UNIVERSIDADE FEDERAL DO PARANÁ

RODRIGO GARCIA EUSTÁQUIO

CLASSE DE MÉTODOS CHEBYSHEV-HALLEY INEXATA LIVRE DE TENSORES COM CONVERGÊNCIA CÚBICA PARA RESOLUÇÃO DE SISTEMAS NÃO LINEARES E UM ESTUDO SOBRE RAIO DE CONVERGÊNCIA

Curitiba

2013

RODRIGO GARCIA EUSTÁQUIO

CLASSE DE MÉTODOS CHEBYSHEV-HALLEY INEXATA LIVRE DE TENSORES COM CONVERGÊNCIA CÚBICA PARA RESOLUÇÃO DE SISTEMAS NÃO LINEARES E UM ESTUDO SOBRE RAIO DE CONVERGÊNCIA

Tese apresentada ao Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, dos Setores de Ciências Exatas e Tecnologia da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Ciências.

Orientador:

Prof. Dr. Ademir Alves Ribeiro.

Co-orientador:

Prof. Dr. Miguel Angel Dumett Canales.

Curitiba

2013

E91c Eustáquio, Rodrigo Garcia
Classe de métodos Chebyshev-Halley inexata livre de tensores com convergência cúbica para resolução de sistemas não lineares e um estudo sobre raio de convergência / Rodrigo Garcia Eustáquio. – Curitiba, 2013.
136f. : il. color. ; 30 cm.

Tese(doutorado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-graduação em Métodos Numéricos em Engenharia, 2013.

Orientador: Ademir Alves Ribeiro -- Co-orientador: Miguel Angel Dumett Canales.
Bibliografia: p. 110-116.

1. Chebyshev, Aproximação de. 2. Sistemas não lineares I. Universidade Federal do Paraná. II. Ribeiro, Ademir Alves. III. Canales, Miguel Angel Dumett. IV. Título.

CDD: 515.55

TERMO DE APROVAÇÃO

RODRIGO GARCIA EUSTÁQUIO

CLASSE DE MÉTODOS CHEBYSHEV-HALLEY INEXATA LIVRE DE TENSORES COM CONVERGÊNCIA CÚBICA PARA RESOLUÇÃO DE SISTEMAS NÃO LINEARES E UM ESTUDO SOBRE RAIO DE CONVERGÊNCIA

Tese de doutorado aprovada como requisito parcial para a obtenção do grau de Doutor em Ciências, no Programa de Pós-Graduação em Métodos Numéricos em Engenharia com área em Programação Matemática da Universidade Federal do Paraná, pela seguinte banca examinadora:

Prof. Dr. Ademir Alves Ribeiro
Departamento de Matemática - UFPR

Prof. Drª. Gislaine Aparecida Periçaro
Departamento de Matemática - UNESPAR

Prof. Dr. Roberto Andreani
Departamento de Matemática Aplicada - UNICAMP

Prof. Dr. Rodolfo Gotardi Begiato
Departamento de Matemática - UTFPR

Prof. Dr. Yuan Jin Yun
Departamento de Matemática - UFPR

Curitiba, 06 de dezembro de 2013.

Parece paradoxal que a pesquisa científica, em vários sentidos uma das mais questionadoras e céticas atividades humanas, dependam da confiança pessoal. Mas o fato é que, sem a confiança, a empreitada da pesquisa não funcionaria.

Arnold S. Relman.
New England Journal Medicines. 1983.

Para minha amada filha Mariana
e minha amada esposa Priscilla.

Agradecimentos

Ao ingressar no curso de doutorado, eu sabia que passaria por várias dificuldades, tanto pessoais como profissionais. No início, minha filha Mariana acabara de nascer e eu havia ingressado via concurso público na Universidade Tecnológica Federal do Paraná como professor. O nascimento de minha filha foi uma grande motivação para que eu continuasse em busca do meu objetivo.

Na escola onde minha filha estuda, eu ouvi diversas vezes alguns pais dizerem que gostariam de dar ao filho tudo que não haviam recebido de seus pais. Eu não tenho nada do que reclamar, eu tive tudo o que meus pais puderam me dar, honestidade, dignidade, respeito e vários outros valores e princípios que apenas os pais que amam seus filhos podem dar. Com esse sentimento, eu gostaria de agradecer algumas pessoas e instituições que me ajudaram a concluir esse trabalho.

Ao Prof. Dr. Ademir Alves Ribeiro e Prof. Dr. Miguel Angel Dumett Canales, pela orientação deste trabalho. Pelos seminários, troca de ideias, ensinamentos, correções e sugestões nas demonstrações dos teoremas e acima de tudo, por confiarem em mim e acreditarem no tema desta tese.

Ao programa de Pós-Graduação em Métodos Numéricos em Engenharia da Universidade Federal do Paraná e ao IMPA pelo financiamento para participar do *IX Brazilian Workshop on Continuous Optimization*.

Ao programa de Pós-Graduação em Matemática Aplicada da Universidade Federal do Paraná pelo financiamento para participar do *II Brazil-China Symposium on Applied and Computational Mathematics* e por me aceitar como aluno em algumas disciplinas e aos colegas Camila Isoton, Geovani Nunes Grapiglia e Adriano Rodrigo Delfino.

Em especial, gostaria de agradecer a Prof. Dr^{a} Elizabeth Wegner Karas e ao Prof. Dr. Ademir Alves Ribeiro pelos ensinamentos desde a época do mestrado, por todo o apoio tanto na minha vida pessoal como profissional e principalmente pela amizade. Meus agradecimentos e admiração.

A alguns professores pesquisadores, que por e-mail puderam esclarecer alguns questionamentos de seus artigos. Em particular, ao Prof. Dr. Hubert Schwetlick da *Dresden University of Technology*, Alemanha.

À Universidade Tecnológica Federal do Paraná, em especial, ao departamento de matemática, pelo afastamento concedido das atividades nesses últimos dois anos.

Aos professores que fizeram parte da banca examinadora. Obrigado Prof. Dr. Roberto Andreani, Prof^{a}. Dr^{a} Gislaine Aparecida Periçaro, Prof. Dr. Rodolfo Gotardi Begiato e

Prof. Dr. Yuan Jin Yun pela leitura da tese, críticas e sugestões que melhoraram bastante este trabalho.

Aos meus pais e ao meu irmão Fernando, pelo incentivo, pelas orações, por acreditarem em mim, por entenderem minha ausência e por sempre me ajudarem quando precisei. Muito obrigado por tudo. Amo vocês.

À minha querida esposa Priscilla, que aceitou esse desafio junto comigo, obrigado por aceitar que eu dividisse minhas frustrações, minhas dificuldades, minhas angústias e por sempre falar e acreditar que eu conseguiria. A conclusão desta tese só foi possível porque eu pude dividir tudo isso com você. Obrigado pelas palavras ditas na minha defesa de tese. Te amo.

À minha linda filha Mariana. Mesmo criança, compreendeu a necessidade de ausentar-me de algumas brincadeiras. Que meu esforço sirva como exemplo em sua vida.

A todos que participaram da minha defesa de tese.

À Deus por permitir que todas essas pessoas e outras, pudessem me ajudar.

Resumo

Esta tese introduz dois novos resultados sobre a Classe Chebyshev-Halley para resolução de sistemas não-lineares. Os métodos dessa classe possuem convergência cúbica, tendo portanto uma taxa de convergência superior a do método de Newton. Em contrapartida, esses métodos são mais caros computacionalmente, por necessitarem de derivadas de segunda ordem.

O primeiro resultado apresentado é um resultado teórico. Introduzimos um novo raio de convergência para a Classe Chebyshev-Halley, ou seja, mostramos que dado qualquer ponto inicial pertencente à uma bola centrada em uma solução com o novo raio, a sequência gerada por qualquer método da Classe Chebyshev-Halley é bem definida e converge para a respectiva solução com taxa de convergência cúbica. Comparamos com o raio utilizado na prova de convergência dada no livro *Numerische Lösung Nichtlinearer Gleichungen* [70] para os métodos Halley, Chebyshev e Super-Halley, através de alguns exemplos. As comparações apresentadas sugerem perpectivas futuras, tais como determinar o raio ótimo de convergência.

O segundo resultado apresentado é a introdução de uma nova classe de métodos, chamada Classe Chebyshev-Halley Inexata livre de tensores, cujo objetivo é baratear o custo computacional da Classe Chebyshev-Halley, no que tange o uso da derivada de segunda ordem e a resolução de dois sistemas lineares. A grosso modo, não utilizamos informações de derivada de segunda ordem e os dois sistemas lineares, necessários para a obtenção do passo, podem ser resolvidos de maneira inexata. Além de apresentar a prova de convergência, mostramos que, dependendo das hipóteses, os métodos dessa classe podem ter taxa de convergência superlinear, quadrática, superquadrática e cúbica. Mostramos também que essas hipóteses são bastante razoáveis.

Por fim, comparações numéricas são apresentadas, mostrando uma melhoria significativa quando se usa a estratégia inexata livre de tensores, proposta nesta tese, nos métodos clássicos da Classe Chebyshev-Halley.

Palavras-chave: Classe Chebyshev-Halley Inexata livre de tensores, raio de convergência, taxa de convergência, sistemas não-lineares.

Abstract

This thesis introduces two new results about the Chebyshev-Halley Class for solving nonlinear systems. The methods in this class have third-order rate of convergence, which means they have a better rate of convergence than Newton's method. In contrast, these methods are computationally expensive, requiring second-order derivatives.

The first result presented is a theoretical result. We introduce a new convergence radius for the Chebyshev-Halley Class, that is, we proved that given any starting point belonging to a ball centered at a solution with the new radius, the sequence generated by any method in the Chebyshev-Halley Class is well defined and converges to that solution with cubic convergence rate. We compared the new radius with the one given in the book *Numerische Lösung Nichtlinearer Gleichungen* [70] for Halley, Super-Halley and Chebyshev methods, using some examples. The comparisons suggest future perspectives, such as determining the optimal radius of convergence.

The second result presented is the introduction of a new class of methods, called Inexact Chebyshev-Halley tensor free Class, whose goal is to reduce the computational cost of the Chebyshev-Halley Class, by not computing the second-order derivatives and by approximately solving two linear systems required for obtaining the necessary intermediate computations. Besides presenting the proof of convergence, we show that, depending on the assumptions, the methods of this class can have superlinear, quadratic, superquadratic and cubic convergence rates. We also show that these assumptions are quite reasonable.

Finally, numerical evidence that shows significant improvement when utilizing the inexact tensor free strategy (in the context of the classical methods of Chebyshev-Halley class) proposed in this thesis is presented.

Key-words: Inexact Chebyshev-Halley tensor free Class, convergence radius, convergence rates, nonlinear systems.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

Introdução

Muitas aplicações de modelagem matemática no mundo real consistem em resolver um sistema de equações, geralmente não lineares. Um sistema de equações não lineares pode ser escrito como $F(x) = 0$, onde F é uma aplicação de $\mathbb{R}^n$ em $\mathbb{R}^m$.

Nesta tese vamos considerar uma aplicação $F : \mathbb{R}^n \to \mathbb{R}^n$ duas vezes continuamente diferenciável cujo objetivo é

$$\text{encontrar um } x^* \in \mathbb{R}^n \text{ tal que } F(x^*) = 0.$$

Os métodos mais utilizados para resolver este problema são os métodos iterativos, pois em geral não é possível encontrar uma solução explícita por meios algébricos. Além disso, existem métodos não iterativos, por exemplo [11].

Dentre os métodos iterativos, podemos destacar o método de Newton. Neste método é resolvido um sistema linear a cada iteração, cuja matriz dos coeficientes é a jacobiana de F avaliada no iterado corrente. Uma das vantagens desse método é a taxa de convergência quadrática (sob condições adequadas). Além disso, é conhecido na literatura o raio ótimo de convergência do método de Newton. Isto significa que, dada uma sequência gerada pelo método de Newton cujo ponto inicial esteja fora da bola de centro em uma solução e raio ótimo, não se tem garantias que esta sequência irá convergir para a respectiva solução. No entanto, tomado qualquer ponto inicial dentro desta bola, não só é garantida a convergência, mas também a taxa de convergência quadrática.

Já nos métodos quase-Newton, não é necessário o uso da jacobiana de F. Esta metodologia é, em termos de número de operações computacionais, mais barata que o método de Newton. Uma contrapartida é a perda da taxa de convergência. Sob hipóteses razoáveis, esses métodos convergem superlinearmente.

Existem métodos que possuem taxa de convergência cúbica, sendo melhores que o método de Newton neste aspecto. Por exemplo, os métodos pertencentes à classe Chebyshev-Halley. Nestes métodos, dada uma estimativa inicial $x^0 \in \mathbb{R}^n$, o próximo iterado é obtido pelo processo iterativo

$$x^{k+1} = x^k - \Big[I + \frac{1}{2}\mathcal{L}(x^k)\Big(I - \alpha\mathcal{L}(x^k)\Big)^{-1}\Big] J_F(x^k)^{-1}F(x^k),$$

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

$$\mathcal{L}(x) = J_F(x)^{-1}\mathcal{T}_F(x)\Big(J_F(x)^{-1}F(x)\Big).$$

O parâmetro α é um número real que indica um método da classe.

Além do cálculo do tensor $\mathcal{T}_F(x)$ ser caro computacionalmente, para obter a matriz $\mathcal{L}(x)$ é necessário resolver $n+1$ sistemas lineares, o que torna os métodos dessa classe impraticáveis. No entanto, foi provado recentemente por Gundersen e Steihaug [37], que para obter o próximo iterado da classe Chebyshev-Halley, basta resolver apenas os dois sistemas

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) \\ \Big(J_F(x^k) + \alpha\mathcal{T}_F(x^k)s^k_{(1)}\Big)s^k_{(2)} &= -\frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)}, \end{aligned} \tag{1}$$

e tomar $x^{k+1} = x^k + s^k_{(1)} + s^k_{(2)}$.

Com esta redução do custo computacional, esta classe de métodos tem sido bastante estudada por alguns pesquisadores. Alguns com o objetivo de resolver problemas de otimização irrestrita, veja por exemplo a referência [38]. Mesmo com esta redução, ainda é necessário o cálculo do tensor, o que demanda um certo esforço computacional.

Em relação aos métodos que utilizam tensores, podemos considerar aqueles que utilizam o modelo quadrático de F em torno de x^k, a saber

$$M_k(s) = F(x^k) + J_F(x^k)s + \frac{1}{2}\mathcal{T}_F(x^k)ss.$$

Duas estratégias considerando este modelo serão apresentadas nesta tese. Uma é dada por Schnabel e Frank [68], que consideraram uma aproximação de posto baixo do tensor $\mathcal{T}_F(x^k)$ e procuraram minimizar a norma de um novo modelo quadrático. Com o avanço das técnicas de otimização e de novos métodos para resolução de sistemas lineares, vários trabalhos têm utilizado técnicas diferentes para minimizar o modelo tensorial proposto por Schnabel e Frank. Veja por exemplo, as recentes referências [4, 6, 7]. A outra estratégia foi publicada em maio de 2013 por Steihaug e Suleiman [73]. Eles procuram encontrar um passo s^k de tal maneira que tenha uma redução do modelo quadrático, ou seja,

$$\left\|\frac{1}{2}\mathcal{T}_F(x^k)s^k s^k + J_F(x^k)s^k + F(x^k)\right\| \leq \eta_k \left\|F(x^k)\right\|$$

para algum $\eta_k \in (0,1)$. Além disso, eles introduziram uma classe de métodos chamada Classe Chebyshev-Halley Inexata para determinar um s^k e um $\eta_k \in (0,1)$ que cumpram tal requisito.

Nesta tese, motivados pelo raio ótimo de convergência do método de Newton, propomos um raio r de convergência para a Classe Chebyshev-Halley. Isto significa que, dada

uma sequência gerada por qualquer método da Classe Chebyshev-Halley, se o ponto inicial estiver na bola de centro em uma solução e raio r, então a sequência converge para a respectiva solução com taxa de convergência cúbica. Além disso, propomos uma nova classe de métodos chamada Classe Chebyshev-Halley Inexata livre de tensores, na qual não utilizamos qualquer informação sobre a segunda derivada da aplicação F, e os dois sistemas lineares necessários para a obtenção do passo, podem ser resolvidos de maneira inexata. A grosso modo, modificamos os dois sistemas lineares (1) onde o produto $\mathcal{T}_F(x^k)s^k_{(1)}$ é substituído por uma matriz que satisfaz uma propriedade e os dois sistemas lineares podem ser resolvidos de maneira inexata. Além disso, mostramos que, dependendo das hipóteses, os métodos dessa classe podem ter taxa de convergência superlinear, quadrática, superquadrática e cúbica. Mostramos também que essas hipóteses são bastante razoáveis.

Este trabalho está organizado da seguinte maneira:

- no Capítulo 1, introduzimos o conceito de tensor de um modo geral. Mostramos que para cada aplicação bilinear, existe um tensor associado. Além disso, estudamos a segunda derivada de uma aplicação $F : \mathbb{R}^n \to \mathbb{R}^m$ como um tensor e em particular, utilizamos as operações elementares de tensores.

- o Capítulo 2 é dedicado ao estudo dos métodos Halley, Chebyshev e Super-Halley unidimensionais, tanto no contexto algébrico como geométrico. Tal estudo é indicado, pois não são tão conhecidos na literatura como o método de Newton.

- no Capítulo 3 são apresentados, em particular, alguns métodos tensoriais conhecidos na literatura, como o proposto por Schnabel e Frank [68], a Classe Chebyshev-Halley e a Classe Chebyshev-Halley Inexata proposta por Steihaug e Suleiman [73].

- no Capítulo 4, apresentamos as contribuições teóricas desta tese. Introduzimos um raio de convergência cúbica e demonstramos um teorema de convergência. Além disso, provamos a convergência da Classe Chebyshev-Halley Inexata livre de tensores proposta nesta tese.

- no Capítulo 5 são realizados experimentos numéricos. Para alguns desses experimentos, modificamos um algoritmo proposto por Steihaug e Suleiman [73] com o objetivo de usar a estratégia livre de tensor. Além desse algoritmo modificado, foi utilizado também o método GMRES.

Capítulo 1

Preliminares

É comum em livros clássicos de análise, estudar a segunda derivada de uma aplicação $F : \mathbb{R}^n \to \mathbb{R}^m$ como uma aplicação bilinear. O propósito deste capítulo é estudar a segunda derivada de uma aplicação $F : \mathbb{R}^n \to \mathbb{R}^m$ como um tensor. Para isso, é necessário entender algumas de suas operações elementares. O autor julga essencial essa preparação, pois não foi tão trivial entender algumas generalizações que serão expostas mais adiante.

A primeira seção deste capítulo apresenta alguns resultados básicos sobre matrizes. A segunda seção trata de um pequeno estudo sobre tensores e algumas de suas propriedades. Na terceira seção são apresentadas algumas propriedades sobre aplicações bilineares e a quarta seção é destinada ao estudo de diferenciabilidade de aplicações. Relacionamos a segunda derivada de uma aplicação com uma aplicação bilinear e então com um tensor de ordem 3.

1.1 Alguns Resultados sobre Matrizes

Os resultados apresentados nesta seção são resultados clássicos na literatura [35, 48].

Lema 1.1 (Lema de Banach) *Se $A \in \mathbb{R}^{n\times n}$ e $\|A\|_p < 1$, então $I - A$ é não singular e*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$$

com

$$\left\|(I - A)^{-1}\right\|_p \leq \frac{1}{1 - \|A\|_p}. \tag{1.1}$$

Prova. Suponha por absurdo que $I-A$ seja singular. Logo existe $\hat{x} \neq 0$ tal que $(I-A)\hat{x} = 0$. Desta forma temos

$$\|A\|_p \geq \frac{\|A\hat{x}\|_p}{\|\hat{x}\|_p} = 1,$$

contradizendo a hipótese. Portanto, $I - A$ é não singular. Observe agora que

$$\Big(\sum_{k=0}^{N} A^k\Big)(I - A) = I - A^{N+1}.$$

Como $\|A\|_p < 1$ e tendo em vista que $\|A^k\|_p \leq \|A\|_p^k$ tem-se que $\lim_{k\to\infty} A^k = 0$. Logo

$$\Big(\lim_{N\to\infty} \sum_{k=0}^{N} A^k\Big)(I - A) = I$$

e portanto $(I - A)^{-1} = \lim_{N\to\infty} \sum_{k=0}^{N} A^k$. Como consequência vemos que

$$\|(I - A)^{-1}\|_p \leq \sum_{k=0}^{\infty} \|A\|_p^k = \frac{1}{1 - \|A\|_p}$$

□

Lema 1.2 *Se A e B são matrizes tais que $\|I - BA\| < 1$, então A e B são não singulares e*

$$\|A^{-1}\| \leq \frac{\|B\|}{1 - \|I - BA\|} \quad e \quad \|B^{-1}\| \leq \frac{\|A\|}{1 - \|I - BA\|}. \tag{1.2}$$

Prova. Seja $M = I - BA$. Pelo Lema 1.1, $I - M = BA$ é não singular. Logo A e B são não singulares. Como $A^{-1} = (BA)^{-1}B$, novamente pelo Lema 1.1, temos que

$$\|A^{-1}\| \leq \|(BA)^{-1}\| \, \|B\| \leq \frac{\|B\|}{1 - \|I - BA\|}.$$

A outra desigualdade se prova de modo análogo. □

Lema 1.3 *Seja A uma matriz $n \times n$. Se $I - A$ é não singular, então*

$$A(I - A)^{-1} = (I - A)^{-1}A. \tag{1.3}$$

Prova. Observe que

$$(I - A)A = A - AA = A(I - A).$$

Daí segue que $A = (I - A)^{-1}A(I - A)$ resultando na equação (1.3). □

1.2 Tensores

Tensores surgem naturalmente em algumas aplicações, tais como quimiometria [72], processamento de sinais [14] e outros. De acordo com [52], para muitas aplicações envol-

vendo tensores de alta ordem, os resultados conhecidos de álgebra matricial pareciam, no século XX, ser insuficientes. Alguns *workshops* e congressos sobre o estudo de tensores têm sido realizados, como por exemplo:

- *Workshop on Tensor Decomposition at the American Institute of Mathematics in Palo Alto, California* em 2004, organizado por Golub, Kolda, Nagy e Van Loan. Detalhes em [34];
- *Workshop on Tensor Decompositions and Applications* em 2005, organizado por Comon e De Lathauwer. Detalhes em [51];
- *Minisymposium on Numerical Multilinear Algebra: A New Beginning* em 2007, organizado por Golub, Comon, De Lathauwer e Lim e realizado em Zurich.

Leitores interessados em decomposição em valores singulares, posto, autovalores e autovetores, bem como outros assuntos de tensores de alta ordem, podem consultar as referências [5, 6, 16, 46, 50, 52]. Para esta tese interessam os tensores de ordem no máximo 3.

Assim, sejam I_1, I_2 e I_3 três números inteiros positivos. Um tensor $\mathcal{T}$ de ordem 3 é uma lista de números $t_{i_1 i_2}^{i_3}$ com $i_1 = 1, \ldots, I_1$, $i_2 = 1, \ldots, I_2$ e $i_3 = 1, \ldots, I_3$ e a n-ésima dimensão do tensor $\mathcal{T}$ é I_n, para $n = 1, 2, 3$. Para exemplificar, a primeira, segunda e terceira dimensões de um tensor $\mathcal{T} \in \mathbb{R}^{2\times 4\times 3}$ são $2, 4, 3$, respectivamente.

Evidentemente, tensores são generalizações de matrizes, ou seja, uma matriz $m \times n$ pode ser vista como um tensor de ordem 2, enquanto que um vetor n-dimensional pode ser visto como um tensor de ordem 1. Dependendo do contexto, um vetor n-dimensional pode ser visto como uma matriz $n \times 1$ e, uma matriz $m \times n$ pode ser vista como um tensor $m \times n \times 1$.

Do ponto de vista algébrico, um tensor $\mathcal{T}$ de ordem 3 é um elemento do espaço vetorial $\mathbb{R}^{I_1 \times I_2 \times I_3}$, enquanto que do ponto de vista geométrico, um tensor $\mathcal{T}$ de ordem 3 pode ser visto como um paralelepípedo [49], com I_1 linhas, I_2 colunas e I_3 tubos. A Figura 1.1 ilustra um tensor $\mathcal{T} \in \mathbb{R}^{2\times 4\times 3}$.

Em álgebra linear, é comum olhar uma matriz através de suas colunas. Se $A \in \mathbb{R}^{m\times n}$, então A pode ser vista como $A = [a_1 \ldots a_n]$, onde $a_j \in \mathbb{R}^m$ representa a j-ésima coluna da matriz A. No caso de tensores de ordem 3, podemos olhá-los através de fibras e camadas. Daí seguem as definições.

Definição 1.4 *Uma fibra de um tensor $\mathcal{T}$ de ordem 3 é um tensor de ordem 1, obtido fixando dois índices.*

Definição 1.5 *Uma camada de um tensor $\mathcal{T}$ de ordem 3 é um tensor de ordem 2, obtido fixando apenas um índice.*

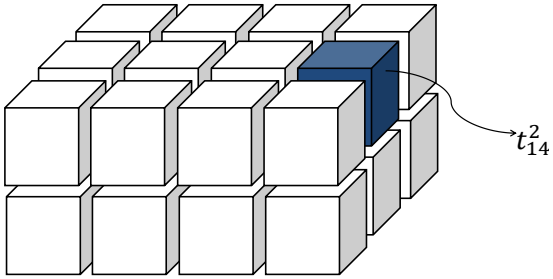

Figura 1.1: Um tensor $\mathcal{T} \in \mathbb{R}^{2\times4\times3}$

A grosso modo, em tensores de ordem 3, uma fibra é um vetor e uma camada é uma matriz. Temos três tipos de fibras:

- fibras colunas (ou fibras modo 1), onde são fixados os índices i_2 e i_3,
- fibras linhas (ou fibras modo 2), onde são fixados os índices i_1 e i_3 e
- fibras tubos (ou fibras modo 3), onde são fixados os índices i_1 e i_2,

e três tipos de camadas:

- camadas horizontais, onde é fixado o índice i_1,
- camadas laterais, onde é fixado o índice i_2 e
- camadas frontais, onde é fixado o índice i_3.

Para exemplificar, considere um tensor $\mathcal{T} \in \mathbb{R}^{2\times4\times3}$ com $i = 1, 2$, $j = 1, 2, 3, 4$ e $k = 1, 2, 3$. A i-ésima camada horizontal, denotada por $\mathcal{T}^{i::}$, é a matriz

$$\mathcal{T}^{i::} = \begin{pmatrix} t^1_{i1} & t^2_{i1} & t^3_{i1} \\ t^1_{i2} & t^2_{i2} & t^3_{i2} \\ t^1_{i3} & t^2_{i3} & t^3_{i3} \\ t^1_{i4} & t^2_{i4} & t^3_{i4} \end{pmatrix},$$

a j-ésima camada lateral, denotada por $\mathcal{T}^{:j:}$, é a matriz

$$\mathcal{T}^{:j:} = \begin{pmatrix} t^1_{1j} & t^2_{1j} & t^3_{1j} \\ t^1_{2j} & t^2_{2j} & t^3_{2j} \end{pmatrix}$$

e a k-ésima camada frontal, denotada por $\mathcal{T}^{::k}$, é a matriz

$$\mathcal{T}^{::k} = \begin{pmatrix} t^k_{11} & t^k_{12} & t^k_{13} & t^k_{14} \\ t^k_{21} & t^k_{22} & t^k_{23} & t^k_{24} \end{pmatrix}. \tag{1.4}$$

As Figuras 1.2 e 1.3 ilustram os três tipos de fibras e camadas, respectivamente, para um tensor $\mathcal{T} \in \mathbb{R}^{2\times4\times3}$.

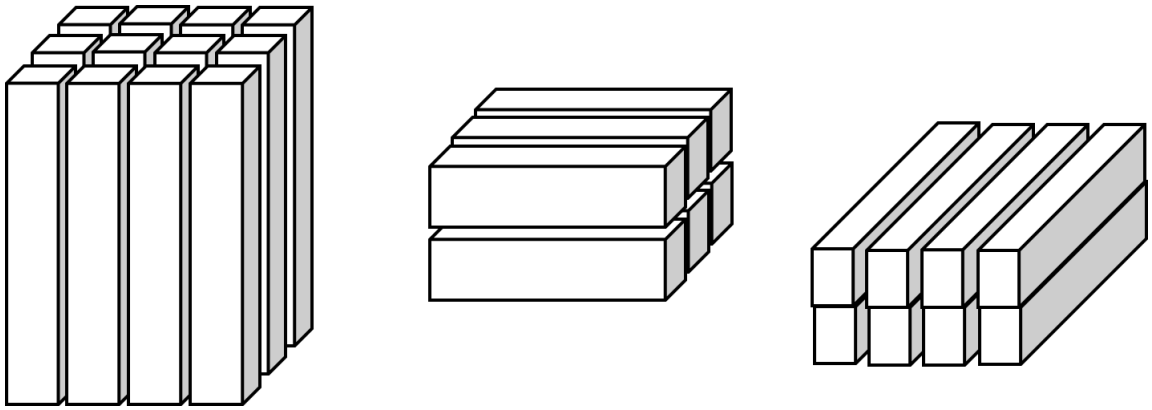

Figura 1.2: Fibras colunas, linhas e tubos, respectivamente.

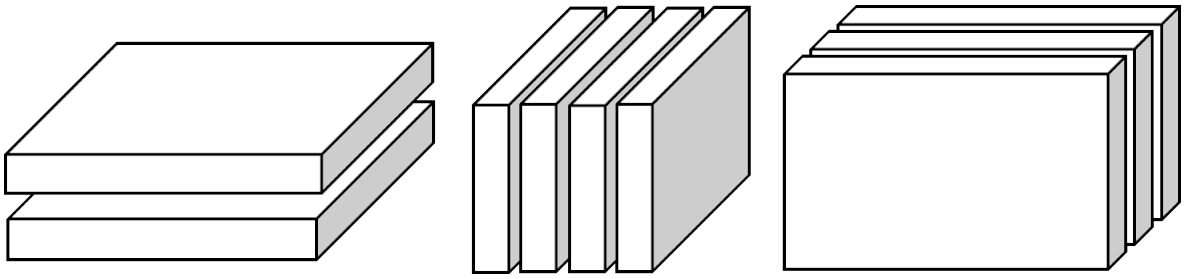

Figura 1.3: Camadas horizontais, laterais e frontais, respectivamente.

1.2.1 Operações com Tensores

A primeira questão a considerar nesta subseção é como efetuar o produto entre tensores e matrizes. Sabemos da álgebra elementar que dadas as matrizes $A \in \mathbb{R}^{m\times n}$ e $B \in \mathbb{R}^{R\times m}$, é possível efetuar o produto BA, pois a primeira dimensão (número de linhas) da matriz A concorda com a segunda dimensão (número de colunas) da matriz B, e cada elemento do produto é resultado do produto interno entre linhas da matriz B e colunas da matriz A.

Como tensores de ordem 3 têm três dimensões (não confundir com a dimensão do espaço vetorial $\mathbb{R}^{I_1\times I_2\times I_3}$), o produto entre tensores de ordem 3 e matrizes ou vetores é um pouco mais complicado. Para obter um elemento do produto entre um tensor e uma matriz, é necessário especificar qual a dimensão do tensor será tomada de modo

a concordar com o número de colunas da matriz, e cada elemento do produto será o resultado do produto interno entre as fibras modo n (coluna, linha ou tubo) e as colunas da matriz. Usaremos a solução adotada por [52], que define o produto *modo-n* entre tensores e matrizes e a solução adotada por [5] que define o produto *modo-n contraído* entre tensores e vetores.

O produto modo-n é útil quando se quer decompor em valores singulares um tensor de alta ordem no sentido de evitar o uso do conceito de transpostas generalizadas. Veja [5, 6, 50, 52].

Definição 1.6 (Produto modo-n entre Tensor e Matriz) *O produto modo-*1 *de um tensor* $\mathcal{T} \in \mathbb{R}^{m\times n\times p}$ *por uma matriz* $A \in \mathbb{R}^{R\times m}$ *é o tensor*

$$\mathcal{Y} = \mathcal{T} \times_1 A \in \mathbb{R}^{R\times n\times p}$$

onde seus elementos são definidos por

$$y^k_{rj} = \sum_{i=1}^{m} t^k_{ij} a_{ri} \quad onde \quad r = 1, \ldots, R,\ j = 1, \ldots, n,\ e\ k = 1, \ldots, p.$$

*O produto modo-*2 *de um tensor* $\mathcal{T} \in \mathbb{R}^{m\times n\times p}$ *por uma matriz* $A \in \mathbb{R}^{R\times n}$ *é o tensor*

$$\mathcal{Y} = \mathcal{T} \times_2 A \in \mathbb{R}^{m\times R\times p}$$

onde seus elementos são definidos por

$$y^k_{ir} = \sum_{j=1}^{n} t^k_{ij} a_{rj} \quad onde \quad i = 1, \ldots, m,\ r = 1, \ldots, R\ e\ k = 1, \ldots, p.$$

*O produto modo-*3 *de um tensor* $\mathcal{T} \in \mathbb{R}^{m\times n\times p}$ *por uma matriz* $A \in \mathbb{R}^{R\times p}$ *é o tensor*

$$\mathcal{Y} = \mathcal{T} \times_3 A \in \mathbb{R}^{m\times n\times R}$$

onde seus elementos são definidos por

$$y^r_{ij} = \sum_{k=1}^{p} t^k_{ij} a_{rk} \quad onde \quad i = 1, \ldots, m,\ j = 1, \ldots, n\ e\ r = 1, \ldots, R.$$

Para entender o produto modo-n em termos de matrizes, considere as matrizes $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{k\times m}$ e $C \in \mathbb{R}^{q\times n}$. De acordo com a Definição 1.6 temos

$$A \times_1 B = BA \in \mathbb{R}^{k\times n} \ \text{ e } \ A \times_2 C = AC^T \in \mathbb{R}^{m\times q}.$$

Desta forma, a decomposição em valores singulares de uma matriz A pode ser escrita como

$$U\Sigma V^T = (\Sigma \times_1 U) \times_2 V = (\Sigma \times_2 V) \times_1 U.$$

O produto modo-n satisfaz a seguinte propriedade [52]:

Propriedade 1 *Dados um tensor $\mathcal{T}$ de ordem 3 e matrizes A e B de tamanhos convenientes, temos para todo $r, s = 1, 2, 3$ que*

$$\begin{aligned}(\mathcal{T} \times_r A) \times_s B &= (\mathcal{T} \times_s B) \times_r A = \mathcal{T} \times_r A \times_s B \quad \text{para } r \neq s \text{ e} \\ (\mathcal{T} \times_r A) \times_r B &= \mathcal{T} \times_r (BA)\end{aligned}$$

A idéia de Bader e Kolda [5] para efetuar o produto entre tensor e vetor, é não usar a dimensão unitária como de costume. Simplesmente calcula-se o produto interno de cada fibra modo n (coluna, linha ou tubo) com o vetor. Não é vantajoso tratar um vetor m-dimensional como uma matriz $m \times 1$. Por exemplo, se tomarmos um tensor $\mathcal{T} \in \mathbb{R}^{m \times n \times p}$ e um vetor $v \in \mathbb{R}^{m \times 1}$, com $m, n, p \neq 1$, de acordo com a Definição 1.6, o produto entre o $\mathcal{T}$ e v não é bem definido, mas é possível efetuar o cálculo $\mathcal{T} \times_1 v^T$.

Definição 1.7 (Produto modo-n contraído entre Tensor e Vetor) *O produto modo-1 contraído de um tensor $\mathcal{T} \in \mathbb{R}^{m \times n \times p}$ por um vetor $v \in \mathbb{R}^m$, é o tensor*

$$A = \mathcal{T} \bar{\times}_1 v \in \mathbb{R}^{n \times p}$$

onde seus elementos são definidos por

$$a_{jk} = \sum_{i=1}^{m} t_{ij}^k v_i \quad \text{onde} \quad j = 1, \ldots, n \text{ e } k = 1, \ldots, p$$

onde v_i é a i-ésima coordenada do vetor v.

O produto modo-2 contraído de um tensor $\mathcal{T} \in \mathbb{R}^{m \times n \times p}$ por um vetor $v \in \mathbb{R}^n$, é o tensor

$$A = \mathcal{T} \bar{\times}_2 v \in \mathbb{R}^{m \times p}$$

onde seus elementos são definidos por

$$a_{ik} = \sum_{j=1}^{n} t_{ij}^k v_j \quad \text{onde} \quad i = 1, \ldots, m \text{ e } k = 1, \ldots, p$$

onde v_j é a j-ésima coordenada do vetor v.

O produto modo-3 contraído de um tensor $\mathcal{T} \in \mathbb{R}^{m \times n \times p}$ por um vetor $v \in \mathbb{R}^p$, é o

tensor

$$A = \mathcal{T} \bar{\times}_3 v \in \mathbb{R}^{m \times n}$$

onde seus elementos são definidos por

$$a_{ij} = \sum_{k=1}^{p} t_{ij}^k v_k \quad onde \quad i = 1, \ldots, m \ e \ j = 1, \ldots, n$$

onde v_k é a k-ésima coordenada do vetor v.

Devemos ter um enorme cuidado ao efetuar o produto entre matrizes e vetores considerando as Definições 1.6 e 1.7. Por exemplo, note que se $A \in \mathbb{R}^{m \times n}$, $u \in \mathbb{R}^n$ e $v \in \mathbb{R}^m$, então $A \bar{\times}_2 u$ e $A \times_2 u^T$ possuem os mesmos elementos, mas

$$A \bar{\times}_2 u \neq A \times_2 u^T,$$

pois $A \bar{\times}_2 u \in \mathbb{R}^m$ (vetor coluna) e $A \times_2 u^T \in \mathbb{R}^{1 \times m}$ (vetor linha). Note que, em relação ao produto matricial, no qual estamos acostumados, temos

$$Au = A \bar{\times}_2 u \tag{1.5}$$

$$v^T A = A \times_1 v^T \neq A \bar{\times}_1 v. \tag{1.6}$$

Em particular, dados um tensor $\mathcal{T} \in \mathbb{R}^{n \times m \times m}$ e um vetor $v \in \mathbb{R}^m$, pela Definição 1.7 e por (1.5) temos $\mathcal{T} \bar{\times}_2 v \in \mathbb{R}^{n \times m}$ e

$$(\mathcal{T} \bar{\times}_2 v) \bar{\times}_2 v = (\mathcal{T} \bar{\times}_2 v) v \in \mathbb{R}^n.$$

O produto modo-n contraído satisfaz a seguinte propriedade [5]:

Propriedade 2 *Dados um tensor $\mathcal{T}$ de ordem 3 e vetores u e v de tamanhos convenientes, temos para todo $r = 1, 2, 3$ e $s = 2, 3$ que*

$$(\mathcal{T} \bar{\times}_r u) \bar{\times}_{s-1} v = (\mathcal{T} \bar{\times}_s v) \bar{\times}_r u \quad para \ r < s.$$

Para exemplificar, considere um tensor $\mathcal{T} \in \mathbb{R}^{2 \times 4 \times 3}$ e denote a k-ésima coluna e a q-ésima linha de uma matriz A por $\text{col}_k(A)$ e $\text{lin}_q(A)$, respectivamente . Note que se

1. $x \in \mathbb{R}^2$ então $\mathcal{T} \bar{\times}_1 x \in \mathbb{R}^{4\times 3}$ e

$$\mathrm{col}_k(\mathcal{T} \bar{\times}_1 x) = \begin{pmatrix} a_{1k} \\ a_{2k} \\ a_{3k} \\ a_{4k} \end{pmatrix} = \begin{pmatrix} t^k_{11} & t^k_{21} \\ t^k_{12} & t^k_{22} \\ t^k_{13} & t^k_{23} \\ t^k_{14} & t^k_{24} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \left(\mathcal{T}^{::k}\right)^T x \text{ e}$$

$$\mathrm{lin}_j(\mathcal{T} \bar{\times}_1 x) = \begin{pmatrix} a_{j1} & a_{j2} & a_{j3} \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} t^1_{1j} & t^2_{1j} & t^3_{11} \\ t^1_{2j} & t^2_{2j} & t^3_{21} \end{pmatrix} = x^T \mathcal{T}^{:j:}$$

2. $x \in \mathbb{R}^4$ então $\mathcal{T} \bar{\times}_2 x \in \mathbb{R}^{2\times 3}$ e

$$\mathrm{col}_k(\mathcal{T} \bar{\times}_2 x) = \begin{pmatrix} a_{1k} \\ a_{2k} \end{pmatrix} = \begin{pmatrix} t^k_{11} & t^k_{12} & t^k_{13} & t^k_{14} \\ t^k_{21} & t^k_{22} & t^k_{23} & t^k_{24} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \left(\mathcal{T}^{::k}\right) x \text{ e}$$

$$\mathrm{lin}_i(\mathcal{T} \bar{\times}_2 x) = \begin{pmatrix} a_{i1} & a_{i2} & a_{i3} \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \end{pmatrix} \begin{pmatrix} t^1_{i1} & t^2_{i1} & t^3_{i1} \\ t^1_{i2} & t^2_{i2} & t^3_{i2} \\ t^1_{i3} & t^2_{i3} & t^3_{i3} \\ t^1_{i4} & t^2_{i4} & t^3_{i4} \end{pmatrix} = x^T \mathcal{T}^{i::}$$

3. $x \in \mathbb{R}^3$ então $\mathcal{T} \bar{\times}_3 x \in \mathbb{R}^{2\times 4}$ e

$$\mathrm{col}_j(\mathcal{T} \bar{\times}_3 x) = \begin{pmatrix} a_{1j} \\ a_{2j} \end{pmatrix} = \begin{pmatrix} t^1_{1j} & t^2_{1j} & t^3_{1j} \\ t^1_{2j} & t^2_{2j} & t^3_{2j} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \left(\mathcal{T}^{:j:}\right) x \text{ e}$$

$$\mathrm{lin}_i(\mathcal{T} \bar{\times}_3 x) = \begin{pmatrix} a_{i1} & a_{i2} & a_{i3} \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} t^1_{i1} & t^1_{i2} & t^1_{i3} & t^1_{i4} \\ t^2_{i1} & t^2_{i2} & t^2_{i3} & t^2_{i4} \\ t^3_{i1} & t^3_{i2} & t^3_{i3} & t^3_{i4} \end{pmatrix} = x^T \left(\mathcal{T}^{i::}\right)^T$$

Este exemplo pode ser facilmente generalizado para dimensões arbitrárias. Em particular, para um tensor $\mathcal{T} \in \mathbb{R}^{m\times n\times n}$ e um vetor $x \in \mathbb{R}^n$, temos

$$\mathrm{lin}_i(\mathcal{T} \bar{\times}_2 x) = x^T \mathcal{T}^{i::} \tag{1.7}$$

$$\mathrm{lin}_i(\mathcal{T} \bar{\times}_3 x) = x^T (\mathcal{T}^{i::})^T \tag{1.8}$$

Lema 1.8 *Seja um tensor $\mathcal{T} \in \mathbb{R}^{n\times n\times n}$. Se $\mathcal{T}^{i::}$ é uma matriz simétrica para todo $i = 1,\ldots,n$, então*

$$(\mathcal{T}\,\bar{\times}_2 u)v = (\mathcal{T}\,\bar{\times}_2 v)u$$

para todo $u, v \in \mathbb{R}^n$.

Prova. Pela Propriedade 2 temos que $(\mathcal{T}\,\bar{\times}_2 u)v = (\mathcal{T}\,\bar{\times}_3 v)u$. Por (1.7), (1.8) e pela simetria de $\mathcal{T}^{i::}$ vemos que $\mathcal{T}\,\bar{\times}_3 v = \mathcal{T}\,\bar{\times}_2 v$. □

1.3 O Espaço das Aplicações Bilineares

Nesta seção, definimos aplicações bilineares sobre espaços vetoriais de dimensão finita, visando relacioná-las com a segunda derivada de uma aplicação duas vezes diferenciável, bem como um tensor de ordem 3.

Definição 1.9 *Sejam espaços vetoriais U, V e W. Uma aplicação $f : U \times V \to W$ é uma aplicação bilinear se*

(i) $f(\lambda u_1 + u_2, v) = \lambda f(u_1, v) + f(u_2, v)$ *para todo* $\lambda \in \mathbb{R}$, $u_1, u_2 \in U$ *e* $v \in V$.

(ii) $f(u, \lambda v_1 + v_2) = \lambda f(u, v_1) + f(u, v_2)$ *para todo* $\lambda \in \mathbb{R}$, $u \in U$ *e* $v_1, v_2 \in V$.

Em outras palavras, uma aplicação $f : U \times V \to W$ é uma aplicação bilinear se for linear em cada uma das variáveis quando deixamos a outra fixa. Denotamos por $\mathcal{B}(U \times V, W)$ o conjunto de todas as aplicações bilineares de $U \times V$ em W. Em particular, se $U = V$ e $W = \mathbb{R}$ na Definição 1.9, então $f : U \times U \to \mathbb{R}$ é uma forma bilinear na qual estamos acostumados em formas quadráticas, por exemplo.

