UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE ENGENHARIA ELÉTRICA

DEPARTAMENTO DE ENGENHARIA DE SISTEMAS

ALGORITMOS DE PROJEÇOES PARALELAS PARA SISTEMAS NÃO LINEARES SOBREDETERMINADOS

MARIA APARECIDA DINIZ EHRHARDT

ORIENTADOR: PROF. DR. JOSÉ MARIO MARTÍNEZ PÉREZ

Tese apresentada à Faculdade de Engenharia Elétrica da Universidade Estadual de Campinas (UNICAMP), como parte dos requisitos para a obtenção do título de **DOUTOR EM ENGENHARIA ELÉTRICA**

DEZEMBRO 1991

Ao João Paulo,

à Débora e Daniela

ao Martínez, pela proposta deste trabalho, pela atenção e orientação, pelo estímulo;

à Márcia e Vera que, além da amizade sempre constante, contribuiram com sugestões e discussões na elaboração desta tese;

à Valéria, minha boa vizinha, pela amizade e apoio;

à Antonieta e Plínio, pela paciência para processar todas estas páginas;

aos funcionários Fátima, Benê e Dorival, pela atenção e disponibilidade;

a todos os amigos do DMA, por tornarem agradáveis as horas de trabalho;

aos meus pais, pelo carinho e incentivo;

enfim, a todos que têm me dado amizade,

Obrigada,

Cheti.

RESUMO

Consideramos sistemas de equações não lineares sobredeterminados $F(x) = 0$, $F : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$. Para "resolver" o sistema, introduzimos uma generalização do método de Cimmino para sistemas lineares, demonstrando, então, resultados de convergência local. Com o objetivo de obter propriedades de convergência global, propomos extensões para o algoritmo inicial. Implementamos computacionalmente cinco algoritmos propostos, para serem aplicados a sistemas grandes e esparsos, e apresentamos experimentos numéricos.

NOTAÇÃO

$\|.\|$	norma euclideana
$<x,y>$	produto interno usual entre os vetores x e y
$B(x,\epsilon)$	bola aberta de centro x e raio ϵ
I	matriz identidade
A^+	pseudo-inversa da matriz A
$R(A)$	espaço imagem da matriz A
$N(A)$	núcleo da matriz A
$S^\perp$	complemento ortogonal do subespaço S

ÍNDICE

CAPÍTULO 1

INTRODUÇÃO

Consideremos o sistema não linear

$$F(x) = 0, \qquad (1.1)$$

onde $F:\mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, m, n grandes.

Muitos problemas práticos envolvem sistemas como (1.1); por exemplo, o problema de tomografia computadorizada, como o descrito por Censor et al [1979]; o problema inverso de Geofísica, como o tratado por Bishop et al [1985]; o problema de recobrimento de materiais, citado em Averick, Carter e Moré [1991-pag.14].

Na maior parte dos casos, uma solução exata de (1.1) não existe; temos então que considerar "soluções aproximadas". O critério mais usual é: escolhida uma norma $\|.\|_p$ em $\mathbb{R}^m$, uma solução aproximada para (1.1) é um minimizador de $\|F(x)\|_p$. Assim, o problema de encontrar uma solução aproximada para (1.1) se transforma no problema de otimização:

$$\underset{x}{Min} \|F(x)\|_p. \qquad (1.2)$$

Por questões de simplicidade, a norma-2 é, na maior parte das vezes, preferida especialmente se F é uma função linear. No caso não linear, os métodos mais eficientes para resolver o problema

$$\underset{x}{Min} \|F(x)\|_2, \qquad (1.3)$$

são variações do Método de Gauss-Newton (Dennis-Schnabel [1983]), o qual na sua forma básica, gera uma sequência $\{x^k\}$ em $\mathbb{R}^n$ definida por:

$$x^{k+1} = x^k - [J(x^k)^T J(x^k)]^{-1} J(x^k)^T F(x^k), \qquad (1.4)$$

onde J(x) é a matriz Jacobiana de F em x e x^o é um ponto inicial arbitrário.

O método definido por (1.4) não é globalmente convergente. A sequência $\{x^k\}$ converge linearmente a x^*, um mínimo local de (1.3), se x^o está em uma vizinhança de x^* e $\|F(x^*)\|_2$ é suficientemente pequeno (Dennis-Schnabel [1983]). $\|F(x^*)\|$ é chamado o **resíduo** de (1.1).

Diferentes estratégias foram incorporadas a (1.4) afim de tornar o método globalmente convergente (Levenberg [1944], Marquardt [1963], Moré [1977], Martínez [1987], Martínez e Santos [1990]).

No entanto, existem alguns problemas intrínsecos ao método: a cada iteração é necessário resolver o sistema linear

$$J(x^k)^T J(x^k)\, y = -J(x^k)^T F(x^k), \qquad (1.5)$$

ou, equivalentemente, obter a solução do problema de quadrados mínimos lineares

$$\min_y \| J(x^k)\, y + F(x^k) \|^2. \qquad (1.6)$$

A obtenção da solução tanto de (1.5) como de (1.6) envolve procedimentos computacionalmente caros, especialmente se m e n são grandes,uma vez que, mesmo sendo $J(x^k)$ uma matriz esparsa, $J(x^k)^T J(x^k)$ pode não ser. Isto faz com que procedimentos como a Decomposição de Cholesky ou fatoração ortogonal, que seriam usados para determinar y em (1.5) e (1.6) respectivamente, sejam não esparsos e consequentemente caros. (Ver: George e Heath [1980], [1983], George e Liu [1981], Heath [1984], Duff, Erisman e Reid [1986]).

Estas dificuldades nos motivaram a estudar novos procedimentos para "resolver" (1.1).

Vamos supor que F seja dividido em s blocos da forma:

$$F = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_s \end{pmatrix},$$

tais que $F_i : \mathbb{R}^n \to \mathbb{R}^{m_i}$ e $\sum_{i=1}^{s} m_i = m$.

Dado $x^o \in \mathbb{R}^n$ arbitrário, os algoritmos que descreveremos no capítulo 2 estão baseados em projeções ortogonais simultâneas da aproximação x^k, k = 0,1,2,..., nas variedades lineares

$$V_i(x^k) = \{y \in \mathbb{R}^n : y = \arg\min_z \| F_i(x^k) + J_i(x^k)\,(z - x^k) \| \}.$$

Em 1938, Cimmino propôs um método de projeções ortogonais paralelas para resolução de equações lineares simultâneas. De Pierro e Yusem [1985] trabalharam com uma generalização do método de Cimmino; propuseram um algoritmo para encontrar um ponto comum de uma familia de conjuntos convexos: a cada iteração, um ponto x é projetado em todos os conjuntos convexos e um novo ponto é determinado a partir da combinação convexa destas projeções. Eles mostraram que, quando a intersecção dos conjuntos é vazia, o ponto limite minimiza uma média ponderada (onde os pesos são os coeficientes da combinação convexa) dos quadrados das distâncias aos conjuntos convexos.

Em 1986, Martínez e Sampaio desenvolveram um método de projeções ortogonais para sistemas não lineares indeterminados baseado nas idéias de Cimmino.

A nossa generalização, dirigida então a sistemas não lineares sobredeterminados, substitui, como no trabalho de Martínez e Sampaio, as variedades não lineares $F_i(x) = 0$, $i = 1,2,\ldots,s$, por suas aproximações lineares em torno de x^k, $k = 0,1,2,\ldots$. Esta simplificação foi motivada por razões práticas: é difícil, em geral, calcular projeções em conjuntos não lineares, tornando-se mais fácil obter as projeções nas suas aproximações lineares.

O primeiro algoritmo que introduzimos possui convergência local linear se o resíduo de (1.1) é suficientemente pequeno e

convergência local quadrática se o resíduo é zero, mas não tem garantida a convergência global. Os resultados de convergência deste algoritmo são os mesmos do método de Gauss-Newton definido por (1.4). Este último é um caso particular do nosso algoritmo básico que corresponde a $s = 1$ e $\text{posto}(J(x^k)) = n$. De uma forma mais geral, o método de Ben-Israel [1966] corresponde a $s = 1$, sem restrição sobre $\text{posto}(J(x^k))$.

Com o objetivo de obter resultados de convergência global introduzimos extensões ao primeiro algoritmo.

Todos os algoritmos estão descritos no capítulo 2. No capítulo 3, demonstramos os resultados de convergência local do algoritmo básico. Os teoremas de convergência global referentes aos algoritmos estendidos, como também alguns outros resultados especiais de convergência destes, estão demonstrados no capítulo 4. No capítulo 5, são colocadas informações sobre a implementação computacional dos vários algoritmos e os resultados obtidos com alguns experimentos numéricos, que nos permitiram fazer uma análise do desempemho computacional dos vários métodos propostos. Finalmente, escrevemos dois apêndices: no primeiro colocamos uma aplicação voltada a revestimento de chapas de aço, e no segundo colocamos definições e resultados relacionados com matriz pseudo-inversa.

CAPÍTULO 2

ALGORITMOS BÁSICOS

Com base no Método de Cimmino [1938] (um método de projeções ortogonais simultâneas para sistemas lineares algébricos), Martínez e Sampaio [1986] propõem um algoritmo para sistemas não lineares indeterminados, demonstrando resultados de convergência local.

Em nosso trabalho, desenvolvemos uma generalização desse algoritmo para sistemas sobredeterminados.

Consideramos o sistema não linear:

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

onde

$$F:D \subseteq \mathbb{R}^n \to \mathbb{R}^m \ , \quad F \in C^1(D) \ ,$$

D um conjunto aberto e convexo. Denotamos f_i como a i-ésima componente de F, i=1,2,...,m, ou seja,

$$F(x)=\begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix}.$$

Especialmente se m ≥ n, não existe, em geral, x tal que F(x)=0. Neste caso, nosso objetivo é, então, encontrar x^* de forma que o resíduo $F(x^*)$ seja o menor possível em algum sentido. Em nosso trabalho apresentamos algoritmos que, neste sentido, procuram "resolver" (2.1).

Apesar de nosso maior interesse estar voltado para sistemas onde m ≥ n, os algoritmos que desenvolvemos também se aplicam a sistemas não lineares m x n, com m < n.

2.1 O Algoritmo de Projeções Paralelas

Consideremos F(x) dividida em s blocos:

$$F(x) = \begin{pmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_s(x) \end{pmatrix}, \qquad (2.2)$$

onde

$$F_i : D \subseteq \mathbb{R}^n \to \mathbb{R}^{m_i} \ , \ i=1,2,\ldots,s,$$

de forma que

$$\sum_{i=1}^{s} m_i = m.$$

Denotamos por $J_i(x)$ o Jacobiano de $F_i(x)$, i=1,2,...,s.

Nesta seção descreveremos o primeiro algoritmo que desenvolvemos. Tanto para este algoritmo, como para os demais que colocaremos neste capítulo, não vamos abordar, por enquanto,

detalhes sobre sua convergência, nem tampouco sobre sua implementação computacional, o que faremos em capítulos subsequentes.

Seja x^k a k-ésima aproximação para x^* obtida pelo algoritmo, k=0,1,2,... . Definiremos $x^{k,i}$ como a projeção ortogonal de x^k na variedade constituída pelos minimizadores do problema:

$$\min_x \| J_i(x^k)(x-x^k) + F_i(x^k) \| ,$$

i=1,2,...,s. Então, uma nova aproximação x^{k+1} é uma combinação convexa de $x^{k,1}$, $x^{k,2}$, ... , $x^{k,s}$.

Algoritmo 2.1

Considerando

. $x^0 \in D$ um ponto inicial arbitrário;

. $\lambda_i > 0$ para i=1,2, ..., s, com $\sum_{i=1}^{s} \lambda_i = 1$;

. k = 0,

os passos do algoritmo são:

Passo 1: calcular para cada i, i=1,2, ..., s, $x^{k,i}$, projeção ortogonal de x^k na variedade

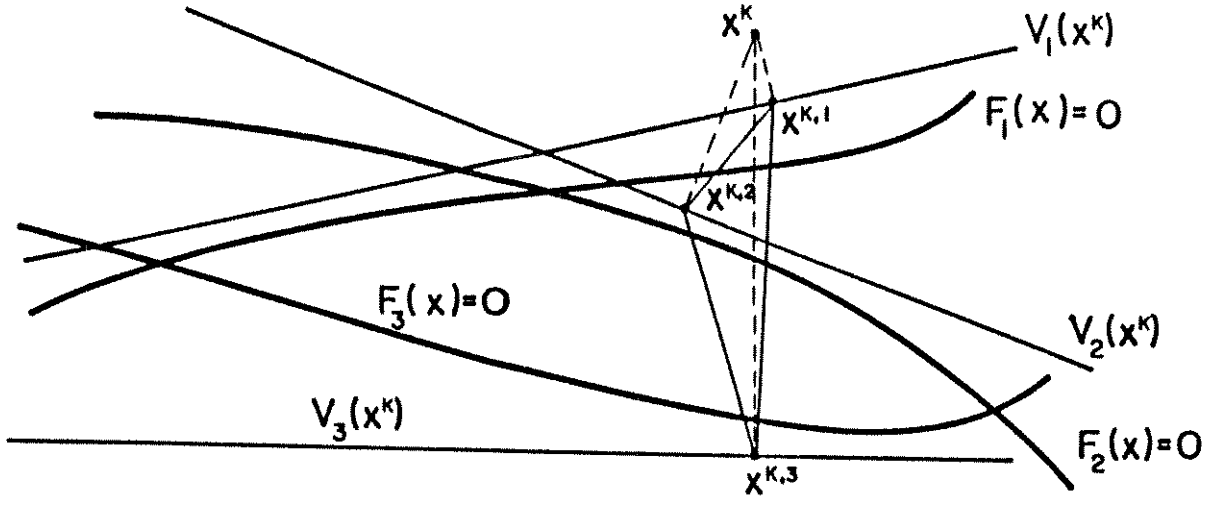
$$V_i(x^k) = \{y : y = arg \min_x \| J_i(x^k)(x-x^k) + F_i(x^k) \| \} ; \qquad (2.3)$$

Passo 2: calcular a combinação convexa das projeções:

$$w_k = \sum_{i=1}^{s} \lambda_i \, x^{k,i} ;$$

Passo 3: fazer $x^{k+1} = w_k$, k=k+1 e voltar ao passo 1.

Graficamente:



Resultados de convergência local do algoritmo vão ser demonstrados no Capítulo 3. Por outro lado, no estudo da convergência global, verificamos que modificações seriam necessárias, uma vez que não podemos garantir que o algoritmo 2.1 gere sempre direções de descida para funções objetivo baseadas em normas adequadas, condição necessária para empregarmos estratégias que garantam a convergência global, conforme colocaremos na seção 2.2.

Os algoritmos seguintes apresentam modificações ao algoritmo 2.1, usando estratégias que os tornam global e localmente convergentes.

2.2 Estratégias de Globalização

Estratégias para tornar certos métodos globalmente convergentes aparecem de uma forma bastante natural quando tratamos de problemas de minimização do tipo:

$$\begin{aligned} &Min\ f(x), \\ &f:\mathbb{R}^n \to \mathbb{R}. \end{aligned} \tag{2.4}$$

Nestes casos, o que pedimos é que, a cada iteração do método, haja um decréscimo no valor da função f.

Existem estratégias correspondentes para sistemas de equações não lineares. No caso do sistema (2.1), o que podemos exigir é que, a cada iteração do algoritmo, haja um decréscimo no valor de alguma norma de F(x).

Em nosso trabalho, escolhemos a norma p de F(x), tal que $1 < p \leq 2$. A escolha se justifica pois, em primeiro lugar, esta norma é diferenciável, e, em segundo lugar, para p próximo de 1, não tem a tendência de ser muito influenciável por "outliers", ou seja, se considerarmos um certo vetor que possua algumas poucas componentes com valores muito distantes dos valores das demais, quando calculamos a norma p deste vetor, tomando p próximo de 1, podemos verificar que a influência dos valores distantes será muito menor do que no cálculo da norma 2, por exemplo.

Consideremos então o problema:

$$\begin{aligned} &Min\ f(x) = \frac{1}{p}\|F(x)\|_p^p, \\ &f:D \subset \mathbb{R}^n \to \mathbb{R} \end{aligned} \tag{2.5}$$

(O termo 1/p foi usado por conveniência nos cálculos)

Devemos notar que:

$$\|F(x)\|_p^p = |f_1(x)|^p + |f_2(x)|^p + \ldots + |f_m(x)|^p$$

e

$$|f_i(x)|^p = \begin{cases} f_i(x)^p \text{ se } f_i(x) \geq 0 , \\ (-f_i(x))^p \text{ se } f_i(x) < 0 , \end{cases}$$
$$i = 1, 2, \ldots, m$$

Assim temos:

$$(|f_i(x)|^p)' = \begin{cases} p f_i(x)^{p-1} f_i'(x) \text{ se } f_i(x) \geq 0 , \\ p(-f_i(x))^{p-1}(-f_i'(x)) \text{ se } f_i(x) < 0 , \end{cases}$$
$$i = 1, 2, \ldots, m$$

e o gradiente de f(x) é:

$$g(x) = \nabla\left(\frac{1}{p}\|F(x)\|_p^p\right) = J^T(x)\overline{F}(x) ,$$

onde J(x) = Jacobiano de F(x) e

$$\overline{F}(x) = (\overline{f}_1(x)\ \overline{f}_2(x)\ \ldots\ \overline{f}_m(x))^T ,$$
$$\text{com } \overline{f}_i(x) = \text{sinal}(f_i(x))\,|f_i(x)|^{p-1} ,$$
$$i = 1, 2, \ldots, m.$$

Observamos que se $x \in D$ resolve (2.1), então x será ponto de mínimo para (2.5). A dificuldade, no entanto, é que os algoritmos que fazem uso de estratégias de globalização garantem a convergência a pontos estacionários de f(x), isto é, a pontos

$x \in D$ tais que

$$g(x) = J^T(x)\overline{F}(x) = 0 \ ,$$

os quais podem ser somente minimizadores locais ou pontos de sela de f(x), mas não fornecem o menor resíduo possível para o sistema.

De qualquer forma, os algoritmos que desenvolvemos usam a estrutura do problema original nos cálculos realizados em cada iteração, ou seja, são algoritmos voltados para a resolução de sistemas como (2.1), e não algoritmos para minimização de funções; a função $f(x) = \frac{1}{p}\|F(x)\|_p^p$ vai ser usada apenas como "função de mérito".

A cada iteração k (k = 0,1,2,...) de um método global, depois de obtida uma direção de descida para a função f(x), a partir de uma certa aproximação x^k, realizamos uma busca nesta direção até obtermos um decréscimo suficiente para f, conforme descreveremos nos algoritmos desta seção.

Uma primeira estratégia de globalização foi aplicada ao algoritmo 2.1. Se a direção $d_k = x^{k+1} - x^k$, gerada por aquele algoritmo, não for uma direção de descida, tomamos a direção oposta a do gradiente de f(x) (direção de máxima descida).

Neste algoritmo, assim como nos seguintes deste capítulo, consideramos:

. dados iniciais: $x^o \in D$, um ponto inicial arbitrário;

$$\theta_1,\ \theta_2,\ \theta_3 \in (0,1);\ 0<\underline{M}<\overline{M}<\infty;$$

. x^k (k=0,1,2,...), a k-ésima aproximação para a solução obtida pelo algoritmo.

