UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS FÍSICAS E MATEMÁTICAS
BACHARELADO EM MATEMÁTICA

Pedro Henrique Silva Passos

Métodos sem derivada para solução de sistemas não lineares monótonos

Florianópolis - SC
2024

Pedro Henrique Silva Passos

Métodos sem derivada para solução de sistemas não lineares monótonos

Trabalho de Conclusão de Curso de Graduação em Bacharelado em Matemática do Centro de Ciências Físicas e Matemáticas da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Matemática.
Orientador: Prof. Douglas Soares Gonçalves

Florianópolis - SC
2024

Ficha de identificação da obra

A ficha de identificação é elaborada pelo próprio autor.

Orientações em:

http://portalbu.ufsc.br/ficha

Pedro Henrique Silva Passos

Métodos sem derivada para solução de sistemas não lineares monótonos

Este Trabalho de Conclusão de Curso de Graduação foi julgado adequado para obtenção do Título de Bacharel em Matemática e aprovado em sua forma final pelo Curso de Bacharelado em Matemática.

Florianópolis - SC, 2024.

Documento assinado digitalmente
Felipe Lopes Castro
Data: 05/08/2024 11:02:22-0300
CPF: ***.060.710-**
Verifique as assinaturas em https://v.ufsc.br

Felipe Lopes Castro
Coordenador do Curso

Banca Examinadora:

Documento assinado digitalmente
Douglas Soares Goncalves
Data: 30/07/2024 08:45:35-0300
CPF: ***.566.318-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Douglas Soares Gonçalves (Orientador)
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Maicon Marques Alves
Data: 29/07/2024 19:18:15-0300
CPF: ***.631.910-**
Verifique as assinaturas em https://v.ufsc.br

Maicon Marques Alves
Universidade Federal de Santa Catarina

Documento assinado digitalmente
EVERTON BOOS
Data: 01/08/2024 19:40:51-0300
CPF: ***.893.329-**
Verifique as assinaturas em https://v.ufsc.br

Everton Boos
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Durante a graduação tive a oportunidade de conhecer inúmeras pessoas, tanto da matemática quanto de outros cursos. Gostaria de agradece-los, pois todos contribuíram com a minha formação acadêmica e pessoal.

Agradeço minha mãe, Analice, e pai, Elder, por me dar a oportunidade de me dedicar a um curso desafiador como a matemática, e a minha família por sempre me apoiar e acreditarem em mim.

Gostaria de agradecer imensamente o professor Douglas Soares Gonçalves, afinal foi com sua orientação que evoluí como jamais esperaria. Quero agradece-lo por sempre encontrar tempo e paciência para me orientar nesses dois anos de Iniciação científica e no Trabalho de Conclusão de Curso.

Agradeço aos professores do Departamento de Matemática da UFSC pela contribuição na minha formação.

Agradeço aos colegas e amigos da matemática, que foram partes essenciais da minha formação, pois estavam sempre me incentivando e ajudando durante todo esse tempo.

RESUMO

Neste trabalho estudaremos alguns métodos para solução de sistemas não lineares monótonos. Encontrar a solução de um sistema não linear significa encontrar $x \in \mathbb{R}^n$ que é raiz simultaneamente de todas as equações não lineares que definem o sistema. Tais equações podem ser organizadas e associadas às componentes de um operador $F : \mathbb{R}^n \to \mathbb{R}^n$ e buscamos x tal que $F(x) = 0$. Sistemas não lineares tem muitas aplicações em física, economia, engenharias e outras áreas. Estudaremos um caso mais específico de sistemas não lineares, que são os sistemas nos quais F é monótono e Lipschitz contínuo. Neste caso podemos usar métodos que não necessitam da derivada do operador, diferentemente de métodos clássicos para o problema, como o método de Newton. Após apresentar com certo detalhe resultados de convergência para três métodos desta classe, avaliaremos o desempenho prático destes em problemas teste da literatura, e em duas aplicações relacionadas a processamento de sinais. Tais experimentos computacionais indicam que os métodos sem derivadas que tiram proveito da monotonia do operador podem ser mais eficientes que métodos clássicos quando a informação de derivada não está disponível e precisa ser aproximada.

Palavras-chave: Sistemas não lineares, operador monótono, Lipschitz contínuo, métodos sem derivada.

ABSTRACT

In this work, we will study some methods for solving monotone nonlinear systems. Finding the solution of a nonlinear system means finding $x \in \mathbb{R}^n$ which is a simultaneous root of all the nonlinear equations defining the system. These equations can be organized and associated with components of an operator $F : \mathbb{R}^n \to \mathbb{R}^n$, and we seek x such that $F(x) = 0$. Nonlinear systems have many applications in physics, economics, engineering, and other fields. We will focus on a more specific case of nonlinear systems, where F is monotone and Lipschitz continuous. In this case, methods can be used that do not require the derivative of the operator, unlike classical methods such as the Newton's method. After presenting convergence results in some detail for three methods of this class, we will evaluate their practical performance on test problems from the literature and on two applications related to signal processing. Computational experiments indicate that derivative-free methods taking advantage of the monotonicity of the operator can be more efficient than classical methods when derivative information is unavailable and needs to be approximated.

Keywords: Non linear systems, monotone operator, Lipschitz continuous, derivative-free methods.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Métodos para solução de sistemas de equações não lineares $F(x) = 0$ são muito importantes, já que diversos fenômenos naturais e físicos são descritos por tais sistemas, tendo aplicações em várias áreas científicas como na economia e circuitos elétricos (KHALIL, 2002). Por isso, desde o século passado muitos estudos sobre métodos iterativos para resolução de sistemas não lineares foram feitos, se aproveitando da evolução dos computadores, inclusive métodos para casos específicos de sistemas não lineares, como os sistemas monótonos.

Para sistemas não lineares diferenciáveis um dos métodos de resolução mais conhecidos é o método de Newton (BERTSEKAS, 2003, Capítulo 1) que faz uso da Jacobiana (matriz de derivadas parciais) para determinar os iterados. No entanto, em certas aplicações a função F não é diferenciável em toda parte, ou ainda a informação de derivada não está disponível, o que demanda o uso de métodos sem derivadas.

Com isso em mente, neste trabalho, estudaremos métodos para sistemas não lineares monótonos que não demandam a Jacobiana do operador F, mas tiram proveito da propriedade de *monotonia*. Tais métodos utilizam apenas a informação $F(x_k)$, i.e. o operador F avaliado no iterado atual x_k, para gerar o novo iterado $x_{k+1} = x_k - t_k F(x_k)$, em que $t_k > 0$ é um tamanho de passo apropriado. Explorando o fato de F ser monótono e uma escolha adequada do tamanho de passo é possível mostrar a convergência da sequência (x_k) para uma solução do problema.

Iremos analisar com certo detalhe a teoria de convergência de três métodos nesta classe e avaliar sua eficiência computacional comparando-os com outros métodos mais genéricos em problemas teste da literatura. Além disso, uma aplicação de sistemas não lineares em processamento de sinais (SMITH, 1997) também será discutida.

O trabalho está organizado da seguinte maneira: o **Capítulo 2** trará noções fundamentais sobre sistemas de equações não lineares, operadores monótonos, Lipschitz contínuos, e resultados auxiliares que serão utilizados no decorrer do trabalho. No **Capítulo 3** apresentamos três algoritmos para resolução de sistemas não lineares que não fazem uso da derivada. Estudaremos em detalhes a boa definição e convergência de tais algoritmos. No **Capítulo 4** avaliaremos a eficiência dos algoritmos estudados em alguns problemas teste da literatura, e também no problema de processamento de sinais esparsos, o qual pode ser escrito como um sistema não linear monótono.

Por fim, no **Capítulo 5** discutiremos as considerações finais e trabalhos futuros.

2 CONCEITOS PRELIMINARES

Neste capítulo revisaremos alguns resultados e definições da teoria de operadores, sequências reais, sistemas não lineares e otimização contínua, que serão utilizados no decorrer deste trabalho. Todo o conteúdo aqui apresentado é baseado nas referências (BAUSCHKE; COMBETTES, 2017), (BERTSEKAS, 2003), (DENNIS; SCHNABEL, 1996), (KHALIL, 2002), (LA CRUZ, 2017), (ROCKAFELLAR; WETS, 1998) e (XIAO; WANG; HU, 2011).