Um exemplo simples de forma bilinear é a função $f : U \times V \to \mathbb{R}$ definida por

$$f(u, v) = h(u)g(v), \tag{1.9}$$

com $h \in U^*$ e $g \in V^*$, onde U^* denota o espaço dual a U. De fato, temos para todo $\lambda \in \mathbb{R}, u_1, u_2 \in U$ e $v \in V$ que

$$f(\lambda u_1 + u_2, v) = h(\lambda u_1 + u_2)g(v) = \Big(\lambda h(u_1) + h(u_2)\Big)g(v) = \lambda f(u_1, v) + f(u_2, v).$$

De forma análoga, é fácil ver que $f(u, \lambda v_1 + v_2) = \lambda f(u, v_1) + f(u, v_2)$ para todo $\lambda \in \mathbb{R}, u \in U$ e $v_1, v_2 \in V$.

O próximo teorema basicamente garante que uma aplicação bilinear $f : U \times V \to W$ fica bem determinada quando se conhece seu aplicado nos pares cujas coordenadas são elementos de uma base de U e V.

Teorema 1.10 *Sejam U, V e W espaços vetoriais, $\{u_1, \ldots, u_m\}$, $\{v_1, \ldots, v_n\}$ bases de U e V, respectivamente e $\{w_{ij} \mid i = 1, \ldots, m \text{ e } j = 1, \ldots, n\}$ um subconjunto de W. Então existe uma única aplicação bilinear $f : U \times V \to W$ tal que $f(u_i, v_j) = w_{ij}$.*

Prova. Sejam $u = \sum_{i=1}^{m} \alpha_i u_i$ e $v = \sum_{j=1}^{n} \beta_j v_j$ elementos arbitrários de U e V, respectivamente. Definimos uma aplicação $f : U \times V \to W$ como sendo

$$f(u, v) = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j w_{ij}.$$

É fácil ver que f é uma aplicação bilinear e que $f(u_i, v_j) = w_{ij}$. Tal aplicação é única, pois se g é uma outra aplicação bilinear satisfazendo $g(u_i, v_j) = w_{ij}$ então

$$\begin{aligned} g(u, v) &= g\Big(\sum_{i=1}^{m} \alpha_i u_i, \sum_{j=1}^{n} \beta_j v_j\Big) = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j g(u_i, v_j) = \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j w_{ij} = f(u, v). \end{aligned}$$

Logo $g = f$. □

O teorema seguinte garante o isomorfismo entre o espaço das aplicações bilineares e o espaço dos tensores de ordem 3.

Teorema 1.11 *Sejam U, V e W espaços vetoriais com dimensões n, p e m respectivamente. Então o espaço $\mathcal{B}(U \times V, W)$ tem dimensão mnp.*

Prova. A ideia da demonstração é exibir uma base para o espaço $\mathcal{B}(U \times V, W)$. Para isso, tome $\{w_1, \ldots, w_m\}$, $\{u_1, \ldots, u_n\}$ e $\{v_1, \ldots, v_p\}$ bases de W, U e V, respectivamente. Para cada tripla (i, j, k), com $i = 1, \ldots, m$, $j = 1, \ldots, n$ e $k = 1, \ldots, p$, definimos uma aplicação bilinear $f_{ij}^k : U \times V \to W$ tal que

$$f_{ij}^k(u_r, v_s) = \begin{cases} w_i & \text{se } r = j \text{ e } s = k \\ 0 & \text{se } r \neq j \text{ ou } s \neq k. \end{cases} \tag{1.10}$$

O Teorema 1.10 garante a existência de f_{ij}^k. Mostraremos então que o conjunto

$$A = \{f_{ij}^k \mid i = 1, \ldots, m,\ j = 1, \ldots, n \text{ e } k = 1, \ldots, p\}$$

é uma base do espaço $\mathcal{B}(U \times V, W)$. Tome $f \in \mathcal{B}(U \times V, W)$. Observe que $f(u_r, v_s)$ pode ser escrito como

$$f(u_r, v_s) = \sum_{i=1}^{m} a_{ir}^s w_i \tag{1.11}$$

para todo $r = 1, \ldots, n$ e $s = 1, \ldots, p$. Considere a aplicação bilinear

$$g = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} a_{ij}^{k} f_{ij}^{k}.$$

Vamos mostrar que $g = f$. Em particular, temos

$$g(u_r, v_s) = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} a_{ij}^{k} f_{ij}^{k}(u_r, v_s) = \sum_{i=1}^{m} a_{ir}^{s} w_i = f(u_r, v_s)$$

para todo $r = 1, \ldots, n$ e $s = 1, \ldots, p$. Portanto $g = f$. O conjunto A é linearmente independente, pois se

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} a_{ij}^{k} f_{ij}^{k} = 0,$$

então

$$0 = \sum_{k=1}^{p} \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^{k} f_{ij}^{k}(u_r, v_s) = \sum_{i=1}^{m} a_{ir}^{s} w_i.$$

Como $\{w_1, \ldots, w_m\}$ é uma base de W, tem-se $a_{ir}^{s} = 0$ para todo $i = 1, \ldots, m$, $r = 1, \ldots, n$ e $k = 1, \ldots, p$. □

Em particular, se as dimensões dos espaços vetoriais U e V são m e n, respectivamente, então o espaço vetorial $\mathcal{B}(U \times V, \mathbb{R})$ tem dimensão mn. Ora, como dois espaços vetoriais de mesma dimensão finita são isomorfos [17], existe uma matriz $m \times n$ associada a cada $f \in \mathcal{B}(U \times V, \mathbb{R})$. Considerando $B = \{u_1, \ldots, u_m\}$ e $C = \{v_1, \ldots, v_n\}$ bases de U e V, respectivamente e se $u = \sum_{i=1}^{m} \alpha_i u_i$ e $v = \sum_{j=1}^{n} \beta_j v_j$, então fazendo $f(u_i, v_j) = a_{ij}$ para todo $i = 1, \ldots, m$ e $j = 1, \ldots, n$ teremos

$$f(u, v) = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i a_{ij} \beta_j$$

que em forma matricial fica $f(u, v) = [u]_B^T A [v]_C$, onde $A = (a_{ij})$ e $[v]_C$ denota as coordenadas do vetor v na base C. Daí segue a seguinte definição:

Definição 1.12 *Sejam U e V espaços vetoriais de dimensão finita. Fixadas bases $B = \{u_1, \ldots, u_m\} \subset U$ e $C = \{v_1, \ldots, v_n\} \subset V$ definimos para cada $f \in \mathcal{B}(U \times V, \mathbb{R})$ a matriz de f em relação às bases ordenadas B e C como sendo $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ cujos elementos são dados por $a_{ij} = f(u_i, v_j)$ com $i = 1, \ldots, m$ e $j = 1, \ldots, n$.*

Considere agora o espaço $\mathcal{B}(\mathbb{R}^m \times \mathbb{R}^n, \mathbb{R}^p)$ e as bases canônicas $\{e_1, \ldots, e_m\}$, $\{\bar{e_1}, \ldots, \bar{e_n}\}$, $\{\hat{e_1}, \ldots, \hat{e_p}\}$ de $\mathbb{R}^m$, $\mathbb{R}^n$ e $\mathbb{R}^p$, respectivamente. Considere $f \in \mathcal{B}(\mathbb{R}^m \times \mathbb{R}^n, \mathbb{R}^p)$. Para

todo $u \in \mathbb{R}^m$ e $v \in \mathbb{R}^n$ temos

$$f(u,v) = \sum_{j=1}^{m}\sum_{k=1}^{n} u_j v_k f(e_j, \bar{e_k})$$

onde u_j e v_k são as coordenadas de u e v nas bases canônicas de $\mathbb{R}^m$ e $\mathbb{R}^n$, respectivamente. Denote a i-ésima coordenada de f por f_i. Observe que $f_i \in \mathcal{B}(\mathbb{R}^m \times \mathbb{R}^n, \mathbb{R})$. Assim para cada $i = 1, \ldots, p$ temos

$$f_i(u,v) = \sum_{j=1}^{m}\sum_{k=1}^{n} u_j v_k f_i(e_j, \bar{e_k}).$$

Pela Definição 1.12, a matriz de f_i em relação as bases canônicas, é a matriz

$$A_i = (t_{ij}^k) \in \mathbb{R}^{m\times n},$$

onde $t_{ij}^k = f_i(e_j, \bar{e_k})$. Assim podemos escrever

$$f_i(u,v) = u^T A_i v.$$

De modo geral, podemos definir p matrizes $m \times n$ e olhá-las, por exemplo, como um tensor $\mathcal{T} \in \mathbb{R}^{p\times m\times n}$, ou seja, as p matrizes podem ser vistas como as camadas horizontais do tensor $\mathcal{T}$. Desta forma, podemos escrever $f(u,v)$ como um produto do tensor $\mathcal{T}$ pelos vetores u e v, isto é,

$$f(u,v) = \begin{pmatrix} u^T A_1 v \\ u^T A_2 v \\ \vdots \\ u^T A_p v \end{pmatrix} = (\mathcal{T} \bar{\times}_2 u)v. \tag{1.12}$$

Desta maneira, podemos generalizar a Definição 1.12 como segue:

Definição 1.13 *Sejam U e V espaços vetoriais de dimensão finita. Fixadas bases $B = \{u_1, \ldots, u_m\}$ e $C = \{v_1, \ldots, v_n\}$ de U e V, respectivamente, definimos para cada $f \in \mathcal{B}(U \times V, \mathbb{R}^p)$ o tensor $\mathcal{T}$ em relação às bases ordenadas B e C como sendo $\mathcal{T} = (t_{ij}^k) \in \mathbb{R}^{p\times m\times n}$ cujos elementos são dados por $t_{ij}^k = f_i(u_j, v_k)$ onde f_i é a i-ésima coordenada de f, isto é, $f_i \in \mathcal{B}(U \times V, \mathbb{R})$, com $i = 1, \ldots, p$, $j = 1, \ldots, m$ e $k = 1, \ldots, n$.*

1.4 Diferenciabilidade

Sejam uma aplicação diferenciável $F : U \subset \mathbb{R}^m \to \mathbb{R}^n$ com U aberto e $a \in U$. Denote $\mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ o conjunto de todas as aplicações lineares de $\mathbb{R}^m$ em $\mathbb{R}^n$. Quando $F' : U \subset \mathbb{R}^m \to \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ for diferenciável em $a \in U$, dizemos que a aplicação F é

duas vezes diferenciável em $a \in U$ e assim temos a transformação linear $F''(a) \in \mathcal{L}(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n))$ que será chamada a segunda derivada de F no ponto $a \in U$.

A norma de $F''(a)$ é definida de maneira natural, isto é, para qualquer $h \in \mathbb{R}^m$,

$$\|F''(a)h\| = \sup_{\|k\|=1} \{\|F''(a)hk\| \text{ com } k \in \mathbb{R}^m\}$$

e então

$$\|F''(a)\| = \sup_{\|h\|=1} \|F''(a)h\| = \sup_{\|h\|=1} \sup_{\|k\|=1} \|F''(a)hk\|.$$

Observe que, pelo Teorema 1.11, os espaços $\mathcal{L}(\mathbb{R}^m, \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n))$ e $\mathcal{B}(\mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^n)$ são isomorfos, o que permite interpretar $F''(a)$ como uma aplicação bilinear no espaço $\mathcal{B}(\mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^n)$. Tal isomorfismo pode ser encontrado em livros clássicos de análise [53, 54]. Por outro lado, pelo mesmo teorema, o espaço das aplicações bilineares $\mathcal{B}(\mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^n)$ é isomorfo ao espaço dos tensores $\mathbb{R}^{n \times m \times m}$. Por esse motivo, interpretaremos $F''(a)$ nesta tese como um tensor no espaço $\mathbb{R}^{n \times m \times m}$. Vamos denotá-la por $\mathcal{T}_F(a)$ e utilizaremos as operações vistas na Seção 1.2.

Resta saber como são formados os elementos do tensor $\mathcal{T}_F(a)$. Para isso, considere $A : \mathbb{R} \to \mathbb{R}^{n \times m}$ e $\alpha \in \mathbb{R}$. Vemos que $A(\alpha)$ é uma matriz com n linhas e m colunas. Seus elementos serão denotados por $a_{ij}(\alpha)$ onde a_{ij} são as funções coordenadas de A com $i = 1, \ldots, n$ e $j = 1, \ldots, m$. Quando $a_{ij} : \mathbb{R} \to \mathbb{R}$ for diferenciável em α para todo $i = 1, \ldots, n$ e $j = 1, \ldots, m$, a derivada de A no ponto α é a matriz

$$A'(\alpha) = (a'_{ij}(\alpha)) \in \mathbb{R}^{n \times m}. \tag{1.13}$$

A definição da derivada de $A(\alpha)$ como em (1.13) é uma definição clássica, veja [35]. Para generalizar (1.13), considere $A : U \subset \mathbb{R}^p \to \mathbb{R}^{n \times m}$ uma aplicação diferenciável em $u \in U$ com funções coordenadas $a_{ij} : \mathbb{R}^p \to \mathbb{R}$ com $i = 1, \ldots, n$ e $j = 1, \ldots, m$. Quando a_{ij} for diferenciável em u para todo $i = 1, \ldots, n$ e todo $j = 1, \ldots, m$, definimos a derivada de A no ponto u como o tensor

$$A'(u) = (\nabla a_{ij}(u)) \in \mathbb{R}^{n \times m \times p}. \tag{1.14}$$

Note que de fato, (1.14) é uma generalização de (1.13). Fixado i e j, $\nabla a_{ij}(u)$ é uma fibra tubo do tensor $A'(u)$, cujos elementos são

$$A'(u)^k_{ij} = \frac{\partial a_{ij}}{\partial x_k}(u) \tag{1.15}$$

para todo $k = 1, \ldots, p$.

Para exemplificar, considere uma aplicação $F : U \subset \mathbb{R}^2 \to \mathbb{R}^3$ duas vezes diferenciável

em $a \in U$ com U aberto. A matriz jacobiana de F no ponto a é

$$J_F(a) = \begin{pmatrix} \nabla f_1(a)^T \\ \nabla f_2(a)^T \\ \nabla f_3(a)^T \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \frac{\partial f_1}{\partial x_2}(a) \\ \frac{\partial f_2}{\partial x_1}(a) & \frac{\partial f_2}{\partial x_2}(a) \\ \frac{\partial f_3}{\partial x_1}(a) & \frac{\partial f_3}{\partial x_2}(a) \end{pmatrix}$$

e sua derivada é, por (1.14), o tensor

$$J'_F(a) = \mathcal{T}_F(a) = \left(\nabla \frac{\partial f_i}{\partial x_j}(a)\right) \in \mathbb{R}^{3\times 2\times 2} \tag{1.16}$$

onde, por (1.15), seus elementos são da forma

$$t_{ij}^k = \frac{\partial^2 f_i}{\partial x_k \partial x_j}(a).$$

Fixado i, vemos facilmente que a i-ésima camada horizontal de $\mathcal{T}_F(a)$ é a matriz Hessiana $\nabla^2 f_i(a)$, isto é, em relação a função coordenada $f_i : U \subset \mathbb{R}^2 \to \mathbb{R}$, temos

$$\nabla^2 f_i(a) = \mathcal{T}_F(a)^{i::} = \begin{pmatrix} \frac{\partial^2 f_i}{\partial x_1 \partial x_1}(a) & \frac{\partial^2 f_i}{\partial x_1 \partial x_2}(a) \\ \frac{\partial^2 f_i}{\partial x_2 \partial x_1}(a) & \frac{\partial^2 f_i}{\partial x_2 \partial x_2}(a) \end{pmatrix}. \tag{1.17}$$

Observe que uma coluna da matriz $\nabla^2 f_i(x)$ é uma fibra linha da i-ésima camada horizontal.

Para os métodos estudados nesta tese, com frequência calculamos o produto do tensor $\mathcal{T}_F(a)$ por vetores do domínio da aplicação F, o que neste exemplo, são vetores em $\mathbb{R}^2$. Segue da Definição 1.7, que é possível realizar os produtos modo-2 contraído e modo-3 contraído. Como as matrizes Hessianas são simétricas, dado $v \in \mathbb{R}^2$, pelo Lema 1.8 e por (1.7) e (1.8), temos

$$\mathcal{T}_F(a) \,\bar{\times}_3 v = \mathcal{T}_F(a) \,\bar{\times}_2 v = \begin{pmatrix} \mathrm{lin}_1(\mathcal{T}_F(a) \,\bar{\times}_2 v) \\ \mathrm{lin}_2(\mathcal{T}_F(a) \,\bar{\times}_2 v) \\ \mathrm{lin}_3(\mathcal{T}_F(a) \,\bar{\times}_2 v) \end{pmatrix} = \begin{pmatrix} v^T \nabla^2 f_1(a) \\ v^T \nabla^2 f_2(a) \\ v^T \nabla^2 f_3(a) \end{pmatrix} \in \mathbb{R}^{3\times 2}.$$

Desta forma, tem-se

$$(\mathcal{T}_F(a) \,\bar{\times}_2 v)u = \begin{pmatrix} v^T \nabla^2 f_1(a) u \\ v^T \nabla^2 f_2(a) u \\ v^T \nabla^2 f_3(a) u \end{pmatrix} \in \mathbb{R}^3 \tag{1.18}$$

qualquer que seja $u, v \in \mathbb{R}^2$.

Isto significa que o tensor $\mathcal{T}_F(a)$ definido como em (1.16) é o tensor associado a aplicação bilinear $F''(a)$, em relação à base canônica de $\mathbb{R}^2$, segundo a Definição 1.13. Sem perda de generalidade, vamos simplesmente denotar em todo este trabalho

$$\mathcal{T}_F(a)\,\bar{\times}_3 v = \mathcal{T}_F(a)\,\bar{\times}_2 v = \mathcal{T}_F(a)v$$

e conforme o Lema 1.8, podemos fazer

$$(\mathcal{T}_F(a)u)v = (\mathcal{T}_F(a)v)u = \mathcal{T}_F(a)vu.$$

Para finalizar esta seção, vamos considerar um caso particular. Sabemos que a k-ésima coluna da jacobiana $J_F(x)$ é o produto $J_F(x)e_k$, onde e_k é o k-ésimo vetor canônico do $\mathbb{R}^n$. Vale a pena identificar qual tipo de camada é a matriz $\mathcal{T}_F(x)e_k$. Por definição, temos

$$\mathcal{T}_F(x)e_k = \begin{pmatrix} e_k^T\nabla^2 f_1(x) \\ e_k^T\nabla^2 f_2(x) \\ \vdots \\ e_k^T\nabla^2 f_n(x) \end{pmatrix} = \begin{pmatrix} \text{lin}_k\nabla^2 f_1(x) \\ \text{lin}_k\nabla^2 f_2(x) \\ \vdots \\ \text{lin}_k\nabla^2 f_n(x) \end{pmatrix}$$

Ora, como $\text{lin}_k\nabla^2 f_i(x)$ é a k-ésima fibra tubo da i-ésima camada horizontal, temos que $\mathcal{T}_F(x)e_k$ é a k-ésima camada lateral ou, por simetria das Hessianas, a transposta da k-ésima camada frontal. Em suma, para uma aplicação $F : U \subset \mathbb{R}^n \to \mathbb{R}^m$ duas vezes diferenciável, temos $\mathcal{T}_F(x) \in \mathbb{R}^{m\times n\times n}$ onde as m camadas horizontais são as Hessianas $\nabla^2 f_i(x)$, com $i = 1, \ldots, m$ e as n camadas laterais e frontais podem ser obtidas pelo produto $\mathcal{T}_F(x)e_k$, com $k = 1, \ldots, n$.

1.4.1 Alguns Resultados Clássicos

Nesta seção, são apresentados alguns resultados clássicos de diferenciabilidade. A principal referência é [62].

Lema 1.14 *Sejam $U \subset \mathbb{R}^n$ aberto e convexo, $F : U \to \mathbb{R}^n$ uma aplicação diferenciável e J_F Lipschitz em U com constante L. Então*

$$\|F(x) - F(y) - J_F(y)(x - y)\| \leq \frac{L}{2}\|x - y\|^2,$$

para todos $x, y \in U$.

Prova. Fazendo $v = x - y$ e utilizando a fórmula de Taylor com resto integral, temos

$$F(x) - F(y) = \int_0^1 J_F(y + tv)v dt.$$

Portanto

$$\|F(x) - F(y) - J_F(y)(x - y)\| \leq \int_0^1 \|(J_F(y + tv) - J_F(y))v\| \, dt \leq \frac{L}{2} \|v\|^2 ,$$

completando a demonstração. □

No Lema 1.14, foi estabelecido um limitante para o erro ao considerar o modelo linear de F em torno de x como uma aproximação para $F(x)$. O mesmo pode ser feito ao considerar o modelo linear de J_F em torno de x como uma aproximação para $J_F(x)$, pois se $F \in \mathcal{C}^2$ em um aberto convexo $U \subset \mathbb{R}^n$ e $\mathcal{T}_F$ é Lipschitz com constante $L_2 > 0$, então

$$J_F(x) - J_F(y) = \int_0^1 \mathcal{T}_F(x + t(y - x))(y - x) dt. \tag{1.19}$$

Veja [62, 3.3.7]. Daí de forma análoga à demonstração do Lema 1.14, temos

$$\|J_F(x) - J_F(y) - \mathcal{T}_F(y)(x - y)\| \leq \frac{L_2}{2} \|x - y\|^2 . \tag{1.20}$$

No próximo lema, vamos estabelecer um limitante para o erro ao considerar o modelo quadrático de F em torno de x como uma aproximação para $F(x)$. O Lema 1.15 pode ser generalizado para $F \in \mathcal{C}^p$ com $p > 2$, veja [62, NR 3.3-3].

Lema 1.15 *Seja $U \subset \mathbb{R}^n$ aberto e convexo e $F : U \to \mathbb{R}^n$ uma aplicação duas vezes diferenciável e $\mathcal{T}_F$ Lipschitz em U com constante L_2. Então*

$$\left\| F(x) - F(y) - J_F(y)(x - y) - \frac{1}{2}\mathcal{T}_F(y)(x - y)(x - y) \right\| \leq \frac{L_2}{6} \|x - y\|^3 ,$$

para todos $x, y \in U$.

Prova. Fazendo $v = x - y$ e utilizando a fórmula de Taylor com resto integral, temos

$$F(x) - F(y) - J_F(y)v = \int_0^1 (1 - t)\mathcal{T}_F(y + tv)vv dt.$$

Portanto

$$F(x) - F(y) - J_F(y)v - \frac{1}{2}\mathcal{T}_F(y)vv = \int_0^1 \left[(1 - t)\mathcal{T}_F(y + tv) - \frac{1}{2}\mathcal{T}_F(y)\right]vv dt.$$

Note que o termo $(1-t)\mathcal{T}_F(y+tv) - \frac{1}{2}\mathcal{T}_F(y)$ pode ser escrito como

$$(1-t)\Big[\mathcal{T}_F(y+tv) - \mathcal{T}_F(y)\Big] + \Big(\frac{1}{2} - t\Big)\mathcal{T}_F(y).$$

Como $\int_0^1 \Big(\frac{1}{2} - t\Big)dt = 0$ temos que

$$\left\|F(x) - F(y) - J_F(y)v - \frac{1}{2}\mathcal{T}_F(y)vv\right\| \leq \int_0^1 \Big[(1-t)\left\|\mathcal{T}_F(y+tv) - \mathcal{T}_F(y)\right\|\Big]\left\|v\right\|^2 dt \leq$$

$$\leq \int_0^1 (1-t)tL_2\left\|v\right\|^3 dt = \frac{L_2}{6}\left\|v\right\|^3.$$

□

Existe uma relação importante sobre uma constante de Lipschitz de uma aplicação e sua derivada. Tal relação é enunciada e demonstrada no Lema 1.16.

Lema 1.16 *Seja $U \subset \mathbb{R}^m$ aberto e convexo e $F : U \to \mathbb{R}^n$ diferenciável. Considere $M > 0$. Temos que $\|J_F(x)\| \leq M$, se e somente se, $\|F(x) - F(y)\| \leq M\|x - y\|$ para todo $x, y \in U$.*

Prova. A primeira afirmação é imediata pela desigualdade do valor médio e pela convexidade. Para provar a segunda afirmação, considere $a \in U$ e $v \in \mathbb{R}^m$ tal que $\|v\| = 1$. Assim, $a+tv \in U$ para $t > 0$ suficientemente pequeno. Além disso, pela diferenciabilidade de F temos

$$F(a+tv) = F(a) + tJ_F(a)v + o(t),$$

ou seja,

$$J_F(a)v = \frac{F(a+tv) - F(a)}{t} - \frac{o(t)}{t}.$$

Utilizando a hipótese temos

$$\|J_F(a)v\| \leq \frac{Mt\|v\|}{t} + \frac{\|o(t)\|}{t}.$$

Passando o limite quando $t \to 0^+$ segue que $\|J_F(a)v\| \leq M$ e portanto

$$\|J_F(a)\| = \sup_{\|v\|=1}\|J_F(a)v\| \leq M.$$

□

Em particular, o Lema 1.16 garante que a menor constante Lipschitz de F é atingida fazendo $M = \sup_{x\in U}\{\|J_F(x)\|\}$. Além disso, analogamente ao que foi feito no Lema 1.16

podemos concluir quer

$$\|\mathcal{T}_F(x)\| \leq M \iff \|J_F(x) - J_F(y)\| \leq M \|x - y\| \tag{1.21}$$

para todo $x, y \in U$.

Capítulo 2

Equações Não Lineares: Caso Unidimensional

Considere neste capítulo o seguinte problema:

$$\text{encontrar um } x^* \in \mathbb{R} \text{ tal que } f(x^*) = 0, \tag{2.1}$$

onde $f : I \subset \mathbb{R} \to \mathbb{R}$ é uma função de classe $\mathcal{C}^2$.

Um método iterativo muito utilizado para resolver o problema (2.1) é o conhecido método de Newton. Dedicamos a primeira seção para uma rápida explanação desse método. Um bom histórico sobre o método de Newton pode ser visto em [78]. Em seguida apresentamos alguns métodos com convergência cúbica, sob hipóteses razoáveis. Descrevemos a construção geométrica e algébrica dos métodos Chebyshev, Halley e Super-Halley. Esses métodos fazem parte da classe de métodos Chebyshev-Halley. Esta classe de métodos foi introduzida por Hernández e Salanova [43] em 1993 para o caso unidimensional e generalizado para espaços de Banach por Hernández and Gutiérrez [42] em 1997.

2.1 Método de Newton

Geralmente, nos métodos iterativos, a cada iteração é construído um modelo para f e toma-se como estimativa para um zero da função um zero do modelo. Obviamente, um modelo pode não ter zeros, o que não é, de certa forma, vantajoso. No método de Newton, dada uma boa estimativa inicial $x^0 \in \mathbb{R}$, o método gera uma sequência (x^k) tal que x^{k+1} é o zero do polinômio de Taylor de primeira ordem em torno do ponto x^k, para todo $k = 0, 1, \ldots$, ou seja, a cada iteração k, toma-se o seguinte modelo para f:

$$m_k(x) = f(x^k) + f'(x^k)(x - x^k).$$

Daí toma-se x^{k+1} como sendo o zero do modelo $m_k(x)$, isto é,

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}.$$

É bem conhecido [22], que sob hipóteses razoáveis, o método de Newton converge quadraticamente.

2.2 Métodos com Convergência Cúbica

Discutimos nesta seção os métodos clássicos com convergência cúbica. Diante do método de Newton, é intuitivo indagar sobre a utilização do polinômio de Taylor de segunda ordem em torno do ponto x^k como sendo um modelo para f, ou seja, ao considerar o modelo

$$m_k(x) = f(x^k) + f'(x^k)(x - x^k) + \frac{1}{2}f''(x^k)(x - x^k)^2 \tag{2.2}$$

e tomar x^{k+1} como sendo um zero deste modelo, devemos ter

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)} \pm \frac{|f'(x^k)|}{f''(x^k)}\sqrt{1 - 2\ell(x^k)}, \tag{2.3}$$

onde

$$\ell(x) = \frac{f(x)f''(x)}{f'(x)^2} \tag{2.4}$$

é o grau de convexidade logarítmica de f avaliado em x. Basicamente, o grau de convexidade logarítmica é uma estimativa do número de vezes que é necessário compor uma certa função convexa, cuja derivada segunda seja estritamente positiva, com a função logarítmica até obter uma função que não seja convexa. Este conceito é apresentado em detalhes no Apêndice A juntamente com as referências no assunto. Uma aplicação importante sobre o grau de convexidade logarítmica será apresentada na subseção 2.2.3 no sentido de estudar a influência da convexidade da função no método de Newton.

Sobre o processo iterativo (2.3), nos deparamos com dois problemas: o primeiro é a escolha do sinal $(+)$ ou $(-)$. Para resolvê-lo, vamos considerar a função de iteração

$$\phi(x) = x - \frac{f'(x)}{f''(x)} \pm \frac{|f'(x)|}{f''(x)}\sqrt{1 - 2\ell(x)}$$

e observar, como em [75], que x^* é um ponto fixo de ϕ se, e somente se, tomarmos o sinal $(+)$ quando $f'(x) > 0$ e o sinal $(-)$ quando $f'(x) < 0$. Com esta escolha teremos

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}\Big(1 - \sqrt{1 - 2\ell(x^k)}\Big). \tag{2.5}$$

O segundo problema está no mau condicionamento de $1-\sqrt{1-2\ell(x^k)}$ quando x^k está próximo da solução x^*. Então reescrevemos (2.5) como

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}\Big(\frac{2}{1+\sqrt{1-2\ell(x^k)}}\Big). \tag{2.6}$$

Cauchy [12] foi o primeiro a estabelecer convergência semilocal do processo iterativo (2.6), além de provar convergência cúbica sob algumas hipóteses. Para outras referências sobre essa convergência, o leitor pode consultar [45, 62, 75].

Observe que para obter x^{k+1}, além de que $f'(x^k)$ deve ser não nulo para todo $k \in \mathbb{N}$, devemos ter

$$\ell(x^k) \leq \frac{1}{2},$$

ou seja, o método é muito restritivo, pois para funções bem simples como $f(x) = x^n$, para $n \geq 3$, vemos facilmente que $\ell(x) > \dfrac{1}{2}$ para todo $x \in \mathbb{R}$. Para essa classe de funções, isto significa que dado qualquer ponto inicial $x^0 \in \mathbb{R}$ não é possível determinar $x^1 \in \mathbb{R}$ pelo processo iterativo (2.6).

2.2.1 Método de Chebyshev

Outros métodos que possuem convergência cúbica são os métodos da classe Chebyshev-Halley que veremos adiante. Em particular, o método de Chebyshev baseia-se no seguinte problema equivalente a (2.1)

$$\text{obter } f^{-1} \text{ e calcular } x^* = f^{-1}(0). \tag{2.7}$$

Observe inicialmente que, se existe $\delta > 0$ tal que $f'(x) \neq 0$ para todo

$$x \in I = (x^* - \delta, x^* + \delta), \tag{2.8}$$

então f possui uma inversa $g = f^{-1}$ em I. Neste sentido, o método de Chebyshev considera o polinômio de Taylor de segunda ordem de g no ponto $y^k = f(x^k)$,

$$p_k(y) = g(y^k) + g'(y^k)(y - y^k) + \frac{1}{2}g''(y^k)(y - y^k)^2, \tag{2.9}$$

onde $x^k \in I$.

Dado $x^k \in I$, obtemos y^k e definimos x^{k+1} como sendo $p_k(0)$, isto é,

$$x^{k+1} = p_k(0) = g(y^k) - g'(y^k)y^k + \frac{1}{2}g''(y^k)(y^k)^2. \tag{2.10}$$

Traub [75] credita este método a Euler, mas na literatura russa ele é atribuído a Chebyshev [13, 25]

Para ilustrar o método de Chebyshev, observe na Figura 2.1 que dado x^k, calculamos $y^k = f(x^k)$ e construímos o polinômio de Taylor de segunda ordem de f^{-1} (em verde) avaliado em y^k e tomamos o próximo iterado como $x^{k+1} = p_k(0)$.

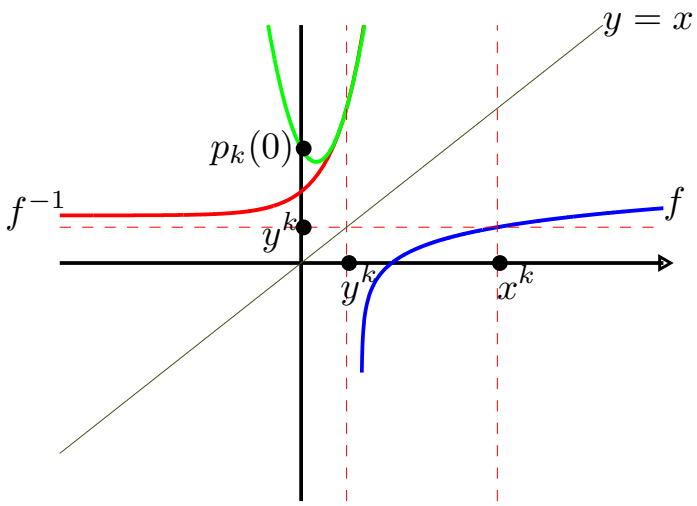

Figura 2.1: Uma iteração do método de Chebyshev.

Como a inversa de uma função nem sempre é disponível e é tão difícil obtê-la quanto resolver o problema (2.1), podemos obter uma expressão para $p_k(0)$ em termos de f. Como

$$g(y) = x, \quad g'(y) = \frac{1}{f'(x)} \text{ e } g''(y) = \frac{-f''(x)}{f'(x)^3},$$

por (2.10), definimos o método de Chebyshev como

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)} - \frac{f''(x^k)f(x^k)^2}{2f'(x^k)^3} = x^k - \frac{f(x^k)}{f'(x^k)}\Big(1 + \frac{1}{2}\ell(x^k)\Big), \tag{2.11}$$

onde $\ell(x)$ é definido em (2.4).

É interessante descrever um modelo para f ao invés de um modelo para f^{-1}. Veremos que um modelo de f em torno de x^k cujo zero é x^{k+1} dado em (2.11), pode ser uma função quadrática, diferente do polinômio de Taylor de segunda ordem de f em x^k, que concorda com f, f' e f'' no ponto x^k.

Por simetria, podemos obter uma curva simétrica à parábola (2.9) em relação à reta $y = x$. Concentraremos nesse instante nossa atenção na parábola tangente *(osculatory)*

ao gráfico de f no ponto x^k, ou seja, a parábola definida por

$$x = am_k(x)^2 + bm_k(x) + c \tag{2.12}$$

que satisfaz as condições

$$m_k(x^k) = f(x^k), \quad m_k'(x^k) = f'(x^k) \text{ e } m_k''(x^k) = f''(x^k). \tag{2.13}$$

Observe que c é o zero da quadrática definida em (2.12). Impondo estas condições, vemos facilmente que

$$a = \frac{-f''(x^k)}{2f'(x^k)^3},$$

$$b = \frac{f'(x^k)^2 + f(x^k)f''(x^k)}{f'(x^k)^3}$$

e

$$c = x^k - \frac{f(x^k)}{f'(x^k)}\Big(1 + \frac{f(x^k)f''(x^k)}{2f'(x^k)^2}\Big).$$

Portanto, x^{k+1} dado em (2.11) é o zero da função (2.12). É ilustrado na Figura 2.2 a simetria entre o polinômio de Taylor de segunda ordem de f^{-1} (em verde) em $f(x^k)$ (2.9) e a parábola tangente a f (em cinza) em x^k (2.12).

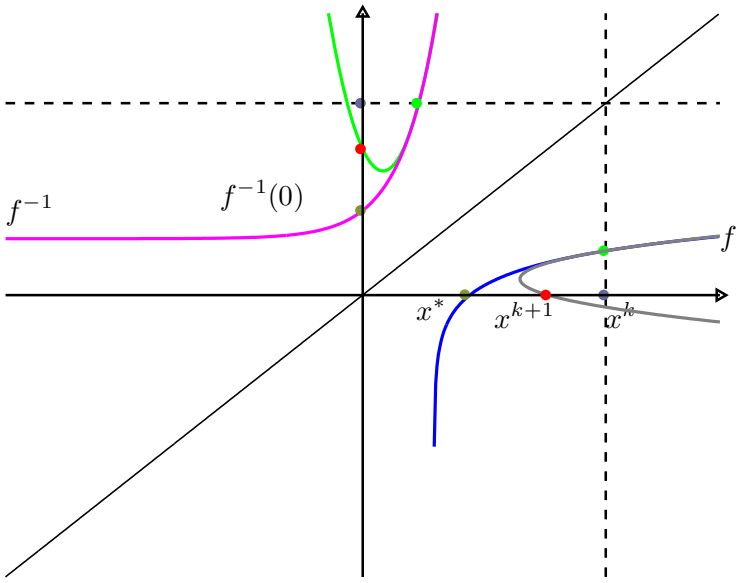

Figura 2.2: Simetria entre as parábolas.

2.2.2 Método de Halley

Outro método com convergência cúbica a ser estudado nesta tese é o método de Halley. Sobre este método, Traub afirma em [75] o seguinte: *"Halley's method must share with the secant method the distinction of being the most frequently rediscovered methods in the literature."*

Já vimos que o processo iterativo (2.6) é muito restritivo. No sentido de evitar problemas numéricos no cálculo da raiz quadrada, substituímos $\sqrt{1-x}$ por sua aproximação de Taylor $1-\frac{1}{2}x$ próximo de $x=0$. Assim obtemos o método de Halley

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}\Big(\frac{2}{2-\ell(x^k)}\Big), \tag{2.14}$$

onde $\ell(x)$ é definido em (2.4).

No sentido de generalizar o método de Halley para o espaço $\mathbb{R}^n$, vamos introduzi-lo de uma maneira mais construtiva. Considere o polinômio de Taylor de segunda ordem da função f no ponto x^k, dado em (2.2). Um fator $x-x^k$ do termo quadrático deste polinômio será aproximado pelo passo de Newton $\dfrac{-f(x^k)}{f'(x^k)}$. Desta forma teremos o seguinte modelo para f em x^k:

$$m_k(x) = f(x^k) + \Big(f'(x^k) - \frac{f''(x^k)f(x^k)}{2f'(x^k)}\Big)(x-x^k). \tag{2.15}$$

A partir de um ponto inicial x^0, o método de Halley gera uma sequência (x^k) tal que x^{k+1} é solução da equação $m_k(x)=0$. Desta forma, temos (2.14).

O método de Halley possui uma interpretação geométrica interessante. Apesar de ter sido descrito pelo modelo (2.15), Salehov [67] aparentemente foi o primeiro a sugerir que o método de Halley poderia ser obtido utilizando uma função racional como modelo para f. Por conveniência, vamos considerar o modelo como sendo uma hipérbole tangente *(osculatory)* [67], isto é, a hipérbole definida pela equação

$$m_k(x) = \frac{(x-x^k)+c}{a(x-x^k)+b} \tag{2.16}$$

deve concordar com f, f' e f'' em x^k, ou seja,

$$m_k(x^k) = f(x^k), \quad m_k'(x^k) = f'(x^k) \text{ e } m_k''(x^k) = f''(x^k).$$

Desta forma, temos

$$a = \frac{-f''(x^k)}{2f'(x^k)^2 - f(x^k)f''(x^k)},$$

$$b = \frac{2f'(x^k)}{2f'(x^k)^2 - f(x^k)f''(x^k)}$$

e

$$c = \frac{2f(x^k)}{f'(x^k)(2 - \ell(x^k))}.$$

Desta forma, x^{k+1} dado em (2.14) é o zero da função definida em (2.16). A Figura 2.3 ilustra uma iteração do método de Halley.

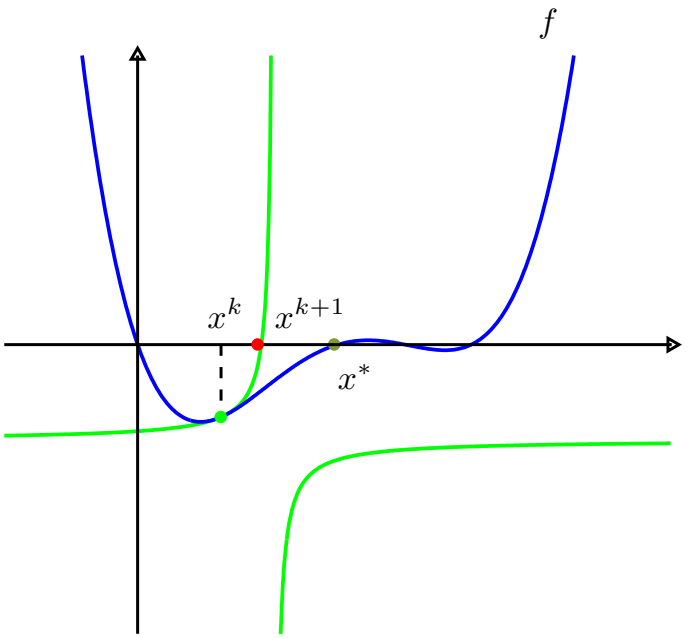

Figura 2.3: Uma iteração do método de Halley

2.2.3 Método Super-Halley

No início desta seção, utilizamos o grau de convexidade logarítmica de uma função f, veja (2.4). Para detalhes veja Apêndice A. Um estudo sobre a influência dessa medida no método de Newton e no método de Halley pode ser encontrado em [41] e [40], respectivamente. Para nossos fins, vamos denotar nesta subseção o grau de convexidade logarítmica de uma função f avaliado em x por

$$\ell_f(x) = \frac{f(x)f''(x)}{f'(x)^2}. \tag{2.17}$$

Vamos analisar, como em [41], a influência desta medida no método de Newton. Para isso, considere uma função $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ duas vezes continuamente diferenciável, convexa e estritamente crescente, tal que $f(a) < 0 < f(b)$. É possível mostrar que a

sequência definida por

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}, \tag{2.18}$$

com $x^0 = b$, converge para o zero $x^* \in [a, b]$ de f. Agora, seja g uma outra função satisfazendo as mesmas condições de f em $[a, b]$ tal que $g(x^*) = 0$. Considere a sequência

$$y^{k+1} = y^k - \frac{g(y^k)}{g'(y^k)} \tag{2.19}$$

com $y^0 = x^0$. Da mesma forma, essa sequência também converge para x^*. Sendo as mesmas condições, gostaríamos de estabelecer uma condição suficiente para que y^k seja estritamente menor que x^k para todo $k \in \mathbb{N}$. Uma sequência (y^k) que cumpra tal propriedade, será chamada de aceleração da sequência (x^k). Essa será a primícia do método Super-Halley [29, 39]. Para isso, considere as funções de iteração

$$\phi_f(x) = x - \frac{f(x)}{f'(x)} \quad \text{e} \quad \phi_g(x) = x - \frac{g(x)}{g'(x)}. \tag{2.20}$$

É fácil ver que $\phi_f' = \ell_f$ e $\phi_g' = \ell_g$.

Lema 2.1 *Considere as sequências* (x^k) *e* (y^k) *definidas em* (2.18) *e* (2.19)*, respectivamente. Se* $|\ell_f(x)| > |\ell_g(x)|$ *para todo* $x \in [a, b] - \{x^*\}$*, então* $y^k < x^k$ *para todo* $k \in \mathbb{N}$*.*

Prova. Como $x^0 = b$, temos que $x^{k+1} > x^*$ para todo $k \geq 1$. Temos que

$$y^1 - x^1 = \phi_g(x^0) - \phi_f(x^0) = (\phi_g - \phi_f)(x^0) - (\phi_g - \phi_f)(x^*).$$

Pelo Teorema do Valor Médio, existe $\xi_0 \in (x^*, x^0)$ tal que

$$y^1 - x^1 = (\ell_g - \ell_f)(\xi_0)(x^0 - x^*). \tag{2.21}$$

Por hipótese, temos que $\ell_g(x) < \ell_f(x)$ para todo $x \in (x^*, b]$. Daí segue por (2.21) que $y^1 < x^1$. Para mostrar que $y^2 < x^2$, primeiro note que ϕ_g é crescente em $(x^*, x^0]$. Daí segue que

$$y^2 - x^2 = \phi_g(y^1) - \phi_f(x^1) < \phi_g(x^1) - \phi_f(x^1) = (\ell_g - \ell_f)(\xi_1)(x^1 - x^*)$$

para algum $\xi_1 \in (x^*, x^1)$. Usando o mesmo argumento temos que $y^2 < x^2$ e por indução segue que $y^k < x^k$ para todo $k \geq 1$. □

Para exemplificar, considere[1]

$$f(x) = \frac{x^3}{216} - 1 \quad \text{e} \quad g(x) = \frac{x^2}{36} - 1$$

duas funções definidas no intervalo $[3, 10]$ cujo zero é $x^* = 6$. Estas funções são estritamente crescentes e convexas em $[3, 10]$. De (2.17), temos que

$$\ell_f(x) = \frac{2}{3} - \frac{144}{x^3} \quad \text{e} \quad \ell_g(x) = \frac{1}{2} - \frac{18}{x^2}.$$

Para ilustrar, note na Figura 2.4, que a hipótese do Lema 2.1 é verificada e, portanto, y^k está mais perto de x^* que x^k para todo $k \in \mathbb{N}$.