Denotamos:

$$\overline{F}_k = \overline{F}(x^k)\ ,\ J_k = J(x^k)\ e\ g_k = g(x^k).$$

Algoritmo 2.2.1

Passo 1: Calcular

$$\overline{w}_k = w_k - x^k\ ,$$

onde w_k é calculado como no algoritmo 2.1.

Passo 2: Se a condição

$$\langle \overline{w}_k, g_k \rangle \leq -\theta_1 \|\overline{w}_k\| \|g_k\| \qquad (2.6)$$

não for satisfeita, fazer $d_k = -g_k$ e ir para o passo 5.

Passo 3: Se

$$\underline{M} \|g_k\| \leq \|\overline{w}_k\| \leq \overline{M} \|g_k\|\ , \qquad (2.7)$$

fazer $d_k = w_k$ e ir para o passo 5.

Passo 4: Se

$$Se\ \|\overline{w}_k\| < \underline{M} \|g_k\|\ fazer$$

$$d_k = \frac{\underline{M} \|g_k\|}{\|\overline{w}_k\|} . \overline{w}_k .$$

$$Se\ \|\bar{w}_k\| > \bar{M}\,\|g_k\|,\ fazer\ d_k = \frac{\bar{M}\,\|g_k\|}{\|\bar{w}_k\|}.\bar{w}_k .$$

Passo 5: Fazer α=1. Enquanto:

$$f(x^k+\alpha d_k) > f(x^k) + \alpha\,\theta_2\,\langle g_k, d_k\rangle,\ fazer:$$

$$\alpha = \bar{\alpha},\ onde\ \bar{\alpha}\ e'\ tal\ que\ \theta_3\alpha \le \bar{\alpha} \le (1-\theta_3)\alpha.$$

Passo 6: Fazer

$$\alpha_k = \alpha,\ s_k = \alpha_k d_k,\ x^{k+1} = x^k + s_k,\ k = k+1.$$

A seguir colocamos outros dois algoritmos que são também extensões do algoritmo 2.1, mas que geram direções de descida a cada iteração, conforme demonstraremos na seção 2.3.

No algoritmo 2.2.2 fazemos uma busca unidimensional na direção de descida gerada pelo método, enquanto que no algoritmo 2.2.3 a busca nesta direção é bidimensional.

Algoritmo 2.2.2

Passo 1: Calcular $v_k \in \mathbb{R}^n$ como solução do problema bidimensional

$$\begin{aligned} Min\ q(v) &= \|J_k v + \bar{F}_k\|^2 \\ s.a\quad v &= \lambda_1 g_k + \lambda_2 \bar{w}_k, \\ \lambda_1, \lambda_2 &\in \mathbb{R}, \end{aligned} \qquad (2.8)$$

onde $\bar{w}_k$ é calculado como no algoritmo 2.2.1.

Passo 2: Se a condição

$$\langle v_k, g_k \rangle \leq -\theta_1 \|v_k\| \|g_k\| \qquad (2.9)$$

não for satisfeita, fazer $d_k = -g_k$ e ir para o passo 5.

Passo 3:

$$Se\ \underline{M}\ \|g_k\| \leq \|v_k\| \leq \overline{M}\ \|g_k\|, \qquad (2.10)$$

fazer $d_k = v_k$ e ir para o passo 5.

Passo 4:

$$Se\ \|v_k\| < \underline{M}\ \|g_k\|,\ \ fazer\ d_k = \frac{\underline{M}\ \|g_k\|}{\|v_k\|}.v_k.$$

$$Se\ \|v_k\| > \overline{M}\ \|g_k\|,\ fazer\ d_k = \frac{\overline{M}\|g_k\|}{\|v_k\|}.v_k$$

Passo 5: Fazer α=1. Enquanto

$$f(x^k + \alpha d_k) > f(x^k) + \alpha\ \theta_2 \langle g_k, d_k \rangle,$$

$$fazer: \alpha = \overline{\alpha},\ onde\ \overline{\alpha}\ é\ tal\ que\ \theta_3\alpha \leq \overline{\alpha} \leq (1-\theta_3)\alpha.$$

Passo 6:

$$\alpha_k = \alpha,\ s_k = \alpha_k d_k,\ x^{k+1} = x^k + s_k,\ k = k+1.$$

Algoritmo 2.2.3

Passo 1: Fazer $d_k^1 = -g_k$.

Passo 2: Calcular v_k por (2.8) e testar a condição (2.9); se esta não for satisfeita, fazer $d_k^2 = d_k^1$ e ir para o passo 5.

Passo 3: Testar a condição (2.10); se esta for satisfeita, fazer $d_k^2 = v_k$ e ir para o passo 5.

Passo 4:

$$Se\ \|v_k\| < \underline{M}\ \|g_k\|,\ fazer$$

$$d_k^2 = \frac{\underline{M}\ \|g_k\|}{\|v_k\|}.v_k. \qquad (2.11a)$$

$$Se\ \|v_k\| > \overline{M}\ \|g_k\|,\ fazer\ d_k^2 = \frac{\overline{M}\ \|g_k\|}{\|v_k\|}.v_k. \qquad (2.11b)$$

Passo 5: Fazer $t = \|d_k^2\|$.

(5.a) Obter d como solução do problema:

$$\begin{aligned} &Min\ q(d) = \|J_k d + \overline{F}_k\|^2 \\ &s.a\ \ d = \lambda_1 d_k^1 + \lambda_2 d_k^2, \\ &\|d\|^2 \leq t, \\ &\lambda_1,\ \lambda_2 \in \mathbb{R}. \end{aligned} \qquad (2.12)$$

(Devemos observar que, conforme as direções d_k^1 e d_k^2 estão definidas, d está necessariamente no cone positivo gerado por d_k^1 e d_k^2, ou seja, $\lambda_1,\ \lambda_2 \geq 0$.)

(5.b) Se

$$f(x^k+d) \leq f(x^k) + \theta_2 \langle g_k, d\rangle, \qquad (2.13)$$

ir para o passo 6.

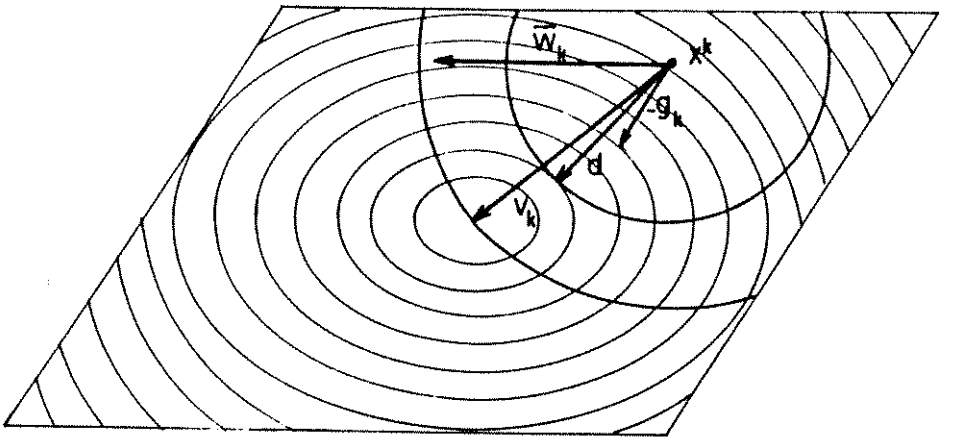
(5.c) Fazer t = t, onde t é tal que

$$\theta_3 t \leq \overline{t} \leq (1-\theta_3)t;$$

voltar para (5.a).

Passo 6: $s_k = d,\ x^{k+1} = x^k + s_k,\ k = k+1.$

Graficamente:



Para mostrarmos que os algoritmos 2.2.1 e 2.2.2 estão bem definidos, basta que demonstremos o seguinte lema:

Lema 2.2.1. Desde que d_k, direção gerada pelo algoritmo 2.2.1 ou 2.2.2, é uma direção de descida, a condição "de Armijo"

$$f(x^k+\alpha d_k) \leq f(x^k) + \alpha\theta_2 \langle g_k, d_k\rangle \qquad (2.14)$$

é satisfeita para α suficientemente pequeno e $\theta_2 \in (0,1)$.

Prova:

Consideremos a derivada direcional de f no ponto x^k, na direção d_k:

$$D_{d_k} f(x^k) = \lim_{\alpha\to 0} \frac{f(x^k+\alpha d_k) - f(x^k)}{\alpha}$$

$$= \langle g_k, d_k\rangle .$$

$$Logo, \quad \lim_{\alpha \to 0} \frac{f(x^k+\alpha d_k) - f(x^k)}{\alpha \langle g_k, d_k \rangle} = 1.$$

Assim, dado $\theta_2 \in (0,1)$, existe $\alpha > 0$ suficientemente pequeno tal que

$$\frac{f(x^k+\alpha d_k) - f(x^k)}{\alpha \langle g_k, d_k \rangle} \geq \theta_2, \text{ para todo } \alpha \in [0, \bar{\alpha}].$$

Mas $\langle g_k, d_k \rangle < 0$, ou seja,

$$f(x^k+\alpha d_k) - f(x^k) \leq \alpha \theta_2 \langle g_k, d_k \rangle \text{ ou}$$
$$f(x^k + \alpha d_k) \leq f(x^k) + \alpha \theta_2 \langle g_k, d_k \rangle,$$

o que mostra que a condição de Armijo é satisfeita para $\alpha > 0$ suficientemente pequeno. ∎

No caso do algoritmo 2.2.3, teremos que a condição (2.13) é satisfeita para t suficientemente pequeno.

Vamos supor, nesta e na próxima seção, que posto(J_k) = n para todo k = 0,1,2,....

Lema 2.2.2.

$$\lim_{t \to 0} \frac{d(t)}{\|d(t)\|} = \frac{-g_k}{\|g_k\|},$$

onde d(t) é a solução do problema (2.12).

Prova: O problema (2.12) pode ser reescrito como:

$$\begin{aligned} &Min\ d^T J_k^T J_k d + 2d^T g_k + \overline{F}_k^T \overline{F}_k \\ &s.a\ \ d=\lambda_1 d_k^1 + \lambda_2 d_k^2, \\ &d^T d \le t, \\ &\lambda_1, \lambda_2 \in \mathbb{R}. \end{aligned}$$

Se $d_k^2 = d_k^1 = -g_k$, então $d(t) = -\lambda g_k$, $\lambda \geq 0$, e o resultado é trivial.

Vamos então considerar o caso em que $d_k^1 \neq d_k^2$. Seja $\{e_1, e_2\}$ uma base ortonormal para o subespaço gerado por d_k^1 e d_k^2. O problema

$$\begin{aligned} &Min\ z^T J_k{}^T J_k z + 2z^T g_k + \overline{F}_k{}^T \overline{F}_k \\ &s.a\ \ z = \alpha_1 e_1 + \alpha_2 e_2, \\ &z^T z \le t, \\ &\alpha_1, \alpha_2 \in \mathbb{R} \end{aligned} \qquad (2.15)$$

é equivalente ao problema (2.12).

Podemos reescrever (2.15) por:

$$\begin{aligned} &Min\ (\alpha_1\alpha_2)\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix} J_k{}^T J_k (e_1 e_2)\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + 2(\alpha_1\alpha_2)\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix} g_k + \overline{F}_k{}^T \overline{F}_k \\ &s.a\ \ \ (\alpha_1\alpha_2)\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix}(e_1 e_2)\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \le t. \end{aligned} \qquad (2.16)$$

As condições de Kuhn-Tucker para (2.16) em relação à solução $\alpha(t)$, são:

$$\begin{aligned} &\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix} J_k^T J_k (e_1 e_2)\alpha(t) + \begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix} g_k + \mu(t)\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix}(e_1 e_2)\alpha(t) = 0, \\ &\mu(t) \in \mathbb{R}^+. \end{aligned}$$

Mas

$$\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix}(e_1\ e_2) = I_{2x2};$$

então temos:

$$\left[\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix} J_k^T J_k (e_1\ e_2) + \mu(t) I\right]\alpha(t) = -\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix} g_k.$$

Consideremos $E = (e_1\ e_2) \in \mathbb{R}^{nx2}$, $H_k = E^T J_k{}^T J_k E \in \mathbb{R}^{2x2}$, e consideremos também a decomposição:

$H_k = Q_k D_k Q_k{}^T$, onde:

. Q_k é uma matriz ortogonal 2 x 2, $Q_k = (q_1{}^k, q_2{}^k)$;

. D_k é uma matriz diagonal, $D_k = \text{diag}\ (\sigma_1{}^k, \sigma_2{}^k)$, com $\sigma_1{}^k \leq \sigma_2{}^k$ os auto-valores de H_k.

Então:

$$\alpha(t) = -(H_k+\mu(t)I)^{-1} E^T g_k$$

$$e \quad \|\alpha(t)\|^2 = \|-(H_k+\mu(t)I)^{-1} E^T g_k\|^2$$

$$= \|-Q_k(D_k+\mu(t)I)^{-1} Q_k^T E^T g_k\|^2$$

$$= \|(D_k+\mu(t)I)^{-1} Q_k^T E^T g_k\|^2$$

$$= \left(\frac{g_k^T E q_1^k}{\sigma_1^k+\mu(t)}\right)^2 + \left(\frac{g_k^T E q_2^k}{\sigma_2^k+\mu(t)}\right)^2.$$

Mas

$$\lim_{t\to 0} \|\alpha(t)\|^2 = 0, \text{ desde que } \|E\alpha(t)\|^2 \leq t.$$

Então:

$$\lim_{t\to 0}\left[\left(\frac{g_k^T E\, q_1^k}{\sigma_1^k+\mu(t)}\right)^2+\left(\frac{g_k^T E\, q_2^k}{\sigma_2^k+\mu(t)}\right)^2\right]=0;$$

portanto, $\lim_{t\to 0}\dfrac{g_k^T E\, q_i^k}{\sigma_i^k+\mu(t)}=0$, *para* $i=1,2$.

Como existe n ϵ {1,2} tal que $g_k{}^T E q_n{}^k \neq 0$, então

$$\lim_{t\to 0}(\sigma_n^k+\mu(t))=\infty, \quad \textit{ou seja,} \quad \lim_{t\to 0}\mu(t)=\infty$$

e, portanto, existe $t_0 > 0$ tal que para todo $t \leq t_0$, $\mu(t) > 0$.

Então, para $t \leq t_0$,

$$\frac{\alpha(t)}{\|\alpha(t)\|}=-\frac{(H_k+\mu(t)I)^{-1}E^T g_k}{\|-(H_k+\mu(t)I)^{-1}E^T g_k\|}$$

$$=-\frac{(H_k/\mu(t)+I)^{-1}E^T g_k}{\|-(H_k/\mu(t)+I)^{-1}E^T g_k\|} \quad e$$

$$\lim_{t\to 0}\frac{\alpha(t)}{\|\alpha(t)\|}=-\frac{E^T g_k}{\|E^T g_k\|}.$$

Para simplificar, vamos tomar $e_1 = -g_k/\|g_k\|$.

Então:

$$-E^T g_k=\begin{pmatrix}-e_1^T\\-e_2^T\end{pmatrix}g_k=\begin{pmatrix}g_k^T g_k/\|g_k\|\\e_2^T e_1/\|g_k\|\end{pmatrix}$$

$$=\begin{pmatrix}\|g_k\|\\0\end{pmatrix},$$

$$e \qquad \|E^T g_k\|=\|g_k\|.$$

Assim,

$$\lim_{t\to 0} \frac{\alpha(t)}{\|\alpha(t)\|} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Mas

$$z(t) = E\,\alpha(t) \text{ e, } portanto, \|z\| = \|E\,\alpha(t)\| = \|\alpha(t)\|.$$

Logo:

$$\lim_{t\to 0} \frac{z(t)}{\|z(t)\|} = E \lim_{t\to 0} \frac{\alpha(t)}{\|\alpha(t)\|}$$

$$= \begin{pmatrix} \frac{-g_k}{\|g_k\|} & e_2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$= \frac{-g_k}{\|g_k\|}.$$

Como z(t) é solução do problema (2.15), o qual é equivalente ao problema (2.12), então d(t), solução de (2.12), também satisfaz:

$$\lim_{t\to 0} \frac{d(t)}{\|d(t)\|} = \frac{-g_k}{\|g_k\|}.$$

■

Lema 2.2.3. A condição

$$f(x^k+d(t)) \leq f(x^k) + \theta_2 < g_k, d(t) >$$

é satisfeita para t suficientemente pequeno, $\theta_2 \in (0,1)$ e d(t) solução do problema (2.12).

Prova: Consideremos

$$\varphi(d) = \frac{f(x^k+d)-f(x^k)}{\|d\|}.$$

Pelo Teorema do Valor Médio, temos que:

$$f(x^k + d)-f(x^k) = < g(x^k + \xi d), d > , \; com \; 0 < \xi < 1.$$

Portanto,

$$\varphi(d) = \frac{< g(x^k+\xi d), d >}{\|d\|} \quad e$$

$$\lim_{t\to 0} \varphi(d(t)) = \lim_{t\to 0} \frac{<g(x^k+\xi d(t)), d(t)>}{\|d(t)\|}$$

$$= \lim_{t\to 0} g(x^k+\xi d(t))^T . \lim_{t\to 0} \frac{d(t)}{\|d(t)\|}$$

$$= g_k^T \frac{(-g_k)}{\|g_k\|}$$

$$= -\|g_k\| ,$$

ou seja,

$$\lim_{t\to 0} \frac{f(x^k+d(t))-f(x^k)}{\|d(t)\|} = -\|g_k\| . \qquad (2.17)$$

Por outro lado:

$$\lim_{t\to 0} < g_k, \frac{d(t)}{\|d(t)\|} > = \lim_{t\to 0} g_k^T \frac{d(t)}{\|d(t)\|}$$

$$= g_k^T \cdot \lim_{t \to 0} \frac{d(t)}{\|d(t)\|}$$

$$= g_k^T \cdot \frac{(-g_k)}{\|g_k\|}$$

$$= -\|g_k\|,$$

ou seja,

$$\lim_{t \to 0} < g_k, \frac{d(t)}{\|d(t)\|} > = -\|g_k\|. \qquad (2.18)$$

De (2.17) e (2.18) temos:

$$\lim_{t \to 0} \frac{f(x^k + d(t)) - f(x^k)}{< g_k, d(t) >} = 1.$$

Portanto, dado $\theta_2 \in (0,1)$, existe $t > 0$ suficientemente pequeno tal que:

$$\frac{f(x^k + d(t)) - f(x^k)}{< g_k, d(t) >} \geq \theta_2 \quad \textit{para todo } t \in [0, \bar{t}].$$

Desde que a direção d, gerada pelo algoritmo 2.2.3, é uma direção de descida, então $< g_k, d > < 0$ e

$$f(x^k + d) - f(x^k) \leq \theta_2 < g_k, d>;$$

isto é, a condição (2.13)

$$f(x^k + d(t)) \leq f(x^k) + \theta_2 < g_k, d(t) >, \ \theta_2 \in (0,1),$$

é satisfeita para t suficientemente pequeno. ∎

2.3. Propriedades dos Algoritmos

Nesta seção demonstramos algumas propriedades que a direção v_k, gerada pelos algoritmos 2.2.2 e 2.2.3, apresenta. Estas propriedades mostram que as condições (2.9) e (2.10) vão ser, teoricamente, satisfeitas em toda iteração k, k=0,1,2,... e, portanto, as buscas se realizam na própria direção v_k, sem que seja preciso substituí-la pela direção de máxima descida. Devemos observar que estas condições são testadas nos algoritmos apenas como medida de segurança.