2.1 SISTEMAS DE EQUAÇÕES NÃO-LINEARES

Seja $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ uma aplicação de $\mathbb{R}^n$ em $\mathbb{R}^n$ tal que $F(x) = (F_1(x), \ldots, F_n(x))^T$, em que $F_i : \mathbb{R}^n \to \mathbb{R}$, para $i = 1, \ldots, n$. Também chamamos F de função vetorial de $\mathbb{R}^n$ em $\mathbb{R}^n$, ou ainda de *operador* de $\mathbb{R}^n$ em $\mathbb{R}^n$. O problema de interesse neste trabalho é o de encontrar (se possível) $x \in \mathbb{R}^n$ tal $F(x) = 0$, isto é, determinar $x \in \mathbb{R}^n$ cuja imagem por F é o vetor nulo de $\mathbb{R}^n$, ou ainda, $x \in \mathbb{R}^n$ que resolve, simultaneamente, as equações não-lineares $F_i(x) = 0$, para $i = 1, 2, \ldots, n$.

O sistema de equações $F(x) = 0$ é chamado *sistema não-linear* e denotaremos o problema de interesse por

$$\text{encontre } x \in \mathbb{R}^n \text{ tal que } F(x) = 0. \qquad \text{P'}$$

É possível transformar o problema P' em um problema de otimização (minimização) equivalente. Note que P' admite solução se, e somente se, o minimizador global [1] de

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

é zero (em que $||\cdot||$ é a norma usual em $\mathbb{R}^n$). De fato, sabemos que $\frac{1}{2}||F(x)||^2 \geq 0$ $\forall x \in \mathbb{R}^n$, e então se x é solução de P', x é minimizador global de $f(x)$. Por outro lado, se z minimizador global de f é tal que $f(z) = 0$ então, $\frac{1}{2}||F(z)||^2 = 0$, logo $||F(z)|| = 0$, e concluímos que $F(z) = 0$ e portanto z é solução de P'. Dada essa equivalência, iremos considerar também o problema:

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2}||F(x)||^2. \qquad \text{P}$$

Para esses problemas existem muitos métodos clássicos da literatura, como por exemplo o método de Newton (BERTSEKAS, 2003, Capítulo 1) e métodos de Gauss-Newton e Levenberg-Marquardt (DENNIS; SCHNABEL, 1996, Capítulo 10). Porém, muitos desses métodos utilizam a Jacobiana do operador para calcular o próximo iterado, e isto pode ser um impeditivo quando o operador não é diferenciável em todo $\mathbb{R}^n$, ou simplesmente não conhecemos sua derivada. Com isso em mente o foco deste trabalho

[1] Sejam $f : \mathbb{R}^n \to \mathbb{R}$ e $y \in \mathbb{R}^n$, se $f(y) \leq f(x)$ $\forall x \in \mathbb{R}^n$, então y é minimizador global de f.

foi estudar algoritmos que não utilizam a Jacobiana do operador. Esses algoritmos são para uma classe particular de sistema não linear, que são os sistemas que tem o operador associado monótono e Lipschitz contínuo, ambos conceitos serão apresentados na seção seguinte.

2.2 OPERADORES MONÓTONOS E LIPSCHITZ CONTÍNUO

Nesta seção veremos a definição de operador monótono e Lipschitz contínuo, trazendo também exemplos desses operadores. A seguir, discutiremos algumas particularidades de sistemas não-lineares definidos por estes tipos de operadores.

Definição 2.2.1. *Seja $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ um operador e $\langle \cdot, \cdot \rangle$ o produto interno usual em $\mathbb{R}^n$. Dizemos que F é um operador monótono se*

$$\langle F(x) - F(y), x - y \rangle \geq 0 \quad \forall x, y \in \mathbb{R}^n.$$

Observação 1. *Se $\langle F(x) - F(y), x - y \rangle > 0 \quad \forall x, y \in \mathbb{R}^n$ com $x \neq y$, diremos que F é estritamente monótono.*

Definição 2.2.2. *Seja $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ um operador e $||\cdot||$ a norma usual em $\mathbb{R}^n$, dizemos que F é Lipschitz contínuo se existe $\ell > 0$ tal que*

$$\|F(x) - F(y)\| \leq \ell \|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

Para esses conceitos ficarem claros, apresentaremos alguns exemplos de tais operadores.

Exemplo 1. *O operador $F(x) = Ax - b$ com $x, b \in \mathbb{R}^n$ e $A \in \mathbb{R}^{n \times n}$ semidefinida positiva é monótono e Lipschitz contínuo.*

De fato, para $x, y \in \mathbb{R}^n$, temos que:

$$\begin{aligned} \langle F(x) - F(y), x - y \rangle &= \langle Ax - b - (Ay - b), x - y \rangle \\ &= \langle Ax - Ay, x - y \rangle \\ &= \langle A(x - y), x - y \rangle. \end{aligned}$$

De A ser semidefinida positiva temos que $\langle Az, z \rangle \geq 0 \ \forall z \in \mathbb{R}^n$. Logo, da igualdade anterior segue que $\langle F(x) - F(y), x - y \rangle \geq 0$. Portanto $F(x)$ é monótono.

Agora para mostrar que é F Lipschitz contínuo veja que:

$$\begin{aligned} \|F(x) - F(y)\| &= \|Ax - b - (Ay - b)\| \\ &= \|Ax - Ay\| \\ &= ||A(x - y)|| \\ &\leq ||A||||x - y|| \end{aligned}$$

(em que $||A||$ é a norma usal de matrizes). Logo $\ell = ||A||$, e portanto $F(x)$ é Lipschitz contínuo.

Exemplo 2. *Se f é convexa e diferenciável em $\mathbb{R}^n$, o operador $F(x) = \nabla f(x)$ é monótono.*

Com efeito, sejam $x, y \in \mathbb{R}^n$. Da convexidade e diferenciabilidade de f temos as seguintes desigualdades (BERTSEKAS, 2003, Apêndice B):

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T(y - x) \quad \forall x, y \in \mathbb{R}^n, \\ f(x) &\geq f(y) + \nabla f(y)^T(x - y) \quad \forall x, y \in \mathbb{R}^n. \end{aligned}$$

Agora, somando tais desigualdades temos:

$$\begin{aligned} f(y) + f(x) &\geq f(x) + f(y) + \nabla f(x)^T(y - x) + \nabla f(y)^T(x - y) \\ 0 &\geq \nabla f(x)^T(y - x) + \nabla f(y)^T(x - y) \\ \nabla f(x)^T(x - y) - \nabla f(y)^T(x - y) &\geq 0 \\ \langle \nabla f(x), x - y \rangle - \langle \nabla f(y), x - y \rangle &\geq 0 \\ \langle \nabla f(x) - \nabla f(y), x - y \rangle &\geq 0 \\ \langle F(x) - F(y), x - y \rangle &\geq 0 \end{aligned}$$

Portanto $F(x)$ é monótono.

Exemplo 3. *Seja $C \subset \mathbb{R}^n$ um conjunto não-vazio, convexo e fechado. Se $F : \mathbb{R}^n \longrightarrow C$ denota a projeção ortogonal sobre C, então F é monótono e Lipschitz contínuo.*

Sejam $x, y \in \mathbb{R}^n$ e $F(x) = [x]^+$ e $F(y) = [y]^+$, em que $[z]^+$ denota a projeção ortogonal de $z \in \mathbb{R}^n$ em C. De $[x]^+$ ser a projeção de x em C temos a seguinte propriedade (BERTSEKAS, 2003, Apêndice B):

$$\langle z - [x]^+, x - [x]^+ \rangle \leq 0 \quad (\forall z \in C).$$

Analogamente para $[y]^+$ temos:

$$\langle z - [y]^+, y - [y]^+ \rangle \leq 0 \quad (\forall z \in C).$$

Como $[y]^+$ e $[x]^+$ pertencem a C substituindo z por $[y]^+$ e $[x]^+$ respectivamente obtemos:

$$\begin{aligned} \langle [y]^+ - [x]^+, x - [x]^+ \rangle &\leq 0, \\ \langle [x]^+ - [y]^+, y - [y]^+ \rangle &\leq 0. \end{aligned}$$

Agora somando ambas inequações:

$$\langle [x]^+ - [y]^+, y - [y]^+ \rangle + \langle [y]^+ - [x]^+, x - [x]^+ \rangle \leq 0. \tag{2}$$

Como

$$\begin{aligned} \langle [y]^+ - [x]^+, x - [x]^+ \rangle &= \langle (-1)([x]^+ - [y]^+), x - [x]^+ \rangle \\ &= -\langle [x]^+ - [y]^+, x - [x]^+ \rangle, \end{aligned} \tag{3}$$

substituindo (3) em (2) temos:

$$\begin{aligned}
\langle [x]^+ - [y]^+, y - [y]^+ \rangle - \langle [x]^+ - [y]^+, x - [x]^+ \rangle &\leq 0 \\
\langle [x]^+ - [y]^+, y - [y]^+ - x + [x]^+ \rangle &\leq 0 \\
\langle [x]^+ - [y]^+, y - x + [x]^+ - [y]^+ \rangle &\leq 0 \\
\langle [x]^+ - [y]^+, [x]^+ - [y]^+ \rangle + \langle [x]^+ - [y]^+, y - x \rangle &\leq 0. \qquad (4)
\end{aligned}$$

