Universidade Federal de Santa Catarina
Curso de Pós-Graduação em Matemática e
Computação Científica

Métodos de Região de Confiança para Sistemas de Equações Não-Lineares com Restrições de Caixa

Mael Sachine
Orientador: Prof. Dr. Mario César Zambaldi

Florianópolis
Fevereiro de 2006

Universidade Federal de Santa Catarina
Curso de Pós-Graduação em Matemática e Computação Científica

Métodos de Região de Confiança para Sistemas de Equações Não-Lineares com Restrições de Caixa

Dissertação apresentada ao Curso de Pós-Graduação em Matemática e Computação Científica, do Centro de Ciências Físicas e Matemáticas da Universidade Federal de Santa Catarina, para a obtenção do grau de Mestre em Matemática, com Área de Concentração em Matemática Aplicada.

Mael Sachine
Florianópolis
Fevereiro de 2006

Métodos de Região de Confiança para Sistemas de Equações Não-Lineares com Restrições de Caixa

por

Mael Sachine

Esta Dissertação foi julgada para a obtenção do Título de "Mestre", Área de Concentração em Matemática Aplicada, e aprovada em sua forma final pelo Curso de Pós-Graduação em Matemática e Computação Científica.

Igor Mozolevski
Coordenador

Comissão Examinadora

Prof. Dr. Mario César Zambaldi (MTM-UFSC-Orientador)

Prof. Dr. Eng. Luciano Vitoria Barboza (UCPel - CEFET/RS)

Prof. Dr. Daniel Noberto Kozakevich (MTM-UFSC)

Prof. Dr. Juliano de Bem Francisco (MTM-UFSC)

Florianópolis, fevereiro de 2006.

À minha família.

Agradecimentos

Agradeço primeiramente às pessoas as quais dedico este trabalho, meus pais Marco Antônio e Joceline, pelo estímulo e apoio incondicional e pela sensatez com que sempre me ajudaram. Agradeço à minha irmã Fernanda pela amizade e alegrias, e à João por todo seu carinho e atenção.

Gostaria de agradecer também ao professor e amigo Mário César Zambaldi pelo constante incentivo, sempre indicando a direção a ser tomada nos momentos de maior dificuldade. Obrigada pela confiança. Meus agradecimentos à José Luiz Rosas Pinho; meu professor, tutor e amigo; por seu caráter e integridade, um exemplo para a vida toda.

Agradeço profundamente aos meus melhores amigos pela ajuda contínua e por todos os momentos de alegria. Agradeço a todas as pessoas que, direta ou indiretamente, contribuíram para a realização desta dissertação.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES - por ter financiado este projeto pelo período de um ano.

Sumário

Resumo

Neste trabalho consideramos o estudo e a implementação de dois métodos de região de confiança para resolução de sistemas não-lineares com restrições de caixa. Estes métodos, que geralmente são desenvolvidos para problemas de minimização, são adaptados para os sistemas não-lineares e para as restrições, em que o escalamento das variáveis desempenha um papel importante permitindo passos robustos. Testes numéricos são feitos para avaliar a metodologia.

Abstract

In this work we consider the study and implementation of two trust region methods for nonlinear systems with box constraints. These methods, that are generally developed to minimization problems, are adapted to the nonlinear systems and to the constraints, which the scaling plays an important role allowing robust steps. Numerical tests are performed to evaluate the methodology.

Introdução

Os sistemas de equações algébricas não-lineares surgem em muitas aplicações das ciências naturais e aplicadas. Os estudos teóricos e computacionais de métodos numéricos para resolução de tais sistemas constituem uma área de pesquisa intensa nas últimas décadas, nas quais o uso de modernas técnicas computacionais tem se tornado uma ferramenta poderosa para a ciência e tecnologia.

No contexto da otimização contínua, os métodos para sistemas não-lineares normalmente empregados buscam obter uma seqüência convergente para um determinado algoritmo partindo de um ponto inicial arbitrário. Duas estratégias que se destacam fazem parte da metodogia clássica: buscas unidirecionais e região de confiança. Métodos que incorporam região de confiança são considerados mais robustos por considerarem um modelo local mais preciso.

Estratégias de região de confiança, normalmente baseadas em minimização de funcionais, devem ser adaptadas para sistemas não-lineares. Outra adaptação que deve ser considerada ocorre quando o problema apresenta restrições. Se as restrições que aparecem na formulação do problema são de certa forma específicas é possível desenvolver uma metodologia própria e eficiente para abordar o problema.

Este trabalho consiste no estudo, desenvolvimento e implementação de métodos de região de confiança para sistemas não-lineares com restrições de caixa. A metodologia é avaliada em termos de testes no ambiente MATLAB considerando problemas padrões e um problema da engenharia elétrica.

Em termos matemáticos o problema consiste em obter um vetor $x \in \mathbb{R}^n$ que satisfaça

$$F(x) = 0;\ x \in \Omega$$

em que $\Omega = \{x \in \mathbb{R}^n; l \leq x \leq u\}$, $F : \mathbb{R}^n \to \mathbb{R}^n$, e o faremos minimizando o quadrado da norma euclidiana de F, sujeito às restrições de caixa, ou seja, resolvendo o problema de otimização

$$\underset{x \in \Omega}{\text{minimizar}} \;\; f(x)$$

em que $f(x) = \frac{1}{2} \|F(x)\|^2$, usando a estrutura do problema original sempre que possível.

Fixando a notação usada no trabalho, o subscrito k é usado como índice para uma seqüência e quando o contexto deixa claro, o argumento da aplicação é omitido. Desta forma, para qualquer aplicação F, a notação F_k é usada para denotar $F(x_k)$ e a i-ésima componente de x_k é denotada por x_{k_i}. A norma euclidiana de $x \in \mathbb{R}^n$ é denotada por $\|x\|$ e a matriz Jacobiana de F em x é denotada por $F'(x)$. Para qualquer vetor $y \in \mathbb{R}^n$, a bola aberta com centro y e raio ρ é indicada por $B_\rho(y)$, ou seja, $B_\rho(y) = \{x; \|x - y\| \leq \rho\}$. Para uma caixa $\Omega = \{x \in \mathbb{R}^n; l \leq x \leq u\}$, o interior estrito de Ω é denotado por $int(\Omega)$.

Capítulo 1

Métodos de Região de Confiança

1.1 Introdução

Na abordagem clássica de região de confiança, minimizamos uma função objetivo sem restrições. A formulação matemática é a seguinte:

$$\underset{x\in\mathbb{R}^n}{\text{minimizar}} \;\; f(x)$$

em que $f : \mathbb{R}^n \to \mathbb{R}$ é uma função continuamente diferenciável.

Os métodos de região de confiança geram uma seqüência $\{x_k\}_{k\in\mathbb{N}}$ com a ajuda do modelo quadrático da função objetivo. Eles definem uma região ao redor da estimativa corrente de modo que dentro desta região podemos confiar que o modelo seja uma representação adequada da função objetivo e então minimiza-se o modelo nesta região. Se o minimizador obtido é aceitável, aumentamos a região de confiança ou a deixamos inalterada e prosseguimos com a minimização a partir da nova estimativa. Se a solução obtida não é aceitável reduzimos o tamanho da região e encontramos um novo minimizador. Em geral, a direção e o passo mudam sempre que o tamanho da região é alterado.

O tamanho da região de confiança é fundamental para a eficácia de cada passo. Se a região é muito pequena, o algoritmo perde a oportunidade de tomar um passo substancial que irá chegar muito mais perto do minimizador da função objetivo. Se é muito grande, o minimizador do modelo pode estar muito longe do minimizador da função objetivo na região, daí temos que reduzir o tamanho da região de confiança e tentar novamente. Na prática, escolhemos o tamanho da região de acordo com o desempenho do algoritmo nas iterações anteriores. Se o modelo é geralmente seguro, produzindo bons passos e predizendo com exatidão o comportamento da função objetivo ao longo desses passos, o

tamanho da região de confiança é constantemente aumentado para permitir passos mais longos e audaciosos. Por outro lado, um passo fracassado indica que o modelo não é uma representação adequada da função objetivo dentro da região de confiança corrente, e então reduzimos o tamanho da região e tentamos novamente.

O modelo quadrático m_k a cada iteração é dado por:

$$m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \tag{1.1}$$

em que $f_k = f(x_k)$, $\nabla f_k = \nabla f(x_k)$ e B_k é uma matriz simétrica. Como

$$f(x_k + p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T \nabla^2 f(x_k + tp) p$$

para $t \in (0, 1)$, e como $m_k(p) = f_k + \nabla f_k^T p + O(\|p\|^2)$, a diferença entre $m_k(p)$ e $f(x_k + p)$ é $O(\|p\|^2)$, e então o erro da aproximação é pequeno quando p é pequeno.

Quando B_k é a Hessiana $\nabla^2 f(x_k)$, o modelo concorda com os três primeiros termos da expansão em série de Taylor de f. O algoritmo obtido fazendo $B_k = \nabla^2 f(x_k)$ é chamado método de Newton para região de confiança. Na prática, a abordagem de região de confiança é muito geral pois precisamos assumir muito pouco de B_k, apenas simetria e limitação uniforme no índice k.

Para obter cada passo, procuramos a solução do subproblema

$$\underset{p \in \mathbb{R}^n}{\text{minimizar}} \;\; m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \qquad \text{s.a. } \|p\| \leq \Delta_k. \tag{1.2}$$

em que $\Delta_k > 0$ é o raio da região de confiança. A solução p_k^* de (1.2) é o minimizador de m_k na bola de raio Δ_k. Portanto, precisamos resolver uma seqüência de subproblemas (1.2) nos quais a função objetivo e a restrição $p^T p \leq \Delta_k^2$ são ambas quadráticas. Quando B_k é positiva definida e $\|B_k^{-1} \nabla f_k\| \leq \Delta_k$, a solução de (1.2) é simplesmente o minimizador irrestrito $p_k^N = -B_k^{-1} \nabla f_k$ do modelo quadrático $m_k(p)$. Neste caso chamamos p_k^N de *passo inteiro*. A solução de (1.2) não é tão óbvia em outros casos, mas pode ser encontrada sem muito esforço. De qualquer maneira, precisamos apenas de uma solução aproximada para obter convergência.

1.2 Algoritmo Padrão

A primeira questão que surge quando vamos definir um método de região de confiança é a estratégia para escolher um raio Δ_k a cada iteração. Baseamos esta escolha na concordância entre o modelo m_k e a função objetivo f nas iterações anteriores. Dada uma direção p_k definimos a razão

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}; \tag{1.3}$$

em que o denominador é chamado de *redução prevista* pelo modelo, ou seja,

$$\text{pred} = m_k(0) - m_k(p_k),$$

e o numerador é chamado de *redução real* da função f, isto é,

$$\text{ared} = f(x_k) - f(x_k + p_k).$$

Note que, como a direção p_k é obtida minimizando o modelo m_k em uma região que inclui a direção $p = 0$, a redução prevista será sempre não negativa. Desta forma, se ρ_k é negativo, o novo valor $f(x_k + p_k)$ é maior do que o valor corrente $f(x_k)$, e então o passo deve ser rejeitado.

Por outro lado, se ρ_k é próximo de 1, existe uma boa concordância entre o modelo m_k e a função f neste passo, e daí é seguro expandir a região de confiança na próxima iteração. Se ρ_k é positivo, mas não próximo de 1, não alteramos a região de confiança, mas se ρ_k é próximo de zero ou negativo, diminuimos a região de confiança.

O seguinte algoritmo descreve o processo.

Algoritmo 1: Método de Região de Confiança

Entrada: x_0, raio máximo $\bar{\Delta} > 0$, raio inicial $\Delta_0 \in (0, \bar{\Delta})$, $\eta \in \left[0, \frac{1}{4}\right)$, ε

Saída: minimizador x^*

$k \longleftarrow 0$
Calcular $\nabla f(x_k)$ e B_k
enquanto $\|\nabla f(x_k)\| > \varepsilon$ **faça**
    $p_k \longleftarrow$ solução de
    $$\underset{p \in \mathbb{R}^n}{\text{minimizar}} \;\; m_k(p) = f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T B_k p \quad \text{s.a.} \; \|p\| \leq \Delta_k$$
    $\rho_k \longleftarrow \dfrac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}$
    se $\rho_k < \frac{1}{4}$ **então**
        $\Delta_{k+1} \longleftarrow \frac{1}{4} \|p_k\|$
    senão
        se $\rho_k > \frac{3}{4}$ *e* $\|p_k\| = \Delta_k$ **então**
            $\Delta_{k+1} \longleftarrow \min(2\Delta_k, \bar{\Delta})$
        senão
            $\Delta_{k+1} \longleftarrow \Delta_k$
        fim
    fim
    se $\rho_k > \eta$ **então**
        $x_{k+1} \longleftarrow x_k + p_k$
        Calcular $\nabla f(x_{k+1})$ e B_{k+1}
    senão
        $x_{k+1} \longleftarrow x_k$
    fim
    $k \longleftarrow k + 1$
fim
$x^* \longleftarrow x_k$

Note que $\bar{\Delta}$ é uma cota global para os passos. Ainda, o raio é aumentado somente se $\|p_k\|$ alcança a fronteira da região de confiança. Se o passo fica estritamente dentro da região de confiança, concluímos que o valor corrente de Δ_k não está interferindo no progresso do algoritmo e deixamos seu valor igual para a próxima iteração. Note ainda que a escolha $\eta \in$

$\left[0, \frac{1}{4}\right)$ é feita por questões práticas [7]; tomando $\eta \in [0, 1)$ as propriedades de convergência não são alteradas.

Para tornar o algoritmo prático, precisamos resolver (1.2). Vamos descrever duas estratégias para encontrar soluções aproximadas que atingem no mínimo tanta redução no modelo m_k quanto a redução alcançada pelo ponto de Cauchy. Este ponto é simplesmente o minimizador de m_k ao longo da direção de máxima descida $-\nabla f_k$, sujeito à fronteira de região de confiança. A primeira estratégia é o *método dogleg*, apropriado quando a matriz B_k é positiva definida. A segunda estratégia, o método CG de Steihaug, é mais apropriado quando B_k é a matriz Hessiana $\nabla^2 f_k$ e quando esta matriz é grande e esparsa.

1.3 O Ponto de Cauchy e Algoritmos Relacionados

Nos métodos de região de confiança, apesar de estarmos procurando a solução ótima do subproblema (1.2), é suficiente para se ter convergência global encontrar uma solução aproximada p_k que esteja dentro da região de confiança e forneça uma *redução suficiente* no modelo. Esta redução pode ser quantificada em termos do ponto de Cauchy, que denotaremos por p_k^C e encontraremos na seqüência.

Seja p_k^S o vetor que resolve uma versão linear de (1.2), isto é,

$$p_k^S = \min_{p \in \mathbb{R}^n} f_k + \nabla f_k^T p \qquad \text{s.a. } \|p\| \leq \Delta_k \tag{1.4}$$

e seja $\tau > 0$ o escalar que minimiza $m_k(\tau p_k^S)$ satisfazendo a fronteira de região de confiança, ou seja,

$$\tau_k = \min_{\tau > 0} m_k(\tau p_k^S) \qquad \text{s.a. } \left\|\tau p_k^S\right\| \leq \Delta_k. \tag{1.5}$$

Logo, $p_k^C = \tau_k p_k^S$. De fato, podemos facilmente escrever uma fórmula fechada para o ponto de Cauchy [7]. A solução de (1.4) é simplesmente

$$p_k^S = -\frac{\Delta_k}{\|\nabla f_k\|} \nabla f_k.$$

Para obter τ_k explicitamente, vamos considerar os casos $\nabla f_k^T B_k \nabla f_k \leq 0$ e

$\nabla f_k^T B_k \nabla f_k > 0$ separadamente. Para o primeiro caso, a função $m_k(\tau p_k^S)$ decresce monotonamente conforme τ cresce, sempre que $\nabla f_k \neq 0$, daí τ_k é simplesmente o maior valor que satisfaz a região de confiança, ou seja, $\tau_k = 1$. Para o caso $\nabla f_k^T B_k \nabla f_k > 0$, $m_k(\tau p_k^S)$ é uma quadrática convexa em τ, então τ_k é ou o minimizador sem restrição da quadrática $\|\nabla f_k\|^3/(\Delta_k \nabla f_k^T B_k \nabla f_k)$, ou o valor de fronteira 1, qual deles vier primeiro. Resumindo, temos

$$p_k^C = -\tau_k \frac{\Delta_k}{\|\nabla f_k\|} \nabla f_k,$$

em que

$$\tau_k = \begin{cases} 1 & \text{se} \quad \nabla f_k^T B_k \nabla f_k \leq 0; \\ \min(\|\nabla f_k\|^3/(\Delta_k \nabla f_k^T B_k \nabla f_k), 1) & \text{caso contrário.} \end{cases} \tag{1.6}$$

O passo de Cauchy p_k^C é barato para calcular, nenhuma fatoração de matriz é necessária e é extremamente importante para decidir se uma solução aproximada do subproblema de região de confiança é aceitável. Especificamente, um método de região de confiança será globalmente convergente se seus passos obtém uma redução suficiente no modelo m_k, isto é, se a redução no modelo m_k é pelo menos algum múltiplo fixo do decréscimo obtido pelo passo de Cauchy a cada iteração.

