1} = H_k$, para um parâmetro TOL suficientemente pequeno.

Semelhante desenvolvimento pode ser feito para BROYDEN1 e CUM. A expressão para BROYDEN1, no exemplo 2.1 (cap. 2), é dada em termos da matriz B_k. A forma $H_k \equiv (B_k^{-1})$, é obtida através da fórmula de Sherman-Morrison [25]:

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)}{s_k^T H_k y_k} s_k^T H_k;. \tag{4.21}$$

Através desta expressão, obtem-se a forma de memória limitada [26]:

$$H_{k+1} v = (I + u_k z_k^T)...(I + u_1 z_1^T)(I + u_l z_l^T) H_l v \tag{4.22}$$

com

$$u_i = \frac{(y_i - H_i s_i)}{z_i^T H_i y_i}, \tag{4.23}$$

onde $z_k = s_k$ para o segundo método de Broyden e $z_k = e_{jk}$, $|s^T e_{jk}| = ||s_k||_\infty$ para o

método CUM.

Assim, temos o seguinte algoritmo, na iteracão k:

Algoritmo 4.3 : Obtenção de $H_k v$ por (4.22 - 4.23) (iteracão k).

Dado um número inteiro m, e um parâmetro β execute:

Passo 1

Se $k \equiv 0 \; mod(m)$,

obtenha H_l, faça $t = H_l v$, e vá para o passo 5

Caso contrário, faça $t = H_l v$,

Passo 2

Para $j = l, \ldots, k-2.$ faça

$t \leftarrow (I + u_j w_j^T)t.$

Passo 3

Faça

$$u_{k-1} = \frac{(s_{k-1} - H_{k-1} y_{k-1})}{z_{k-1}^T H_{k-1} y_{k-1}}$$

Passo 4

Faça

$$t \leftarrow t + u_k w_k^T t.$$

Passo 5

Obtenha $\alpha = min\{1, \frac{\beta}{||t||_\infty}\}$

e faça $s_k^Q = \alpha t$. □

A condição para que H_{k+1} seja não singular é que $z_k^T H_k y_k \neq 0$. Considerando $r = H_k y_k$, se

$$|z_k^T r|| < TOL||z_k||||r||.$$

fazemos $H_{k+1} = H_k$.

4.4 O GMRES

Faremos, agora, uma breve introdução sobre o método iterativo linear que utilizamos nos testes numéricos, o GMRES [51], dando ênfase aos aspectos práticos.

Para resolver um sitema linear $n \times n$, não singular,

$$Ax = b \tag{4.24}$$

busca-se a solucão aproximada x_k da forma $x_k = x_0 + z_k$, onde x_0 é a aproximação inicial e z_k pertence ao subespaço de Krylov: $K_k = [r_0, Ar_0, ..., A^{k-1}r_0]$, com $r_0 = b_0 - Ax_0$. O método consiste em gerar uma base ortonormal de K_k e, usando isto, minimizar a norma do resíduo neste subespaço:

$$\min_{z \in K_k} ||b - A[x_0 - z]|| = \min_{z \in K_k} ||r_0 - Az||. \tag{4.25}$$

Após o processo de ortogonalização (passo 2 do algoritmo 4.4 que se segue), obtem-se um sistema ortonormal V_{k+1} e uma matriz Hessenberg $(k+1) \times k, R_k$. Os vetores v_i e a matriz R_k satizfazem a relação

$$AV_k = V_{k+1}R_k \tag{4.26}$$

e, considerando $z = V_k y$, podemos reescrever (4.25),

$$\min J(y) = \min_y ||\beta e_1 - R_k y|| \tag{4.27}$$

onde e_1 é o vetor canônico de $\mathbb{R}^{n+1}$ e $\beta = ||r_0||$.

Portanto, a fase seguinte consiste em resolver o problema de quadrados mínimos (4.27).

A aproximação corrente será dada por

$$x_k = x_0 + V_k y_k \tag{4.28}$$

onde y_k minimiza $J(y)$ em R^n.