Como $\langle [x]^+ - [y]^+, [x]^+ - [y]^+ \rangle \geq 0$, então:

$$\begin{aligned}
\langle [x]^+ - [y]^+, y - x \rangle &\leq 0 \\
\langle [x]^+ - [y]^+, (-1)(x - y) \rangle &\leq 0 \\
(-1)\langle [x]^+ - [y]^+, x - y \rangle &\leq 0.
\end{aligned}$$

De onde segue que:

$$\begin{aligned}
\langle [x]^+ - [y]^+, x - y \rangle &\geq 0 \\
\langle F(x) - F(y), x - y \rangle &\geq 0.
\end{aligned}$$

Portanto F é monótono. Agora para mostrar que é Lipschitz, partindo de (4) temos

$$\begin{aligned}
\langle [x]^+ - [y]^+, [x]^+ - [y]^+ \rangle + \langle [x]^+ - [y]^+, y - x \rangle &\leq 0 \\
\langle [x]^+ - [y]^+, y - x \rangle &\leq \langle [x]^+ - [y]^+, [x]^+ - [y]^+ \rangle \\
\langle [x]^+ - [y]^+, y - x \rangle &\leq || [x]^+ - [y]^+ ||^2 \\
|| [x]^+ - [y]^+ ||^2 &\leq \langle [x]^+ - [y]^+, x - y \rangle
\end{aligned}$$

e da desigualdade de Cauchy-Schwarz, e supondo $|| [x]^+ - [y]^+ || > 0$, segue que:

$$\begin{aligned}
|| [x]^+ - [y]^+ ||^2 \leq \langle [x]^+ - [y]^+, y - x \rangle &\leq || [x]^+ - [y]^+ || \, ||x - y|| \\
|| [x]^+ - [y]^+ ||^2 &\leq || [x]^+ - [y]^+ || \, ||x - y|| \\
|| [x]^+ - [y]^+ || &\leq ||x - y|| \\
|| F(x) - F(y) || &\leq ||x - y||.
\end{aligned}$$

Portanto F é Lipschitz contínuo com $\ell = 1$.

Vale destacar que operadores monótonos generalizam a noção de gradiente e derivadas, através do uso de subderivadas, subgradientes e inequações variacionais (ROCKAFELLAR; WETS, 1998, Capítulo 12). Essas ferramentas proporcionam a solução de muitos problemas de otimização sem utilizar a derivada do operador.

Se F for estritamente monótono e diferenciável, então é possível mostrar que se x não é solução de P', então $-F(x)$ fornece uma *direção de descida* para $\|F(x)\|^2$ a partir de x.

Proposição 2.2.1. *Seja $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ um operador estritamente monótono, diferenciável e $f(x) = \frac{1}{2}||F(x)||^2$. Se $F(x) \neq 0$, então $-F(x)$ é uma direção de descida* [2] *para f.*

Demonstração. Para mostrar que $-F(x)$ é direção de descida basta mostrar que

$$\nabla f(x)^T(-F(x)) < 0.$$

Como $\nabla f(x) = J(x)^T F(x)$ temos que:

$$\nabla f(x)^T(-F(x)) = (J(x)^T F(x))^T(-F(x)) = -F(x)^T J(x)F(x).$$

Agora, vamos mostrar que $\langle J(x)v, v\rangle > 0 \ \forall x \in \mathbb{R}^n$ e $\forall v \in \mathbb{R}^n \setminus \{0\}$. Para isso assuma por contradição que existe um $v \in \mathbb{R}^n\setminus\{0\}$ tal que $\langle J(x)v, v\rangle \leq 0$. Note que o sinal do produto interno não muda para qualquer vetor não nulo u múltiplo positivo de v. Usando a aproximação de Taylor de primeira ordem com u sendo múltiplo de v temos:

$$F(x + u) = F(x) + J(x)u + o(||u||).$$

Disso temos:

$$\begin{aligned}\langle F(x + u) - F(x), u\rangle &= \langle F(x) + J(x)u + o(||u||) - F(x), u\rangle \\ &= \langle J(x)u, u\rangle + o||u||^2\end{aligned}$$

Agora dividindo tudo por $||u||^2$:

$$\frac{\langle F(x + u) - F(x), u\rangle}{||u||^2} = \frac{\langle J(x)u, u\rangle}{||u||^2} + \frac{o||u||^2}{||u||^2}. \tag{5}$$

Como $\frac{\langle J(x)u,u\rangle}{||u||^2} = \alpha \leq 0$, pois u é múltiplo não negativo de v, tomando um u com norma suficientemente pequena, o termo $\frac{o||u||^2}{||u||^2}$ fica menor em módulo que α e assim o lado direito de (5) fica não positivo, mas $\langle F(x + u) - F(x), u\rangle > 0$ por F ser estritamente monótona, chegando a uma contradição (que veio de assumirmos que $\forall x \in \mathbb{R}^n$ existe um $v \in \mathbb{R}^n\setminus\{0\}$ tal que $\langle J(x)v, v\rangle \leq 0$). Logo, $\langle J(x)v, v\rangle > 0 \ \forall x \in \mathbb{R}^n$ e $\forall v \in \mathbb{R}^n$. Portanto $J(x)$ é definida positiva, e finalmente:

$$\nabla f(x)^T(-F(x)) = -F(x)^T J(x)F(x) < 0.$$

Com isso concluímos que $-F(x)$ é uma direção de descida para $f(x)$. □

Os algoritmos que estudaremos no próximo capítulo farão uso de um múltiplo positivo da direção $-F(x)$ no intuíto de reduzir o valor de $f(x)$, embora tal redução possa não ocorrer em toda iteração.

[2] Sejam $f : \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$ e $d \in \mathbb{R}^n$. Se existir $\epsilon > 0$ tal que $f(x + td) < f(x)$, com $t \in [0, \epsilon]$, então d é uma direção de descida.

2.3 RESULTADOS AUXILIARES

Encerramos este capítulo com alguns resultados técnicos de análise que serão úteis em capítulos posteriores. Tais resultados foram extraídos de (LA CRUZ, 2017) e (XIAO; WANG; HU, 2011), mas apresentamos suas demonstrações por uma questão de completude.

Lema 2.3.1. *Sejam (a_k) e (b_k) sequências positivas em $\mathbb{R}$ tais que:*

$$a_{k+1} \leq (1+b_k)a_k + b_k, \tag{6}$$
$$\sum_{k=0}^{\infty} b_k = \beta < \infty.$$

Então a sequência (a_k) converge.

Demonstração. Vamos mostrar que (a_k) é limitada superiormente. Seja $c_k = \prod_{i=1}^{k-1}(1+b_i)$. Então $c_k \geq 1$, e note que $\prod_{i=1}^{\infty}(1+b_i)$ converge se $\sum_{k=0}^{\infty} b_k < \infty$, pois $(1+b_k) \leq e^{b_k}$ e fazendo o produtório em ambos os lados temos:

$$\prod_{k=0}^{\infty}(1+b_k) < e^{\beta}.$$

Portanto (c_k) é limitada superiormente e como (c_k) é crescente então (c_k) converge. De (6) segue que:

$$\begin{aligned}\frac{a_{k+1}}{c_{k+1}} &\leq \frac{(1+b_k)a_k}{c_{k+1}} + \frac{b_k}{c_{k+1}} \\ &= \frac{a_k}{c_k} + \frac{b_k}{c_{k+1}} \leq \frac{a_k}{c_k} + b_k.\end{aligned}$$

E aplicando (6) recursivamente para $\frac{a_k}{c_k} + b_k$ obtemos:

$$\frac{a_{m+1}}{c_{m+1}} \leq \frac{a_1}{c_1} + \sum_{k=1}^{m} b_k \leq \gamma := \frac{a_1}{c_1} + \beta.$$