Lema 2.3.1. Sejam $G(y) \in \mathbb{R}^{nxn}$ e $b(y) \in \mathbb{R}^n$ funções contínuas de $y \in A \subset \mathbb{R}^n$, A compacto. Consideremos ainda G(y) simétrica e definida positiva $\forall\, y \in A$. Definimos

$$q_y(x) = \frac{1}{2}x^T G(y)x + b(y)^T x \quad e$$

$$g_y(x) = \nabla_x q_y(x) = G(y)x + b(y).$$

Então existe $\gamma > 0$ tal que, se $q_y(\omega) \leq q_y(z)$, temos $\|g_y(\omega)\| \leq \gamma \|g_y(z)\|$.

Prova: Ver Martínez [1987].

Para os lemas seguintes vamos supor que toda a sequência $\{x^k\}$ gerada pelos algoritmos 2.2.2 ou 2.2.3 está contida em um compacto $S \subset D$.

Lema 2.3.2. Para todo k = 0,1,2..

$$\underline{M}\, \|g_k\| \leq \|v_k\| \leq \bar{M}\, \|g_k\|,$$

com $0 < \underline{M} < M < \infty$.

Prova:

(i) vamos mostrar inicialmente que

$$\|v_k\| \leq \bar{M} \|g_k\|, \ \bar{M} > 0.$$

Como v_k resolve o problema (2.8) então:

$$\|J_k v_k + \bar{F}_k\| \leq \|J_k g_k + \bar{F}_k\|.$$

Pelo lema 2.3.1, existe $\gamma > 0$ tal que

$$\|J_k^T J_k v_k + g_k\| \leq \gamma \|J_k^T J_k g_k + g_k\|,$$

$$\|(J_k^T J_k)^{-1}\| \cdot \|J_k^T J_k v_k + g_k\| \leq \gamma \|(J_K^T J_K)^{-1}\| \|J_k^T J_k g_k + g_k\|$$

$$\|v_k + (J_k^T J_K)^{-1} g_k\| \leq \gamma \|(J_k^T J_k)^{-1}\| \cdot \|J_k^T J_k g_k + g_k\|$$

$$\|v_k\| - \|(J_k^T J_k)^{-1} g_k\| \leq \gamma \|(J_k^T J_k)^{-1}\| \|J_k^T J_k g_k + g_k\|$$

$$\|v_k\| \leq \|(J_k^T J_k)^{-1} g_k\| + \gamma \|(J_k^T J_k)^{-1}\| \|J_k^T J_k g_k + g_k\|$$

$$\leq cond\,(J_k^T J_k) \left[\frac{1+\gamma}{\|J_k^T J_k\|} + \gamma \right] \|g_k\|.$$

Com a hipótese de que a sequência $\{x^k\}$ está contida em um conjunto compacto, garantimos que $J_k = J(x^k)$ é uniformemente limitada neste conjunto. Assim, fazendo

$$\bar{M}_k = cond(J_k^T J_k) \left[\frac{1+\gamma}{\|J_k^T J_k\|} + \gamma \right],$$

temos que existe M tal que

$$\bar{M} = \sup_k \bar{M}_k \quad e \quad \|v_k\| \leq \bar{M} \|g_k\|,$$

para todo k = 0,1,2,... ■

(ii) Vamos agora mostrar que

$$\|v_k\| \geq \underline{M} \|g_k\|, \ \underline{M} > 0.$$

Por (2.8) temos que

$$\|g_k\|^2 = -\langle g_k, J_k^T J_k v_k \rangle.$$

Portanto,

$$\langle g_k, J_k^T J_k v_k \rangle < 0 \quad e,$$

usando a desigualdade de Schwarz:

$$\|g_k\|^2 \leq \|g_k\| \cdot \|J_k^T J_k v_k\|,$$

$$\|J_k^T J_k v_k\| \geq \|g_k\|,$$

$$\|v_k\| \geq \frac{\|g_k\|}{\|J_k^T J_k\|}.$$

Assim, com as mesmas considerações feitas na parte (i), se tomarmos

$$\underline{M} = \inf_k \underline{M}_k = \inf_k \frac{1}{\|J_k^T J_k\|},$$

teremos satisfeita a desigualdade

$$\|v_k\| \geq \underline{M} \|g_k\|, \ para \ k = 0,1,2,...$$

Agora, só falta mostrar que $\underline{M} < \bar{M}$.

$$\bar{M}_k = cond\ (J_k^T J_k)\left(\frac{1+\gamma}{\|J_k^T J_k\|} + \gamma\right)$$

$$\geq \frac{1+\gamma}{\|J_k^T J_k\|} + \gamma$$

$$\geq \frac{1+\gamma}{\|J_k^T J_k\|}$$

$$\geq \frac{1}{\|J_k^T J_k\|}.$$

Mas $\underline{M}_k = \dfrac{1}{\|J_k^T J_k\|}$ e, assim, $\bar{M}_k > \underline{M}_k$, para todo k = 0,1,2,..., e $\underline{M} < \bar{M}$. ∎

Lema 2.3.3. Existe $\theta_1 > 0$ tal que

$$\langle v_k, g_k \rangle \leq -\theta_1 \|v_k\| \|g_k,$$

para todo k = 0,1,2...

Prova: Seja $P \in \mathbb{R}^{nx2}$ uma matriz cujas colunas formam uma base ortonormal para o subespaço gerado por g_k e $\bar{w}_k$.

Assim,

$$v_k = P\alpha,\ \alpha \in \mathbb{R}^2,\ e$$

$$\alpha = -(P^T J_k^T J_k P)^{-1} P^T g_k.$$

Fazendo $B_k = P^T J_k^T J_k P$, consideremos a decomposição:

$$B_k = Q_k D_k Q_k^T,$$

onde:

.$Q_k \in \mathbf{R}^{2x2}$ é uma matriz ortogonal;

.D_k é uma matriz diagonal,

$$D_k = diag(\sigma_1^k, \sigma_2^k), \; com \; \sigma_1^k \leq \sigma_2^k$$

os auto valores de B_k.

Então, sendo $z_k = P^T g_k$, temos:

$$\frac{z_k^T B_k^{-1} z_k}{\|B_k^{-1} z_k\| \cdot \|z_k\|} = \frac{z_k^T Q_k D_k^{-1} Q_k^T z_k}{\|D_k^{-1} Q_k^T z_k\| \cdot \|z_k\|}$$

$$= \frac{\omega_k^T D_k^{-1} \omega_k}{\|D_k^{-1} \omega_k\| \cdot \|\omega_k\|},$$

onde $\omega_k = Q_k^T z_k$.

Portanto,

$$\frac{z_k^T B_k^{-1} z_k}{\|B_k^{-1} z_k\| \cdot \|z_k\|} = \frac{\frac{\omega_1^2}{\sigma_1} + \frac{\omega_2^2}{\sigma_2}}{\left(\frac{\omega_1^2}{\sigma_1^2} + \frac{\omega_2^2}{\sigma_2^2}\right)^{1/2} (\omega_1^2 + \omega_2^2)^{1/2}}$$

$$\geq \frac{\omega_1^2 + \omega_2^2}{\sigma_2} \cdot \frac{\sigma_1}{\omega_1^2 + \omega_2^2}$$

$$\geq \frac{\sigma_1}{\sigma_2}$$

$$\geq \frac{1}{cond(B_k)},$$

ou seja,

Então,

$$\frac{(P^T g_k)^T (P^T J_k^T J_k P)^{-1} P^T g_k}{\| (P^T J_k^T J_k P)^{-1} P^T g_k \| . \| P^T g_k \|} \geq \frac{1}{cond(P^T J_k^T J_k P)}$$

$$\geq \frac{1}{cond(J_k^T J_k)} .$$

$$\frac{-\alpha^T P^T g_k}{\| \alpha \| . \| P^T g_k \|} \geq \frac{1}{cond(J_k^T J_k)} ,$$

$$<P\alpha, g_k> \leq \frac{-\| \alpha \| . \| P^T g_k \|}{cond(J_k^T J_k)} ,$$

$$<v_k, g_k> \leq \frac{-\| P\alpha \| . \| g_k \|}{cond(J_k^T J_k)} ,$$

$$\text{e} \quad <v_k, g_k> \leq \frac{-1}{cond(J_k^T J_k)} . \| v_k \| . \| g_k \| .$$

Neste caso, fazemos $\theta_1^k = \frac{1}{cond(J_k^T J_k)}$ e $\theta_1 = \inf_k \theta_1^k$,

e a tese está provada, para todo k = 0,1,2,... ∎

2.4. Estratégias "Tolerantes"

O último algoritmo que descrevemos neste capítulo faz uso do que chamamos estratégias "tolerantes". Nosso procedimento está inspirado na estratégia global introduzida por Gomes-Ruggiero, Martínez e Moretti [1991] para sistemas de equações não lineares n x n.

Quando aplicamos o algoritmo 2.1, nosso objetivo é obter uma sequência tal que $\lim_{k\to\infty} \|x^{k+1}-x^k\| = 0$. Para problemas fortemente não lineares, e que possuem resíduo grande, sabemos que a convergência deste algoritmo "puro" é menos provável. Já nos casos de resíduo pequeno, demonstramos no capítulo 3 que o algoritmo 2.1 converge localmente, mas não apresenta condições para convergência global. No entanto, observamos que as direções $\overline{w}_k = x^{k+1} - x^k$ muitas vezes representam direções de descida para a função de mérito f definida por (2.5). Também observamos que, em muitos casos, a convergência do algoritmo 2.1 ocorre, mas $\|\overline{w}_k\|$ não decresce monotonicamente.

Estas observações motivaram a aplicação de estratégia tolerante, isto é, aceitamos que $\|\overline{w}_k\|$ aumente, mas usamos um passo "global" que faça decrescer o valor de f se $\|\overline{w}_k\|$ não decresce depois de um número pré-determinado de iterações.

Chamamos "iteração ordinária" a cada iteração produzida pelo método local e "iteração especial" às iterações produzidas pelos algoritmos com estratégias de globalização (qualquer um dos algoritmos da seção 2.2). A combinação das iterações ordinárias e especiais pela estratégia tolerante gera o seguinte algoritmo:

Algoritmo 2.4.

$$Seja\ \|\overline{w}_k^a\| = \min\{\|\overline{w}_0\|, \|\overline{w}_1\|, \ldots, \|\overline{w}_k\|\}. \qquad (2.19)$$

Definimos $\|\overline{w}_k^a\| = \infty$ *se* $k < 0$. Sejam:

. q, um inteiro positivo;

. $\gamma \in (0,1)$;

. $x^0 \in D$ uma aproximação inicial arbitrária.

Inicialização: k = 0; FLAG = 1.

Passo 1:

Se FLAG = 1, obter x^{k+1} através de uma iteração ordinária.

Se FLAG = -1, obter x^{k+1} através de uma iteração especial.

Passo 2:

Se $\|\bar{w}^a_{k+1}\| \leq \gamma \|\bar{w}^a_{k+1-q}\|$, fazer FLAG = 1, k = k + 1.

Se $\|\bar{w}^a_{k+1}\| > \gamma \|\bar{w}^a_{k+1-q}\|$, fazer FLAG = -1 e k = k + 1.

No capítulo 4 demonstramos a convergência global deste algoritmo. Mostramos que, de fato, duas possibilidades podem ocorrer: ou $\|\bar{w}^a_k\|$ converge a zero ou, se x^* é um ponto limite da sequência $\{x^k\}$ gerada pelo algoritmo, então $g(x^*) = \nabla f(x^*) = 0$.

CAPÍTULO 3

RESULTADOS DE CONVERGÊNCIA LOCAL

Neste capítulo vamos mostrar a convergência local do algoritmo 2.1 descrito no capítulo anterior. O algoritmo pode ser visto como um processo iterativo do tipo $x^{k+1} = \phi(x^k)$ e, então, demonstramos que $\lim_{k\to\infty} x^k = x^*$, se x^* é um ponto fixo do operador $\phi(x)$, $\|F(x^*)\|$ é pequeno e x^0 está suficientemente perto de x^*.

Abordamos também o caso $m < n$, usando as idéias de Meyn [1983].

3.1. Os Operadores ϕ e G

Consideremos F(x) como em (2.1) - (2.2) e seja $J_i(x) \in \mathbb{R}^{m_i \times n}$ o Jacobiano de $F_i(x)$, $i = 1,2,\ldots,s$.

No algoritmo 2.1 vimos que uma nova aproximação x^{k+1} é obtida a partir da combinação convexa das projeções ortogonais de x^k, $k = 0,1,2,\ldots$, nas variedades (2.3). Assim, se definirmos

$$x^i = x - J_i(x)^+ F_i(x), \quad i = 1,2,\ldots,s,$$

e

$$\phi(x) = \sum_{i=1}^{s} \lambda_i x^i \tag{3.1}$$

com $\lambda_i > 0$ para $i = 1,2,\ldots,s$ e $\sum_{i=1}^{s} \lambda_i = 1$, temos que $x^{k+1} = \phi(x^k)$, para $k = 0,1,2,\ldots$

Consideramos x^* um ponto fixo de $\phi(x)$ e usamos um operador auxiliar definido por:

$$G(x) = \sum_{i=1}^{s} \lambda_i \bar{x}^i, \tag{3.2}$$

onde $\bar{x}^i = x - J_i(x^*)^+ F_i(x), \quad i = 1,2,\ldots,s.$

Observamos que x^* é também um ponto fixo de G (desde que $\sum_{i=1}^{s} \lambda_i J_i^+(x^*) F_i(x^*) = 0$) e que $G'(x)$ existe e é contínua $\forall\, x \in D$:

$$G'(x) = \sum_{i=1}^{s} \lambda_i (I - J_i(x^*)^+ J_i(x)). \tag{3.3}$$

Algumas hipóteses devem ser consideradas neste capítulo.

Vamos supor que $J_i(x)$ é Lipschitz - contínua em x^*, isto é, existe $L > 0$ tal que para todo $x \in D$,

$$\|J_i(x) - J_i(x^*)\| \leq L\|x - x^*\|. \tag{3.4}$$

A vantagem de trabalharmos com a hipótese de continuidade de

Lipschitz é que não precisamos exigir que $F \in C^2$ para obtermos os resultados desejados.

A desigualdade (3.4) implica que para todo x, z $\in$ D, i = 1,2 ,..., s,

$$\|F_i(z) - F_i(x) - J_i(x^*)(z-x)\| \leq L\|z-x\| \max\{\|z-x^*\|, \|x-x^*\|\} \quad (3.5)$$

(Ver Broyden, Dennis e Moré [1973]).

Vamos também supor que

$$posto(J_i(x)) = m_i' \leq m_i, \quad (3.6)$$

para todo x em alguma vizinhança de x^*, 1 = 1,2,...,s. Então podemos também supor que existe M > 0 tal que

$$\|J_i(x)^+\| \leq M, \quad (3.7)$$

para todo x na mesma vizinhança de x^*, i = 1,2,...,s. (Ver Golub e Van Loan [1983]).

Lema 3.1.1. Sejam ϕ e G definidos por (3.1) - (3.2). Existem $\theta \in [0,1)$ e $\epsilon_1 > 0$ tais que se $\|F(x^*)\| \leq \theta$ então, para todo $x \in B(x^*, \epsilon_1)$ temos:

$$\|G(x) - \phi(x)\| \leq \alpha_1\|x-x^*\|, \text{ com } \alpha_1 \in [0,1).$$

Prova: $$\|G(x) - \phi(x)\| = \|\sum_{i=1}^{s} \lambda_i \bar{x}^i - \sum_{i=1} \lambda_i x^i\|$$

$$= \| \sum_{i=1} \lambda_i (\bar{x}^i - x^i) \|$$

$$\leq \sum_{i=1}^{s} \lambda_i \| \bar{x}^i - x^i \|$$

$$\leq \sum_{i=1}^{s} \lambda_i \| J_i(x^*)^+ F_i(x) - J_i(x)^+ F_i(x) \|$$

$$\leq \sum_{i=1}^{s} \lambda_i \| J_i(x^*)^+ - J_i(x)^+ \| \, \| F_i(x) \| . \qquad (3.8)$$

Mas, por (3.4), (3.7) e pelo teorema A.4 do apêndice, temos que

$$\| J_i(x^*)^+ - J_i(x)^+ \| \leq 2\sqrt{n}\, ML \, \| x - x^* \| .$$

Assim, de (3.8):

$$\| G(x) - \phi(x) \| \leq 2\sqrt{n}\, ML \, \| x - x^* \| \sum_{i=1}^{s} \lambda_i \| F_i(x) \| . \qquad (3.9)$$

Supomos então que ϵ_1 e θ são tais que

$$2\sqrt{n}\, ML \sum_{i=1}^{s} \lambda_i \| F_i(x) \| \leq \alpha_1 < 1,$$

para todo $x \in B(x^*, \epsilon_1)$. Assim,

$$\| G(x) - \phi(x) \| \leq \alpha_1 \| x - x^* \| ,$$

para todo $x \in B(x^*, \epsilon_1)$. ∎

3.2. O Caso m ≥ n

Nesta seção vamos considerar o algoritmo 2.1 aplicado à resolução do sistema (2.1) se **m ≥ n**, ou seja, se o número de

equações é maior ou igual que o número de variáveis.

Demonstramos que, se o resíduo do sistema (2.1) for suficientemente pequeno, obtemos a convergência local linear do algoritmo. Tratamos de uma forma especial o caso em que o resíduo é zero e os blocos são tais que posto($J_i(x^*)$) = n, quando então a convergência local quadrática é mostrada.

Lema 3.2.1. Seja A uma matriz m x n tal que posto(A) = n. Vamos dividir a matriz A em blocos:

$$A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_s \end{pmatrix},$$

de modo que cada matriz A_i tenha dimensão m_i x n, para i = 1,2,...s e $\sum_{i=1}^{s} m_i = m$. Consideremos $P_i \in \mathbb{R}^{n \times n}$ a projeção ortogonal sobre $R(A_i^T)$, para i = 1,2,...,s e a matriz

$$\bar{A} = \sum_{i=1}^{s} \lambda_i (I - P_i),$$

com $\lambda_i > 0$, para i = 1,2,...,s e $\sum_{i=1}^{s} \lambda_i = 1$.

Então $\|\bar{A}\| < 1$.

Prova: Pela definição de norma matricial temos que

$$\|\bar{A}\| = \max_{\|y\|=1} \|\bar{A}y\|, \; y \in \mathbb{R}^n.$$

Mas,

$$\|\bar{A}y\| = \|\sum_{i=1}^{s} \lambda_i (I-P_i)y\|$$

$$\leq \sum_{i=1}^{s} \lambda_i \|(I-P_i)y\| .$$

Agora, $(I - P_i)y$ é a projeção de y em $N(A_i)$, para i = 1,2,...,s e, portanto, $\|(I-P_i)y\| \leq \|y\|$, onde a igualdade somente se verifica se $y \in N(A_i)$ para todo i = 1,2,...,s simultâneamente, ou seja, $y \in N(A)$, o que não é possível desde que A tem posto completo e $y \neq 0$. Logo,

$$\|\bar{A}y\| < \|y\| .$$

Como $\{y \in \mathbb{R}^n : \|y\| = 1\}$ é um conjunto compacto, $\|\bar{A}y\|$ possui um máximo neste conjunto e $\max\|\bar{A}y\| < \|y\|$, *ou seja* $\|\bar{A}\| < 1$. ∎

Consideremos novamente o sistema (2.1) dividido em blocos como em (2.2). Vamos supor que $J(x^*)$ tem posto completo (=n).