O processo de ortogonalização é o passo crítico do método. Walker [57] propõe o emprego das transformações de Householder na construção das matrizes. Embora esta técnica conduza a bons resultados numéricos em alguns casos, é cerca de três vezes mais cara que o procedimento usual de Gram-Schimdt modificado [51].

Como comentamos anteriormente, a convergência finita se baseia na acumulação de um vetor adicional por iteração. Neste sentido, os recomeços ou, alternativamente, o truncamento da ortogonalização [57], são necessários, considerando n grande. Usaremos, na implementação do GMRES, os recomeços a cada, digamos, mg iterações. A escolha

de mg requer alguma técnica experimental adequada ao problema que se deseja resolver. Um valor reduzido de mg pode resultar em convergência lenta ou até mesmo na não convergência do método. Em [34] experimentos para escolha deste parâmetro são apresentados e discutidos.

O problema de quadrados mínimos relacionado é peculiar. De fato, tem a dimensão do parâmetro de recomeços mg, e é resolvido parcialmente durante o procedimento de ortogonalização, dada sua estrutura especial. Isto permite a obtenção da norma do resíduo sem explicitar a solução corrente [49].

O algoritmo 4.4 descreve a implementação do método GMRES, com a matriz de precondicionamento H, obtida dos Algoritmos 4.1 ou 4.2.

Algoritmo 4.3 : GMRES(mg)

Passo 1 (Inicialização)

Dado $x_0 \in \mathbb{R}^n$ e $H \in \mathbb{R}^{n \times n}$, faça

$r = H(b_0 - Ax_0)$ e $v_1 = \frac{r}{\|r\|}$.

Passo 2 (Ortogonalização)

Para $j = 1, \ldots, mg$, faça

$w = HAv_j$

Para $i = 1, \ldots, j$, faça

$$h_{i,j} = (w, v_i),$$

$$w = w - h_{i,j}v_i,$$

$h_{j+1,j} = ||w||,$

$v_{j+1} = w/h_{j+1,j}.$

Passo 3 (Solução corrente)

Obtenha a solução aproximada

$$x_{mg} = x_0 + V_{mg}y_{mg},$$

onde $y_{mg} = \min_y \ ||\beta e_1 - R_{mg}y||$, $y \in \mathbb{R}^n$.

Passo 4 (Recomeço)

Calcule $r = H(b_{mg} - Ax_{mg})$.

Se $||r|| < EPS$, pare.

Senão, faça $x_0 = x_{mg}$ e $v_1 = \frac{r}{||r||}$.

Vá para o passo 2. □

Observemos que, em nosso contexto, $EPS = \theta_k||F(x_k)||$.

Finalmente, embora o algoritmo apresente certo grau de simplicidade, quanto à implementação, deve-se considerar que os problemas que justificam emprego dos métodos

iterativos são, em geral, grandes e esparsos, fazendo da implementação algo não trivial. Cabe ainda salientar a possibilidade de paralelismo e/ou vetorização do método.

4.5 OS PROBLEMAS

Nesta seção descrevemos a formulação de dois problemas práticos ligados à engenharia:

O Problema de Fluxo de Carga [22]

Fluxo de Carga ou Fluxo de Potência, é a solução para a condição de um sistema de transmissão de potência elétrica.

O objetivo fundamental do cálculo de fluxo de carga, é a determinação das tensões e das injeções de potência em todos os nós do sistema de transmissão (rede), sob determinadas condições de geração de carga.

As equações que modelam o comportamento dos principais componentes da rede são dadas por:

$$P_k(\alpha, v) = v_k \sum_{m \in K_k} v_m (G_{km} cos\alpha_{km} + B_{km} sen\alpha_{km}) \tag{4.29}$$

$$Q_k(\alpha, v) = v_k \sum_{m \in K_k} v_m (G_{km} sen\alpha_{km} - B_{km} cos\alpha_{km}) \tag{4.30}$$

onde $P_k(\alpha, v)$ e $Q_k(\alpha, v)$ representam as injeções liquidas de potência ativa e reativa respectivamente, v_k representa a magnitude da tensão no nó k; G_{km} e B_{km} são as componentes da matriz de admitancia: $Y_{km} = Gkm + iB_{km}$; $\alpha = \alpha_k - \alpha_m$ é a abertura do ramo km e K_k é o conjunto dos nós vizinhos ao nó k.