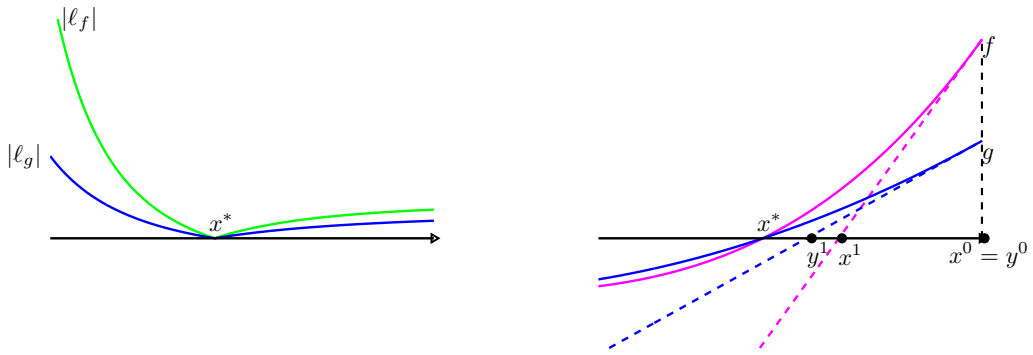

Figura 2.4: A sequência (y^k) é uma aceleração da sequência (x^k).

A questão é como determinar uma função g que cumpra a hipótese do Lema 2.1. Ora, um exemplo trivial é a função

$$g(x) = f'(x^*)(x - x^*).$$

Em [28], é apresentado uma família de funções que cumprem a hipótese do Lema 2.1. Em particular, com algumas hipóteses sobre ℓ_f e $\ell_{f'}$, a função

$$g(x) = \frac{f(x)}{\sqrt{f'(x)}}$$

cumpre tal hipótese. Essa função é interessante pois o método de Halley (2.14) pode ser obtido aplicando o método de Newton em g, veja [1].

O método Super-Halley é baseado na seguinte aproximação proveniente da expansão

[1] Veja referência [28, Example 1].

de Taylor de segunda ordem para f em torno de x^*:

$$f'(x^*)(x-x^*) \approx f(x) - \frac{f''(x^*)}{2}(x-x^*)^2$$

para x próximo de x^*. Como feito em [39, 41], defina

$$g(x) = f(x) - \frac{f''(x^*)}{2}(x-x^*)^2. \tag{2.22}$$

Note que x^* é desconhecido.

Como a ideia é acelerar a sequência (x^k), devemos obter uma sequência (y^k) tal que y^k esteja mais próximo de x^* que x^k para todo $k \in \mathbb{N}$. Para isso, faça

$$f''(x^*)(x^k-x^*)^j \approx f''(x^k)(x^k-x^{k+1})^j = f''(x^k)\frac{f(x^k)^j}{f'(x^k)^j} \tag{2.23}$$

para $j = 1, 2$. Desta forma, de (2.22) e (2.23), obtemos

$$g(x^k) \approx f(x^k) - \frac{f''(x^k)}{2}\frac{f(x^k)^2}{f'(x^k)^2} \quad \text{e} \tag{2.24}$$

$$g'(x^k) \approx f'(x^k) - f''(x^k)\frac{f(x^k)}{f'(x^k)}. \tag{2.25}$$

Usando (2.24) e (2.25), o método Super-Halley é definido fazendo $x^{k+1} \approx \phi_g(x^k)$, ou seja,

$$x^k - \frac{g(x^k)}{g'(x^k)} \approx x^k - \frac{f(x^k)}{f'(x^k)}\left[1 + \frac{\ell_f(x^k)}{2(1-\ell_f(x^k))}\right] \stackrel{def}{=} x^{k+1}. \tag{2.26}$$

Para ver que (2.26) é uma aceleração de (2.18), é suficiente mostrar que

$$\lim_{k\to\infty} \frac{\left\|x^{k+1}-x^*\right\|}{\|\phi_f(x^k)-x^*\|} = 0,$$

onde x^{k+1} é dado por (2.26) e ϕ_f é dado por (2.20). Esta prova é dada em [39, 41].

Apesar do método Super-Halley ter taxa de convergência cúbica, veja [39, Teorema 2.5], esse método possui propriedades interessantes quando f é um polinômio quadrático. Neste caso, um passo do método Super-Halley equivale a dois passos do método de Newton, veja [39, Teorema 2.7]. Isso garante que a taxa de convergência para essa classe de funções é 4, veja [39, Teorema 2.6]. A Figura 2.5 ilustra uma iteração do método Super-Halley. A curva em azul representa o polinômio de Taylor de segunda ordem de f avaliado em x^k.

Por fim, Hernández e Salanova [43] definem uma família de métodos chamada classe

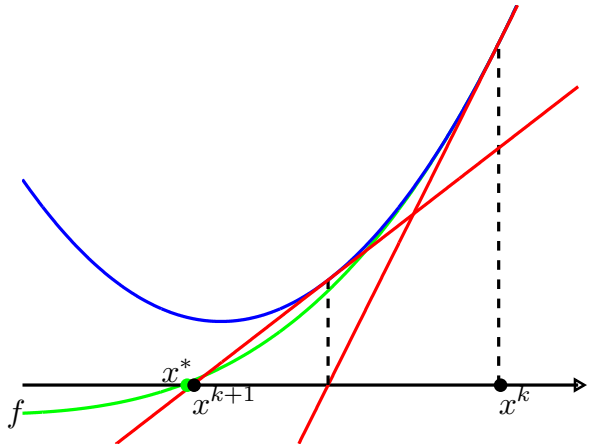

Figura 2.5: Uma iteração do método Super-Halley.

Chebyshev-Halley como sendo

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}\Big(1 + \frac{\ell_f(x^k)}{2(1-\alpha\ell_f(x^k))}\Big) \tag{2.27}$$

com $\alpha \in \mathbb{R}$. Observe que quando $\alpha = 0$ temos o método de Chebyshev (2.11), quando $\alpha = \dfrac{1}{2}$ temos o método de Halley (2.14) e quando $\alpha = 1$ temos o método Super-Halley (2.26). Propriedades de convergência podem ser dadas com hipóteses sobre $\ell_{f'}$. Uma generalização dessa classe no espaço $\mathbb{R}^n$ será vista na Seção 3.4.2. O leitor pode consultar também a referência [15].

Capítulo 3

Sistemas Não Lineares

Muitas aplicações de modelagem matemática no mundo real [24, 58, 60] consistem em resolver um sistema de equações, geralmente não lineares. Um sistema de equações não lineares pode ser escrito como $F(x) = 0$, onde F é uma aplicação de $\mathbb{R}^n$ em $\mathbb{R}^m$.

Nesta tese, vamos considerar uma aplicação $F : \mathbb{R}^n \to \mathbb{R}^n$ duas vezes continuamente diferenciável. O objetivo é

$$\text{encontrar um } x^* \in \mathbb{R}^n \text{ tal que } F(x^*) = 0. \tag{3.1}$$

Os métodos mais utilizados para resolver este problema são os métodos iterativos, pois em geral não é possível encontrar uma solução explícita por meios algébricos. Porém, existem outros métodos de aproximação diferentes, veja por exemplo [11].

Dentre os métodos iterativos, talvez o mais conhecido seja o método de Newton. O método de Newton é uma importante ferramenta, não apenas aplicada, mas também teórica, tendo um vasto campo de aplicações em matemática pura [32].

Basicamente, dado $x^k \in \mathbb{R}^n$, o método de Newton consiste em resolver o sistema linear

$$J_F(x^k)s^k = -F(x^k) \tag{3.2}$$

a cada iteração e tomar o próximo iterado como sendo $x^{k+1} = x^k + s^k$.

Resultados de convergência sobre o método de Newton são muito bem conhecidos [62]. No entanto, a título de motivação, exibiremos não só a prova de convergência quadrática desse método, mas o raio ótimo de convergência. Isso será apresentado na Seção 4.1 do Capítulo 4.

Embora o método de Newton seja teoricamente muito atrativo, é difícil usá-lo na prática. Observe que a cada passo, o sistema linear (3.2) deve ser resolvido de forma exata. O custo para resolvê-lo é de $O\left(\frac{n^3}{3}\right)$ operações quando se usa decomposição LU, veja [57]. Isto significa que, quando n é grande e o problema não possui nenhuma estrutura

especial, como por exemplo esparsidade da jacobiana, resolver o sistema (3.2) de forma exata torna-se inviável. Além disso, devemos calcular n^2 derivadas para a obtenção da jacobiana.

Algumas modificações do método de Newton são bem conhecidas. Podemos, por exemplo, resolver o sistema (3.2) de forma inexata, ou seja, resolvê-lo por algum método iterativo impondo uma precisão, como no método de Newton Inexato, ou aproximar a jacobiana $J_F(x^k)$, utilizando diferenças finitas, ou ainda, substituir a jacobiana $J_F(x^k)$ por outra matriz com alguma propriedade, como nos métodos quase-Newton.

Apresentaremos algumas dessas variações do método de Newton de maneira sucinta, pois elas serviram de inspiração e motivação para diminuirmos o custo computacional dos métodos da Classe Chebyshev-Halley.

3.1 Método de Newton Discreto

A ideia geral do método de Newton discreto é utilizar certas aproximações para a jacobiana. Essas aproximações são baseadas na seguinte definição:

Definição 3.1 *Seja $F : \mathbb{R}^n \to \mathbb{R}^n$ diferenciável. Dizemos que $A : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{n\times n}$ é uma aproximação consistente para J_F se*

$$\lim_{h\to 0} A(x,h) = J_F(x) \quad \textit{uniformemente.} \tag{3.3}$$

Além disso, se existem constantes $c, r \geq 0$ tais que

$$\|J_F(x) - A(x,h)\| \leq c|h|,$$

para todo $x \in \mathbb{R}^n$ e para todo h tal que $|h| \leq r$, então $A(x,h)$ é uma aproximação fortemente consistente para J_F.

A maneira mais natural de obter uma aproximação consistente para J_F é simplesmente aproximar a jacobiana $J_F(x)$ por diferenças finitas. Se a matriz $A(x,h)$ é tal que

$$\mathrm{col}_j(A(x,h)) = \frac{F(x+he_j) - F(x)}{h} \tag{3.4}$$

para todo $j = 1, \ldots, n$, onde h é um parâmetro de discretização, então A é uma aproximação consistente para J_F. Além disso, sob hipótese Lipschitz sobre J_F, A é uma aproximação fortemente consistente para J_F, veja [62, 11.2.5]. Note que h pode ser diferente para cada derivada parcial.

Utilizando a aproximação dada em (3.4), é possível mostrar que se $h_k \to 0$, então a taxa de convergência da sequência gerada pelo método de Newton discreto

$$x^{k+1} = x^k - A(x^k, h_k)^{-1} F(x^k)$$

é superlinear. Além disso, se $h_k = O(\|F(x^k)\|)$, então a taxa de convergência é quadrática. Detalhes podem ser vistos em [22, 62].

Embora o método de Newton discreto não exija o cálculo da jacobiana $J_F(x^k)$, ele pode se tornar caro computacionalmente dependendo da dimensão do problema. Se (3.4) é usado para discretizar a jacobiana, então são necessárias $n+1$ avaliações de função. Uma maneira mais eficiente de baratear o método de Newton, referente a jacobiana, é utilizar outras matrizes mais fáceis de serem construídas, como nos métodos quase-Newton.

3.2 Métodos Quase-Newton

Métodos quase-Newton são muito utilizados na prática devido as simplicidades de implementação e por terem boas propriedades de convergência. Esses métodos utilizam matrizes que são atualizadas a cada iteração. A rigor, eles não procuram aproximar a jacobiana a cada iteração como o método de Newton discreto, mas procuram satisfazer a condição de Dennis-Moré, que será vista mais adiante, para garantir taxa de convergência superlinear.

Dados $x^k \in \mathbb{R}^n$ e $B_k \in \mathbb{R}^{n\times n}$, esses métodos consistem em resolver o sistema

$$B_k s^k = -F(x^k) \tag{3.5}$$

e tomar o próximo iterado como sendo $x^{k+1} = x^k + s^k$. A matriz B_{k+1} deve ser escolhida de forma que a equação secante

$$B_{k+1} s^k = y^k, \tag{3.6}$$

onde

$$s^k = x^{k+1} - x^k \quad \text{e} \quad y^k = F(x^{k+1}) - F(x^k),$$

seja satisfeita.

Quando $n > 1$, existe uma infinidade de matrizes B_{k+1} que cumprem a condição secante (3.6). No entanto, é natural (por exemplo, por razões de estabilidade numérica) pedir que a mudança entre B_{k+1} e B_k, isto é, a diferença $B_{k+1} - B_k$ seja "mínima"em algum sentido.

Dados $B \in \mathbb{R}^{n\times n}$, $y \in \mathbb{R}^n$ e $s \in \mathbb{R}^n$ não-nulo, o Teorema 4.1 do artigo [47], garante

que a única solução do problema

$$\begin{array}{ll} \text{minimizar} & \left\|\hat{B} - B\right\|_F \\ \text{sujeito a} & \hat{B}s = y, \end{array} \tag{3.7}$$

onde $\|.\|_F$ denota a norma de Frobenius, é a matriz

$$B^+ = B + \frac{(y - Bs)s^T}{\langle s, s\rangle}. \tag{3.8}$$

A atualiação B^+ é chamada de atualização de Broyden. Neste sentido, o método de Broyden para resolver o problema (3.1), consiste em resolver a cada iteração o sistema (3.5) atualizando as matrizes como em (3.8), ou seja, determinar um x^{k+1} tal que

$$B_k(x^{k+1} - x^k) = -F(x^k)$$

e

$$B_{k+1} = B_k + \frac{(y^k - B_k s^k){s^k}^T}{\langle s^k, s^k\rangle}.$$

Em relação à convergência, uma estratégia clássica é mostrar que o método de Broyden satisfaz a condição de Dennis-Moré [21], ou seja, exigir que a sequência de matrizes (B_k) convirja para $J_F(x^*)$ é uma exigência um tanto forte e de certa forma desnecessária quando o objetivo é gerar uma sequência (x^k) que convirja para x^* com taxa superlinear. Basicamente, a condição de Dennis-Moré garante que a sequência (x^k) gerada pelo processo iterativo

$$x^{k+1} = x^k - B_k^{-1}F(x^k) \tag{3.9}$$

converge para uma solução x^* com taxa superlinear se, e somente se,

$$\lim_{k\to\infty} \frac{\left\|(B_k - J_F(x^*))s^k\right\|}{\|s^k\|} = 0. \tag{3.10}$$

Detalhes podem ser vistos em [22, 47, 57]. Perceba que o que deve tender para zero é $B_k - J_F(x^*)$ na direção incremental $\dfrac{s^k}{\|s^k\|}$, e não simplesmente $B_k - J_F(x^*)$. Isto significa que uma sequência (x^k) gerada pelo processo iterativo (3.9) pode convergir superlinearmente, mesmo que a sequência de matrizes (B_k) não convirja para $J_F(x^*)$. Para ver isso, considere o Lema 8.2.7 do livro [22]. Outros métodos quase-Newton para sistemas não-lineares podem ser consultados também em [22].

3.3 Método de Newton Inexato

Já mencionamos que resolver o sistema linear

$$J_F(x^k)s^k = -F(x^k) \tag{3.11}$$

de forma exata, por exemplo por decomposição LU, é caro computacionalmente quando o número de variáveis é "grande"e o esforço computacional não se justifica quando x^k está "longe"de uma solução. Neste sentido, Dembo, Eisenstat e Steihaug [19] propuseram uma classe de métodos que obtém uma solução aproximada para o sistema (3.11) de maneira que o resíduo

$$r^k = J_F(x^k)s^k + F(x^k)$$

satisfaça

$$\left\|r^k\right\| \leq \eta_k \left\|F(x^k)\right\|, \tag{3.12}$$

onde o *termo forçante* $\eta_k \in [0, 1)$ é usado para controlar a precisão do passo. A forma de se obter um s^k que satisfaça (3.12) não é especificada, ou seja, cada método que obtém um s^k satisfazendo (3.12) é um método de Newton inexato diferente. A grosso modo, um método de Newton inexato é qualquer método, onde dado uma aproximação inicial x^0, é gerada uma sequência (x^k) tal que (3.12) seja verificada. Neste sentido, temos o seguinte algoritmo:

Algoritmo 3.1: Método de Newton Inexato

Dado: $x^0 \in \mathbb{R}^n$
$k = 0$
REPITA enquanto $F(x^k) \neq 0$
    Escolha $\eta_k \in [0, 1)$
    Obtenha s^k tal que $\left\|J_F(x^k)s^k + F(x^k)\right\| \leq \eta_k \left\|F(x^k)\right\|$
    Faça $x^{k+1} = x^k + s^k$
    $k = k + 1$
FIM

Usualmente, a iteração que gera o iterado x^k é chamada de *iteração externa* e a iteração que gera o passo s^k é chamada de *iteração interna*. Neste mesmo artigo [19], é provado a convergência local do método de Newton inexato, sob hipóteses razoáveis, como por exemplo jacobiana Lipschitz contínua. Além disso, é provado também que

- se $\eta_k \to 0$, então a taxa de convergência é superlinear e
- se $\eta_k = O\Big(\left\|F(x^k)\right\|\Big)$, então tem-se taxa de convergência quadrática.

Esses resultados são importantes, pois permitem escolher uma sequência (η_k) de modo que o método tenha uma determinada taxa de convergência.

Quando η_k é muito próximo de zero, ocorre um fenômeno chamado de *oversolving*. Isto significa que um grande número de iterações internas serão realizadas a cada iteração externa e pode resultar em um pequeno ou até mesmo nenhum decréscimo da norma de F. Neste sentido, Eisenstat e Walker [27], propuseram duas escolhas para o termo forçante que procuram evitar esse fenômeno.

A primeira escolha reflete a concordância entre a função e seu modelo linear local, isto é, η_k será menor quanto melhor for esta concordância.

Escolha 1: Dado $\eta_0 \in [0, 1)$, escolha

$$\eta_k = \frac{\left\| F(x^k) - F(x^{k-1}) - J_F(x^{k-1})s^{k-1} \right\|}{\|F(x^{k-1})\|}, \quad k = 1, 2, \ldots. \tag{3.13}$$

ou

$$\eta_k = \frac{\left| \, \|F(x^k)\| - \left\| F(x^{k-1}) + J_F(x^{k-1})s^{k-1} \right\| \, \right|}{\|F(x^{k-1})\|}, \quad k = 1, 2, \ldots. \tag{3.14}$$

É demonstrado em [27] a convergência superlinear do Algoritmo 3.1 usando a Escolha 1.

A segunda escolha, não reflete diretamente a concordância entre a função e seu modelo linear local, mas experimentos realizados em [27], sugerem poucos *oversolvings*. Esta escolha depende do decréscimo da norma de F.

Escolha 2: Dados $\gamma \in [0, 1]$, $\beta \in (1, 2]$ e $\eta_0 \in [0, 1)$, escolha

$$\eta_k = \gamma \left(\frac{\|F(x^k)\|}{\|F(x^{k-1})\|} \right)^{\beta}, \quad k = 1, 2, \ldots. \tag{3.15}$$

Também é demonstrado em [27] a convergência do Algoritmo 3.1 usando a Escolha 2. Além disso, se $\gamma < 1$, então a taxa de convergência é da ordem β.

Na prática, é necessário impor salvaguardas de modo a assegurar que os termos da sequência (η_k) não se tornem muito pequenos rapidamente. Em [27] é introduzido para a Escolha 1, a seguinte salvaguarda:

$$\eta_k \leftarrow \max\{\eta_k, \eta_{k-1}^{(1+\sqrt{5})/2}\} \quad \text{quando} \quad \eta_{k-1}^{(1+\sqrt{5})/2} > 0.1,$$

e para a Escolha 2:

$$\eta_k \leftarrow \max\{\eta_k, \gamma\eta_{k-1}^{\beta}\} \quad \text{quando} \quad \gamma\eta_{k-1}^{\beta} > 0.1.$$

É necessário também, impor salvaguardas adicionais de modo que $\eta_k \in [0, 1)$ para cada k.

Vimos que cada método que obtém s^k satisfazendo (3.12) é um método de Newton inexato diferente. Apresentaremos na próxima subseção o método GMRES que obtém uma solução aproximada de um sistema linear. Este método é muito utilizado no método de Newton inexato, sendo conhecido como Newton-GMRES.

3.3.1 GMRES

Apresentaremos nesta seção um dos métodos mais aplicados na prática para obter uma solução aproximada, em um certo sentido, de um sistema linear

$$Ax = b, \tag{3.16}$$

onde $A \in \mathbb{R}^{n\times n}$ e $b \in \mathbb{R}^n$.

É bem conhecido que caso a matriz A seja simétrica definida positiva, o método gradiente conjugado pode ser aplicado e é bastante empregado na prática [35, 65]. Esse método faz parte dos que são chamados métodos sobre espaços de Krylov.

Dados um vetor $v \in \mathbb{R}^n$, uma matriz $A \in \mathbb{R}^{n\times n}$ e um escalar m, definimos o espaço de Krylov $\mathcal{K}_m(A, v)$ como

$$\mathcal{K}_m(A, v) = \text{span}\{v, Av, A^2v, \ldots, A^{m-1}v\}.$$

A princípio, dada uma aproximação inicial x^0, uma solução aproximada x^m pertencente à variedade $x^0 + \mathcal{K}_m(A, v)$ que minimiza a norma residual, pode ser obtida de várias maneiras. Entretanto, sem a hipótese de simetria e positividade da matriz A, se faz necessário a utilização de um método mais geral do que o gradiente conjugado, por exemplo. O objetivo desta seção é apresentar uma síntese do método proposto por Saad e Schultz [66], o GMRES *(Generalized Minimum Residual Method)*, que também pertence a classe de métodos sobre espaços de Krylov.

Dada uma aproximação inicial x^0 e o resíduo $r^0 = b - Ax^0$, defina

$$v_1 = \frac{r^0}{\|r^0\|_2}.$$

O método GMRES é um método de projeção [65], onde a solução aproximada x^m, que minimiza a norma residual, pertence à variedade $x^0 + \mathcal{K}_m(A, v_1)$ e

$$b - Ax^m \perp A\mathcal{K}_m(A, v_1).$$

Isto significa que um vetor z^m deve ser determinado no subespaço $\mathcal{K}_m(A, v_1)$ de modo que

$$z^m = \underset{z \in \mathcal{K}_m(A,v_1)}{\arg\min} \left\| b - A(x^0 + z) \right\|. \tag{3.17}$$

Daí, a solução aproximada dada por GMRES é

$$x^m = x^0 + z^m. \tag{3.18}$$

A implementação do GMRES dada em [66] determina z^m construindo uma base ortonormal $\{v_1, \ldots, v_m\}$ de $\mathcal{K}_m(A, v_1)$ através do método de Arnoldi [2]. Os detalhes serão vistos mais adiante.

O método de Arnoldi é um método que constrói uma base ortonormal $\{v_1, \ldots, v_m\}$ para o espaço $\mathcal{K}_m(A, v_1)$ baseado no processo de Gram-Schmidt. Em cada passo j, o algoritmo multiplica o vetor v_j por A e toma $\hat{v}_j$ como a diferença entre Av_j e a projeção ortogonal de Av_j sobre o subespaço gerado por $\{v_1, \ldots, v_j\}$. O Algoritmo de Arnoldi é apresentado abaixo.

Algoritmo 3.2: Método de Arnoldi

Escolha um vetor v_1 tal que $\|v_1\|_2 = 1$.
PARA $j = 1, 2, \ldots, m$
  PARA $i = 1, 2, \ldots, j$
    $h_{ij} = \langle Av_j, v_i \rangle$
  FIM
  $\hat{v}_j = Av_j - \sum_{i=1}^{j} h_{ij} v_i$
  $h_{j+1,j} = \|\hat{v}_j\|_2$
  Se $h_{j+1,j} = 0$, então pare.
  $v_{j+1} = \dfrac{\hat{v}_j}{h_{j+1,j}}$
FIM

Observe que o algoritmo pode parar na iteração j caso a norma de $\hat{v}_j$ seja zero. Neste caso, o vetor v_{j+1} não pode ser calculado.

Proposição 3.2 *Denote por V_m a matriz $n \times m$ cujas colunas são os vetores $v_1, \ldots, v_m$, e $\bar{H}_m$ a matriz de Hessenberg superior $(m+1) \times m$ cujas entradas não nulas são h_{ij} definidos no Algoritmo 3.2. Definindo $V_{m+1} = \begin{pmatrix} V_m & v_{m+1} \end{pmatrix}$, temos*

$$AV_m = V_{m+1}\bar{H}_m. \tag{3.19}$$

Prova. Pelo Algoritmo 3.2, temos para $j = 1, 2, \ldots, m$ que

$$\begin{aligned} Av_j &= \hat{v}_j + \sum_{i=1}^{j} h_{ij} v_i \\ &= v_{j+1} h_{j+1,j} + \sum_{i=1}^{j} h_{ij} v_i \\ &= \sum_{i=1}^{j+1} h_{ij} v_i \\ &= V_{m+1} \mathrm{col}_j(\bar{H}_m) \end{aligned}$$

Daí segue que $AV_m = V_{m+1}\bar{H}_m$. □

O procedimento de Gram-Schmidt é numericamente instável. Por esse motivo, é comum substituir o procedimento de Gram-Schmidt no Algoritmo 3.2 por Gram-Schmidt modificado [35] ou por transformações Householder [65, 74, 77]. O Algoritmo de Arnoldi com Gram-Schmidt modificado é apresentado a seguir.

Algoritmo 3.3: Método de Arnoldi com Gram-Schmidt modificado

Escolha um vetor v_1 tal que $\|v_1\|_2$=1.
PARA $j = 1, 2, \ldots, m$
  $\hat{v}_j = Av_j$
  PARA $i = 1, 2, \ldots, j$
    $h_{ij} = \langle \hat{v}_j, v_i \rangle$
    $\hat{v}_j = \hat{v}_j - h_{ij} v_i$
  FIM
  $h_{j+1,j} = \|\hat{v}_j\|_2$
  Se $h_{j+1,j} = 0$, então pare.
  $v_{j+1} = \dfrac{\hat{v}_j}{h_{j+1,j}}$
FIM

Agora, para determinar z^m dado em (3.17), considere a matriz V_m dada na Proposição 3.2. Vamos obter um vetor $y^m \in \mathbb{R}^m$ de tal modo que

$$z^m = V_m y^m.$$

Para isso, observe que se $x \in x^0 + \mathcal{K}_m(A, v_1)$, da relação (3.19), temos que

$$\begin{aligned} b - Ax &= b - A(x^0 + V_m y) \\ &= r_0 - AV_m y \\ &= \left\|r^0\right\|_2 v_1 - V_{m+1}\bar{H}_m y \\ &= V_{m+1}(\left\|r^0\right\|_2 e_1 - \bar{H}_m y) \end{aligned}$$

para algum $y \in \mathbb{R}^m$. Como as colunas de V_{m+1} são ortonormais, temos

$$\|b - Ax\|_2 = \left\|\left\|r^0\right\|_2 e_1 - \bar{H}_m y\right\|_2 .$$

Desta maneira, tomando

$$y^m = \underset{y \in \mathbb{R}^m}{\arg\min} \left\|\left\|r^0\right\|_2 e_1 - \bar{H}_m y\right\|_2 \tag{3.20}$$

temos

$$x^m = x^0 + z^m = x^0 + V_m y^m.$$

Esta abordagem é muito vantajosa, pois ao invés de obter um $\tilde{x} \in \mathbb{R}^n$ que minimiza $\|b - Ax\|_2$, obtemos um $x^m \in x^0 + \mathcal{K}_m(A, v_1)$, com $m \ll n$ que é solução do problema

$$\begin{array}{ll} \text{minimizar} & \|b - Ax\|_2 \\ \text{sujeito a} & x \in x^0 + \mathcal{K}_m(A, v_1) \\ & b - Ax \perp A\mathcal{K}_m(A, v_1) \end{array} \tag{3.21}$$

Obter o vetor y^m dado em (3.20), é de certa forma uma tarefa fácil. Como a matriz $\bar{H}_m$ é Hessenberg-superior, podemos decompô-la em $\bar{H}_m = Q_m \bar{R}_m$, onde

$$Q_m \in \mathbb{R}^{(m+1)\times(m+1)}$$

é um produto de matrizes de rotação de Givens e

$$\bar{R}_m \in \mathbb{R}^{(m+1)\times m}$$

é triangular superior. Para ver isso, denote Ω_i a matriz de rotação de Givens

$$\Omega_i = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c_i & s_i & & & & \text{i} \\ & & & -s_i & c_i & & & & \text{i+1} \\ & & & & & 1 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \quad \begin{matrix} \\ \\ \\ \text{i} \\ \text{i+1} \\ \\ \\ \\ \end{matrix}$$

onde $c_i^2 + s_i^2 = 1$. Desta maneira, pré-multiplicamos a matriz de Hessenberg superior $\bar{H}_m$ e o vetor $g_0 = \|r^0\|_2 e_1$ por m matrizes de Givens, ou seja, definindo

$$Q_m^T = \Omega_m \Omega_{m-1} \dots \Omega_1 \tag{3.22}$$

temos que

$$\bar{R}_m \stackrel{def}{=} Q_m^T \bar{H}_m \tag{3.23}$$

é triangular superior e

$$\bar{g}_m \stackrel{def}{=} Q_m^T(\left\|r^0\right\|_2 e_1) = (\gamma_1, \dots, \gamma_{m+1})^T. \tag{3.24}$$

Como Q_m é ortogonal,

$$\left\|\left\|r^0\right\|_2 e_1 - \bar{H}_m y\right\|_2 = \left\|\bar{g}_m - \bar{R}_m y\right\|_2. \tag{3.25}$$

É importante observar que a última linha da matriz $\bar{R}_m$ é nula. Definimos então a matriz $R_m \in \mathbb{R}^{m \times m}$ tal que

$$\bar{R}_m = \begin{pmatrix} R_m \\ 0 \end{pmatrix} \tag{3.26}$$

e o vetor $g_m \in \mathbb{R}^m$ tal que

$$\bar{g}_m = \begin{pmatrix} g_m \\ \gamma_{m+1} \end{pmatrix}. \tag{3.27}$$

A próxima proposição mostra que y^m é obtido resolvendo um sistema triangular superior e fornece uma expressão para o resíduo no passo m.

Proposição 3.3 *Sejam $m \leq n$, Ω_i, $i = 1, \dots, m$ matrizes de rotação de Givens usadas*

para transformar a matriz de Hessenberg $\bar{H}_m$ em uma matriz triangular superior $\bar{R}_m$, $\bar{g}_m$ dada por (3.24) *e g_m dada por* (3.27)*. Temos*

1. *O vetor y^m que minimiza $\left\| \|r^0\|_2 e_1 - \bar{H}_m y \right\|$ é dado por*

$$y^m = R_m^{-1} g_m.$$

2. *O resíduo no passo m é*

$$\|b - Ax^m\|_2 = |\gamma_{m+1}|. \tag{3.28}$$

Prova. Para provar a primeira parte, observe que

$$\left\| \|r^0\|_2 e_1 - \bar{H}_m y \right\|^2 = \left\| \bar{g}_m - \bar{R}_m y \right\|_2^2 = \left\| \begin{pmatrix} g_m - R_m y \\ \gamma_{m+1} \end{pmatrix} \right\|_2^2 = |\gamma_{m+1}|^2 + \|g_m - R_m y\|_2^2 .$$

Como γ_{m+1} não depende de y, o mínimo é atingido quando $\|g_m - R_m y\|_2 = 0$. Ora, como R_m é não singular, temos que $y^m = R_m^{-1} g_m$. Desta maneira, a segunda parte é imediata. □

Se o resíduo (3.28) não satisfaz a precisão requerida, então um passo a mais no Algoritmo de Arnoldi é necessário, ou seja, fazemos $j = m+1$ no Algoritmo de Arnoldi. Desta forma, obtemos uma base V_{m+1} para o espaço $\mathcal{K}_{m+1}(A, v_1)$ e a $(m+1)$-ésima coluna da matriz $\bar{H}_{m+1}$. Vemos que

$$\mathrm{col}_{m+1}(\bar{H}_{m+1}) = \begin{pmatrix} h_{1,m+1} \\ h_{2,m+1} \\ \vdots \\ h_{m+1,m+1} \\ h_{m+2,m+1} \end{pmatrix}, \tag{3.29}$$

onde $h_{m+2,m+1} = \|\hat{v}_{m+1}\|_2$. Não é necessário efetuar todos os cálculos novamente. Os próprios autores de [66] sugerem anexar a $(m+1)$-ésima coluna de $\bar{H}_{m+1}$ à matriz $\bar{R}_m$ completando os restantes dos elementos da última linha por zeros. Desta maneira tem-se

$$\begin{pmatrix} & & & h_{1,m+1} \\ & \bar{R}_m & & \vdots \\ & & & h_{m+1,m+1} \\ 0 & \dots & 0 & h_{m+2,m+1} \end{pmatrix}. \tag{3.30}$$

Para construir a matriz triangular superior $\bar{R}_{m+1}$ aplica-se as m matrizes de rotação de Givens, $\Omega_1, \dots, \Omega_m$, obtidas anteriormente, no vetor dado em (3.29). Feito isso, apenas o

elemento $h_{m+2,m+1}$ não foi zerado. Definimos então a nova matriz de rotação de Givens Ω_{m+1} e efetuamos o cálculo necessário para zerar $h_{m+2,m+1}$. O mesmo é feito para construir o vetor $\bar{g}_{m+1}$. Para combinar a dimensão, anexamos um zero na última linha de $\bar{g}_m$ e pré-multiplicamos a rotação de Givens Ω_{m+1} por esse novo vetor. Desta forma, temos obtido a matriz triangular superior $\bar{R}_{m+1}$ e o vetor

$$\bar{g}_{m+1} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \\ c_{m+1}\gamma_{m+1} \\ -s_{m+1}\gamma_{m+1} \end{pmatrix}. \tag{3.31}$$

Se a norma residual $|-s_{m+1}\gamma_{m+1}|$ for suficientemente pequena, então calculamos y_{m+1}, solução do sistema triangular superior

$$R_{m+1}y = g_{m+1},$$

onde R_{m+1} e g_{m+1} são definidos de forma semelhante a (3.26) e (3.27), respectivamente. Daí, a solução aproximada $x_{m+1} = x^0 + V_{m+1}y_{m+1}$ é calculada.

Em [66], é provado que o método GMRES falha, se na iteração j do Algoritmo de Arnoldi ocorrer $h_{j+1,j} = 0$. Isto ocorre, se e somente se, x_j for a solução exata do sistema linear (3.16). Este fato é chamado de *lucky breakdown*, veja [66]. Uma observação importante é que a convergência do método GMRES se dá no máximo em n iterações. A discussão acima sobre o método GMRES está sintetizada no Algoritmo 3.4.

Algoritmo 3.4: Método GMRES

Calcule $r^0 = b - Ax^0$.
Faça $v_1 = \dfrac{r^0}{\|r^0\|_2}$.
Construa as matrizes V_m e $\bar{H}_m$ usando o método de Arnoldi (Algoritmo 3.2 ou 3.3) começando com v_1.
Calcule y^m que minimiza $\left\|\|r^0\|_2 e_1 - \bar{H}_m y\right\|$.
Faça $x^m = x^0 + V_m y^m$.

Em problemas de grande porte, o método GMRES torna-se impraticável, pois enquanto a norma residual não satisfaz a precisão requerida, aumenta-se a dimensão do espaço de Krylov e consequentemente o custo de memória. Existem duas possibilidades para contornar esse problema. A primeira é utilizar pré-condicionadores, cujo objetivo é

reduzir o número de iterações necessárias para atingir a convergência. A segunda possibilidade é executar um *restart* ou recomeço. Isto significa que a cada m iterações, o método GMRES recomeça formando um novo ciclo com $x^0 = x^m$, ou seja, a nova aproximação inicial passa a ser x^m calculada no ciclo anterior e o resíduo r_m será usado para gerar o novo espaço de Krylov.

É bem conhecido [66], que o método GMRES com recomeços pode estagnar quando a matriz dos coeficientes não é definida positiva. Além disso, nem sempre há convergência.

3.4 Métodos Tensoriais

Até agora, todos os métodos que apresentamos neste capítulo, são métodos consagrados na literatura. Embora de longa data, poucas pesquisas sobre métodos tensoriais têm sido publicadas. Esses métodos foram, de certa forma, esquecidos pela comunidade científica. Um dos motivos pode ter sido o alto custo computacional e a dificuldade de diferenciação, por exemplo. Mas com o avanço tecnológico, novos métodos de resolução de sistemas lineares, novos métodos de otimização e novas técnicas de diferenciação como diferenciação automática, por exemplo, os métodos tensoriais têm sido retomados por alguns pesquisadores. Nossa pesquisa insere-se nesse contexto.

Vimos que dada uma boa estimativa inicial $x^0 \in \mathbb{R}^n$, o método de Newton gera uma sequência (x^k), tal que x^{k+1} é o zero do modelo linear

$$M_k(x) = F(x^k) + J_F(x^k)(x - x^k),$$

onde $J_F(x^k)$ denota a jacobiana de F avaliada em x^k.

Se derivadas de alta ordem são embutidas no modelo, temos o que chamamos de modelos tensoriais. O mais simples é o modelo quadrático,

$$M_k(x) = F(x^k) + J_F(x^k)(x - x^k) + \frac{1}{2}\mathcal{T}_F(x^k)(x - x^k)(x - x^k), \tag{3.32}$$

onde o tensor $\mathcal{T}_F(x^k) \in \mathbb{R}^{n\times n\times n}$ denota a segunda derivada de F em x^k.

Existem várias desvantagens quando se toma esse modelo para F em x^k. Por exemplo:

- calcular n^3 derivadas de segunda ordem a cada iteração;
- problemas de armazenamento, pelo menos da ordem de $n^3/2$;
- encontrar um zero do modelo de n equações quadráticas e n variáveis.
- o modelo pode não ter um zero.

Seguindo a primícia dos métodos quase-Newton, é desejável formar um tensor $\mathcal{T}_k$ que seja mais barato computacionalmente que o tensor $\mathcal{T}_F(x^k)$, de modo que evite as desvantagens listadas anteriormente ou pelo menos algumas. Um dos primeiros trabalhos nesse sentido, foi introduzido por Schnabel e Frank [68].

3.4.1 Método Tensorial de Schnabel e Frank

No artigo de Schnabel e Frank [68], é introduzido um novo modelo tensorial para F em x^k, onde o custo computacional para formá-lo é no máximo $O(n^{2.5})$ multiplicações e adições por iteração, ou seja, é um modelo um pouco mais barato que o modelo (3.32).

A estratégia adotada em [68] é escolher $\mathcal{T}_k \in \mathbb{R}^{n\times n\times n}$ de maneira que o modelo quadrático

$$\hat{M}_k(x) = F(x^k) + J_F(x^k)(x - x^k) + \frac{1}{2}\mathcal{T}_k(x - x^k)(x - x^k), \tag{3.33}$$

interpole F em $p \leq \sqrt{n}$ iterados já determinados, $x^{-1}, \ldots, x^{-p}$, não necessariamente consecutivos, ou seja,

$$\hat{M}(x^{-c}) = F(x^{-c})$$

para $c = 1, \ldots, p$. Isto significa que o sistema

$$F(x^{-c}) = F(x^k) + J_F(x^k)s_c + \frac{1}{2}\mathcal{T}_k s_c s_c \tag{3.34}$$

onde $s_c = x^{-c} - x^k$ para $c = 1, \ldots, p$, deve ser verificado.

Note que (3.34) é formado por $np \leq n^{1.5}$ equações lineares e n^3 incógnitas. Como (3.34) é sobredeterminado, procura-se escolher $\mathcal{T}_k$ tal que seja solução do seguinte problema de minimização:

$$\begin{aligned} &\text{minimizar} && \|\mathcal{T}_k\|_F \\ &\text{sujeito a} && \mathcal{T}_k s_c s_c = z_c, \quad c = 1, \ldots p, \end{aligned} \tag{3.35}$$

onde $\|\mathcal{T}_k\|_F$ denota a norma de Frobenius de $\mathcal{T}_k$ definida por

$$\|\mathcal{T}_k\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n}(t_{ij}^r)^2,$$

onde t_{ij}^r são os elementos do tensor $\mathcal{T}_k$ e

$$z_c = 2\Big(F(x^{-c}) - F(x^k) - J_F(x^k)s_c)\Big).$$

Como demonstrado em [68], a solução deste problema é

$$\mathcal{T}_k = \sum_{c=1}^{p} a_c \otimes s_c \otimes s_c \tag{3.36}$$

onde $\otimes$ denota o produto de Kronecker, a_c denota a c-ésima coluna da matriz

$$A = ZM^{-1},$$

onde os elementos de $M \in \mathbb{R}^{p\times p}$ são definidos por $m_{ij} = (s_i^T s_j)^2$ para $1 \leq i, j \leq p$ e as p colunas da matriz $Z \in \mathbb{R}^{n\times p}$ são os vetores z_c.

A expressão (3.36) significa que $\mathcal{T}_k$ é uma soma de p tensores de posto 1. Além disso, as camadas horizontais de $\mathcal{T}_k$ são simétricas, o que é desejável do ponto de vista de armazenamento. O leitor interessado em posto de tensores bem como produto de Kronecker pode consultar, respectivamente, as referências [5, 52] e [55].

Substituindo (3.36) em (3.33), tem-se o modelo tensorial proposto por [68], que é

$$M_T(x^k + d) = F(x^k) + J_F(x^k)d + \frac{1}{2}\sum_{c=1}^{p} a_c(d^T s_c)^2. \tag{3.37}$$

Desta maneira procura-se encontrar um $d \in \mathbb{R}^n$ tal que $M_T(x^k + d) = 0$. Os autores mostram que o custo computacional para formar o modelo tensorial (3.37) é no máximo $O(n^{2.5})$ multiplicações e adições por iteração e que para formar o tensor (3.36) são necessárias $n^2p + O(np^2)$ multiplicações e adições.

Os algoritmos propostos em [68, 69], procuram resolver o problema

$$\underset{d\in\mathbb{R}^n}{\text{minimizar}} \left\| M_T(x^k + d) \right\|_2, \tag{3.38}$$

ou seja, quando o modelo M_T não possui um zero real, os algoritmos encontram um minimizador do modelo tensorial M_T. Detalhes sobre alguns procedimentos para resolver o subproblema (3.38) podem ser consultados em [33, 68, 69] e análise de convergência em [30].

Outras estratégias, de certa forma mais modernas, para resolver o subproblema (3.38) têm sido publicadas, por exemplo, em [3, 4, 9, 10, 31].

Apesar de serem relativamente antigos, os métodos tensoriais baseados em Schnabel e Frank [68], têm sido retomados recentemente na literatura por alguns pesquisadores.

O algoritmo tensor-GMRES proposto por Dan Feng e Thomas H. Pulliam [31], pode ser visto como uma extensão do método de Newton Inexato usando GMRES. O passo é calculado utilizando informações do espaço de Krylov gerado pelo passo de Newton

Inexato. Já no trabalho de Brett W. Bader [3, 4], três métodos baseados no modelo (3.37) com $p = 1$ foram propostos. Eles procuram resolver o problema de minimização (3.38) de maneira inexata, utilizando métodos sobre espaços de Krylov. Mais especificamente, procuram

$$\underset{d\in\mathcal{K}_m}{\text{minimizar}} \left\| F(x^k) + J_F(x^k) + \frac{1}{2}a_k(s_k^T d)^2 \right\|_2, \tag{3.39}$$

onde

$$a_k = \frac{2(F(x^{k-1}) - F(x^k) - J_F(x^k)s_k)}{(s_k^T s_k)^2} \quad \text{e} \quad s_k = x^{k-1} - x^k$$

e $\mathcal{K}_m$ é um subespaço de Krylov. O método GMRES também é utilizado nesses métodos.

A performance de alguns métodos tensoriais em problemas mal-condicionados ou singulares, tem sido estudada e verificada recentemente por Bader e Schnabel em [7]. Nesses tipos de problemas, métodos baseados em Newton apresentam convergência muito lenta.

3.4.2 Classe Chebyshev-Halley: Caso Multidimensional

Os métodos da Classe Chebyshev-Halley podem ser vistos como métodos tensoriais por fazerem uso do tensor. Assim como no método de Newton, os métodos de Chebyshev (2.11), Halley (2.14) e Super-Halley (2.26) unidimensionais , podem ser facilmente estendidos para o espaço $\mathbb{R}^n$. Os métodos Halley e Chebyshev foram estendidos por Mertvecova [59] em 1953 e Nečepuerenko [61] em 1954, respectivamente.

Denote I a matriz identidade e, para cada $x \in \mathbb{R}^n$, considere a matriz

$$\mathcal{L}(x) = J_F(x)^{-1}\mathcal{T}_F(x)\Big(J_F(x)^{-1}F(x)\Big). \tag{3.40}$$

A matriz $\mathcal{L}(x)$ generaliza o grau de convexidade logarítmica definido em (2.4).