Então $a_{m+1} \leq (c_{m+1})\gamma \leq \gamma e^{\beta} =: a$, logo (a_k) também será limitada e portanto terá ao menos um ponto limite. Agora suponha que existam subsequências (a_{k_j}) e (a_{k_i}) de (a_k) que convergem respectivamente para w e v. Como $a_k \leq a$, temos que se $k_j \geq k_i$, usando (6) e uma soma telescópica, obtemos $a_{k_j} - a_{k_i} \leq (1+a)\sum_{k=k_i}^{k_j} b_k$, fazendo $k_j \to \infty$ obtemos:

$$w - a_{k_i} \leq (1+a)\sum_{k=k_i}^{\infty} b_k.$$

Agora fazendo $k_i \to \infty$:

$$w - v \leq 0.$$

Fazendo analogamente se $k_i \geq k_j$ teremos $v - w \leq 0$. Logo $w = v$ e portanto (a_k) converge. □

Lema 2.3.2. *Sejam $a, b, c, d \in \mathbb{R}$, então*

$$\min(a, b) - \min(c, d) = (1 - m)(a - c) + m(b - d) \tag{7}$$

em que

$$m = \begin{cases} 0 & ,se \;\; b \geq a, d \geq c \\ 1 & ,se \;\; a \geq b, c \geq d \\ \frac{\min(a,b)-\min(c,d)+c-a}{b-d+c-a} & ,caso\ contrário. \end{cases}$$

Demonstração. Vamos dividir a demonstração em quatro casos.
Caso(i): $b \geq a, d \geq c$, nesse caso $m = 0$, substituindo em (7)

$$(1 - 0)(a - c) + 0(b - d) = a - c = \min(a, b) - \min(c, d).$$

Caso(ii): $a \geq b, c \geq d$, nesse caso $m = 1$, substituindo em (7)

$$(1 - 1)(a - c) + 1(b - d) = b - d = \min(a, b) - \min(c, d).$$

Caso(iii): $b \geq a, c \geq d$, nesse caso $m = \frac{\min(a,b)-\min(c,d)+c-a}{b-d+c-a} = \frac{c-d}{b-d+c-a}$, substituindo em (7)

$$\begin{aligned}(1 - \frac{c - d}{b - d + c - a})(a - c) + \frac{c - d}{b - d + c - a}(b - d) &= \frac{b - d + c - a + d - c}{b - d + c - a}(a - c) + \frac{c - d}{b - d + c - a}(b - d) \\ &= \frac{b - a}{b - d + c - a}(a - c) + \frac{cb - cd - bd + d^2}{b - d + c - a}.\end{aligned}$$

Agora note que

$$\begin{aligned}\frac{b - a}{b - d + c - a}(a - c) + \frac{cb - cd - bd + d^2}{b - d + c - a} &= \frac{ba - bc - a^2 + ac + cb - cd - bd + d^2}{b - d + c - a} \\ &= \frac{a(b + c - a) - d(b - d + c) + da - da}{b - d + c - a} \\ &= \frac{a(b - d + c - a) - d(b - d + c - a)}{b - d + c - a} \\ &= a - d = \min(a, b) - \min(c, d).\end{aligned}$$

Caso(iv): $a \geq b, d \geq c$, nesse caso $m = \frac{\min(a,b)-\min(c,d)+c-a}{b-d+c-a} = \frac{b-a}{b-d+c-a}$, substituindo em (7)

$$\begin{aligned}(1 - \frac{b - a}{b - d + c - a})(a - c) + \frac{b - a}{b - d + c - a}(b - d) &= \frac{b - d + c - a + d - c}{b - d + c - a}(a - c) + \frac{c - d}{b - d + c - a}(b - d) \\ &= \frac{c - d}{b - d + c - a}(a - c) + \frac{b^2 - bd - ab + ad}{b - d + c - a}.\end{aligned}$$

Agora note que

$$\begin{aligned}\frac{c-d}{b-d+c-a}(a-c)+\frac{b^2-bd-ab+ad}{b-d+c-a} &= \frac{-c^2+ca-da+cd+b^2-bd-ab+ad}{b-d+c-a}\\ &= \frac{b(b-d-a)-c(-d+c-a)+cb-cb}{b-d+c-a}\\ &= \frac{b(b-d+c-a)-c(b-d+c-a)}{b-d+c-a}\\ &= b-c=\min(a,b)-\min(c,d).\end{aligned}$$

Com isso mostramos que a equação (7) é valida para todos os casos possíveis. □

3 ALGORITMOS SEM DERIVADA PARA SOLUÇÃO DE SISTEMAS NÃO-LINEARES MONÓTONOS

Neste capítulo apresentamos os métodos sem derivada para sistemas não-lineares monótonos, baseado nos artigos (LA CRUZ, 2017), (SOLODOV; SVAITER, 1999) e (YU *et al.*, 2009), bem como a teoria de convergência para tais métodos.

Os três métodos utilizam a direção de busca $d_k = -\alpha F(x_k)$, $\alpha > 0$, e procuram garantir um decréscimo, ainda que não-monótono, no valor de $\|F(x)\|^2$ ao longo das iterações.

Para os resultados apresentados neste capítulo iremos assumir as seguintes hipóteses:

H1: Existe $x^* \in \mathbb{R}^n$ tal que $F(x^*) = 0$.

H2: F é monótono e Lipschitz contínuo.

3.1 ALGORITMO DFSANME

Primeiro apresentamos o algoritmo para a solução de sistemas não-lineares DFSANME (Derivative-Free Spectral Algorithm for Nonlinear Monotone Equations) desenvolvido por La Cruz em (LA CRUZ, 2017).

Algoritmo 1: DFSANME

Inicialização. Escolha $x_0 \in \mathbb{R}^n$, $1 \ll \alpha_{max} < \infty$, $\alpha_0 \in (0, \alpha_{max})$, $\gamma, \sigma \in (0,1)$ e uma sequência positiva (η_k) que satisfaz $\sum_{k=0}^{\infty} \eta_k = \eta < \infty$. Faça $k := 0$.

Passo 1. Se $||F(x_k)|| = 0$, pare.

Passo 2. Tome $\lambda := 1$.

Passo 3. Enquanto $||F(x_k - \lambda\alpha_k F(x_k))||^2 > ||F(x_k)||^2 + \eta_k - \gamma\lambda^2||F(x_k)||^2$ faça $\lambda := \sigma\lambda$.

Passo 4. Defina $\lambda_k = \lambda$, $x_{k+1} = x_k - \lambda_k\alpha_k F(x_k)$ e

$$\alpha_{k+1} = \begin{cases} \frac{||x_{k+1}-x_k||^2}{\langle x_{k+1}-x_k, F(x_{k+1})-F(x_k)\rangle}, & \text{se } \frac{\langle F(x_{k+1}),F(x_k)\rangle}{||F(x_k)||^2} \leq 1 - \frac{\alpha_k}{\alpha_{max}} \\ \alpha_{max}, & \text{caso contrário.} \end{cases}$$

Passo 5. Faça $k := k + 1$ e volte para o Passo 1.

No Algoritmo 1 temos as seguintes definições:

$$x_{k+1} = x_k - \lambda_k\alpha_k F(x_k) \tag{8}$$

$$||F(x_k - \lambda_k \alpha_k F(x_k))||^2 = ||F(x_{k+1})||^2 \leq ||F(x_k)||^2 + \eta_k - \gamma \lambda_k^2 ||F(x_k)||^2 \tag{9}$$

em que (8) é a definição do próximo iterado x_{k+1}, e (9) é uma busca linear que utiliza uma sequência η_k para possibilitar uma busca não monótona pelo tamanho de passo λ_k, não exigindo um decréscimo estrito no valor da norma do resíduo. A sequência η_k ser somável, significa que a sequência deve ir para zero suficientemente rápido. Esta é uma condição técnica necessária na prova de convergência do algoritmo, segundo (LA CRUZ, 2017). Para mostrar as propriedades de convergência do Algoritmo 1, e que este está bem definido, vamos assumir além das hipóteses enunciadas no começo deste capítulo a seguinte condição:

H3: Assuma que o conjunto $\Omega_0 = \{x \in \mathbb{R}^n : ||F(x)||^2 \leq ||F(x_0)||^2 + \eta\}$ é limitado.

Com essas hipóteses podemos começar a apresentar os lemas necessários para mostrar que o algoritmo está bem definido e sua convergência global.