Na formulação mais simples do problema, a cada nó estão associadas quatro variáveis P_k, Q_k, α_k e v_k. Os nós do sistema são classificados em três tipos, dependendo das quantidades que entram com incógnitas:

Nós folga : nós k onde v_k e α_k são dados.
Nós do tipo A : nós k onde P_k e α_k são dados.
Nós do tipo B : nós k onde Q_k e v_k são dados.

Se N representa o total de nós, o sistema de potência é formado por $2N$ equações para as quais as variáveis são precisamente as incógnitas em cada nó.

Após algumas simplificações algébricas, as equações correspondentes aos nós folga podem ser eliminadas, assim como pode ser eliminada uma equação para cada nó do tipo B. Supondo que temos S nós do tipo B e T nós folga, teremos um sistema de $n = 2N - S - 2T$ equações e variáveis.

O Problema da Cavidade [35]

Um exemplo clássico em mecânica dos fluidos, é a modelagem do problema da Cavidade, o qual é formulado em relação à função corrente, originando uma equação diferencial parcial não linear de quarta ordem que, discretizada, produz um sistema algébrico não linear.

As equações de Navier-Stokes, para um fluido imcompressível estacionário em duas dimensões origina a seguinte equação bi-harmônica:

$$P(\psi) = \Delta^2\psi + Re[\varphi_x(\Delta\psi)_y - \varphi_y(\Delta\psi)_x] = 0 \tag{4.31}$$

onde ψ é a função corrente, Re é o número de Reynolds e Δ o Laplaciano. Acrescentando

condições de fronteira apropriadas, completa-se a formulação do problema de contorno.

A cavidade é considerada como sendo o quadrado $R = \{(x,y) \in I\!R^2 \text{ tq. } 0 \leq x,y \leq 1\}$ com o lado superior aberto em contato com o fluido, com as condições de fronteira como descrita em [35].

A solução de (4.31) é aproximada por um esquema de diferenças finitas (veja [35]) e o sistema resultante, considerando uma malha quadrada de $n \times n$ nós, pode ser escrito como

$$F(x, Re) = L(x) + ReG(x) = 0 \tag{4.32}$$

onde $L(x)$ corresponde à discretização de $\nabla^2\psi$ e $G(x)$ corresponde à discretização do respectivo termo seguinte de (4.31), com $F : (f_1, f_2, ..., f_{n\times n})$ e $X : (x_1, x_2, ..., x_{n\times n})$.

Usando tal discretização, e variando o número de Reynolds, obtemos diferentes sistema não lineares do tipo (4.31), onde a não linearidade aumenta com Re.

4.6 EXPERIMENTOS NUMÉRICOS

Nesta seção apresentamos os resultados experimentais obtidos da implementação computacional do algoritmo 4.1 para os dois problemas descritos acima. O método iterativo linear utilizado é o GMRES descrito na seção 4.4. Fazemos a comparação com o método de Newton Inexato puro (sem precondicionamento), e com o precondicionador gerado pela fatoração do Jacobiano truncado, que iremos referenciar simplesmente como fatoração truncada.

Esta fatoração é feita na matriz resultante da seleção de um determinado número de diagonais à direita e à esquerda da diagonal principal de $J(x_k)$ (caracterizando a banda da matriz), e descartando as restantes.

Definimos os seguintes parâmetros:

θ_k : Parâmetro de tolerância, usado para o critério de parada do método iterativo (veja Algoritmo 4.1).

mg : Parâmetro de recomeços do GMRES (veja algoritmo 4.2).

ban : Número de diagonais à esquerda/direita da diagonal principal da matriz jacobiana consideradas na fatoração truncada.

tempo : tempo de execução em segundos de CPU.

Os métodos serão indicados por:

NIP : Método de Newton Inexato puro.

NIFT : Método de Newton Inexato precondicionado por fatoração truncada.