Lema 3.2.2. Existe $\alpha_2 \in [0,1)$ tal que $\|G'(x^*)\| \leq \alpha_2 < 1$.

Prova: De (3.3) vemos que:

$$G'(x^*) = \sum_{i=1}^{s} \lambda_i (I-J_i(x^*)^+ J_i(x^*)) .$$

Mas $P_i = J_i(x^*)^+ J_i(x^*)$ é a projeção ortogonal sobre $R(J_i(x^*)^T)$, para todo i = 1,2,...,s. (ver apêndice). Portanto, usando o Lema 3.2.1 temos demonstrado que

$$\|G'(x^*)\| \leq \alpha_2 < 1. \qquad \blacksquare$$

Lema 3.2.3. Sejam $\epsilon_2 > 0$ e $\alpha_2 \in [0,1)$. Então

$$\|G(x)-G(y)\| \leq \alpha_2\|x-y\|, \ \forall x,y \in B(x^*,\epsilon_2).$$

Prova: Como G(x) é contínuo, com derivadas contínuas $\forall x \in D$, pelo Teorema do Valor Médio temos que:

$$G(x) = G(y) + G'(\xi)(x-y),$$

onde $\xi = x-t(x-y)$, $t \in (0,1)$.

Então:

$$\|G(x)-G(y)\| \leq \|G'(\xi)\|\,\|x-y\|.$$

Seja $\epsilon_2 > 0$ e $x,y \in B(x^*,\epsilon_2)$; temos que $\xi \in B(x^*,\epsilon_2)$ e assim $\|G'(\xi)\| \leq \alpha_2 < 1$ (pelo lema 3.2.2, levando-se em conta a continuidade de G'). Logo:

$$\|G(x)-G(y)\| \leq \alpha_2\,\|x-y\|. \qquad \blacksquare$$

Lema 3.2.4. Existem $\epsilon > 0$ e $\alpha_3 \in [0,1)$ tais que

$$\|\phi(x)-x^*\| \leq \alpha_3\,\|x-x^*\|, \ \forall x \in B(x^*,\epsilon).$$

Prova:

$$\|\phi(x)-x^*\| \leq \|\phi(x)-G(x)\| + \|G(x)-x^*\|.$$

Pelos lemas 3.1.1 e 3.2.3, se ϵ_1 e ϵ_2 forem suficientemente

pequenos e $\epsilon \leq \min\{\epsilon_1, \epsilon_2\}$, podemos mostrar que existe α_3 tal que $0 \leq \alpha_1 + \alpha_2 \leq \alpha_3 < 1$ e, para todo $x \in B(x^*, \epsilon)$, temos:

$$\|\phi(x) - x^*\| \leq \alpha_3 \|x - x^*\|. \qquad \blacksquare$$

Teorema 3.2.5. Existem $\epsilon > 0$ e $\theta \in [0,1)$, tais que se $x^0 \in B(x^*, \epsilon)$ e $\|F(x^*)\| \leq \theta < 1$, então a sequência definida por $x^{k+1} = \phi(x^k)$ está em $B(x^*, \epsilon)$ e converge linearmente a x^* com taxa de convergência menor ou igual a $\alpha_3 \in [\alpha_2, 1)$.

Prova: Definamos ϵ como no lema 3.2.4.

(a) Vamos mostrar que $\{x^k\} \subset B(x^*, \epsilon)$, para todo $k = 0,1,2,\ldots$ Esta parte da demonstração vai ser feita por indução.

Por hipótese temos que $x^0 \in B(x^*, \epsilon)$. Vamos então supor que $x^k \in B(x^*, \epsilon)$ e queremos mostrar que $x^{k+1} \in B(x^*, \epsilon)$.

$$\begin{aligned} \|x^{k+1} - x^*\| &= \|\phi(x^k) - x^*\| \\ &\leq \alpha_3 \|x^k - x^*\| \\ &\leq \alpha_3 \epsilon \\ &< \epsilon, \end{aligned}$$

e assim temos demonstrado que toda a sequência $\{x^k\}$ está em $B(x^*, \epsilon)$.

(b) Vamos agora mostrar que $\{x^k\}$ converge a x^*.

$$\begin{aligned} \|x^1 - x^*\| &= \|\phi(x^0) - x^*\| \leq \alpha_3 \|x^0 - x^*\| \\ \|x^2 - x^*\| &= \|\phi(x^1) - x^*\| \leq \alpha_3 \|x^1 - x^*\| \leq \alpha_3^2 \|x^0 - x^*\| \\ &\vdots \\ \|x^k - x^*\| &= \|\phi(x^{k-1}) - x^*\| \leq \alpha_3^k \|x^0 - x^*\| \end{aligned}$$

$$\lim_{k\to\infty} \|x^k - x^*\| \leq \lim_{k\to\infty} \alpha_3^k \|x^0 - x^*\|$$

Mas $\lim_{k\to\infty} \alpha_3^k = 0$, desde que $\alpha_3 \in [0,1)$. Logo

$$\lim_{k\to\infty} \|x^k - x^*\| = 0 \quad e \quad \lim_{k\to\infty} x^k = x^*.$$

Além disso, vemos que a sequência $\{x^k\}$ converge linearmente a x^*, com taxa de convergência menor ou igual a α_3. ∎

No seguinte teorema colocamos condições sob as quais convergência quadrática é obtida.

Teorema 3.2.6. Vamos supor que $F(x^*) = 0$ e que posto$(J_i(x^*)) = n$, para $i = 1,2,\ldots,s$. Então existem $\epsilon > 0$, $c > 0$ tais que, se $x^0 \in B(x^*,\epsilon)$, a sequência gerada por $x^{k+1} = \phi(x^k)$, $k = 0,1,2,\ldots$, converge a x^* e satisfaz

$$\|x^{k+1} - x^*\| \leq c \|x^k - x^*\|^2. \qquad (3.10)$$

Prova: Definamos ϵ como no teorema 3.2.5. Convergência de $\{x^k\}$ a uma taxa linear segue-se deste teorema. De (3.1) temos que:

$$\|x^{k+1} - x^*\| = \|x^k - x^* - \sum_{i=1}^{s} \lambda_i J_i(x^k)^+ F_i(x^k)\|.$$

Consideremos:

$$A_1 = \| x^k - x^* - \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ J_i(x^k)(x^k - x^*) \|,$$

$$A_2 = \| \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ J_i(x^*)(x^k - x^*) - \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ F_i(x^k) \|,$$

$$A_3 = \| \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ F_i(x^k) - \sum_{i=1}^{s} \lambda_i J_i(x^k)^+ F_i(x^k) \|.$$

Assim,

$$\| x^{k+1} - x^* \| \leq A_1 + A_2 + A_3.$$

Temos que:

$$A_1 = \| x^k - x^* - \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ J_i(x^*)(x^k - x^*) \|$$

$$\leq \| I - \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ J_i(x^*) \| \, \| x^k - x^* \|.$$

Mas, desde que posto($J_i(x^*)$) = n, temos que $N(J_i(x^*)) = \{0\}$ e, portanto,

$$\| I - \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ J_i(x^*) \| = 0,$$

ou seja, $A_1 = 0$.

$$A_2 = \| \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ J_i(x^*)(x^k - x^*) - \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ F_i(x^k) \|$$

$$= \| \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ [F_i(x^k) - J_i(x^*)(x^k - x^*)] \|$$

$$= \| \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ [F_i(x^k) - F_i(x^*) - J_i(x^*)(x^k - x^*)] \|$$

$$\leq \sum_{i=1}^{s} \lambda_i ML \, \|x^k - x^*\|^2$$

$$\leq ML \, \|x^k - x^*\|^2,$$

usando (3.4) e (3.5).

$$A_3 = \| \sum_{i=1}^{s} \lambda_i J_i(x^*)^+ F_i(x^k) - \sum_{i=1}^{s} \lambda_i J_i(x^k)^+ F_i(x^k) \|$$

$$= \|G(x^k) - \phi(x^k)\|$$

$$\leq 2\sqrt{n}\, ML \, \|x^k - x^*\| \sum_{i=1}^{s} \lambda_i \|F_i(x^k)\|,$$

usando (3.9).

Mas, por (3.5), temos que:

$$\|F_i(x^k) - F_i(x^*)\| - \|J_i(x^*)(x^k - x^*)\| \leq L\|x^k - x^*\|^2$$

$$\|F_i(x^k) - F_i(x^*)\| \leq [L\|x^k - x^*\| + \|J_i(x^*)\|] \, \|x^k - x^*\|.$$

Por hipótese, $F_i(x^*) = 0$ para i =1,2,...,s, e como, pelo teorema anterior, temos que $\{x^k\}$ converge a x^*, obtemos:

$$A_3 \leq 2\sqrt{n}\, ML \, \|x^k - x^*\|^2.$$

Logo,

$$\|x^{k+1} - x^*\| \leq c \, \|x^k - x^*\|^2,$$

onde $c = (2\sqrt{n} + 1)ML$. ■

3.3. Resultados de convergência na presença de uma variedade de pontos fixos(caso m < n) (Sampaio [1985], Martínez e Sampaio [1986])

Consideremos novamente o sistema não linear (2.1), mas agora com **m < n**, isto é, o número de equações menor que o número de variáveis. Neste caso, as funções ø e G podem ter uma variedade de pontos fixos.

Seja

$$E = \{u : u = G(u)\} \subset D,$$

uma variedade de pontos fixos da aplicação G, definida por (3.2). Por questões de notação vamos reescrever (3.2) como

$$G(x) = \sum_{i=1}^{s} \lambda_i (x - J_i(u^0)^+ F_i(x)), \qquad (3.11)$$

onde $u^0 \in E$.

Vamos supor nesta seção que $J(u^0)$ tem posto completo (= m), $\forall\ u^0 \in E$.

Lema 3.3.1. u^0 é ponto fixo de G se e somente se $F(u^0) = 0$.

Prova:

(1) Vamos supor que u^0 é ponto fixo de G.

$$G(u^0) = \sum_{i=1}^{s} \lambda_i (u^0 - J_i(u^0)^+ F_i(u^0)).$$

Como $G(u^0) = u^0$, então:

$$\sum_{i=1}^{s} \lambda_i J_i(u^0)^+ F_i(u^0) = 0, \quad ou\ seja,$$

$$[J_1(u^0)^+ \ldots\ldots J_s(u^0)^+]\, LF(u^0) = 0, \qquad (3.12)$$

onde

$$L = \begin{pmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & & \\ & & \lambda_1 & & & & \\ & & & \lambda_2 & & & \\ & & & & \ddots & & \\ & & & & & \lambda_s & \\ & & & & & & \ddots \\ & & & & & & & \lambda_s \end{pmatrix} \begin{matrix} \uparrow \\ m_1 \\ \downarrow \\ \uparrow \\ m_2 \\ \downarrow \\ \\ \uparrow \\ m_s \\ \downarrow \end{matrix}$$

Desde que $J(u^0)$ tem posto completo m, então $J_i(u^0)$ tem posto completo m_i, para todo i = 1,2,...,s, e podemos reescrever (3.12) por:

$$\left[J_1(u^0)^T.(J_1(u^0)J_1(u^0)^T)^{-1} \ldots\ldots J_s(u^0)^T.(J_s(u^0)J_s(u^0)^T)^{-1}\right] LF(u^0) = ,$$

ou ainda,

$$J(u^0)^T \begin{pmatrix} (J_1(u^0)J_1(u^0)^T)^{-1} \\ \vdots \\ (J_s(u^0)J_s(u^0)^T)^{-1}) \end{pmatrix} LF(u^0) = 0. \qquad (3.13)$$

Vamos chamar

$$Y = \begin{pmatrix} (J_1(u^0)J_1(u^0)^T)^{-1} \\ \vdots \\ (J_s(u^0)J_s(u^0)^T)^{-1} \end{pmatrix} LF(u^0),$$

$y \in \mathbb{R}^m$. Então o sistema (3.13) pode ser reescrito por:

$$J(u^0)^T y = 0,$$

o qual só possui a solução trivial $y = 0$.
Assim:

$$\begin{pmatrix} (J_1(u^0)J_1(u^0)^T)^{-1} \\ \vdots \\ (J_s(u^0)J_s(u^0)^T)^{-1} \end{pmatrix} LF(u^0) = 0.$$

Como J_i tem posto completo e $\lambda_i > 0$, para $i = 1,2,\ldots,s$, então $F(u^0) = 0$.
(2) Vamos supor que u^0 é tal que $F(u^0) = 0$. Então

$$G(u^0) = \sum_{i=1}^{s} \lambda_i u^0$$

e, portanto, $G(u^0) = u^0$. ∎

Corolário 3.3.2. u^0 é ponto fixo de ϕ (definida em (3.1)) se e somente se $F(u^0) = 0$.

Trabalhando com sistema não lineares indeterminados ($m < n$), possivelmente não temos $\|G'(u^0)\| < 1$, como demonstramos na seção anterior. Baseados no trabalho de Meyn [1983], obtemos resultados de convergência local partindo de $\|G'(u^0)|_{R(J(u^0)^T)}\| < 1$.

como vamos demonstrar no lema a seguir.

Lema 3.3.3. $\|G'(u^0)|_{R(J(u^0)^T)}\| < 1$.

Prova: Sabemos que

$$\|G'(u^0)|_{R(J(u^0)^T)}\| = \max_{\substack{y \in R(J(u^0)^T), \\ |y|=1}} \|G'(u^0)y\|.$$

Seja então $y \in R(J(u^0)^T)$. De (3.3), vemos que:

$$G'(u^0) = \sum_{i=1}^{s} \lambda_i (I - J_i(u^0)^+ J_i(u^0)),$$

e, portanto,

$$\|G'(u^0)y\| = \|\sum_{i=1}^{s} \lambda_i (I - P_i)y\|,$$

onde $P_i y = J_i(u^0)^+ J_i(u^0)y$ é a projeção ortogonal de y sobre $R(J_i(u^0)^T)$, para i = 1,2,...,s, e $(I - P_i)y$ é a projeção ortogonal de y sobre $N(J_i(u^0))$, para i = 1,2,...,s (ver apêndice). Assim,

$$\|G'(u^0)y\| \leq \|y\|, \qquad (3.14)$$

onde a igualdade em (3.14) só ocorre se $y \in N(J(u^0))$. Neste caso, isto é impossível pois, por hipótese, $y \in R(J(u^0)^T)$. Portanto,

$$\|G'(u^0)y\| < \|y\|.$$

Como $\|y\| = 1$ define um conjunto compacto, $\|G'(u^0)y\|$ possui um máximo neste conjunto e

$$\max \|G'(u^0)y\| < \|y\|,$$

ou seja,

$$\| G'(u^0) |_{R(J(u^0)^T)} \| < 1. \qquad \blacksquare$$

Vamos agora mostrar que existe um conjunto aberto $D_1 \subset D$ tal que definindo

$$v : D_1 \to E \quad tal\ que$$

$$\| x - v(x) \| = \min_{u \in E} \| x - u \|, \tag{3.15}$$

então vale o seguinte lema.

Lema 3.3.4. Se $F \in C^2(D)$, temos que v é uma aplicação bem definida e contínua em $D_1 \subset D$.

Prova: v(x) pode ser definida como solução do problema

$$\begin{aligned} &\min_{u} \| x - u \|^2 \\ &s.a \quad F(u) = 0. \end{aligned} \tag{3.16}$$

As condições de Lagrange para (3.16) nos levam ao sistema:

$$\begin{aligned} (x-u) + J(u)^T \gamma &= 0 \\ F(u) &= 0, \ \gamma \in \mathbb{R}^m, \end{aligned} \tag{3.17}$$

o qual possui (n+m) equações e (2n+m) variáveis (x,u,γ).

Uma solução particular para o sistema (3.17) pode ser: $(u^0,u^0,0) \in \mathbb{R}^{2n+m}$. Além disso, com a hipótese que $F \in C^2(D)$, existe a matrix Jacobiana $\bar{J}$, $(n+m)x(n+m)$, do sistema (3.17), em relação às variáveis (u,γ), a qual avaliada em $(u^0,u^0,0)$ é dada por:

$$\bar{J}(u^0,u^0,0) = \begin{pmatrix} -I & \vdots & J(u^0)^T \\ J(u^0) & \vdots & 0 \\ \leftarrow\ n\ \rightarrow & & \leftarrow\ m\ \rightarrow \end{pmatrix}$$

Podemos mostrar também que $\bar{J}(u^0,u^0,0)$ é não singular:

-consideremos o vetor $\begin{pmatrix} y \\ z \end{pmatrix} \in \mathbb{R}^{n+m}$, e o sistema

$$\bar{J}(u^0,u^0,0)\begin{pmatrix} y \\ z \end{pmatrix} = 0 \qquad (3.18)$$

Então:

$$-y + J(u^0)^T z = 0 \qquad (3.19)$$

$$J(u^0)y = 0 \qquad (3.20)$$

De (3.19) temos que $y = J(u^0)^T z$, e substituindo y em (3.20) obtemos $J(u^0)J(u^0)^T z = 0$. Como estamos supondo que $J(u^0)$ tem posto completo, então $z = 0$, o que implica $y = 0$. Portanto, o sistema (3.18) tem uma única solução $\begin{pmatrix} y \\ z \end{pmatrix} = 0$, o que prova que $\bar{J}(u^0,u^0,0)$ é não singular.

Dessa forma temos todas as hipóteses necessárias para que o Teorema da Função Implícita nos garanta que existe um conjunto aberto $D_1 \subset D$, com $u^0 \in D_1$, tal que para $x \in D_1$, existem funções $v_i(x)$, $i = 1,2,\ldots,n+m$, tais que $v_i \in C^1$, $i = 1,2,\ldots,n+m$, $v(u^0) = (u_1^0,\ldots,u_n^0,0,\ldots,0)$ e $(u^0,v(u^0))$ é solução para o sistema (3.17).

Em particular demonstramos que para $x \in D_1$ existe a função $v(x)$, $v(x) = (v_1(x),v_2(x),\ldots,v_n(x))$, tal que $v \in C^1$ e $v(u^0) = u^0$.