No caso multidimensional[1], o método de Chebyshev (2.11) pode ser escrito como

$$x^{k+1} = x^k - \Big[I + \frac{1}{2}\mathcal{L}(x^k)\Big]J_F(x^k)^{-1}F(x^k), \tag{3.41}$$

o método de Halley (2.14) como

$$x^{k+1} = x^k - \Big[I + \frac{1}{2}\mathcal{L}(x^k)\Big(I - \frac{1}{2}\mathcal{L}(x^k)\Big)^{-1}\Big]J_F(x^k)^{-1}F(x^k) \tag{3.42}$$

e o método Super-Halley (2.26) como

$$x^{k+1} = x^k - \Big[I + \frac{1}{2}\mathcal{L}(x^k)\Big(I - \mathcal{L}(x^k)\Big)^{-1}\Big]J_F(x^k)^{-1}F(x^k). \tag{3.43}$$

[1] Observamos que os métodos Chebyshev e Halley, podem ser obtidos modificando o modelo quadrático (3.32), substituíndo o termo $\mathcal{T}_F(x^k)(x - x^k)(x - x^k)$ por $\mathcal{T}_F(x^k)(-J_F(x^k)^{-1}F(x^k))(-J_F(x^k)^{-1}F(x^k))$ e $\mathcal{T}_F(x^k)(x - x^k)(-J_F(x^k)^{-1}F(x^k))$, respectivamente.

No artigo de Hernández e Gutiérrez [42] é definida, para $\alpha \in [0,1]$, a seguinte classe de métodos:

$$x^{k+1} = x^k - \Big[I + \frac{1}{2}\mathcal{L}(x^k)\Big(I - \alpha\mathcal{L}(x^k)\Big)^{-1}\Big] J_F(x^k)^{-1}F(x^k), \tag{3.44}$$

para espaços de Banach, o que generaliza a classe de Hernández e Salanova [43] dada por (2.27). Em (3.44), tem-se o método de Chebyshev (3.41) pondo $\alpha = 0$, o método de Halley (3.42) pondo $\alpha = \dfrac{1}{2}$ e o método Super-Halley (3.43) pondo $\alpha = 1$. Essa classe de métodos é chamada pelos próprios autores [42] de Classe Chebyshev-Halley. Além disso, foi estabelecido convergência semilocal segundo hipóteses tipo Kantorovich.

No entanto, é apresentada no livro *Numerische Lösung Nichtlinearer Gleichungen* do Professor Hubert Schwetlick, veja referência [70], a seguinte classe de métodos parametrizados por um escalar $\gamma \in \mathbb{R}$ e um $i \in \mathbb{N}$:

$$\begin{aligned} &F(x^k) + J_F(x^k)(y^{k,i+1} - x^k) + \frac{\gamma}{2}\mathcal{T}_F(x^k)(y^{k,i} - x^k)(y^{k,i+1} - x^k) + \\ &+\frac{1-\gamma}{2}\mathcal{T}_F(x^k)(y^{k,i} - x^k)(y^{k,i} - x^k) = 0 \end{aligned} \tag{3.45}$$

com

$$y^{k,0} = x^k \quad \text{e} \quad x^{k+1} = y^{k,i+1}.$$

Em [70] é definido, pondo $x^{k+1} = y^{k,2}$ em (3.45), o seguinte algoritmo:

Algoritmo 3.5: Algoritmo de Schwetlick

Dados: $x^0 \in \mathbb{R}^n, \gamma \in \mathbb{R}$
$k = 0$
REPITA enquanto $F(x^k) \neq 0$
  Calcule y^k a partir da equação
  $F(x^k) + J_F(x^k)(y^k - x^k) = 0$
  Calcule x^{k+1} a partir da equação
  $F(x^k) + \Big[J_F(x^k) + \dfrac{\gamma}{2}\mathcal{T}_F(x^k)(y^k - x^k)\Big](x^{k+1} - x^k) +$
  $+\dfrac{1-\gamma}{2}\mathcal{T}_F(x^k)(y^k - x^k)(y^k - x^k) = 0$
  $k = k + 1$
FIM

Note que se $\gamma = 0$ tem-se o método de Chebyshev (3.41), o método de Halley (3.42) é obtido com $\gamma = 1$ e o método Super-Halley (3.43) é obtido com $\gamma = 2$. Essas equivalências são facilmente verificadas usando a igualdade

$$I + \frac{1}{2}\mathcal{L}(x^k)(I - \alpha\mathcal{L}(x^k))^{-1} = (I - \alpha\mathcal{L}(x^k))^{-1}\Big(I + \Big(\frac{1}{2} - \alpha\Big)\mathcal{L}(x^k)\Big).$$

Em [26, 70] é provado, sob hipóteses razoáveis, a convergência cúbica da sequência (x^k) gerada pelo Algoritmo 3.5 para qualquer valor real γ. Sendo assim, em particular os métodos de Chebyshev, Halley e Super-Halley convergem cubicamente.

Pois bem, com relação à Classe Chebyshev-Halley (3.44), Gundersen e Steihaug [37], mostraram recentemente, que para $i = 1$ em (3.45), ou seja, $x^{k+1} = y^{k,2}$, a Classe Chebyshev-Halley (3.44), agora com $\alpha \in \mathbb{R}$, e a classe de métodos baseada nos modelos dados em (3.45) são equivalentes. Logo, em particular, a taxa de convergência da sequência (x^k) gerada por qualquer método da Classe Chebyshev-Halley (3.44), com $\alpha \in \mathbb{R}$, é cúbica. Além disso, mostraram que a Classe Chebyshev-Halley pode ser escrita como:

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) \\ \Big(J_F(x^k) + \alpha\mathcal{T}_F(x^k)s^k_{(1)}\Big)s^k_{(2)} &= -\frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)}, \\ x^{k+1} &= x^k + s^k_{(1)} + s^k_{(2)}. \end{aligned} \tag{3.46}$$

A abordagem (3.46) é extremamente importante, pois o passo s^k pode ser decomposto como a soma $s^k_{(1)} + s^k_{(2)}$. Esta é uma forma bem mais eficiente que a abordagem (3.44), pois não é necessário obter a matriz $\mathcal{L}(x^k)$ a cada iteração, o que é extremamente caro computacionalmente, já que seria necessário resolver $n + 1$ sistemas lineares, enquanto que em (3.46), apenas dois sistemas lineares são necessários. Note que no método de Chebyshev ($\alpha = 0$), a matriz dos coeficientes dos dois sistemas lineares (3.46) é a jacobiana $J_F(x^k)$. Isto significa que se os sistemas lineares são resolvidos via decomposição, por exemplo LU, apenas uma decomposição será necessária. Por esse motivo, o método de Chebyshev tem sido utilizado com mais frequência, principalmente em problemas de otimização irrestrita [20, 79]. Nesses problemas a matriz dos coeficientes é a Hessiana da função objetivo e os dois sistemas lineares são resolvidos, preferencialmente via método gradiente-conjugado.

A prova dada por Gundersen e Steihaug [37], é basicamente a que segue.

Lema 3.4 *Considere a Classe Chebyshev-Halley* (3.44). *Esta classe pode ser escrita como* (3.46).

Prova. Pelo Lema 1.3, podemos observar que

$$\begin{aligned} I + \frac{1}{2}\mathcal{L}(x^k)(I - \alpha\mathcal{L}(x^k))^{-1} &= I + \frac{1}{2}(I - \alpha\mathcal{L}(x^k))^{-1}\mathcal{L}(x^k) \\ &= (I - \alpha\mathcal{L}(x^k))^{-1}\Big(I - \alpha\mathcal{L}(x^k) + \frac{1}{2}\mathcal{L}(x^k)\Big) \\ &= (I - \alpha\mathcal{L}(x^k))^{-1}\Big(I + \Big(\frac{1}{2} - \alpha\Big)\mathcal{L}(x^k)\Big). \end{aligned} \tag{3.47}$$

Agora defina

$$s^k_{(1)} = -J_F(x^k)^{-1}F(x^k)$$

e

$$s^k_{(2)} = x^{k+1} - x^k - s^k_{(1)}.$$

Daí segue que

$$\begin{aligned}(I - \alpha\mathcal{L}(x^k))(s^k_{(1)} + s^k_{(2)}) &= (I - \alpha\mathcal{L}(x^k))(x^{k+1} - x^k) \\ &= (I - \alpha\mathcal{L}(x^k))(I - \alpha\mathcal{L}(x^k))^{-1}\Big(I + \Big(\frac{1}{2} - \alpha\Big)\mathcal{L}(x^k)\Big)s^k_{(1)} \\ &= (I - \alpha\mathcal{L}(x^k))s^k_{(1)} + \frac{1}{2}\mathcal{L}(x^k)s^k_{(1)}.\end{aligned}$$

Com isso

$$(I - \alpha\mathcal{L}(x^k))s^k_{(2)} = \frac{1}{2}\mathcal{L}(x^k)s^k_{(1)}.$$

Multiplicando por $J_F(x^k)$ em ambos os lados e usando a definição de $\mathcal{L}(x)$ dada em (3.40), obtemos

$$(J_F(x^k) + \alpha\mathcal{T}_F(x^k)s^k_{(1)})s^k_{(2)} = -\frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)}.$$

Desta forma, temos (3.46). □

3.4.3 Algumas Variações da Classe Chebyshev-Halley

Apesar dos métodos da Classe Chebyshev-Halley serem muito atrativos para resolver o problema (3.1), por terem taxa de convergência cúbica, eles são computacionalmente caros, basicamente por dois motivos:

1. necessidade de se obter o tensor $\mathcal{T}_F(x^k)$ a cada iteração e
2. resolver de forma exata dois sistemas lineares.

Apresentaremos nesta subseção uma aproximação para o tensor $\mathcal{T}_F(x^k)$ utilizando diferenças finitas baseada no trabalho [26], e um algoritmo baseado no trabalho de Steihaug e Suleiman [73], que procura encontrar um zero aproximado para o modelo quadrático de F em torno de x^k utilizando ideias da Classe Chebyshev-Halley.

Classe Chebyshev-Halley Discreta

Algumas aproximações para o tensor $\mathcal{T}_F(x)$, relativamente antigas, foram publicadas em [26, 63, 76]. Aqui vamos nos restringir ao artigo [26] de Ehle e Schwetlick de 1976. Para isso, considere a seguinte definição:

Definição 3.5 *Seja $F : \mathbb{R}^n \to \mathbb{R}^n$ duas vezes diferenciável. Dizemos que $\mathcal{B} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{n\times n\times n}$ é uma aproximação fortemente consistente para $\mathcal{T}_F$ quando existem $c, r \geq 0$ tais que*

$$\|\mathcal{B}(x,h) - \mathcal{T}_F(x)\| \leq c|h|$$

para todo $x \in \mathbb{R}^n$ e para todo h tal que $|h| < r$.

Uma definição mais geral pode ser vista em [26, Definição 3.1].

Com hipótese Lipschitz sobre $\mathcal{T}_F$, um exemplo de aproximação fortemente consistente para $\mathcal{T}_F$, é aproximar $\mathcal{T}_F(x)$ usando diferenças finitas, ou seja, construímos um tensor $\mathcal{B}(x,h)$ tal que a q-ésima camada lateral de $\mathcal{B}(x,h)$ é

$$\mathcal{B}^{:q:} = \mathcal{B}e_q = \frac{J_F(x+he_q) - J_F(x)}{h} \tag{3.48}$$

para todo $q = 1 \ldots n$, onde h é um parâmetro de discretização, podendo ser diferente para cada derivada parcial. Este fato pode ser visto em [26].

A expressão (3.48) significa que cada camada horizontal $\nabla^2 f_i(x)$ com $i = 1, \ldots, n$ do tensor $\mathcal{T}_F(x)$, pode ser aproximada por diferenças de gradientes, ou seja,

$$\mathrm{col}_q(\nabla^2 f_i(x)) \approx \frac{\nabla f_i(x+he_q) - \nabla f_i(x)}{h}.$$

Neste sentido, definimos a Classe Chebyshev-Halley discreta como

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) \\ \Big(J_F(x^k) + \alpha\mathcal{B}(x^k,h_k)s^k_{(1)}\Big)s^k_{(2)} &= -\frac{1}{2}\mathcal{B}(x^k,h_k)s^k_{(1)}s^k_{(1)}, \\ x^{k+1} &= x^k + s^k_{(1)} + s^k_{(2)}. \end{aligned} \tag{3.49}$$

Algumas aproximações fortemente consistentes para o tensor $\mathcal{T}_F(x^k)$ foram utilizadas em um algoritmo proposto em [26], em particular a aproximação (3.48). Com essa abordagem, foi provado a convergência quadrática da Classe Chebyshev-Halley discreta (3.49). Além disso, se $h_k \to 0$, então a taxa de convergência é superquadrática, no sentido que

$$\left\|x^{k+1} - x^*\right\| \leq \varepsilon_k \left\|x^k - x^*\right\|^2 \quad \text{com} \quad \varepsilon_k \to 0 \tag{3.50}$$

e se

$$h_k = O\Big(\left\|F(x^k)\right\|\Big),$$

então a convergência cúbica é garantida. Veja [26, Teorema 3.3].

Classe Chebyshev-Halley Inexata

A Classe Chebyshev-Halley Inexata introduzida no artigo de Steihaug e Suleiman [73], publicado recentemente, foi motivada pelo fato de que um passo do método Super-Halley utilizado para encontrar um zero de F é equivalente a dois passos do método de Newton aplicados na aproximação quadrática de F em torno de x^k [37]. Para ver isso, note que um passo s^k do método Super-Halley é encontrado resolvendo o sistema (3.46) com $\alpha = 1$, ou seja,

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) \\ \Big(J_F(x^k) + \mathcal{T}_F(x^k)s^k_{(1)}\Big)s^k_{(2)} &= -\frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)}, \\ s^k &= s^k_{(1)} + s^k_{(2)}. \end{aligned}$$

Os vetores $s^k_{(1)}$ e $s^k_{(2)}$ são unicamente determinados supondo que as matrizes $J_F(x^k)$ e $J_F(x^k) + \mathcal{T}_F(x^k)s^k_{(1)}$ sejam não singulares.

O método de Newton aplicado duas vezes na quadrática

$$M_k(s) = F(x^k) + J_F(x^k)s + \frac{1}{2}\mathcal{T}_F(x^k)ss, \tag{3.51}$$

começando com $s^0 = 0$, gera os sistemas

$$\begin{aligned} J_{M_k}(0)s^1 &= -M_k(0) \\ J_{M_k}(s^1)s^2 &= -M_k(s^1), \end{aligned} \tag{3.52}$$

onde J_{M_k} denota a jacobiana de M_k. Vamos mostrar que $s^k_{(1)} = s^1$ e $s^k_{(2)} = s^2$. Temos que

$$J_{M_k}(s) = J_F(x^k) + \mathcal{T}_F(x^k)s \quad \text{e} \quad \mathcal{T}_{M_k}(s) = \mathcal{T}_F(x^k).$$

Como $J_{M_k}(0) = J_F(x^k)$ e $M_k(0) = F(x^k)$, vemos facilmente que $s^k_{(1)} = s^1$ e de

$$J_{M_k}(s^1) = J_F(x^k) + \mathcal{T}_F(x^k)s^1$$

e $M_k(s^1) = \dfrac{1}{2}\mathcal{T}_F(x^k)s^1s^1$, temos $s^k_{(2)} = s^2$.

Isto significa que o método Super-Halley pode ser definido como um método que, a cada iteração, obtém um *"zero aproximado"* do modelo quadrático (3.51) usando dois passos do método de Newton no modelo quadrático (3.51).

Determinar os zeros do modelo quadrático (3.51) não é tarefa fácil devido as desvantagens citadas no início da Seção 3.4, principalmente porque os zeros podem nem existir.

Como visto na Seção 3.4.1, na estratégia adotada por Schnabel e Frank [68], o tensor $\mathcal{T}_F(x^k)$ é aproximado pelo tensor $\mathcal{T}_k$ dado em (3.36), e então procura-se um zero para o modelo quadrático (3.37) resolvendo o problema de minimização (3.38). Ao contrário

dessa estratégia, com o objetivo de resolver o problema (3.1), Steihaug e Suleiman [73] propuseram um algoritmo que consiste em encontrar um zero aproximado para o modelo quadrático (3.51) a cada iteração, de modo que o resíduo

$$r^k = \frac{1}{2}\mathcal{T}_F(x^k)s^k s^k + J_F(x^k)s^k + F(x^k)$$

satisfaça

$$\left\|r^k\right\| \leq \eta_k \left\|F(x^k)\right\|, \tag{3.53}$$

onde $\eta_k \in [0, 1)$ é o termo forçante, também usado para controlar a precisão do passo como no método de Newton inexato. Observe que nenhuma aproximação para o tensor $\mathcal{T}_F(x^k)$ é utilizada. Neste sentido, segue adiante o algoritmo de Steihaug e Suleiman.

Algoritmo 3.6: Algoritmo de Steihaug e Suleiman [73]

Dado: $x^0 \in \mathbb{R}^n$

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

Encontrar uma solução aproximada s^k para $M_k(s) = 0$ tal que para $\eta_k \leq \eta < 1$, $\left\|\frac{1}{2}\mathcal{T}_F(x^k)s^k s^k + J_F(x^k)s^k + F(x^k)\right\| \leq \eta_k \left\|F(x^k)\right\|$.

Faça $x^{k+1} = x^k + s^k$

$k = k + 1$

FIM

Podemos entender o Algoritmo 3.6 como uma extensão do método de Newton inexato, pois ao invés de exigir um decréscimo suficiente no modelo linear, é exigido um decréscimo suficiente no modelo quadrático. Destacamos uma diferença sutil a respeito do termo forçante entre o Algoritmo 3.6 de Steihaug e Suleiman e o método de Newton inexato. No método de Newton inexato, η_k pode ser dado a priori e no Algoritmo 3.6 não, ou seja, dado um $\eta_k \in [0, 1)$, nem sempre é possível obter um s^k tal que

$$\left\|\frac{1}{2}\mathcal{T}_F(x^k)s^k s^k + J_F(x^k)s^k + F(x^k)\right\| \leq \eta_k \left\|F(x^k)\right\|. \tag{3.54}$$

Para ver isso, note na Figura 3.1 que qualquer que seja $\eta_k \in [0, 0.36)$, não existe $s^k \in \mathbb{R}$ tal que a condição (3.54) seja verificada.

Como é de se esperar, assim como no método de Newton inexato, o termo forçante η_k tem um papel fundamental na taxa de convergência da sequência (x^k) gerada pelo Algoritmo 3.6, conforme estabelece o seguinte resultado.

Teorema 3.6 *Sejam $x^* \in \mathbb{R}^n$ um zero de $F : \mathbb{R}^n \to \mathbb{R}^n$, $\eta_k \leq \eta < 1$ para todo $k \in \mathbb{N}$. Suponha que F seja três vezes continuamente diferenciável e que $J_F(x^*)$ seja não*

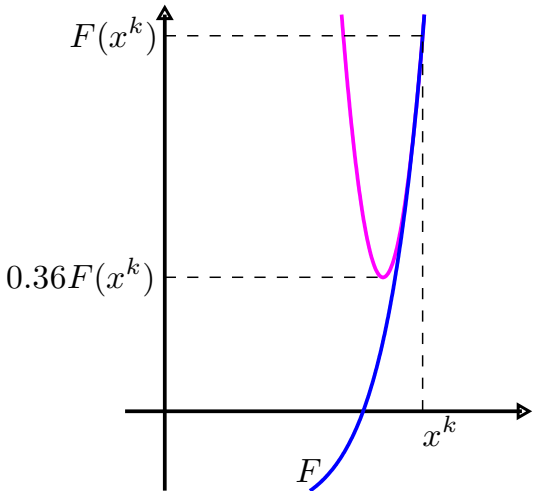

Figura 3.1: A condição do resíduo (3.54) não é verificada.

singular. Se $\|s^k\| = O\Big(\|F(x^k)\|\Big)$ *para todo* $k \in \mathbb{N}$, *então existe* $\varepsilon > 0$ *tal que para todo* $x^0 \in B(x^*, \varepsilon)$, *o Algoritmo 3.6 gera uma sequência* (x^k) *tal que* $x^k \in B(x^*, \varepsilon)$ *para todo* $k \in \mathbb{N}$ *e* $x^k \to x^*$ *linearmente no sentido que*

$$\left\|x^{k+1} - x^*\right\|_* \leq \rho \left\|x^k - x^*\right\|_*,$$

para algum $\rho \in (\eta, 1)$, *onde, por definição,* $\|y\|_* = \|J_F(x^*)y\|$. *Além disso, a taxa de convergência é*

1. *superlinear se* $\eta_k \to 0$.
2. *quadrática se* $\eta_k = O\Big(\|F(x^k)\|\Big)$.
3. *cúbica se* $\eta_k = O\Big(\|F(x^k)\|^2\Big)$.
4. $\min\{\hat{p}, 3\}$ *se* $\eta_k = O\Big(\|F(x^k)\|^{\hat{p}-1}\Big)$, *com* $\hat{p} > 1$.

Prova. [73, Teorema 1]. □

É necessário ter um algoritmo que obtém um zero aproximado s^k para o modelo quadrático (3.51) de maneira que a condição do resíduo (3.53) seja verificada para algum $\eta_k \in [0, 1)$ e que $\|s^k\| = O\Big(\|F(x^k)\|\Big)$. Neste sentido e motivados pelo método Super-Halley, Steihaug e Suleiman também propuseram em [73] uma classe de métodos chamada Classe Chebyshev-Halley Inexata. Esta abordagem inexata consiste em aplicar o método de Newton inexato nos dois sistemas (3.52), ou seja, dados $\eta_k^{(1)}, \eta_k^{(2)} \in [0, 1)$, obter $s_{(1)}^k$ e

$s^k_{(2)}$ tais que

$$\begin{aligned} \left\|\tilde{r}^k_1\right\| &\leq \eta^{(1)}_k \left\|M_k(0)\right\| \\ \left\|\tilde{r}^k_2\right\| &\leq \eta^{(2)}_k \left\|M_k(s^k_{(1)})\right\| \end{aligned} \tag{3.55}$$

onde

$$\tilde{r}^k_1 = J_{M_k}(0)s^k_{(1)} + M_k(0) \tag{3.56}$$

$$\tilde{r}^k_2 = J_{M_k}(s^k_{(1)})s^k_{(2)} + M_k(s^k_{(1)}). \tag{3.57}$$

Como

$$\begin{aligned} J_{M_k}(0) &= J_F(x^k), \\ M_k(0) &= F(x^k), \\ M_k(s^k_{(1)}) &= \tilde{r}^k_1 + \frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)}, \\ J_{M_k}(s^k_{(1)}) &= J_F(x^k) + \mathcal{T}_F(x^k)s^k_{(1)}, \end{aligned}$$

então (3.56) e (3.57) podem ser escritos, respectivamente, como

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) + \tilde{r}^k_1 \\ \Big(J_F(x^k) + \mathcal{T}_F(x^k)s^k_{(1)}\Big)s^k_{(2)} &= -\tilde{r}^k_1 - \frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)} + \tilde{r}^k_2 \end{aligned} \tag{3.58}$$

Os dois sistemas dados em (3.58) fornecem o passo $s^k = s^k_{(1)} + s^k_{(2)}$, o qual será chamado passo do método Super-Halley Inexato [38]. Desta forma, em [73] é proposto a Classe Chebyshev-Halley Inexata incorporando em (3.58) o parâmetro $\alpha \in \mathbb{R}$ da seguinte maneira

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) + \tilde{r}^k_1 \\ \Big(J_F(x^k) + \alpha\mathcal{T}_F(x^k)s^k_{(1)}\Big)s^k_{(2)} &= -\tilde{r}^k_1 - \frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)} + \tilde{r}^k_2 \\ x^{k+1} &= x^k + s^k_{(1)} + s^k_{(2)} \end{aligned} \tag{3.59}$$

Teorema 3.7 *Sejam $x^* \in \mathbb{R}^n$ um zero de $F : \mathbb{R}^n \to \mathbb{R}^n$, $s^k_{(1)}$ e $s^k_{(2)}$ soluções dos sistemas dados em* (3.59) *tais que as condições dos resíduos* (3.55) *sejam verificadas. Então os métodos da Classe Chebyshev-Halley Inexata são localmente convergentes. Além disso, supondo que*

$$\left\|\tilde{r}^k_1\right\| = O\Big(\left\|F(x^k)\right\|^{1+p}\Big) \quad e \quad \left\|\tilde{r}^k_2\right\| = O\Big(\left\|M_k(s^k_{(1)})\right\|^{1+q}\Big)$$

para algum $p, q \in (0, 1]$*, temos que a taxa de convergência é*

$$\min\{(1+p)(1+q), 3\}, \quad \textit{para } \alpha = 1$$

e

$$\min\{(1+p)(1+q), 2+p, 3\}, \quad \textit{para } \alpha \neq 1$$

Prova. Em [73, Teorema 3], a ideia da demonstração é mostrar que para k suficientemente grande existe $\eta_k \in (0, 1)$, tal que

$$\left\| \frac{1}{2}\mathcal{T}_F(x^k)s^k s^k + J_F(x^k)s^k + F(x^k) \right\| \leq \eta_k \left\| F(x^k) \right\|,$$

onde $s^k = s^k_{(1)} + s^k_{(2)}$ e que $\left\| s^k \right\| = O\Big(\left\| F(x^k) \right\| \Big)$. Daí a convergência segue do Teorema 3.6. O restante da prova pode ser vista em [73, Teorema 3]. □

Note no teorema anterior, que a convergência cúbica da Classe Chebyshev-Halley Inexata (3.59) para $\alpha \neq 1$ é atingida, quando $p = 1$ e $0.5 \leq q \leq 1$ e para $\alpha = 1$ é suficiente escolher $p = q = \sqrt{3} - 1$.

Os autores de [73] também propuseram uma modificação da Classe Chebyshev-Halley Inexata. Nesta classe, o passo de Newton é calculado de maneira exata e então o segundo sistema de (3.59) é resolvido por algum método iterativo, ou seja,

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) && (3.60) \\ \Big(J_F(x^k) + \alpha\mathcal{T}_F(x^k)s^k_{(1)}\Big)s^k_{(2)} &= -\frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)} + \tilde{r}^k_2 && (3.61) \\ x^{k+1} &= x^k + s^k_{(1)} + s^k_{(2)} \end{aligned}$$

Esta classe é denominada em [73] como Classe Chebyshev-Halley Inexata Modificada. A estratégia adotada em [73] para obter $s^k_{(1)}$ é resolver o sistema (3.60) via decomposição LU. Para resolver o sistema (3.61), os autores não utilizam a decomposição da matriz $J_F(x^k) + \alpha\mathcal{T}_F(x^k)s^k_{(1)}$. Eles reutilizam a decomposição LU da jacobiana $J_F(x^k)$ e executam algumas iterações do método do ponto fixo linear baseado em *splittings* [2] de

$$J_F(x^k) + \alpha\mathcal{T}_F(x^k)s^k_{(1)}.$$

Mais especificamente, fixado k e fazendo

$$B_k = J_F(x^k),$$

[2] Estratégias desta natureza são clássicas na literatura, como por exemplo os métodos de Jacobi e Gauss-Seidel.

$$C_k = -\alpha \mathcal{T}_F(x^k) s^k_{(1)}$$

e

$$b = -\frac{1}{2}\mathcal{T}_F(x^k) s^k_{(1)} s^k_{(1)},$$

o sistema (3.61) pode ser reescrito, como

$$Bw = Cw + b,$$

onde $w = s^k_{(2)}$. Dada uma estimativa inicial w^0, o processo iterativo

$$Bw^l = Cw^{l-1} + b$$

é construído para todo $l = 1, 2, 3, \ldots$. Desta forma, o Algoritmo 3.7 é proposto em [73].

Algoritmo 3.7: Cálculo de $s^k_{(2)}$ e $\tilde{r}^k_2$

Defina $A = J_F(x^k) + \alpha \mathcal{T}_F(x^k) s^k_{(1)}$, $b = -\frac{1}{2}\mathcal{T}_F(x^k) s^k_{(1)} s^k_{(1)}$.
Dados $w^0 = 0$ e $r^0 = b$.
PARA $l = 1, 2, \ldots$
  Defina z^{l-1} a solução do sistema $J_F(x^k) z^{l-1} = r^{l-1}$
  Atualize $w^l = w^{l-1} + z^{l-1}$
  Atualize $r^l = b - Aw^l$
FIM
$s^k_{(2)} = w^l$, $\tilde{r}^k_2 = r^l$ e $j = l$.

Note que apenas a decomposição de $J_F(x^k)$ é necessária no Algoritmo 3.7. Calculando $s^k_{(2)}$ pelo Algoritmo 3.7, os autores mostraram a convergência da Classe Chebyshev-Halley Inexata Modificada. Isto pode ser constatado no próximo teorema.

Teorema 3.8 *Sejam $x^* \in \mathbb{R}^n$ um zero de $F : \mathbb{R}^n \to \mathbb{R}^n$, $s^k_{(1)}$ e $s^k_{(2)}$ soluções de* (3.60) *e* (3.61)*, respectivamente, tais que a condição do resíduo* (3.55) *seja verificada. Suponha que o Algoritmo 3.7 termine em j iterações. Então os métodos da Classe Chebyshev-Halley Inexata Modificada são localmente convergentes e a taxa de convergência é* $\min\{3, j+2\}$ *para qualquer $\alpha \in \mathbb{R}$.*

Prova. Assim como na prova do Teorema 3.7, a ideia da demonstração é mostrar que para k suficientemente grande existe $\eta_k \in (0, 1)$, tal que

$$\left\| \frac{1}{2}\mathcal{T}_F(x^k) s^k s^k + J_F(x^k) s^k + F(x^k) \right\| \leq \eta_k \left\| F(x^k) \right\|,$$

onde $s^k = s^k_{(1)} + s^k_{(2)}$ e que $\left\|s^k\right\| = O\Big(\left\|F(x^k)\right\| \Big)$. Daí a convergência segue do Teorema 3.6. A taxa de convergência é obtida escolhendo $\eta_k = O\Big(\left\|F(x^k)\right\|^{\min\{2,j+1\}} \Big)$ no Teorema 3.6. Os detalhes da demonstração podem ser vistos em [73, Teorema 2]. □

Capítulo 4

Contribuições da Tese I - Teoria

4.1 Teorema de Raio de Convergência Cúbica da Classe Chebyshev-Halley

Quando começamos a estudar os métodos da Classe Chebyshev-Halley, a primeira preocupação foi com a prova de convergência, principalmente a taxa de convergência. Motivados pelo conhecimento do raio ótimo de convergência do método de Newton, pensamos em construir um raio de convergência dos métodos da Classe Chebyshev-Halley. Até então, desconhecíamos os trabalhos de Schwetlick [26, 70] e já tínhamos desenvolvido alguns resultados. Em um certo momento, a pedido, o Professor Schwetlick gentilmente enviou uma cópia de algumas páginas escaneadas de seu livro [70] de 1979. Depois de estudá-las e com os resultados que já havíamos desenvolvido, conseguimos enfim construir um raio de convergência para os métodos da Classe Chebyshev-Halley. Essa é uma das contribuições teóricas desta tese.

Em [70], é exibido um raio de convergência para a classe de métodos baseada nos modelos dados em (3.45) com $i = 1$. Mas como apontado na Seção 3.4.2, Gundersen e Steihaug [37] mostraram que a Classe Chebyshev-Halley (3.44), com $\alpha \in \mathbb{R}$, é equivalente a classe de métodos baseada nos modelos dados em (3.45) com $i = 1$. Isto significa que o raio dado em [70] é um raio de convergência para a Classe Chebyshev-Halley.

Considerando o raio dado em [70], é exigido na prova de convergência dada em [70] que o iterado de Newton

$$x^k - J_F(x^k)^{-1}F(x^k)$$

pertença à bola de centro em uma solução e tal raio. Salientamos que não é feita qualquer exigência sobre o iterado de Newton ao provar que o raio proposto nesta tese é de fato um raio de convergência para a Classe Chebyshev-Halley, podendo permitir um aumento no raio dado em [70]. Além disso, apresentamos uma comparação entre o raio de convergência

dado em [70] e o proposto nesta tese através de exemplos numéricos.

Para fixar as ideias, entendemos como raio de convergência o que segue na Definição 4.1.

Definição 4.1 *Considere x^* um zero de uma aplicação F e Ω um método iterativo para determinar x^*. Um número real $r > 0$ é dito ser um raio de convergência para Ω, quando tomado qualquer $x^0 \in B(x^*, r)$, todos os termos da sequência (x^k) gerada pelo método Ω pertencem à bola $B(x^*, r)$ e $x^k \to x^*$. O raio r^* será chamado ótimo quando for o maior possível, isto é, quando dado qualquer $r' > r^*$ existe um $x^0 \in B(x^*, r')$ tal que a sequência (x^k) gerada a partir de x^0 não converge para x^*.*

Com o objetivo de construir uma bola onde o método de Newton é bem definido, ou seja, que a jacobiana J_F avaliada em qualquer ponto desta bola seja não singular, vamos supor que J_F seja Lipschitz em uma vizinhança de x^*. Em seguida, vamos exibir o raio ótimo de convergência do método de Newton sob hipótese Lipschitz sobre a jacobiana. Vale salientar que o raio ótimo de convergência do método de Newton também pode ser atingido sob condições mais fracas que Lipschitz sobre a jacobiana, ver [32].

Lema 4.2 *Sejam $x^* \in \mathbb{R}^n$ uma solução do problema (3.1), J_F Lipschitz com constante L em uma bola $B(x^*, \tilde{\delta})$. Suponha que $J_F(x^*)$ seja não singular. Dado $t \in (0, 1)$, defina*

$$\bar{\delta} = \min\left\{\tilde{\delta}, \frac{t}{L\left\|J_F(x^*)^{-1}\right\|}\right\}.$$

Então, $J_F(x)$ é não singular e

$$\left\|J_F(x)^{-1}\right\| \leq \frac{\left\|J_F(x^*)^{-1}\right\|}{1-t} \tag{4.1}$$

para todo $x \in B(x^*, \bar{\delta})$.

Prova. Para facilitar a notação, faça $p = \left\|J_F(x^*)^{-1}\right\|$. Para todo $x \in B(x^*, \bar{\delta})$, temos que

$$\begin{aligned}\left\|I - J_F(x^*)^{-1}J_F(x)\right\| &= \left\|J_F(x^*)^{-1}(J_F(x^*) - J_F(x))\right\| \\ &\leq p\left\|J_F(x^*) - J_F(x)\right\| \\ &\leq pL\left\|x - x^*\right\| < pL\frac{t}{Lp} = t < 1\end{aligned}$$

Pelo Lema 1.2, $J_F(x)$ é não singular e

$$\left\|J_F(x)^{-1}\right\| \leq \frac{\|J_F(x^*)^{-1}\|}{1 - \|I - J_F(x^*)^{-1}J_F(x)\|} \leq \frac{\left\|J_F(x^*)^{-1}\right\|}{1-t}.$$

□

Teorema 4.3 *Sejam $x^* \in \mathbb{R}^n$ uma solução do problema (3.1), J_F Lipschitz em uma bola $B(x^*, \tilde{\delta})$, $L > 0$ a menor constante Lipschitz de J_F. Suponha que $J_F(x^*)$ seja não singular. Tome*

$$\delta = \min\left\{\tilde{\delta}, \frac{2}{3L\left\|J_F(x^*)^{-1}\right\|}\right\}. \tag{4.2}$$

Se $x^0 \in B(x^*, \delta)$ então o método de Newton gera uma sequência (x^k) tal que $x^k \in B(x^*, \delta)$ para todo $k \in \mathbb{N}$ e $x^k \to x^*$ com taxa de convergência quadrática. Além disso, δ é o maior raio de convergência possível.

Prova. Se $x^k \in B(x^*, \delta)$, então existe $t < \dfrac{2}{3}$ tal que

$$\left\|x^k - x^*\right\| < \frac{t}{L\left\|J_F(x^*)^{-1}\right\|}. \tag{4.3}$$

Pelo Lema 4.2, o passo de Newton está bem definido. Além disso, como $F(x^*) = 0$, temos

$$x^{k+1} - x^* = J_F(x^k)^{-1}\Big(F(x^*) - F(x^k) - J_F(x^k)(x^* - x^k)\Big).$$

Aplicando agora os Lemas 1.14 e 4.2 e usando (4.3), obtemos

$$\left\|x^{k+1} - x^*\right\| \leq \frac{\left\|J_F(x^*)^{-1}\right\|}{1-t}\frac{L}{2}\left\|x^k - x^*\right\|^2 \leq \frac{t}{2(1-t)}\left\|x^k - x^*\right\|. \tag{4.4}$$

Como $t < \dfrac{2}{3}$, temos $\dfrac{t}{2(1-t)} < 1$ e isto prova que a sequência (x^k) está bem definida, que $x^k \in B(x^*, \delta)$ para todo $k \in \mathbb{N}$ e que $x^k \to x^*$. A convergência quadrática decorre da primeira desigualdade na relação (4.4), completando a primeira parte da demonstração.

Para mostrar que δ é o maior raio de convergência possível, vamos considerar um caso particular de um exemplo dado em [32]. Considere $F : \mathbb{R} \to \mathbb{R}$ dada por

$$F(t) = \begin{cases} -t^2 - t & \text{se } t \leq 0 \\ t^2 - t & \text{se } t > 0. \end{cases} \tag{4.5}$$

Note que 0 é um zero de F e que $F'(t) = 2|t| - 1$ para todo $t \in \mathbb{R}$. Note que $|F'(0)| = 1$. Temos que

$$\Big|F'(u) - F'(v)\Big| \leq 2\Big||u| - |v|\Big| \leq 2|u - v|$$

para todo $u, v \in R$. Desta forma, F' é Lipschitz com constante 2 em todo $\mathbb{R}$ e desta

forma as hipóteses do teorema são satisfeitas. Afirmamos que

$$\delta = \frac{1}{3}$$

é o maior raio de convergência possível. Ora, como já provado, se $t_0 \in (-\delta, \delta)$, a sequência

$$t_{k+1} = t_k - \frac{F(t_k)}{F'(t_k)} \tag{4.6}$$

gerada pelo método de Newton está bem definida e converge para $t_* = 0$. Por outro lado, iniciando com

$$t_0 = -\frac{1}{3}$$

a sequência (t_k) dada em (4.6) não converge, pois

$$t_1 = \frac{1}{3} \quad \text{e} \quad t_2 = -\frac{1}{3}.$$

Desta maneira, o método de Newton produz a sequência alternada

$$\left(-\frac{1}{3}, \frac{1}{3}, -\frac{1}{3}, \dots\right).$$

Isso mostra que δ dado em (4.2) é o maior raio de convergência possível. □

A Figura 4.1 ilustra que o método de Newton falha na tentativa de encontrar um zero da função F definida em (4.5), tomando como ponto inicial $t_0 = -\frac{1}{3}$.

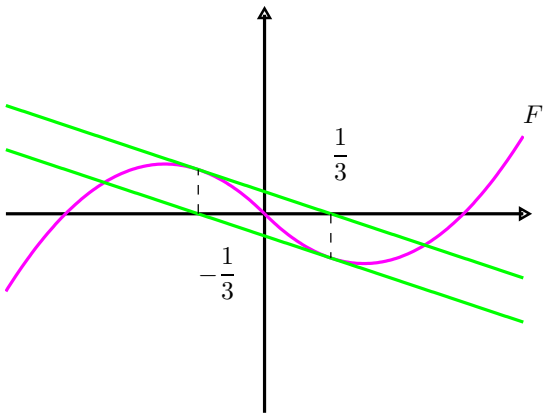

Figura 4.1: Exemplo de raio ótimo de convergência do método de Newton.

No entanto, vale ressaltar que podemos tomar pontos iniciais fora da bola de raio ótimo e também obter convergência. A grosso modo, podemos ter convergência tomando

pontos iniciais em regiões que não são descritas por bolas. Para esse propósito, vamos definir bacia ou região de convergência.

Definição 4.4 *Considere x^* um zero de uma aplicação F e Ω um método iterativo para determinar x^*. A bacia de convergência $\mathcal{R}(x^*)$ de um método Ω é o conjunto de pontos $x^0 \in \mathbb{R}^n$ tal que a sequência (x^k) gerada pelo método Ω converge para x^*, isto é,*

$$\mathcal{R}(x^*) = \{x^0 \in \mathbb{R}^n \mid x^k \to x^*\}.$$

A estrutura de uma bacia de convergência não tem nenhum padrão específico. Pode ser, por exemplo, um conjunto desconexo. Para ver isso, reformulamos um exemplo apresentado em [71][1] para o espaço $\mathbb{R}^2$.

Exemplo 4.5 *Considere $F : \mathbb{R}^2 \to \mathbb{R}^2$ dada por*

$$F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^3 - 3xy^2 - 1 \\ 3x^2y - y^3 \end{pmatrix},$$

cujos zeros são $x^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $x^{**} = \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix}$ e $x^{***} = \begin{pmatrix} -1/2 \\ -\sqrt{3}/2 \end{pmatrix}$. A Figura 4.2 ilustra

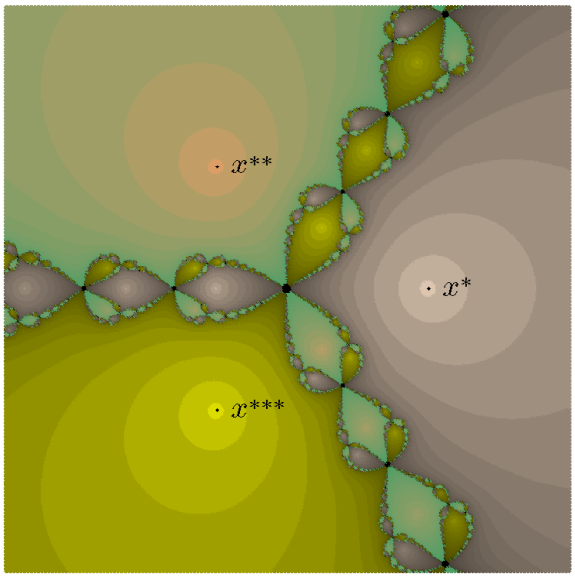

Figura 4.2: Bacia de convergência do método de Newton para o Exemplo 4.5

as 3 bacias de convergência nas cores cinza, laranja esverdeado e verde para o método de Newton. A região colorida de cinza é a bacia de convergência $\mathcal{R}(x^*)$, a região colorida de laranja esverdeado é a bacia de convergência $\mathcal{R}(x^{**})$ e a região colorida de verde é a bacia

[1] Em [71] é considerado $p : \mathbb{C} \to \mathbb{C}$ definida por $p(z) = z^3 - 1$.

de convergência $\mathcal{R}(x^{***})$. A fronteira entre as bacias de convergência, que está colorida de preto, é o conjunto de Julia, ou seja, é o conjunto de todos os pontos x^0 para os quais o método falha. As diferentes tonalidades indicam a quantidade de iterações realizadas para atingir a precisão requerida. As mais claras representam menos iterações e as mais escuras indicam uma quantidade maior de iterações.

Os próximos lemas são lemas puramente técnicos, exclusivamente para atingir nosso objetivo nesta seção, ou seja, exibir um raio de convergência para a Classe Chebyshev-Halley (3.44) para qualquer α real sob hipótese Lipschitz sobre $\mathcal{T}_F$.

Lema 4.6 *Considere as hipóteses e $\bar{\delta} > 0$ do Lema 4.2. Para todo $x \in B(x^*, \bar{\delta})$, temos a seguinte estimativa*

$$\|F(x)\| \leq \left(\frac{t}{\left\| J_F(x^*)^{-1} \right\|} + \|J_F(x^*)\| \right) \|x - x^*\| .$$

Prova. Note inicialmente que

$$\|J_F(x)\| \leq \|J_F(x) - J_F(x^*)\| + \|J_F(x^*)\| \leq L\bar{\delta} + \|J_F(x^*)\| \tag{4.7}$$

para todo $x \in B(x^*, \bar{\delta})$. Pela fórmula de Taylor com resto integral tem-se que

$$\|F(x)\| \leq \int_0^1 \|J_F(x^* + \tau(x - x^*))\| \, \|x - x^*\| \, d\tau$$

para todo $x \in B(x^*, \bar{\delta})$. Então por (4.7) e pela definição de $\bar{\delta}$ no Lema 4.2 tem-se

$$\|F(x)\| \leq (L\bar{\delta} + \|J_F(x^*)\|) \, \|x - x^*\| \leq \left(\frac{t}{\left\| J_F(x^*)^{-1} \right\|} + \|J_F(x^*)\| \right) \|x - x^*\| .$$

□

Lema 4.7 *Considere as hipóteses e $\bar{\delta} > 0$ do Lema 4.2. Dado $x \in B(x^*, \bar{\delta})$, defina*

$$y = x - J_F(x)^{-1} F(x). \tag{4.8}$$

Temos as seguintes estimativas:

$$\|y - x\| \leq \left(\frac{t + \left\| J_F(x^*)^{-1} \right\| \|J_F(x^*)\|}{1 - t} \right) \|x - x^*\| \tag{4.9}$$

e

$$\|y - x^*\| \leq \frac{\left\| J_F(x^*)^{-1} \right\| L}{2(1 - t)} \|x - x^*\|^2 . \tag{4.10}$$

Prova. Pelo Lema 4.2, a matriz $J_F(x)$ é não singular para todo $x \in B(x^*, \bar{\delta})$. Com isso, y dado em (4.8) está bem definido. Observe agora que a desigualdade (4.9) decorre diretamente da limitação de $J_F(x)^{-1}$ dado no Lema 4.2 e do Lema 4.6. Basta observar que

$$\|y - x\| \leq \left\|J_F(x)^{-1}\right\| \|F(x)\|.$$

Agora, observando que

$$y - x^* = J_F(x)^{-1}\Big[F(x^*) - F(x) - J_F(x)(x^* - x)\Big],$$

a desigualdade (4.10) decorre diretamente do Lema 1.14 e também da limitação de $J_F(x)^{-1}$ dado no Lema 4.2. □

Neste momento, vale a pena observar que como $x \in B(x^*, \bar{\delta})$, então

$$\|y - x^*\| \leq \frac{t}{2(1-t)} \|x - x^*\|.$$

Isso não garante que $y \in B(x^*, \bar{\delta})$, a menos que t seja menor que $\dfrac{2}{3}$, como observado no Teorema 4.3, mais especificamente, na desigualdade (4.4).