Referimo-nos a um precondicionador secante com Newton-Inexato identificando o precondicionador. Por exemplo, quando referenciarmos *broyden*1, significa que este precondicionador foi usado com Newton-Inexato.

Os números contidos no corpo principal das tabelas referem-se ao número de iterações GMRES, utilizados numa iteração de Newton Inexato. Por exemplo, para o primeiro problema, na tabela 1, na linha que representa o método de Newton Inexato puro, NIP, a primeira iteração requer 1 iteração GMRES, a segunda 6, e assim por diante, até a convergência, obtida na sétima iteração com 29 iterações GMRES. Ficam assim subentendido os "brancos" das tabelas.

Os testes que executamos envolveram um número relativamente "pequeno" de iterações principais. Isto parece caracterizar os métodos de Newton-Inexatos. Por este motivo, não foram necessários recomeços (representados pelo parâmetro m nos algorit-

mos 4.1 e 4.2) na obtenção dos precondicionadores, contrariamente ao que ocorreu no experimento numérico do cap. 3.

Os testes foram executados em uma SUN Workstation SPARC 2, usando o compilador Fortran 77, no Laboratório do Departamento de Matemática Aplicada da UNICAMP.

As tabelas abaixo, mostram os resultados obtidos para o problema de fluxo de Carga.

Resultados para $n = 54$, (30 nós).
$ban = 4$.

	ITERAÇÃO								
MÉTODO	1	2	3	4	5	6	7	8	TEMPO
NIP	1	6	20	24	27	30	29		0.54
NIFT	0	3	5	6	7	7	10	14	0.43
BROYDEN 1	0	3	4	2	4	7	4	6	0.33
BROYDEN 2	0	3	4	3	5	6	7	2	0.30
CUM	0	3	4	3	6	5	6	5	0.30
ICUM	0	3	4	2	5	8	6		0.27

Tabela 1: $\theta_k \equiv 0.9/k$

	ITERAÇÃO					
MÉTODO	1	2	3	4	5	TEMPO
NIP	24	29	32	32	32	0.40
NIFT	8	5	6	10	8	0.23
BROYDEN 1	8	5	6	9		0.18
BROYDEN 2	8	5	6	9		0.19
CUM	8	5	6	8	6	0.23
ICUM	8	5	6	9	6	0.23

Tabela 2: $\theta_k \equiv 0.1$

	ITERAÇÃO							
MÉTODO	1	2	3	4	5	6	7	TEMPO
NIP	16	21	22	25	127	27		0.41
NIFT	7	3	4	6	8	8	8	0.32
BROYDEN 1	7	3	4	9	3	4	5	0.28
BROYDEN 2	7	3	4	9	4	3	7	0.30
CUM	7	3	4	9	3	3	4	0.26
ICUM	7	3	4	10	5	4	7	0.29

Tabela 3: $\theta_k \equiv 0.2$

Resultados para $n = 182$, (118 nós).
$ban = 4$.

	ITERAÇÃO									
MÉTODO	1	2	3	4	5	6	7	8	9	TEMPO
NIP	1	5	14	31	32	37	48	54		4.08
NIFT	não converge									
BROYDEN 1	0	1	3	5	14	13	9	14	17	2.97
BROYDEN 2	0	1	3	6	15	13	16	13		2.63
CUM	0	1	4	5	13	11	11	14	18	2.85
ICUM	0	1	3	6	14	15	13	15		2.44

Tabela 4: $\theta_k \equiv 0.9/k$

	ITERAÇÃO				
MÉTODO	1	2	3	4	TEMPO
NIP	48	57	85	98	6.89
NIFT	12	17	25	23	1.77
BROYDEN 1	12	22	23		1.36
BROYDEN 2	12	22	23		1.26
CUM	12	22	23		1.35
ICUM	12	22	23		1.39

Tabela 5: $\theta_k \equiv 0.01$

	ITERAÇÃO					
MÉTODO	1	2	3	4	5	TEMPO
NIP	15	40	31	53	53	3.20
NIFT	não converge					
BROYDEN 1	5	7	16	17	12	1.67
BROYDEN 2	5	7	16	16	13	1.70
CUM	5	7	16	16	14	1.63
ICUM	5	7	16	17	10	1.61

Tabela 6: $\theta_k \equiv 0.1$

A não convergência indicada nas tabelas para NIFT, deram-se em função da instabilidade da fatoração do Jacobiano truncado.