■

Lema 3.3.5. Seja $u^0 \in E$. Existem $\alpha \in [0,1)$ e $\epsilon > 0$ tais que para todo $x \in B(u^0,\epsilon)$,

$$\|G(x) - v(x)\| \leq \alpha\|x - v(x)\|.$$

Prova: Por (3.16) vemos que para todo $x \in D$ satisfazendo $v(x) = u^0$ temos que $x - u^0 \in R(J(u^0)^T)$. Então:

$$\|G'(u^0)(x-u^0)\| \leq \beta\|x-u^0\|, \ \beta \in [0,1).$$

Desde que $G'(v(x))$ é contínua, existem $\epsilon > 0$, $\beta_1 \in [0,1)$, tais que:

$$\|G'(v(x))(x - v(x))\| \leq \beta_1\|x - v(x)\|,$$

para todo $x \in B(u^0,\epsilon)$. Para ϵ suficientemente pequeno, existe $\gamma < 1-\beta_1$ tal que

$$\|G(x) - G(v(x)) - G'(v(x))(x - v(x))\| \leq \gamma\|x - v(x)\|.$$

Portanto:

$$\begin{aligned}\|G(x) - v(x)\| &= \|G(x) - G'(v(x))(x-v(x)) + G'(v(x))(x-v(x)) - v(x)\| \\ &\leq \|G(x) - v(x) - G'(v(x))(x-v(x))\| + \|G'(v(x))(x-v(x))\| \\ &\leq (\gamma + \beta_1)\|x - v(x)\|,\end{aligned}$$

e a tese se verifica para $\alpha = \gamma + \beta_1$. ∎

Lema 3.3.6. Seja $u^0 \in E$. Existem $\beta \in [0,1)$ e $\epsilon > 0$ tais que, para todo $x \in B(u^0,\epsilon)$,

$$\|\phi(x) - v(x)\| \leq \beta\|x - v(x)\|.$$

Prova:

$$\|\phi(x) - v(x)\| \leq \|\phi(x) - G(x)\| + \|G(x) - v(x)\|.$$

Pelo lema 3.3.5 existem $\alpha \in [0,1)$ e $\epsilon > 0$ tais que, para todo $x \in B(u^0,\epsilon)$,

$$\|G(x) - v(x)\| \leq \alpha \|x - v(x)\|.$$

Agora, por (3.9),

$$\|\phi(x) - G(x)\| \leq 2\sqrt{n}ML \|x - u^0\| \sum_{i=1}^{s} \lambda_i \|F_i(x)\|, \qquad (3.20)$$

e por (3.5):

$$\|F_i(x) - F_i(v(x))\| - \|J_i(u^0)\| \|x-v(x)\| \leq L\|x-v(x)\| M_x,$$

onde

$$M_x = \max\{\|x-u^0\|, \|v(x)-u^0\|\}.$$

Como $F_i(v(x)) = 0$, para todo $i = 1,2,\ldots,s$, então:

$$\|F_i(x)\| \leq [LM_x + \|J_i(u^0)\|] \|x - v(x)\|. \qquad (3.21)$$

Tomemos $\alpha_2 \in [0,1)$, com $\alpha_2 + \alpha \leq \beta < 1$ e $\epsilon_2 > 0$ tais que, de (3.20) e (3.21),

$$\|\phi(x) - G(x)\| \leq \alpha_2 \|x - v(x)\|,$$

para todo $x \in B(u^0,\epsilon_2)$.

Portanto, existe $\epsilon \leq \min\{\epsilon,\epsilon_2\}$ tal que

$$\|\phi(x) - v(x)\| \leq \beta \|x - v(x)\|,$$

para todo $x \in B(u^0,\epsilon)$. ■

Teorema 3.3.7. Sejam $u^0 \in E$, $0 < \delta < \epsilon$. Existe $B(u^0,\delta) \subset B(u^0,\epsilon)$ tal que para todo $x^0 \in B(u^0,\delta)$, a sequência definida por $x^{k+1} = \phi(x^k)$ converge para um ponto de E.

Prova: (Ver Meyn [1983]) Pelo lema 3.3.6, para $\beta < 1$ existe $\epsilon > 0$ tal que

$$\|\phi(x) - v(x)\| \leq \beta \|x - v(x)\|, \ \forall x \in B(u^0,\bar{\epsilon}).$$

Sabemos então que $\frac{1+\beta}{1-\beta} \geq 1$ e, portanto,

$$\|x^0 - v(x^0)\| \leq \frac{1+\beta}{1-\beta} \|x^0 - v(x^0)\|.$$

Assim, δ pode ser escolhido tal que, para todo $x^0 \in B(u^0,\delta)$,

$$\frac{1+\beta}{1-\beta} \|x^0 - v(x^0)\| + \|v(x^0) - u^0\| < \delta < \bar{\epsilon}.$$

A primeira parte da demonstração consiste em:

(a) mostrar que $\{x^k\} \subset B(u^0,\bar{\epsilon})$, $k = 0,1,2,\ldots$

(a.1) mostrar que

$$\|x^k - v(x^{k-1})\| \leq \beta^k \|x^0 - v(x^0)\|.$$

Prova por indução:

. Se $k = 1$:

$$\|x^1 - v(x^0)\| = \|\phi(x^0) - v(x^0)\| \leq \beta \|x^0 - v(x^0)\|.$$

. Vamos supor que

$$\|x^{k-1} - v(x^{k-2})\| \leq \beta^{k-1} \|x^0 - v(x^0)\|.$$

. Então:

$$\begin{aligned}\|x^k - v(x^{k-1})\| &= \|\phi(x^{k-1}) - v(x^{k-1})\| \\ &\leq \beta \|x^{k-1} - v(x^{k-1})\| \\ &\leq \beta \|x^{k-1} - v(x^{k-2})\| \\ &\leq \beta^k \|x^0 - v(x^0)\| .\end{aligned}$$

(a.2) mostrar que

$$\|x^k - u^0\| \leq (1+\beta) \sum_{i=0}^{k-1} [\beta^i \|x^0 - v(x^0)\|] + \|v(x^0) - u^0\| .$$

Prova por indução:

. Se k = 1:

$$\begin{aligned}\|x^1-u^0\| &\leq \|x^1-v(x^0)\| + \|v(x^0)-u^0\| \\ &\leq \beta \|x^0-v(x^0)\| + \|v(x^0) - u^0\| \\ &\leq (1+\beta)\|x^0-v(x^0)\| + \|v(x^0)-u^0\| .\end{aligned}$$

. Vamos supor que

$$\|x^k-u^0\| \leq (1+\beta) \sum_{i=0}^{k-1} [\beta^i\|x^0-v(x^0)\|] + \|v(x^0) - u^0\| .$$

. Então:

$$\begin{aligned}\|x^{k+1} - u^0\| &\leq \|x^{k+1} - v(x^k)\| + \|v(x^k) - x^k\| + \|x^k - u^0\| \\ &\leq \beta^{k+1}\|x^0 - v(x^0)\| + \|x^k - v(x^{k-1})\| + \|x^k - u^0\| \\ &\leq \beta^{k+1}\|x^0 - v(x^0)\| + \beta^k\|x^0 - v(x^0)\| + \|x^k-u^0\| \\ &\leq \beta^{k+1}\|x^0-v(x^0)\| + \beta^k\|x^0-v(x^0)\| + \\ &\quad (1+\beta) \sum_{i=0}^{k-1} [\beta^i\|x^0 - v(x^0)\|] + \|v(x^0) - u^0\| \\ &\leq (1+\beta)\beta^k\|x^0 - v(x^0)\| + (1+\beta) \sum_{i=0}^{k-1} \beta^i\|x^0 - v(x^0)\| + \\ &\quad \|v(x^0) - u^0\|\end{aligned}$$

$$\leq (1+\beta) \sum_{i=0}^{k} [\beta^i \|x^0 - v(x^0)\|] + \|v(x^0) - u^0\|$$

$$\leq \frac{1+\beta}{1-\beta} \|x^0 - v(x^0)\| + \|v(x^0) - u^0\|$$

$$< \bar{\epsilon}.$$

Assim fica demonstrado que $\{x^k\} \subset B(u^0, \bar{\epsilon})$, para todo $k = 0,1,2,\ldots$, e todo ponto de acumulação de $\{x^k\}$ é um ponto fixo de G.

(b) Mostrar que $\{x^k\}$ é uma sequência de Cauchy.

$$\|x^p - x^q\| \leq \|x^p - v(x^{p-1})\| + \|x^{p-1} - v(x^{p-1})\| + \|x^{p-1} - x^q\|, \; p-1 > q.$$

$$\|x^{p-1} - x^q\| \leq \|x^{p-1} - v(x^{p-2})\| + \|x^{p-2} - v(x^{p-2})\| + \|x^{p-2} - x^q\|, \; p-2 > q.$$

Repetindo este argumento para

$$\|x^{p-i} - x^q\|, \; i = 2,3,\ldots, p-q-1, \text{ temos:}$$

$$\|x^p - x^q\| \leq \sum_{i=1}^{p-q} [\|x^{q+1} - v(x^{q+i-1})\| + \|x^{q+i-1} - v(x^{q+i-1})\|]$$

$$\leq \sum_{i=1}^{p-q} [\beta^{q+i} \|x^0 - v(x^0)\| + \beta^{q+i-1} \|x^0 - v(x^0)\|]$$

$$\leq \frac{2\beta^q}{1-\beta} \|x^0 - v(x^0)\|$$

$$\leq \frac{2\beta^q}{1-\beta} \delta.$$

Assim, dado $\epsilon > 0$, pedimos que $\frac{2\beta^q}{1-\beta} \delta < \epsilon$, ou seja,

$$q > \frac{\ln\left(\frac{\epsilon(1-\beta)}{2\delta}\right)}{\ln\beta}.$$

Portanto, existe $N = \dfrac{\ln\left(\frac{\epsilon(1-\beta)}{2\delta}\right)}{\ln\beta}$ tal que, se $p > q > N$, $\| x^p - x^q \| < \epsilon$, o que mostra que $\{x^k\}$ tem um único ponto de acumulação. ■

CAPÍTULO 4

RESULTADOS DE CONVERGÊNCIA USANDO ESTRATÉGIAS GLOBALIZADORAS

Vamos considerar novamente o problema (2.5):

$$Min_{x\epsilon D\subset\mathbb{R}^n} f(x) = \frac{1}{p}\|F(x)\|_p^p \ , \ 1<p\leq 2.$$

Neste capítulo vamos mostrar que os algoritmos da seção 2.2 possuem convergência global, como também, sob certas condições, apresentam convergência local. Porém tratamos de convergência local "fraca", no sentido que pode ser necessária mais de uma avaliação de função por iteração.

4.1 Convergência Global

Teorema 4.1.1. Se x^* é um ponto de acumulação da sequência $\{x^k\}$ gerada por um dos algoritmos da seção 2.2, então $g(x^*)=\nabla f(x^*)=0$.

Prova: (Ver Martínez [1987]) Se x^* é um ponto limite de $\{x^k\}$ então $\exists\ K_1 \subset \mathbb{N}$ e a subsequência $\{x^k\}_{k\epsilon K_1}$ tal que $\lim_{k\epsilon K_1} x^k = x^*$.

Seja $B=\{x^k : k\epsilon K_1\}$; B é um subconjunto limitado de D. Como $f\epsilon C^1(D)$, então

$$\|g(x^k)\| \leq C_1 \ , \ \forall\, x^k \epsilon B.$$

Vamos supor que $g(x^*) \neq 0$.

Como $\lim_{k\epsilon K_1} g(x^k) = g(x^*)$, então $\exists\ K_2 \subset K_1$ tal que

$$\|g(x^k)\| \geq C_2 > 0 \ , \ \forall\, k \epsilon K_2.$$

Com estas considerações iniciais, vamos então desenvolver a demonstração em duas partes:

(1) - relativa aos algoritmos 2.2.1 e 2.2.2

Sabemos que as direções d_k, k=0,1,2,..., geradas por qualquer um destes algoritmos, satisfazem:

$$\underline{M}\ \|g(x^k)\| \leq \|d_k\| \leq \overline{M}\ \|g(x^k)\| \ , \ para\ 0<\underline{M}<\overline{M} \ , \qquad (4.1)$$

e

$$<d_k, g(x^k)> \leq -\theta_1\ \|d_k\|\ \|g(x^k)\| \ , \ para\ \theta_1 \epsilon (0,1) \qquad (4.2)$$

Então

$$\underline{M}C_2 \leq \underline{M}\ \|g(x^k)\| \leq \|d_k\| \leq \overline{M}\ \|g(x^k)\| \leq \overline{M}C_1, \forall k \epsilon K_2. \qquad (4.3)$$

Assim, $\exists\ K_3 \subset K_2$ tal que $\lim_{k\epsilon K_3} d_k = d$.

Tomando limites, para $k \epsilon K_3$, em ambos os lados de (4.1) e (4.2), obtemos:

$$\underline{M}\ \|g(x^*)\| \leq \|d\| \leq \overline{M}\ \|g(x^*)\| \ e$$

$$<d, g(x^*)> \leq -\theta_1\ \|d\|\ \|g(x^*)\| .$$

Agora, considerando a sequência $\{s_k\} = \{\alpha_k\ d_k\}$, vemos que duas possibilidades podem ocorrer:

(1.a) $\exists \beta>0$ *tal que* $\|s_k\| \geq \beta\ \|g(x^K)\|\ \forall k \epsilon K_3$.

(1.b) *o oposto de* (1.a).

Vamos analisar cada uma destas possibilidades:

(1.a) $\beta C \leq \beta\ \|g(x^K)\| \leq \|\alpha_k d_k\| = \alpha_k\ \|d_k\| \leq \alpha_k \bar{M} C_1\ , \forall k \epsilon K_3$.

Então $\exists\ K_4 \subset K_3$ tal que $\lim_{k \epsilon K_4} s_k = s \neq 0$.

Mas, desde que a condição (4.2) é satisfeita, podemos ver que $s_k = \alpha_k d_k$ também satisfaz a condição

$$<s_k, g(x^k)> \leq -\theta_1\ \|s_k\|\ \|g(x^k)\|\ . \qquad (4.4)$$

Então, aplicando limites em ambos os lados de (4.4) para $k \epsilon K_4$, obtemos:

$$<s, g(x^*)> \leq -\theta_1\ \|s\|\ \|g(x^*)\|\ ,$$

ou seja,

$$<s, g(x^*)> \leq -\theta_1 \gamma\ ,\ \gamma>0.$$

Como s_k satisfaz:

$$f(x^k+s_k) \leq f(x^k) + \theta_2 <g(x^k, s_k>\ ,\ para\ \theta_2 \epsilon (0,1)\ ,$$

aplicando limites em ambos os lados para $k \epsilon K_4$ temos:

$$\begin{aligned} f(x^*+s) &\leq f(x^*) + \theta_2 <g(x^*, s> \\ &\leq f(x^*) - \theta_1 \theta_2 \gamma \\ &< f(x^*)\ . \end{aligned}$$

Assim, $\exists\ k_0 \in N$ tal que para $k \geq k_0$, $k \in K_4$,

$$f(x^k + s_k) < f(x^*) \ ,$$

o que é absurdo.

(1.b) Neste caso $s_k \neq d_k$ e para todo $k \in K_5 \subset K_3$, $\lim s_k = 0$.

Então, para todo $k \in K_5$,

$$\exists\ \bar{s}_k = \bar{\alpha} d_k \ , \ \bar{\alpha} \in (0,1) \ , \ tal\ que$$
$$\|\bar{s}_k\| \leq \mu \ \|s_k\| \ , \ e$$

$$f(x^k + \bar{s}_k) > f(x^k) + \theta_2 \langle g(x^k), \bar{s}_k \rangle . \qquad (4.5)$$

Vemos ainda que $\lim_{k \in K_5} \bar{s}_k = 0$.

Pelo Teorema do Valor Médio e pela desigualdade (4.5) temos:

$$\langle g(x^k + \xi_k \bar{s}_k), \bar{s}_k \rangle \ > \ \theta_2 \ \langle g(x^k), \bar{s}_k \rangle \ , \ \xi_k \in (0,1) ;$$

$$\langle g(x^k + \xi_k \bar{s}_k), \frac{\bar{s}_k}{\|\bar{s}_k\|} \rangle \ > \ \theta_2 \ \langle g(x^k), \frac{\bar{s}_k}{\|\bar{s}_k\|} \rangle . \qquad (4.6)$$

Seja $K_6 \subset K_5$ tal que

$$\lim_{k \in K_6} \frac{\bar{s}_k}{\|\bar{s}_k\|} = z .$$

Aplicando limites para $k \in K_6$ em ambos os lados de (4.6) obtemos:

$$\langle g(x^*), z \rangle \ \geq \ \theta_2 \ \langle g(x^*), z \rangle .$$

Como $\theta_2 \in (0,1)$, esta desigualdade só ocorre se $\langle g(x^*), z \rangle \geq 0$.

Mas,

$$\begin{aligned} \langle g(x^k), \bar{s}_k \rangle &= \langle g(x^k), \bar{\alpha}_k d_k \rangle \\ &= \bar{\alpha}_k \langle g(x^k), d_k \rangle \\ &\leq -\theta_1 \bar{\alpha}_k \, \| g(x^k) \| \, \| d_k \| \\ &\leq -\theta_1 \, \| g(x^k) \| \, \| \bar{s}_k \| . \end{aligned}$$

Logo

$$\langle g(x^k), \frac{\bar{s}_k}{\| \bar{s}_k \|} \rangle \leq -\theta_1 \, \| g(x^k) \| . \qquad (4.7)$$

Aplicando limites para $k \in K_6$ em ambos os lados de (4.7) obtemos:

$$\langle g(x^*), z \rangle \leq -\theta_1 \, \| g(x^*) \| < 0,$$

o que é uma contradição.

Portanto, se x^* é um ponto limite da sequência $\{x^k\}$ gerada pelo algoritmo 2.2.1 ou 2.2.2, então $g(x^*) = 0$.

(2) - relativa ao algoritmo 2.2.3

Este algoritmo gera direções d_k^i $i=1,2$, $k=0,1,2,\ldots$, que satisfazem

$$\underline{M} \, \| g(x^k) \| \leq \| d_k^i \| \leq \bar{M} \| g(x^k) \| . \qquad (4.8)$$

Então temos que:

$$\underline{M} C_2 \leq M \, \| g(x^k) \| \leq \| d_k^i \| \leq \bar{M} \, \| g(x^k) \| \leq \bar{M} C_1 ,$$
$$\forall k \in K_2 , i=1,2.$$

Logo,

$$\exists K_3 \subset K_2 \text{ tal que } \lim_{k\in K_3} d_k{}^i = d^i \text{ , } i=1,2 \text{ e}$$

$$\underline{M} \, \|g(x^*)\| \leq \|d^i\| \leq \overline{M} \, \|g(x^*)\|, \; 1=1,2.$$

Além disso, como a condição

$$\langle d_k{}^i, g(x^k)\rangle \leq -\theta_1 \, \|d_k{}^i\| \, \|g(x^k)\| \text{ , para } \theta_1 \in (0,1), \qquad (4.9)$$

deve ser satisfeita, então

$$\langle d^i, g(x^*)\rangle \leq -\theta_1 \, \|d^i\| \, \|g(x^*)\|.$$

Temos mais uma vez, duas possibilidades, considerando a sequência $\{s_k\}$ gerada ao final do passo 6 do algoritmo:

(2.a) $\exists \beta>0$ *tal que* $\|s_k\| \geq \beta \, \|g(x^k)\|$, $\forall k \in K_3$;

(2.b) *o oposto de* (2.a).

Analisando ambas possibilidades:

(2.a) $$\beta C_1 \leq \beta \, \|g(x^k)\| \leq \|s_k\| \leq \|d_k{}^2\| \leq \overline{M} \, C_1 \text{ , } \forall k \in K_3.$$

Então $\exists$ $K_4 \subset K_3$ tal que $\lim_{k\in K4} s_k = s \neq 0$.