Note que se sempre tomarmos o ponto de Cauchy como nosso passo, estaremos simplesmente implementando o método do Gradiente com uma escolha particular do tamanho do passo. Mas este método tem um desempenho pobre mesmo se os passos ótimos forem usados a cada iteração. Por esta razão devemos aperfeiçoar o ponto de Cauchy para obter uma solução aproximada melhor do subproblema.

O ponto de Cauchy não depende fortemente da matriz B_k, que é usada apenas no cálculo do passo. Convergência rápida pode ser esperada apenas se B_k têm a função de determinar tanto a direção do passo assim como seu tamanho.

Vários algoritmos que geram soluções aproximadas p_k para o subproblema (1.2) começam calculando o ponto de Cauchy e então tentam melhorá-lo. A estratégia de aperfeiçoamento é frequentemente projetada para que o passo inteiro $p_k^N = -B_k^{-1} \nabla f_k$ seja escolhido sempre que B_k é positiva definida e $\left\|p_k^N\right\| \leq \Delta_k$.

Vamos considerar dois métodos que apresentam as características acima citadas.

1.3.1 O Método Dogleg

Sabemos que a solução do subproblema (1.2) depende do raio da região de confiança Δ. Quando B_k é positiva definida, o minimizador irrestrito do modelo m_k é o passo inteiro $p^N = -B_k^{-1}\nabla f_k$. Se este ponto é viável, ele será uma solução. Assim, a solução $p^*(\Delta)$ é dada por

$$p^*(\Delta) = p^N, \quad \text{quando} \quad \left\|p^N\right\| \leq \Delta.$$

Quando Δ é pequeno, a restrição $\|p_k\| \leq \Delta$ assegura que o termo quadrático no modelo m_k tem pouco efeito na solução de (1.2). A solução verdadeira $p(\Delta)$ é aproximadamente a mesma solução que obteríamos minimizando a função linear $f_k + \nabla f_k^T p_k$ sobre $\|p_k\| \leq \Delta$, isto é,

$$p^*(\Delta) \approx -\Delta \frac{\nabla f_k}{\|\nabla f_k\|}, \quad \text{quando} \quad \Delta \text{ é pequeno.}$$

Para valores intermediários de Δ, a solução $p^*(\Delta)$ segue a trajetória curva como na figura (1.1).

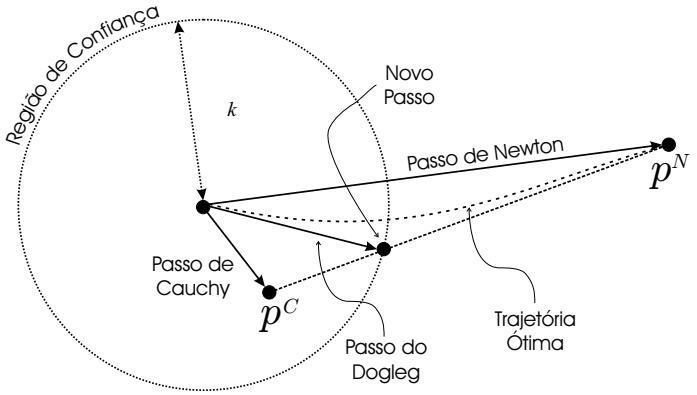

Figura 1.1: O passo do método dogleg.

O método dogleg encontra uma solução aproximada substituindo a trajetória curva por um caminho constituído por dois segmentos. O primeiro segmento vai da origem ao minimizador irrestrito sobre a direção de máxima descida definido por

$$p^C = -\frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T B_k \nabla f_k} \nabla f_k$$

enquanto o segundo segmento vai de p^C até p^N. Formalmente, denotamos esta trajetória por $\tilde{p}(\tau)$ para $\tau \in [0, 2]$ em que

$$\tilde{p}(\tau) = \begin{cases} \tau p^C, & 0 \leq \tau \leq 1, \\ p^C + (\tau - 1)(p^N - p^C), & 1 \leq \tau \leq 2. \end{cases} \tag{1.7}$$

O método dogleg escolhe p que minimiza o modelo m ao longo do caminho, sujeito à região de confiança. De fato, não é necessário nem realizar uma busca, pois o caminho dogleg intersecta a fronteira da região de confiança no máximo uma vez e a intersecção pode ser calculada analiticamente. Vamos provar estas afirmações no seguinte Lema.

Lema 1.1. *Seja B definida positiva, então:*

(i) $\|\tilde{p}(\tau)\|$ *é uma função crescente de* τ*, e*

(ii) $m(\tilde{p}(\tau))$ *é uma função decrescente de* τ*.*

Prova: É fácil mostrar que (i) e (ii) valem para $\tau \in [0, 1]$, assim voltamos nossa atenção para o caso de $\tau \in [1, 2]$.

Para (i), defina $h(\alpha)$ por:

$$\begin{aligned} h(\alpha) &= \frac{1}{2} \|\tilde{p}(1 + \alpha)\|^2 \\ &= \frac{1}{2} \left\|p^C + \alpha(p^N - p^C)\right\|^2 \\ &= \frac{1}{2} \left\|p^C\right\|^2 + \alpha p^{C^T}(p^N - p^C) + \frac{1}{2}\alpha^2 \left\|p^N - p^C\right\|^2. \end{aligned}$$

O lema é provado se mostrarmos que $h'(\alpha) \geq 0$ para $\alpha \in (0, 1)$. Agora, tomando $g = \nabla f(x_k)$ temos

$$\begin{aligned} h'(\alpha) &= -p^{C^T}(p^C - p^N) + \alpha \left\| p^C - p^N \right\|^2 \\ &\geq -p^{C^T}(p^C - p^N) \\ &= \frac{g^T g}{g^T B g} g^T \left(-\frac{g^T g}{g^T B g} g + B^{-1} g \right) \\ &= g^T g \frac{g^T B^{-1} g}{g^T B g} \left(1 - \frac{(g^T g)^2}{(g^T B g)(g^T B^{-1} g)} \right) \\ &\geq 0 \end{aligned}$$

em que a última desigualdade vem do lema 1.2.

Para (ii), definimos $\hat{h}(\alpha) = m(\tilde{p}(1+\alpha))$, isto é,

$$\hat{h}(\alpha) = f + g^T(p^C + \alpha(p^N - p^C)) + \frac{1}{2}(p^C + \alpha(p^N - p^C))^T B(p^C + \alpha(p^N - p^C))$$

e mostramos que $\hat{h}'(\alpha) \leq 0$ para $\alpha \in (0, 1)$.

$$\begin{aligned} \hat{h}'(\alpha) &= (p^N - p^C)^T(g + Bp^C) + \alpha(p^N - p^C)^T B(p^N - p^C) \\ &\leq (p^N - p^C)^T(g + Bp^C + B(p^N - p^C)) \\ &= (p^N - p^C)^T(g + Bp^N) = 0 \end{aligned}$$

concluindo a prova.

■

Lema 1.2. *Seja $g \in \mathbb{R}^n$ e seja $B \in \mathbb{R}^{n \times n}$ uma matriz definida positiva. Então*

$$g^T g \frac{g^T B^{-1} g}{g^T B g} \left(1 - \frac{(g^T g)^2}{(g^T B g)(g^T B^{-1} g)} \right) \geq 0.$$

Prova: Sabemos que $g^T g = \|g\|^2 \geq 0, \forall g \in \mathbb{R}^n$. Como B é definida positiva, temos que $g^T B^{-1} g$ e $g^T B g$ são valores positivos. Sejam $u = B^{1/2} g$ e $v = B^{-1/2} g$, assim

$$
\begin{aligned}
\left\|(B^{1/2}g)^T B^{-1/2}g\right\|^2 &\leq [(B^{1/2}g)^T(B^{1/2}g)][(B^{-1/2}g)^T(B^{-1/2}g)] \\
\left\|g^T B^{1/2}B^{-1/2}g\right\|^2 &\leq (g^T Bg)(g^T B^{-1}g) \\
(g^T g)^2 &\leq (g^T Bg)(g^T B^{-1}g) \\
\frac{(g^T g)^2}{(g^T Bg)(g^T B^{-1}g)} &\leq 1 \\
1 - \frac{(g^T g)^2}{(g^T Bg)(g^T B^{-1}g)} &\geq 0
\end{aligned}
$$

Portanto,

$$
g^T g \frac{g^T B^{-1}g}{g^T Bg}\left(1 - \frac{(g^T g)^2}{(g^T Bg)(g^T B^{-1}g)}\right) \geq 0,
$$

concluindo a demonstração do lema.

■

Segue do Lema 1.1 que o caminho $\tilde{p}(\tau)$ intersecta a fronteira da região de confiança $\|p\| = \Delta$ em exatamente um ponto se $\left\|p^N\right\| \geq \Delta$, e em nenhum ponto caso contrário. Visto que m é decrescente ao longo do caminho, o valor escolhido de p será em p^N se $\left\|p^N\right\| \leq \Delta$.

Se $\left\|p^N\right\| > \Delta$, escolhe-se o ponto de intersecção do dogleg com a fronteira da região de confiança. Para este caso, calcula-se o valor de τ resolvendo a equação:

$$
\left\|p^C + \lambda d\right\|^2 = \Delta^2
$$

em que $\lambda = \tau - 1$ e $d = p^N - p^C$.

Observe que resolver a equação acima é o mesmo que resolver

$$
\left\|p^C\right\|^2 + 2\lambda {p^C}^T d + \lambda^2 \|d\|^2 = \Delta^2
$$

ou ainda,

$$
\left(\|d\|^2\right)\lambda^2 + \left(2{p^C}^T d\right)\lambda + \left(\left\|p^C\right\|^2 - \Delta^2\right) = 0.
$$

As soluções para esta equação de segundo grau são dadas por

$$
\lambda = \frac{-2{p^C}^T d \pm \sqrt{\left(2{p^C}^T d\right)^2 - 4\|d\|^2\left(\|p^C\|^2 - \Delta^2\right)}}{2\|d\|^2}.
$$

Como $\left\|p^C\right\| < \Delta$, é fácil ver que há sempre uma raiz positiva e uma negativa. A solução positiva é a que corresponde com a intersecção do dogleg com o limite da região de confiança. Assim, a intersecção do dogleg com o limite da região de confiança acontece quando

$$\tau = \frac{-2p^{C^T}(p^N - p^C) + \sqrt{\left(2p^{C^T}(p^N - p^C)\right)^2 - 4\left\|p^N - p^C\right\|^2\left(\|p^C\|^2 - \Delta^2\right)}}{2\left\|p^N - p^C\right\|^2} + 1.$$

1.3.2 O Método CG de Steihaug

O método dogleg descrito anteriormente requer a solução de um sistema linear envolvendo a matriz B a cada iteração. Quando B é grande, esta operação pode ter um alto custo computacional, de forma que devemos considerar outras técnicas para encontrar uma solução aproximada de (1.2) que não exija a solução exata de um sistema linear mas ainda assim produza uma melhora em relação ao ponto de Cauchy. O método CG de Steihaug é uma técnica com estas propriedades e é baseada no algoritmo gradiente conjugado, um algoritmo iterativo para resolver sistemas lineares com matrizes simétricas positivas definidas. O algoritmo gradiente conjugado (CG) está descrito no Apêndice deste trabalho, neste momento vamos apenas comentar as diferenças entre o CG e o método de Steihaug; que são essencialmente que o algoritmo de Steihaug termina ou quando encontra uma direção tal que $\|p\| \leq \Delta$ ou quando encontra uma direção de curvatura negativa em B. O método CG de Steihaug pode ser declarado formalmente como segue:

Algoritmo 2: CG de Steihaug

Entrada: $\varepsilon > 0$

Saída: direção p

Faça $p_0 = 0$, $r_0 = g$, $d_0 = -r_0$
$k \longleftarrow 0$
enquanto $\|r_k\| \geq \varepsilon$ **faça**
    se $d_k^T B d_k \leq 0$ **então**
        Encontre τ tal que $p = p_k + \tau d_k$ minimiza $m(p)$ e satizfaz $\|p\| = \Delta$;
        Retorna p
    senão
        $\alpha_k = r_k^T r_k / d_k^T B d_k$
        $p_{k+1} = p_k + \alpha_k d_k$
    fim
    se $\|p_{k+1}\| \geq \Delta$ **então**
        Ache $\tau \geq 0$ tal que $p = p_k + \tau d_k$ satisfaz $\|p\| = \Delta$;
        Retorna p
    senão
        $r_{k+1} = r_k + \alpha_k B d_k$
    fim
    se $\|r_{k+1}\| < \varepsilon \|r_0\|$ **então**
        Retorna $p = p_{k+1}$
    senão
        $\beta_{k+1} = r_{k+1}^T r_{k+1} / r_k^T r_k$
        $d_{k+1} = -r_{k+1} + \beta_{k+1} d_k$
    fim
    $k \longleftarrow k + 1$
fim

Relacionando este algoritmo com o Algoritmo CG do Apêndice, notamos que o vetor p toma o lugar de x, a matriz B toma o lugar de A e o vetor $-g$ toma o lugar de b. A mudança de sinal na substituição $b \rightarrow -g$ propaga-se através do algoritmo.

O algoritmo CG de Steihaug difere do CG padrão nos dois primeiros comandos **se** dentro do loop **enquanto**. O primeiro comando **se** encerra o método se a

direção corrente d_k é uma direção de curvatura nula ou curvatura negativa em B. O segundo comando **se** termina o método se p_{k+1} viola a fronteira da região de confiança. Em ambos os casos um ponto final p é encontrado intersectando a direção corrente com a fronteira da região de confiança.

A inicialização $p_0 = 0$ é uma característica decisiva do método. Após a primeira iteração (assumindo $\|r_0\| \geq \varepsilon$), temos

$$p_1 = \alpha_0 d_0 = \frac{r_0^T r_0}{d_0^T B d_0} d_0 = -\frac{g^T g}{g^T B g} g,$$

que é exatamente o ponto de Cauchy. Como cada iteração do método gradiente conjugado reduz o modelo $m(\cdot)$, este algoritmo cumpre a condição necessária para convergência global.

Outra propriedade importante do método é que cada estimativa p_k tem norma maior que a estimativa anterior. Esta propriedade é outra conseqüência da inicialização $p_0 = 0$. Sua principal aplicação é que devemos parar de iterar assim que a fronteira da região de confiança for atingida, pois nenhuma outra iteração que forneça um valor mais baixo em $m(\cdot)$ estará dentro desta região. Declaramos esta propriedade formalmente no seguinte Teorema.

Teorema 1.1. *A seqüência de vetores gerada pelo Algoritmo CG de Steihaug satisfaz*

$$0 = \|p_0\|_2 < \cdots < \left\|p_j\right\|_2 < \left\|p_{j+1}\right\|_2 < \cdots < \|p\|_2 \leq \Delta.$$

Prova: Veja NOCEDAL [7].

■

Deste Teorema, notamos que as iterações do Agoritmo CG de Steihaug percorrem pontos p_k que movem-se em algum caminho interpolador de p_1 até p, um caminho em que cada passo aumenta sua distância do ponto inicial. Quando B é positiva definida, este caminho pode ser comparado ao caminho do método dogleg, pois ambos os métodos se movem do passo de Cauchy p^C ao passo inteiro p^N, até que a fronteira da região de confiança intervenha.

1.4 Escalamento

Problemas de otimização freqüentemente apresentam um escalamento ruim - a função objetivo f é altamente sensível a pequenas mudanças em certas componentes do vetor x e relativamente insensível a mudanças em outras componentes. Um sintoma de escalamento ruim é que o minimizador x^* encontra-se em um vale estreito, e então os contornos da função $f(\cdot)$ próximo de x^* tendem a elipses alongadas. Os algoritmos podem apresentar um desempenho fraco a menos que compensem o escalamento ruim.

Lembrando a definição de região de confiança - uma região ao redor da estimativa corrente dentro da qual o modelo $m_k(\cdot)$ é uma representação adequada da função objetivo $f(\cdot)$ - podemos notar que uma região de confiança esférica não é apropriada no caso de funções com um escalamento ruim. Podemos confiar que o modelo $m_k(\cdot)$ seja razoavelmente exato apenas para distâncias pequenas ao longo das direções altamente sensíveis, mas que seja seguro para grandes distâncias ao longo das direções menos sensíveis. Visto que a forma da região de confiança deve ser tal que nossa confiança no modelo seja mais ou menos a mesma em todos os pontos da fronteira da região, devemos considerar regiões de confiança elípticas nas quais os eixos são menores nas direções mais sensíveis e maiores nas direções menos sensíveis. Regiões de confiança elípticas podem ser definidas por

$$\|Dp\| \leq \Delta,$$

em que D é uma matriz diagonal com elementos diagonais positivos, fornecendo o seguinte subproblema de região de confiança escalado:

$$\underset{p\in\mathbb{R}^n}{\text{minimizar}} \;\; m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \qquad \text{s.a. } \|Dp\| \leq \Delta_k. \tag{1.8}$$

Quando $f(x)$ é altamente sensível ao valor da i-ésima componente x_i, tomamos o elemento diagonal correspodente d_{ii} de D grande. O valor de d_{ii} deve ser próximo de zero para componentes x_i menos sensíveis.