Lema 3.1.1. *Se **H1**, **H2** e **H3** são satisfeitas, então as iterações do Algoritmo 1 estão bem-definidas e $x_k \in \Omega_0$ $\forall k \geq 0$, em que (x_k) é a sequência gerada pelo algoritmo.*

Demonstração. Primeiro vamos mostrar que $\alpha_k \leq \alpha_{max}$ para $k \geq 0$. Como F é monótono e $x_{k+1} = x_k - \lambda_k \alpha_k F(x_k)$, para $k \geq 0$ temos:

$$0 \leq \langle x_{k+1} - x_k, F(x_{k+1}) - F(x_k) \rangle = \lambda_k \alpha_k (||F(x_k)||^2 - \langle F(x_{k+1}), F(x_k) \rangle). \tag{10}$$

Substituindo (10) em α_{k+1} temos:

$$\alpha_{k+1} = \begin{cases} \frac{\lambda_k \alpha_k ||F(x_k)||^2}{||F(x_k)||^2 - \langle F(x_{k+1}), F(x_k) \rangle} & , se \ \frac{\langle F(x_{k+1}), F(x_k) \rangle}{||F(x_k)||^2} \leq 1 - \frac{\alpha_k}{\alpha_{max}} \\ \\ \alpha_{max} & , \text{caso contrário.} \end{cases}$$

Pela expressão acima é fácil ver que:

$$\alpha_{k+1} = \alpha_{max}, \quad \text{se } \frac{\langle F(x_{k+1}), F(x_k) \rangle}{||F(x_k)||^2} > 1 - \frac{\alpha_k}{\alpha_{max}}.$$

Agora se $\frac{\langle F(x_{k+1}), F(x_k) \rangle}{||F(x_k)||^2} \leq 1 - \frac{\alpha_k}{\alpha_{max}}$ temos:

$$\frac{\alpha_k}{1 - \frac{\langle F(x_{k+1}), F(x_k) \rangle}{||F(x_k)||^2}} \leq \alpha_{max}. \tag{11}$$

Novamente pela definição de α_{k+1}:

$$\alpha_{k+1} = \frac{\lambda_k \alpha_k ||F(x_k)||^2}{||F(x_k)||^2 - \langle F(x_{k+1}), F(x_k) \rangle} = \lambda_k \left(\frac{\alpha_k}{1 - \frac{\langle F(x_{k+1}), F(x_k) \rangle}{||F(x_k)||^2}} \right).$$

Como $\lambda_k \in (0,1)$ de (11) segue que $\alpha_{k+1} \leq \alpha_{max}$, e portanto $\alpha_k \leq \alpha_{max}$ para todo k. Agora como F é contínua, $\eta_k > 0$ e $||F(x_0)|| < \infty$, usando um argumento indutivo assuma que a busca linear:

$$||F(x_k - \lambda_k \alpha_k F(x_k))||^2 \leq ||F(x_k)||^2 + \eta_k - \gamma\lambda^2||F(x_k)||^2$$

é satisfeita após um número finito de reduções de λ. Assuma por contradição que a busca linear não é satisfeita para $k+1$ em um número finito de reduções de λ, então teríamos:

$$||F(x_{k+1} - \lambda\alpha_{k+1} F(x_{k+1}))||^2 > ||F(x_{k+1})||^2 + \eta_{k+1} - \gamma\lambda^2||F(x_{k+1})||^2$$

para todo $\lambda > 0$. Então para $\lambda \to 0^+$ obtemos, graças a continuidade de F e de $\|\cdot\|$:

$$||F(x_{k+1})||^2 > ||F(x_{k+1})||^2 + \eta_{k+1}.$$

Como $\eta_k > 0$ temos uma contradição que veio de assumirmos que a busca linear não é satisfeita após finitas reduções de λ. Portanto a busca linear (9) é satisfeita para todo k, logo o método DFSANME está bem definido, pois gera x_{k+1}, para todo $k \geq 0$.
Por fim usando (9) e um argumento indutivo é fácil notar que:

$$||F(x_k)||^2 \leq ||F(x_0)||^2 + \sum_{i=0}^{k-1} \eta_i \text{ , para } k \geq 1.$$

Disso e como $\sum_{k=0}^{\infty} \eta_k = \eta < \infty$, temos:

$$||F(x_k)|| \leq ||F(x_0)||^2 + \eta \text{ para } k \geq 0.$$

Portanto a sequência (x_k) gerada pelo Algoritmo 1 pertence a Ω_0 como queríamos mostrar. □

Nos próximos resultados assumiremos que o algoritmo não para, isto é, $\|F(x_k)\| > 0$ para todo k, de modo que o algoritmo gera uma sequência infinita (x_k).

Lema 3.1.2. *Se **H1**, **H2** e **H3** são satisfeitas e (x_k) é a sequência gerada pelo Algoritmo 1, então as seguintes propriedades são satisfeitas:*

(i) A sequência $(||F(x_k)||)$ converge.
(ii) Existe uma constante $\alpha_{min} > 0$, tal que: $\alpha_{min} \leq \alpha_k$, para $k \geq 0$.
(iii) $\lim_{k\to\infty} \lambda_k ||F(x_k)|| = 0$.
(iv) A sequência (x_k) satisfaz $\lim_{k\to\infty} ||x_{k+1} - x_k|| = 0$.
(v) Existe um conjunto $K \subset \mathbb{N}$ que contém uma sequência infinita de índices tal que:

$$\langle F(x_{k+1}), F(x_k)\rangle = \left(1 - \frac{\lambda_k \alpha_k}{\alpha_{k+1}}\right) ||F(x_{k+1})||^2, \quad \forall k \in K. \tag{12}$$

Demonstração. (i) Seja $a_k = ||F(x_k)||^2$ e $b_k = \eta_k$. Como $||F(x_k)||^2 \geq 0$, $(1 + \eta_k) \geq 1$ e $x_{k+1} = x_k - \lambda_k \alpha_k F(x_k)$, substituindo isso na busca linear (9) temos:

$$||F(x_{k+1})||^2 \leq ||F(x_k)||^2 + \eta_k \leq (1 + \eta_k)||F(x_k)||^2 + \eta_k.$$

Então pelo Lema 2.3.1 e a definição de (η_k) a sequência $||F(x_k)||^2$ converge, logo $||F(x_k)||$ também converge como queríamos mostrar.
(ii) Defina $\alpha_{min} = \min\{\ell^{-1}, \alpha_0\}$. Como F é Lipschitz contínuo com constante ℓ, usando a desigualdade de Cauchy-Schwarz obtemos:

$$\frac{\langle x_{k+1}-x_k, F(x_{k+1})-F(x_k)\rangle}{||x_{k+1}-x_k||^2} \leq \frac{||x_{k+1}-x_k||\ ||F(x_{k+1})-F(x_k)||}{||x_{k+1}-x_k||^2} \leq \ell.$$

Disso e da definição de α_{k+1} temos que:

$$\ell^{-1} \leq \frac{||x_{k+1}-x_k||^2}{\langle x_{k+1}-x_k, F(x_{k+1})-F(x_k)\rangle} = \alpha_{k+1} \text{ para } k \geq 0.$$

Portanto ℓ^{-1} é menor igual que todo α_k para $k \geq 1$, logo $\alpha_{min} = \min\{\ell^{-1}, \alpha_0\}$ é cota inferior de α_k para $k \geq 0$.
(iii) Da busca linear (9) para $k \geq 0$ temos:

$$\lambda_k^2||F(x_k)||^2 \leq \frac{\eta_k}{\gamma} + \frac{||F(x_k)||^2 - ||F(x_{k+1})||^2}{\gamma}.$$

Como $\sum_{k=0}^{\infty} \eta_k = \eta < \infty$, fazendo o somatório em ambos os lados obtemos:

$$\sum_{k=0}^{\infty} \lambda_k^2||F(x_k)||^2 \leq \gamma^{-1}(\eta + ||F(x_0)||^2) < \infty.$$

Veja que $\lambda_k \in (0, 1]$, pois da definição do DFSANME $\lambda_0 = 1$ e $\lambda_k = \sigma^k\lambda$ com $\sigma \in (0, 1)$, portanto $\lim_{k\to\infty} \lambda_k||F(x_k)|| = 0$.
(iv) Como $x_{k+1} - x_k = -\lambda_k\alpha_k F(x_k)$, então

$$||x_{k+1} - x_k|| = || - \lambda_k\alpha_k F(x_k)||$$

e de α_k ser limitado e de (iii) temos que $\lim_{k\to\infty} ||x_{k+1} - x_k|| = 0$.