Por (2.12), s_k está no cone positivo determinado por $d_k{}^1$ e $d_k{}^2$. Podemos mostrar que, então, s_k satisfaz a condição:

$$\langle s_k, g(x^k)\rangle \leq -\theta_1 \, \|s_k\| \, \|g(x^k)\|,$$

desde que:

$$s_k = \lambda_1 \, d_k{}^1 + \lambda_2 \, d_k{}^1, \; \lambda_1, \; \lambda_2 \geq 0,$$

$$\begin{aligned}
\langle s_k, g(x^k)\rangle &= \langle \lambda_1 d_k^{\,1} + \lambda_2 d_k^{\,2}, g(x^k)\rangle \\
&= \langle \lambda_1 d_k^{\,1}, g(x^k)\rangle + \langle \lambda_2 d_k^{\,2}, g(x^k)\rangle \\
&\leq -\theta_1 \lambda_1 \|d_k^{\,1}\| \, \|g(x^k)\| - \theta_1 \lambda_2 \|d_k^{\,2}\| \, \|g(x^k)\| \\
&\leq -\theta_1 \|g(x^k)\| \, [\lambda_1 \|d_k^{\,1}\| + \lambda_2 \|d_k^{\,2}\|] \\
&\leq -\theta_1 \|s_k\| \, \|g(x^k)\|,
\end{aligned}$$

e a demonstração segue análoga a (1.a), levando a uma contradição.

(2.b) Neste caso, existe $K_5 \subset K_3$ tal que

$$\lim_{k \in K_5} s_k = 0 \text{ e } s_k \neq d_k^{\,2} \ \forall k \in K_5.$$

Então existe

$$\bar{s}_k \in C(d_k^{\,1}, d_k^{\,2}) \text{ tal que}$$

$$\|\bar{s}_k\| \leq \mu \|s_k\| \text{ e } f(x^k + \bar{s}_k) > f(x^k) + \theta_2 \langle g(x^k), \bar{s}_k \rangle.$$

Novamente, pelos mesmos argumentos usados em (1.b), chegamos a uma contradição.

Concluimos então que, se x^* é um ponto limite da sequência $\{x^k\}$ gerada por um dos algoritmos da seção 2.2, então $g(x^*) = 0$. ■

Devemos observar que o teorema 4.1.1 é um resultado de convergência global mais geral, no sentido de que se aplica a qualquer algoritmo que gere uma sequência $\{x^k\}$ por

$$x^{k+1} = x^k + s_k \ , \ k=0,1,2,\ldots,$$

onde s_k satisfaz:

$$\langle g(x^k), s_k \rangle \leq -\theta_1 \|g(x^k)\| \|s_k\|,$$

$$\underline{M} \|g(x^k)\| \leq \|s_k\| \leq M \|g(x^k)\| \quad e$$

$$f(x^k) + s_k \leq f(x^k) + \theta_2 \langle g(x^k), s_k \rangle,$$

com θ_1, $\theta_2 \in (0,1)$, $0<\underline{M}<\overline{M}<\infty$, $k=0,1,2,\ldots$

Teorema 4.1.2. Seja $\{x^k\}$ uma sequência limitada gerada pelo algoritmo 2.4. Então:

$$(i)\ \lim_{k\to\infty} \|\overline{w}_k^a\| = 0 \ ou$$

$$(ii)\ \inf_k \|g(x^k)\| = 0.$$

Prova: Considerando a sequência $\{x^k\}$ gerada pelo algoritmo 2.4, duas possibilidades podem ocorrer:
(i) $\{x^k\}$ é gerada por um número infinito de iterações ordinárias. Supomos então que

$$K_1 = \{k_1, k_2, k_3, \ldots\} \subset \mathbb{N}, \text{ com } k_1 \geq q \text{ e } k_i < k_i + 1,$$
$$\forall i = 1,2,3,\ldots,$$

é um conjunto de índices de iterações ordinárias. Portanto,

$$\|\overline{w}_{k_i}^a\| \leq \gamma \|\overline{w}_{k_i - q}^a\|, \qquad (4.10)$$
$$\forall i = 1,2,3,\ldots$$

Definimos indutivamente

$$k_i' \ , i = 1,2,3,\ldots, \text{ por :}$$

$$(a)\ k_1' = k_1 - q; \qquad (4.11)$$

$$(b)\ \text{para } i = 1,2,3,\ldots,$$

$$\text{se} \quad k'_i = k_j - q \quad \text{então}$$

$$k'_{i+1} = k_l - q, \tag{4.12}$$

para algum l tal que $k_l - q \geq k_j$.

Por (4.10) - (4.12) temos:

$$\|\bar{w}^a_{k'_{i+1}}\| \leq \gamma \, \|\bar{w}^a_{k_i}\|, \tag{4.13}$$

para todo i=1,2,3,... .

Mas, por (2.19),

$$\|\bar{w}^a_{k+1}\| \leq \|\bar{w}^a_k\|, \tag{4.14}$$

para todo k=0,1,2,... .

Lembrando que $\gamma \in (0,1)$, temos, então, por (4.13) e (4.14), que:

$$\lim_{k\to\infty} \|\bar{w}^a_k\| = 0.$$

(ii) existe K tal que para todo $k \geq K$ as aproximações x^k serão obtidas por um dos algoritmos da seção 2.2. Neste caso, o Teorema 4.1.1 garante o resultado desejado. ∎

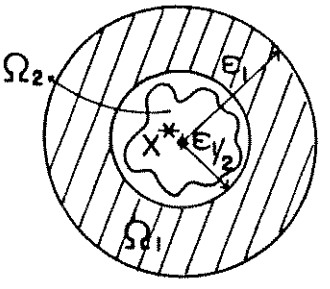
4.2 - Alguns resultados especiais de convergência

Vamos considerar nesta seção a sequência $\{x^k\}$ gerada por qualquer um dos algoritmos da seção 2.2.

Teorema 4.2.1 Seja x^* um mínimo local de f em D e um ponto estacionário isolado de g(x). Existe $\epsilon > 0$, tal que se $x^0 \in B(x^*,\epsilon)$, então $\lim x^k = x^*$.

Prova: Como x^* é um mínimo local estrito de f em D, então $\exists$ $\epsilon_1 > 0$ tal que $f(x) > f(x^*)$ se $0 < \|x-x^*\| \leq \epsilon_1$.
Consideremos o conjunto

$$\Omega_1 = \{x : \epsilon_1/2 \leq \|x-x^*\| \leq \epsilon_1\}.$$



Como f é uma função contínua e Ω_1 um conjunto compacto, então f tem um mínimo em Ω_1. Seja $m = \min_{x\in\Omega_1} f(x)$.

Consideremos agora o conjunto

$$\Omega_2 = \{x \in B(x^*,\epsilon_1) : f(x) < m\}.$$

Observamos que:

. $\Omega_2 \neq \emptyset$, desde que $x^* \in \Omega_2$;

. $\Omega_2 \subset B(x^*,\epsilon_1/2)$ e é aberto, pois, caso contrário, poderia haver $x \in \Omega_1 \cap \Omega_2$ tal que $f(x) \geq m$ e $f(x) < m$. (absurdo)

Vamos supor que para todo $x \in \Omega_2$ temos $\bar{M}\,\|g(x)\| < \epsilon_1/2$; e escolhamos $\epsilon > 0$ tal que $B(x^*,\epsilon) \subset \Omega_2$.

Por hipótese, $x^0 \in B(x^*,\epsilon) \subset \Omega_2$.

Vamos supor que $x^k \in \Omega_2$, e queremos mostrar que $x^{k+1} \in \Omega_2$, $\forall$ k= 0,1,2,... .

Sabemos que $\|s_k\| = \|x^{k+1} - x^k\|$ satisfaz: $\|s^k\| \leq M\,\|g(x^k)\|$, para k=0,1,2,..., qualquer que seja o algoritmo da seção 2.2 aplicado. Assim

$$\begin{aligned}\|x^{k+1}-x^*\| &\leq \|x^{k+1}-x^k\| + \|x^k-x^*\| \\ &\leq \bar{M}\,\|g(x^k)\| + \|x^k-x^*\| < \epsilon_1/2 + \epsilon_1/2,\end{aligned}$$

ou seja, $\|x^{k+1} - x^*\| < \epsilon_1$ e, então, $x^{k+1} \in B(x^*,\epsilon_1)$. Mas $f(x^{k+1}) < f(x^k)$ e $f(x^k) < m$, desde que $x^k \in \Omega_2$; portanto, $f(x^{k+1}) < m$ e, então, mostramos que $x^{k+1} \in \Omega_2$, $\forall$ k=0, 1, 2, ...

Assim, toda a sequência $\{x^k\}$ está em $\Omega_2 \subset B(x^*,\epsilon_1/2)$ e todo ponto de acumulação de $\{x^k\}$ está em $B(x^*,\epsilon_1/2)$. Seja $\hat{x}$ um ponto de acumulação. Pelo Teorema 4.1.1, $\hat{x}$ é tal que $g(\hat{x})=0$; como, por hipótese, x^* é um ponto estacionário isolado de $g(x)$, então $\hat{x} \equiv x^*$, ou seja, a sequência $\{x^k\}$ converge para x^*, um mínimo local de f. ∎

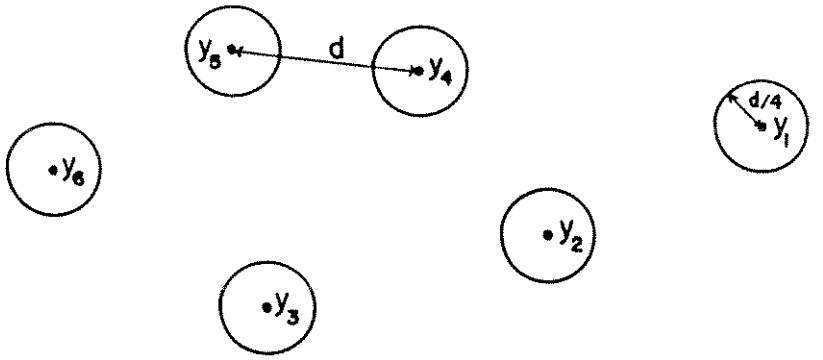
Teorema 4.2.2. Dado $x^0 \in D$, consideremos os conjuntos:

$$\Omega_1 = \{x: f(x) \leq f(x^0)\},$$
$$\Omega_2 = \{x: g(x) = 0\}.$$

Se Ω_1 é compacto e $\Omega_1 \cap \Omega_2$ tem um número finito de pontos, então a sequência $\{x^k\}$ converge.

Prova: $\{x^k\} \subset \Omega_1$, $\forall$ k= 0,1 ,2, ... Como, por hipótese, Ω_1 é compacto, então $\{x^k\}$ possui uma subsequência convergente cujo limite x é tal que g(x)=0 (pelo Teorema 4.1.1). Como, também por hipótese, $\Omega_1 \cap \Omega_2$ é finito, $\{x^k\}$ possui um conjunto finito de pontos limites, que vamos chamar: y_1, y_2,...y_l. Consideremos $d = \min_{i \neq j} d(y_i,y_j)$, i,j = 1, 2, ..., l, e consideremos as bolas abertas $B(y_i,d/4)$, i = 1, 2, ..., l; seja

$$B = \bigcup_{i=1}^{l} B(y_i,d/4).$$



Podemos afirmar que o conjunto Ω_1 - B possui um número finito de pontos de $\{x^k\}$, k = 0, 1, 2, ..., pois se houver infinitos, ou seja, se o conjunto $K_1 = \{k : x^k \in \Omega_1\text{-}B\}$ é infinito, a subsequência $\{x^k\}$, $k \in K_1$, que está contida em um conjunto compacto , possui um ponto de acumulação diferente de y_i, i = 1, 2, ..., l, o que é uma contradição.

Seja então k o maior índice a partir do qual não há mais pontos de $\{x^k\}$ em Ω_1-B. Se para $k \geq \bar{k}$, $x^k \in B_i$ e $x^{k+1} \in B_j$, $i, j = 1, 2, \ldots, l$, então:

$$\| x^{k+1} - x^k \| \geq d/4.$$

Se $l > 1$, isto acontece infinitas vezes; mas então $\| x^{k+1} - x^k \| \not\to 0$, o que é uma contradição.

Portanto, a sequência $\{x^k\}$ possui um único ponto de acumulação. ∎

CAPITULO 5

IMPLEMENTAÇÃO COMPUTACIONAL

Os algoritmos descritos no capítulo 2, com exceção do algoritmo 2.2.3., foram implementados computacionalmente no VAX 11/785 - Sistema Operacional VMS - da UNICAMP, em linguagem FORTRAN 77, com aritmética de precisão simples. A implementação está voltada para sistemas de grande porte, com estrutura esparsa da matria Jacobiana.

5.1. Divisão em Blocos

Trabalhamos com os sistemas divididos em blocos quadrados, ou seja, $m_i = n$ para todo $i = 1,2,...,s$. Com isso, o cálculo de cada projeção ortogonal reduz-se à resolução de um sistema linear quadrado. Para obter a solução de cada um destes sistemas lineares usamos a subrotina MA28 de Harwell Subroutine Library (Duff [1977]). Esta subrotina implementa o método da Decomposição LU para sistemas grandes e esparsos, usando estratégia de pivotamento aplicada de forma a manter a esparsidade da matriz a ser decomposta, mas controlando também a perda de precisão por

erros de arredondamento. Num certo estágio r do processo de decomposição, r = 0,1,2,..., o elemento $a_{rr}^{(r)}$ escolhido para pivô deve satisfazer

$$|a_{rr}^{(r)}| \geq u \max_{j \geq r} |a_{rj}^{(r)}|$$

$$ou \quad |a_{rr}^{(r)}| \geq u \max_{i \geq r} |a_{ir}^{(r)}|,$$

onde $u \in [0,1]$ deve ser fornecido pelo usuário (observamos então que conforme u fica mais perto de zero, mais a preservação da esparsidade é priorizada em relação à estabilidade).

Nos algoritmos que incorporam estratégias de globalização (algoritmos 2.2.1, 2.2.2 e 2.4), o parâmetro α usado na busca linear é determinado por um procedimento de interpolação quadrática seguida por interpolação cúbica, descrito em Dennis e Schnabel [1983, pag. 126-129].

Ao aplicarmos o algoritmo 2.2.2 há necessidade de resolvermos, a cada iteração k, um sistema linear 2x2, para a obtenção da direção v_k, conforme (2.8). Supondo que a matriz Jacobiana $J(x^k)$ tem posto completo para todo k, a matriz de coeficientes do sistema é definida positiva e, assim, usamos Decomposição de Cholesky. (Como medida de proteção, a execução do programa é interrompida quando a matriz não é definida positiva, e uma mensagem é impressa).

5.2. Critérios de Parada

No algoritmo 2.1, ao qual nos referimos por método "local", usamos como critério de convergência

$$\|x^{k+1} - x^k\|_\infty \leq \epsilon_1, \tag{5.1}$$

$$ou \quad \frac{\|x^{k+1} - x^k\|_\infty}{\|x^0\|_\infty} \leq \epsilon_1, \tag{5.2}$$

onde ϵ_1 é um parâmetro real, próximo de zero, fornecido pelo usuário, o qual deve optar pelo emprego do critério absoluto (5.1) ou do relativo (5.2).

Nos algoritmos "globais", o objetivo é encontrar x^* tal que $\| g(x^*) \| \approx 0$. No entanto, este objetivo pode ser difícil de ser atingido, especialmente quando os elementos da matriz Jacobiana, em pontos próximos de x^*, forem grandes. Assim, declaramos convergência nos métodos "globais" quando um dos tres critérios a seguir é satisfeito:

$$\|x^{k+1} - x^k\|_\infty \leq \epsilon_1$$

(que pode ser substituído pelo critério relativo (5.2)),

$$\|g(x^k)\|_\infty \leq \epsilon_2, \tag{5.3}$$

$$f(x^k) \leq \epsilon_3, \tag{5.4}$$

onde ϵ_2 e ϵ_3, como ϵ_1, são parâmetros reais, próximos de zero, fornecidos pelo usuário, e f(x) é a função objetivo definida por (2.5).

A execução do programa também pode ser interrompida se um número máximo de iterações, fixado pelo usuário, for excedido.

5.3. Experimentos Numéricos

Para analisar e comparar o desempenho computacional dos métodos, usamos como testes os seguintes problemas:

Problema 1 (Martínez [1987])

$$
\begin{aligned}
f_1(x) &= (3-2x_1)x_1-2x_2+1,\\
f_i(x) &= (3-2x_i)x_i-x_{i-1}-2x_{i+1}+1,\ i=2,\ldots,n-1,\\
f_n(x) &= (3-2x_n)x_n-x_{n-1}+1,\\
f_{n+i}(x) &= f_i(x)+r,\ i=1,2,\ldots,n.
\end{aligned}
$$

Trabalhamos com n = 1000, r = 0.5, 1.0, 10.0.

Problema 2 (Martínez [1987])

$$
\begin{aligned}
f_1(x) &= (3-2x_1)x_1-2x_2+1+r,\\
f_i(x) &= (3-2x_i)x_i-x_{i-1}-2x_{i+1}+1+r,\ i=2,\ldots,n-1,\\
f_n(x) &= (3-2x_n)x_n-x_{n-1}+1+r,\\
f_{n+i}(x) &= f_i(x)+cx_{\alpha(i)},\ i=1,2,\ldots,n,
\end{aligned}
$$

onde $\alpha(i)$ é um inteiro aleatório entre 1 e r e $\alpha(i)\notin\{i-1,i,i+1\}$.

Trabalhamos com n = 1000, r = 10.0, c = 0.5, 1.0, 10.0.

Problema 3

$$
\begin{aligned}
f_1(x) &= (3-2x_1)x_1-2x_2+1,\\
f_i(x) &= (3-2x_i)x_i-x_{i-1}-2x_{i+1}+1,\ i=2,\ldots,n-1,\\
f_n(x) &= (3-2x_n)x_n-x_{n-1}+1,
\end{aligned}
$$

$$f_{n+i}(x) = f_i(x) + 0.5x_{\alpha(i)},$$
$$f_{2n+i}(x) = f_{n+i}(x) + 0.5$$
$\alpha(i)$ *como no problema* 2, $i = 1,\ldots,n$.

Trabalhamos com n = 1500.

Problema 4

$$f_{(j-1)n+1}(x) = (3-2x_1)x_1 - 2x_2 + 1 + cx_{\alpha(1,j)},$$
$$f_{(j-1)n+i}(x) = (3-2x_i)x_i - x_{i-1} - 2x_{i+1} + 1 + cx_{\alpha(i,j)},$$
$$f_{(j-1)n+n}(x) = (3-2x_n)x_n - x_{n-1} + 1 + cx_{\alpha(n,j)},$$
$$i = 2,\ldots,n-1$$

onde $\alpha(i,j)$ é um inteiro aleatório entre 1 e n, e $\alpha(i,j) \notin \{i-1,i,i+1\}$ para $j = 1,2,\ldots,s$.

Trabalhamos com diferentes valores de n, s e c.