Todos os algoritmos discutidos até o momento podem ser modificados para o caso de regiões de confiança elípticas. No procedimento para o cálculo do ponto de Cauchy, por exemplo, são necessárias mudanças em (1.4) e (1.5). A seguir descrevemos a versão generalizada.

Encontrar o vetor p_k^S que resolve

$$p_k^S = \min_{p \in \mathbb{R}^n} f_k + \nabla f_k^T p \qquad \text{s.a. } \|Dp\| \leq \Delta_k; \tag{1.9}$$

Calcular o escalar $\tau_k > 0$ que minimiza $m_k(\tau p_k^S)$ satisfazendo a fronteira de região de confiança, ou seja,

$$\begin{aligned} \tau_k &= \min_{\tau > 0} m_k(\tau p_k^S) \qquad \text{s.a. } \left\|\tau D p_k^S\right\| \leq \Delta_k. \\ p_k^C &= \tau_k p_k^S. \end{aligned} \tag{1.10}$$

Para esta versão escalada, encontramos

$$p_k^S = -\frac{\Delta_k}{\left\|D^{-1}\nabla f_k\right\|} D^{-2}\nabla f_k,$$

e o tamanho do passo τ_k é obtido da seguinte modificação de (1.6):

$$\tau_k = \begin{cases} 1 & \text{se} \quad \nabla f_k^T D^{-2} B_k D^{-2} \nabla f_k \leq 0 \\ \min(\|D^{-1}\nabla f_k\|^3 / (\Delta_k \nabla f_k^T D^{-2} B_k D^{-2} \nabla f_k), 1) & \text{caso contrário.} \end{cases}$$

Uma alternativa mais simples para ajustar a definição de ponto de Cauchy e os algoritmos descritos até agora à região de confiança elíptica é reescalar as variáveis p no subproblema (1.8) de modo que a região seja elíptica nas variáveis escaladas. Definindo

$$\widetilde{p} = Dp$$

e substituindo em (1.8), obtemos

$$\underset{p \in \mathbb{R}^n}{\text{minimizar}} \ \tilde{m}_k(\tilde{p}) = f_k + \nabla f_k^T D^{-1} \tilde{p} + \frac{1}{2} \tilde{p}^T D^{-1} B_k D^{-1} \tilde{p} \qquad \text{s.a. } \|\tilde{p}\| \leq \Delta_k.$$

A teoria e os algoritmos podem então ser derivados da maneira usual substituindo $\tilde{p}$ por p, $D^{-1}\nabla f_k$ por ∇f_k, $D^{-1} B_k D^{-1}$ por B_k e assim por diante.

Capítulo 2

Sistemas Não-Lineares com Restrições de Caixa

Vamos considerar o problema de encontrar a solução de um conjunto de equações não-lineares com restrições de caixa, isto é, encontrar um vetor $x \in \mathbb{R}^n$ tal que

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

em que $F : X \to \mathbb{R}^n$ é uma aplicação continuamente diferenciável e $X \subseteq \mathbb{R}^n$ é um conjunto aberto contendo a caixa n-dimensional

$$\Omega = \{x \in \mathbb{R}^n / \ l \leq x \leq u\}.$$

Os vetores $l \in (\mathbb{R} \cup -\infty)^n$ e $u \in (\mathbb{R} \cup +\infty)^n$ são cotas inferiores e superiores sobre as variáveis de modo que o conjunto Ω tenha interior não vazio.

Com a intenção de resolver este problema, olhamos primeiro para o problema irrestrito

$$F(x) = 0, \qquad x \in \mathbb{R}^n. \tag{2.2}$$

Analisando os vários métodos na literatura clássica que tentam resolver o sistema de equações não-lineares sem restrições (2.2), notamos que a maioria deles é baseada no método de Newton.

O método de Newton para sistemas de equações não-lineares é um procedimento iterativo que pode ser resumido da seguinte maneira: dado $x_0 \in \mathbb{R}^n$, em cada

iteração resolve-se

$$\begin{aligned} F'(x_k)p_k &= -F(x_k) \\ x_{k+1} &= x_k + p_k \end{aligned} \tag{2.3}$$

Este método tem a vantagem da convergência quadrática quando começando de uma boa estimativa inicial se a matriz Jacobiana de F avaliada na solução x^*, dada por $F'(x^*)$, é não-singular. Mas também tem as desvantagens de não apresentar convergência global para muitos problemas, de avaliar para cada k a matriz $F'(x_k)$ e de requerer em cada iteração a solução do sistema de equações lineares (2.3) que pode ser singular ou mal condicionado. Os métodos existentes para resolver sistemas não-lineares irrestritos tentam contornar estas desvantagens para obter convergência.

Métodos globalmente convergentes para o problema irrestrito $F(x) = 0$ podem ser inadequados para resolver o problema com restrição de caixa (2.1). De fato, tais métodos são propensos a encontrar soluções falsas, isto é, vetores que satisfazem $F(x) = 0$ mas não pertencem à Ω. A existência de soluções falsas pode afetar desfavoravelmente o desempenho de algoritmos numéricos, e a maioria dos métodos numéricos irá convergir tanto para soluções falsas como para soluções significativas. Ainda, mesmo que uma solução significativa esteja no interior estrito do conjunto viável, tomar um ponto inicial na vizinhança desta solução não é garantia de evitar uma solução falsa.

Neste capítulo vamos apresentar um método que combina idéias do método clássico de região de confiança para resolver o sistema de equações irrestrito $F(x) = 0, x \in \mathbb{R}^n$, e a abordagem afim escala para a solução de problemas de otimização restrita dada por Coleman e Li [6]. Grande parte da teoria deste capítulo está fundamentada em [1] e [2] de Bellavia, Macconi e Morini.

Basicamente, o método usa regiões de confiança elípticas definidas por uma mudança de escala. A escala é determinada pela proximidade da estimativa corrente à fronteira da caixa, permitindo que um passo mais longo seja tomado dentro da região viável.

Uma propriedade importante do método descrito a seguir, é a exigência de que todas as iterações estejam no interior estrito de Ω. Para manter a viabilidade estrita, são realizadas limitações apropriadas do passo escolhido, se necessário. Desta forma, nosso método pode trabalhar com problemas nos quais a função F não está definida fora de Ω. Por outro lado, nosso método se reduz a um método de região de confiança padrão para sistemas não-lineares irrestritos quando $\Omega = \mathbb{R}^n$.

2.1 Descrição do Método

Nesta seção vamos generalizar a idéia de região de confiança para sistemas de equações não-lineares irrestritos para o problema com restrições limitadas (2.1) e propor um método que gera aproximações da solução estritamente viáveis. Observamos que a exigência da viabilidade estrita é fundamental pois a violação de uma restrição implica em que a função F não retorna nenhum valor.

Dado $x_k \in int(\Omega)$ e uma direção de busca p_k, olhamos ao longo de p_k para a próxima iteração x_{k+1} em Ω. Seja $\lambda(p_k)$ o tamanho do passo ao longo de p_k até a fronteira, isto é

$$\lambda(p_k) = \begin{cases} \infty & se \quad \Omega = \mathbb{R}^n, \\ \min_i \Lambda_i(p_k) & se \quad \Omega \subset \mathbb{R}^n. \end{cases} \tag{2.4}$$

em que, para cada $i = 1, 2, ..., n$, $\Lambda_i(p_k)$ é dado por

$$\Lambda_i(p_k) = \begin{cases} \max\{\frac{l_i - (x_k)_i}{(p_k)_i}, \frac{u_i - (x_k)_i}{(p_k)_i}\} & se \quad (p_k)_i \neq 0, \\ \infty & se \quad (p_k)_i = 0. \end{cases} \tag{2.5}$$

Note que se $\lambda(p_k) > 1$, então $x_k + p_k$ está dentro de Ω; caso contrário, uma redução no passo ao longo de p_k será necessária para ficar no interior de Ω. Seja $\theta \in (0, 1)$ uma constante fixa, $\zeta(p_k)$ é dado por

$$\zeta(p_k) = \begin{cases} 1 & \text{se } \lambda(p_k) > 1, \\ \max\{\theta, 1 - \|p_k\|\}\lambda(p_k) & \text{caso contrário} \end{cases} \tag{2.6}$$

e

$$\alpha(p_k) = \zeta(p_k) p_k. \tag{2.7}$$

Então, para garantir que a nova iteração seja estritamente viável em relação às restrições de caixa, tomamos

$$x_{k+1} = x_k + \alpha(p_k).$$

Vamos agora considerar o problema de escolher a direção de busca p_k. No

contexto de sistemas não-lineares irrestritos, se x_k é uma boa aproximação de uma solução, o método de Newton pode ser aplicado e p_k é tomada igual à solução p_k^N da equação de Newton

$$F'_k p_k^N = -F_k. \tag{2.8}$$

No entanto, para definir um processo iterativo robusto, o método de Newton pode ser incorporado em um esquema de região de confiança globalmente convergente. Na abordagem clássica de região de confiança, uma região ao redor da estimativa corrente x_k é definida. Dentro desta região, o modelo quadrático

$$m_k(p) = \frac{1}{2}\|F'_k p + F_k\|^2 = \frac{1}{2}\|F_k\|^2 + F_k^T F'_k p + \frac{1}{2} p^T F_k'^T F'_k p = f_k + \nabla f_k^T p + \frac{1}{2} p^T F_k'^T F'_k p$$

é confiável para representar adequadamente a função de mérito

$$f(x) = \frac{1}{2}\|F(x)\|^2.$$

Portanto, a direção de busca p_k é o vetor solução do subproblema

$$\min_p \{m_k(p);\ \|p\| \leq \Delta_k\}, \tag{2.9}$$

para um raio da região de confiança Δ_k dado. Quando o sistema não-linear é restrito, temos que considerar que a exigência da viabilidade estrita pode levar a reduções no passo escolhido p_k. Em particular, se a direção do passo aponta para uma restrição que está próxima, uma fração excessivamente pequena de p_k deve ser tomada para ficar no interior de Ω e isto pode impedir a convergência da seqüência $\{x_k\}$ para uma solução de (2.1). Para prevenir estes problemas, usamos a aplicação afim escala proposta por Coleman e Li no contexto de minimização não-linear com restrições limitadas [5, 6]. Seguindo estas referências, vamos considerar o gradiente $F'^T(x)F(x)$ da função de mérito f, a função vetorial $v(x)$ com componentes $v_i(x)$, $i = 1, 2, ..., n$, dadas por

$$
\begin{array}{rcll}
v_i(x) & = & x_i - u_i & \text{se} \quad (F'^T(x)F(x))_i < 0 \quad \text{e} \quad u_i < \infty \\
v_i(x) & = & x_i - l_i & \text{se} \quad (F'^T(x)F(x))_i \geq 0 \quad \text{e} \quad l_i > -\infty \\
v_i(x) & = & -1 & \text{se} \quad (F'^T(x)F(x))_i < 0 \quad \text{e} \quad u_i = \infty \\
v_i(x) & = & 1 & \text{se} \quad (F'^T(x)F(x))_i \geq 0 \quad \text{e} \quad u_i = -\infty
\end{array} \tag{2.10}
$$

e $D(x)$ a matriz diagonal de escala tal que

$$D(x) = diag(|v_1(x)|^{-1/2}, |v_2(x)|^{-1/2}, \ldots, |v_n(x)|^{-1/2}).$$

Observe que embora $D(x)$ possa não estar definida na fronteira de Ω, $D^{-1}(x)$ pode ser estendida continuamente à fronteira. Denotaremos esta extensão por $D^{-1}(x)$ para todo $x \in \Omega$.

Analisando então a região de confiança elíptica

$$\|D_k p\| \leq \Delta_k,$$

em vez de considerar o subproblema de região de confiança (1.2), vamos considerar o seguinte subproblema de região de confiança elíptico

$$\min_p \{m_k(p);\ \|D_k p\| \leq \Delta_k\}. \tag{2.11}$$

Para este subproblema, o ponto de Cauchy é o ponto que minimiza m_k ao longo da direção de máxima descida escalada $d_k = -D_k^{-2} F_k'^T F_k$ sujeito à fronteira da região de confiança, ou seja,

$$p^C = \tau_k d_k = -\tau_k D_k^{-2} F_k'^T F_k \tag{2.12}$$

em que $\tau_k = \min_{\tau > 0} m_k(\tau d_k)$ s.a. $\|\tau D_k d_k\| \leq \Delta_k$. Pode-se verificar que τ_k tem a forma

$$\tau_k = \min \left\{ \frac{\left\| D_k^{-1} \nabla f_k \right\|^2}{\left\| F_k' D_k^{-2} \nabla f_k \right\|^2}, \frac{\Delta_k}{\left\| D_k^{-1} \nabla f_k \right\|} \right\}. \tag{2.13}$$

Ainda, a respeito da solução p_k do problema (2.11), sabemos de [13], Lema 6.4.1, que existe um parâmetro $\mu_k \geq 0$ tal que p_k resolve o sistema linear

$$(D_k^{-1} F_k'^T F_k' D_k^{-1} + \mu_k I) D_k p_k = -D_k^{-1} F_k'^T F_k. \tag{2.14}$$

O uso da matriz de escala permite que um passo eficaz seja tomado mesmo quando a direção escolhida está próxima de uma fronteira da caixa. Consequentemente as restrições não irão impedir que um passo relativamente grande ao longo de d_k seja tomado. Desta forma, o subproblema de minimização (2.11) trabalha com as restrições implicitamente através da matriz diagonal D_k.

Note que o tamanho da região de confiança Δ_k não tem a intenção de fazer cumprir as restrições. Portanto, para manter a viabilidade estrita, uma redução no passo ao longo da solução p_k de (2.11) pode ser necessário. Este fato origina um passo da forma (2.7).

Para obter convergência global, é suficiente encontrar um vetor p_k tal que $\alpha(p_k)$ forneça uma redução suficiente no modelo quadrático m_k. Esta redução novamente será quantificada em termos do ponto de Cauchy p^C, ou seja, levando em conta as restrições dadas, verificamos se a condição

$$\rho_k^C(p_k) = \frac{m_k(0) - m_k(\alpha(p_k))}{m_k(0) - m_k(\alpha(p^C))} \geq \beta_1 \tag{2.15}$$

é satisfeita para uma constante fixa $\beta_1 \in (0, 1]$. Note que a solução do subproblema de região de confiança (2.11) satisfaz a condição (2.15) mas se uma redução no passo for necessária, isto é, se $\alpha(p_k) \neq p_k$, então a condição (2.15) não é necessariamente satisfeita.

A condição (2.15) não garante uma boa concordância entre o modelo quadrático m_k e a função objetivo f. Desta forma, exigimos que p_k satisfaça

$$\rho_k^f(p_k) = \frac{f(x_k) - f(x_k + \alpha(p_k))}{m_k(0) - m_k(\alpha(p_k))} \geq \beta_2 \tag{2.16}$$

em que β_2 é uma constante dada tal que $\beta_2 \in (0, 1)$.

As condições (2.15) e (2.16) são analisadas da seguinte maneira: se (2.15) não é satisfeita, deixamos a direção corrente p_k e tomamos $p_k = p^C$. Então procedemos como na estratégia clássica de região de confiança, isto é, se a condição (2.16) é satisfeita, tomamos $x_k + \alpha(p_k)$ como próxima estimativa. Caso contrário, o passo $\alpha(p_k)$ é rejeitado e diminuímos o tamanho da região de confiança fazendo

$$\Delta_k = \min\{\alpha_1 \Delta_k, \alpha_2 \|D_k \alpha(p_k)\|\}, \tag{2.17}$$

para constantes α_1, α_2 tal que $0 < \alpha_1 \leq \alpha_2 < 1$ e um novo passo deve ser calculado.

O mecanismo descrito acima de aceitação do passo não termina o algoritmo, no sentido de que um passo $\alpha(p_k)$ aceitável é determinado em um número finito de reduções do raio da região de confiança, como veremos adiante.

Uma vez que o passo $\alpha(p_k)$ é aceito, o raio da região de confiança é atualizado de acordo com as regras padrões. Sabemos que se o raio da região de confiança é muito pequeno em relação à concordância entre o modelo e a função objetivo, o método perde a oportunidade de dar um passo que obteria uma melhora substancial na estimativa. Portanto, ao final de cada iteração, testamos a condição

$$\rho_k^f(p_k) = \frac{f(x_k) - f(x_k + \alpha(p_k))}{m_k(0) - m_k(\alpha(p_k))} \geq \beta_3 \tag{2.18}$$

em que $\beta_3 \in (0, 1]$ é uma constante dada tal que $\beta_2 < \beta_3 < 1$.

Se a condição (2.18) é satisfeita, permitimos um aumento no raio da região de confiança e tomamos

$$\Delta_{k+1} = \max\{\Delta_k, 2\|D_k \alpha(p_k)\|\},$$

caso contrário, o raio da região de confiança é mantido o mesmo.