Lema 4.8 *Considere as hipóteses e $\bar{\delta} > 0$ do Lema 4.2. Dado $x \in B(x^*, \bar{\delta})$, considere y definido no Lema 4.7 e defina os vetores*

$$u = \mathcal{T}_F(x)(x^* - x)(x^* - x) - \mathcal{T}_F(x)(y - x)(y - x)$$

e

$$v = \mathcal{T}_F(x)(y - x)(y - x) + \mathcal{T}_F(x)(y - x)(x - x^*)$$

Nestas condições, temos que

$$\|u\| \leq \frac{\left\|J_F(x^*)^{-1}\right\| L^2}{2(1-t)^2}\Big(1 + \left\|J_F(x^*)^{-1}\right\| \|J_F(x^*)\|\Big) \|x - x^*\|^3 \tag{4.11}$$

e

$$\|v\| \leq \frac{\left\|J_F(x^*)^{-1}\right\| L^2}{2(1-t)^2}(t + \left\|J_F(x^*)^{-1}\right\| \|J_F(x^*)\|) \|x - x^*\|^3.$$

Prova. Para facilitar a notação, faça $p = \left\|J_F(x^*)^{-1}\right\|$ e $c = \|J_F(x^*)\|$. Pelo Teorema de Schwarz para aplicações [54], temos que as camadas horizontais do tensor $\mathcal{T}_F(x)$ são

matrizes simétricas e assim, pelo Lema 1.8 podemos escrever u como

$$\begin{aligned} u &= \mathcal{T}_F(x)(x^*-x)(x^*-x) - \mathcal{T}_F(x)(x^*-x)(y-x) + \\ &+ \mathcal{T}_F(x)(y-x)(x^*-x) - \mathcal{T}_F(x)(y-x)(y-x) \\ &= \mathcal{T}_F(x)\Big[(x^*-x)(x^*-x-y+x) + (y-x)(x^*-x-y+x)\Big] \\ &= \mathcal{T}_F(x)\Big[(x^*-x)(x^*-y) + (y-x)(x^*-y)\Big]. \end{aligned}$$

Como J_F é Lipschitz com constante L, podemos utilizar (1.21) para concluir que

$$\|\mathcal{T}_F(x)\| \leq L. \tag{4.12}$$

Logo, por (4.12) e pelo Lema 4.7 temos que

$$\begin{aligned} \|u\| &\leq \|\mathcal{T}_F(x)\| \Big[\|x^*-x\| \|x^*-y\| + \|y-x\| \|x^*-y\| \Big] \\ &\leq L\Big[\frac{pL}{2(1-t)} \|x-x^*\|^3 + \Big(\frac{t+pc}{1-t}\Big)\frac{pL}{2(1-t)} \|x-x^*\|^3 \Big] \\ &= \frac{pL^2}{2(1-t)}\Big[1+\frac{t+pc}{1-t}\Big] \|x-x^*\|^3 \\ &= \frac{pL^2}{2(1-t)}\Big[\frac{1+pc}{1-t}\Big] \|x-x^*\|^3 . \end{aligned}$$

Daí segue (4.11). Para mostrar o que falta, escrevemos v como

$$v = \mathcal{T}_F(x)(y-x)\Big[(y-x)+(x-x^*)\Big] = \mathcal{T}_F(x)(y-x)(y-x^*).$$

Usando novamente (4.12) e o Lema 4.7 temos

$$\begin{aligned} \|v\| &\leq \|\mathcal{T}_F(x)\| \|y-x\| \|y-x^*\| \\ &\leq L\Big(\frac{t+pc}{1-t}\Big) \|x-x^*\| \frac{pL}{2(1-t)} \|x-x^*\|^2 \\ &= \frac{pL^2}{2(1-t)^2}(t+pc) \|x-x^*\|^3 , \end{aligned}$$

completando a demonstração. □

O próximo lema define uma bola onde a Classe Chebyshev-Halley é bem definida, ou seja, qualquer que seja o ponto x desta bola, as matrizes $J_F(x)$ e $I - \alpha\mathcal{L}(x)$ são não singulares.

Lema 4.9 *Considere as hipóteses e $\bar{\delta} > 0$ do Lema 4.2 e $\alpha \neq 0$. Defina*

$$\hat{\delta} = \min \left\{ \bar{\delta}, \frac{t(1-t)^2}{|\alpha| L \left\| J_F(x^*)^{-1} \right\| \left(t + \left\| J_F(x^*)^{-1} \right\| \|J_F(x^*)\| \right)} \right\}.$$

Então, a matriz $I - \alpha \mathcal{L}(x)$ é não singular e

$$\left\| \left(I - \alpha \mathcal{L}(x) \right)^{-1} \right\| \leq \frac{1}{1-t} \tag{4.13}$$

para todo $x \in B(x^*, \hat{\delta})$.

Prova. Como J_F é Lipschitz com constante $L > 0$, temos de (1.21) que

$$\|\mathcal{T}_F(x)\| \leq L \tag{4.14}$$

para todo $x \in B(x^*, \hat{\delta})$. Agora, para facilitar a notação, faça $p = \left\| J_F(x^*)^{-1} \right\|$ e $c = \|J_F(x^*)\|$. Assim, usando os Lemas 4.2 e 4.6 e (4.14), temos que

$$\begin{aligned} \|\alpha \mathcal{L}(x)\| &\leq |\alpha| \left\| J_F(x)^{-1} \right\|^2 \|\mathcal{T}_F(x)\| \|F(x)\| \\ &\leq |\alpha| \frac{p^2}{(1-t)^2} L \left(\frac{t}{p} + c \right) \|x - x^*\| \\ &\leq \frac{|\alpha| p L (t + pc)}{(1-t)^2} \|x - x^*\| \\ &\leq t < 1 \end{aligned}$$

para todo $x \in B(x^*, \hat{\delta})$. Pelo Lema 1.1, $I - \alpha \mathcal{L}(x)$ é não singular e vale (4.13) para todo $x \in B(x^*, \hat{\delta})$. □

Agora que conhecemos uma bola onde a Classe Chebyshev-Halley está bem definida, temos condições de estabelecer um raio de convergência para esta classe. Um raio de convergência possível é dado pelo Teorema 4.10 para todos os métodos da Classe Chebyshev-Halley, com exceção do método de Chebyshev ($\alpha = 0$). Para o método de Chebyshev, um raio de convergência será apresentado adiante.

Teorema 4.10 *Sejam $x^* \in \mathbb{R}^n$ uma solução do problema* (3.1)*, com $J_F(x^*)$ não singular, $\mathcal{T}_F$ Lipschitz em uma bola $B(x^*, \delta_1)$, $L_2 > 0$ a menor constante Lipschitz de $\mathcal{T}_F$. Denote $p = \left\| J_F(x^*)^{-1} \right\|$ e $c = \|J_F(x^*)\|$. Considere*

$$L = \sup_{x \in B(x^*, \delta_1)} \{ \|\mathcal{T}_F(x)\| \}. \tag{4.15}$$

Dados $t \in (0,1)$ *e* $\alpha \neq 0$, *defina*

$$\hat{\delta} = \min\left\{\frac{t}{Lp}, \frac{t(1-t)^2}{|\alpha|Lp(t+pc)}\right\}$$

e

$$\delta^* = \min\left\{\delta_1, \hat{\delta}, \frac{t(1-t)^2}{p}, \frac{12(1-t)^2}{2L_2(1-t)^2 + 3pL^2\Big((1+pc)+2|\alpha|(t+pc)\Big)}\right\}.$$

Se $x^0 \in B(x^*, \delta^*)$, *então a Classe Chebyshev-Halley* (3.44) *gera uma sequência* (x^k) *tal que* $x^k \in B(x^*, \delta^*)$ *para todo* $k \in \mathbb{N}$ *e* $x^k \to x^*$ *com taxa de convergência cúbica.*

Prova. Observe inicialmente que

$$\|\mathcal{T}_F(x)\| \leq L \tag{4.16}$$

para todo $x \in B(x^*, \delta_1)$. Segue diretamente de (1.21) que J_F é Lipschitz na bola $B(x^*, \delta_1)$, sendo L a menor constante de Lipschitz de J_F. Pelo Lema 4.2, $J_F(x)$ é não singular para todo $x \in B(x^*, \delta^*)$. Dado $x^k \in B(x^*, \delta^*)$, considere

$$\begin{aligned} y &= x^k - J_F(x^k)^{-1}F(x^k) \\ u &= \mathcal{T}_F(x^k)(x^* - x^k)(x^* - x^k) - \mathcal{T}_F(x^k)(y - x^k)(y - x^k) \\ v &= \mathcal{T}_F(x^k)(y - x^k)(y - x^k) + \mathcal{T}_F(x^k)(y - x^k)(x^k - x^*). \end{aligned} \tag{4.17}$$

Temos que

$$\mathcal{L}(x^k) = -J_F(x^k)^{-1}\mathcal{T}_F(x^k)(y - x^k).$$

Para facilitar a notação, faça

$$A_k = I - \alpha\mathcal{L}(x^k).$$

Pelo Lema 4.9, $I - \alpha\mathcal{L}(x)$ é não singular para todo $x \in B(x^*, \delta^*)$. Assim, (3.44) pode ser escrita como

$$x^{k+1} = x^k + (y - x^k) + \frac{1}{2}\mathcal{L}(x^k)A_k^{-1}(y - x^k) \tag{4.18}$$

Pelo Lema 1.3,

$$\mathcal{L}(x^k)A_k^{-1} = A_k^{-1}\mathcal{L}(x^k).$$

Desta forma, de (4.18), temos

$$\begin{aligned}
x^{k+1} - x^* &= x^k - x^* + (y - x^k) + \frac{1}{2}A_k^{-1}\mathcal{L}(x^k)(y - x^k) \\
&= A_k^{-1}\Big[A_k(x^k - x^*) + A_k(y - x^k) + \frac{1}{2}\mathcal{L}(x^k)(y - x^k)\Big] \\
&= A_k^{-1}\Big[A_k(x^k - x^*) + A_k(y - x^k) - \frac{1}{2}J_F(x^k)^{-1}\mathcal{T}_F(x^k)(y - x^k)(y - x^k)\Big] \\
&= A_k^{-1}J_F(x^k)^{-1}\Big[J_F(x^k)A_k(x^k - x^*) + J_F(x^k)A_k(y - x^k) - \frac{1}{2}\mathcal{T}_F(x^k)(y - x^k)(y - x^k)\Big].
\end{aligned} \tag{4.19}$$

Como

$$J_F(x^k)A_k = J_F(x^k) + \alpha\mathcal{T}_F(x^k)(y - x^k),$$

e pela definição de y, u e v dados em (4.17), escrevemos a expressão dentro do colchetes de (4.19), como

$$\begin{aligned}
& J_F(x^k)(x^k - x^*) + \alpha\mathcal{T}_F(x^k)(y - x^k)(x^k - x^*) + J_F(x^k)(y - x^k) + \\
+ \quad & \alpha\mathcal{T}_F(x^k)(y - x^k)(y - x^k) - \frac{1}{2}\mathcal{T}_F(x^k)(y - x^k)(y - x^k) \\
= \quad & J_F(x^k)(x^k - x^*) + \alpha\Big[\mathcal{T}_F(x^k)(y - x^k)(x^k - x^*) + \mathcal{T}_F(x^k)(y - x^k)(y - x^k)\Big] + \\
+ \quad & J_F(x^k)(-J_F(x^k)^{-1}F(x^k)) - \frac{1}{2}\mathcal{T}_F(x^k)(y - x^k)(y - x^k) \\
= \quad & J_F(x^k)(x^k - x^*) + \alpha v - F(x^k) - \frac{1}{2}\Big[\mathcal{T}_F(x^k)(x^* - x^k)(x^* - x^k) - u\Big].
\end{aligned}$$

Daí, a expressão dentro do colchetes de (4.19) fica

$$F(x^*) - F(x^k) - J_F(x^k)(x^* - x^k) - \frac{1}{2}\mathcal{T}_F(x^k)(x^* - x^k)(x^* - x^k) + \frac{1}{2}u + \alpha v. \tag{4.20}$$

Aplicando os Lemas 1.15 e 4.8 temos que

$$\begin{aligned}
& \left\|F(x^*) - F(x^k) - J_F(x^k)(x^* - x^k) - \frac{1}{2}\mathcal{T}_F(x^k)(x^* - x^k)(x^* - x^k) + \frac{1}{2}u + \alpha v\right\| \\
\leq \quad & \Big[\frac{L_2}{6} + \frac{pL^2}{4(1-t)^2}(1 + pc) + |\alpha|\frac{pL^2}{2(1-t)^2}(t + pc)\Big]\left\|x^k - x^*\right\|^3 \\
\leq \quad & \Big[\frac{L_2}{6} + \frac{pL^2}{4(1-t)^2}\Big((1 + pc) + 2|\alpha|(t + pc)\Big)\Big]\left\|x^k - x^*\right\|^3.
\end{aligned}$$

Com isso, de (4.19) e usando os Lemas 4.2 e 4.9, temos

$$\begin{aligned}\left\|x^{k+1}-x^*\right\| &\leq \frac{p}{(1-t)^2}\left[\frac{L_2}{6}+\frac{pL^2}{4(1-t)^2}\Big((1+pc)+2|\alpha|(t+pc)\Big)\right]\left\|x^k-x^*\right\|^3 \\ &\leq \frac{p}{(1-t)^2}\left[\frac{2L_2(1-t^2)+3pL^2\Big((1+pc)+2|\alpha|(t+pc)\Big)}{12(1-t)^2}\right]\left\|x^k-x^*\right\|^3.\end{aligned} \tag{4.21}$$

Pela definição de δ^* e usando (4.21) temos que

$$\left\|x^{k+1}-x^*\right\| \leq t\left\|x^k-x^*\right\|$$

e isto prova que a sequência (x^k) está bem definida, que $x^k \in B(x^*,\delta^*)$ para todo $k \in \mathbb{N}$ e que $x^k \to x^*$. A convergência cúbica decorre de (4.21), completando a demonstração. □

Para o método de Chebyshev (3.41), ou seja, quando $\alpha = 0$, a única matriz que deve ser não singular é $J_F(x^k)$. De forma inteiramente análoga como demonstrado no Teorema 4.10, um raio de convergência para esse método é

$$\delta^* = \min\left\{\delta_1, \frac{t}{pL}, \frac{t(1-t)^2}{p}, \frac{12(1-t)^2}{2L_2(1-t)^2+3pL^2(1+pc)}\right\}.$$

Considerando o Exemplo 4.5, as Figuras 4.3, 4.4 e 4.5, ilustram as 3 bacias de convergência quando a sequência (x^k) é gerada pelo método Chebyshev ($\alpha = 0$), Halley ($\alpha = 1/2$) e Super-Halley ($\alpha = 1$), respectivamente.

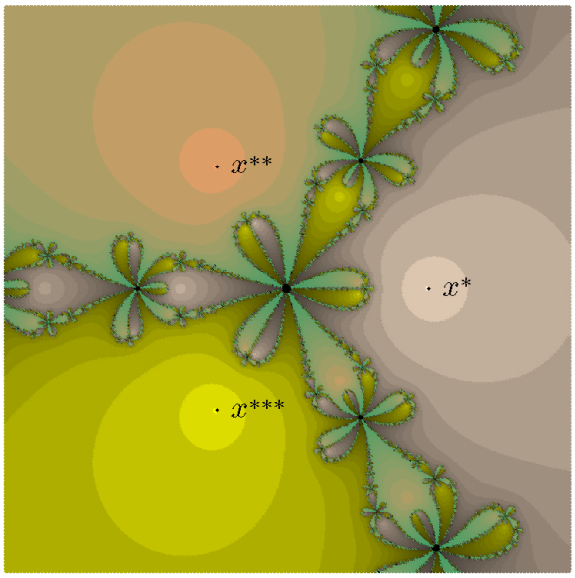

Figura 4.3: Bacia de convergência do método de Chebyshev para o Exemplo 4.5

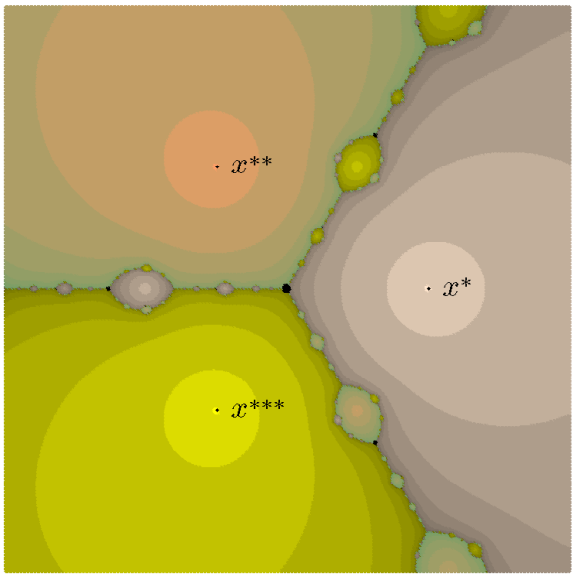

Figura 4.4: Bacia de convergência do método de Halley para o Exemplo 4.5

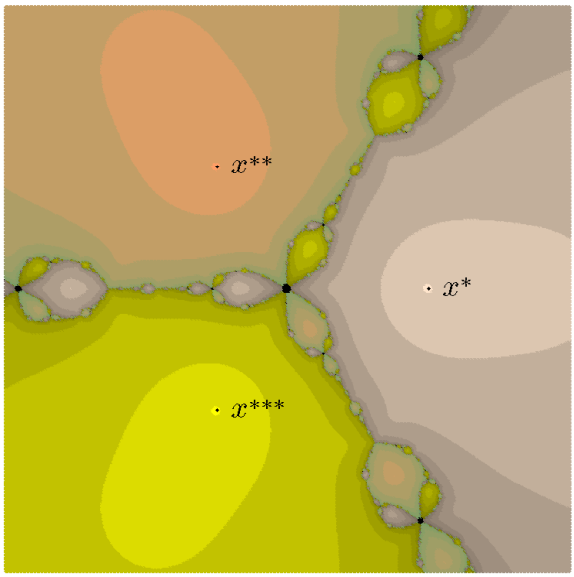

Figura 4.5: Bacia de convergência do método Super-Halley para o Exemplo 4.5

Como mencionado no início desta seção, em [70] também é exibido um raio de convergência para a Classe Chebyshev-Halley. Este raio é exibido na demonstração do Teorema 5.7.5 do Capítulo 5 em [70]. Este teorema é enunciado abaixo.

Teorema 4.11 *Sejam $x^* \in \mathbb{R}^n$ uma solução do problema* (3.1)*, com $J_F(x^*)$ não singular, $\mathcal{T}_F$ Lipschitz em uma bola $B(x^*, \delta_1)$ com constante L_2. Para cada $\gamma \in \mathbb{R}$, existe $\delta_0 > 0$ tal que qualquer que seja $x^0 \in B(x^*, \delta_0)$, x^k e y^k gerados pelo Algoritmo 3.5 pertencem a bola $B(x^*, \delta_0)$ para todo $k \in \mathbb{N}$. Além disso, $x^k \to x^*$ com taxa de convergência cúbica.*

Prova. [70, Teorema 5.7.5] □

O raio de convergência δ_0 do Teorema 4.11 é exibido na demonstração do teorema em [70], a saber,

$$\delta_0 = \min\Big\{\delta, \frac{1}{C_1}, \frac{2\delta}{|\alpha|L_1C_0}, \sqrt{\frac{t}{C_3}}\Big\},$$

onde

$$\begin{aligned}
&t \in (0,1),\\
&L_1 = \|\mathcal{T}_F(x^*)\| + L_2\delta_1,\\
&M > \left\|J_F(x^*)^{-1}\right\|,\\
&\delta = \min\Big\{\delta_1, \frac{M - \left\|J_F(x^*)^{-1}\right\|}{(1+L_1)\left\|J_F(x^*)^{-1}\right\| M}, \frac{2t}{(2+L_1)M}\Big\},\\
&C_0 = M(\|J_F(x^*)\| + L_1\delta_1),\\
&C_1 = \frac{ML_1}{2},\\
&C_2 = \frac{L_1C_1}{2}(1 + C_0 + |\alpha|C_0),\\
&C_3 = M\Big(\frac{L_2}{6} + C_2\Big).
\end{aligned}$$

Como já apontamos no início desta seção, a prova do Teorema 4.11 dada em [70] nos ajudou a estabelecer o Teorema 4.10. Destacamos duas diferenças nas demonstrações dos teoremas. A primeira é que não exigimos que o iterado de Newton

$$x^k - J_F(x^k)^{-1}F(x^k)$$

esteja na bola $B(x^*, \delta^*)$, enquanto que no Teorema 4.11 o iterado de Newton deve pertencer à bola $B(x^*, \delta_0)$. É importante notar que o fato de não exigirmos que o iterado de Newton esteja na bola $B(x^*, \delta^*)$ indica que o raio δ_0 proposto por Schwetlick [70] pode ser aumentado. A segunda diferença é na definição de L e L_1. Ambas as constantes servem para limitar o tensor na bola $B(x^*, \delta_1)$, ou seja,

$$\|\mathcal{T}_F(x)\| \leq L_1 \quad \text{e} \quad \|\mathcal{T}_F(x)\| \leq L$$

para todo $x \in B(x^*, \delta_1)$. Evidentemente, $L \leq L_1$.

Com o propósito de comparar o raio δ_0 do Teorema 4.11 e o raio δ^* do Teorema 4.10 proposto nesta tese, apresentaremos a seguir 4 exemplos onde estimamos que

$$\delta^* > \delta_0$$

quando se usam os métodos Halley e Super-Halley e, quando se aplica o método de Chebyshev, estimamos que

$$\delta^* < \delta_0.$$

Exemplo 4.12 *Considere* $F : (2,4) \to \mathbb{R}$ *dada por* $F(t) = \dfrac{t^3}{3} - 9$. *Note que* 3 *é o zero de* F *e que* $F'(t) = t^2$ *e* $F''(t) = 2t$. *Vemos imediatamente que* F'' *é Lipschitz no domínio de* F, *sendo* $L_2 = 2$ *a menor constante Lipschitz de* F''. *Note que* $\delta_1 = 1$ *e* $L = 8$ *por definição.*

Exemplo 4.13 *Considere* $F : (0,2) \to \mathbb{R}$ *dada por* $F(t) = \dfrac{t^4}{4} - \dfrac{1}{4}$. *Note que* 1 *é o zero de* F *e que* $F'(t) = t^3$ *e* $F''(t) = 3t^2$. *Como* $F'''(t) = 6t$, *vemos imediatamente que* F'' *é Lipschitz no domínio de* F, *sendo* $L_2 = 12$ *a menor constante Lipschitz de* F''. *Note que* $\delta_1 = 1$ *e* $L = 12$ *por definição.*

Exemplo 4.14 *Considere* $F : \left(\dfrac{-\pi}{2}, \dfrac{\pi}{2}\right) \to \mathbb{R}$ *dada por* $F(t) = -sen(t)$. *Note que* 0 *é o zero de* F *e que* $F'(t) = -cos(t)$ *e* $F''(t) = sen(t)$. *Como* $F'''(t) = cos(t)$, *vemos imediatamente que* F'' *é Lipschitz no domínio de* F, *sendo* $L_2 = 1$ *a menor constante Lipschitz de* F''. *Note que* $\delta_1 = \dfrac{\pi}{2}$ *e* $L = 1$ *por definição.*

Exemplo 4.15 *Considere* $F : (0,2) \to \mathbb{R}$ *dada por* $F(t) = e^t - e$. *Note que* 1 *é o zero de* F *e que* $F'(t) = F''(t) = F'''(t) = e^t$. *Vemos imediatamente que* F'' *é Lipschitz no domínio de* F, *sendo* $L_2 = e^2$ *a menor constante Lipschitz de* F''. *Note que* $\delta_1 = 1$ *e* $L = e^2$ *por definição.*

Por estes exemplos, não podemos afirmar qual raio é maior quando se permite variar o parâmetro que determina o método, mas eles apresentam alguns indícios para pesquisa futura, como por exemplo, propor o raio ótimo de convergência para os métodos da Classe Chebyshev-Halley ou pelo menos para os métodos clássicos desta classe.

Para fazer uma comparação de δ^* e δ_0, fizemos o seguinte procedimento:

1. fixamos um valor para α (parâmetro que determina o método);
2. calculamos δ^* com t variando de 0.01 até 0.99 com incremento 0.01;
3. calculamos δ_0 com t variando de 0.01 até 0.99 com incremento 0.01 e M variando de $1.01 \left\|J_F(x^*)^{-1}\right\|$ até $20 \left\|J_F(x^*)^{-1}\right\|$ com incremento 0.01.

Os valores estão listados na Tabela 4.1.

Método	Exemplo	δ^*	δ_0
Chebyshev	4.12	0.181998843445359	0.502487562189055
	4.13	0.010296026052251	0.018793706293706
	4.14	0.148137	0.205227938437902
	4.15	0.062539504999145	0.102444476255497
Halley	4.12	0.133406835722161	0.066176470588235
	4.13	0.007335056450084	0.000260416666667
	4.14	0.148137	0.071418112747264
	4.15	0.045462020433983	0.005131730404308
Super-Halley	4.12	0.096720606212916	0.033088235294118
	4.13	0.005443148967833	0.000130208333334
	4.14	0.1134	0.035709056373632
	4.15	0.033014395313883	0.002565865202154

Tabela 4.1: Comparação do raio de convergência proposto nesta tese e outro conhecido na literatura.

4.2 Classe Chebyshev-Halley Livre de Tensores: Uma Abordagem Inexata

Vimos na Seção 3.4.3 que a necessidade de se obter o tensor $\mathcal{T}_F(x^k)$ a cada iteração e resolver de forma exata dois sistemas lineares, inviabiliza o uso dos métodos da Classe Chebyshev-Halley. De certa forma, a Classe Chebyshev-Halley Inexata (3.59) proposta em [73], como apresentada também na Seção 3.4.3, reduz o custo computacional da classe Chebyshev-Halley. Recorde que ela foi introduzida com o objetivo de encontrar um zero aproximado para o modelo quadrático de F em torno de x^k. No entanto, essa redução não é muito significativa, pois é necessário o uso do tensor $\mathcal{T}_F(x^k)$ a cada iteração e, além disso, do ponto de vista prático, não é possível controlar a precisão do passo, pois o termo forçante não pode ser dado a priori.

Em particular, outras versões do método de Chebyshev inexato, para problemas de otimização sem restrições, têm sido propostas da forma

$$\begin{aligned} \nabla^2 f(x^k)s^k_{(1)} &= -\nabla f(x^k) + r^k_{(1)} \\ \nabla^2 f(x^k)s^k_{(2)} &= -\frac{1}{2}\nabla^3 f(x^k)s^k_{(1)}s^k_{(1)} + r^k_{(2)} \\ x^{k+1} &= x^k + s^k_{(1)} + s^k_{(2)} \end{aligned} \tag{4.22}$$

onde f é a função objetivo que é minimizada, $\nabla f, \nabla^2 f$ e $\nabla^3 f$ são, repectivamente, os operadores gradiente, Hessiana e tensor de f. Esses dois sistemas podem ser resolvidos via método gradiente conjugado pré-condicionado, veja [20, 79]. Ao contrário da Classe Chebyshev-Halley Inexata proposta em [73], aqui o resíduo $r^k_{(2)}$ do segundo sistema linear de (4.22) não depende do resíduo $r^k_{(1)}$ do primeiro sistema linear.

Neste sentido, propomos nesta tese, uma maneira mais eficiente de tornar os métodos da Classe Chebyshev-Halley mais econômicos computacionalmente. Ao invés de encontrar um zero aproximado do modelo quadrático de F em torno de x^k usando a Classe Chebyshev-Halley Inexata (3.59), como feita em [73], vamos definir uma nova classe de métodos baseado em ideias *matrix-free* para o método de Newton inexato. Esta classe será chamada Classe Chebyshev-Halley Inexata livre de tensores.

Considere então uma aplicação contínua $C : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ tal que

$$\|C(x)\| = O\Big(\|F(x)\| \Big). \tag{4.23}$$

Uma matriz trivial que cumpre (4.23) é a matriz nula. De qualquer maneira, iremos mostrar adiante exemplos de matrizes que cumprem a condição (4.23) para x suficientemente próximo de um zero de F.

A Classe Chebyshev-Halley Inexata livre de tensores consiste em resolver de forma inexata os dois sistemas lineares

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) \\ \Big(J_F(x^k) + \alpha C(x^k)\Big)s^k_{(2)} &= -\frac{1}{2}C(x^k)s^k_{(1)}, \end{aligned} \tag{4.24}$$

baseado na ideia do método de Newton inexato, ou seja, dado $x^k \in \mathbb{R}^n$, obtemos $\eta^{(1)}_k \in [0,1)$, $\eta^{(2)}_k \in [0,1)$, $s^k_{(1)}$ e $s^k_{(2)}$ tais que

$$\left\|r^k_{(1)}\right\| \leq \eta^{(1)}_k \left\|F(x^k)\right\| \tag{4.25}$$

$$\left\|r^k_{(2)}\right\| \leq \eta^{(2)}_k \left\|\frac{1}{2}C(x^k)s^k_{(1)}\right\| \tag{4.26}$$

onde

$$r^k_{(1)} = J_F(x^k)s^k_{(1)} + F(x^k) \tag{4.27}$$

$$r^k_{(2)} = \Big(J_F(x^k) + \alpha C(x^k)\Big)s^k_{(2)} + \frac{1}{2}C(x^k)s^k_{(1)} \tag{4.28}$$

e tomar o próximo iterado como sendo $x^{k+1} = x^k + s^k_{(1)} + s^k_{(2)}$.

Observe que, a cada iteração, a matriz $\mathcal{T}_F(x^k)s^k_{(1)}$ na Classe Chebyshev-Halley (3.46) é substituída por uma matriz $C(x^k)$ que cumpre a condição (4.23). Além disso, podemos controlar os termos forçantes $\eta^{(1)}_k$ e $\eta^{(2)}_k$ a cada iteração, de modo a acelerar a convergência, ao contrário da Classe Chebyshev-Halley Inexata proposta em [73]. Note também que a exigência (4.26) do resídulo $r^k_{(2)}$ é diferente da exigência (3.55) requerida do resídulo $\tilde{r}^k_2$ na Classe Chebyshev-Halley Inexata proposta por Steihaug e Suleiman [73].

Para $\alpha = 0$ temos o método Chebyshev inexato livre de tensor, para $\alpha = \dfrac{1}{2}$ temos o

método Halley inexato livre de tensor e para $\alpha = 1$ temos o método Super-Halley inexato livre de tensor. Note que $s^k_{(1)}$ é um passo do método de Newton inexato.

Daí segue o algoritmo proposto nesta tese.

Algoritmo 4.1: Classe Chebyshev-Halley Inexata Livre de Tensores

---

Dados: $x^0 \in \mathbb{R}^n$ e $C : \mathbb{R}^n \to \mathbb{R}^{n\times n}$
$k = 0$
REPITA enquanto $F(x^k) \neq 0$
  Escolha $\eta_k^{(1)} \in [0, 1)$
  Obtenha $s^k_{(1)}$ tal que $\left\| J_F(x^k)s^k_{(1)} + F(x^k) \right\| \leq \eta_k^{(1)} \left\| F(x^k) \right\|$
  Escolha $\eta_k^{(2)} \in [0, 1)$
  Obtenha $s^k_{(2)}$ tal que $\left\| \Big(J_F(x^k) + \alpha C(x^k) \Big) s^k_{(2)} + \frac{1}{2} C(x^k)s^k_{(1)} \right\| \leq \eta_k^{(2)} \left\| \frac{1}{2} C(x^k)s^k_{(1)} \right\|$
  Faça $x^{k+1} = x^k + s^k_{(1)} + s^k_{(2)}$
  $k = k + 1$
FIM

Observe que não impomos qualquer maneira de obter $s^k_{(1)}$ e $s^k_{(2)}$, deixando livre para o uso de qualquer procedimento para resolver um sistema linear.

4.2.1 Análise de Convergência

Vimos que o termo forçante no método de Newton inexato e no Algoritmo 3.6 de Steihaug e Suleiman, além de controlar a precisão do passo, tem uma grande influência na taxa de convergência. A convergência quadrática no método de Newton inexato pode ser atingida fazendo $\eta_k = O\Big(\left\| F(x^k) \right\| \Big)$ na condição do resíduo

$$\left\| J_F(x^k)s^k + F(x^k) \right\| \leq \eta_k \left\| F(x^k) \right\|,$$

e no Algoritmo 3.6 de Steihaug e Suleiman, a taxa de convergência quadrática também é atingida ao fazer $\eta_k = O\Big(\left\| F(x^k) \right\| \Big)$ na condição do resíduo

$$\left\| \frac{1}{2}\mathcal{T}_F(x^k)s^k s^k + J_F(x^k)s^k + F(x^k) \right\| \leq \eta_k \left\| F(x^k) \right\|.$$

Além disso, a taxa de convergência cúbica pode ser atingida no Algoritmo 3.6 de Steihaug e Suleiman.

O objetivo desta seção é provar a convergência da sequência (x^k) gerada pela Classe Chebyshev-Halley Inexata livre de tensores. Veremos que os termos forçantes influenciam na taxa de convergência. Para isso apresentaremos alguns resultados preliminares. O Lema 4.16 é bem conhecido na literatura, veja [62].

Lema 4.16 *Sejam $U \subset \mathbb{R}^n$ aberto e convexo, $\bar{x} \in U$ e $A : U \to \mathbb{R}^{n\times n}$ contínua em $\bar{x}$. Se $A(\bar{x})$ é não singular, então existe um $\tilde{\varepsilon} > 0$ e um $\gamma > 0$ tais que $A(x)$ seja não singular e $\left\|A(x)^{-1}\right\| \leq \gamma$ para todo $x \in B(\bar{x}, \tilde{\varepsilon})$.*

Prova. Considere $t \in (0, 1)$ e $c = \|A(\bar{x})\|$. Por continuidade, existe $\tilde{\varepsilon} > 0$ tal que

$$\|A(x) - A(\bar{x})\| \leq \frac{t}{\left\|A(\bar{x})^{-1}\right\|}$$

para todo $x \in B(\bar{x}, \tilde{\varepsilon})$. Daí,

$$\|I - A(\bar{x})A(x)\| = \left\|A(\bar{x})^{-1}\Big(A(\bar{x}) - A(x)\Big)\right\| \leq t < 1.$$

Logo, tomando $\gamma = \dfrac{c}{1-t}$, pelo Lema 1.2, $A(x)$ é não singular e $\left\|A(x)^{-1}\right\| \leq \gamma$ para todo $x \in B(\bar{x}, \tilde{\varepsilon})$. □

O resultado do Lema 4.17 merece um destaque especial. Ele é uma ferramenta muito importante em nossa análise de convergência, principalmente na taxa de convergência. Tal resultado também foi utilizado nos trabalhos [19, 38, 73].

Lema 4.17 *Seja $x^* \in \mathbb{R}^n$ uma raiz de $F : U \to \mathbb{R}^n$ diferenciável em U aberto e convexo. Suponha que J_F seja Lipschitz com constante L e que $J_F(x^*)$ seja não singular. Para qualquer $\delta \in (0, 1)$, existe um $\hat{\varepsilon} > 0$ tal que*

$$(1-\delta)\left\|J_F(x^*)(x-x^*)\right\| \leq \|F(x)\| \leq (1+\delta)\left\|J_F(x^*)(x-x^*)\right\| \tag{4.29}$$

para todo $x \in B(x^*, \hat{\varepsilon})$.

Prova. Inicialmente, note que

$$F(x) = F(x) - F(x^*) - J_F(x^*)(x-x^*) + J_F(x^*)(x-x^*).$$

Utilizando o Lema 1.14, vemos que

$$\|F(x)\| \leq \left\|J_F(x^*)(x-x^*)\right\| + \frac{L}{2}\|x-x^*\|^2. \tag{4.30}$$

Dado $\delta \in (0, 1)$, defina

$$\hat{\varepsilon} = \min_{\|u\|=1}\left\{\frac{2\delta}{L}\left\|J_F(x^*)u\right\|\right\}.$$

Observe que $\hat{\varepsilon} > 0$, pois $J_F(x^*)$ é não singular e u pertence a esfera unitária. Assim, se $\|x - x^*\| \leq \hat{\varepsilon}$, então

$$\frac{L}{2}\|x-x^*\|^2 \leq \frac{L}{2}\hat{\varepsilon}\|x-x^*\| \leq \frac{L}{2}\|x-x^*\|\frac{2\delta}{L}\left\|J_F(x^*)\frac{x-x^*}{\|x-x^*\|}\right\| = \delta\left\|J_F(x^*)(x-x^*)\right\|.$$

Daí, por (4.30) temos

$$\|F(x)\| \leq \|J_F(x^*)(x-x^*)\| + \delta \|J_F(x^*)(x-x^*)\|$$

com $\|x - x^*\| \leq \hat{\varepsilon}$. Para mostrar a outra desigualdade, note que

$$J_F(x^*)(x-x^*) = F(x) - \Big[F(x) - F(x^*) - J_F(x^*)(x-x^*)\Big].$$

Usando novamente o Lema 1.14, temos

$$\|J_F(x^*)(x-x^*)\| \leq \|F(x)\| + \frac{L}{2}\|x-x^*\|^2.$$

Logo, para todo x tal que $\|x - x^*\| \leq \hat{\varepsilon}$, temos que

$$\|J_F(x^*)(x-x^*)\| \leq \|F(x)\| + \delta \|J_F(x^*)(x-x^*)\|$$

completando a demonstração. □

Como já observado, a primícia do Algoritmo 4.1 é não usar tensores e resolver de forma inexata dois sistemas lineares. Para isso, as matrizes dos coeficientes desses sistemas devem ser não singulares. Além disso, devemos ter um certo controle no tamanho no passo s^k, mais especificamente, no tamanho das soluções inexatas $s^k_{(1)}$ e $s^k_{(2)}$ desses sistemas. Os próximos lemas fornecem condições suficientes para atender estes quesitos.

Lema 4.18 *Seja $x^* \in \mathbb{R}^n$ uma solução do problema (3.1), com $J_F(x^*)$ não singular. Dados $\bar{\eta} \in (0,1)$, considere $\tilde{\varepsilon} > 0$ e $\gamma > 0$ dados no Lema 4.16 e uma aplicação $r_1 : B(x^*, \tilde{\varepsilon}) \to \mathbb{R}^n$ tal que*

$$\|r_1(x)\| \leq \bar{\eta}\|F(x)\|$$

para todo $x \in B(x^*, \tilde{\varepsilon})$. Seja $s_1 : B(x^*, \tilde{\varepsilon}) \to \mathbb{R}^n$ tal que

$$s_1(x) = J_F(x)^{-1}\Big(-F(x) + r_1(x)\Big). \tag{4.31}$$

Nestas condições, temos

$$\|s_1(x)\| \leq 2\gamma\|F(x)\|. \tag{4.32}$$

Prova. Observe inicialmente que s_1 está bem definido, pois como J_F é contínua em x^*, o Lema 4.16 garante que $J_F(x)$ seja não singular para todo $x \in B(x^*, \tilde{\varepsilon})$. Além disso, $\left\|J_F(x)^{-1}\right\| \leq \gamma$ para todo $x \in B(x^*, \tilde{\varepsilon})$. Desta forma,

$$\|s_1(x)\| \leq \gamma\Big(\|F(x)\| + \bar{\eta}\|F(x)\|\Big)$$

donde segue (4.32). □

Lema 4.19 *Seja $x^* \in \mathbb{R}^n$ uma solução do problema (3.1), com $J_F(x^*)$ não singular. Considere $\tilde{\varepsilon} > 0$ e $\gamma > 0$ dados pelo Lema 4.16 e $\alpha \in \mathbb{R}$. Considere a aplicação $C : B(x^*, \tilde{\varepsilon}) \to \mathbb{R}^{n\times n}$ cumprindo a condição (4.23) e a aplicação $A : B(x^*, \tilde{\varepsilon}) \to \mathbb{R}^{n\times n}$, tal que*

$$A(x) = J_F(x) + \alpha C(x).$$

Então, existe $\bar{\varepsilon} \in (0, \tilde{\varepsilon}]$ tal que $A(x)$ é não singular e

$$\left\|A(x)^{-1}\right\| \leq \gamma$$

para todo $x \in B(x^*, \bar{\varepsilon})$.

Prova. Observe que $A(x^*) = J_F(x^*)$, pois por (4.23), $C(x^*) = 0$. Como A é contínua e $A(x^*)$ é não singular, podemos aplicar o Lema 4.16 para completar a demonstração. □

Lema 4.20 *Seja $x^* \in \mathbb{R}^n$ uma solução do problema (3.1), com $J_F(x^*)$ não singular. Considere $\bar{\varepsilon} > 0$ do Lema 4.19 e $\gamma > 0$ do Lema 4.16. Sejam A e C as aplicações dadas no Lema 4.19 e s_1 a aplicação dada no Lema 4.18, $\bar{\eta} \in (0, 1)$ e uma aplicação $r_2 : B(x^*, \bar{\varepsilon}) \to \mathbb{R}^n$ tal que*

$$r_2(x) \leq \bar{\eta} \left\| \frac{1}{2} C(x) s_1(x) \right\|$$

para todo $x \in B(x^*, \bar{\varepsilon})$. Seja $s_2 : B(x^*, \bar{\varepsilon}) \to \mathbb{R}^n$ tal que

$$s_2(x) = A(x)^{-1}\Big(-\frac{1}{2} C(x) s_1(x) + r_2(x)\Big).$$

Existe $M > 0$ tal que

$$\|s_2(x)\| \leq M \|F(x)\|^2$$

para todo $x \in B(x^*, \bar{\varepsilon})$.

Prova. Observe que s_2 está bem definido, pois $A(x)$ e $J_F(x)$ são não singulares na bola $B(x^*, \bar{\varepsilon})$. Além disso, existe $p > 0$ tal que

$$\|C(x)\| \leq p \|F(x)\|$$

para todo $x \in B(x^*, \bar{\varepsilon})$. Portanto, usando (4.32), temos que

$$\left\| \frac{1}{2} C(x) s_1(x) \right\| \leq p\gamma \|F(x)\|^2$$

para todo $x \in B(x^*, \bar{\varepsilon})$. Definindo $M = 2p\gamma^2$, obtemos

$$\|s_2(x)\| \leq \gamma\Big(\left\| \frac{1}{2} C(x)s_1(x) \right\| + \bar{\eta} \left\| \frac{1}{2} C(x)s_1(x) \right\| \Big) \leq M \|F(x)\|^2 .$$

□

Diante dos lemas apresentados, podemos agora estabelecer uma das contribuições principais desta tese: a prova de convergência da Classe Chebyshev-Halley Inexata livre de tensores. Tal resultado é dado no próximo teorema.

Teorema 4.21 *Sejam $x^* \in \mathbb{R}^n$ uma solução do problema (3.1), $\mathcal{T}_F$ Lipschitz com constante L_2 em uma bola $B(x^*, \tilde{\varepsilon})$ e uma aplicação C satisfazendo (4.23). Suponha que $J_F(x^*)$ seja não singular. Existem $0 < \bar{\eta} < \tilde{\eta} < 1$, $\varepsilon > 0$ tais que, se $\eta_k^{(i)} \leq \bar{\eta} < \tilde{\eta} < 1$ com $i = 1, 2$, então para todo $x^0 \in B(x^*, \varepsilon)$, o Algoritmo 4.1 gera uma sequência (x^k) tal que $x^k \in B(x^*, \varepsilon)$ para todo $k \in \mathbb{N}$ e $x^k \to x^*$ linearmente. Além disso, a taxa de convergência é*

1. *superlinear se $\eta_k^{(1)} \to 0$.*
2. *quadrática se $\eta_k^{(1)} = O\Big(\|F(x^k)\| \Big)$.*

Se adicionalmente

$$\eta_k^{(1)} = O\Big(\|F(x^k)\|^2 \Big),\ \eta_k^{(2)} = O\Big(\|F(x^k)\| \Big) \ \ e \ \ \|\mathcal{T}_F(x)s_{(1)}^k - C(x^k)\| = O\Big(\|F(x^k)\|^w \Big)$$

para $w \in (1, 2]$, então a taxa de convergência é

3. *superquadrática se $1 < w < 2$.*
4. *cúbica se $w = 2$.*

Prova. Seja

$$\mu = \max\{\|J_F(x^*)\|, \|J_F(x^*)^{-1}\|\}. \tag{4.33}$$

Considere

$$0 < \bar{\eta} < \tilde{\eta} < \min\left\{1, \frac{1}{\mu^2}\right\},$$

$\delta \in (0, 1)$ suficientemente pequeno tal que

$$\frac{\tilde{\eta}(1+\delta)\mu^2}{1-\delta} < 1, \tag{4.34}$$

$\bar{\varepsilon} > 0$ dado no Lema 4.19, $\gamma > 0$ dado no Lema 4.16 e $\hat{\varepsilon} > 0$ dado no Lema 4.17. Tome

$$\varepsilon_1 = \min\{\bar{\varepsilon}, \hat{\varepsilon}, \tilde{\varepsilon}\}.$$

Pelos Lemas 4.16 e 4.19, temos que

$$\left\|J_F(x)^{-1}\right\| \leq \gamma, \tag{4.35}$$

$$\left\|\Big(J_F(x)+\alpha C(x)\Big)^{-1}\right\| \leq \gamma. \tag{4.36}$$

para todo $x \in B(x^*, \varepsilon_1)$. Como $F \in \mathcal{C}^2$, podemos supor, sem perda de generalidade, que

$$\|\mathcal{T}_F(x)\| \leq \gamma \tag{4.37}$$

para todo $x \in B(x^*, \varepsilon_1)$.