(v) Como (12) é satisfeito se $||F(x_k)|| = 0$, assuma que $0 < ||F(x_k)|| < \infty$. Agora defina o seguinte conjunto:

$$K = \left\{ k \in \mathbb{N} : \frac{\langle F(x_{k+1}), F(x_k)\rangle}{||F(x_k)||^2} \leq \max\left(\frac{\alpha_k}{\alpha_{max}}, 1 - \frac{\alpha_k}{\alpha_{max}}\right)\right\}. \tag{13}$$

Vamos provar que K possui infinitos índices tal que (12) é satisfeito. Como F é monótona , $\alpha_k \geq 0$ e $\lambda_k \in (0, 1]$, então por (10) temos:

$$\langle F(x_{k+1}), F(x_k)\rangle \leq ||F(x_k)||^2 \tag{14}$$

para todo $k \geq 0$. Suponha que $K = \emptyset$, então temos que:

$$\frac{\langle F(x_{k+1}), F(x_k)\rangle}{||F(x_k)||^2} > \max\left(\frac{\alpha_k}{\alpha_{max}}, 1 - \frac{\alpha_k}{\alpha_{max}}\right) \text{ para } k \geq 0. \tag{15}$$

Então pela definição de α_{k+1} e (15) temos que $\alpha_k = \alpha_{max}$ e daí:

$$\frac{\langle F(x_{k+1}), F(x_k)\rangle}{||F(x_k)||^2} > \frac{\alpha_k}{\alpha_{max}} = 1,$$
$$\langle F(x_{k+1}), F(x_k)\rangle > ||F(x_k)||^2 \text{ para } k \geq 0,$$

contradizendo (14). Então K é não vazio. Se supomos que K é finito usando o mesmo argumento de antes, chegaríamos numa contradição quando $\alpha_k = \alpha_{max}$ para algum k que não pertence a K. Agora resta provar que (12) é satisfeito para todo $k \in K$. É claro que $\alpha_k < \alpha_{max}$ para todo $k \in K$. Disso, de (10) e da construção de α_{k+1} temos que:

$$\alpha_{k+1} = \frac{\lambda_k \alpha_k ||F(x_k)||^2}{||F(x_k)||^2 - \langle F(x_{k+1}), F(x_k)\rangle} \text{ para } k \in K,$$

isto é,

$$\frac{\lambda_k \alpha_k}{\alpha_{k+1}} = 1 - \frac{\langle F(x_{k+1}), F(x_k)\rangle}{||F(x_k)||^2} \text{ para } k \in K.$$

Portanto (12) é satisfeito para todo $k \in K$. □

Teorema 3.1.3. *Se **H1, H2 e H3** são satisfeitas e (x_k) é a sequência gerada pelo Algoritmo 1, então*

$$\lim_{k\to\infty} ||F(x_k)|| = 0.$$

Demonstração. Do Lema 3.1.2(i) temos que $(\|F(x_k)\|)$ converge e pela definição do algoritmo $(\lambda_k) \in (0, 1]\ \forall k \geq 0$.

Se $\lambda_k \geq \underline{\lambda} > 0$ para todo $k \in K$, em que K é o conjunto definido por (13), então do Lema 3.1.2(iii) segue que $\lim_{k\in K} ||F(x_k)|| = 0$. Mas como $(\|F(x_k)\|)$ converge, temos que $\lim_{k\to\infty} ||F(x_k)|| = 0$.

Caso contrário, existe uma subsequência $(\lambda_k)_{K_1}$ com $K_1 \subset K$ tal que $\lim_{k\in K_1} \lambda_k = 0$. Como para $k \in K_1$ suficientemente grande, $\lambda_k < 1$ satisfaz a busca linear (9), então $\widetilde{\lambda_k} = \frac{\lambda_k}{\sigma}$ não satisfaz (9), ou seja, $\widetilde{\lambda_k}$ seria o último tamanho de passo que falhou na iteração k. Logo

$$\begin{aligned} ||F(x_k - \widetilde{\lambda_k}\alpha_k F(x_k))||^2 &> ||F(x_k)||^2 + \eta_k - \gamma\widetilde{\lambda_k}^2 ||F(x_k)||^2 \\ &> ||F(x_k)||^2 - \gamma\widetilde{\lambda_k}^2 ||F(x_k)||^2. \end{aligned} \tag{16}$$

Defina

$$\begin{aligned} x_k - \widetilde{\lambda_k}\alpha_k F(x_k) &= \widetilde{x_k}, \\ x_k - \lambda_k \alpha_k F(x_k) &= x_{k+1}. \end{aligned}$$

Como $(||F(x_k)||)$ converge, $||F(x_k)|| \leq M$ para algum $M > 0$ e por F ser Lipschitz contínuo temos que:

$$\begin{aligned} ||F(\widetilde{x_k}) - F(x_{k+1})||^2 &\leq \ell^2 ||\widetilde{x_k} - x_{k+1}||^2 \\ &= \ell^2 ||x_k - \widetilde{\lambda_k}\alpha_k F(x_k) - x_k + \lambda_k \alpha_k F(x_k)|| \\ &= \ell^2 ||(\lambda_k - \widetilde{\lambda_k})\alpha_k F(x_k)||^2 \\ &\leq \ell^2 \alpha_{max}^2 ||F(x_k)||^2 |\lambda_k - \widetilde{\lambda_k}|^2 \\ &\leq \ell^2 \alpha_{max}^2 M^2 |\lambda_k - \frac{\lambda_k}{\sigma}|^2 \\ &= \ell^2 \alpha_{max}^2 M^2 \lambda_k^2 |1 - \frac{1}{\sigma}|^2. \end{aligned}$$

Disso e pela desigualdade triangular reversa obtemos:

$$||F(\widetilde{x_k})|| - ||F(x_{k+1})|| \leq ||F(\widetilde{x_k}) - F(x_{k+1})|| \leq \lambda_k \left(\ell \alpha_{max} M |1 - \tfrac{1}{\sigma}|\right).$$

Mas veja que:

$$\begin{aligned} ||F(\widetilde{x_k})||^2 - ||F(x_{k+1})||^2 &= (||F(\widetilde{x_k})|| - ||F(x_{k+1})||)(||F(\widetilde{x_k})|| + ||F(x_{k+1})||) \\ &\leq (||F(\widetilde{x_k})|| - ||F(x_{k+1})||)2M \\ &\leq \left(2M^2 \ell \alpha_{max} |1 - \frac{1}{\sigma}|\right) \lambda_k. \end{aligned}$$

Portanto podemos escrever:

$$||F(\widetilde{x_k})||^2 = ||F(x_{k+1})||^2 + \lambda_k^2 \epsilon_k \qquad \text{para } k \geq 0$$

para uma sequência limitada adequada (ϵ_k). Assim a inequação (16) se torna:

$$||F(x_{k+1})||^2 - ||F(x_k)||^2 > -\widetilde{\lambda_k}^2 (\gamma ||F(x_k)||^2 + \sigma^2 \epsilon_k). \tag{17}$$

Agora usando o fato de F ser Lipschitz contínuo e a igualdade (12), para $k \in K_1 \subset K$ podemos escrever:

$$\begin{aligned} \ell^2 \lambda_k^2 \alpha_k^2 ||F(x_k)||^2 &\geq ||F(x_{k+1}) - F(x_k)||^2 \\ &= ||F(x_{k+1})||^2 - 2\langle F(x_{k+1}), F(x_k)\rangle + ||F(x_k)||^2 \\ &= ||F(x_{k+1})||^2 - 2\left(1 - \frac{\lambda_k \alpha_k}{\alpha_{k+1}}\right) ||F(x_k)||^2 + ||F(x_k)||^2 \\ &= ||F(x_{k+1})||^2 - ||F(x_k)||^2 + 2\left(\frac{\lambda_k \alpha_k}{\alpha_{k+1}}\right) ||F(x_k)||^2. \end{aligned} \tag{18}$$

Agora, do Lema 3.1.2(ii) e da demonstração do Lema 3.1.1, existem α_{min} e α_{max} tais que $\alpha_{min} \leq \alpha_k \leq \alpha_{max}$ para $k \geq 0$, disso temos que:

$$\frac{\alpha_k}{\alpha_{k+1}} \geq \frac{\alpha_{min}}{\alpha_{max}} > 0 \ \text{ para } k \geq 0. \tag{19}$$

Assim, de (19), (17) e (18) para $k \in K_1$ obtemos:

$$\begin{aligned} \ell^2 \lambda_k^2 \alpha_k^2 ||F(x_k)||^2 &> -\widetilde{\lambda_k}^2 (\gamma ||F(x_k)||^2 + \sigma^2 \epsilon_k) + 2\left(\frac{\lambda_k \alpha_k}{\alpha_{k+1}}\right) ||F(x_k)||^2 \\ &\geq -\widetilde{\lambda_k}^2 (\gamma ||F(x_k)||^2 + \sigma^2 \epsilon_k) + 2\lambda_k \left(\frac{\alpha_{min}}{\alpha_{max}}\right) ||F(x_k)||^2. \end{aligned}$$