Estas considerações nos levam ao seguinte algoritmo:

Algoritmo 3: Método de Região de Confiança Escalado

Entrada: $x_0 \in int(\Omega)$, $\Delta_0 > 0$, $\theta \in (0,1)$, $0 < \alpha_1 \leq \alpha_2 < 1$, $\beta_1 \in (0,1]$, $0 < \beta_2 < \beta_3 < 1$

Saída: solução x^*

$k \longleftarrow 0$
Calcule a matriz D_k
enquanto $\rho_k^f(p_k) < \beta_2$ **faça**
    Calcule $p_k = \min_{\|D_k p\| \leq \Delta_k} m_k(p)$
    Calcule p^C usando (2.12) e (2.13)
    Calcule $\alpha(p_k)$ e $\alpha(p^C)$ usando (2.6) e (2.7)
    Calcule $\rho_k^C(p_k)$ usando (2.15)
    se $\rho_k^C(p_k) < \beta_1$ **então**
        $p_k = p^C$
    fim
    Tome $\Delta_k^* = \Delta_k$ e diminua Δ_k usando (2.17)
    Calcule $\rho_k^f(p_k)$ usando (2.16)
fim
Tome $x_{k+1} = x_k + \alpha(p_k)$ e $\Delta_k = \Delta_k^*$
se $\rho_k^f(p_k) \geq \beta_3$ **então**
    Tome $\Delta_{k+1} = \max\{\Delta_k, 2\,\|D_k \alpha(p_k)\|\}$
senão
    Tome $\Delta_{k+1} = \Delta_k$
fim
$k \longleftarrow k + 1$
$x^* \longleftarrow x_k$

Para tornar este algoritmo prático, precisamos resolver o item 4. Na próxima seção, descreveremos duas estratégias para encontrar soluções aproximadas, que atingem no mínimo tanta redução no modelo m_k quanto a redução alcançada pelo ponto de Cauchy.

2.2 O Subproblema de Região de Confiança Elíptico

Para determinar a estimativa x_{k+1} a partir de uma estimativa corrente $x_k \in int(\Omega)$, precisamos resolver o subproblema de região de confiança elíptico

$$\min_p \{m_k(p);\ \|D_k p\| \leq \Delta_k\}, \tag{2.19}$$

em que Δ_k é o raio da região de confiança, D_k é a matriz diagonal de escala e $m_k(p)$ é o modelo quadrático da função objetivo dado por

$$m_k(p) = \frac{1}{2}\|F'_k p + F_k\|^2 = f_k + \nabla f_k^T p + \frac{1}{2} p^T F_k'^T F'_k p$$

A primeira estratégia que vamos considerar com a intenção de resolver este problema é o método dogleg. Uma vez que já analisamos como este método resolve o subproblema de região de confiança padrão, a idéia é simplesmente fazer uma mudança de variável para resolver o subproblema elíptico.

Já sabemos que o passo de Newton p_k^N dado pela solução da equação

$$F'_k p_k^N = -F_k$$

resolve (2.19) se Δ_k é grande o suficiente para que $\left\|D_k p_k^N\right\| \leq \Delta_k$ seja satisfeito.

Por outro lado, se $\left\|D_k p_k^N\right\| > \Delta_k$, calculamos uma solução aproximada do problema (2.19) reescalando a variável p de forma que a região de confiança seja esférica na variável escalada. De fato, definindo $\tilde{p} = D_k p$ e substituindo em (2.19) obtemos

$$\min_{\tilde{p}} \{\tilde{m}_k(\tilde{p}) = f_k + \nabla f_k^T D_k^{-1}\tilde{p} + \frac{1}{2}\tilde{p}^T (D_k^{-1} F_k'^T F'_k D_k^{-1})\tilde{p};\quad \|\tilde{p}\| \leq \Delta_k\}. \tag{2.20}$$

Se $\left\|\tilde{p}_k^N\right\| = \left\|D_k p_k^N\right\| > \Delta_k$, calculamos uma solução aproximada usando o método dogleg, ou seja, aproximamos a trajetória curva ótima por um caminho constituído de dois segmentos. O primeiro segmento vai da origem ao minimizador irrestrito $\tilde{p}_k^u$ do modelo $\tilde{m}_k(\tilde{p})$ ao longo da direção de máxima descida $D_k^{-1}\nabla f_k$:

$$\tilde{p}_k^u = -\frac{\left\|D_k^{-1}\nabla f_k\right\|^2}{\left\|F'_k D_k^{-2}\nabla f_k\right\|^2} D_k^{-1}\nabla f_k \tag{2.21}$$

enquanto o segundo segmento conecta $\tilde{p}_k^u$ à $\tilde{p}_k^N$. O método dogleg aproxima a solução $\tilde{p}_k$ de (2.20) calculando o minimizador do modelo $\tilde{m}_k$ ao longo deste caminho, isto é,

$$\tilde{p}_k = \begin{cases} \Delta_k D_k^{-1} \nabla f_k / \left\| D_k^{-1} \nabla f_k \right\| & \text{se } \left\| \tilde{p}_k^u \right\| \geq \Delta_k, \\ \tilde{p}_k^u + (1-\mu)(\tilde{p}_k^N - \tilde{p}_k^u) & \text{caso contrário,} \end{cases}$$

em que μ é a solução positiva da seguinte equação quadrática

$$\left\| \tilde{p}_k^u + (1-\mu)(\tilde{p}_k^N - \tilde{p}_k^u) \right\|^2 = \Delta_k^2. \tag{2.22}$$

Por fim, para voltar ao espaço original e calcular uma solução aproximada p_k para (2.19), basta tomar $p_k = D_k^{-1} \tilde{p}_k$.

As considerações anteriores levam ao seguinte algoritmo:

Algoritmo 4: Método Dogleg

Entrada: p_k^N, ∇f_k, D_k e Δ_k
Saída: direção p_k

se $\left\| D_k p_k^N \right\| \leq \Delta_k$ **então**
    $p_k = p_k^N$ e pare.
fim
Calcule $\tilde{p}_k^u$ usando (2.21).
se $\left\| \tilde{p}_k^u \right\| \geq \Delta_k$ **então**
    $\tilde{p}_k = \Delta_k D_k^{-1} \nabla f_k / \left\| D_k^{-1} \nabla f_k \right\|$
senão
    $\tilde{p}_k = D_k p_k^N$
    Calcule μ resolvendo (2.22) e tome $\tilde{p}_k = \tilde{p}_k^u + (1-\mu)(\tilde{p}_k^N - \tilde{p}_k^u)$.
fim
Tome $p_k = D_k^{-1} \tilde{p}_k$.

A segunda estratégia que vamos analisar para resolver o problema (2.19) é o método CG de Steihaug. Como já vimos anteriormente, a implementação de Steihaug utiliza o algoritmo gradiente conjugado para encontrar uma solução aproximada para o sistema (2.8) sujeito à região de confiança. No caso do subproblema de região de confiança elíptico, novamente a idéia é fazer a mudança de variável $\tilde{p} = D_k p$ e trabalhar com o

Algoritmo 2 no espaço escalado. Desta forma obtemos o seguinte algoritmo:

Algoritmo 5: CG de Steihaug Escalado

Entrada: $\varepsilon > 0$

Saída: direção p

Faça $p_0 = 0$, $r_0 = D^{-1}g$, $d_0 = -r_0$
$k \longleftarrow 0$
enquanto $\|r_k\| \geq \varepsilon$ **faça**
    se $d_k^T D^{-1} F_k'^T F_k' D^{-1} d_k \leq 0$ **então**
        Encontre τ tal que $p = p_k + \tau d_k$ minimiza $m(p)$ e satizfaz $\|p\| = \Delta$;
        Retorna p
    senão
        $\alpha_k = r_k^T r_k / d_k^T D^{-1} F_k'^T F_k' D^{-1} d_k$
        $p_{k+1} = p_k + \alpha_k d_k$
    fim
    se $\|p_{k+1}\| \geq \Delta$ **então**
        Ache $\tau \geq 0$ tal que $p = p_k + \tau d_k$ satisfaz $\|p\| = \Delta$;
        Retorna p
    senão
        $r_{k+1} = r_k + \alpha_k D^{-1} F_k'^T F_k' D^{-1} d_k$
    fim
    se $\|r_{k+1}\| < \varepsilon \|r_0\|$ **então**
        Retorna $p = p_{k+1}$
    senão
        $\beta_{k+1} = r_{k+1}^T r_{k+1} / r_k^T r_k$
        $d_{k+1} = -r_{k+1} + \beta_{k+1} d_k$
    fim
    $k \longleftarrow k + 1$
fim

No capítulo seguinte vamos analisar a convergência do método. Esta é feita sob certas hipóteses que são padrões no contexto de sistemas de equações não-lineares. O método é globalmente convergente e a taxa de convergência para uma solução no interior de Ω é quadrática.

Capítulo 3

Resultados de Convergência

Neste capítulo vamos investigar as propriedades de convergência dos métodos descritos anteriormente. Primeiro, sejam $r > 0$ e $L = \bigcup_{k=0}^{\infty}\{x \in \Omega; \|x - x_k\| \leq r\}$ uma vizinhança de toda a seqüência $\{x_k\}$ gerada pelo método. Então, sob as hipóteses:

(A1) F' é Lipschitz contínua em L, com constante de Lipschitz $2\gamma_L$;
(A2) $\|F'(x)\|$ é limitada superiormente em L;

vamos declarar os seguintes resultados principais:

- se $\{x_k\}$ é limitada, então todos os seus pontos limites são pontos estacionários para o problema $\min_{\Omega} f$
- se $\{x_k\}$ é limitada e existe um ponto limite isolado x^* tal que $F'(x^*)$ é invertível e $F(x^*) = 0$, então
  (a) $\|F_k\| \to 0$ e $x_k \to x^*$;
  (b) se $\left\|D_k p_k^N\right\| \to 0$ e $\zeta(p_k^N)$ é limitado fora do zero, então $\|F_k\| \to 0$ q-linearmente;
  (c) se $\left\|D_k p_k^N\right\| \to 0$ e $\zeta(p_k^N) \to 1$, então $\{x_k\} \to x^*$ q-superlinearmente;
  (d) se o ponto limite $x^* \in int(\Omega)$, então $x_k \to x^*$ q-quadraticamente.

Para discutir as propriedades teóricas do método, vamos usar a notação:

$$\text{ared}(p) = f(x_k) - f(x_k + \alpha(p)), \qquad \text{pred}(p) = m_k(0) - m_k(\alpha(p)).$$

Os próximos dois lemas fornecem a relação entre a redução atual ared na função f e a redução prevista pred pelo modelo quadrático m_k.

Lema 3.1. *Assuma que (A1) é satisfeito e seja $\varepsilon(x_k, p_k)$ dado por*

$$\varepsilon(x_k, p_k) = \gamma_L \|F_k\| + \frac{1}{2}\gamma_L^2 \|\alpha(p_k)\|^2 . \tag{3.1}$$

Então todo vetor p_k tal que $\left\|F_k'\alpha(p_k) + F_k\right\| \leq \|F_k\|$ satisfaz

$$ared(p_k) \geq pred(p_k) - \varepsilon(x_k, p_k) \|\alpha(p_k)\|^2 . \tag{3.2}$$

Prova: Como $F(x_k + \alpha(p_k)) = F(x_k) + \int_0^1 F'(x_k + \xi\alpha(p_k))\alpha(p_k)\mathrm{d}\xi$, temos

$$\|F(x_k + \alpha(p_k))\|^2 = \|F(x_k) + F'(x_k)\alpha(p_k) + \omega(x_k, p_k)\|^2$$

em que

$$\omega(x_k, p_k) = \int_0^1 (F'(x_k + \xi\alpha(p_k)) - F'(x_k))\alpha(p_k)\mathrm{d}\xi.$$

Portanto, obtemos

$$\begin{aligned} |m_k(\alpha(p_k)) \quad - \quad & f(x_k + \alpha(p_k))| = \frac{1}{2}| \|F(x_k) + F'(x_k)\alpha(p_k)\|^2 - \|F(x_k + \alpha(p_k))\|^2 | \\ \leq \quad & \|F(x_k) + F'(x_k)\alpha(p_k)\| \|\omega(x_k, p_k)\| + \frac{1}{2} \|\omega(x_k, p_k)\|^2 . \end{aligned} \tag{3.3}$$

Pela continuidade de Lipschitz de F', obtemos

$$\|\omega(x_k, p_k)\| \leq \gamma_L \|\alpha(p_k)\|^2 ,$$

e usando $\|F'(x_k)\alpha(p_k) + F(x_k)\| \leq \|F(x_k)\|$ e (3.1), a desigualdade (3.3) se torna

$$|m_k(\alpha(p_k)) - f(x_k + \alpha(p_k))| \leq \varepsilon(x_k, p_k) \|\alpha(p_k)\|^2 .$$

Como $m_k(0) = f(x_k)$, esta desigualdade implica em que

$$\begin{aligned} \text{ared}(p_k) &= f(x_k) - m_k(\alpha(p_k)) + m_k(\alpha(p_k)) - f(x_k + \alpha(p_k)) \\ &\geq \text{pred}(p_k) - |m_k(\alpha(p_k)) - f(x_k + \alpha(p_k))| \geq \text{pred}(p_k) - \varepsilon(x_k, p_k) \|\alpha(p_k)\|^2 . \end{aligned}$$

■

Lema 3.2. *Assuma que (A1) é satisfeito e F'_k é invertível. Seja $\alpha(p_k)$ o passo tomado na k-ésima iteração do método e $\varepsilon(x_k, p_k)$ dado por (3.1). Então, a seguinte desigualdade é satisfeita:*

$$\frac{ared(p_k)}{pred(p_k)} \geq 1 - 2\varepsilon(x_k, p_k) \left\| F_k'^{-1} \right\|^2 .$$

Prova: Veja BELLAVIA, MACCONI e MORINI [1].

■

Vetores p_k que satisfazem (2.15) podem ser caracterizados como vamos mostrar no seguinte lema.

Lema 3.3. *Se p_k satisfaz (2.15) então*

$$pred(p_k) \geq \frac{1}{2}\beta_1 \left\| D_k^{-1} F_k'^T F_k \right\| \min \left\{ \Delta_k, \frac{\left\| D_k^{-1} F_k'^T F_k \right\|}{\left\| D_k^{-1} F_k'^T F_k' D_k^{-1} \right\|}, \frac{\theta \left\| D_k^{-1} F_k'^T F_k \right\|}{\left\| F_k'^T F_k \right\|_\infty} \right\} \tag{3.4}$$

em que θ é a constante usada em (2.6).

Prova: Veja BELLAVIA, MACCONI e MORINI [1].

■

Na seqüência vamos mostrar que cada iteração do método está bem definida.

Lema 3.4. *Assuma que (A1) é satisfeito. Na k-ésima iteração do método, se F'_k é não-singular e $F_k \neq 0$, então o loop-enquanto termina.*

Prova: Seja Δ_k tal que

$$\Delta_k \leq \min\left\{ \frac{\left\|D_k^{-1}F_k'^T F_k\right\|}{\left\|D_k^{-1}F_k'^T F_k' D_k^{-1}\right\|}, \frac{\theta\left\|D_k^{-1}F_k'^T F_k\right\|}{\left\|F_k'^T F_k\right\|_\infty} \right\}. \tag{3.5}$$

Como $\left\|F_k'\alpha(p_k) + F_k\right\| \leq \|F_k\|$, sabemos do Lema 3.1 que a desigualdade (3.2) vale. Para provar este lema vamos explorar este fato e relacionar $\|\alpha(p_k)\|$ e pred(p_k).

Note que $\left\|D_k^{-1}F_k'^T F_k\right\| \neq 0$, (3.5) e (3.4) implica $\Delta_k \leq \tilde{C}_k \text{pred}(p_k)$, em que $\tilde{C}_k = 2/(\beta_1 \left\|D_k^{-1}F_k'^T F_k\right\|)$. Lembrando que $\|D_k p_k\| \leq \Delta_k$, obtemos

$$\|\alpha(p_k)\| \leq \|p_k\| \leq \left\|D_k^{-1}p_k\right\| \Delta_k \tag{3.6}$$

e portanto

$$\|\alpha(p_k)\| \leq \|p_k\| \leq \left\|D_k^{-1}p_k\right\| \tilde{C}_k \text{pred}(p_k). \tag{3.7}$$

Então, por (3.2), (3.6) e (3.7) temos

$$\begin{aligned} \text{ared}(p_k) &\geq \text{pred}(p_k) - (\gamma_L \|F_k\| \left\|D_k^{-1}\right\| \Delta_k + (\gamma_L^2/2)\left\|D_k^{-1}\right\|^3 \Delta_k^3) \|\alpha(p_k)\| \\ &\geq \text{pred}(p_k)(1 - \left\|D_k^{-1}\right\| \tilde{C}_k(\gamma_L \|F_k\| \left\|D_k^{-1}\right\| \Delta_k + (\gamma_L^2/2)\left\|D_k^{-1}\right\|^3 \Delta_k^3)), \end{aligned}$$

e podemos verificar que existe um $\Delta^* > 0$ tal que

$$(1 - \left\|D_k^{-1}\right\| \tilde{C}_k(\gamma_L \|F_k\| \left\|D_k^{-1}\right\| \Delta_k + (\gamma_L^2/2)\left\|D_k^{-1}\right\|^3 \Delta_k^3)) \geq \beta_2,$$

para todo $\Delta_k \leq \Delta^*$. Evidentemente, para tais Δ_k's obtemos

$$\text{ared}(p_k) \geq \beta_2 \text{pred}(p_k),$$

isto é, (2.16) vale para $\Delta_k \leq \min\left\{\Delta^*, \frac{\|D_k^{-1}F_k'^T F_k\|}{\|D_k^{-1}F_k'^T F_k' D_k^{-1}\|}, \frac{\theta\|D_k^{-1}F_k'^T F_k\|}{\|F_k'^T F_k\|_\infty}\right\}$. Desta forma, tomando $\bar{\Delta}_k$ como o valor inicial de Δ_k no Passo 3 do Algoritmo 3, o loop enquanto termina com

$$\Delta_k \geq \min\{\bar{\Delta}_k, \tilde{\Delta}_f^k\},$$

em que

$$\tilde{\Delta}_f^k = \min\left\{\delta_1\Delta^*, \frac{\left\|D_k^{-1}F_k'^T F_k\right\|}{\left\|D_k^{-1}F_k'^T F_k' D_k^{-1}\right\|}, \frac{\theta\left\|D_k^{-1}F_k'^T F_k\right\|}{\left\|F_k'^T F_k\right\|_\infty}\right\}$$

com $\delta_1 = (1/\Delta_k)\min\{\alpha_1\Delta_k, \alpha_2\|D_k\alpha(p_k)\|\}$.