Nos métodos "globais", usamos:

$$.\theta_1 = 10^{-4};$$
$$.\theta_2 = 10^{-4};$$
$$.\theta_3 = 0.1\ ;$$
$$.\underline{M} = 10^{-4};$$
$$.\overline{M} = 10^{4}.$$

Os resultados dos problemas 1 a 4 estão mostrados na tabela 5.1. Os pontos iniciais usados para estes testes foram:

$$I\ : x_0 = (-1,\ldots,-1);$$
$$II\ : x_0 = (-10,\ldots-10);$$
$$III: x_0 = (-10^2,\ldots,-10^2).$$

O desempenho dos vários algoritmos está assim representado:

-coluna ALG2.1 (relativa ao algoritmo 2.1)

(IT,TEMPO), onde:

IT = nº de iterações efetuadas;

TEMPO = tempo de CPU (em segundos);

-colunas ALG2.2.1 e ALG2.2.2 (relativas aos algoritmos 2.2.1 e 2.2.2 respectivamente)

(IER,IT,NF,TEMPO,P), onde:

IER = razão de parada;

-se IER = 1, foi declarada a convergência por (5.1) ou (5.2);

-se IER = 2, foi declarada a convergência por (5.3);

-se IER = 3, foi declarada a convergência por (5.4);

-se IER = 4, no algoritmo 2.2.2, a matriz de coeficientes do sistema 2x2, que deve ser resolvido para calcular v_k, não é definida positiva, e o programa é interrompido;

NF = número de avaliações da função f realizadas;

P indica a norma que está sendo usada para determinar a função objetivo f;

-coluna ALG2.4 (relativa ao algoritmo 2.4, com $\gamma = 0.9$ e $q = 3$)

(IER,ITOR,ITESP,NF,TEMPO), onde:

ITOR = número de iterações ordinárias efetuadas;

ITESP = número de iterações especiais efetuadas.

Para alguns testes (quando ITESP > 0) acrescentamos a norma P que foi usada.

PR	R	C	x^0	n	s	ALG2.1	ALG2.2.1	ALG2.2.2	ALG2.4
1	0.5	1.0	I	1000	2	(4,21.30)	(1,5,21,32.98,1.5)	(1,6,19,41.55,1.5)	(1,4,0,4,22.58)
			III			(10,52.15)	(1,13,35,84.96,1.5)	(4,16,17,98.97,1.5)	(1,10,0,10,62.89)
	1.0		I			(4,22.81)	(1,4,16,27.73,1.5)	(1,7,16,46.85,1.5))	(1,4,0,4,22.50)
			III			(10,56.61)	(1,13,38,77.65,1.5)	(4,17,18,103.19,1.5)	(1,10,0,10,63.11)
	10.0		I			(4,21.21)	(1,4,6,22.98,1.5)	(1,10,18,59.0,1.5)	(1,4,0,4,22.54)
			III			(9,52.43)	(1,11,27,73.43,1.5)	(4,18,19,123.48,1.5)	(1,9,0,9,57.64)
2	10.0	0.5	I	1000	2	(4,45.74)	(1,4,6,45.75,1.5)	(1,6,14,68.64,1.5)	(1,4,0,4,49.41)
			III			(9,114.86)	(1,10,21,119.07,1.5)	(4,18,19,264.1,1.5)	(1,9,0,9,118.18)
		1.0	I			(4,34.68)	(1,4,6,53.69,1.5)	(1,8,16,106.62,1.5)	(1,4,0,4,33.04)
			III			(9,117.32)	(1,15,61,160.95,1.5)	(4,16,17,243.48,1.5)	(1,9,0,9,119.74)
		10.0	I			(4,48.46)	(1,7,33,64.87,1.5)	(1,8,37,93.28,1.5)	(1,4,0,4,49.14)
			III			(6,126.90)	(1,11,30,105.1,1.5)	(4,16,17,241.55,1.5)	(1,6,0,6,126.31)
3			I	1500	3	(4,99.99)	(1,9,38,222.71,1.5)	(1,13,35,433.94,1.5)	(1,4,0,4,96.84)
			III			(11,258.45)	(1,15,36,422.65,1.5)	(1,178,193,6256.57,1.5)	(1,11,0,11,247.81)
4		0.5	I	1000	10	(5,227.29)	(1,6,14,372.28,1.5)	(1,51,115,3275.1,1.5)	(1,5,0,5,226.87)
					20	(4,427.59)	(1,6,17,731.4,1.5)	(1,11,35,1519.21,1.5)	(1,4,0,4,501.25)
		100.	I	500	2	excedeu 1000 iter.	(1,3,10,207.94,1.1)	(1,33,70,2300.13,1.1)	(1,4,22,36,853.46,1.1)
			II	300	5	(99,899.88)	(1,4,11,304.48,1.1)	(1,47,74,3037.36,1.1)	(1,5,22,31,593.18,1.1)
						(IT,TEMPO)	(IER,IT,NF,TEMPO,P)	(IER,IT,NF,TEMPO,P)	(IER,ITOR,ITESP,NF,T,(P))

TABELA 5.1

PR	R	C	x^0	n	s	ALG2.1	ALG2.4
1	0.5		IV	1000	2	(12,65.07)	(1,12,0,12,74.08)
			V			(14,78.09)	(1,14,0,14,87.01)
	1.0		IV			(12,66.93)	(1,12,0,12,71.33)
			V			(14,77.73)	(1,14,0,14,82.21)
	10.0		IV			(12,65.90)	(1,12,0,12,70.80)
			V			(13,74.06)	(1,13,0,13,78,13)
2	10.0	0.5	IV	1000	2	(12,158.75)	(1,12,0,12,163.88)
			V			(13,169.16)	(1,13,0,13,173.21)
		1.0	IV			(12,146.05)	(1,12,0,12,165.27)
			V			(13,160.30)	(1,13,0,13,171.45)
		10.0	IV			(11,183.44)	(1,11,0,11,192.59)
			V			(13,227.11)	(1,13,0,13,225.04)
4		0.5	I		2	(4,39.22)	(1,4,0,4,37.88)
					3	(4,61.37)	(1,4,0,4,64.55)
			II		3	(7,101.25)	(1,7,0,7,102.39)
		1.0	I		3	(5,73.20)	(3,4,0,4,60.68)
			II		10	(7,308.31)	(1,7,0,7,364.97)
						(IT,T)	(IER,ITOR,ITESP,NF,T)

TABELA 5.2

A tabela 5.2 mostra resultados obtidos para os problemas 1 e 2 a partir dos pontos iniciais:

$$IV: x_0 = (-10^3, \ldots, -10^3);$$
$$V: x_0 = (-10^4, \ldots, -10^4);$$

e outros resultados obtidos para o problema 4. No entanto, nesta tabela, é exibido somente o desempenho dos algoritmos 2.1 e 2.4.

5.4. Comentários e Conclusões

Em primeiro lugar, devemos observar que o tempo de CPU gasto na resolução dos vários problemas pode reduzir-se consideravelmente em uma implementação feita em computador com processadores paralelos, quando, então, as projeções ortogonais nos diversos blocos podem ser feitas simultâneamentes. No entanto vemos que a implementação particular que fizemos usando blocos quadrados apresenta resultados satisfatórios mesmo em um computador com arquitetura sequencial.

Observando os resultados obtidos, vemos que os algoritmos globais que realizam buscas em todas as iterações forneceram, em geral, os piores resultados em termos de número de iterações realizadas, em especial o algoritmo 2.2.2. Também verificamos que o algoritmo que emprega estratégia tolerante, na maior parte dos testes não se diferenciou do algoritmo local. As exceções a estes resultados se mostraram nas duas últimas linhas da tabela 5.1, as quais se referem a problemas cujo resíduo é grande. Para estes sistemas, os métodos com estratégias globalizadoras convergiram

muito mais rapidamente que o método local; os bons resultados apresentados pelo algoritmo 2.2.1 para estes testes devem ser ressaltados.

Os gráficos 5.1, 5.2. e 5.3 destacam alguns resultados colocados na tabela 5.1.

O gráfico 5.1 representa os resultados obtidos com o problema 2, em termos de número de iterações, a partir de $x^0 = (-1,...,-1)$. Nos tres diferentes testes gerados pelos diferentes valores atribuídos à constante c, todos os algoritmos atingiram a convergência, sendo que o algoritmo básico (ALG1) e o algoritmo com estratégia tolerante (ALG4) sempre efetuaram o menor número de iterações (ALG4 teve exatamente o mesmo comportamento que o algoritmo básico).

O gráfico 5.2 mostra os resultados obtidos com o sistema 4 com dimensões 10000 x 1000 e 20000 x 1000, e com a constante $c = 0.5$. Novamente ALG1 e ALG4 se mostraram os algoritmos mais eficientes, e o número de iterações realizadas pelo algoritmo 2.2.2 (ALG3) se tornou ainda maior nestes problemas de dimensões maiores.

No entanto, para o problema 4 com $c = 100$, o gráfico 5.3 destaca a ineficiência do algoritmo básico, mostrando que nestes testes, nos quais o resíduo do problema é maior, e as dimensões do sistema não são tão grandes, o algoritmo 2.2.1 (ALG2) é o que efetuou o menor número de iterações.

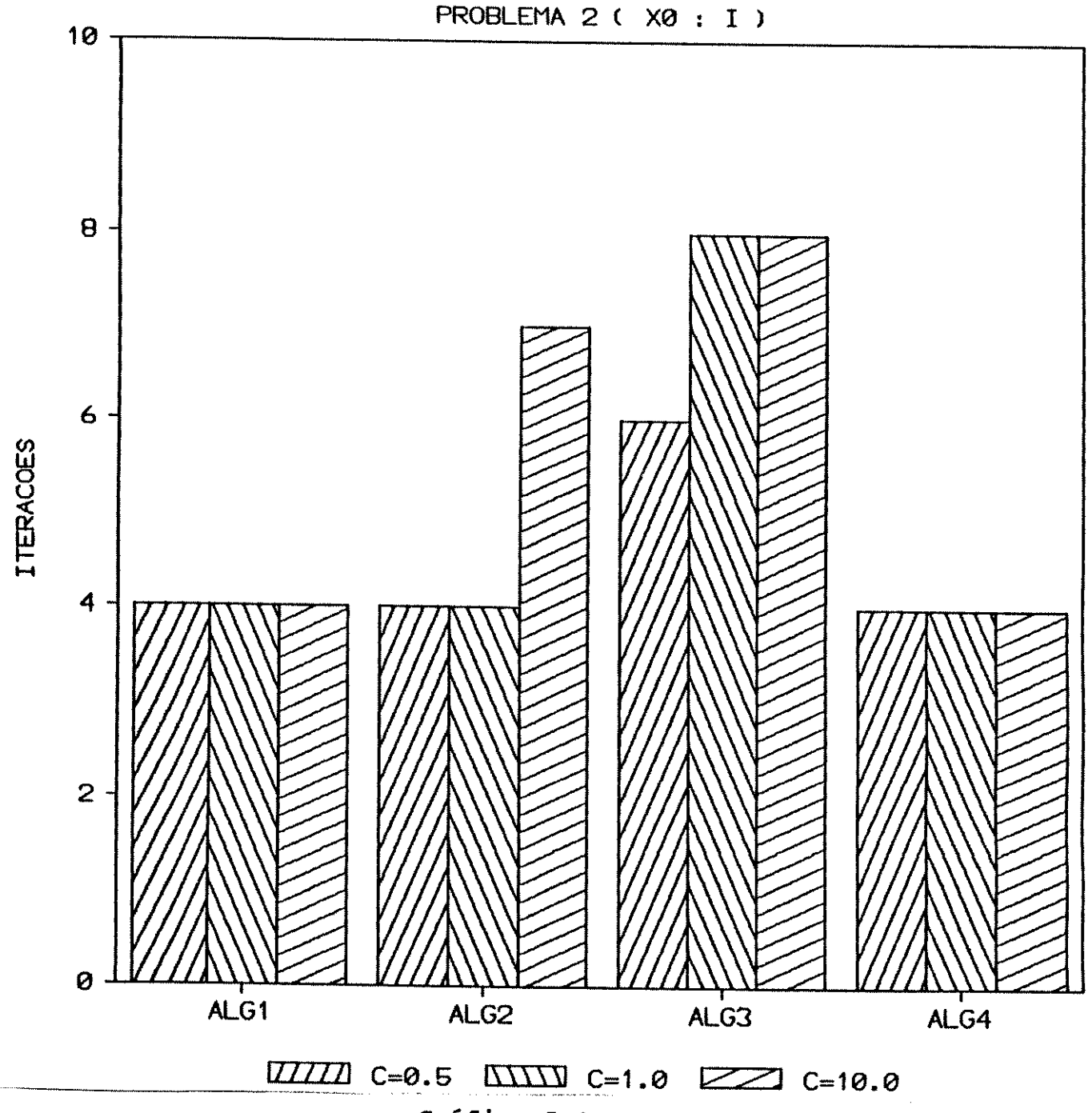

PROBLEMA 2 (X0 : I)
10
8
6
4
2
0
ITERACOES
ALG1
ALG2
ALG3
ALG4
C=0.5
C=1.0
C=10.0

Gráfico 5.1

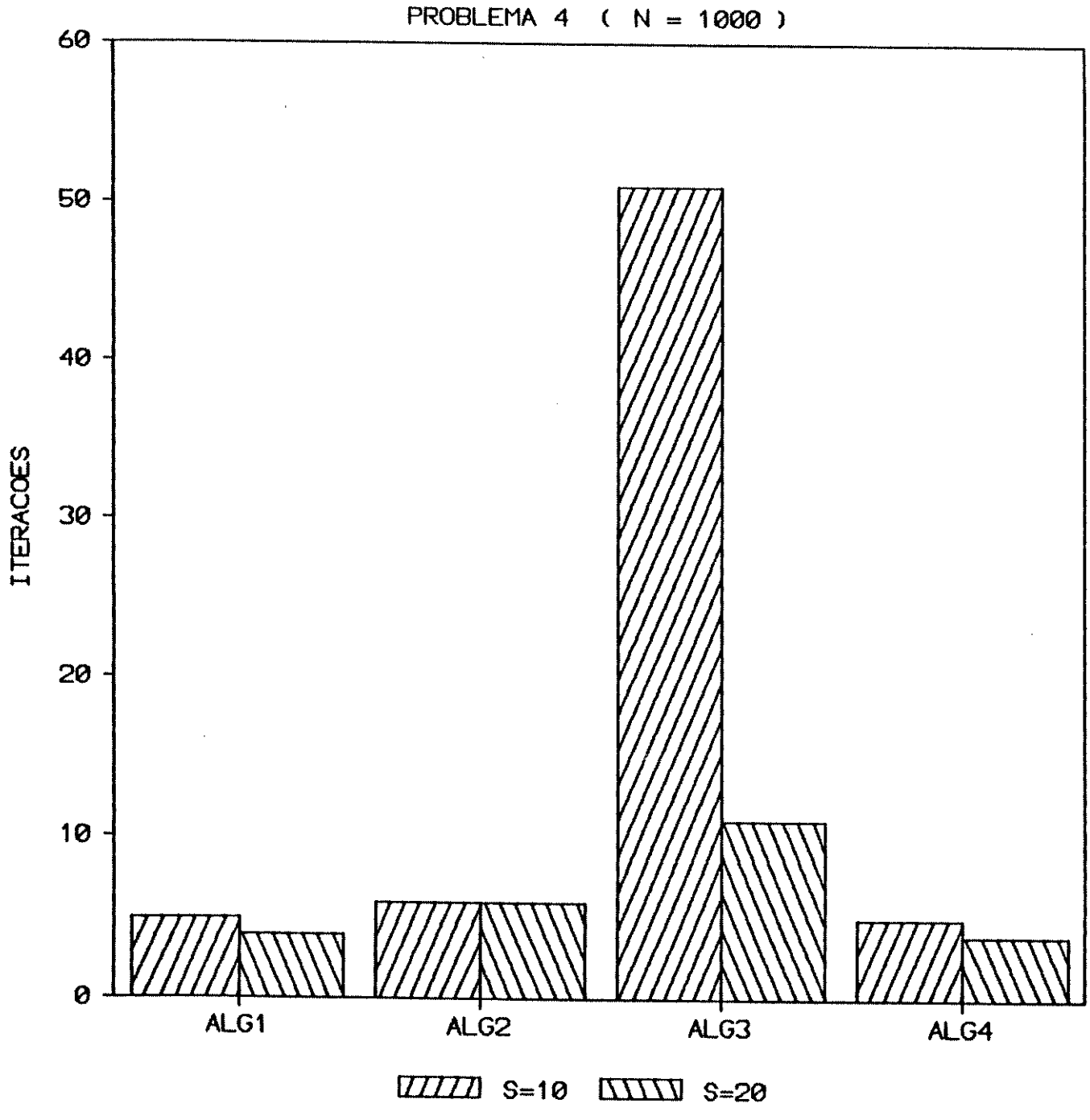

PROBLEMA 4 (N = 1000)
ITERACOES
60
50
40
30
20
10
0
ALG1
ALG2
ALG3
ALG4
S=10
S=20

Gráfico 5.2

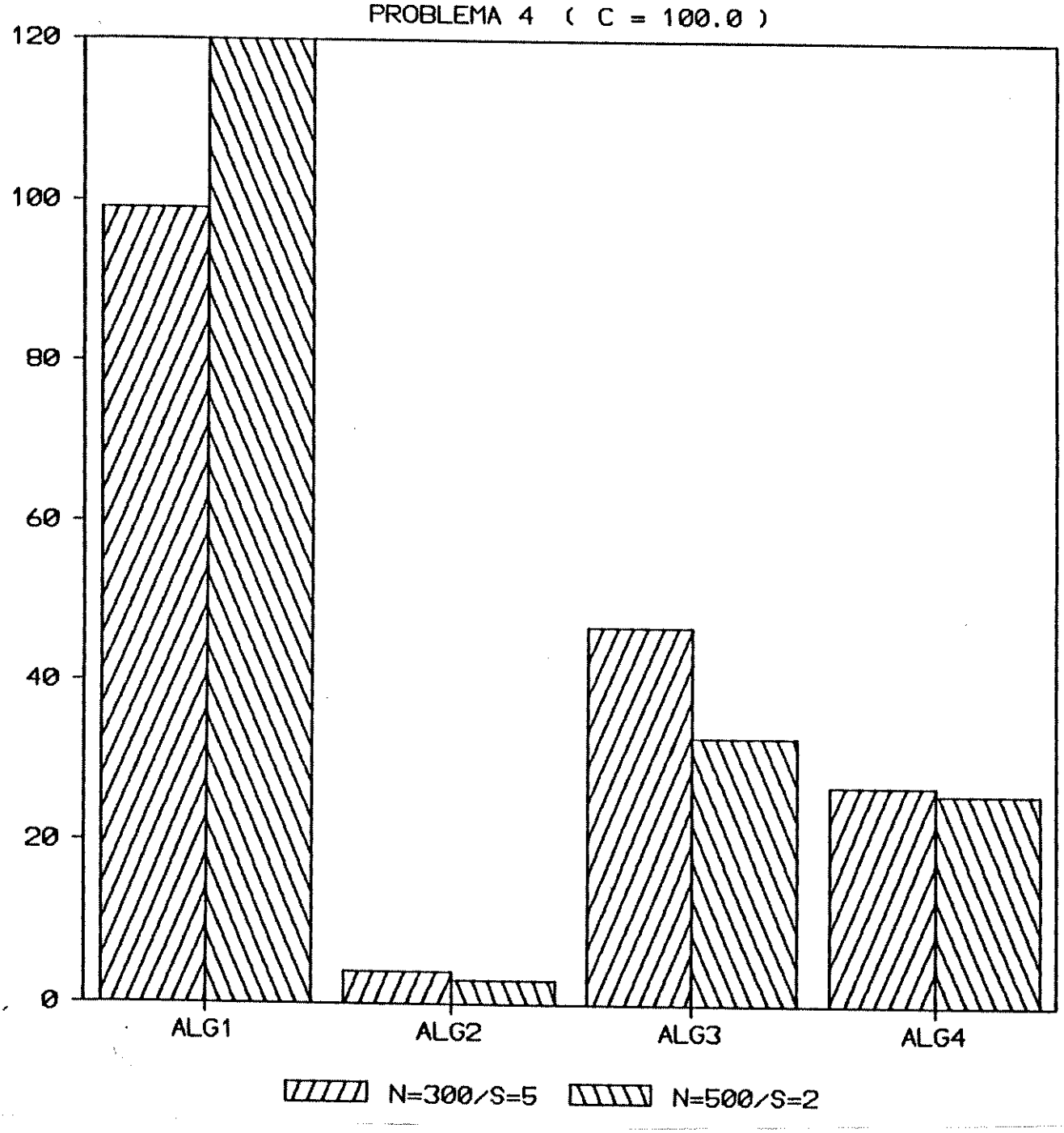

PROBLEMA 4 (C = 100.0)
120
100
80
60
40
20
0
ALG1
ALG2
ALG3
ALG4
N=300/S=5
N=500/S=2

Gráfico 5.3

Comparamos os resultados numéricos dos problemas 1 e 2 com os obtidos por Martínez [1987] em uma implementação esparsa do método de Gauss-Newton. Observamos que, nestes testes, sempre pelo menos um de nossos métodos compete em termos de número de iterações com o algoritmo de Gauss-Newton esparso. No entanto, este último gasta, em geral, menos tempo de CPU, o que pode ser justificado porque o problema não é resolvido precisamente: um algoritmo de gradientes conjugados pré-condicionado é usado em cada iteração para obter uma solução aproximada do problema de quadrados mínimos. Por outro lado, o esquema de fatoração usado em MA28 torna-se bastante caro em questão de tempo. Finalmente, para problemas nos quais $m >> n$, como o problema 4 com $s = 10$ e $s = 20$, o método de Gauss-Newton torna-se completamente ineficiente, enquanto os nosso algoritmos apresentaram resultados bastante satisfatórios.