Daí substituindo $\widetilde{\lambda_k} = \frac{\lambda_k}{\sigma}$ e dividindo tudo por λ_k temos:

$$\ell^2 \lambda_k \alpha_{max}^2 ||F(x_k)||^2 > -\frac{\lambda_k}{\sigma^2} (\gamma ||F(x_k)||^2 + \sigma^2 \epsilon_k) + 2\left(\frac{\alpha_{min}}{\alpha_{max}}\right) ||F(x_k)||^2 \tag{20}$$

para $k \in K_1$. Agora, como $\lim_{k\in K_1} \lambda_k = 0$, fazendo o limite em ambos os lados de (20) obtemos:

$$0 \geq 2\left(\frac{\alpha_{min}}{\alpha_{max}}\right) \lim_{k\in K_1} ||F(x_k)||^2.$$

Uma vez que a sequência $(||F(x_k)||)$ converge, temos portanto que $\lim_{k\to\infty} ||F(x_k)|| = 0$. □

3.2 UM ALGORITMO BASEADO EM PROJEÇÃO

Agora veremos outro algoritmo para solução de sistemas não lineares monótonos, desenvolvido por Solodov e Svaiter em (SOLODOV; SVAITER, 1999), que utiliza uma estratégia um pouco diferente daquela do Algoritmo 1. A ideia é explorar um hiperplano que separa o ponto atual do conjunto solução para trazer o iterado corrente para mais próximo do conjunto solução através de sua projeção em tal hiperplano.

Para facilitar a exposição, no Algoritmo 2 faremos algumas simplificações em relação ao trabalho original: a direção d_k é definida originalmente pela solução do sistema linear $0 = F(x_k) + (G_k + \mu_k I)d_k + e_k$, em que G_k é uma matriz positiva semidefinida e e_k representa o erro na solução do sistema linear. Aqui, assumiremos que $e_k = 0$ e $G_k = 0$, ou seja, o vetor d_k é obtido de forma exata e dado por um múltiplo de $-F(x_k)$.

Algoritmo 2: Método da Projeção

Inicialização. Escolha $x_0 \in \mathbb{R}^n$, $\beta, \delta \in (0,1)$ e (μ_k) em $\mathbb{R}_{++}$.
Faça $k := 0$.

Passo 1. Se $F(x_k) = 0$, pare.

Passo 2. Faça $d_k = -\frac{1}{\mu_k}F(x_k)$.

Passo 3. Defina $\gamma_k = \beta^{m_k}$ com m_k sendo o menor inteiro tal que

$$-\langle F(x_k + \beta^m d_k), d_k\rangle \geq \delta\mu_k||d_k||^2,$$

e faça

$$y_k = x_k + \gamma_k d_k.$$

Passo 4. Compute $x_{k+1} = x_k - \frac{\langle F(y_k), x_k - y_k\rangle}{||F(y_k)||^2}F(y_k)$.

Passo 5. Faça $k := k + 1$ e volte para o Passo 1.

Do Algoritmo 2 temos as seguintes definições:

$$x_{k+1} = x_k - \frac{\langle F(y_k), x_k - y_k\rangle}{||F(y_k)||^2}F(y_k), \tag{21}$$

$$-\langle F(x_k + \beta^m d_k), d_k\rangle \geq \delta\mu_k||d_k||^2, \tag{22}$$

$$0 = F(x_k) + \mu_k d_k. \tag{23}$$

Na análise a seguir também iremos assumir que $\mu_k > 0$ é limitada e afastada de zero (existe $\delta > 0$ tal que $\delta < \mu_k \ \forall k \geq 0$). Deste modo, d_k é a mesma nos Algoritmos 1 e 2.

Para mostrar as propriedades de convergência e que o Algoritmo 2 está bem definido, começaremos com o seguinte lema (SOLODOV; SVAITER, 1999, Lema 2.1).

Lema 3.2.1. *Seja F um operador monótono e $x, y \in \mathbb{R}^n$ tais que*

$$\langle F(y), x - y \rangle > 0. \tag{24}$$

Se

$$x^+ = x - \frac{\langle F(y), x-y \rangle}{||F(y)||^2} F(y)$$

então para qualquer $\bar{x} \in \mathbb{R}^n$ tal que $F(\bar{x}) = 0$ temos que:

$$||x^+ - \bar{x}||^2 \leq ||x - \bar{x}||^2 - ||x^+ - x||^2.$$

Demonstração. Seja $\bar{x} \in \mathbb{R}^n$ tal que $F(\bar{x}) = 0$. Por F ser monótono segue que para qualquer $y \in \mathbb{R}^n$:

$$\langle F(y), \bar{x} - y \rangle \leq 0.$$

Então segue de (24) que o hiperplano

$$H := \{s \in \mathbb{R}^n | \langle F(y), s - y \rangle = 0\}$$

separa estritamente x de $\bar{x}$. Também é fácil ver que x^+ é a projeção de x no semiespaço $\{s \in \mathbb{R}^n \mid \langle F(y), s - y \rangle \leq 0 \}$. Como $\bar{x}$ pertence a esse semiespaço, segue das propriedades do operador projeção que $\langle x - x^+, x^+ - \bar{x} \rangle \geq 0$. Daí temos que

$$\begin{aligned} ||x - \bar{x}||^2 &= ||x - x^+ + x^+ - \bar{x}||^2 \\ &= ||x - x^+||^2 + ||x^+ - \bar{x}||^2 + 2\langle x - x^+, x^+ - \bar{x} \rangle \\ &\geq ||x - x^+||^2 + ||x^+ - \bar{x}||^2 \\ ||x - \bar{x}||^2 - ||x - x^+||^2 &\geq ||x^+ - \bar{x}||^2, \end{aligned}$$

como queríamos mostrar. □

Agora com esse lema podemos mostrar a convergência global com o seguinte teorema.

Teorema 3.2.2. *Seja F um operador contínuo, monótono e (x_k) uma sequência gerada pelo Algoritmo 2. Para qualquer $\bar{x}$ tal que $F(\bar{x}) = 0$, temos que*

$$||x_{k+1} - \bar{x}||^2 \leq ||x_k - \bar{x}||^2 - ||x_{k+1} - x_k||^2.$$

Em particular (x_k) é limitada. Além disso, a sequência (x_k) é finita e o último iterado é a solução, ou a sequência é infinita e

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

Se a sequência (x_k) é infinita e (μ_k) é limitada e afastada de zero, então (x_k) converge para algum $\bar{x}$ tal que $F(\bar{x}) = 0$.

Demonstração. Primeiro note que se a sequência termina em algum iterado k, então $F(x_k) = 0$, portanto estamos na solução. Por isso a partir de agora assumiremos que $d_k \neq 0$ para todo k. Agora mostraremos que o método está bem definido e uma sequência infinita (x_k) é gerada.

Primeiro mostraremos que a busca linear (22) sempre termina com um tamanho de passo positivo γ_k. Suponha por contradição que para algum iterado k isso não é verdade, ou seja, para todos os inteiros m temos que:

$$-\langle F(x_k + \beta^m d_k), d_k\rangle < \delta\mu_k||d_k||^2. \tag{25}$$

Porém aplicando o limite para $m \to \infty$ em (25) e o fato de $d_k = -\frac{1}{\mu_k}F(x_k)$ temos

$$\begin{aligned}
\delta\mu_k||d_k||^2 &\geq \lim_{m\to\infty} -\langle F(x_k + \beta^m d_k), d_k\rangle \\
&= -\langle F(x_k), d_k\rangle \\
&= -\langle F(x_k), -\frac{1}{\mu_k}F(x_k)\rangle \\
&= \frac{1}{\mu_k}||F(x_k)||^2 \\
&= \frac{\mu_k}{\mu_k^2}||F(x_k)||^2 \\
&= \mu_k||d_k||^2
\end{aligned}$$

implicando em

$$(1 - \delta)\mu_k||d_k||^2 \leq 0,$$

que é um absurdo, pois $\delta \in (0, 1)$, $\mu_k > 0$ e $d_k \neq 0$. Portanto a busca linear (e o algoritmo) está bem definida.