■

Para os resultados de convergência do método, vamos assumir que a seqüência $\{x_k\}$ é limitada. Desta forma, falhas devido à falta de pontos limite são evitados.

Note que se $\{x_k\}$ é limitada, então existe uma constante $\chi_D > 0$ tal que

$$\left\|D^{-1}(x)\right\| < \chi_D, \quad x \in L. \tag{3.8}$$

Ainda, se a hipótese (A1) é satisfeita, $F'(x)^T F(x)$ é Lipschitz contínua em L com constante $2\gamma_L\tilde{\gamma} + \tilde{\gamma}^2$ em que $\tilde{\gamma} = \max\{\sup_{x\in L} f(x), \sup_{x\in L}\|F'(x)\|\}$ e se a hipótese (A2) é satisfeita existe um escalar positivo χ_g tal que o gradiente $F_k'^T F_k$ da função de mérito f satisfaz

$$\left\|F_k'^T F_k\right\|_\infty < \chi_g, \tag{3.9}$$

para todo $x \in L$.

Vamos começar mostrando que se a seqüência $\{x_k\}$ é limitada, então a seqüência $\left\{\left\|D_k^{-1}F_k'^T F_k\right\|\right\}$ converge para zero, isto é, todos os pontos limite de $\{x_k\}$ são pontos estacionários para o problema $\min_\Omega f$.

Teorema 3.1. *Assuma que as hipóteses (A1) e (A2) são satisfeitas. Suponha que a seqüência $\{x_k\}$ das iterações geradas pelo método é limitada superiormente. Então*

$$\lim_{k\to\infty}\left\|D_k^{-1}F_k'^T F_k\right\| = 0.$$

Prova: A prova será feita por contradição. Primeiro, da hipótese (A2) segue que existe uma constante positiva $\chi_B > 0$ tal que $\left\|F_k'^T F_k'\right\| < \chi_B,\ \forall x_k \in L$. Então, de [6] Teorema 3.4, deduzimos que

$$\liminf_{k\to\infty} \left\| D_k^{-1} F_k'^T F_k \right\| = 0. \tag{3.10}$$

Note que de (3.8) e (3.9) existe uma constante $\chi_f > 0$ tal que

$$\left\| D(x)^{-1} F'(x)^T F'(x) D(x)^{-1} \right\| < \chi_f, \quad x \in L.$$

Portanto, usando (3.4) temos

$$\text{pred}(p_k) \geq \frac{1}{2}\beta_1 \left\| D_k^{-1} F_k'^T F_k \right\| \min\left\{ \Delta_k, \frac{\left\| D_k^{-1} F_k'^T F_k \right\|}{\chi_f}, \frac{\theta \left\| D_k^{-1} F_k'^T F_k \right\|}{\chi_g} \right\},$$

e (2.16) fornece

$$\begin{aligned} f(x_k) \;-\; & f(x_k + \alpha(p_k)) \geq \frac{1}{2}\beta_2 \text{pred}(p_k) \\ \geq\; & \frac{1}{2}\beta_1\beta_2 \left\| D_k^{-1} F_k'^T F_k \right\| \min\left\{ \Delta_k, \frac{\left\| D_k^{-1} F_k'^T F_k \right\|}{\chi_f}, \frac{\theta \left\| D_k^{-1} F_k'^T F_k \right\|}{\chi_g} \right\}. \end{aligned} \tag{3.11}$$

Agora vamos supor que exista uma seqüência $\{m_i\}$ tal que $\left\| D_{m_i}^{-1} F_{m_i}'^T F_{m_i} \right\| \geq \varepsilon_1$ para algum $\varepsilon_1 \in (0, 1)$. Usando (3.10) podemos afirmar que para qualquer $\varepsilon_2 \in (0, \varepsilon_1)$ existe uma subseqüência de $\{m_i\}$, sem perda de generalidade vamos assumir que é a seqüência inteira, e uma seqüência $\{l_i\}$ tal que

$$\left\| D_k^{-1} F_k'^T F_k \right\| \geq \varepsilon_2, \quad m_i \leq k < l_i \quad \left\| D_{l_i}^{-1} F_{l_i}'^T F_{l_i} \right\| < \varepsilon_2. \tag{3.12}$$

Então (3.11) fornece

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2}\beta_1\beta_2\varepsilon_2 \min\left\{ \Delta_k, \frac{\varepsilon_2}{\chi_f}, \frac{\theta\varepsilon_2}{\chi_g} \right\}, \quad m_i \leq k < l_i,$$

e como de (3.6) temos $\|x_{k+1} - x_k\| \leq \chi_D \Delta_k$, concluímos que

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2}\beta_1\beta_2\varepsilon_2 \min\left\{ \frac{\|x_{k+1} - x_k\|}{\chi_D}, \frac{\varepsilon_2}{\chi_f}, \frac{\theta\varepsilon_2}{\chi_g} \right\}, \quad m_i \leq k < l_i. \tag{3.13}$$

A seqüência $\{f(x_k)\}$ converge pois é não-crescente e limitada inferiormente por zero. Portanto, $f(x_k) - f(x_{k+1})$ tende à zero. De (3.13) segue que

$$f(x_k) - f(x_{k+1}) \geq \varepsilon_3 \|x_{k+1} - x_k\|, \quad m_i \leq k < l_i,$$

para i suficientemente grande e $\varepsilon_3 = \frac{1}{2}\beta_1\beta_2\varepsilon_2/\chi_D$. Então, usando a desigualdade triangular obtemos

$$f(x_{m_i}) - f(x_{k_i}) \geq \varepsilon_3 \left\|x_{m_i} - x_{k_i}\right\|, \quad m_i \leq k_i < l_i, \tag{3.14}$$

e podemos concluir que $\left\|x_{m_i} - x_{k_i}\right\|$ tende à zero. Ainda, pela continuidade de Lipschitz de $F'^T F$ e do fato de que $\left\|x_{m_i} - x_{k_i}\right\|$ tende à zero segue que

$$\left\|F'^T_{m_i} F_{m_i} - F'^T_{k_i} F_{k_i}\right\| \leq \varepsilon_2, \tag{3.15}$$

para i suficientemente grande.

Sem perda de generalidade, assuma que toda a seqüência x_{l_i} converge para um ponto, digamos x^*. De (3.14) temos que $\{x_{m_i}\}$ converge para x^* também.

Se $(F'(x^*)^T F(x^*))_j \neq 0$ para algum $1 \leq j \leq n$, então (2.10) implica

$$|(v_{m_i})_j - (v_{l_i})_j| \leq |(x_{m_i})_j - (x_{l_i})_j|$$

para i suficientemente grande. Consequentemente $\left\|(D^{-1}_{m_i} - D^{-1}_{l_i})F'^T_{l_i} F_{l_i}\right\| \to 0$ e portanto

$$\left\|(D^{-1}_{m_i} - D^{-1}_{l_i})F'^T_{l_i} F_{l_i}\right\| \leq \varepsilon_2, \tag{3.16}$$

para i suficientemente grande. Finalmente, de $\left\|D^{-1}_{m_i} F'^T_{m_i} F_{m_i}\right\| \geq \varepsilon_1$, (3.12), (3.15) e (3.16) e

$$\left\|D^{-1}_{m_i} F'^T_{m_i} F_{m_i}\right\| \leq \left\|D^{-1}_{m_i}\right\| \left\|F'^T_{m_i} F_{m_i} - F'^T_{l_i} F_{l_i}\right\| + \left\|(D^{-1}_{m_i} - D^{-1}_{l_i})F'^T_{l_i} F_{l_i}\right\| + \left\|D^{-1}_{l_i} F'^T_{l_i} F_{l_i}\right\|,$$

temos

$$\varepsilon_1 \leq (\chi_D + 2)\varepsilon_2,$$

isto é, uma contradição pois $\varepsilon_2 \in (0, \varepsilon_1)$ pode ser arbitrariamente pequeno.

■

O seguinte resultado é uma aplicação direta do teorema anterior.

Corolário 3.1. *Assuma que as hipóteses (A1) e (A2) são satisfeitas. Se a seqüência* $\{x_k\}$ *gerada pelo método é limitada e existe um ponto limite* x^* *de* $\{x_k\}$ *tal que* $x^* \in int(\Omega)$ *e* $F'(x^*)$ *é não-singular, então* $\|F_k\| \to 0$ *e todos os pontos de acumulação de* $\{x_k\}$ *resolvem o problema (2.1).*

Prova: Veja BELLAVIA, MACCONI e MORINI [1].

■

Observe que pode acontecer de nenhum ponto limite de $\{x_k\}$ satisfazer $F(x) = 0$. Neste caso, a matriz $D(x)^{-1}F'(x)^T$ é singular em cada ponto limite x^*. O fato de $D(x^*)^{-1}F'(x^*)^T$ ser singular ocorre ou se $F'(x^*)$ é singular ou se $D(x^*)^{-1}$ é singular. Este último caso ocorre quando x^* está na fronteira de Ω.

Na seqüência vamos formalizar as propriedades de convergência do método quando $\{x_k\}$ tem pelo menos um ponto de acumulação x^* tal que $F(x^*) = 0$ e $F'(x^*)$ é invertível.

Teorema 3.2. *Assuma que as hipóteses (A1) e (A2) são satisfeitas e que a seqüência de iterações* $\{x_k\}$ *geradas pelo método é limitada. Se* x^* *é um ponto limite isolado de* $\{x_k\}$ *tal que* $F(x^*) = 0$ *e* $F'(x^*)$ *é não-singular, então* $\{x_k\}$ *converge para* x^*.

Prova: Primeiro, note que se $x^* \in int(\Omega)$, as hipóteses $F(x^*) = 0$ e x^* ser um ponto limite isolado são redundantes. De fato, a hipótese $F'(x^*)$ ser não-singular e o Corolário 3.1 implicam em $F(x^*) = 0$. Ainda, da invertibilidade de $F'(x^*)$ e da hipótese (A1) sabemos que existe uma vizinhança de x^* que pertence a $int(\Omega)$ em que $F'(x)$ é não-singular e $\|F(x)\| > 0$ se $x \neq x^*$. Portanto, do fato de todos os vetores $x \neq x^*$ em tal vizinhança verificarem $D(x)^{-1}F'(x)^T F(x) \neq 0$ e do Teorema 3.1, podemos concluir que x^* é um ponto limite isolado de $\{x_k\}$.

Como a seqüência $\{\|F_k\|\}$ é monótona decrescente, esta é também convergente. A hipótese $F(x^*) = 0$ implica $\|F_k\| \to 0$.

Tome $K = \left\|F'(x^*)^{-1}\right\|$ e seja $\rho > 0$ suficientemente pequeno de forma que quando $x \in B_\rho(x^*) \cap \Omega$, $F'(x)^{-1}$ existe e $\left\|F'(x)^{-1}\right\| \leq 2K$. Também, seja $\{x_{k_j}\}$ uma subseqüência tal que $x_{k_j} \to x^*$ e j_0 o índice tal que $x_{k_j} \in B_\rho(x^*) \cap \Omega$ quando $k_j \geq k_{j_0}$, Assuma $k_j \geq k_{j_0}$.

Em cada passo do método, a nova estimativa x_{k+1} tem a forma $x_{k+1} = x_k + \alpha(p_k)$, o passo $\alpha(p_k)$ é tal que $\|\alpha(p_k)\| \leq \|p_k\|$ e p_k pode ser ou a solução do problema (2.11) ou o ponto de Cauchy (2.12). Para provar o teorema, vamos mostrar que, para toda subseqüência $\{x_{k_j}\}$ convergindo para x^*, $\lim_{k_j \to \infty} p_{k_j} = 0$. Então, aplicando [14] Lema 4.10, a prova está concluída.

Logo, $F_{k_j} \to 0$ e $\left\|F'^{-1}_{k_j}\right\| \leq 2K$ fornece $\lim_{k_j \to \infty} p^N_{k_j} = 0$.

Quanto ao ponto de Cauchy $p^C_{k_j}$, temos $p^C_{k_j} = -\tau_{k_j} D^{-2}_{k_j} F'^T_{k_j} F_{k_j}$ em que τ_k está definido em (2.13). Então, obtemos a seguinte desigualdade:

$$\left\|p^C_{k_j}\right\| \leq \frac{\left\|D^{-1}_{k_j} F'^T_{k_j} F_{k_j}\right\|^2}{\left\|F'_{k_j} D^{-2}_{k_j} F'^T_{k_j} F_{k_j}\right\|^2} \left\|D^{-2}_{k_j} F'^T_{k_j} F_{k_j}\right\| \leq \frac{\left\|D^{-1}_{k_j} F'^T_{k_j} F_{k_j}\right\|^2}{\left\|D^{-2}_{k_j} F'^T_{k_j} F_{k_j}\right\|^2} \left\|F'^{-1}_{k_j}\right\|^2.$$

Como $\left\|F'^{-1}_{k_j}\right\| \leq 2K$, para provar que $\left\|p^C_{k_j}\right\| \to 0$ temos que mostrar

$$\left\|D^{-1}_{k_j} F'^T_{k_j} F_{k_j}\right\|^2 \Big/ \left\|D^{-2}_{k_j} F'^T_{k_j} F_{k_j}\right\|^2 \to 0.$$

Com este propósito, note que

$$\left\|D^{-1}_{k_j} F'^T_{k_j} F_{k_j}\right\| \leq \sqrt{n} \left\|D^{-1}_{k_j} F'^T_{k_j} F_{k_j}\right\|_\infty = \sqrt{n} \sqrt{|v_{i*}(x_{k_j})|}\, |(F'^T_{k_j} F_{k_j})_{i*}|,$$

para um índice $1 \leq i* \leq n$ e

$$\left\|D^{-2}_{k_j} F'^T_{k_j} F_{k_j}\right\| \geq \left\|D^{-2}_{k_j} F'^T_{k_j} F_{k_j}\right\|_\infty \geq |v_{i*}(x_{k_j})|\, |(F'^T_{k_j} F_{k_j})_{i*}|.$$

Então

$$\frac{\left\|D^{-1}_{k_j} F'^T_{k_j} F_{k_j}\right\|^2}{\left\|D^{-2}_{k_j} F'^T_{k_j} F_{k_j}\right\|} \leq n \frac{|v_{i*}(x_{k_j})|(|(F'^T_{k_j} F_{k_j})_{i*}|)^2}{|v_{i*}(x_{k_j})||(F'^T_{k_j} F_{k_j})_{i*}|},$$

e de $F'^T_{k_j} F_{k_j} \to 0$ concluímos que $\left\|p^C_{k_j}\right\| \to 0$.

Agora, vamos nos concentrar no caso em que p_{k_j} é a solução do problema (2.11) e é diferente de $p^N_{k_j}$. Neste caso, sabemos que p_{k_j} resolve o sistema linear (2.14) com $k = k_j$. Então, temos que

$$(F'^T_{k_j} F'_{k_j} + \mu_{k_j} D^2_{k_j}) p_{k_j} = F'^T_{k_j} F_{k_j},$$

em que $\mu_{k_j} > 0$. Isto fornece

$$p^T_{k_j} F'^T_{k_j} F'_{k_j} p_{k_j} + \mu_{k_j} p^T_{k_j} D^2_{k_j} p_{k_j} = -p^T_{k_j} F'^T_{k_j} F_{k_j}. \tag{3.17}$$

Portanto, observando que

$$\frac{\left\|p^2_{k_j}\right\|}{\left\|F'^{-1}_{k_j}\right\|^2} \leq \left\|F'_{k_j} p_{k_j}\right\|^2 \leq p^T_{k_j} F'^T_{k_j} F'_{k_j} p_{k_j} + \mu_{k_j} p^T_{k_j} D^2_{k_j} p_{k_j},$$

e usando (3.17) temos

$$\frac{\left\|p^2_{k_j}\right\|}{\left\|F'^{-1}_{k_j}\right\|^2} \leq -p^T_{k_j} F'^T_{k_j} F_{k_j} \leq \left\|p_{k_j}\right\| \left\|F'_{k_j}\right\| \left\|F_{k_j}\right\|.$$

Portanto, de (A2) e $\left\|F_{k_j}\right\| \to 0$ concluímos que $\left\|p_{k_j}\right\| \to 0$.