Nos testes para este problema não foram necessário os recomeços para o GMRES, em virtude do reduzido número de iterações exigido por este método iterativo.

Outro teste ao nosso alcance para este problema, é a formulação para $n = 2190$. Neste caso não obtivemos convergência para Newton-Inexato.

As tabelas abaixo mostram os resultados obtidos para o problema da Cavidade.

Nos testes, resolvemos o sistema (4.32) para $Re = 250, 500$ usando a solução de $F(x, Re) = 0$ como ponto inicial para a resolução de $F(x, Re + 250) = 0$. O teste de parada é o mesmo de [35]. Os recomeços GMRES foram feitos a cada 150 iterações ($mg = 150$). Também usamos $ban = 2$.

Semelhante comportamento ocorrereu para diferentes malhas. Apresentamos aqui os resultados obtidos dividindo o intervalo $[0, 1]$ em 32 subintervalos. Assim, o sistema não linear tem $29 \times 29 = 841$ variáveis e equações.

	ITERAÇÃO								
MÉTODO	1	2	3	4	5	6	7	8	TEMPO
NIP	1	3	12	108	127	140	145	246	123.09
NIFT	0	1	5	65	95	103	107	110	69.52
BROYDEN 1	0	2	6	71	91	101	104	110	73.05
BROYDEN 2	0	2	7	73	89	75	96	102	63.09
CUM	0	2	6	66	92	96	98	94	63.08
ICUM	0	2	5	65	97	102	95	105	68.52

Tabela 7: Re = 0

	ITERAÇÃO									
MÉTODO	1	2	3	4	5	6	7	8	9	TEMPO
NIP	2	64	113	181	825	250	1650	1594		809.15
NIFT	0	27	51	72	237	84	275	277		165.82
BROYDEN 1	0	12	32	71	122	131	124	327	473	288.21
BROYDEN 2	0	5	40	73	129	266	248	300	396	270.56
CUM	0	3	30	68	97	260	111	407	412	242.46
ICUM	0	2	58	71	145	125	274	118	378	205.32

Tabela 8: Re = 250

	ITERAÇÃO										
MÉTODO	1	2	3	4	5	6	7	8	9	10	TEMPO
NIP	1	34	134	449	1650	1650	1650	1650	1650	1650	1886.84
NIFT	0	14	54	104	293	136	447	742	718		463.87
BROYDEN 1	0	6	40	98	270	281	446	446	472		397.74
BROYDEN 2	0	5	42	105	275	245	585	521			337.19
CUM	0	5	12	101	314	137	0	692			388.08
ICUM	0	6	41	98	426	872	436	723	568		597.30

Tabela 9: Re = 500

4.7 CONCLUSÕES

Os resultados numéricos nos mostram a efetividade das técnicas precondicionadoras quando usamos métodos de Newton-Inexatos. Os precondicionadores secantes, nesse sentido, apresentam bons resultados, mostrando ser uma boa alternativa de precondicionamento de sistemas não lineares. De fato, como podemos observar pelos experimentos, o custo de obter tais precondicionadores é muito reduzido em relação ao benefício que podem trazer, reduzindo significativamente o número de iterações do método iterativo linear.

Para o problema de fluxo de cargas, os precondicionadores secantes mostram melhor desempenho, mesmo quando comparado ao precondicionador por fatoração do Jacobiano truncado. Para este problema, a não convergência, referenciada anteriormente, se deu na medida do aumento de k. Como a fatoração inicial não apresenta problemas, os precondicionadores secantes foram capazes de manter $B_k \approx J(x_k)$, garantindo a convergência do processo.

No problema da Cavidade, os precondicionadores apresentaram resultados semelhantes. A vantagem, verificada na tabela 8, para NIFT, sugere que alguma característica intrínseca do problema favoreça ou não uma ou outra técnica.