Além disso, seja $M > 0$ dado no Lema 4.20 e suponha também, para todo $x \in B(x^*, \varepsilon_1)$ e para todo $k \geq 0$, que

$$\|F(x)\| \leq \gamma, \tag{4.38}$$

$$\eta_k^{(1)} + p\gamma\eta_k^{(2)}\|F(x)\| + a(x) < \tilde{\eta}, \tag{4.39}$$

onde

$$a(x) = \gamma(2\gamma^2+p)\|F(x)\| + \left[pM|\alpha| + 2\gamma^2 M + \frac{L_2}{6}\Big(2\gamma+\gamma M\Big)^3\right]\|F(x)\|^2 + \frac{1}{2}\gamma M^2\|F(x)\|^3. \tag{4.40}$$

Considere as aplicações s_1 e s_2 dos Lemas 4.18 e 4.20, respectivamente. Como $s_{(1)}^k = s_1(x^k)$ e $s_{(2)}^k = s_2(x^k)$ e o passo s^k gerado pelo Algoritmo 4.1 é $s_{(1)}^k + s_{(2)}^k$, usando (4.38) e os Lemas 4.18 e 4.20, temos que

$$\left\|s^k\right\| \leq \left[2\gamma + M\left\|F(x^k)\right\|\right]\left\|F(x^k)\right\| \tag{4.41}$$

$$\leq \Big[2\gamma+\gamma M\Big]\left\|F(x^k)\right\| \tag{4.42}$$

para todo $x^k \in B(x^*, \varepsilon_1)$. Agora, de (4.28), temos

$$J_F(x^k)s_{(2)}^k = r_{(2)}^k - \frac{1}{2}C(x^k)s_{(1)}^k - \alpha C(x^k)s_{(2)}^k \tag{4.43}$$

para todo $k \geq 0$. Pela definição de $r_{(1)}^k$ e por (4.43), o modelo quadrático de F em torno

de s^k pode ser escrito como

$$\begin{aligned} M_k(s^k) &= F(x^k) + J_F(x^k)s^k + \frac{1}{2}\mathcal{T}_F(x^k)s^k s^k \\ &= F(x^k) + J_F(x^k)s^k_{(1)} + J_F(x^k)s^k_{(2)} + \frac{1}{2}\mathcal{T}_F(x^k)s^k_{(1)}s^k_{(1)} + \mathcal{T}_F(x^k)s^k_{(1)}s^k_{(2)} + \\ &+ \frac{1}{2}\mathcal{T}_F(x^k)s^k_{(2)}s^k_{(2)} \\ &= r^k_{(1)} + r^k_{(2)} + \frac{1}{2}\Big[\mathcal{T}_F(x^k)s^k_{(1)} - C(x^k)\Big]s^k_{(1)} - \alpha C(x^k)s^k_{(2)} + \mathcal{T}_F(x^k)s^k_{(1)}s^k_{(2)} + \\ &+ \frac{1}{2}\mathcal{T}_F(x^k)s^k_{(2)}s^k_{(2)} \end{aligned}$$

Como existe $p > 0$ tal que

$$\left\|C(x^k)\right\| \leq p\left\|F(x^k)\right\|, \tag{4.44}$$

por (4.26) e pelo Lema 4.18, temos

$$\left\|r^k_{(2)}\right\| \leq p\gamma\eta^{(2)}_k \left\|F(x^k)\right\|^2. \tag{4.45}$$

Daí, por (4.25), (4.37), (4.44), (4.45) e pelos Lemas 4.18 e 4.20, temos que

$$\begin{aligned} \left\|M_k(s^k)\right\| \leq\ & \eta^{(1)}_k \left\|F(x^k)\right\| + p\gamma\eta^{(2)}_k \left\|F(x^k)\right\|^2 + \gamma\left\|\mathcal{T}_F(x^k)s^k_{(1)} - C(x^k)\right\| \left\|F(x^k)\right\| + \\ &+ (pM|\alpha| + 2\gamma^2 M)\left\|F(x^k)\right\|^3 + \frac{1}{2}\gamma M^2\left\|F(x^k)\right\|^4 \end{aligned} \tag{4.46}$$

para todo $x^k \in B(x^*, \varepsilon_1)$. Para mostrar a convergência, não há necessidade de nenhuma hipótese adicional sobre a aplicação C. Apenas vamos observar que

$$\|\mathcal{T}_F(x)s_1(x) - C(x)\| \leq \|\mathcal{T}_F(x)\|\,\|s_1(x)\| + \|C(x)\| \leq (2\gamma^2 + p)\,\|F(x)\| \tag{4.47}$$

para todo $x \in B(x^*, \varepsilon_1)$. Desta maneira, fazendo

$$F(x^k + s^k) = M_k(s^k) + F(x^k + s^k) - M_k(s^k),$$

pelo Lema 1.15 e usando (4.39), (4.40), (4.42), (4.46) e (4.47) temos, para todo $x^k \in$

$B(x^*, \varepsilon_1)$, que

$$\begin{aligned}
\left\|F(x^k+s^k)\right\| &\leq \left\|M_k(s^k)\right\| + \left\|F(x^k+s^k) - M_k(s^k)\right\| \\
&\leq \eta_k^{(1)} \left\|F(x^k)\right\| + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\|^2 + \gamma \left\|\mathcal{T}_F(x^k)s_{(1)}^k - C(x^k)\right\| \left\|F(x^k)\right\| + \\
&+ (pM|\alpha| + 2\gamma^2 M) \left\|F(x^k)\right\|^3 + \frac{1}{2}\gamma M^2 \left\|F(x^k)\right\|^4 + \frac{L_2}{6} \left\|s^k\right\|^3 \leq \\
&\leq \left\{\eta_k^{(1)} + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\| + \gamma \left\|\mathcal{T}_F(x^k)s_{(1)}^k - C(x^k)\right\| + \right. \qquad (4.48) \\
&+ (pM|\alpha| + 2\gamma^2 M) \left\|F(x^k)\right\|^2 + \frac{1}{2}\gamma M^2 \left\|F(x^k)\right\|^3 + \\
&+ \left. \frac{L_2}{6}(2\gamma + \gamma M)^3 \left\|F(x^k)\right\|^2 \right\} \left\|F(x^k)\right\| \leq \\
&\leq \left\{\eta_k^{(1)} + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\| + \gamma(2\gamma^2 + p) \left\|F(x^k)\right\| + \right. \qquad (4.49) \\
&+ \left[pM|\alpha| + 2\gamma^2 M + \frac{L_2}{6}(2\gamma + \gamma M)^3\right] \left\|F(x^k)\right\|^2 + \\
&+ \left. \frac{1}{2}\gamma M^2 \left\|F(x^k)\right\|^3 \right\} \left\|F(x^k)\right\| = \\
&= \left\{\eta_k^{(1)} + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\| + a(x^k)\right\} \left\|F(x^k)\right\| \leq \qquad (4.50) \\
&\leq \tilde{\eta} \left\|F(x^k)\right\| \qquad (4.51)
\end{aligned}$$

Por continuidade, existe $\varepsilon_2 \in (0, \varepsilon_1]$ tal que

$$\|F(x)\| \leq \frac{\varepsilon_1}{2\Big[2\gamma + \gamma M\Big]}$$

para todo $x \in B(x^*, \varepsilon_2)$. Desta maneira, usando (4.42), temos

$$\left\|s^k\right\| \leq \frac{\varepsilon_1}{2}$$

para todo $x^k \in B(x^*, \varepsilon_2)$. Defina $\varepsilon = \dfrac{\varepsilon_2}{2}$. Afirmamos que

$$\text{se } x^k \in B(x^*, \varepsilon), \text{ então } x^k + s^k \in B(x^*, \varepsilon_1). \qquad (4.52)$$

De fato,

$$\left\|x^k + s^k - x^*\right\| \leq \left\|x^k - x^*\right\| + \left\|s^k\right\| \leq \varepsilon + \frac{\varepsilon_1}{2} \leq \varepsilon_1.$$

A relação (4.52) permite aplicar o resultado do Lema 4.17 para os pontos x^k e $x^k + s^k$ simultaneamente usando (4.51). Vamos mostrar que $x^k + s^k$ pertence a bola $B(x^*, \varepsilon)$, o que caracterizará a boa definição da sequência (x^k). De fato, seja a norma-$J_F(x^*)$ definida como

$$\|y\|_* = \|J_F(x^*)y\|$$

para todo $y \in \mathbb{R}^n$. Note que pela definição de μ dada em (4.33), temos

$$\begin{aligned} \|y\| &\leq \mu \|y\|_* \\ \|y\|_* &\leq \mu \|y\| \end{aligned}$$

para todo $y \in \mathbb{R}^n$. Daí segue que, dado $\delta \in (0,1)$ e notando que $\varepsilon_1 \leq \hat{\varepsilon}$, temos pelo Lema 4.17 e por (4.51) que

$$\begin{aligned} (1-\delta)\left\|x^k + s^k - x^*\right\| \leq (1-\delta)\mu \left\|x^k + s^k - x^*\right\|_* &\leq \mu \left\|F(x^k + s^k)\right\| \\ &\leq \mu\tilde{\eta} \left\|F(x^k)\right\| \\ &\leq \mu\tilde{\eta}(1+\delta) \left\|x^k - x^*\right\|_* \\ &\leq \mu^2\tilde{\eta}(1+\delta) \left\|x^k - x^*\right\|. \end{aligned}$$

Como $x^{k+1} = x^k + s^k$, temos

$$\left\|x^{k+1} - x^*\right\| \leq \frac{\tilde{\eta}(1+\delta)\mu^2}{1-\delta} \left\|x^k - x^*\right\| \tag{4.53}$$

Por (4.34) e (4.53), concluímos que a sequência (x^k) gerada pelo Algoritmo 4.1 está bem definida, que $x^k \in B(x^*, \varepsilon)$ para todo $k \in \mathbb{N}$ e que $x^k \to x^*$ linearmente.

Para provar a convergência superlinear, observe que da relação (4.50) e usando o Lema 4.17, dado $\delta \in (0,1)$, temos que

$$\begin{aligned} (1-\delta)\left\|x^{k+1} - x^*\right\| &\leq (1-\delta)\mu \left\|x^{k+1} - x^*\right\|_* \\ &\leq \mu \left\|F(x^{k+1})\right\| \\ &\leq \mu\left[\eta_k^{(1)} + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\| + a(x^k)\right] \left\|F(x^k)\right\| \\ &\leq \mu\left[\eta_k^{(1)} + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\| + a(x^k)\right](1+\delta) \left\|x^k - x^*\right\|_* \\ &\leq \mu^2\left[\eta_k^{(1)} + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\| + a(x^k)\right](1+\delta) \left\|x^k - x^*\right\|. \end{aligned}$$

Como $\eta_k^{(1)} \to 0$ e $a(x^k) \to 0$, temos que

$$\frac{\left\|x^{k+1} - x^*\right\|}{\|x^k - x^*\|} \leq \frac{\mu^2\left[\eta_k^{(1)} + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\| + a(x^k)\right](1+\delta)}{1-\delta} \to 0.$$

Para provar a convergência quadrática, sejam $p_1 > 0$ tal que

$$\eta_k^{(1)} \leq p_1 \left\|F(x^k)\right\|$$

e

$$\rho_1 = p_1 + p\gamma\bar{\eta} + \gamma(2\gamma^2 + p) + \Big[pM|\alpha| + 2\gamma^2 M + \frac{L_2}{6}(2\gamma + \gamma M)^3\Big]\gamma + \frac{1}{2}\gamma^3 M^2.$$

Observe que $\rho_1 > 0$. Da relação (4.49) e usando (4.38) temos

$$\begin{aligned}
\left\|F(x^k + s^k)\right\| &\leq \Big\{p_1 \left\|F(x^k)\right\| + p\gamma\eta_k^{(2)} \left\|F(x^k)\right\| + \gamma(2\gamma^2 + p) \left\|F(x^k)\right\| + \\
&+ \Big[pM|\alpha| + 2\gamma^2 M + \frac{L_2}{6}(2\gamma + \gamma M)^3\Big] \left\|F(x^k)\right\|^2 + \\
&+ \frac{1}{2}\gamma M^2 \left\|F(x^k)\right\|^3 \Big\} \left\|F(x^k)\right\| \leq \\
&\leq \Big\{p_1 + p\gamma\eta_k^{(2)} + \gamma(2\gamma^2 + p) + \\
&+ \Big[pM|\alpha| + 2\gamma^2 M + \frac{L_2}{6}(2\gamma + \gamma M)^3\Big] \left\|F(x^k)\right\| + \\
&+ \frac{1}{2}\gamma M^2 \left\|F(x^k)\right\|^2 \Big\} \left\|F(x^k)\right\|^2 \\
&\leq \rho_1 \left\|F(x^k)\right\|^2.
\end{aligned}$$

Como $x^{k+1} = x^k + x^k$, pelo Lema 4.17, dado $\delta \in (0, 1)$, temos que

$$\begin{aligned}
(1 - \delta)\left\|x^{k+1} - x^*\right\| &\leq (1 - \delta)\mu \left\|x^{k+1} - x^*\right\|_* \\
&\leq \mu \left\|F(x^{k+1})\right\| \\
&\leq \mu\rho_1 \left\|F(x^k)\right\|^2 \\
&\leq \mu\rho_1(1 + \delta)^2 \left\|x^k - x^*\right\|_*^2 \\
&\leq \mu^3\rho_1(1 + \delta)^2 \left\|x^k - x^*\right\|^2.
\end{aligned}$$

Daí segue que

$$\left\|x^{k+1} - x^*\right\| \leq \frac{\mu^3\rho_1(1 + \delta)^2}{1 - \delta} \left\|x^k - x^*\right\|^2.$$

Observe que na prova da convergência superlinear e quadrática, usamos apenas o fato que $\|C(x)\| = O\Big(\left\|F(x)\right\|\Big)$ e (4.47). Para provar o que falta, além das hipóteses sobre os termos forçantes, vamos também utilizar a hipótese que

$$\left\|\mathcal{T}_F(x^k)s_{(1)}^k - C(x^k)\right\| = O\Big(\left\|F(x^k)\right\|^w\Big) \quad \text{para } w \in (1, 2]. \tag{4.54}$$

Para isso, sejam $p_2, q_2, q_3 > 0$ tais que

$$\eta_k^{(1)} \leq p_2 \left\|F(x^k)\right\|^2, \quad \eta_k^{(2)} \leq q_2 \left\|F(x^k)\right\| \quad \text{e} \quad \left\|\mathcal{T}_F(x^k)s_{(1)}^k - C(x^k)\right\| \leq q_3 \left\|F(x^k)\right\|^w.$$

Defina

$$\rho_2 = p_2 + p\gamma q_2 + pM|\alpha| + 2\gamma^2 M + \frac{1}{2}\gamma^2 M^2 + \frac{L_2}{6}(2\gamma + \gamma M)^3.$$

Da relação (4.48) e usando (4.38), temos

$$\begin{aligned}
\left\|F(x^k+s^k)\right\| &\leq \left[p_2\left\|F(x^k)\right\|^2 + p\gamma q_2\left\|F(x^k)\right\|^2 + \gamma q_3\left\|F(x^k)\right\|^w + \right.\\
&+ (pM|\alpha| + 2\gamma^2 M)\left\|F(x^k)\right\|^2 + \frac{1}{2}\gamma M^2\left\|F(x^k)\right\|^3 + \\
&+ \left.\frac{L_2}{6}(2\gamma+\gamma M)^3\left\|F(x^k)\right\|^2\right]\left\|F(x^k)\right\| \leq \\
&\leq \left[p_2 + p\gamma q_2 + pM|\alpha| + 2\gamma^2 M + \frac{1}{2}\gamma M^2\left\|F(x^k)\right\| + \right.\\
&+ \left.\frac{L_2}{6}(2\gamma+\gamma M)^3\right]\left\|F(x^k)\right\|^3 + \gamma q_3\left\|F(x^k)\right\|^{w+1} \leq \\
&\leq \rho_2\left\|F(x^k)\right\|^3 + \gamma q_3\left\|F(x^k)\right\|^{w+1}.
\end{aligned}$$

Novamente pelo Lema 4.17, dado $\delta \in (0,1)$, temos que

$$\begin{aligned}
(1-\delta)\left\|x^{k+1}-x^*\right\| &\leq (1-\delta)\mu\left\|x^{k+1}-x^*\right\|_* \\
&\leq \mu\left\|F(x^{k+1})\right\| \\
&\leq \mu\rho_2\left\|F(x^k)\right\|^3 + \mu\gamma q_3\left\|F(x^k)\right\|^{w+1} \\
&\leq \mu\rho_2(1+\delta)^3\left\|x^k-x^*\right\|_*^3 + \mu\gamma q_3(1+\delta)^{w+1}\left\|x^k-x^*\right\|_*^{w+1} \\
&\leq \mu^4\rho_2(1+\delta)^3\left\|x^k-x^*\right\|^3 + \mu^{w+2}\gamma q_3(1+\delta)^{w+1}\left\|x^k-x^*\right\|^{w+1}.
\end{aligned}$$

Daí, segue que

$$\left\|x^{k+1}-x^*\right\| \leq \frac{\left[\mu^4\rho_2(1+\delta)^3\left\|x^k-x^*\right\| + \mu^{w+2}\gamma q_3(1+\delta)^{w+1}\left\|x^k-x^*\right\|^{w-1}\right]\left\|x^k-x^*\right\|^2}{1-\delta}. \tag{4.55}$$

A convergência superquadrática decorre de (4.55) observando que

$$\frac{\left\|x^{k+1}-x^*\right\|}{\left\|x^k-x^*\right\|^2} \leq \frac{\left[\mu^4\rho_2(1+\delta)^3\left\|x^k-x^*\right\| + \mu^{w+2}\gamma q_3(1+\delta)^{w+1}\left\|x^k-x^*\right\|^{w-1}\right]}{1-\delta} \to 0$$

e basta tomar $w = 2$ em (4.55) para garantir a convergência cúbica. □

Agora vamos mostrar que existem matrizes, além da matriz nula, que cumprem a condição (4.23) na bola $B(x^*,\varepsilon)$ onde ε é dado no Teorema 4.21. Uma matriz também trivial é

$$C(x) = \mathcal{T}_F(x)s_1(x),$$

pois usando (4.37) e o Lema 4.18, temos $\|C(x)\| \leq 2\gamma^2\|F(x)\|$. Na verdade, qualquer matriz pertencente ao conjunto

$$\mathbf{C} = \{\mathcal{B}s_1(x) \mid \mathcal{B} \in U \subset \mathbb{R}^{n\times n\times n}(U \text{ limitado}) \text{ e } x \in B(x^*,\varepsilon)\}$$

satisfaz a condição (4.23).

Assim, mostramos que a condição (4.23) pode ser facilmente verificada. Por outro lado, apenas a limitação do tensor $\mathcal{B}$ não é suficiente para mostrar que a hipótese (4.54) seja satisfeita. Mostraremos no próximo lema, que as hipóteses (4.23) e (4.54) podem ser verificadas para uma determinada matriz.

Lema 4.22 *Sejam $x^* \in \mathbb{R}^n$ uma solução do problema (3.1), $\mathcal{T}_F$ Lipschitz com constante $L_2 > 0$ na bola $B(x^*, \varepsilon)$, onde $\varepsilon > 0$ é dado no Teorema 4.21 e s_1 dado no Lema 4.18. Dado $h > 0$, defina $C : B(x^*, \varepsilon) \to \mathbb{R}^{n\times n}$ por*

$$C(x) = \frac{J_F(x + hs_1(x)) - J_F(x)}{h}.$$

Temos que $\|C(x)\| = O\Big(\|F(x)\|\Big)$ e $\|\mathcal{T}_F(x)s_1(x) - C(x)\| = O\Big(\|F(x)\|^2\Big)$.

Prova. Como $\mathcal{T}_F$ é Lipschitz na bola $B(x^*, \varepsilon)$, temos

$$\|\mathcal{T}_F(x)\| \leq \|\mathcal{T}_F(x) - \mathcal{T}_F(x^*)\| + \|\mathcal{T}_F(x^*)\| \leq L_2\varepsilon + \|\mathcal{T}_F(x^*)\| \stackrel{def}{=} M$$

para todo $x \in B(x^*, \varepsilon)$. Daí segue de (1.21) que J_F é Lipschitz com constante M. Assim, utilizando o Lema 4.18, temos que

$$\|C(x)\| \leq \frac{1}{h}\|J_F(x + hs_1(x)) - J_F(x)\| \leq \frac{1}{h}Mh\|s_1(x)\| \leq 2\gamma M\|F(x)\|$$

para todo $x \in B(x^*, \varepsilon)$. Para mostrar o que falta, pela desigualdade (1.20) temos que

$$\|J_F(x + hs_1(x)) - J_F(x) - \mathcal{T}_F(x)hs_1(x)\| \leq \frac{L_2}{2}\|hs_1(x)\|^2.$$

Dividindo esta expressão por h, obtemos

$$\left\|\frac{J_F(x + hs_1(x)) - J_F(x)}{h} - \mathcal{T}_F(x)s_1(x)\right\| \leq \frac{L_2}{2}h\|s_1(x)\|^2.$$

Desta forma, pelo Lema 4.18, temos que

$$\|C(x) - \mathcal{T}_F(x)s_1(x)\| \leq 2L_2h\gamma^2\|F(x)\|^2.$$

□

Com isso, mostramos que as hipóteses exigidas no Teorema 4.21 são hipóteses razoáveis, ou seja, elas podem ser verificadas.

Capítulo 5

Contribuições da Tese II - Implementação

No Capítulo 4 apresentamos uma modificação da classe Chebyshev-Halley com o objetivo de reduzir seu custo computacional. Introduzimos uma nova classe de métodos chamada classe Chebyshev-Halley Inexata livre de tensores, a qual não faz uso do tensor $\mathcal{T}_F(x^k)$ a cada iteração e os dois sistemas lineares, que fornecem o passo, podem ser resolvidos de maneira inexata.

O objetivo deste capítulo é analisar o desempenho computacional do Algoritmo 4.1 proposto nesta tese, aplicada aos métodos clássicos da classe Chebyshev-Halley, nos quesitos eficiência e robustez. Para isso, utilizamos o conjunto de problemas utilizados em La Cruz, Martínez e Raydan [18] e os da Seção 4 de Lukšan e Vlček [56]. Estes problemas, bem como os pontos iniciais adotados, podem ser consultados no Apêndice B. As estatísticas (medidas de desempenho) que coletamos foram número de iterações, número de avaliações de função e tempo computacional e todos os métodos foram implementados em MATLAB R2010b em um notebook Dell XPS15 (L502X), 2,5 GHz, RAM de 6 Gb, processador Intel® Core^{TM} i5-2450M .

Para comparar os métodos, utilizamos a ferramenta *performance profile* proposta por Dolan e Moré [23]. A ideia é basicamente comparar uma medida de desempenho de um determinado algoritmo na resolução de um problema, com a melhor medida de desempenho determinada entre todos os algoritmos. Para isso, é definido o índice de desempenho do algoritmo s na resolução do problema p como sendo

$$r_{p,s} = \begin{cases} \dfrac{m_{p,s}}{\min\{m_{p,j} \mid j \in S\}}, & \text{se o algoritmo } s \text{ resolveu o problema } p \\ r_M, & \text{caso contrário,} \end{cases}$$

onde $m_{p,j}$ é uma medida de desempenho avaliada pelo algoritmo $j \in S$ na resolução

do problema p, S é um conjunto de algoritmos aplicados na resolução do problema p e $r_M \geq \max\{r_{p,s}\}$ é um parâmetro definido previamente.

Quanto à eficiência, um algoritmo $\bar{s}$ será mais eficiente na resolução do problema p quando $r_{p,\bar{s}} = 1$ e quanto maior for esse valor, pior o desempenho do respectivo algoritmo.

Além disso, em [23] é introduzida a função distribuição de probabilidade $\rho_s : [1, \infty) \to [0, 1]$, para cada algoritmo $s \in S$, definida por

$$\rho_s(\tau) = \frac{\text{card}\{p \in P \mid r_{p,s} \leq \tau\}}{\text{card}\{P\}},$$

onde P é um conjunto de problemas que estão sendo resolvidos pelo algoritmo $s \in S$. Note que os algoritmos com maiores valores para $\rho_s(1)$ são os mais eficientes. A grosso modo, o valor $\rho_s(\tau)$ significa a porcentagem de problemas que o algoritmo s resolve em τ vezes o valor da medida de desempenho do algoritmo mais eficiente.

Quanto à robustez, devemos observar o valor de τ para o qual $\rho_s(\tau) = 1$. Quanto menor for esse valor, mais robusto será o algoritmo. Assumimos que $r_{p,s} \in [1, r_M]$ e que $r_{p,s} = r_M$ somente quando o algoritmo s não resolveu o problema p. Isto significa que $\rho_s(r_M) = 1$. Desta forma, a probabilidade de um algoritmo s resolver um problema é medido como

$$\rho_s^* = \lim_{\tau \to r_M^-} \rho_s(\tau).$$

Agora vamos justificar numericamente a necessidade de modificar os métodos da classe Chebyshev-Halley comparando-os com o método de Newton. Aqui não estamos preocupados com métodos diretos (exatos) para resolução de sistemas lineares. Por esse motivo, quando necessário, usaremos um recurso próprio do MATLAB para resolver um sistema linear que é o operador \. Problemas em que o MATLAB detectou singularidade de matrizes foram declarados como problemas não resolvidos.

Em todos os testes que apresentados, foram considerados os problemas citados no Apêndice B. Em sua maioria, as dimensões consideradas foram $n = 30$ e $n = 50$ com 3 pontos iniciais para cada dimensão, totalizando assim 276 problemas.

Declaramos falha nos algoritmos quando

$$\left\|F(x^k)\right\|_\infty > 10^{20}$$

para algum $k \in \mathbb{N}$ ou quando atinge o número máximo de iterações $k = 200$. O critério de parada adotado foi

$$\left\|F(x^k)\right\|_\infty \leq 10^{-8}. \tag{5.1}$$

Em uma primeira análise, a Tabela 5.1 mostra a porcentagem de problemas resolvidos pelos algoritmos testados.

	% de problemas resolvidos
Newton	72,46%
Halley	76,44%
Chebyshev	66,66%
Super-Halley	67,39%

Tabela 5.1: Percentual de problemas resolvidos pelos métodos Newton, Halley, Chebyshev e Super-Halley

Podemos perceber que o método de Halley atingiu o critério de parada (5.1) em $76,44\%$ dos problemas, enquanto que para o método de Newton a porcentagem foi de $72,46\%$. Isto mostra que o método de Halley foi mais robusto que o método de Newton para os problemas considerados. Os métodos Chebyshev e Super-Halley foram os menos robustos.

Quanto à eficiência, vamos analisar as três medidas de desempenho citadas no início deste capítulo.

Como os métodos pertencentes à classe Chebyshev-Halley possuem taxa de convergência cúbica, é de se esperar que eles sejam mais eficientes que Newton em relação ao número de iterações. O gráfico de desempenho do número de iterações está ilustrado na Figura 5.1.

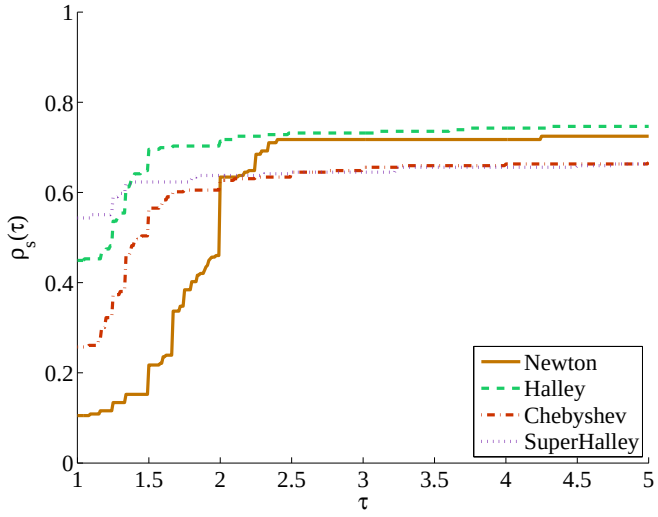

Figura 5.1: Gráfico de desempenho do número de iterações dos métodos Newton, Halley, Chebyshev e Super-Halley.

Vemos que os métodos clássicos pertencentes à classe Chebyshev-Halley foram mais eficientes que o método de Newton. O melhor algoritmo foi o método Super-Halley, que resolveu aproximadamente $54,34\%$ dos problemas com o menor número de iterações, enquanto que os métodos Halley, Chebyshev e Newton resolveram, respectivamente, em torno de $44,92\%$, $25,72\%$ e $10,5\%$ dos problemas com o menor número de iterações.

Apesar do método Super-Halley ter sido mais eficiente que o método de Halley, pode-

mos observar que para $\tau = 1,34$, ambos os algoritmos resolveram $61,95\%$ dos problemas e para valores de τ superiores a $1,34$, o método de Halley foi o mais eficiente, além de ter sido o mais robusto. Em relação ao método de Newton, podemos constatar que ele teve o mesmo desempenho que o método Super-Halley considerando o dobro do número de iterações do melhor algoritmo. Isto corresponde aproximadamente a $63,5\%$ dos problemas.

Evidentemente, o bom desempenho dos métodos clássicos da classe Chebyshev-Halley, no quesito número de iterações, não refletem diretamente na eficiência como um todo.

Para comparar o número de avaliações de função em todos os testes apresentados nesta tese, nos baseamos no trabalho de Griewank, Juedes e Utke [36], onde afirma que o custo da avaliação de derivadas é no máximo 5 vezes o custo da avaliação da função. Escolhendo o peso 3, definimos uma medida que leva em consideração o número de avaliações de cada função coordenada de F $(\#f_i, i = 1, \ldots, n)$, dos gradientes de cada função coordenada de F $(\#g_i, i = 1, \ldots, n)$ e das jacobianas (Hessianas) de cada gradiente respectivo $(\#H_i, i = 1, \ldots, n)$, dada por

$$nf = n\#f_1 + 3n\#g_1 + 3n\#H_1.$$

Usando esta metodologia, podemos notar na Figura 5.2 que o método de Newton foi o mais eficiente. Isso era de certa forma esperado, pois o cálculo do tensor nos outros métodos é excessivamente caro computacionalmente. O método de Newton resolveu aproximadamente $53,25\%$ dos problemas com o menor número de avaliações de função, enquanto que os métodos Halley, Chebyshev e Super-Halley resolveram, respectivamente, em torno de $11,59\%$, $0,7\%$ e $21,73\%$ dos problemas com o menor número de avaliações de função. O método de Newton manteve um desempenho superior aos demais para valores de $\tau \in [1, 2]$ e teve desempenho muito similar ao método de Halley para $\tau > 2$.

Observamos que usando um pouco menos que o dobro de vezes de nf do melhor algoritmo, mais especificamente para $\tau = 1,75$, os métodos Halley e Newton resolveram aproximadamente $71,73\%$ dos problemas. Em relação aos métodos Halley e Super-Halley, foram resolvidos aproximadamente $51,08\%$ dos problemas para $\tau = 1,32$. O método de Chebyshev se mostrou inferior aos demais para valores de $\tau \in [1, 1.58]$, alcançando um desempenho igual ou levemente superior ao método Super-Halley para valores de τ maiores do que 1.58.

Além do número de avaliações de função, o tempo computacional também é um fator que torna os métodos da Classe Chebyshev-Halley impraticáveis, devido ao tempo gasto para o cálculo do tensor e para resolver dois sistemas lineares de forma exata.

Através do gráfico de desempenho do tempo computacional, ilustrado na Figura 5.3, percebemos claramente que o método de Newton é o mais eficiente resolvendo aproximadamente $60,14\%$ dos problemas no menor tempo, enquanto que os métodos Halley, Chebyshev e Super-Halley resolveram, respectivamente, em torno de $4,7\%$, $2,8\%$ e $12,68\%$

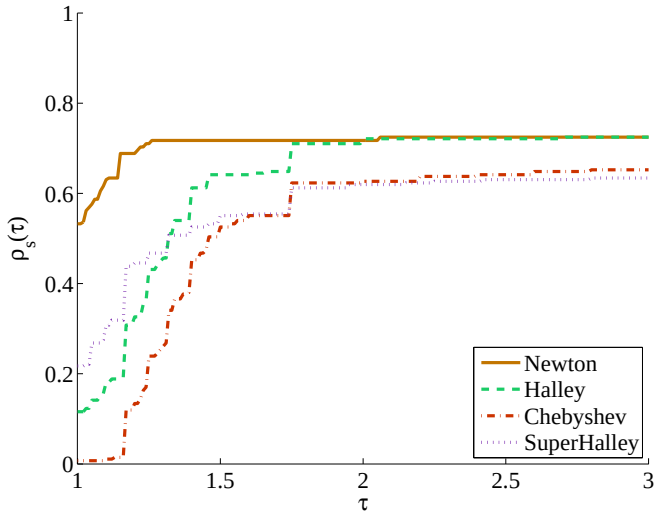

Figura 5.2: Gráfico de desempenho do número de avaliações de função dos métodos Newton, Halley, Chebyshev e Super-Halley.

dos problemas com o menor tempo computacional.

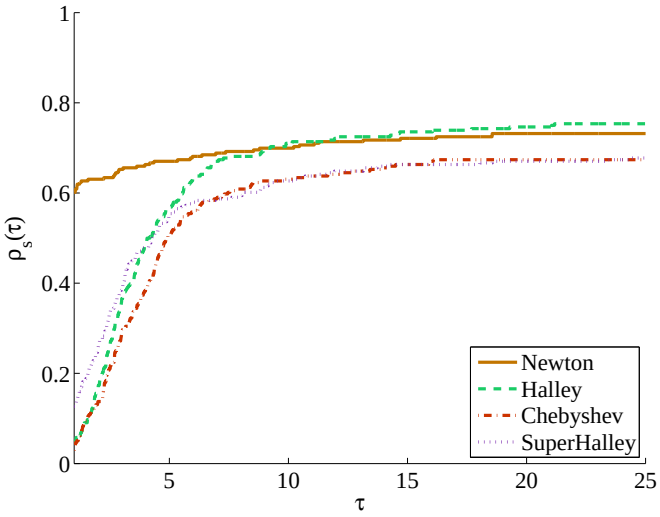

Figura 5.3: Gráfico de desempenho do tempo computacional dos métodos Newton, Halley, Chebyshev e Super-Halley.

5.1 Resultados Numéricos

A metodologia proposta nesta tese consiste em substituir o cálculo do produto $\mathcal{T}_F(x)s_1(x)$ por uma matriz $C(x)$, tal que

$$\|C(x)\| = O\Big(\|F(x)\| \Big) \tag{5.2}$$

para todo x suficientemente próximo de x^* na classe Chebyshev-Halley. Além disso, os dois sistemas lineares necessários para a obtenção do passo, podem ser resolvidos de maneira inexata.

Se os dois sistemas lineares forem resolvidos de forma exata, a hipótese (5.2) é suficiente para mostrar a convergência quadrática do Algoritmo 4.1, mas não a convergência cúbica. Neste caso, ao agregarmos a hipótese

$$\|\mathcal{T}_F(x)s_1(x) - C(x)\| = O\Big(\|F(x)\|^2 \Big), \tag{5.3}$$

então a convergência cúbica é garantida pelo Teorema 4.21.

O resultado do Lema 4.22, mostra que estas hipóteses não são impossíveis de serem verificadas, exibindo uma matriz que cumpre as hipóteses (5.2) e (5.3) para x suficientemente próximo de x^*, a saber

$$C(x) = \frac{J_F(x + hs_1(x)) - J_F(x)}{h} \tag{5.4}$$

para um dado $h > 0$.

Em nossos testes computacionais, utilizamos a matriz dada em (5.4) e o valor de h o mesmo usado por Bellavia e Morini [8] no método Newton-GMRES para a aproximação

$$J_F(x)v \approx \frac{F(x + hv) - F(x)}{h}$$

no processo de ortogonalização de Arnoldi, ou seja,

$$h = \sqrt{\epsilon}\frac{\left\|x^k\right\|_2}{\left\|s^k_{(1)}\right\|_2}, \tag{5.5}$$

onde ϵ é a precisão da máquina.

Inicialmente, nossa intenção é observar os efeitos que os métodos Halley, Chebyshev e Super-Halley sofrem ao substituir o produto $\mathcal{T}_F(x^k)s^k_{(1)}$ por $C(x^k)$ a cada iteração. Para isso, resolvemos os sistemas lineares de maneira exata usando o operador $\backslash$ do MATLAB como já comentado neste capítulo.

Para facilitar a notação, abreviamos como método HTF o método Halley livre de tensor, como método CTF o método Chebyshev livre de tensor e como SHTF o método Super-Halley livre de tensor.

Podemos observar na Tabela 5.2, que a robustez dos métodos Halley, Chebyshev e Super-Halley praticamente não sofreu alteração ao usar a estratégia livre de tensor com h dado em (5.5).

Na estratégia livre de tensor, praticamente trocamos o custo computacional do cálculo

	% de problemas resolvidos
Halley	76,44%
HTF	75,72%
Chebyshev	66,66%
CTF	65,94%
Super-Halley	67,39%
SHTF	66,30%

Tabela 5.2: Percentual dos problemas resolvidos indicando que a robustez dos métodos Halley, Chebyshev e Super-Halley praticamente não sofreu alteração ao usar a estratégia livre de tensor.

do tensor $\mathcal{T}_F(x^k)$ e do produto $\mathcal{T}_F(x^k)s^k_{(1)}$ por uma avaliação a mais da jacobiana, a saber, $J_F(x^k + hs^k_{(1)})$, já que $J_F(x^k)$ foi avaliada no primeiro sistema linear. Sendo assim, é de se esperar uma melhoria significativa no tempo computacional e no número de avaliações de função. O número de iterações não deve ter uma mudança significativa, já que o erro ao aproximar $\mathcal{T}_F(x^k)s^k_{(1)}$ por $C(x^k)$ dada por (5.4) é, por Taylor, $O(h)$. Depois de realizados os testes, podemos observar estes resultados nos gráficos de desempenho ilustrados nas Figuras 5.4, 5.5 e 5.6.

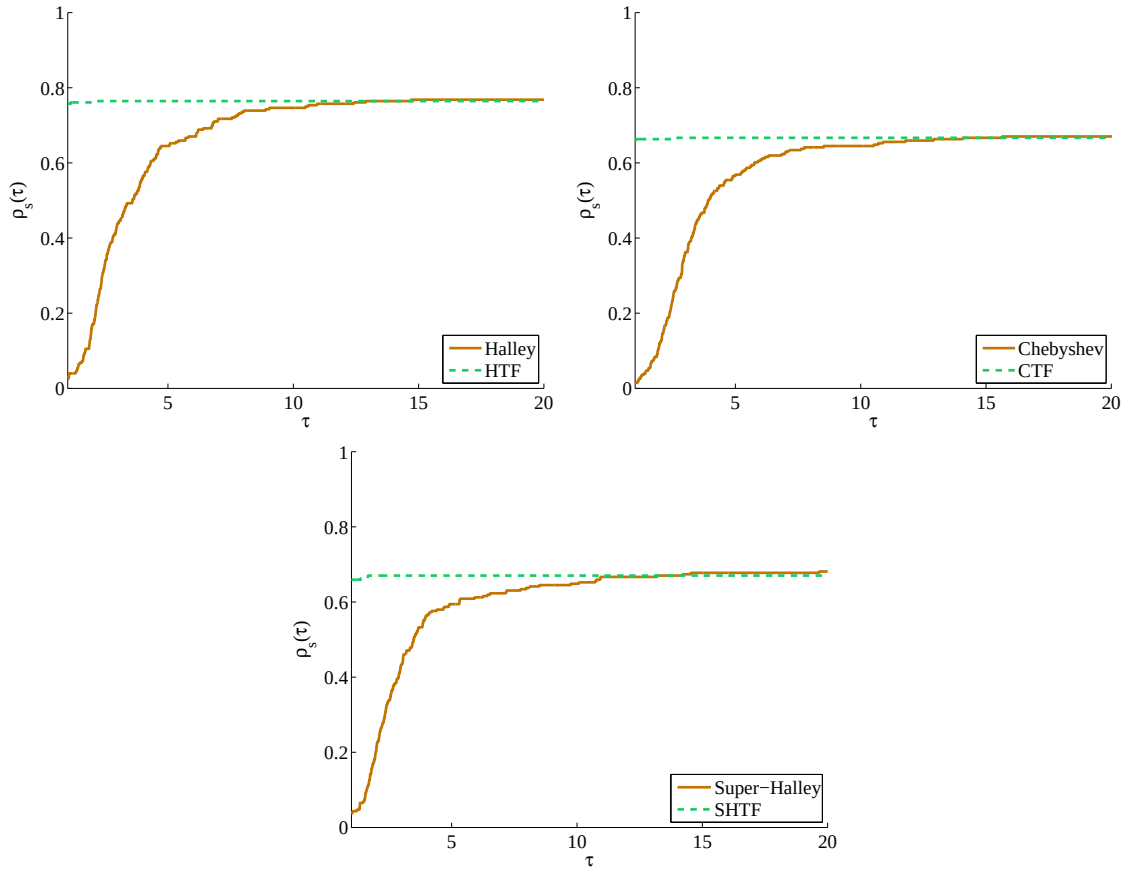

Figura 5.4: Gráficos de desempenho do tempo computacional dos métodos Halley e HTF, Chebyshev e CTF e Super-Halley e SHTF.

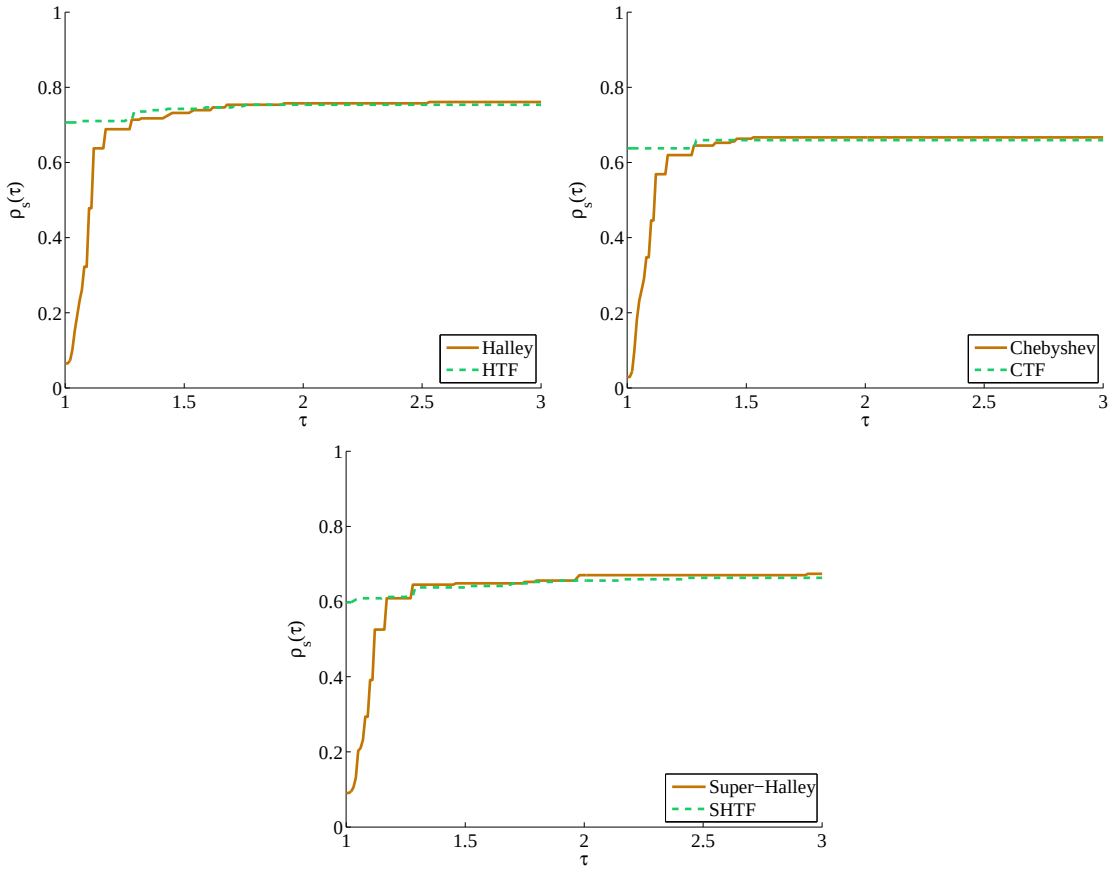

Figura 5.5: Gráficos de desempenho do número de avaliações de função dos métodos Halley e HTF, Chebyshev e CTF e Super-Halley e SHTF.