Resumindo, mostramos que o vetor p_{k_j} tende a zero. Daí, cada passo $\alpha(p_{k_j})$ tende à zero e usando [14] Lema 4.10, concluímos que $\{x_k\}$ converge para x^*.

■

No seguinte teorema vamos analisar a taxa de convergência assintótica da seqüência $\{x_k\}$.

Teorema 3.3. *Assuma que as hipóteses (A1) e (A2) são satisfeitas e que existe uma solução x^* de (2.1) tal que $F'(x^*)$ é não-singular. Se a seqüência $\{x_k\}$ gerada pelo método converge para x^*, então a seqüência $\{\Delta_k\}$ é limitada fora do zero. Ainda,*

(I) Se $\left\|D_k p^N_k\right\| \to 0$ e, eventualmente, p^N_k satisfaz

$$\zeta(p^N_k) \geq 1 - \sqrt{1 - \beta_1},$$

então existe $c \in (0, 1)$ tal que

$$\|F'(x^*)(x_{k+1} - x^*)\| \leq c \|F'(x^*)(x_k - x^*)\|,$$

e $\|F_k\| \to 0$ *q-linearmente.*

(II) Se $\left\|D_k p_k^N\right\| \to 0$ *e* $\zeta(p_k^N) \to 1$*, então* $x_k \to x^*$ *q-superlinearmente.*

(III) Se $\left\|D_k p_k^N\right\| \to 0$ *e, eventualmente,* p_k^N *satisfaz* $\zeta(p_k^N) = 1$*, então* $x_k \to x^*$ *q-quadraticamente.*

Prova: Tome $K = \left\|F'(x^*)^{-1}\right\|$. Seja $\rho > 0$ suficientemente pequeno para que $F'(x)^{-1}$ exista, $\left\|F'(x)^{-1}\right\| \leq 2K$ sempre que $x \in B_\rho(x^*) \cap \Omega$. Seja x_m tal que $x_m \in B_\rho(x^*) \cap \Omega$ e toda a seqüência $\{x_k\}_{k>m}$ pertença à $B_\rho(x^*) \cap \Omega$. Assuma $k > m$.

Note que nossas hipóteses implicam em que $\|F_k\| \to 0$ e $\|\alpha(p_k)\| \to 0$.

Para provar o teorema precisamos mostrar que existe um $\bar{\Delta}$ independente de k tal que $\Delta_k > \bar{\Delta}$ para k suficientemente grande. Ou seja, ao final do Passo 14 do Algoritmo 3, precisamos de estimativas de Δ_k que sejam independentes de k. Com este propósito, vamos mostrar que $\rho_k^f(p_k) \geq \beta_3$ é satisfeito para k suficientemente grande. Do Lema 3.2 temos

$$\frac{\text{ared}(p_k)}{\text{pred}(p_k)} \geq 1 - 2\varepsilon(x_k, p_k)\left\|F_k'^{-1}\right\|^2,$$

em que $\varepsilon(x_k, p_k)$ é dado por (3.1). Então obtemos

$$\frac{\text{ared}(p_k)}{\text{pred}(p_k)} \geq 1 - 8\varepsilon(x_k, p_k)K^2.$$

Como $\|F_k\| \to 0$ e $\|\alpha(p_k)\| \to 0$ temos que $\varepsilon(x_k, p_k) \to 0$. Portanto, existem $0 < \delta \leq \rho$ e $\xi > 0$ tal que

$$\varepsilon(x_k, p_k) \leq \frac{(1-\beta_3)}{8K^2},$$

sempre que $x_k \in B_\delta(x^*)$ e $\|\alpha(p_k)\| \leq \xi$. Neste caso, temos

$$\frac{\text{ared}(p_k)}{\text{pred}(p_k)} \geq \beta_3, \tag{3.18}$$

isto é, $\rho_k^f(p_k) \geq \beta_3$ para k suficientemente grande, e a condição (2.16) é satisfeita.

Agora vamos estabelecer uma estimativa de Δ_k^* no fim do loop enquanto, isto é, Passo 3 do Algoritmo 3. Primeiro, vamos denotar por $\bar{\Delta}_k$ o valor inicial de Δ_k. Assuma que $\Delta_k \leq \frac{\xi}{\chi_D}$ em que χ_D está definido em (3.8). Como $\|D_k p_k\| \leq \Delta_k$, temos

$$\|p_k\| \leq \left\|D_k^{-1}\right\| \|D_k p_k\| \leq \chi_D \Delta_k \leq \xi$$

e $\|\alpha(p_k)\| \leq \xi$. Claro que se o loop enquanto não diminui Δ_k, temos $\Delta_k^* = \bar{\Delta}_k$ no final; mas se o loop enquanto diminui Δ_k então o penúltimo valor de Δ_k^* é pelo menos $\frac{\xi}{\chi_D}$, daí $\Delta_k^* \geq \frac{\delta_1 \xi}{\chi_D}$ no final. Isto é, existe $k_0 \geq m$ tal que para $k \geq k_0$, o loop enquanto termina com

$$\Delta_k^* \geq \min\left\{\bar{\Delta}_k, \frac{\delta_1 \xi}{\chi_D}\right\}.$$

Ainda, se no final do Passo 3 , $\Delta_k^* \leq \frac{\xi}{\chi_D}$, devido à (3.18), temos $\Delta_{k+1} > \Delta_k = \Delta_k^*$ no Passo 15. Por outro lado, se no final do Passo 3, $\Delta_k^* > \frac{\xi}{\chi_D}$, então no Passo 15 temos $\Delta_{k+1} \geq \Delta_k > \delta_1 \Delta_k = \delta_1 \Delta_k^* > \delta_1 \frac{\xi}{\chi_D}$. Logo, ao final de cada iteração temos

$$\Delta_{k+1} \geq \min\left\{\bar{\Delta}_{k_0}, \frac{\delta_1 \xi}{\chi_D}\right\},$$

e, por indução, temos

$$\liminf_{k\to\infty} \Delta_k > 0.$$

Portanto, $\{\Delta_k\}$ é limitada fora do zero.

Como temos por hipótese que $\left\|D_k p_k^N\right\| \to 0$, segue que para k suficientemente grande p_k^N é a solução do subproblema de região de confiança e $\rho_k^f(p_k^N) \geq \beta_3$.

Agora, se provarmos que $\alpha(p_k^N)$ satisfaz a condição (2.15), podemos concluir que para k suficientemente grande o passo de Newton truncado é usado. Levando em conta que $\text{pred}(p_k^N) = \frac{1}{2}\zeta(p_k^N)(2 - \zeta(p_k^N)) \|F_k\|^2$ e $\text{pred}(p_k^C) \leq \frac{1}{2}\|F_k\|^2$, obtemos

$$\frac{\text{pred}(p_k^N)}{\text{pred}(p_k^C)} \geq \zeta(p_k^N)(2 - \zeta(p_k^N)).$$

Como por hipótese $\zeta(p_k^N) \geq 1 - \sqrt{1-\beta_1}$, segue que

$$\frac{\text{pred}(p_k^N)}{\text{pred}(p_k^C)} \geq \beta_1,$$

e (2.15) é satisfeito.

Resumindo, para k suficientemente grande, x_{k+1} tem a forma

$$x_{k+1} = x_k + \alpha(p_k^N). \tag{3.19}$$

Como $\left\|F'_k(p_k^N) + F_k\right\| = (1 - \zeta(p_k^N))\,\|F_k\|$, $\alpha(p_k^N)$ pode ser interpretado como um passo de Newton Inexato e as afirmações (I) e (II) seguem do Teorema 6.4.1 e da Proposição 6.1.1 em [12].

Finalmente, a afirmação (III) segue imediatamente de $\alpha(p_k^N) = p_k^N$ e (3.19).

■

Por último vamos mostrar um resultado que caracteriza a taxa de convergência da seqüência $\{x_k\}$ quando o ponto limite x^* pertence a $int(\Omega)$. Este resultado é uma conseqüência direta do teorema anterior.

Corolário 3.2. *Assuma que as hipóteses (A1) e (A2) são satisfeitas. Seja $\{x_k\}$ a seqüência das iterações geradas pelo método e assuma que $x_k \to x^*$. Se $x^* \in int(\Omega)$ e $F'(x^*)$ é invertível, então o passo de Newton inteiro p_k^N é eventualmente tomado e a taxa de convergência é quadrática.*

Prova: Como $x^* \in int(\Omega)$ e $F'(x^*)$ é invertível, o Corolário 3.1 implica em $F(x^*) = 0$. Seja $\rho \in (0, 2]$ suficientemente pequeno de modo que $B_\rho(x^*) \subset int(\Omega)$ e $F'(x)$ é não-singular para $x \in B_\rho(x^*)$. Seja x_m tal que $x_m \in B_{\rho/2}(x^*)$ e toda a seqüência $\{x_k\}, k > m$ pertença à $B_{\rho/2}(x^*)$. Assuma $k > m$. Então $|l_i - (x_k)_i| > \rho/2$ e $|u_i - (x_k)_i| > \rho/2$ para $i = 1, 2, ..., n$ e consequentemente $\|D_k\| \leq \sqrt{2/\rho}$ com $\sqrt{2/\rho} \geq 1$. Também, de (A2) e $p_k^N = F'(x_k)^{-1}F(x_k)$ segue que $p_k^N \to 0$. Isto implica que $\left\|D_k p_k^N\right\| \to 0$ e usando (2.5) e (2.6) também temos que $\zeta(p_k^N) = 1$ para k suficientemente grande. Desta forma as hipóteses da afirmação (III) do Teorema 3.3 são verificadas e a tese do corolário segue diretamente do Teorema 3.3.

No caso do método CG de Steihaug, para obter a convergência quadrática, basta exigir que o critério de parada torne-se mais estrito ao longo das iterações. Desta maneira, a taxa de convergência quadrática é obtida do ponto de vista teórico, mas na prática esta estratégia não é eficiente. [7]

■

Capítulo 4

Resultados Numéricos

Neste capítulo vamos resumir os resultados dos experimentos numéricos realizados com a intenção de verificar características e capacidades dos métodos estudados. Os testes foram executados usando problemas disponíveis na literatura surgidos de modelos matemáticos relacionados à fenômenos físicos [3]. Na seqüência, apresentaremos o problema do fluxo de cargas em redes de energia elétrica, mostrando suas equações, variáveis, parâmetros e formulações e em seguida descreveremos os resultados obtidos na resolução deste problema. Finalmente, vamos comparar estes resultados com o método de Newton clássico.

4.1 Características da Implementação e Resultados

Os problemas inicialmente relacionados têm dimensões que vão de duas à catorze equações e evidentemente todos têm soluções que pertencem ao interior de Ω. Cada problema possui várias estimativas iniciais, ou seja, cada problema fornece vários testes; e como o método exige viabilidade estrita, não foram consideradas as estimativas iniciais não viáveis. Em [3] cada problema possui um nome de identificação que está associado com a dimensão do problema; de maneira que vamos nos referir aos problemas usando estes nomes.

Este conjunto de problemas fornece vários tipos de sistemas restritos . De fato, várias aplicações possuem discontinuidades no conjunto Ω. Ainda, estão relacionados sistemas com solução na fronteira do conjunto viável, sistemas com variáveis livres, sistemas somente com limitante inferior (superior), sistemas com variáveis limitadas tanto acima como abaixo.

Na implementação do método, a convergência é atingida quando a seguinte

condição é satisfeita:

$$\|F_{k+1}\| \leq \text{atol} + \text{rtol}\,\|F_0\|,$$

em que atol e rtol são constantes de tolerância dadas.

Também, falhas são declaradas se alguma das seguintes situações ocorre:
ERRO 1: o número máximo de iterações maxit foi atingido.
ERRO 2: o número máximo de avaliações da função F maxnf foi atingido.
ERRO 3: o tamanho da região de confiança tornou-se muito pequeno.
ERRO 4: nenhuma melhora no resíduo não-linear foi obtida, ou seja,

$$\|F_{k+1} - F_k\| \leq \varepsilon\,\|F_k\|.$$

ERRO 5: a norma do gradiente escalado da função de mérito tornou-se muito pequena, isto é,

$$\left\|D_k^{-1}\nabla f_k\right\| < \varepsilon.$$

ERRO 6: a matriz de escala D_k não pode ser calculada pois a seqüência de iterações está se aproximando de uma fronteira da caixa.

O método foi programado de forma que poucos dados são exigidos pelo programa, apenas a função F, os vetores l e u que especificam as cotas inferior e superior, uma estimativa inicial x_0, critérios de parada atol e rtol e o número máximo permitido de iterações e avaliações da função F. Considerando o raio inicial da região de confiança Δ_0, pode-se escolher

$$\Delta_0 = 1, \quad \text{ou} \quad \Delta_0 = \left\|D_0^{-1}\nabla f_0\right\|.$$

Se o problema não disponibiliza a matriz Jacobiana F', o método a aproxima usando diferenças finitas [13].

No final, o algoritmo indica o sucesso ou a falha do procedimento e retorna, além da estimativa corrente, as seguintes informações:

- o número de iterações realizadas;
- o número de F-avaliações realizadas;
- a norma do valor corrente de $F(x)$;

- a norma do valor corrente do gradiente escalado $D^{-1}(x)\nabla f(x)$;
- o número de reduções do raio da região de confiança.

Os resultados obtidos com $\Delta_0 = 1$ estão apresentados na Tabela 4.1, em que para cada problema está listado o número NT de testes realizados e o número NS de testes resolvidos com sucesso. Ainda, para os testes bem sucedidos, relatamos a média MIT de iterações realizadas e a média MAF de avaliações da função F desempenhadas.

As observações baseadas na tabela 4.1 são as seguintes: de um total de 112 testes, 80 foram bem sucedidos com o método dogleg e 67 com a abordagem CG de Steihaug; oito problemas foram resolvidos para apenas uma das estimativas iniciais com o dogleg e nove com o CG de Steihaug; e 17 problemas foram resolvidos para todas as estimativas iniciais utilizadas usando dogleg e 13 usando CG de Steihaug. Os problemas 22 e 27 que falharam em ambos os métodos são classificados como problemas de alta dificuldade.

A maioria dos problemas foi resolvida com poucas iterações em pelo menos um dos métodos testados. Note que MAF é menor que 40 para 28 problemas em cada um dos métodos. Note ainda que MIT é quase igual a MAF na maioria dos casos, isto é, a maior parte dos problemas foi resolvida com poucas reduções do raio da região de confiança.

O comportamento do método foi ligeiramente afetado pela escolha do raio inicial da região de confiança. Escolhendo $\Delta_0 = \left\|D_0^{-1}\nabla f_0\right\|$ foram resolvidos 82 testes com o método dogleg e 75 com a abordagem CG de Steihaug. Estes dados encontram-se na tabela 4.2.

Observe que, para o método dogleg, a escolha $\Delta_0 = \left\|D_0^{-1}\nabla f_0\right\|$ resultou, de um modo geral, em um menor número médio de iterações e de avaliações da função F do que a escolha $\Delta_0 = 1$. De qualquer maneira, testes adicionais com outros problemas deveriam ser realizados para estabelecer a efetividade entre as escolhas do raio inicial de região de confiança.

		Método Dogleg			Método CG de Steihaug		
Problemas	NT	NS	MIT	MAF	NS	MIT	MAF
TWOEQ2	4	2	8	9	2	8	9
TWOEQ3	5	3	11	12	3	11	12
TWOEQ4a	3	3	4	5	3	4	5
TWOEQ4b	3	3	4	5	3	4	5
TWOEQ5a	4	4	4	5	4	4	5
TWOEQ5b	4	4	6	7	4	6	7
TWOEQ6	4	4	9	15	4	9	15
TWOEQ7	4	4	7	8	4	7	8
TWOEQ8	4	2	4	5	2	4	5
TWOEQ9	4	4	307	343	4	307	343
TWOEQ10	4	4	6	7	4	6	7
THREEQ1	4	4	21	27	4	343	363
THREEQ2	4	1	5	6	1	5	6
THREEQ3	4	4	14	15	4	14	15
THREEQ4a	4	1	5	6	1	5	6
THREEQ4b	4	4	7	8	4	7	8
THREEQ5	4	1	20	28	1	15	17
THREEQ6	4	4	103	144	2	68	94
THREEQ8	1	1	6	7	1	6	7
FOUREQ1	5	5	9	11	3	17	20
FIVEQ1	4	2	9	10	1	9	11
SIXEQ1	4	0	-	-	0	-	-
SIXEQ2a	2	1	3	4	1	3	4
SIXEQ2b	2	1	4	5	1	4	5
SIXEQ2c	2	1	3	4	1	3	4
SIXEQ3	4	2	9	11	0	-	-
SIXEQ4a	3	0	-	-	0	-	-
SIXEQ4b	4	3	343	385	0	-	-
SEVENEQ1	3	3	16	45	2	28	40
SEVENEQ2a	3	1	5	6	1	5	6
TENEQ1a	2	2	13	14	2	13	14
14EQ1	2	2	10	11	0	-	-

Tabela 4.1: Performance do método com $\Delta_0 = 1$.