Testes preliminares com o problema de simulação de Reservatórios [35], mostram que o método de Newton-Inexato, com as técnicas de precondicionamento aqui descritas, apresentam desempenho muito superior quando comparados aos métodos quase-newtonianos em [26], contrariamente ao que ocorreu no problema da Cavidade.

Observamos algumas dificuldades práticas para a verificação do terorema 4.2. O processo de convergência se dá em poucas iterações, e uma análise assintótica dificilmente abrange este fato com propriedade. De qualquer maneira, acreditamos que testes adicionais possam nos conduzir a resultados superiores no uso dos precondicionadores secantes, na direção do teorema (4.2). É neste sentido que estamos direcionando nossas pesquisas

Bibliografia

[1] Axelsson, O. [1977] Solution of linear systems of equations: Iterative Methods, Lectures Notes in Math., vol. 572, Springer-Verlag, Berlin and N.Y. pp. 1-51.

[2] Axelsson, O. [1980] A generalized conjugate direction method and its application to a singular perturbation problem. Lecture Notes in Math. vol. 773, Springer-Verlag, Berlin and N.Y. pp 1-11.

[3] Barnes, J.G.P. [1965]: An algorithm for solving nonlinear equations based on the secant method, Computer Journal 8, pp. 66-72.

[4] Brown, P. [1987] A local convergence theory for combined Inexact-Newton finite - difference projection methods. SIAM J. Numer. Anal. vol. 24, Nº 2.

[5] Brown, P.N. [1990]: Hybrid Krylov methods or nonlinear systems of equations, SIAM J. Sci Stat. Comput. 11, pp. 450-481.

[6] Brown, Saad, Y. [1989]: Globally convergent techniques in nonlinear Newton-Krylov algorithms, Technical Report, Lawrence National Laboratory, UCRL 102434.

[7] Broyden, C.G. [1965]: A class of methods for solving nonlinear simultaneous equations, Math. Comput. 19, pp. 577-593.

[8] Broyden, C.G. [1971]: The convergence of an algorithm for solving sparse nonlinear systems, Math. Comput. 25, pp. 285-294.

[9] Broyden, C.G.; Dennis, J.E., Jr; Moré, J.J. [1973]: On the local and superlinear convergence of quasi-Newton methods, J. Inst. Math. Appl. 12, pp. 223-245.

[10] Chen, X. [1990]: On the convergence of Broyden-like methods for nonlinear equations with nondifferentiable terms, Annals of the Institute of Statistics and Mathematics 42, pp. 387-401.

[11] Chen, X. and Yamamoto, T. [1989]: Convergence domains of certain iterative methods for solving nonlinear equations, Numerical Functional Analysis and Optimization 10, pp. 37-48.

[12] Davidon, W.C. [1959], Variable metric method for minimization, Rep. ANL-5990 Rev. Argonne National Laboratories, Argone, Ill.

[13] Dennis, J.E. Jr. [1971]: Towards a unified convergence theory for Newton-like methods. Nonlinear Functional Analysis and Applications, Academic Press, New York, 425-472.

[14] Dennis, J.E.; Martínez, J.M. [1990]: Numerical methods for solving nonlinear systems, em Handbook of Numerical Analysis, P.G. Ciarlet and J.L. Lions (editors), Elsevier-North-Holland (em preparação).

[15] Dennis, J.E., Jr.; Marwil, E.S. [1982]: Direct secant updates of matrix factorizations, Math. Comput. 38, pp. 459-476.

[16] Dennis, J.E., Jr.; Moré, J.J. [1974]: A characterization of superlinear convergence and its application to quasi-Newton methods, Math. Comp. 28, pp. 549-560.

[17] Dennis, J.E. Jr. Moré, J.J. [1977]: Quasi-Newton methods, motivation and theory, SIAM Review 19, pp. 46-89.

[18] Dennis, J.E., Jr.; Schnabel, R.B. [1983]: Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice Hall, Englewood Cliffs, N.J.

[19] Dennis, J.E., Jr. and Walker, H.F. [1981]: Convergence theorems for least-change secant update methods, SIAM J. on Numer. Anal. 18, pp. 949-987.