Para problemas com n pequeno, podemos ainda fazer uma comparação quanto ao número de operações gastas pelo método de Gauss-Newton e pelo nosso algoritmo local: Gauss-Newton usa $O(mn^2 + n^3/6)$ flops por iteração e a implementação do algoritmo de projeções paralelas com blocos quadrados ($m = sn$) usa somente $O(sn^3/3)$ flops por iteração. Assim, para uma implementação paralela o tempo de CPU gasto seria o correspondente a $O(n^3/3)$ flops por iteração, excluindo os custos de comunicação.

Podemos concluir que o algoritmo local mostrou um bom desempenho na resolução dos diversos problemas, mesmo quando os pontos iniciais estavam mais distantes da melhor solução obtida x^*. O método apresentou dificuldades quando o resíduo do sistema é grande; neste caso, os algoritmos globais propostos, em

especial o algoritmo 2.2.1 e o algoritmo 2.4, podem representar boas alternativas para determinar uma solução aproximada do sistema (2.1).

5.5 Conclusões Finais e Trabalhos Futuros

Neste trabalho introduzimos, inicialmente, um método de projeções ortogonais para sistemas não lineares sobredeterminados. Demonstramos resultados de convergência local linear se o resíduo do sistema é pequeno e convergência local quadrática nos casos de resíduo zero. Introduzimos então tres outros algoritmos com estratégias globalizadoras, para os quais demonstramos a convergência global, como também local. Finalmente, um algoritmo com estratégia "tolerante" é proposto, também com resultados de convergência global.

Implementamos computacionalmente os algoritmos propostos, com exceção de um deles, e obtivemos resultados, de uma forma geral, satisfatõrios com os experimentos numéricos que realizamos.

Como trabalhos futuros, pretendemos calcular as projeções ortogonais nos vários blocos através de um método iterativo, implementando assim um método do tipo "Newton-Inexato". (Dembo, Eisenstat e Steihaug [1982]). Neste caso, o uso de pré-condicionadores é indispensável. (Golub e Van Loan [1989], Martínez [1990]).

Além disso, planejamos fazer uma implementação dos vários algoritmos usando processadores em paralelo, desenvolvendo também variações assíncronas destes.

APÊNDICE I

UMA APLICAÇÃO:

ESTIMATIVA DA CAMADA CHUMBO-ESTANHO NO RECOBRIMENTO DE CHAPAS

Um modelo para calcular a espessura da camada de chumbo-estanho, usada para recobrir certos materiais padrões resume-se em um problema de ajuste de dados. (Averick, Carter e Moré [1991])

Em uma certa superfície, toma-se n_0 pontos isolados e, em cada um destes pontos, mede-se:

- y_i = espessura da camada de chumbo-estanho no ponto i;
- y_{i+n_0} = quantidade relativa, em peso do estanho na liga;
- $(\xi_1,\xi_2)i$ = coordenadas do ponto i; $i = 1,2,\ldots,n_0$.

Todos estes valores estão sujeitos a erros.

As funções

$$z_1(\xi_1,\xi_2) = x_1 + x_2\xi_1 + x_3\xi_2 + x_4\xi_1\xi_2,$$
$$z_2(\xi_1,\xi_2) = x_5 + x_6\xi_1 + x_7\xi_2 + x_8\xi_1\xi_2,$$

modelam, respectivamente, a espessura da camada e a concentração relativa do estanho.

Para estimar os parâmetros $x_1, \ldots, x_8$ e pequenas perturbações $x_9, \ldots, x_{8+2n_0}$ para as coordenadas medidas $(\xi_1,\xi_2)i$ chega-se às equações

$$f_i(x) = z_1(\xi_{1,i}+x_{8+i}, \xi_{2,i}+x_{8+i+n_0}) - y_i, \quad 1 \le i \le n_0,$$

$$f_{i+n_0}(x) = z_2(\xi_{1,i}+x_{8+i}, \xi_{2,i}+x_{8+1+n_0}) - y_{i+n_0}, \quad 1 \le i \le n_0,$$

$$f_{i+2n_0}(x) = w_i\, x_{8+i}, \quad 1 \le i \le 2n_0,$$

onde w_i, $i = 1,2,\ldots,2n_0$, é um conjunto de pesos.

Os dados y_i, $i = 1,2,\ldots, n_0$, são da ordem de 10μ (10^{-2}mm) e y_{i+n_0}, $1 = 1,2,\ldots,n_0$, variam de 3 a 15%. (Ver V. Chiaverini [1977]).

Temos um sistema não linear com $m = 4n_0$ equações e $n = 8 + 2n_0$ variáveis. A tabela A.1.1 apresenta o desempenho dos nossos algoritmos na resolução deste sistema, com $n_0 = 63$, de forma que $n = 134$ e $m = 252$. Fizemos os experimentos com dois pontos iniciais diferentes:

$$.x_I^0 \text{ tal que } \|F(x_I^0)\|_{1.5} = 9.309604;$$

$$.x_{II}^0 \text{ tal que } \|F(x_{II}^0\|_{1.5} = 294.2647.$$

Os resultados estão representados por (IER, IT, NF, TEMPO), de acordo com a nomenclatura usada nas tabelas 5.1 e 5.2.

ALG	I	II
ALG2.1	excedeu 50 iter.	excedeu 50 iter.
ALG2.2.1	(1,2,15,2.11)	(1,20,123,19.38)
ALG2.2.2	(1,2,15,2.23)	(1,7,44,6.89)
ALG2.4	OVERFLOW	excedeu 50 iter.

Tabela A.1.1

Observamos que os algoritmos 2.1 e 2.4 não convergiram a partir dos pontos iniciais dados. Em um dos testes, depois de um certo número de iterações, a execução foi interrompida por "OVERFLOW". Nos demais testes com estes algoritmos foram excedidas 50 iterações sem que houvesse decréscimo satisfatório no valor de f(x). Concluimos que o sistema linear, que tem que ser resolvido a cada iteração, é mal condicionado, gerando direções "ruins". Os algoritmos que realizam buscas em todas as iterações "corrigem" estas direções, e levam à convergência. Com o ponto x_{II}^{0} devemos notar a superioridade do algoritmo 2.2.2 em termos de número de iterações e, consequentemente, de número de avaliações de função e tempo de CPU.

APÊNDICE II

PSEUDO-INVERSA

Teorema A.2.1. (Decomposição em Valor Singular). Seja $A \in \mathbb{R}^{mxn}$ e seja $k = \min\{m,n\}$. Existem matrizes ortogonais $U = [u^1,\ldots,u^m] \in \mathbb{R}^{mxm}$ e $V = [v^1,\ldots,v^n] \in \mathbb{R}^{nxn}$, tais que

$$U^T AV = D = diag(\sigma_1,\ldots,\sigma_k) \in \mathbb{R}^{mxn}$$

onde $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \geq 0$

Prova: Ver Golub-Van Loan [1989].

As quantidades $\sigma_1,\ldots,\sigma_k$ são chamadas os *valores singulares* de A. Se definirmos r por

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_k = 0,$$

então posto (A) = r.

Definição A.2.2. Seja $S \subseteq \mathbb{R}^n$ um subespaço. $P \in \mathbb{R}^{nxn}$ é a *projeção ortogonal* sobre S se $R(P) = S$, $P^2 = P$ e $P^T = P$. Desta definição concluimos que se $x \in \mathbb{R}^n$ então $Px \in S$ e $(I - P)x \in S^{\perp}$.

Se as colunas de $W = [w^1,\ldots,w^r]$ formam uma base ortonormal para um subespaço S, então $P = WW^T$ é a única projeção ortogonal sobre S.

Existem muitas projeções ortogonais importantes associadas com a Decomposição em Valor Singular. Vamos supor que $A = UDV^T \in \mathbb{R}^{mxn}$ e que posto(A) = r. Se particionamos as matrizes U e V por:

$$U = \underset{r}{[U_r} \quad \underset{m-r}{\bar{U}_r]} \quad e \quad V = \underset{r}{[V_r} \quad \underset{n-r}{\bar{V}_r]}, \tag{A.2.1}$$

então:

$V_r V_r^T$ = projeção sobre $R(A^T)$;

$\bar{V}_r \bar{V}_r^T$ = projeçao sobre $N(A)$;

$U_r U_r^T$ = projeção sobre $R(A)$;

$\bar{U}_r \bar{U}_r^T$ = projeção sobre $N(A^T)$.

Vamos agora relacionar a Decomposição em Valor Singular com o problema de *Quadrados Mínimos Lineares*

$$Min_{x \in \mathbb{R}^n} \|Ax-b\|_2, \tag{A.2.2}$$

onde $A \in \mathbb{R}^{mxn}$, $b \in \mathbb{R}^m$, $m > n$. Se x resolve o problema (A.2.2) então $A^T(Ax-b) = 0$.

Se $z \in \mathbb{R}^n$ é tal que x + z é também um minimizador local de (A.2.2), então podemos concluir que $z \in N(A)$.

Portanto, se A tem posto coluna completo, $N(A) = \{0\}$ e, assim, x é a única solução de (A.2.2), a qual vamos chamar de x_{QM}; x_{QM} resolve o sistema linear simétrico definido positivo

$$A^T A\, x_{QM} = A^T b, \tag{A.2.3}$$

ou seja,

$$x_{QM} = (A^T A)^{-1} A^T b. \qquad (A.2.4)$$

Chamamos as equações do sistema (A.2.3) de *equações normais*. Como

$$\nabla(\frac{1}{2}\|Ax-b\|^2) = A^T(Ax-b),$$

então vemos que resolver as equações normais é o mesmo que resolver a equação

$$\nabla(\frac{1}{2}\|Ax-b\|^2) = 0.$$

Vamos denotar

$$r_{QM} = b - A\,x_{QM}$$

o *resíduo mínimo* e

$$\rho_{QM} = \|Ax_{QM} - b\|.$$

Por outro lado, se posto(A) = r < n, o problema (A.2.2) tem infinitas soluções, pois, neste caso, N (A) $\neq$ {0} e, portanto, se x é minimizador e z $\in$ N (A), então x + z é também minimizador.

Tomemos então:

X = conjunto de todos os minimizadores

= {x $\in$ $\mathbb{R}^n$: ‖Ax - b‖ = min}.

O conjunto X é convexo, pois se x_1, x_2 $\in$ X e λ $\in$ [0, 1] então:

$$\|A(\lambda x_1+(1-\lambda)x_2)-b\| \leq \lambda\|Ax_1-b\|+(1-\lambda)\|Ax_2-b\|$$
$$= \min\|Ax-b\|,$$

ou seja, $x = \lambda x_1 + (1-\lambda)x_2 \in X$. Assim vemos que X tem um único elemento que possui norma -2 mínima, o qual chamamos novamente de x_{QM}.

Teorema A.2.3. Consideremos o problema (A.2.2) e a decomposição em valor singular $U^TAV = D$ de A. Se $U = [u^1,\ldots,u^m]$ e $V = [v^1,\ldots,v^n]$, então

$$x_{QM} = \sum_{i=1}^{r} \frac{(u^i)^T b}{\sigma_i} v^i \qquad (A.2.5)$$

e

$$\rho^2_{QM} = \|Ax_{QM} - b\|^2 = \sum_{i=r+1}^{m} [(u^i)^T b]^2.$$

Prova: Ver Golub-Van Loan [1989].

Se definirmos a matriz $A^+ \in \mathbb{R}^{nxm}$ por $A^+ = VD^+U^T$, onde

$$D^+ = diag(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_r},0,\ldots,0) \in \mathbb{R}^{nxm},$$

então $x_{QM} = A^+b$ e $\rho_{QM} = \|(I - AA^+)b\|$. Chamamos A^+ de *pseudo-inversa* de A. Se posto(A) = n, então $A^+ = (A^TA)^{-1}A^T$ e x_{QM} é dado por (A.4). Se m = n = posto(A), então $A^+ = A^{-1}$. A^+ é a única matriz $X \in \mathbb{R}^{nxm}$ que satisfaz as *condições de Moore-Penrose*:

(i) $AXA = A$ (iii) $(AX)^T = AX$

(ii) $XAX = X$ (iv) $(XA)^T = XA$.

A partir destas condições podemos ver que AA^+ e A^+A são projeções ortogonais sobre $R(A)$ e $R(A^T)$ respectivamente. De fato, temos que $AA^+ = U_rU_r^T$ e $A^+A = V_rV_r^T$, onde U_r e V_r são as partições de U e V definidas em (A.2.1).

Para finalizar este apêndice, colocamos um resultado que se refere ao efeito de perturbações na matriz pseudo-inversa.

Teorema A.2.4. Se A e $\delta A \in \mathbb{R}^{mxn}$, com posto(A) = posto(A+δA) = r, então:

$$\| (A+\delta A)^{+} - A^{+} \|_F \leq 2 \| \delta A \|_F \max\{ \| A^{+} \|_2^2 , \| A+\delta A)^{+} \|_2^2 \},$$

onde

$$\| A \|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$$

é a norma de Frobenius da matriz A.

Prova: Ver Wedin [1973].

REFERÊNCIAS

[1] Averick, B. M. ; Carter, R. G. ; Moré, J. J. [1991]: The Minpack-2 Test Problem Collection, Technical Memorandum nº 150, Argonne National Laboratory, Argonne, Illinois.

[2] Ben-Israel, A. [1966]: A Newton-Raphson method for the solution of systems of equations, J. Mathematical Analysis and Applications **15**, pp. 243-252.

[3] Bishop, T. N. et al [1985]: Tomographic determination of velocity and depth in laterallly varying media, Geophisics **50**, pp. 903-923.

[4] 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.

[5] Censor, Y. ; Gustafson, D. E. ; Lent, A. ; Ivy, H. [1979]: A new approach to the emission computerized tomography problem: simultaneous calculation of attenuation and activity coefficients, IEEE Trans. Nuclear Sci **NS-26**, pp. 2775-2779.

[6] Cimmino, G. [1938]: Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari, La Ricerca Scientifica Ser II, Anno **IV**, **1**, pp. 326-333.

[7] Dembo, R.S.; Eisenstat, S.C.; Steihaug, T. [1982]: Inexact Newton methods, SIAM J. Numer. Anal. **19**, pp. 400 - 408.

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

[9] De Pierro, A. R. ; Iusem, A. N. [1985]: A parallel projection method of finding a common point of a family of convex sets, Pesquisa Operacional **5**, pp. 1-20.

[10] Duff, I. S. [1977]: MA-28 - a set of FORTRAN subroutines for sparse unsymetric linear equations. AERE R8730, HMSO, Londres.

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

[12] George, A. ; Heath, M. T. [1980]: Solution of sparse linear least squares using Givens rotations, Linear Algebra and its Applications **34**, pp. 69-83.

[13] George, A. ; Heath, M. T. ; Ng, E. [1983]: A comparison of some methods for solving sparse least-square problems, SIAM J. Sci. Stat. Comput. **4**, pp.177-187.

[14] George, J. A. ; Liu, J. W. [1981]: *Computer Solution of Large Sparse Positive Definite Systems*, Prentice Hall, Englewood Cliffs, N.J.

[15] Golub, G. H. ; Van Loan, Ch. I. [1989]: *Matrix Computations*, The Johns Hopkins University Press, Baltimore e Londres.

[16] Gomes-Ruggiero, M. A. ; Martínez, J. M. ; Moretti, A. C. [1991]: Comparing Algorithms for solving sparse nonlinear systems of equations, SIAM J. Sci. Stat. Comput. (por aparecer)

[17] Heath, M. T. [1984]: Numerical methods for large sparse linear least squares problems, SIAM J. Sci. Comput. **5**, pp. 497-513.

[18] Levenberg, K. [1944]: A method for the solution of certain problems in least squares, Quarterly Journal of Applied Mathematics, **2**, pp. 164-168.

[19] Marquardt, D. [1963]: An algorithm for least squares estimation of nonlinear parameters, SIAM J. Applied Mathematics, **11**, pp. 431-441.

[20] Martínez, J. M. [1987]: An algorithm for solving sparse nonlinear least squares problems, Computing **39**, pp. 307-325.

[21] Martínez, J. M. [1990]: Local convergence theory of inexact Newton methods based on structured least change updates, Math. Comput. 5, nº 191.

[22] Martínez, J. M. [1991]; Quasi-Newton methods for solving underdetermined nonlinear simultaneous equations, Journal of Computational and Applied Mathematics **34**, pp. 171-190.

[23] Martínez, J. M. ; Sampaio, R. J. [1986]: Parallel and sequential Kaczmarz methods for solving underdetermined nonlinear equations, Journal of Computational and Applied Mathematics **15**, pp. 311-321.

[24] Martínez, J. M. ; Santos, R. F. [1990]: An algorithm for solving nonlinear least squares problems with a new curvilinear search, Computing **44**, pp. 83-90.

[25] Meyn, K. H. [1983]: Solution of underdetermined nonlinear equations by stationary iteration methods, Numerische Mathematik **42**, pp. 161-172.

[26] Moré, J. J. [1977]: The Levenberg-Marquardt algorithm: implementation and theory, em *Numerical Analysis, Dundee 1977,* Lecture Notes in Mathematics **630**, G. A. Watson ed., Springer-Verlag, Berlim, pp. 105-116.

[27] Sampaio, R.J.B. [1985]: Métodos do tipo Kaczmarz para solução de sistemas não lineares indeterminados, Tese de Mestrado, Departamento de Matemática Aplicada, UNICAMP, Campinas, Brasil.

[28] Vicente Chiaverini [1977]: *Aços e Ferros Fundidos: características gerais, tratamento térmico*, ABM, São Paulo

[29] Wedin, P. A. [1973]: Perturbation theory for pseudo-inverses, BIT **13**, pp. 217-232.