		Método Dogleg			Método CG de Steihaug		
Problemas	NT	NS	MIT	MAF	NS	MIT	MAF
TWOEQ2	4	2	5	6	2	5	6
TWOEQ3	5	4	8	9	4	8	9
TWOEQ4a	3	3	5	6	3	5	6
TWOEQ4b	3	3	6	7	3	6	7
TWOEQ5a	4	5	7	5	4	5	7
TWOEQ5b	4	4	8	9	4	8	9
TWOEQ6	4	4	8	12	4	8	12
TWOEQ7	4	4	8	10	4	8	10
TWOEQ8	4	2	4	5	2	4	5
TWOEQ9	4	4	310	348	4	310	348
TWOEQ10	4	4	8	10	4	8	10
THREEQ1	4	4	27	34	4	311	335
THREEQ2	4	1	5	6	1	5	6
THREEQ3	4	4	5	6	4	5	6
THREEQ4a	4	1	5	6	1	5	6
THREEQ4b	4	4	6	8	4	6	8
THREEQ5	4	3	6	7	3	6	8
THREEQ6	4	4	74	109	3	6	7
THREEQ8	1	1	5	6	1	5	6
FOUREQ1	5	4	5	6	4	5	6
FIVEQ1	4	2	13	14	1	14	16
SIXEQ1	4	0	-	-	0	-	-
SIXEQ2a	2	1	3	4	1	3	4
SIXEQ2b	2	1	4	5	1	4	5
SIXEQ2c	2	1	3	4	1	3	4
SIXEQ3	4	3	7	9	2	7	8
SIXEQ4a	3	0	-	-	0	-	-
SIXEQ4b	4	3	7	8	1	10	11
SEVENEQ1	3	3	10	12	2	5	6
SEVENEQ2a	3	1	4	5	1	4	5
TENEQ1a	2	2	8	9	2	8	9
14EQ1	2	1	5	7	0	-	-

Tabela 4.2: Performance do método com $\Delta_0 = \left\|D_0^{-1}\nabla f_0\right\|$.

4.2 Fluxo de Carga - Aspectos Gerais

O cálculo do fluxo de carga em uma rede de energia elétrica consiste essencialmente na determinação do estado da rede, da distribuição dos fluxos e de algumas outras grandezas de interesse. Neste tipo de problema a modelagem do sistema é estática, ou seja, a rede é representada por um conjunto de equações e inequações algébricas.

Os componentes de um sistema de energia elétrica podem ser classificados em dois grupos:

- **barras** - geradores, cargas, reatores e capacitores
- **circuitos** - elementos que interligam as barras (linhas de transmissão e transformadores)

As equações básicas do fluxo de carga são obtidas exigindo-se a conservação das potências ativa e reativa em cada barra, isto é, a potência líquida injetada deve ser igual à soma das potências que fluem pelos componentes internos conectados à barra. Isso equivale a se impor a primeira lei de Kirchhoff.

O problema do fluxo de carga pode ser formulado por um sistema de equações e inequações algébricas não-lineares que correspondem, respectivamente, às leis de Kirchhoff e a um conjunto de restrições operacionais da rede elétrica e de seus componentes. Na formulação mais simples do problema, para cada barra da rede são associadas quatro variáveis, sendo que duas delas entram no problema como dados e duas como incógnitas:

V_k - magnitude da tensão nodal na k-ésima barra

θ_k - ângulo de fase da tensão nodal na k-ésima barra

P_k - injeção líquida de potência ativa na k-ésima barra

Q_k - injeção líquida de potência reativa na k-ésima barra

A tensão complexa na barra k é dada por $E_k = V_k e^{j\theta_k}$, em que $j = \sqrt{-1}$. Dependendo de quais variáveis nodais entram como dados e quais são consideradas como incógnitas, definem-se três tipos de barras:

PQ - são dados P_k e Q_k, e calculados V_k e θ_k

PV - são dados P_k e V_k, e calculados Q_k e θ_k

Folga - são dados V_k e θ_k, e calculados P_k e Q_k

As barras do tipo PQ e PV são utilizadas para representar, respectivamente, barras de carga e barras de geração. A barra de folga tem uma dupla função: fornecer a referência angular e fechar o balanço de potência do sistema, considerando as perdas de transmissão não conhecidas antes de se ter a solução final do problema.

O conjunto de equações do problema do fluxo de carga é formado por duas equações para cada barra, cada uma delas representando o fato de as potências ativa e reativa injetadas em uma barra serem iguais à soma dos fluxos correspondentes que deixam a barra através de linhas de transmissão, transformadores, etc. Essas equações são representadas por:

$$P_k = \sum_{m \in \Omega_k} P_{km}(V_k, V_m, \theta_k, \theta_m)$$

$$Q_k = \sum_{m \in \Omega_k} Q_{km}(V_k, V_m, \theta_k, \theta_m)$$

em que $k = 1, \ldots, nb$, sendo nb o número de barras e Ω_k o conjunto das barras vizinhas à barra k. Duas barras são vizinhas quando existe um circuito interligando-as.

4.2.1 Modelagem de Linhas de Transmissão

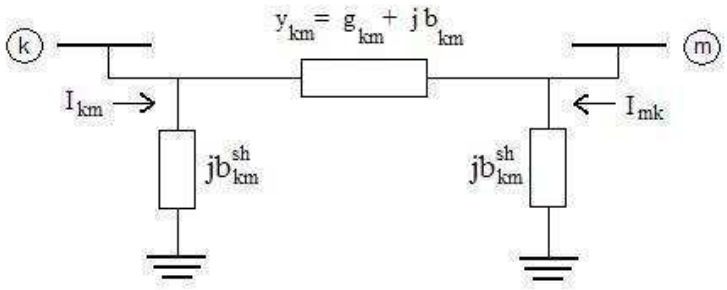

Figura 4.1: Linha de transmissão k-m

Seja I_{km} a corrente em uma linha de transmissão (que liga a barra k à barra m). A injeção líquida de corrente na barra k é obtida aplicando-se a primeira lei de Kirchhoff:

$$I_k + I_k^{sh} = \sum_{m \in \Omega_k} I_{km}, \qquad k = 1, \ldots, nb$$

Esta expressão, utilizando as equações nodais da rede elétrica, pode ser escrita na seguinte forma matricial:

$$I = YE,$$

em que I é o vetor das injeções de corrente, cujas componentes são I_k, $k = 1, \ldots, nb$. O vetor E representa as tensões nodais cujas componentes são $E_k = V_k e^{j\theta_k}$. A matriz $Y = G + jB$ é denominada matriz de admitância nodal, sendo G a matriz de condutância e B a matriz de susceptância. Os elementos da matriz Y são obtidos da seguinte maneira:

$$Y_{km} = -y_{km} \qquad Y_{kk} = jb_k^{sh} + \sum_{m \in \Omega_k} (jb_{km}^{sh} + y_{km})$$

em que b_k^{sh} corresponde à susceptância de equipamentos reativos conectados à barra k, b_{km}^{sh} é a metade da susceptância shunt do circuito conectando as barras k e m; e y_{km} é a admitância série do circuito conectando as barras k e m. Em geral a matriz Y é esparsa pois $Y_{km} = 0$ sempre que entre as barras k e m não existirem circuitos conectando-as.

As equações de potências ativa e reativa são deduzidas aplicando as leis de Kirchhoff, e são dadas respectivamente por

$$P_k^{cal} = G_{kk}V_k^2 + V_k \sum_{m\in\Omega_k} V_m[G_{km}\cos(\theta_k - \theta_m) + B_{km}\,\mathrm{sen}(\theta_k - \theta_m)] \tag{4.1}$$

$$Q_k^{cal} = -B_{kk}V_k^2 + V_k \sum_{m\in\Omega_k} V_m[G_{km}\,\mathrm{sen}(\theta_k - \theta_m) - B_{km}\cos(\theta_k - \theta_m)] \tag{4.2}$$

em que $k = 1, \ldots, nb$ e Ω_k é o conjunto de índices das barras vizinhas à barra k excluindo a própria barra k.

4.2.2 Formulação do Problema

Considere inicialmente um problema no qual são dados P_k e Q_k para as barras PQ; P_k e V_k para as barras PV; e V_k e θ_k para a barra Vθ (folga). Pede-se para calcular V_k e θ_k nas barras PQ; θ_k nas barras PV; e P_k e Q_k para a barra de folga. Sejam NPQ e NPV o número de barras PQ e PV da rede, respectivamente (será considerada a existência de apenas uma barra de folga). Desta forma o problema formulado anteriormente pode ser decomposto em dois subsistemas de equações algébricas, conforme indicado a seguir:

Subsistema 1: (dimensão 2NPQ +NPV)

Neste subproblema são dados P_k e Q_k nas barras PQ, e P_k e V_k nas barras PV. Pretende-se calcular V_k e θ_k nas barras PQ e θ_k nas barras PV. Ou seja, trata-se de um sistema de (2NPQ+NPV) equações algébricas não-lineares com o mesmo número de incógnitas, isto é:

$$\begin{aligned} P_k^{dado} - P_k^{cal} &= 0 \quad \text{para as barras PQ e PV;} \\ Q_k^{dado} - Q_k^{cal} &= 0 \quad \text{para as barras PQ.} \end{aligned}$$

Subsistema 2: (dimensão NPV+2)

Resolvido o Subsistema 1, e portanto, já sendo conhecidos V_k e θ_k para todas as barras, deseja-se calcular P_k e Q_k na barra de folga. Trata-se de um sistema com NPV+2 equações algébricas não-lineares com o mesmo número de incógnitas, no qual todas as incógnitas aparecem de forma explícita, o que torna trivial o processo de resolução. O mesmo não ocorre com o Subsistema 1, no qual as incógnitas são implícitas, o que exige um processo de iteração para resolvê-las. Os dois subsistemas correspondem

ao *problema de Fluxo de Carga*.

As incógnitas do Subsistema 1 podem ser agrupadas no vetor x dado a seguir:

$$x = \begin{bmatrix} V \\ \theta \end{bmatrix}$$

em que V é o vetor das magnitudes das tensões das barras PQ e θ é o vetor dos ângulos das tensões das barras PQ e PV. As expressões que formam o Subsistema 1 podem ser reescritas da seguinte maneira:

$$\begin{aligned} \Delta P_k &= P_k^{dado} - P_k^{cal}(V,\theta) = 0 \quad \text{para as barras PQ e PV} \\ \Delta Q_k &= Q_k^{dado} - Q_k^{cal}(V,\theta) = 0 \quad \text{para as barras PQ,} \end{aligned}$$

em que ΔP_k e ΔQ_k são, respectivamente, os balanços de potências ativa e reativa na barra k. As funções ΔP_k e ΔQ_k podem ser colocadas na forma vetorial

$$\begin{aligned} \Delta P &= P^{dado} - P^{cal}(V,\theta) \\ \Delta Q &= Q^{dado} - Q^{cal}(V,\theta), \end{aligned}$$

em que $P^{cal}(V,\theta)$ e $Q^{cal}(V,\theta)$ são os vetores das injeções de potências ativa e reativa calculadas a partir das variáveis de estado.

Considerando a função vetorial

$$F(x) = \begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix}$$

o Subsistema 1 pode ser colocada na forma

$$F(x) = 0.$$

Ainda, consideraremos limites máximo e mínimo para a variável V de maneira que obteremos um sistema de equações não-lineares com restrição de caixa.

4.2.3 Resultados Numéricos

Nesta seção vamos descrever os resultados numéricos obtidos na resolução do problema do fluxo de carga usando os métodos estudados nos capítulos anteriores e o método de Newton puro.

O critério de parada utilizado foi $\|F(x)\|_2 \leq 10^{-8}$ e os sistemas teste usados foram IEEE6, IEEE30 e IEEE118 barras. As dimensões dos Jacobianos dos sistemas testados estão na tabela (4.3).

Tabela 4.3: Dimensões dos Jacobianos

Sistema	IEEE6	IEEE30	IEEE118
Dimensão	9×9	53×53	201×201

A performance de cada método está descrita nas tabelas (4.4) e (4.5), em que para cada sistema, está declarado o número de iterações que foram necessárias para que o critério de parada fosse atingido. Para a tabela (4.4), foram utilizados os seguintes dados:

Dados iniciais I:

- a estimativa inicial foi 2.4 para a variável V e nula para a variável θ,
- a variável V teve cota inferior -1 e cota superior 3
- a variável θ foi deixada livre

Tabela 4.4: Performance dos métodos com dados I

	Método Dogleg	Abordagem CG de Steihaug	Newton Puro
Sistema de 6 barras	7	7	6
Sistema de 30 barras	7	8	6
Sistema de 118 barras	11	9	não convergiu

Já para a tabela (4.5), foram utilizados:

Dados iniciais II:

- a estimativa inicial foi 3 para a variável V e nula para a variável θ,
- a variável V teve cota inferior -1 e cota superior 4
- a variável θ foi deixada livre

Tabela 4.5: Performance dos métodos com dados II

	Método Dogleg	Abordagem CG de Steihaug	Newton Puro
Sistema de 6 barras	7	7	7
Sistema de 30 barras	8	9	6
Sistema de 118 barras	ERRO 3	11	não convergiu

Estes resultados foram obtidos tomando como raio inicial da região de confiança $\Delta_0 = 1$. Para o raio inicial $\Delta_0 = \left\|D_0^{-1}\nabla f_0\right\|$, em geral, não houve diferença entre os métodos; e por isso estes resultados serão omitidos.

Analisando as tabela (4.4), podemos observar que para o sistema de 118 barras, o método de Newton não obteve convergência enquanto os métodos de região de confiança resolveram o problema em uma média de 10 iterações. Analisando agora a tabela (4.5), novamente o método de Newton não convergiu para o sistema de 118 barras. Também o método dogleg não obteve convergência para este problema, acusando que o raio da região de confiança tornou-se muito pequeno.

Os dados iniciais considerados para o problema do fluxo de carga em redes de energia elétrica não correspondem à realidade, mas foram tomados para avaliar as características e capacidades dos métodos estudados.

É interessante notar na figura (4.2) a estrutura esparsa da matriz Jacobiana para o sistema de 118 barras. Observe a presença de muitos elementos não-nulos longe da diagonal principal. Estes fatos sugerem que o método CG de Steihaug pode ser mais adequado para este problema.

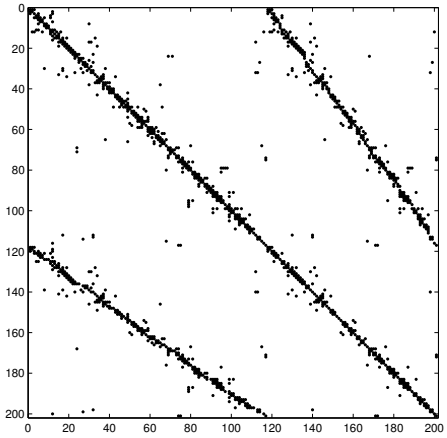

Figura 4.2: Estrutura do Jacobiano para 118 barras

Conclusão

Neste trabalho, abordamos dois métodos de região de confiança para sistemas não-lineares. Ambos os métodos foram adaptados para as restrições de caixa em que o escalamento desempenha um importante papel proporcionando, em geral, passos robustos. O método CG de Steihaug apresenta a vantagem de obter direções com menor custo computacional, já que não necessariamente atinge a direção de Newton, como faz o método Dogleg. Como seqüência natural deste trabalho, sugerimos testes com problemas de grande porte em que técnicas específicas para obtenção da direção de Newton ou aproximação para a mesma devem ser empregadas. Outra possibilidade importante é abordar os sistemas indeterminados em que várias soluções são possíveis ou mesmo sistemas sobredeterminados. Estes últimos têm uma formulação importante no contexto do problema das equações da rede elétrica.

Referências Bibliográficas

[1] BELLAVIA, S.; MACCONI, M.; MORINI, B. *An affine scaling trust-region method approach to bound-constrained nonlinear systems*. Applied Numerical Mathematics, vol.44, pp.257-280, 2003.

[2] BELLAVIA, S.; MACCONI, M.; MORINI, B. *STRSCNE: A Scaled Trust-Region Solver for Constrained Nonlinear Equations*. Computational Optimization and Applications, vol. 28, pp. 31-50, 2004.

[3] SHACHAM M., CUTLIP M.B., BRAUNER N. **POLYMATH**. A Web-Based Library for Testing the Performance of Numerical Software. Disponível em: <http://www.polymath-software.com>. Acesso em: 10 fevereiro 2006.

[4] MONTICELLI, A. J. *Fluxo de Carga em Redes de Energia Elétrica*. Edgard Blücher, São Paulo, 1983.

[5] COLEMAN, T. F.; LI, Y. *On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds*. Math. Programming, vol.67, pp.189-224, 1994.

[6] COLEMAN, T. F.; LI, Y. *An interior trust-region approach for nonlinear minimization subject to bounds*. SIAM Journal on Optimization, vol.6, pp.418-445, 1996.