[20] Dembo, R.S., Eisenstat, S. C., Steihaug, T.[1982]: Inexact Newton methods, SIAM J. Num. Anal. 14, pp. 400-408.

[21] Duff, I.S.; Erisman, A.M.; Reid, J.K. [1986]: Direct Methods for Sparse Matrices, Clarendon Press, Oxford.

[22] Duran, A.C. [1990]: Resolução de sistemas não lineares esparsos: sua aplicação na resolução do problema de fluxo de carga em redes de energia elétrica, Tese de Mestrado, Depto. de Matemática Aplicada, IMECC-UNICAMP, Campinas, Brasil.

[23] Gay, D.M. [1979]: Some convergence properties of Broyden's method SIAM J. Numer. Anal. 16, pp. 623-630.

[24] Glowinski, R.; Keller, H.B.; Reinhard, L. [1985]: Continuation conjugate-gradient methods for least squares solution of nonlinear boundary value problems, SIAM J. Sci. Stat. Comput. 6, pp. 793-832.

[25] Golub, G.H.; Van Loan, Ch. F. [1989]: Matrix Computations, The Johns Hopkins University Press, 2nd. edition, Baltimore and London.

[26] Gomes-Ruggiero, M.A. [1992]: Métodos quase-Newton para resolução de sistemas não lineares esparsos e de grande porte. Tese de Doutorado. FEE-UNICAMP-Campinas, Brasil.

[27] Gomes-Ruggiero, M.A.; Martínez, J.M. [1992]: The Column-Updating Method for solving nonlinear equations in Hilbert space, Mathematical Modelling and Numerical Analysis 26, pp 309-330.

[28] Gomes-Ruggiero, M.A., Martínez, J.M. and Moretti, A.C. [1991]. Comparing Algorithms for Solving Sparse Nonlinear Systems of Equations, SIAM J. Sci. Stat. Comput. V. 13 n. 2, pp. 459-483.

[29] Gragg, W.B. Stewart, G.W. [1976]: A stable variant of the secant method for solving nonlinear equations, SIAM Journal on Numer. Anal. 13, pp. 127-140.

[30] Griewank, A. [1986]: The Solution of boundary value problems by Broyden based secant methods, CTAC-85, Proc. of the Computational Techniques and Applications Conference, University of Melbourne, August 1985, J. Noye and R. May, eds., North Holland, Amsterdam.

[31] Griewank, A. [1992]: Achieving logarithmic growth of temporal and spacial complexity in reverse automatic differentiation, Optimization methods and sotftware, pp. 35 - 34.

[32] Hageman, A.L. and Young, D.M. [1981]: Applied Iterative Methods, Academic Press, New York.

[33] Hestenes, M.R.; Stiefel, E. [1952]: Methods of conjugate gradients for solving linear systems, Journal of Reserch of the National Bureau of Standards B49, pp. 409-436.

[34] Huang, Y. and Van Der Vorst, H.A. [1989]: Some observations on the convergence behavior of GMRES. Report 89-09, Delft University of Technology.

[35] Kozakevich, D.N. [1993]: resolução de sistemas não lineares da Física e da Engenharia. Tese de Doutorado, Depto. de Matemática Aplicada, IMECC-UNICAMP-Campinas, Brasil.

[36] Martínez, J.M. [1979a]: Three new algorithms based on the sequential method, BIT 19, pp. 236-243.

[37] Martínez, J.M. [1983]: A quasi-Newton method with a new updating for the LDU factorization of the approximate Jacobianm, Matemática Aplicada e Computacional 2, pp. 131-142.

[38] Martínez, J.M. [1987]: Quasi-Newton Methods with Factorization Scaling for Solving Sparse Nonlinear Systems of Equations, Computing 38, 133-141.

[39] Martínez, J.M. [1984]: A quasi-Newton method with modification of one column per iteration, Computing, 33, pp. 353-362.

[40] Martínez, J.M. [1990]: Local convergence theory of inexact Newton methods based on structured least change updates, Math. Comput., 5, $N^{\underline{o}}$ 191.