O método HTF comparado com o método de Halley, resolveu aproximadamente $75,72\%$ dos problemas no menor tempo (Figura 5.4). Observando os dados da Tabela 5.2 e a menos de erros de arrendondamento nos percentuais, isso significa que o tempo de execução, em praticamente todos os problemas por ele resolvidos, foi menor. Além disso, ele resolveu aproximadamente $70,65\%$ dos problemas com o menor número de avaliações de função (Figura 5.5) e o percentual de problemas resolvidos com o menor número de iterações é praticamente o mesmo que o método de Halley. Em relação ao número de iterações (Figura 5.6), o mesmo ocorre quando se compara os métodos CTF e Chebyshev. O método CTF resolveu aproximadamente $65,94\%$ dos problemas no menor tempo, significando também que o tempo de execução, em praticamente todos os problemas por ele resolvidos, foi menor, e $63,77\%$ dos problemas com o menor número de avaliações de função. Por fim, quando comparado com o método Super-Halley, o método SHTF resolveu aproximadamente $65,94\%$ dos problemas no menor tempo e $59,78\%$ dos problemas com o menor número de avaliações de função. No entanto, é visível uma alteração, não tão significativa, do número de iterações. O método Super-Halley resolveu aproximadamente $64,49\%$ dos problemas com o menor número de iterações, enquanto que o método SHTF

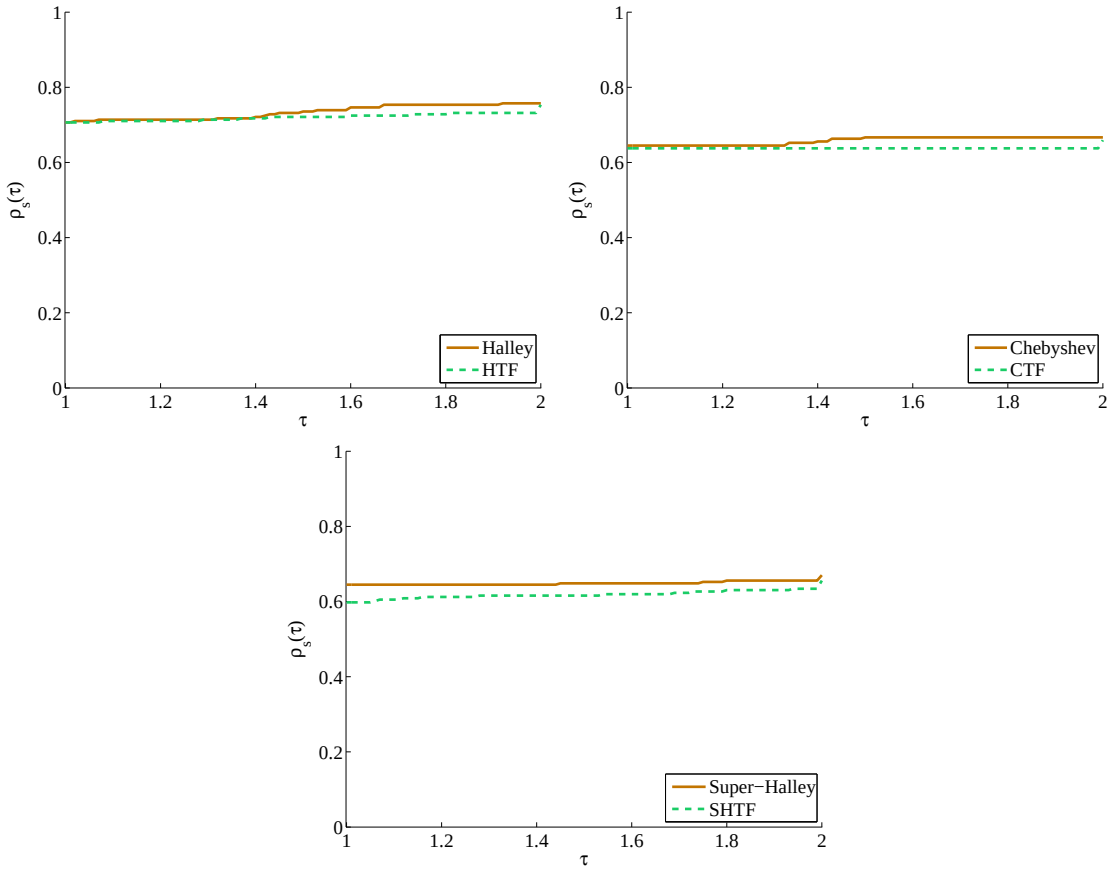

Figura 5.6: Gráficos de desempenho do número de iterações dos métodos Halley e HTF, Chebyshev e CTF e Super-Halley e SHTF.

resolveu aproximadamente $59,78\%$ dos problemas com o menor número de iterações.

Observando que a robustez não teve alterações significativas quando se compara os métodos com suas respectivas modificações, com a análise realizada até agora, podemos perceber uma melhoria significativa em todos os aspectos ao usar a matriz (5.4) na estratégia livre de tensor. Com isso atingimos o primeiro objetivo desta análise numérica.

A segunda estratégia é considerar a resolução dos sistemas lineares, necessários para a obtenção do passo, de maneira inexata, mantendo a estratégia livre de tensor. Particularmente, como comparamos inicialmente os métodos Halley, Chebyshev e Super-Halley com o método de Newton, agora em nossa análise apenas o segundo sistema linear será resolvido de maneira inexata, ou seja, obtemos $s^k_{(1)}$ e $s^k_{(2)}$ tais que

$$\begin{aligned} J_F(x^k)s^k_{(1)} &= -F(x^k) \\ \Big(J_F(x^k) + \alpha C(x^k)\Big)s^k_{(2)} &= -\frac{1}{2}C(x^k)s^k_{(1)} + r^k_{(2)} \end{aligned} \tag{5.6}$$

e

$$\left\|r_{(2)}^k\right\| \leq \eta_k^{(2)} \left\|\frac{1}{2}C(x^k)s_{(1)}^k\right\| \tag{5.7}$$

para algum $\eta_k^{(2)} \in (0,1)$. O passo é definido como sendo $s^k = s_{(1)}^k + s_{(2)}^k$.

Note que (5.6) corresponde à classe Chebyshev-Halley Inexata Modificada dada em (3.60) e (3.61) com $r_{(2)}^k = \tilde{r}_2^k$ e substituindo $\mathcal{T}_F(x^k)s_{(1)}^k$ pela matriz $C(x^k)$.

Na prova do Teorema 4.21, não exigimos que seja utilizado um determinado método para resolver um sistema linear de maneira inexata. Para nossos experimentos, propomos uma adaptação do Algoritmo 3.7 (Cálculo de $s_{(2)}^k$ e $\tilde{r}_2^k$) de Steihaug e Suleiman [73], baseado em ponto fixo para o cálculo de $s_{(2)}^k$. A vantagem dessa abordagem é que a decomposição da matriz $J_F(x^k)$ utilizada para a obtenção de $s_{(1)}^k$ poderá ser reutilizada para a obtenção de $s_{(2)}^k$, ou seja, não há necessidade de decompor a matriz $J_F(x^k)+\alpha C(x^k)$. Detalhes podem ser vistos na Seção 3.4.3. Optamos por usar a decomposição LU. A adaptação a qual nos referimos, consiste basicamente em substituir o cálculo $\mathcal{T}_F(x^k)s_{(1)}^k$ pela matriz $C(x^k)$ para todo $k \in \mathbb{N}$ no Algoritmo 3.7. Desta forma, propomos o Algoritmo 5.1.

Algoritmo 5.1: Cálculo de $s_{(2)}^k$ e $r_{(2)}^k$ - livre de tensor

---

Defina $A = J_F(x^k) + \alpha C(x^k)$, $b = -\dfrac{1}{2}C(x^k)s_{(1)}^k$.
Dados $w^0 = 0$ e $r^0 = b$.
PARA $l = 1, 2, \ldots$
  Defina z^{l-1} a solução do sistema $J_F(x^k)z^{l-1} = r^{l-1}$
  Atualize $w^l = w^{l-1} + z^{l-1}$
  Atualize $r^l = b - Aw^l$
FIM
$s_{(2)}^k = w^l$, $r_{(2)}^k = r^l$ e $j = l$.

É importante relembrar que o Algoritmo 3.7 foi utilizado para à obtenção de um passo s^k de modo que o critério do resíduo

$$\left\|\frac{1}{2}\mathcal{T}_F(x^k)s^k s^k + J_F(x^k)s^k + F(x^k)\right\| \leq \eta_k \left\|F(x^k)\right\| \tag{5.8}$$

fosse verificado para algum $\eta_k \in (0,1)$ e que $\left\|s^k\right\| = O\Big(\left\|F(x^k)\right\|\Big)$, cuja garantia de convergência foi estabelecida no Teorema 3.8.

Conjecturamos que ao usar o Algoritmo 5.1 para formar o passo $s^k = s_{(1)}^k + s_{(2)}^k$, o critério do resíduo (5.8) deve ser verificado para algum $\eta_k \in (0,1)$ e que $\left\|s^k\right\| = O\Big(\left\|F(x^k)\right\|\Big)$. Sendo assim, a convergência da sequência (x^k), tal que $x^{k+1} = x^k + s^k$, é garantida pelo Teorema 3.6.

Evidentemente, se $s^k_{(2)}$ é determinado pelo Algoritmo 5.1, não estaremos mais exigindo a condição do resíduo (5.7), mas estaremos exigindo que o passo s^k cumpra a condição do resíduo (5.8) para k suficientemente grande. No entanto, como apresentado no Capítulo 3, η_k não pode ser dado a priori, pois pode não existir um s^k que cumpra a condição (5.8) para o respectivo η_k dado, ou seja, a precisão do passo não pode ser controlada.

Em [73], foram realizados experimentos numéricos comparando apenas o número de iterações dos métodos Halley, Chebyshev, Super-Halley e Super-Halley Inexato Modificado usando $j = 1, 2, 3$ iterações internas do Algoritmo 3.7. O método Super-Halley foi o mais eficiente neste quesito e foi observado uma pequena diferença entre os métodos Super-Halley e Super-Halley Inexato Modificado com $j = 3$ iterações internas do Algoritmo 3.7. Por esse motivo, optamos por usar apenas $j = 3$ iterações internas do Algoritmo 5.1, não só no método Super-Halley Inexato livre de tensor, mas também em Halley e Chebyshev Inexatos livres de tensores. Salientamos que nenhuma estratégia para aproximar o tensor $\mathcal{T}_F(x^k)$ foi utilizada em [73].

Já o método GMRES para quando encontrar um $s^k_{(2)}$ que cumpra (5.7). No entanto, para cada iteração externa, o método GMRES pode gastar no máximo n iterações para encontrar tal $s^k_{(2)}$. Sendo assim, vamos limitar o método GMRES a realizar no máximo 10 iterações internas. Isto significa que a solução fornecida pelo GMRES pode não satisfazer o critério (5.7).

Consideramos em nossa implementação $\eta^{(2)}_0 = 0.01$ e o atualizamos como

$$\eta^{(2)}_k = \min\left\{\frac{1}{k+2}, \left\|F(x^k)\right\|_\infty\right\}$$

de modo que $\eta^{(2)}_k \in [10^{-8}, 10^{-2}]$. O ponto inicial foi $s^0_{(2)} = 0$.

Utilizamos uma modificação do algoritmo implementado por Kelley [48] que pode ser encontrado em `www.siam.org/books/kelley/fr16/matlabcode.php`, que utiliza o Algoritmo 3.3 (Método de Arnoldi com Gram-Schmidt modificado) incluindo uma estratégia de reortogonalização, caso seja detectado uma perda de ortogonalidade após obter um novo vetor v_{j+1} no Algoritmo de Arnoldi. O critério utilizado para detectar perda de ortogonalidade foi

$$\left\|(J_F(x^k) + \alpha C(x^k))v_j\right\| + 0.001\left\|v_{j+1}\right\| = \left\|(J_F(x^k) + \alpha C(x^k))v_j\right\|$$

e a estratégia de reortogonalização é embutida no Algoritmo 3.3 da seguinte maneira

1. Para $i = 1, \ldots, j$

   (a) Defina $h_{tmp} = \langle v_{j+1}, v_i\rangle$;

   (b) Faça $h_{i,j} = h_{i,j} + h_{tmp}$ e $v_{j+1} = v_{j+1} - h_{tmp}v_i$;

2. Redefina $h_{j+1,j} = \|v_{j+1}\|_2$ e $v_{j+1} = \dfrac{v_{j+1}}{\|v_{j+1}\|_2}$.

Novamente, para facilitar a notação, abreviamos por HTF-GMRES o método HTF onde apenas o segundo sistema linear é resolvido pelo método GMRES e por HTF-PONTO FIXO o método HTF onde o segundo sistema linear é resolvido pelo Algoritmo 5.1. As notações dos outros métodos seguem de maneira análoga.

Podemos ver na Tabela 5.3 os percentuais de problemas resolvidos pelos métodos Halley, Chebyshev e Super-Halley utilizando estratégia livre de tensor e onde o segundo sistema linear é resolvido por GMRES ou pelo Algoritmo 5.1.

	% de problemas resolvidos
HTF-GMRES	72,10%
HTF-PONTO FIXO	65,21%
CTF-GMRES	66,30%
CTF-PONTO FIXO	65,94%
SHTF-GMRES	65,21%
SHTF-PONTO FIXO	63,76%

Tabela 5.3: Percentual dos problemas resolvidos pelos métodos HTF-GMRES, HTF-PONTO FIXO, CTF-GMRES, CTF-PONTO FIXO, SHTF-GMRES e SHTF-PONTO FIXO

Em termos de robustez, percebemos uma ligeira vantagem ao utilizar a estratégia GMRES nos métodos CTF e SHTF. No método HTF essa diferença foi maior. Os métodos CTF-GMRES e CTF-PONTO FIXO resolveram quase o mesmo percentual de problemas que o método de Chebyshev. Além disso o método CTF-PONTO FIXO resolveu exatamente o mesmo percentual de problemas que o método CTF.

Na Figura 5.7, podemos ver claramente que os métodos Chebyshev e CTF-PONTO FIXO foram um pouco mais eficientes em termos de número de iterações que o método CTF-GMRES. Mais especificamente, ambos resolveram aproximadamente $61,59\%$ dos problemas com o menor número de iterações, enquanto que o método CTF-GMRES resolveu aproximadamente $57,24\%$ dos problemas com o menor número de iterações, quando comparados entre si. Já nas variantes dos métodos Halley e Super-Halley, o método GMRES foi mais eficiente em termos do número de iterações do que o Algoritmo 5.1. Os métodos HTF-GMRES e HTF-PONTO FIXO resolveram, respectivamente, em torno de $58,69\%$ e $52,53\%$ dos problemas com o menor número de iterações quando comparados com o método de Halley e os métodos SHTF-GMRES e SHTF-PONTO FIXO resolveram, respectivamente, em torno de $51,08\%$ e $42,39\%$ dos problemas com o menor número de iterações quando comparados com o método Super-Halley.

Na Figura 5.8, percebemos que o método CTF-PONTO FIXO foi também mais eficiente que o método CTF-GMRES no quesito número de avaliações de função. Quando

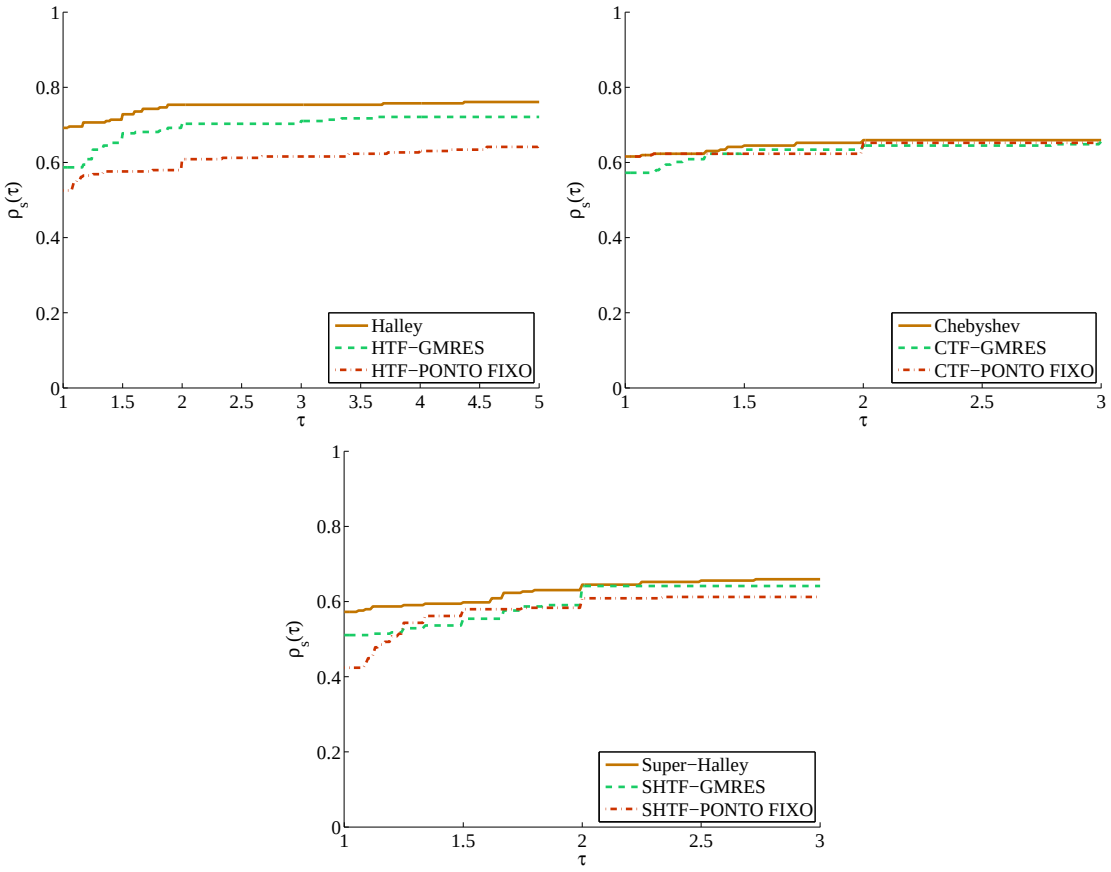

Figura 5.7: Gráficos de desempenho do número de iterações dos métodos Halley, HTF-GMRES e HTF-PONTO FIXO, Chebyshev, CTF-GMRES e CTF-PONTO FIXO e Super-Halley, SHTF-GMRES e SHTF-PONTO FIXO.

comparado com o método de Chebyshev, ele resolveu aproximadamente $61,59\%$ dos problemas com o menor número de avaliações de função enquanto que o método CTF-GMRES resolveu em torno de $57,25\%$. Nas variantes dos métodos Halley e Super-Halley, o método GMRES se mostrou mais eficiente em relação ao número de avaliações de função que o Algoritmo 5.1. Os métodos HTF-GMRES e HTF-PONTO FIXO resolveram, respectivamente, em torno de $58,69\%$ e $52,53\%$ dos problemas com o menor número de avaliações de função quando comparados com o método de Halley e os métodos SHTF-GMRES e SHTF-PONTO FIXO resolveram, respectivamente, em torno de $51,27\%$ e $42,18\%$ dos problemas com o menor número de avaliações de função quando comparados com o método Super-Halley.

Por fim, na Figura 5.9, podemos perceber que o Algoritmo 5.1 foi bem mais eficiente que o método GMRES em todos os métodos, no quesito tempo computacional. Quando comparado com o método de Halley, o método HTF-PONTO FIXO resolveu aproximadamente $56,15\%$ dos problemas no menor tempo enquanto que o método HTF-GMRES resolveu, aproximadamente, apenas $12,31\%$ dos problemas no menor tempo. Como já

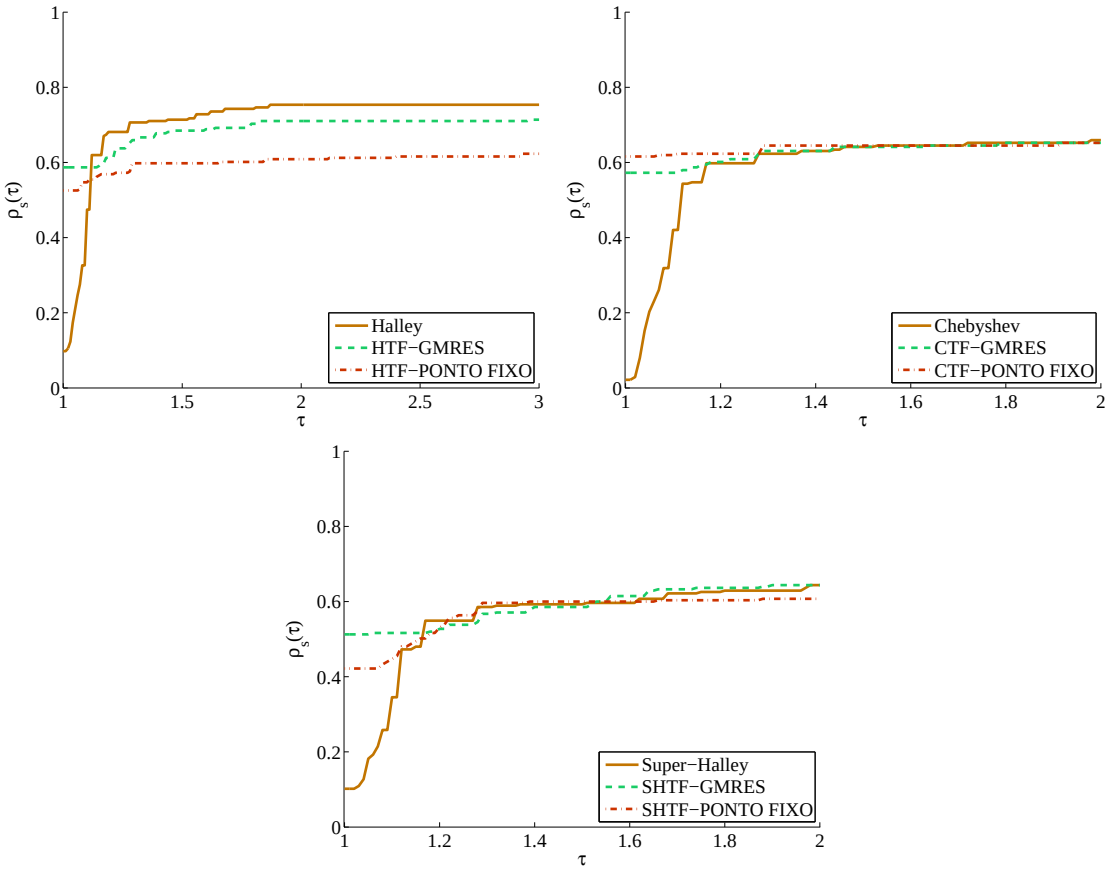

Figura 5.8: Gráficos de desempenho do número de avaliações de função dos métodos Halley, HTF-GMRES e HTF-PONTO FIXO, Chebyshev, CTF-GMRES e CTF-PONTO FIXO e Super-Halley, SHTF-GMRES e SHTF-PONTO FIXO.

destacado, a diferença de robustez entre os método HTF-GMRES e HTF-PONTO FIXO foi mais acentuada. Já os métodos CTF-PONTO FIXO e CTF-GMRES resolveram, respectivamente, em torno de $63,04\%$ e $5,43\%$ dos problemas no menor tempo quando comparado com o método Chebyshev e os métodos SHTF-PONTO FIXO e SHTF-GMRES resolveram, respectivamente, em torno de $56,88\%$ e $13,40\%$ dos problemas no menor tempo quando comparado com o método Super-Halley.

5.2 Conclusões dos Resultados Numéricos

Diante dos resultados numéricos obtidos na seção anterior, podemos perceber que o método de Halley foi o mais robusto entre os métodos de Newton, Chebyshev e Super-Halley, considerando os problemas testados. No entanto, comparando apenas os métodos analisados da classe Chebyshev-Halley, o método Super-Halley foi o mais eficiente em todos os quesitos analisados, ou seja, em número de iterações, número de avaliações de função e tempo computacional. O método de Chebyshev foi o menos eficiente em todos

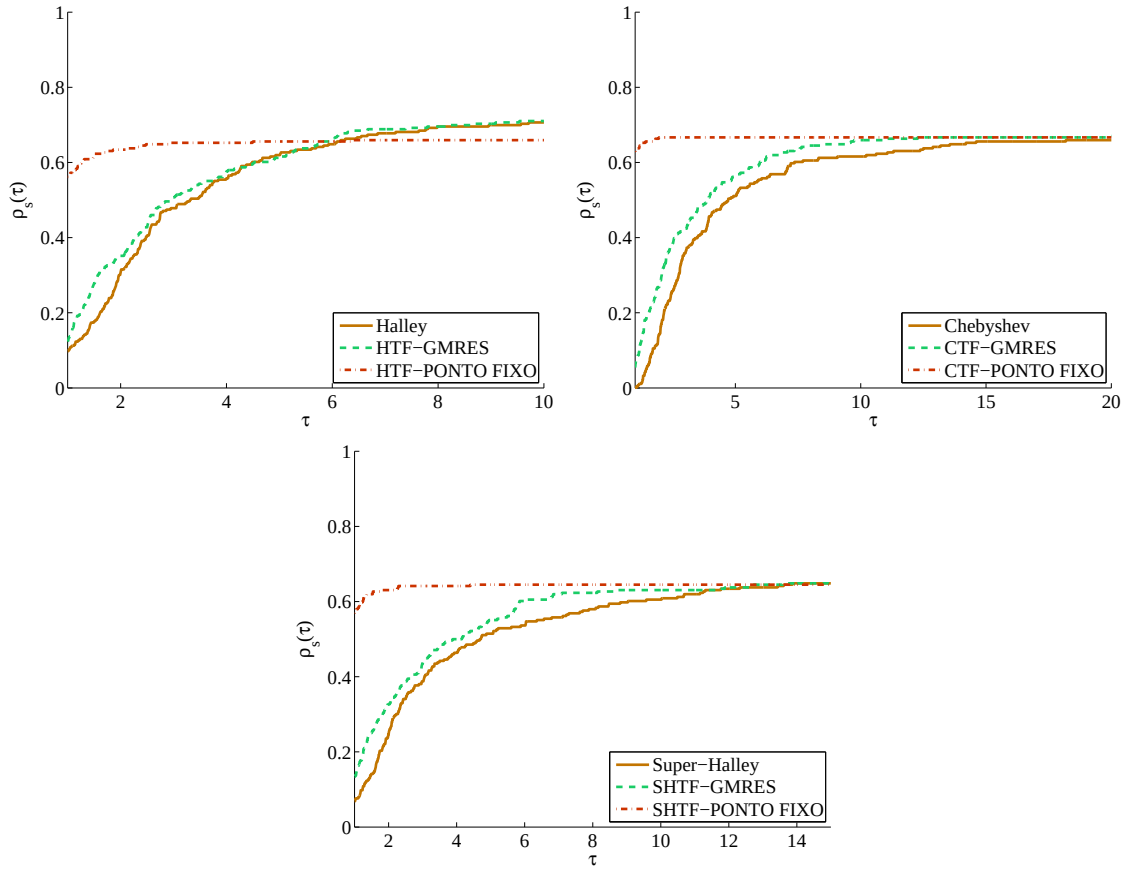

Figura 5.9: Gráficos de desempenho do tempo computacional dos métodos Halley, HTF-GMRES e HTF-PONTO FIXO, Chebyshev, CTF-GMRES e CTF-PONTO FIXO e Super-Halley, SHTF-GMRES e SHTF-PONTO FIXO.

esses quesitos, além de também perder em robustez.

Ao utilizar a estratégia livre de tensor, resolvendo os dois sistemas lineares de forma exata, percebemos uma certa semelhança na robustez, comparando os métodos e suas respectivas modificações, evidentemente. E ao mesmo tempo, o número de avaliações de função e tempo computacional foram significativamente melhores. Esperávamos, em contrapartida, uma redução na eficiência do número de iterações, mas apenas o método Super-Halley obteve tal desvantagem. Concluímos então, que a proposta livre de tensor foi, em geral, melhor que o método de Newton, no quesito número de iterações e também reduziu significativamente o tempo computacional e o número de avaliações de função dos métodos Halley, Chebyshev e Super-Halley, atingindo nossas expectativas.

Na segunda análise realizada, utilizamos dois métodos diferentes para resolver o segundo sistema linear de maneira inexata, o método GMRES e uma adaptação que propomos do Algoritmo 3.7, a saber, o Algoritmo 5.1. Em relação ao tempo computacional e do número de avaliações de função, a estratégia inexata livre de tensor, tanto usando o método GMRES ou o Algoritmo 5.1, foi mais eficiente, destacando o Algoritmo 5.1

quanto à eficiência do tempo computacional. Os métodos que utilizaram como subrotina o Algoritmo 5.1, apesar de terem sido menos robustos que aqueles que utilizam o método GMRES, foram significativamente mais eficientes que aqueles que usaram o GMRES. Em contrapartida, esses que utilizaram o método GMRES foram mais eficientes que aqueles que utilizaram o Algoritmo 5.1 nos quesitos número de iterações e no número de avaliações de função, com exceção do método CTF-GMRES.

Por fim, concluímos como consequência da redução do número de operações realizadas pelos métodos da Classe Chebyshev-Halley, principalmente no cálculo de derivadas de segunda ordem, que as modificações dos métodos Halley, Chebyshev e Super-Halley, propostas nesta tese, foram bastante contundentes, abrindo caminhos para pesquisas futuras.

Conclusões

Apresentamos nesta tese, dois novos resultados sobre a Classe Chebyshev-Halley. O primeiro é um resultado teórico. Introduzimos um novo raio de convergência para a Classe Chebyshev-Halley e comparamos com o raio utilizado na prova de convergência dada no livro *Numerische Lösung Nichtlinearer Gleichungen* [70] para os métodos Halley, Chebyshev e Super-Halley, através de alguns exemplos. Neste exemplos, observamos que o raio introduzido nesta tese é maior que o introduzido em [70] para os métodos Halley e Super-Halley. No entanto, ele é menor para o método de Chebyshev. Essa comparação sugere indícios de pesquisa futura, como por exemplo, estabelecer o raio ótimo de convergência para a Classe Chebyshev-Halley, ou para algum método em particular.

A segunda contribuição consiste em uma modificação da Classe Chebyshev-Halley. Esta modificação é justificada, pois o cálculo do tensor é caro computacionalmente. Além disso, dois sistemas lineares devem ser resolvidos de maneira exata. Pensando em estratégias *matrix-free* aplicadas no método de Newton inexato, introduzimos uma nova classe de métodos, chamada Classe Chebyshev-Halley Inexata livre de tensores, a qual não utiliza informação de derivadas de segunda ordem e os dois sistemas lineares necessários para a obtenção do passo, podem ser resolvidos de maneira inexata.

Concluímos na prova de convergência da Classe Chebyshev-Halley Inexata livre de tensores, que sob hipóteses razoáveis, além de convergirem localmente, os métodos dessa classe podem atingir taxa de convergência superlinear, quadrática, superquadrática e cúbica. Além disso, exibimos uma matriz que cumpre a hipótese exigida para garantir a convergência cúbica desta classe. Com essa matriz, realizamos testes computacionais, com o objetivo de observar se houve melhorias em número de iterações, número de avaliações de função e tempo computacional, em relação aos métodos Halley, Chebyshev e Super-Halley. Para isso, introduzimos uma modificação de um algoritmo proposto em [73], baseado em ponto fixo, para obter uma solução aproximada do segundo sistema linear, necessário para a obtenção do passo. Este algoritmo é vantajoso em termos de custo computacional, pois há necessidade de decompor apenas a matriz jacobiana, que já fora obtida na resolução do primeiro sistema linear. Foram executadas apenas três iterações deste algoritmo, ou seja, para cada iteração externa, três iterações internas foram executadas. Além disso, também

utilizamos o método GMRES, que é um método consagrado na literatura para resolver de maneira inexata um sistema linear, cuja solução pertence a um espaço de Krylov. Diante dos resultados numéricos obtidos, concluímos que a estratégia proposta nesta tese melhorou de maneira contundente esses métodos nos quesitos número de avaliações de função e tempo computacional, sendo que o número de iterações não teve alterações significativas.

Sugestões para Trabalhos Futuros

Utilizamos na Classe Chebyshev-Halley Inexata livre de tensores uma matriz $C(x)$ que cumpre a propriedade

$$\|C(x)\| = O\Big(\|F(x)\| \Big) \tag{5.9}$$

para x suficientemente próximo de uma solução x^*. Garantimos a convergência local da sequência (x^k) gerada por qualquer método desta classe. Além disso, mostramos que qualquer matriz do conjunto

$$\mathbf{C} = \{\mathcal{B}s_1(x) \mid \mathcal{B} \in U \subset \mathbb{R}^{n\times n\times n}(U \text{ limitado}) \text{ e } x \in B(x^*, \varepsilon)\} \tag{5.10}$$

cumpre (5.9). Dado $h > 0$, a matriz

$$C(x) = \frac{J_F(x + hs_1(x)) - J_F(x)}{h} \tag{5.11}$$

foi introduzida de modo a verificar a hipótese

$$\|\mathcal{T}_F(x)s_1(x) - C(x)\| = O\Big(\|F(x)\|^w \Big) \quad \text{para } w \in (1, 2] \tag{5.12}$$

exigida para garantir a taxa de convergência superquadrática, caso $w \in (1, 2)$, e a cúbica, caso $w = 2$, da sequência (x^k) gerada por qualquer método da Classe Chebyshev-Halley Inexata livre de tensores. Apesar de ser mais econômica computacionalmente, comparado com o cálculo do tensor $\mathcal{T}_F(x^k)$, ela exige um cálculo a mais de uma jacobiana, a saber, $J_F(x + hs_1(x))$, já que $J_F(x^k)$ é conhecida pelo primeiro sistema linear.

No sentido de evitar um cálculo a mais de uma jacobiana, podemos definir outras matrizes que cumpram pelo menos a hipótese (5.9). Como dito anteriormente, qualquer matriz do conjunto $\mathbf{C}$, definido em (5.10), verifica a hipótese (5.9). No entanto, gostaríamos que a hipótese (5.12) também fosse verificada pelo menos para algum $w \in (1, 2)$, garantindo assim uma taxa de convergência superquadrática.

A primeira ideia que tivemos em nossa pesquisa, foi utilizar estratégias semelhantes a dos métodos quase-Newton, ou seja, construir uma sequência $(\mathcal{B}_k)$ de tensores com alguma

propriedade e usar regras do tipo Broyden, BFGS, DFP entre outras. Mas preferimos deixar como trabalho futuro, um estudo sobre atualizações de tensores de modo a garantir, se possível, que a hipótese (5.12) seja verificada.

Em relação ao estudo sobre raio de convergência, pode ser muito difícil ou até mesmo impossível, estabelecer o raio ótimo de convergência da Classe Chebyshev-Halley. Neste sentido, deixamos também como trabalho futuro um estudo sobre o raio ótimo de convergência apenas dos métodos Halley, Chebyshev e Super-Halley.

Referências Bibliográficas

[1] G. Alefeld. On the convergence of Halley's method. *The American Mathematical Monthly*, 88(7):530–536, 1981.

[2] W. E. Arnoldi. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, 9:17–29, 1951.

[3] B. W. Bader. *Tensor-Krylov Methods for Solving Large-Scale Systems of Nonlinear Equations*. PhD thesis, University of Colorado, Boulder, Department of Computer Science, 2003.

[4] B. W. Bader. Tensor-Krylov methods for solving large-scale systems of nonlinear equations. *SIAM Journal on Numerical Analysis*, 43(3):1321–1347, 2006.

[5] B. W. Bader and T. G. Kolda. Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4):635–653, December 2006.

[6] B. W. Bader and T. G. Kolda. Efficient MATLAB computations with sparse and factored tensors. Technical Report SAND2006-7592, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, December 2006.

[7] B. W. Bader and R. B. Schnabel. On the performance of tensor methods for solving ill-conditioned problems. *SIAM Journal on Scientific Computing*, 29(6):2329–2351, October 2007.

[8] S. Bellavia and B. Morini. A globally convergent Newton-GMRES supspace method for system of nonlinear equations. *SIAM Journal on Scientific Computing*, 23:940–960, 2001.

[9] A. Bouaricha. *Solving large sparse systems of nonlinear equations and nonlinear least squares problems using tensor methods on sequential and parallel computers.* PhD thesis, University of Colorado, Boulder, Department of Computer Science, 1992.

[10] A. Bouaricha and R. B. Schnabel. Algorithm 768: TENSOLVE: A software package for solving systems of nonlinear equations and nonlinear least-squares problems using tensor methods. *ACM Transactions of Mathematical Software*, 23:174–195, 1997.

[11] J. P. Boyd. Finding the zeros of a univariate equation: proxy roothfinders, Chebyshev interpolation, and the companion matrix. *SIAM review*, 55(2):375–396, 2013.

[12] A. L. Cauchy. Sur la détermination approximative des racines d'une équation algébrique ou transcendante. Leçons sur le Calcul Differentiel, Buré fréres, Paris, 1829.

[13] P. L. Chebyshev. Complete collected works. *Izdateĺstvo Akademii Nauk SSR*, V, 1951.

[14] B. Chen, A. Petropulu, and L. De Lathauwer. Blind identification of convolutive MIMO systems with 3 sources and 2 sensors. *Applied Signal Processing*, 5:487–496, 2002. Special Issue Space-time Coding and Its Applications - Part II.

[15] D. Chen, I. K. Argyros, and Q. S. Qian. A local convergence theorem for the super-Halley method in a Banach space,. *Applied Mathematics Letters*, 7(5):49–52, 1994.

[16] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari. *Nonegative Matrix and Tensor Factorizations: Applications to Exploratory Multiway Data Analysis and Blind Source Separation*. John Wiley Sons, Ltd, 2009.

[17] F. U. Coelho and M. L. Lourenço. *Um Curso de Álgebra Linear*. Editora da Universidade de São Paulo, São Paulo, 2007.

[18] W. La Cruz, J. M. Martínez, and M. Raydan. Spectral residual method without gradient information for solving large-scale nonlinear systems of equations. *Mathematics of Computations*, 75:1429–1448, 2006.

[19] R. Dembo, S. C. Eisenstat, and T. Steihaug. Inexact newton methods. *SIAM Journal on Numerical Analysis*, 19(2):400–408, April 1982.

[20] N. Deng and H. Zhang. Theoretical efficiency of a new inexact method of tangent hyperbolas. *Optimization Methods and Software*, 19:247–265, 2004.

[21] J. E. Dennis and J. J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of Computation*, 28:546–560, 1974.

[22] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, 1983.

[23] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91:201–213, 2002.

[24] M. A. Dumett and J. P. Keener. The pyrite iron cycle catalyzed by *Acidithiobacillus ferroxidans*. *Journal of Mathematical Biology*, July 2013. DOI 10.1007/s00285-013-0708-0.

[25] P. Butzer e F. Jongmans. P. L. Chebyshev: A guide to his life and work. *Journal of Approximation Theory*, 96:111–138, 1999.

[26] G. P. Ehle and H. Schwetlick. Discretized Euler-Chebyshev multistep methods. *SIAM Journal on Numerical Analysis*, 13(3):432–447, 1976.

[27] S. C. Eisenstat and H. F. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing*, 17(1):16–32, January 1996.

[28] J. A. Ezquerro and M. A. Hernández. Differente acceleration procedures of Newton's method. *Novi Sad Journal of Mathematics*, 27(1):1–17, 1997.

[29] J. A. Ezquerro and M. A. Hernández. On a convex acceleration of Newton's method. *Journal of Optimization Theory and Applications*, 100(2):311–326, February 1999.

[30] D. Feng, P. D. Frank, and R. B. Schnabel. Local convergence analysis of tensor methods for nonlinear equations. Technical report, Departament of Computer Science, University of Colorado at Boulder, April 1992. CU-CS-591-92.

[31] D. Feng and T. H. Pulliam. Tensor-GMRES method for large systems of nonlinear equations. *SIAM Journal on Optimization*, 7:757–779, 1997.

[32] O. P. Ferreira. Local convergence of Newton's method in Banach space from the viewpoint of the majorant principle. *IMA Journal of Numerical Analysis*, 29:746–759, 2009.

[33] P. D. Frank. *Tensor methods for solving systems of nonlinear equations*. PhD thesis, Departament of Computer Science, University of Colorado at Boulder, 1984.

[34] G. H. Golub, T. G. Kolda, J. G. Nagy, and C. F. Van Loan. Workshop on tensor decompositions. American Institute of Mathematics, Palo Alto, California, 2004. http://www.aimath.org/WWN/tensordecomp/.

[35] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3 edition, 1996.

[36] A. Griewank, D. Juedes, and J. Utke. Algorithm 755: Adol-c: A package for the automatic differentation of algorithms written in c/c++. *ACM Transactions on Mathematical Software*, 22(2):135–167, 1996.

[37] G. Gundersen and T. Steihaug. On large scale unconstrained optimization problems and higher order methods. *Optimization Methods and Software*, 25(3):337–358, 2010.

[38] G. Gundersen and T. Steihaug. On diagonally structured problems in unconstrained optimization using an inexact super Halley method. *Journal of Computational and Applied Mathematics*, 236(15):3685–3695, September 2012.

[39] J. M. Gutiérrez and M. A. Hernández. An acceleration of Newton's method: super-Halley method. *Applied Mathematics and Computation*, 117(2-3):223–239, 2001.

[40] M. A. Hernandez. A note on Halley's method. *Numerische Mathematik*, 59(1):273–276, 1991.

[41] M. A. Hernandez. Newton-Raphson's method and convexity. *Zb. Rad. Prirod.-Mat. Fak. Ser.Mat.*, 22(1):159–166, 1993.

[42] M. A. Hernández and J. M. Gutiérrez. A family of Chebyshev-Halley type methods in Banach spaces. *Bulletin - Australian Mathematical Society*, 55:113–130, 1997.

[43] M. A. Hernández and M. A. Salanova. A family of Chebyshev-Halley type methods. *International Journal of Computer Mathematics*, 47:59–63, 1993.

[44] M. A. Hernández and M. A. Salanova. *La Convexidad en la Resolución de Ecuaciones Escalares no Lineales*. University de La Rioja, 2011.

[45] S. Hitotumatu. A method of successive approximation based on the expansion of second order. *Math. Japon.*, (7):31–50, 1962.

[46] M. Ishteva. *Numerical methods for the best low multilinear rank approximation of higher-order tensors*. PhD thesis, Katholieke Universiteit Leuven, Faculty of Engineering, Belgium, 2009.

[47] Jr. J. E. Dennis and J . J. Moré. Quasi-Newton methods, motivation and theory. *SIAM Review*, 19(1):46–89, January 1977.

[48] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. SIAM, Philadelphia, 1995.

[49] H. A. L. Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, (14):105–122, 2000.

[50] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September 2009.

[51] P. Comon. L. De Lathauwer. Workshop on tensor decompositions and applications. Luminy, Marseille, France, August-September 2005. http://www.etis.ensea.fr/wtda/.

[52] L. De Lathauwer, B. De Moor, and J. VandeWalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis Applications*, 21(4):1253–1278, 2000.

[53] E. L. Lima. *Análise no Espaço* $\mathbb{R}^n$. Editora Universidade de Brasília, São Paulo, 1970.

[54] E. L. Lima. *Curso de Análise*, volume 2. IMPA, Rio de Janeiro, Brasil, 1981.

[55] C. F. Van Loan. The ubiquitous kronecker product. *Journal of Computational and Applied Mathematics*, (123):85–100, 2000.

[56] L. Lukšan and J. Vlček. Sparse and partially separable test problems for unconstrained and equality constrained optimization. Technical Report V-767, Institute of Computer Science, Academy of Sciences of the Czech Replubic, Prague, 1999.

[57] J. M. Martínez and S. A. Santos. Métodos computacionais de otimização. $20.^0$ Colóquio Brasileiro de Matemática - IMPA, July 1995. In Portuguese.

[58] K. Meintjes and A. P. Morgan. Chemical equilibrium systems as numerical test problems. *ACM Transactions on Mathematical Software*, (16):143–151, 1990.

[59] M. A. Mertvecova. Analogue of the process of tangent hyperbolas for general functional equations. *Doklady Akademii Nauk SSSR(NS)*, 88:611–614, 1953.

[60] A. P. Morgan. *Solving Polynomial Systems Using Continuation for Scientific and Engineering Problems*. Prentice-Hall,Englewood Cliffs,NJ, 1987.

[61] M. I. Nečepuerenko. On Čebyšev's method for functional equations. *Uspehi Matematicheskih Nauk*, 9(2):163–170, 1954.

[62] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.

[63] F. A. Potra. On an iterative algorithm of order 1.839... for solving nonlinear operator equations. *Numerical Functional Analysis and Optimization*, 7(1):75–106, 1984.

[64] A. W. Robert and D. E. Varberg. *Convex functions*. New York, London: Academic Press, 1973.

[65] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2 edition, 2003.

[66] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, July 1986.

[67] G. S. Salehov. On the convergence of the process of tangent hyperbolas (in russian). *Doklady Akademii Nauk SSSR*, 82:525–528, 1952.