[7] NOCEDAL, J.; WRIGHT, S. J. *Numerical Optimization*. New York: Springer-Verlag New York, Inc., 1999. (Springer Series in Operations Research).

[8] FRANCISCO, J. B. *Métodos Numéricos Aplicados à Resolução das Equações da Rede Elétrica*. Dissertação de Mestrado. Departamento de Matemática. Universidade Federal de Santa Catarina, 2002.

[9] GAVA, G.; SACHINE, M.; ZAMBALDI, M. C. *Métodos de Região de Confiança para Sistemas Não-Lineares Esparsos com Restrições de Caixa*. Anais do 62^o Seminário Brasileiro de Análise, Rio de Janeiro, 2005.

[10] CONN, A. R.; GOULD, N. I. M.; TOINT, P. L. *Trust-Region Methods*. Philadelphia: Society for Industrial and Applied Mathematics, 2000.

[11] KELLEY, C. T. *Iterative Methods for Optimization*. Philadelphia: Society for Industrial and Applied Mathematics, 1999.(Frontiers in Applied Mathematics).

[12] KELLEY, C. T. *Iterative Methods for Solving Linear and Nonlinear Equations*. Philadelphia: Society for Industrial and Applied Mathematics, vol. 16, 1995.(Frontiers in Applied Mathematics).

[13] DENNIS, J. E.; SCHNABEL, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: Society for Industrial and Applied Mathematics, 1996.

[14] MORÉ, J. J.; SORENSEN, D. C. *Computing a trust region step*. SIAM Journal on Scientific Computing, vol.4, pp.553-572, 1983.

[15] GILL, P. E.; MURRAY, W.; WRIGHT, M. H. *Practical Optimization*. Academic Press, 1981.

[16] GOLUB, G. H.; VAN LOAN, C. F. *Matrix Computations*. London: The Johns Hopkins University Press, 3rd. Edition, 1996.

[17] CONN, A. R.; GOULD, N. I. M.; TOINT, Ph.L. *LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization*. Berlin: Springer-Verlag, 1992.

[18] CEREDA, R. L. D.; MALDONADO, J. C. *Introdução ao FORTRAN 77 para Microcomputadores*. São Paulo: McGraw-Hill, 1987.

[19] HEHL, M. E. *Linguagem de Programação Estruturada: FORTRAN 77*. São Paulo: McGraw-Hill, 1987.

Apêndice A

Método Gradiente Conjugado

Neste apêndice, vamos analisar o método gradiente conjugado linear, o procedimento mais útil para resolver sistemas de equações lineares de grande porte. Esta técnica foi proposta por Hestenes e Stiefel na década de 50 como um método iterativo para resolver sistemas lineares com a matriz dos coeficientes definida positiva e deste modo, é uma alternativa para a eliminação Gaussiana. Na seqüência vamos derivar o método gradiente conjugado linear e discutir as principais propriedades de convergência. Para simplificar, vamos deixar o termo "linear"através do texto.

O método gradiente conjgado é um método iterativo para resolver um sistema de equações lineares

$$Ax = b, \tag{A.1}$$

em que A é uma matriz $n \times n$ simétrica definida positiva. O problema (A.1) pode ser equivalentemente declarado como o seguinte problema de minimização:

$$\phi(x) = \frac{1}{2}x^T Ax - x^T b, \tag{A.2}$$

isto é, tanto (A.1) quanto (A.2) têm a mesma e única solução. Esta equivalência nos permite interpretar o método gradiente conjugado ou como um algoritmo para resolver sistemas lineares ou como uma técnica para minimização de funções quadráticas convexas. Neste momento, é importante notar que o gradiente da função ϕ é igual ao resíduo do sistema linear, ou seja,

$$\nabla\phi(x) = Ax - b = r(x). \tag{A.3}$$

A.1 Métodos de Direções Conjugadas

Uma das propriedades mais importantes do método gradiente conjugado é a habilidade de gerar, de uma maneira muito econômica, um conjunto de vetores conjugados. Um conjunto de vetores não nulos $\{p_0, p_1, \ldots, p_l\}$ é dito ser conjugado em relação à matriz simétrica definida positiva A se

$$p_i^T A p_j = 0, \qquad \forall i \neq j. \tag{A.4}$$

A importância da conjugacidade está no fato de que podemos minimizar a função $\phi(\cdot)$ em n passos minimizando-a sucessivamente ao longo das direções individuais em um conjunto conjugado. Para verificar esta afirmação vamos considerar o seguinte método das *direções conjugadas*: dado $x_0 \in \mathbb{R}^n$ e um conjunto de direções conjugadas $\{p_0, p_1, \ldots, p_{n-1}\}$, vamos gerar uma seqüência $\{x_k\}$ tomando

$$x_{k+1} = x_k + \alpha_k p_k, \tag{A.5}$$

em que α_k é o minimizador da função $\phi(\cdot)$ ao longo de $x_k + \alpha p_k$, dado explicitamente por

$$\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}. \tag{A.6}$$

Temos o seguinte resultado.

Teorema A.1. *Para qualquer $x_0 \in \mathbb{R}^n$ a seqüência $\{x_k\}$ gerada pelo algoritmo de direções conjugadas (A.5), (A.6) converge para a solução x^* do sistema linear (A.1) em no máximo n passos.*

Prova: Como as direções $\{p_i\}$ são linearmente independentes, pois são conjugadas, elas geram todo o espaço $\mathbb{R}^n$. Assim, podemos escrever a diferença entre x_0 e a solução x^* da seguinte maneira:

$$x^* - x_0 = \sigma_0 p_0 + \sigma_1 p_1 + \cdots + \sigma_{n-1} p_{n-1},$$

para alguma escolha de escalares σ_k. Pré-multiplicando esta expressão por $p_k^T A$ e usando a propriedade (A.4), obtemos

$$\sigma_k = \frac{p_k^T A(x^* - x_0)}{p_k^T A p_k}. \tag{A.7}$$

O teorema está provado se mostrarmos que estes coeficientes σ_k coincidem com os tamanhos do passo α_k gerados por (A.6).

Se x_k é gerado pelo algoritmo (A.5) e (A.6) temos

$$x_k = x_0 + \alpha_0 p_0 + \alpha_1 p_1 + \cdots + \alpha_{k-1} p_{k-1}.$$

Pré-multiplicando esta expressão por $p_k^T A$ e usando a propriedade da conjugacidade, obtemos

$$p_k^T A(x_k - x_0) = 0,$$

e portanto

$$p_k^T A(x^* - x_0) = p_k^T A(x^* - x_k) = p_k^T (b - A x_k) = -p_k^T r_k.$$

Comparando este resultado com (A.6) e (A.7), obtemos $\sigma_k = \alpha_k$, concluindo a demonstração do teorema.

■

Uma simples interpretação das propriedades das direções conjugadas é que se a matriz A em (A.2) é diagonal, as curvas de nível da função $\phi(\cdot)$ são elipses cujos eixos são paralelos aos eixos coordenados. Assim, podemos encontrar o minimizador desta função realizando minimizações sucessivas ao longo das direções coordenadas $e_1, e_2, \ldots, e_n$.

Quando A não é diagonal, as curvas de nível ainda são elipses, mas não são paralelas aos eixos coordenados. A estratégia de minimizações sucessivas ao longo destas direções não nos levam à solução em n iterações. Neste caso, basta fazer uma mudança de variável para tornar a matriz A diagonal e então minimizar ao longo das

direções coordenadas.

Outra propriedade interessante do método das direções conjugadas é que quando a matriz Hessiana da quadrática é diagonal, cada minimização ao longo dos eixos coordenados determina corretamente uma das componentes da solução x^*. Ou seja, após k minimizações, a quadrática foi minimizada no subespaço gerado por $e_1, e_2, \ldots, e_k$. O seguinte teorema prova este resultado para o caso geral em que a matriz Hessiana não é necessariamente diagonal. Ao provar o resultado, vamos usar a seguinte expressão que pode ser verificada a partir das relações (A.3) e (A.5):

$$r_{k+1} = r_k + \alpha_k A p_k. \tag{A.8}$$

Teorema A.2. *Seja $x_0 \in \mathbb{R}^n$ arbitrário e suponha que a seqüência $\{x_k\}$ é gerada pelo algoritmo das direções conjugadas (A.5) e (A.6). Então*

$$r_k^T p_i = 0, \qquad i = 0, 1, \ldots, k-1, \tag{A.9}$$

e x_k é o minimizador de $\phi(x) = \frac{1}{2}x^T A x - x^T b$ sobre o conjunto

$$\{x; x = x_0 + span\{p_0, p_1, \ldots, p_{k-1}\}\}. \tag{A.10}$$

Prova: Primeiramente vamos mostrar que um ponto $\tilde{x}$ minimiza ϕ sobre o conjunto (A.10) se e somente se $r(\tilde{x})^T p_i = 0$, para cada $i = 0, 1, \ldots, k-1$. Defina $h(\sigma) = \phi(x_0 + \sigma_0 p_0 + \cdots + \sigma_{k-1} p_{k-1})$, em que $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_{k-1})^T$. Como $h(\sigma)$ é uma quadrática estritamente convexa, possui um único minimizador σ^* que satisfaz

$$\frac{\partial h(\sigma^*)}{\partial \sigma_i} = 0, \quad i = 0, 1, \ldots, k-1.$$

Pela regra da cadeia, isto implica que

$$\nabla\phi(x_0 + \sigma_0^* p_0 + \cdots + \sigma_{k-1}^* p_{k-1})^T p_i = 0, \quad i = 0, 1, \ldots, k-1.$$

Logo, obtemos o resultado desejado usando (A.3).

Agora, usando indução, vamos mostrar que x_k satisfaz (A.9). Como α_k é

sempre o minimizador unidimensional, temos que $r_1^T p_0 = 0$. Seja a hipótese de indução $r_{k-1}^T p_i = 0$ para $i = 0, \ldots, k-2$. Por (A.8) obtemos

$$r_k = r_{k-1} + \alpha_{k-1} A p_{k-1},$$

e então

$$p_{k-1}^T r_k = p_{k-1}^T r_{k-1} + \alpha_{k-1} p_{k-1}^T A p_{k-1} = 0,$$

pela definição (A.6) de α_{k-1}. Por outro lado, para os outros vetores p_i, $i = 0, 1, \ldots, k-2$, temos

$$p_i^T r_k = p_i^T r_{k-1} + \alpha_{k-1} p_i^T A p_{k-1} = 0$$

pela hipótese de indução e a conjugacidade de p_i. Logo, concluímos que $r_k^T p_i = 0$ para $i = 0, 1, \ldots, k-1$, completando a prova.

■

O fato do resíduo corrente ser ortogonal a todas as direções anteriores, como está expresso em (A.9), é uma propriedade que será amplamente usada neste apêndice.

A discussão até agora tem sido geral, no sentido de que se aplica a um método de direção conjugada baseado em uma escolha arbitrária do conjunto de direções conjugadas $\{p_0, p_1, \ldots, p_{n-1}\}$. Existem várias maneiras de se escolher o conjunto de direções conjugadas, por exemplo, usar os autovalores da matriz A, ou então modificar o processo de Gram-Schmidt. No entanto, estas abordagens têm um alto custo computacional pois armazenam todo o conjunto de direções.

A.2 Propriedades do Método Gradiente Conjugado

O método gradiente conjugado é um método de direções conjugadas que possui uma importante propriedade: ao gerar o conjunto de vetores conjugados, pode calcular um novo vetor p_k usando apenas o vetor anterior p_{k-1}. O método não necessita conhecer todas as direções anteriores $p_0, p_1, \ldots, p_{k-2}$ do conjunto conjugado; p_k é automaticamente conjugado a estes vetores. Esta propriedade implica que o método exige pouco armazenamento e cálculo de dados.

Cada direção p_k é escolhida como sendo uma combinação linear da direção de máxima descida $-\nabla\phi(x_k)$ (que é igual ao resíduo negativo $-r_k$) e da direção anterior

p_{k-1}. Temos

$$p_k = -r_k + \beta_k p_{k-1}, \tag{A.11}$$

em que o escalar β_k é determinado pela exigência de que p_{k-1} e p_k devem ser conjugadas em relação à matriz A. Pré-multiplicando (A.11) por $p_{k-1}^T A$ e impondo a condição $p_{k-1}^T A p_k = 0$, temos

$$\beta_k = \frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}.$$

É intuitivo escolher a primeira direção p_0 como sendo a direção de máxima descida no ponto inicial x_0. Como no método das direções conjugadas, são realizadas sucessivas minimizações unidimensionais ao longo de cada direção. Desta forma, obtemos o seguinte algoritmo:

Algoritmo 6: Gradiente Conjugado - Versão Preliminar

Entrada: x_0

Saída: solução x^*

Faça $r_0 = Ax_0 - b$ e $p_0 = -r_0$

$k \longleftarrow 0$

enquanto $r_k \neq 0$ **faça**

  $\alpha_k \longleftarrow -\dfrac{p_k^T r_k}{p_k^T A p_k}$

  $x_{k+1} \longleftarrow x_k + \alpha_k p_k$

  $r_{k+1} \longleftarrow A x_{k+1} - b$

  $\beta_{k+1} \longleftarrow \dfrac{r_{k+1}^T A p_k}{p_k^T A p_k}$

  $p_{k+1} \longleftarrow -r_{k+1} + \beta_{k+1} p_k$

  $k \longleftarrow k + 1$

fim

$x^* \longleftarrow x_{k+1}$

Posteriormente, vamos apresentar uma versão mais eficiente do método gradiente conjugado; a versão acima é útil para entender as propriedades essenciais do

método. Vamos mostrar primeiro que as direções $p_0, p_1, \ldots, p_{n-1}$ são de fato conjugadas, o que pelo Teorema A.1 implica finalização em n passos. O teorema abaixo estabelece esta propriedade e outras duas propriedades importantes; os resíduos r_i são mutuamente ortogonais e cada direção p_k e resíduo r_k está contido no subespaço de Krylov de grau k para r_0, definido por

$$\mathcal{K}(r_0; k) = span\{r_0, Ar_0, \ldots, A^k r_0\}.$$

Teorema A.3. *Suponha que a k-ésima iteração gerada pelo método gradiente conjugado não é a solução x^*. As seguintes propriedades são verdadeiras:*

$$\begin{aligned} r_k^T r_i &= 0, \quad para \quad i = 0, \ldots, k-1, \\ span\{r_0, r_1, \ldots, r_k\} &= span\{r_0, Ar_0, \ldots, A^k r_0\}, \\ span\{p_0, p_1, \ldots, p_k\} &= span\{r_0, Ar_0, \ldots, A^k r_0\}, \\ p_k^T A p_i &= 0, \quad para \quad i = 0, \ldots, k-1. \end{aligned}$$

Além disso, a solução x^* é obtida em no máximo n iterações.

Prova: Veja NOCEDAL [7].

■

Note que, como os gradientes r_k são mutuamente ortogonais, a designação "método gradiente conjugado" é incorreta. São as direções, e não os gradientes, que são conjugados em relação à matriz A.

Uma forma mais econômica do método gradiente conjugado pode ser derivada usando os resultados dos Teoremas A.2 e A.3. Primeiro, usando (A.9) e a equação do passo 8 do Algoritmo 6, podemos substituir a fórmula para α_k por

$$\alpha_k = \frac{r_k^T r_k}{p_k^T A p_k}.$$

Também, temos de (A.8) que $\alpha_k A p_k = r_{k+1} - r_k$, e usando novamente (A.9) e a equação do passo 8 do Algoritmo 6, obtemos

$$\beta_{k+1} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}.$$

Usando estas fórmulas e a igualdade (A.8) obtemos a seguinte forma para o método do gradiente conjugado.

Algoritmo 7: Gradiente Conjugado

Entrada: x_0

Saída: solução x^*

Faça $r_0 = Ax_0 - b$ e $p_0 = -r_0$

$k \longleftarrow 0$

enquanto $r_k \neq 0$ **faça**

$$\alpha_k \longleftarrow \frac{r_k^T r_k}{p_k^T A p_k}$$

$x_{k+1} \longleftarrow x_k + \alpha_k p_k$

$r_{k+1} \longleftarrow r_k + \alpha_k A p_k$

$$\beta_{k+1} \longleftarrow \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$$

$p_{k+1} \longleftarrow -r_{k+1} + \beta_{k+1} p_k$

$k \longleftarrow k + 1$

fim

$x^* \longleftarrow x_{k+1}$

Em qualquer ponto do Algoritmo 7, nunca precisamos conhecer os vetores x, r e p apenas da iteração corrente e da anterior. Os cálculos com mais alto custo computacional realizados em cada iteração são o produto matriz-vetor Ap_k, os produtos internos $p_k^T(Ap_k)$ e $r_{k+1}^T r_{k+1}$, e três somas vetoriais. O método CG é recomendado apenas para problemas grandes, caso contrário a eliminação Gaussiana ou qualquer outro algoritmo de fatoração deve ser utilizado. Para grandes problemas, o método CG tem a vantagem de não alterar a matriz dos coeficientes, e algumas vezes aproxima a solução rapidamente [7].