[41] Martínez, J.M. [1990]: A family of quasi-Newton methods for nonlinear equations with direct secant updates of matrix factorizations, SIAM J. Anal., (por aparecer).

[42] Martínez [1992]: On the relation between two local convergence theories of least change secant update methods, por aparecer em Math. Comp..

[43] Martínez [1992]: A theory of secant preconditioners, por aparecer em Math. Comp.

[44] Martínez, J.M.; Zambaldi, M.C. [1992]: An inverse Column-Updating Method for solving Large-Scale Nonlinear Systems of Equations, Optimization Methods and Software, V.1 ,pp. 129-140.

[45] Martínez, J.M.; Zambaldi, M.C. [1992]: Least Change Update methods for Nonlinear Systems with Nondifferentiable terms, RT n. 9/92 IMECC-UNICAMP, por aparecer em Numerical Functional Anaysis and Optimization.

[46] Matthies, H and Strang, G. [1970]: The solution of nonlinear finite element equations, Int. J. on Numerical Methods in Engineering 14, pp. 1613-1626.

[47] Moré, J.J. [1989]: A collection of nonlinear model problems, Preprint MCS - P60 - 0289, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois.

[48] Ortega, J.M. Rheinbolt, W.C. [1970]: Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York.

[49] Osterby, O., Zlatev, Z. [1982]: Direct Methods for Sparse Matrices, Lectures Notes in Computer Science, Nº 157, Springer-Verlag.

[50] Saad, Y. [1989]: Krylov Subspace methods on supercomputers, SIAM J. Sci. Stat. Comput. 10, pp. 1200-1232.

[51] Saad, Y. and Shultz, M.H. [1986]: GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM J. Numer. Anal, 7, pp. 859-869.

[52] Schwandt, H. [1984]: An interval arithmetic approach for the construction of an almost globally convergent method for the solution of the nonlinear Poisson equation on the unit square, SIAM J. Sci. Stat. Comput. 5, pp. 427-452.

[53] Sherman [1978] On Newton-iterative methods for the solution of systems of nonlinear equations, SIAM J. Anal. 15, pp. 755-771.

[54] Tewarson, R.P. [1988]: A New quasi-Newton Algorithm, Applied Mathematics Letters 1, pp. 101-104.

[55] Tewarson, R.P. and Zhang Y. [1987]: Sparse quasi-Newton LDU update, Int. J. Numerical Methods in Engineering 24, pp. 1093-1100.

[56] Toint, Ph. L. [1986]: Numerical solution of large sets of algebraic nonlinear equations, Math. Comp. 16, pp. 175-189.

[57] Walker H.F. [1988]: Implementation of the GMRES Method using Honseholder Transformations. SIAM J. Sci. Stat. Comput., 9:152-163.

[58] Wolfe, P. [1959]: The secant method for solving nonlinear equations, Communications ACM, 12, pp. 12-13.

[59] Yamamoto, T. [1987]: A note on a posteriori error bound of Zabrejko and Nguen for Zincenko's iteration, Numerical Functional ANalysis and Optimization 9, pp. 987-994.

[60] Yamamoto, T. and Chen, X. [1990]: Ball-convergence theorems and error estimates for certain iterative methods for nolinear equations, Japan Journal on Applied Mathematics, 7, pp. 131-143.

[61] Young, D.M. [1989]: A historical overview of iterative methods, Computer Physics Communications 53, pp. 1-18.

[62] Zabrejko, P.P. and Nguen, D.F. [1987]: The majorant in the theory of Newton-Kantorovich approximations and the Pták error estimates, Numerical Functional Analysis and Optimization 9, pp. 671-684.

[63] Zambaldi, M.C. [1990]: Estuturas estáticas e dinâmicas para resolver Sistemas Nâo Lineares esparsos, Tese de Mestrado, Departamento de Matemática Aplicada, UNICAMP, Campinas, Brasil.

[64] Zlatev, Z. [1980]: On some pivotal strategies in Gaussian elimination by sparse techniques, SIAM J. Numer. Anal., 17, pp.18-30.