[68] R. B. Schnabel and P. D. Frank. Tensor methods for nonlinear equations. *SIAM Journal on Numerical Analysis*, 21:815–843, 1984.

[69] R. B. Schnabel and P. D. Frank. Solving systems of nonlinear equations by tensor methods. In A. Iserles and M. J. D Powell, editors, *The State of the Art in Numerical Analysis*, pages 245–271. Clarendon Press, Oxford, 1987.

[70] H. Schwetlick. *Numerische Lösung Nichtlinearer Gleichungen*. R. Oldenbourg Verlag, München-Wien, 1979.

[71] C. P. Serra and E. W. Karas. *Fractais Gerados por Sistemas Dinâmicos Complexos*. Ed. Champagnat, 1997.

[72] A. Smilde, R. Bro, and P. Geladi. *Multi-Way Analysis: Applications in the Chemical Sciences*. Wiley, 2004.

[73] T. Steihaug and S. Suleiman. Rate of convergence of higher order methods. *Applied Numerical Mathematics*, 67:230–242, May 2013.

[74] G. W. Stewart. *Introduction to Matrix Computations*. Academic Press, New York, 1973.

[75] J. F. Traub. *Iterative methods for the solution of equations*. Chelsea Publishing Company, 1982.

[76] S. Y. Ul'm. Iteration methods with divided differences of the second order. *Soviet Mathematics Doklady*, 5:1187–1190, 1964.

[77] H. F. Walker. Implementation of the GMRES method using Householder transformations. *SIAM Journal on Scientific Computing*, 9(1):152–163, January 1988.

[78] T. Yamamoto. Historical developments in convergence analysis for Newton's and Newton-like methods. *Journal of Computational and Applied Mathematics*, 124:1–23, 2000.

[79] G. F. Yan and X. Tian. An inexact Halley's method. *Journal of Beijing Institute of Technology*, 14:340–343, 2005.

Apêndice A

Os métodos Halley, Chebyshev e Super-Halley unidimensionais, foram introduzidos no Capítulo 2 através do grau de convexidade logarítmica de f, a saber,

$$\ell_f(x) = \frac{f(x)f''(x)}{f'(x)^2}.$$

Neste apêndice vamos introduzir melhor este conceito. As principais referências bibliográficas vistas sobre este assunto foram [40, 41, 44, 64].

A ideia principal para medir, de certa forma, a convexidade de uma função é calcular o número de vezes que devemos aplicar um operador côncavo em uma função convexa com derivada segunda estritamente positiva, até obter uma função que não seja convexa. Em particular, o operador logaritmo pode ser aplicado.

Definição A.1 *Considere $I \subset \mathbb{R}$ um intervalo. Dizemos que uma função $f : I \to (0, \infty)$ é log-convexa em I, quando a função $h = log(f)$ for convexa em I.*

De maneira equivalente[1], uma função $f : I \subset \mathbb{R} \to (0, \infty)$ é log-convexa quando

$$f[(1-\lambda)x + \lambda y] \leq f(x)^{1-\lambda} f(y)^{\lambda} \tag{A.1}$$

para todo $x, y \in I$ e para todo $\lambda \in (0, 1)$. Um exemplo simples de função log-convexa é $f : \mathbb{R} \to \mathbb{R}$ dada por $f(x) = e^{x^2}$.

É fácil ver que se $f \in \mathcal{C}^2$ é log-convexa em I, então f é convexa em I. De fato, como

$$0 \leq h''(x) = \frac{f''(x)f(x) - f'(x)^2}{f(x)^2} \tag{A.2}$$

para todo $x \in I$, observando que, por definição, $f(x) > 0$, temos que $f''(x) \geq 0$ para todo $x \in I$. A recíproca não é verdadeira, basta considerar $f : \mathbb{R} \to \mathbb{R}$ dada por $f(x) = e^x - 1$.

Para os nossos propósitos, vamos considerar um intervalo aberto $I \subset \mathbb{R}$ e uma função

[1] Essa equivalência é de fácil verificação.

$f : I \subset \mathbb{R} \to (0, \infty)$ tal que $f \in \mathcal{C}^2$ e que

$$f''(x) > 0 \quad \text{para todo } x \in I, \tag{A.3}$$

ou seja, consideramos as funções estritamente convexas que cumprem $(A.3)$. Considere também $\bar{x}$ um ponto no domínio de f. Queremos analisar a convexidade de f em uma vizinhança $\bar{I} \subset I$ de $\bar{x}$. Isto significa que consideraremos a restrição de f ao intervalo $\bar{I}$.

Para simplificar os cálculos, suponha em todo este apêndice que $f(\bar{x}) = 1$, exceto quando mencionado o contrário. Definindo

$$\mathcal{C}^2_{(x,r)}(I) = \left\{ g \in \mathcal{C}^2(I) \mid g(x) = r \right\},$$

considere uma sequência de funções $(F_n) \subset \mathcal{C}^2_{(\bar{x},0)}(\bar{I})$ tal que

$$\begin{aligned} F_1(x) =& \log f(x) \\ F_{n+1}(x) =& \log[1 + F_n(x)]. \end{aligned} \tag{A.4}$$

Teorema A.2 *A sequência (F_n) está bem definida e cumpre $F_n(x) \geq 0$ para todo $x \in \bar{I}$ se, e somente se, $f(x) \geq 1$ para todo $x \in \bar{I}$.*

Prova. Suponha por absurdo que exista $\hat{x} \in \bar{I}$ tal que $f(\hat{x}) < 1$. Desta forma, segue que $F_1(\hat{x}) < 0$. Queremos mostrar que existe um $n \in \mathbb{N}$ tal que $F_n(\hat{x}) \leq -1$. Caso $F_1(\hat{x}) \leq -1$, não há o que provar. Considere então $-1 < F_1(\hat{x}) < 0$. Neste caso, $F_2(\hat{x})$ é bem definido e vale $F_2(\hat{x}) < 0$. Analogamente, considere o caso em que $-1 < F_2(\hat{x}) < 0$. A ideia é mostrar que é um absurdo considerar $-1 < F_n(\hat{x}) < 0$ para todo $n \in \mathbb{N}$. Para isso, considere $g : (0, \infty) \to \mathbb{R}$ tal que $g(x) = \log x$ e sua aproximação linear em torno de $x = 1$, a saber, $h(x) = x - 1$. Como g é uma função côncava, segue que $g(x) < h(x)$ para todo $x \in \mathbb{R} - \{1\}$. Afirmamos que $(F_n(\hat{x}))$ é uma sequência estritamente decrescente. De fato,

$$F_{n+1}(\hat{x}) - F_n(\hat{x}) = g(1 + F_n(\hat{x})) - F_n(\hat{x}) < h(1 + F_n(\hat{x})) - F_n(\hat{x}) = 0.$$

Disto e do fato que $F_n(\hat{x}) \in (-1, 0)$ para todo $n \in \mathbb{N}$, segue que a sequência $(F_n(\hat{x}))$ é convergente. Digamos que $F_n(\hat{x}) \to a$, onde $a \in [-1, 0)$. Assim, $F_{n+1}(\hat{x}) \to a$. Por outro lado, $F_{n+1}(\hat{x}) \to \log(1 + a)$. Logo $a = 0$. Absurdo!

Para mostrar a recíproca, note que para todo $x \in \bar{I}$ que cumpre $f(x) \geq 1$, temos $F_1(x) \geq 0$. Por indução, suponha que $F_n(x)$ é bem definido e que $F_n(x) \geq 0$ para todo $x \in \bar{I}$. Assim, temos que $1 + F_n(x) \geq 1$. Logo $F_{n+1}(x)$ está bem definido e $F_{n+1}(x) \geq 0$.

□

O Teorema A.2 basicamente garante a boa definição da sequência de funções (F_n)

quando $\bar{x}$ for minimizador de f.

Definição A.3 *Dizemos que f é n-logaritmicamente convexa em I quando $F_n''(x) > 0$ para todo $x \in I$. Dizemos que f é infinitamente logaritmicamente convexa quando para todo $n \in \mathbb{N}$, existir um intervalo I_n tal que f seja n-logaritmicamente convexa em I_n.*

Teorema A.4 *Se a função $F_n : I_n \subset \mathbb{R} \to \mathbb{R}$ é bem definida, então*

(i) *$F_j'(\bar{x}) = f'(\bar{x})$ e $F_j''(\bar{x}) = f''(\bar{x}) - jf'(\bar{x})^2$ para todo $j = 1, \ldots, n$.*

(ii) *Fixado $j = 1, \ldots, n$, f é j-logaritmicamente convexa em uma vizinhança $\bar{I}_j \subset I$ de $\bar{x}$ se, e somente se, $f''(\bar{x}) > jf'(\bar{x})^2$.*

Prova. A prova é baseada em [40, Teorema 1]. Inicialmente, note que F_j está bem definida para todo $j = 1, \ldots, n$, pois como $F_{n-1}(x) > -1$ para todo $x \in I_n$, segue que $F_{n-2}(x) > -1 + e^{-1} > -1$. Desta maneira, é fácil ver que $F_j(x) > -1$ para todo $j = 1, \ldots, n-1$ e para todo $x \in I_n$. Além disso,

$$F_j(\bar{x}) = 0 \quad \text{para todo } j = 1, \ldots, n. \tag{A.5}$$

É óbvio que (i) vale para $j = 1$, pois $f(\bar{x}) = 1$. Como

$$F_j'(x) = \frac{F_{j-1}'(x)}{1 + F_{j-1}(x)}$$

para todo $j = 2, \ldots, n$, da igualdade $F_1'(\bar{x}) = f'(\bar{x})$ e de $(A.5)$, segue o primeiro resultado do item (i). Por outro lado, observe que

$$F_j''(x) = \frac{F_{j-1}''(x)[1 + F_{j-1}(x)] - [F_{j-1}'(x)]^2}{[1 + F_{j-1}(x)]^2}$$

para todo $j = 2, \ldots, n$. Da relação acima e da igualdade $F_1''(\bar{x}) = f''(\bar{x}) - f'(\bar{x})^2$, segue o segundo resultado do item (i). No item (ii), se f é j-logaritmicamente convexa em uma vizinhança $\bar{I}_j$ de $\bar{x}$, então $F_j''(x) > 0$ para todo $x \in \bar{I}_j$. Em particular, $F_j''(\bar{x}) > 0$. Logo, pelo item (i), $f''(\bar{x}) > jf'(\bar{x})^2$. Para mostrar a recíproca, considere $j = 1, \ldots, n$. Como $F_j''(\bar{x}) = f''(\bar{x}) - jf'(\bar{x})^2 > 0$, existe uma vizinhança $\bar{I}_j \subset I$ de $\bar{x}$ tal que $F_j''(x) > 0$ para todo $x \in \bar{I}_j$. □

Definição A.5 *Definimos o grau de convexidade logarítmica de f em $\bar{x}$, com $f(\bar{x}) = 1$, como*

$$\ell_f(\bar{x}) = \frac{f''(\bar{x})}{f'(\bar{x})^2}$$

quando $\bar{x}$ não for minimizador de f. Caso contrário, fazemos $\ell_f(\bar{x}) = +\infty$.

Pelo item (i) do Teorema A.4, é suficiente mostrar que

$$\ell_f(\bar{x}) < n$$

para que F_n seja côncava em uma vizinhança $\bar{I}_n \subset I$ de $\bar{x}$. A grosso modo, o grau de convexidade logarítmica de f em $\bar{x}$ com $f(\bar{x}) = 1$, é uma estimativa do número de vezes que é necessário a aplicação do operador logaritmo até obter uma função F_n que não seja convexa qualquer que seja a vizinhança de $\bar{x}$.

Para exemplificar, considere

$$f(x) = x - 3\cos(x-1) + 3. \tag{A.6}$$

Observe que para $\bar{x} = 1$, temos $f(\bar{x}) = 1$. Além disso, $f'(1) = 1$ e $f''(1) = 3$. Isto significa que f é estritamente convexa em uma vizinhança $\bar{I}$ de $\bar{x} = 1$.

Pelo item (i) do Teorema A.4, temos que

$$\begin{aligned} F_1''(\bar{x}) &= 2 > 0 \\ F_2''(\bar{x}) &= 1 > 0 \\ F_3''(\bar{x}) &= 0. \end{aligned}$$

Isto significa que aplicamos 3 vezes o operador logaritmo até obtermos uma função F_n com derivada segunda não positiva. Com um pouco mais de cálculo, é possível verificar que a função F_3 não é convexa em qualquer vizinhança de $\bar{x}$. Por outro lado, observe que

$$\ell_f(\bar{x}) = \frac{f''(1)}{f'(1)^2} = 3.$$

Obviamente, $\ell_f(\bar{x})$ pode não ser um número natural. Denote $E(r)$ a parte inteira do número real r.

Teorema A.6 **(i)** *caso* $r = \ell_f(\bar{x}) \notin \mathbb{N}$.

- **(a)** *para* $p = E(r) \in \mathbb{N}$, *temos que* f *é* p*-logaritmicamente convexa em uma vizinhança de* $\bar{x}$ *e não é* $(p+1)$*-logaritmicamente convexa qualquer que seja a vizinhança de* $\bar{x}$.
- **(b)** f *é* n*-logaritmicamente convexa em uma vizinhança de* $\bar{x}$ *para* $n \leq E(r)$.

(ii) *caso* $r = \ell_f(\bar{x}) \in \mathbb{N}$.

- **(a)** f *é* $(r-1)$*-logaritmicamente-convexa em uma vizinhança de* $\bar{x}$ *e* f *não é* $(r+1)$*-logaritmicamente convexa qualquer que seja a vizinhança de* $\bar{x}$.

(b) *f é r-logaritmicamente convexa em uma vizinhança de $\bar{x}$ se, e somente se, existe $k \in \mathbb{N}$ par tal que*

$$F_{r-1}^{(t)}(\bar{x}) = f'(\bar{x})^t, \quad 2 \leq t \leq k-1 \quad e \quad F_{r-1}^{(k)}(\bar{x}) > f'(\bar{x})^k$$

Prova. [44, Teorema 1.6]. □

Considerando ainda f dada em $(A.6)$, observamos que $F_2''(1) = f'(1)^2 = 1$ e $-11 = F_2'''(1) \neq f'(1)^3$. Pelo item (ii) do Teorema A.6, temos que f não é 3-logaritmicamente convexa qualquer que seja a vizinhança de $\bar{x} = 1$.

Agora, vamos definir o grau de convexidade logarítmica de uma função em um ponto qualquer. Para isso, vamos considerar um intervalo aberto $I \subset \mathbb{R}$ e uma função $f : I \subset \mathbb{R} \to (0, \infty)$ tal que $f \in \mathcal{C}^2$ e que

$$f''(x) > 0 \quad \text{para todo } x \in I,$$

e $\bar{x} \in I$ arbitrário. Defina

$$\ell_f(\bar{x}) = \ell_{f^*}(\bar{x}), \quad \text{onde} \quad f^*(x) = \frac{f(x)}{f(\bar{x})}. \tag{A.7}$$

Note que $f^*(\bar{x}) = 1$. Então pela Definição A.5 e por $(A.7)$, temos que o grau de convexidade de f em um ponto $\bar{x}$ é dado por

$$\ell_f(\bar{x}) = \frac{f(\bar{x})f''(\bar{x})}{f'(\bar{x})^2}.$$

Propriedades e outros detalhes podem ser encontrados em [44].

Apêndice B

Apresentamos neste apêndice os problemas compilados em La Cruz, Martínez e Raydan [18] e os da Seção 4 de Lukšan e Vlček [56] utilizados nesta tese, bem como os pontos iniciais adotados e a dimensão do problema. Denotamos como $x^0_{\text{padrão}}$ como o ponto inicial adotado nesses trabalhos.

Para os números inteiros positivos k e l, usamos a notação $\text{div}(k,l)$ para divisão inteira, isto é, o máximo inteiro não maior que k/l, e $\text{mod}(k,l) = l(k/l - \text{div}(k,l))$.

1. Countercurrent reactors problem 1 (modified)

$$
\begin{aligned}
f_k(x) &= \alpha - (1-\alpha)x_{k+2} - x_k(1+4x_{k+1}), \quad k=1\\
f_k(x) &= -(2-\alpha)x_{k+2} - x_k(1+4x_{k-1}), \quad k=2\\
f_k(x) &= \alpha x_{k-2} - (1-\alpha)x_{k+2} - x_k(1+4x_{k+1}), \quad \text{mod}(k,2)=1,\ 2<k<n-1\\
f_k(x) &= \alpha x_{k-2} - (2-\alpha)x_{k+2} - x_k(1+4x_{k-1}), \quad \text{mod}(k,2)=0,\ 2<k<n-1\\
f_k(x) &= \alpha x_{k-2} - x_k(1+4x_{k+1}), \quad k=n-1\\
f_k(x) &= \alpha x_{k-2} - (2-\alpha) - x_k(1+4x_{k-1}), \quad k=n\\
\alpha &= 0.5.
\end{aligned}
$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}}$ tal que

$$
\begin{aligned}
x^0_{\text{padrão}}(i) &= 0.1, \quad \text{mod}(i,8)=1\\
x^0_{\text{padrão}}(i) &= 0.2, \quad \text{mod}(i,8)=2 \text{ ou } \text{mod}(i,8)=0\\
x^0_{\text{padrão}}(i) &= 0.3, \quad \text{mod}(i,8)=3 \text{ ou } \text{mod}(i,8)=7\\
x^0_{\text{padrão}}(i) &= 0.4, \quad \text{mod}(i,8)=4 \text{ ou } \text{mod}(i,8)=6\\
x^0_{\text{padrão}}(i) &= 0.5, \quad \text{mod}(i,8)=5,
\end{aligned}
$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n=30$ e $n=50$.

2. Countercurrent reactors problem 2 (modified)

$$\begin{aligned}
f_k(x) &= x_1 - (1-x_1)x_{k+2} - \alpha(1+4x_{k+1}), \quad k=1\\
f_k(x) &= -(1-x_1)x_{k+2} - \alpha(1+4x_k), \quad k=2\\
f_k(x) &= \alpha x_1 - (1-x_1)x_{k+2} - x_k(1+4x_{k-1}), \quad k=3\\
f_k(x) &= x_1x_{k-2} + (1-x_1)x_{k+2} - x_k(1+4x_{k-1}), \quad 3<k<n-1\\
f_k(x) &= x_1x_{k-2} + x_k(1+4x_{k-1}), \quad k=n-1\\
f_k(x) &= x_1x_{k-2} - (1-x_1) - x_k(1+4x_{k-1}), \quad k=n\\
\alpha &= 0.414214.
\end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}}$ tal que

$$\begin{aligned}
x^0_{\text{padrão}}(i) &= 0.1, \quad \text{mod}(i,8)=1\\
x^0_{\text{padrão}}(i) &= 0.2, \quad \text{mod}(i,8)=2\\
x^0_{\text{padrão}}(i) &= 0.3, \quad \text{mod}(i,8)=3\\
x^0_{\text{padrão}}(i) &= 0.4, \quad \text{mod}(i,8)=4\\
x^0_{\text{padrão}}(i) &= 0.5, \quad \text{mod}(i,8)=5\\
x^0_{\text{padrão}}(i) &= 0.4, \quad \text{mod}(i,8)=6\\
x^0_{\text{padrão}}(i) &= 0.3, \quad \text{mod}(i,8)=7\\
x^0_{\text{padrão}}(i) &= 0.2, \quad \text{mod}(i,8)=0,
\end{aligned}$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n=30$ e $n=50$.

3. Trigonometric system

$$\begin{aligned}
f_k(x) &= 5 - (l+1)(1-\cos(x_k)) - \sin(x_k) - \sum_{j=5l+1}^{5l+5}\cos(x_j)\\
l &= \text{div}(k-1,5).
\end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = \left(\frac{1}{n},\ldots,\frac{1}{n}\right)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n=30$ e $n=50$.

4. Trigonometric - exponential system (trigexp 1)

$$\begin{aligned} f_k(x) &= 3x_k^3 + 2x_{k+1} - 5 + \sin(x_k - x_{k+1})\sin(x_k + x_{k+1}), \quad k = 1 \\ f_k(x) &= 3x_k^3 + 2x_{k+1} - 5 + \sin(x_k - x_{k+1})\sin(x_k + x_{k+1}) \\ &+ 4x_k - x_{k-1}\exp(x_{k-1} - x_k) - 3, \quad 1 < k < n \\ f_k(x) &= 4x_k - x_{k-1}\exp(x_{k-1} - x_k) - 3, \quad k = n. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (0, \ldots, 0)^T$, $x^0 = (1, \ldots, 1)^T$ e $x^0 = (2, \ldots, 2)^T$ com dimensões $n = 30$ e $n = 50$.

5. Trigonometric - exponential system (trigexp 2)

$$\begin{aligned} f_k(x) &= 3(x_k - x_{k+2})^3 - 5 + 2x_{k+1} \\ &+ \sin(x_k - x_{k+1} - x_{k+2})\sin(x_k + x_{k+1} - x_{k+2}), \quad \text{mod}(k, 2) = 1,\ k = 1 \\ f_k(x) &= -6(x_{k-2} - x_k)^3 + 10 - 4x_{k-1} \\ &- 2\sin(x_{k-2} - x_{k-1} - x_k)\sin(x_{k-2} + x_{k-1} - x_k) \\ &+ 3(x_k - x_{k+2})^3 - 5 + 2x_{k+1} \\ &+ \sin(x_k - x_{k+1} - x_{k+2})\sin(x_k + x_{k+1} - x_{k+2}), \quad \text{mod}(k, 2) = 1,\ 1 < k < n \\ f_k(x) &= -6(x_{k-2} - x_k)^3 + 10 - 4x_{k-1} \\ &- 2\sin(x_{k-2} - x_{k-1} - x_k)\sin(x_{k-2} + x_{k-1} - x_k), \quad \text{mod}(k, 2) = 1,\ k = n \\ f_k(x) &= 4x_k - (x_{k-1} - x_{k+1})\exp(x_{k-1} - x_k - x_{k+1}) - 3, \quad \text{mod}(k, 2) = 0. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1, \ldots, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 27$ e $n = 49$.

6. Singular Broyden problem

$$\begin{aligned} f_k(x) &= ((3 - 2x_k)x_k - 2x_{k+1} + 1)^2, \quad k = 1 \\ f_k(x) &= ((3 - 2x_k)x_k - x_{k-1} - 2x_{k+1} + 1)^2, \quad 1 < k < n \\ f_k(x) &= ((3 - 2x_k)x_k - x_{k-1} + 1)^2, \quad k = n. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (-1, \ldots, -1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

7. Tridiagonal system

$$\begin{aligned} f_k(x) &= 4(x_k - x_{k+1}^2), \quad k = 1 \\ f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) + 4(x_k - x_{k+1}^2), \quad 1 < k < n \\ f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k), \quad k = n. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (12, \ldots, 12)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

8. Five-diagonal system

$$\begin{aligned} f_k(x) &= 4(x_k - x_{k+1}^2) + x_{k+1} - x_{k+2}^2, \quad k = 1 \\ f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) \\ &+ 4(x_k - x_{k+1}^2) + x_{k+1} - x_{k+2}^2, \quad k = 2 \\ f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) \\ &+ 4(x_k - x_{k+1}^2) + x_{k-1}^2 - x_{k-2} + x_{k+1} - x_{k+2}^2, \quad 2 < k < n - 1 \\ f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) \\ &+ 4(x_k - x_{k+1}^2) + x_{k-1}^2 - x_{k-2}, \quad k = n - 1 \\ f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) + x_{k-1}^2 - x_{k-2}, \quad k = n. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (-2, \ldots, -2)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

9. Seven-diagonal system

$$
\begin{aligned}
f_k(x) &= 4(x_k - x_{k+1}^2) + x_{k+1} - x_{k+2}^2 + x_{k+2} - x_{k+3}^2, \quad k = 1 \\
f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) \\
&+ 4(x_k - x_{k+1}^2) + x_{k-1}^2 + x_{k+1} - x_{k+2}^2 + x_{k+2} - x_{k+3}^2, \quad k = 2 \\
f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) \\
&+ 4(x_k - x_{k+1}^2) + x_{k-1}^2 - x_{k-2} + x_{k+1} - x_{k+2}^2 \\
&+ x_{k-2}^2 + x_{k+2} - x_{k+3}^2, \quad k = 3 \\
f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) \\
&+ 4(x_k - x_{k+1}^2) + x_{k-1}^2 - x_{k-2} + x_{k+1} - x_{k+2}^2 \\
&+ x_{k-2}^2 + x_{k+2} - x_{k-3} - x_{k+3}^2, \quad 3 < k < n - 2 \\
f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) \\
&+ 4(x_k - x_{k+1}^2) + x_{k-1}^2 - x_{k-2} + x_{k+1} - x_{k+2}^2 \\
&+ x_{k-2}^2 + x_{k+2} - x_{k-3}, \quad k = n - 2 \\
f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) \\
&+ 4(x_k - x_{k+1}^2) + x_{k-1}^2 - x_{k-2} + x_{k+1} \\
&+ x_{k-2}^2 - x_{k-3}, \quad k = n - 1 \\
f_k(x) &= 8x_k(x_k^2 - x_{k-1}) - 2(1 - x_k) + x_{k-1}^2 - x_{k-2} \\
&+ x_{k-2}^2 - x_{k-3}, \quad k = n.
\end{aligned}
$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (-3, \ldots, -3)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

10. Structured Jacobian problem

$$
\begin{aligned}
f_k(x) &= -2x_k^2 + 3x_k - 2x_{k+1} + 3x_{n-4} - x_{n-3} \\
&- x_{n-2} + 0.5x_{n-1} - x_n + 1, \quad k = 1 \\
f_k(x) &= -2x_k^2 + 3x_k - x_{k-1} - 2x_{k+1} + 3x_{n-4} - x_{n-3} \\
&- x_{n-2} + 0.5x_{n-1} - x_n + 1, \quad 1 < k < n \\
f_k(x) &= -2x_k^2 + 3x_k - x_{k-1} + 3x_{n-4} - x_{n-3} \\
&- x_{n-2} + 0.5x_{n-1} - x_n + 1, \quad k = n.
\end{aligned}
$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (-1, \ldots, -1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

11. Extended Freudenstein and Roth function

$$\begin{aligned} f_k(x) &= x_k + ((5 - x_{k+1})x_{k+1} - 2)x_{k+1} - 13, \quad \mathrm{mod}(k,2) = 1 \\ f_k(x) &= x_{k-1} + ((x_k + 1)x_k - 14)x_k - 29, \quad \mathrm{mod}(k,2) = 0. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}}$ tal que

$$\begin{aligned} x^0_{\text{padrão}}(i) &= 90, \quad \mathrm{mod}(i,2) = 1 \\ x^0_{\text{padrão}}(i) &= 60, \quad \mathrm{mod}(i,2) = 0, \end{aligned}$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

12. Extended Powell singular problem

$$\begin{aligned} f_k(x) &= x_k + 10x_{k+1}, \quad \mathrm{mod}(k,4) = 1 \\ f_k(x) &= \sqrt{5}(x_{k+1} - x_{k+2}), \quad \mathrm{mod}(k,4) = 2 \\ f_k(x) &= (x_{k-1} - 2x_k)^2, \quad \mathrm{mod}(k,4) = 3 \\ f_k(x) &= \sqrt{10}(x_{k-3} - x_k)^2, \quad \mathrm{mod}(k,4) = 0. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}}$ tal que

$$\begin{aligned} x^0_{\text{padrão}}(i) &= 3, \quad \mathrm{mod}(i,4) = 1 \\ x^0_{\text{padrão}}(i) &= -1, \quad \mathrm{mod}(i,4) = 2 \\ x^0_{\text{padrão}}(i) &= 0, \quad \mathrm{mod}(i,4) = 3 \\ x^0_{\text{padrão}}(i) &= 1, \quad \mathrm{mod}(i,4) = 0, \end{aligned}$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 28$ e $n = 48$.

13. Extended Cragg and Levy problem

$$\begin{aligned} f_k(x) &= (\exp(x_k) - x_{k+1})^2, \quad \mathrm{mod}(k,4) = 1 \\ f_k(x) &= 10(x_k - x_{k+1})^3, \quad \mathrm{mod}(k,4) = 2 \\ f_k(x) &= \tan^2(x_k - x_{k+1}), \quad \mathrm{mod}(k,4) = 3 \\ f_k(x) &= x_k - 1, \quad \mathrm{mod}(k,4) = 0. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}}$ tal que

$$\begin{aligned} x^0_{\text{padrão}}(i) &= 1, \quad \text{mod}(i,4) = 1 \\ x^0_{\text{padrão}}(i) &= 2, \quad \text{mod}(i,4) \neq 1, \end{aligned}$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 28$ e $n = 48$.

14. Broyden tridiagonal problem

$$\begin{aligned} f_k(x) &= x_k(0.5x_k - 3) + 2x_{k+1} - 1, \quad k = 1 \\ f_k(x) &= x_k(0.5x_k - 3) + x_{k-1} + 2x_{k+1} - 1, \quad 1 < k < n \\ f_k(x) &= x_k(0.5x_k - 3) - 1 + x_{k-1}, \quad k = n. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (-1, \ldots, -1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

15. Generalized Broyden banded problem

$$\begin{aligned} f_k(x) &= (2 + 5x_k^2)x_k + 1 + \sum_{i=k_1}^{k_2} x_i(1 + x_i) \\ k_1 &= \max\{1, k-5\} \\ k_2 &= \min\{n, k+1\}. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (-1, \ldots, -1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

16. Extended Powell badly scaled function

$$\begin{aligned} f_k(x) &= 10000x_kx_{k+1} - 1, \quad \text{mod}(k,2) = 1 \\ f_k(x) &= \exp(-x_{k-1}) + \exp(-x_k) - 1.0001, \quad \text{mod}(k,2) = 2. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}}$ tal que

$$\begin{aligned} x^0_{\text{padrão}}(i) &= 0, \quad \text{mod}(i,2) = 1 \\ x^0_{\text{padrão}}(i) &= 1, \quad \text{mod}(i,2) = 0, \end{aligned}$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

17. Extended Wood problem

$$\begin{aligned} f_k(x) &= -200x_k(x_{k+1} - x_k^2) - (1 - x_k), \quad \text{mod}(k,4) = 1 \\ f_k(x) &= 200(x_k - x_{k-1}^2) + 20(x_k - 1) + 19.8(x_{k+2} - 1), \quad \text{mod}(k,4) = 2 \\ f_k(x) &= -180x_k(x_{k+1} - x_k^2) - (1 - x_k), \quad \text{mod}(k,4) = 3 \\ f_k(x) &= 180(x_k - x_{k-1}^2) + 20.2(x_k - 1) + 19.8(x_{k-2} - 1), \quad \text{mod}(k,4) = 4. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}}$ tal que

$$\begin{aligned} x^0_{\text{padrão}}(i) &= -3, \quad \text{mod}(i,2) = 1 \\ x^0_{\text{padrão}}(i) &= -1, \quad \text{mod}(i,2) = 0, \end{aligned}$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 24$ e $n = 48$.

18. Tridiagonal exponential problem

$$\begin{aligned} f_k(x) &= x_k - \exp(\cos(h(x_k + x_{k+1}))), \quad k = 1 \\ f_k(x) &= x_k - \exp(\cos(h(x_{k-1} + x_k + x_{k+1}))), \quad 1 < k < n \\ f_k(x) &= x_k - \exp(\cos(h(x_{k-1} + x_k))), \quad k = n \\ h &= \frac{1}{n+1}. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1.5, \ldots, 1.5)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

19. Discrete boundary value problem

$$\begin{aligned} f_k(x) &= 2x_k + 0.5h^2(x_k + hk)^3 - x_{k+1}, \quad k = 1 \\ f_k(x) &= 2x_k + 0.5h^2(x_k + hk)^3 - x_{k-1} - x_{k+1}, \quad 1 < k < n \\ f_k(x) &= 2x_k + 0.5h^2(x_k + hk)^3 - x_{k-1}, \quad k = n \\ h &= \frac{1}{n+1} \end{aligned}$$

Os pontos iniciais adotados foram:

$$x^0_{\text{padrão}} = (h(h-1), 2h(2h-1), \ldots, nh(nh-1))^T,$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

20. Brent problem

$$\begin{aligned} f_k(x) &= 3x_k(x_{k+1} - 2x_k) + x_{k+1}^2/4, \quad k = 1 \\ f_k(x) &= 3x_k(x_{k+1} - 2x_k + x_{k-1}) + (x_{k+1} - x_{k-1})^2/4, \quad 1 < k < n \\ f_k(x) &= 3x_k(20 - 2x_k + x_{k-1}) + (20 - x_{k-1})^2/4, \quad k = n. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (10, \ldots, 10)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

21. Troesch problem

$$\begin{aligned} f_k(x) &= 2x_k + \rho h^2 \sinh(\rho x_k) - x_{k+1}, \quad k = 1 \\ f_k(x) &= 2x_k + \rho h^2 \sinh(\rho x_k) - x_{k-1} - x_{k+1}, \quad 1 < k < n \\ f_k(x) &= 2x_k + \rho h^2 \sinh(\rho x_k) - x_{k-1} - 1, \quad k = n \\ \rho &= 10 \\ h &= \frac{1}{n+1}. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1, \ldots, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

22. Exponential function 1

$$\begin{aligned} f_1(x) &= e^{x_1 - 1} - 1 \\ f_i(x) &= i(e^{x_i - 1} - x_i), \quad 1 < i \leq n \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (\frac{n}{n-1}, \ldots, \frac{n}{n-1})^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

23. Exponential function 2

$$\begin{aligned} f_1(x) &= e^{x_1} - 1 \\ f_i(x) &= \frac{i}{10}(e^{x_i} + x_{i-1} - 1), \quad 1 < i \leq n. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (\frac{1}{n^2}, \ldots, \frac{1}{n^2})^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

24. Exponential function

$$\begin{aligned} f_i(x) &= \frac{i}{10}(1 - x_i^2 - e^{-x_i^2}), \quad i = 2, \ldots, n-1 \\ f_n(x) &= \frac{n}{10}(1 - e^{-x_n^2}). \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (\frac{1}{4n^2}, \frac{2}{4n^2}, \ldots, \frac{n}{4n^2})^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

25. Diagonal function premultiplied by a quasi-orthogonal matrix (n is a multiple of 3)

$$\begin{aligned} \text{For} \quad i &= 1, 2, \ldots, n/3 \\ f_{3i-2}(x) &= 0.6x_{3i-2} + 1.6x_{3i-2}^3 - 7.2x_{3i-1}^2 + 9.6x_{3i-1} - 4.8 \\ f_{3i-1}(x) &= 0.48x_{3i-2} - 0.72x_{3i-1}^3 + 3.24x_{3i-1}^2 - 4.32x_{3i-1} - x_{3i} + 0.2x_{3i}^3 + 2.16 \\ f_{3i}(x) &= 1.25x_{3i} - 0.25x_{3i}^3. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = \Big(-1, \frac{1}{2}, -1, \ldots, -1, \frac{1}{2}, -1\Big)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 24$ e $n = 48$.

26. Extended Rosenbrock function (n is even)

$$\begin{aligned} \text{For} \quad i &= 1, 2, \ldots, n/2 \\ f_{2i-1}(x) &= 10(x_{2i} - x_{2i-1}^2) \\ f_{2i}(x) &= 1 - x_{2i-1}. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (5, 1, \ldots, 5, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

27. Chandrasekhar's H-equation

$$F_6(H)(\mu) = H(\mu) - \Big(1 - \frac{c}{2}\int_0^1 \frac{\mu H(\nu)}{\mu + \nu} d\mu\Big)^{-1} = 0.$$

The discretized version is:

$$f_i(x) = x_i - \Big(1 - \frac{c}{2n}\sum_{j=1}^{n} \frac{\mu_i x_j}{\mu_i + \mu_j}\Big)^{-1}, \quad \text{for} \quad i = 1, \ldots, n$$

with $c \in [0, 1)$ and $\mu_i = (i - 1/2)/n$, for $1 \leq i \leq n$. (In our experiments we take $c = 0.9$). Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1, \ldots, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

28. Trigonometric function

$$f_i(x) = 2\Big(n + i(1 - \cos(x_i)) - \sin(x_i) - \sum_{j=1}^{n} \cos(x_j)\Big)(2\sin(x_i) - \cos(x_i)).$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (\frac{101}{100n}, \ldots, \frac{101}{100n})^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

29. Singular function

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

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1, \ldots, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

30. Logarithmic function

$$f_i(x) = \ln(x_i + 1) - \frac{x_i}{n}, \quad i = 1, 2, \ldots, n.$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1, \ldots, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

31. Variable band function

$$\begin{aligned} f_1(x) &= -2x_1^2 + 3x_1 - 2x_2 + 0.5x_{\alpha_1} + 1 \\ f_i(x) &= -2x_1^2 + 3x_i - x_{i-1} - 2x_{i+1} + 0.5x_{\alpha_i} + 1, \quad i = 2, \ldots, n-1 \\ f_n(x) &= -2x_n^2 + 3x_n - x_{n-1} + 0.5x_{\alpha_n} + 1, \end{aligned}$$

and α_i is a random integer number in $[\alpha_{i_{min}}, \alpha_{i_{max}}]$, where $\alpha_{i_{min}} = \max\{1, i-2\}$ and $\alpha_{i_{max}} = \min\{n, i+2\}$, for all i.
Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (0, \ldots, 0)^T$, $x^0 = (1, \ldots, 1)^T$ e $x^0 = (2, \ldots, 2)^T$ com dimensões $n = 30$ e $n = 50$.

32. Variable band function 2

$$\begin{aligned} f_1(x) &= -2x_1^2 + 3x_1 - 2x_2 + 0.5x_{\alpha_1} + 1 \\ f_i(x) &= -2x_1^2 + 3x_i - x_{i-1} - 2x_{i+1} + 0.5x_{\alpha_i} + 1, \quad i = 2, \ldots, n-1 \\ f_n(x) &= -2x_n^2 + 3x_n - x_{n-1} + 0.5x_{\alpha_n} + 1, \end{aligned}$$

and α_i is a random integer number in $[\alpha_{i_{min}}, \alpha_{i_{max}}]$, where $\alpha_{i_{min}} = \max\{1, i-10\}$ and $\alpha_{i_{max}} = \min\{n, i+10\}$, for all i.
Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (0, \ldots, 0)^T$, $x^0 = (1, \ldots, 1)^T$ e $x^0 = (2, \ldots, 2)^T$ com dimensões $n = 30$ e $n = 50$.

33. Function 15

$$\begin{aligned} f_1(x) &= -2x_1^2 + 3x_1 + 3x_{n-4} - x_{n-3} - x_{n-2} + 0.5x_{n-1} - x_n + 1 \\ f_i(x) &= -2x_i^2 + 3x_i - x_{i-1} - 2x_{i+1} + 3x_{n-4} - x_{n-3} - x_{n-2} + 0.5x_{n-1} \\ &\quad -x_n + 1, \quad i = 2, \ldots, n-1 \\ f_n(x) &= -2x_n^2 + 3x_n - x_{n-1} + 3x_{n-4} - x_{n-3} - x_{n-2} + 0.5x_{n-1} - x_n + 1. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (-1, \ldots, -1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

34. Strictly convex function
$F(x)$ is the gradient of $h(x) = \sum_{i=1}^{n}(e^{x_1} - x_i)$.

$$f_i(x) = e^{x_i} - 1, \quad i = 1, 2, \ldots, n.$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (\frac{1}{n}, \frac{2}{n}, \ldots, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

35. Strictly convex funtion 2
$F(x)$ is the gradient of $h(x) = \sum_{i=1}^{n} \frac{i}{10}(e^{x_1} - x_i)$.

$$f_i(x) = \frac{i}{10}(e^{x_i} - 1), \quad i = 1, 2, \ldots, n.$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1, \ldots, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

36. Function 18 (n is a multiple of 3)

$$\begin{aligned} \text{For} \quad i &= 1, 2, \ldots, n/3 \\ f_{3i-2}(x) &= x_{3i-2}x_{3i-1} - x_{3i}^2 - 1 \\ f_{3i-1}(x) &= x_{3i-2}x_{3i-1}x_{3i} - x_{3i-2}^2 + x_{3i-1}^2 - 2 \\ f_{3i}(x) &= e^{-x_{3i-2}} - e^{-x_{3i-1}}. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (0, \ldots, 0)^T$, $x^0 = (1, \ldots, 1)^T$ e $x^0 = (2, \ldots, 2)^T$ com dimensões $n = 24$ e $n = 48$.

37. Zero Jacobian function

$$\begin{aligned} f_1(x) &= \sum_{j=1}^{n} x_j^2 \\ f_i(x) &= -2x_1x_i, \quad i = 2, \ldots, n. \end{aligned}$$

Os pontos iniciais adotados foram:

$$x^0_{\text{padrão}} = (\frac{100(n-100)}{n}, \frac{(n-1000)(n-500)}{(60n)^2}, \ldots, \frac{(n-1000)(n-500)}{(60n)^2})^T,$$

$x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

38. Function 21 (n is a multiple of 3)

$$\begin{aligned} \text{For} \quad i &= 1, 2, \ldots, n/3 \\ f_{3i-2}(x) &= x_{3i-2}x_{3i-1} - x_{3i}^2 - 1 \\ f_{3i-1}(x) &= x_{3i-2}x_{3i-1}x_{3i} - x_{3i-2}^2 + x_{3i-1}^2 - 2 \\ f_{3i}(x) &= e^{-x_{3i-2}} - e^{-x_{3i-1}}. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1, \ldots, 1)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 24$ e $n = 48$.

39. Linear function-full rank

$$f_i(x) = x_i - \frac{2}{n}\sum_{j=1}^{n} x_j + 1.$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (100, \ldots, 100)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

40. Linear function-rank 2

$$\begin{aligned} f_1(x) &= x_1 - 1 \\ f_i(x) &= i\sum_{j=1}^{n} jx_j - i, \quad i = 2, 3, \ldots, n. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1, \frac{1}{n}, \ldots, \frac{1}{n})^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

41. Penalty I function

$$\begin{aligned} f_i(x) &= \sqrt{10^{-5}}(x_i - 1), \quad i = 1, 2, \ldots, n-1 \\ f_n(x) &= \frac{1}{4n}\sum_{j=1}^{n} x_j^2 - \frac{1}{4}. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (\frac{1}{3}, \ldots, \frac{1}{3})^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

42. Brown almost function

$$\begin{aligned} f_i(x) &= x_i + \sum_{j=1}^{n} x_j - (n+1), \quad i = 1, 2, \ldots, n-1 \\ f_n(x) &= \prod_{j=1}^{n} x_j - 1. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (\frac{n-1}{n}, \ldots, \frac{n-1}{n})^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

43. Variable dimensioned function

$$\begin{aligned} f_i(x) &= x_i - 1, \quad i = 1, 2, \ldots, n-2 \\ f_{n-1}(x) &= \sum_{j=1}^{n-2} j(x_j - 1) \\ f_n(x) &= \Big(\sum_{j=1}^{n-2} j(x_j - 1)\Big)^2. \end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1 - \frac{1}{n}, 1 - \frac{2}{n}, \ldots, 0)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

44. Tridimensional valley function (n is a multiple of 3)

$$\begin{aligned}
\text{For} \quad i &= 1, 2, \ldots, n/3 \\
f_{3i-2}(x) &= (c_2 x_{3i-2}^3 + c_1 x_{3i-2})\exp\Big(\frac{-x_{3i-2}^2}{100}\Big) - 1 \\
f_{3i-1}(x) &= 10(\sin(x_{3i-2}) - x_{3i-1}) \\
f_{3i}(x) &= 10(\cos(x_{3i-2}) - x_{3i}) \\
c_1 &= 1.003344481605351 \\
c_2 &= -3.344481605351171\,\text{x}\,10^{-3}.
\end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (2, 1, 2, 1, \ldots)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 48$.

45. Complementary function (n is even)

$$\begin{aligned}
\text{For} \quad i &= 1, 2, \ldots, n/2 \\
f_{2i-1}(x) &= \Big(x_{2i-1}^2 + \Big(x_{2i-1}e^{x_{2i-1}} - \frac{1}{n}\Big)^2\Big)^{1/2} - x_{2i-1} \\
&\quad -x_{2i-1}e^{x_{2i-1}} + \frac{1}{n} \\
f_{2i}(x) &= \Big(x_{2i}^2 + (3x_i + \sin(x_{2i}) + e^{x_{2i}})^2\Big)^{1/2} - x_{2i} \\
&\quad -3x_{2i} - \sin(x_{2i}) - e^{x_{2i}}.
\end{aligned}$$

Os pontos iniciais adotados foram: $x^0_{\text{padrão}} = (0.5, \ldots, 0.5)^T$, $x^0 = 2x^0_{\text{padrão}}$ e $x^0 = 5x^0_{\text{padrão}}$ com dimensões $n = 30$ e $n = 50$.

46. Minimal function

$$f_i(x) = \frac{(\ln(x_i) + \exp(x_i)) - \sqrt{(\ln(x_i) - \exp(x_i))^2 + 10^{-10}}}{2}.$$

pontos iniciais adotados foram: $x^0_{\text{padrão}} = (1.5, \ldots, 1.5)^T$, $x^0 = (2, \ldots, 2)^T$ e $x^0 = (5, \ldots, 5)^T$ com dimensões $n = 30$ e $n = 50$.