Agora, por (22) temos

$$\langle F(y_k), x_k - y_k\rangle = -\gamma_k\langle F(y_k), d_k\rangle \geq \delta\mu_k\gamma_k||d_k||^2 > 0. \tag{26}$$

Seja $\bar{x}$ um ponto qualquer tal que $F(\bar{x}) = 0$. Por (21) e (26) e o Lema 3.2.1 segue que:

$$||x_{k+1} - \bar{x}||^2 \leq ||x_k - \bar{x}||^2 - ||x_{k+1} - x_k||^2. \tag{27}$$

Logo, a sequência $(||x_k - \bar{x}||)$ é não-crescente e convergente. Portanto, a sequência (x_k) é limitada, e

$$\lim_{k\to\infty} ||x_{k+1} - x_k|| = 0. \tag{28}$$

Por (23), μ_k ser limitada e o fato de F ser contínua segue que

$$||F(x_k)|| = \mu_k ||d_k||,$$
$$\frac{||F(x_k)||}{\mu_k} = ||d_k||.$$

Assim, (d_k) é limitada e, consequentemente, (y_k) também será. Novamente pela continuidade de F, existe uma constante $C > 0$ tal que $\delta||F(y_k)||^{-1} \geq C$. Usando (21) e (26), obtemos

$$||x_{k+1} - x_k|| = \frac{\langle F(y_k), x_k - y_k\rangle}{||F(y_k)||} \geq C\mu_k\gamma_k||d_k||^2$$

e de (28) temos que

$$0 = \lim_{k\to\infty} \mu_k\gamma_k||d_k||^2. \tag{29}$$

Com isso consideramos duas possibilidades

$$\lim_{k\to\infty} ||F(x_k)|| = 0 \quad \text{ou} \quad \lim_{k\to\infty} ||F(x_k)|| > 0.$$

No primeiro caso a continuidade de F implica que a sequência (x_k) possui um ponto de acumulação $\hat{x}$ tal que $F(\hat{x}) = 0$. Como $\bar{x}$ foi uma solução arbitrária, podemos escolher $\bar{x} = \hat{x}$ em (27). A sequência $(||x_k - \hat{x}||)$ converge, e como $\hat{x}$ é um ponto de acumulação de (x_k), então (x_k) converge para $\hat{x}$.

Agora considerando o segundo caso, como $\mu_k \geq \mu_{\min} > 0$ para todo k,

$$\lim_{k\to\infty} \inf \mu_k > 0$$

e

$$||F(x_k)|| = \mu_k||d_k||$$

segue que $\lim_{k\to\infty} \inf ||d_k|| > 0$. Disso e de (29) temos que:

$$\lim_{k\to\infty} \gamma_k = 0.$$

Então similarmente $m_k \to \infty$. Pela regra do tamanho de passo temos que (22) não é satisfeito para $\beta^{m_{k-1}}$, ou seja:

$$-\langle F(x_k + \beta^{m_{k-1}} d_k), d_k \rangle < \delta \mu_k ||d_k||^2.$$

Como as sequências $(x_k), (\mu_k)$ e (d_k) são limitadas e usando uma subsequência se necessário, quando $k \to \infty$ temos:

$$-\langle F(\hat{x}), \hat{d} \rangle \leq \delta \hat{\mu} ||\hat{d}||^2$$

onde $\hat{x}, \hat{\mu}$ e $\hat{d}$ são os limites das subsequências correspondentes. Mas por (22) e um argumento já usado temos que:

$$-\langle F(\hat{x}), \hat{d} \rangle = \hat{\mu} ||\hat{d}||^2.$$

Dessas duas últimas relações temos uma contradição já que $\delta \in (0, 1)$. Portanto o caso $\lim_{k\to\infty} ||F(x_k)|| > 0$ não é possível, logo só vale $\lim_{k\to\infty} ||F(x_k)|| = 0$ como desejado.

□

Note que α_k^{-1}, com α_k como definido no Algoritmo 1, satisfaz as hipóteses necessárias de μ_k para a demonstração de convergência, logo podemos fazer

$$\mu_k = \frac{1}{\alpha_k}. \tag{30}$$

Essa estratégia será usada no algoritmo da próxima seção.

3.3 UM ALGORITMO COM PROJEÇÃO E PASSO ESPECTRAL

Por fim, veremos o terceiro algoritmo estudado neste trabalho, apresentado no artigo (YU *et al.*, 2009). Tal algoritmo (Algoritmo 3) consiste no Algoritmo 2 com uma busca linear um pouco diferente e empregando a ideia do parâmetro espectral (apresentado no Algoritmo 1) com uma salva-guarda.

Algoritmo 3: Método da Projeção com parâmetro espectral

Inicialização. Escolha $x_0 \in \mathbb{R}^n$, $\beta, \sigma \in (0,1)$ e $r > 0$ e defina $\theta_0 = 1$. Faça $k := 0$

Passo 1. Se $||F(x_k)|| = 0$, pare.

Passo 2. Compute a direção $d_k = -\theta_k F(x_k)$.

Passo 3. Defina $\gamma_k = \beta^{m_k}$ com m_k sendo o menor inteiro em que

$$-\langle F(x_k + \beta^m d_k), d_k \rangle \geq \sigma \beta^m ||d_k||^2$$

e então faça $z_k = x_k + \gamma_k d_k$.

Passo 4. Atualize

$$x_{k+1} = x_k - \frac{\langle F(z_k), x_k - z_k \rangle}{||F(z_k)||^2} F(z_k).$$

Passo 5. Compute $s_k = x_{k+1} - x_k$,

$$y_k = F(x_{k+1}) - F(x_k) + r s_k \quad \text{e} \quad \theta_{k+1} = \frac{\langle s_k, s_k \rangle}{\langle s_k, y_k \rangle}.$$

Passo 6. Faça $k := k + 1$ e volte para o Passo 1.

Do Algoritmo 3 temos as seguintes definições:

$$-\langle F(x_k + \beta^m d_k), d_k \rangle \geq \sigma \beta^m ||d_k||^2, \tag{31}$$

$$x_{k+1} = x_k - \frac{\langle F(z_k), x_k - z_k \rangle}{||F(z_k)||^2} F(z_k), \tag{32}$$

$$y_k = F(x_{k+1}) - F(x_k) + r s_k \quad \text{e} \quad \theta_{k+1} = \frac{\langle s_k, s_k \rangle}{\langle s_k, y_k \rangle}. \tag{33}$$

Observação 2. *Note que o parâmetro espectral é similar ao do Algoritmo 1, exceto pelo termo rs_k no produto interno do denominador.*

A busca linear também é similar a do Algoritmo 2, o que muda é que ao invés de usar o μ_k no lado direito da inequação, é usado β^m.

Com os lemas apresentados a seguir podemos mostrar a convergência global do algoritmo. O primeiro lema consiste em mostrar que a busca linear está bem definida e o segundo é para mostrar que o parâmetro espectral é limitado.

Lema 3.3.1. *Se **H2** é satisfeita, então existe um número não negativo m_k tal que* (31) *é satisfeita.*

Demonstração. Suponha por contradição que na iteração k a desigualdade (31) não é satisfeita para nenhum m_k, isto é

$$-\langle F(x_k + \beta^m d_k), d_k\rangle < \sigma\beta^m ||d_k||^2, \quad \forall m \geq 1.$$

Fazendo $m \to \infty$ obtemos

$$-\langle F(x_k), d_k\rangle \leq 0 \tag{34}$$

e pelos passos 1,2 e 5 temos que $F(x_k) \neq 0$ e $d_k \neq 0$.

Como F é monótono segue que

$$y_{k-1}^T s_{k-1} = \langle F(x_k) - F(x_{k-1}), x_k - x_{k-1}\rangle + r s_{k-1}^T s_{k-1} \geq r s_{k-1}^T s_{k-1} > 0$$

e por F ser Lipschitz contínuo, usando a desigualdade de Cauchy-Scharwz temos

$$y_{k-1}^T s_{k-1} = \langle F(x_k) - F(x_{k-1}), x_k - x_{k-1}\rangle + r s_{k-1}^T s_{k-1} \leq (\ell + r) s_{k-1}^T s_{k-1}. \tag{35}$$

Agora de (33) e (35) obtemos

$$\theta_k = \frac{s_{k-1}^T s_{k-1}}{y_{k-1}^T s_{k-1}} \geq \frac{1}{\ell + r} > 0.$$

Logo

$$-\langle F(x_k), d_k\rangle = \theta_k \langle F(x_k), F(x_k)\rangle > 0.$$

Mas essa última inequação contradiz (34), portanto a busca linear está bem definida. □

Lema 3.3.2. *O parâmetro θ_k é limitado.*

Demonstração. Já vimos na demonstração do Lema 3.3.1 que $\theta_k \geq \frac{1}{\ell+r}$. Para mostrar uma cota superior vamos olhar para a definição de θ_{k+1}:

$$\begin{aligned}
\theta_{k+1} &= \frac{\langle s_k, s_k\rangle}{\langle s_k, y_k\rangle} \\
&= \frac{||x_{k+1} - x_k||^2}{\langle x_{k+1} - x_k, F(x_{k+1}) - F(x_k) + r s_k\rangle} \\
&= \frac{||x_{k+1} - x_k||^2}{\langle x_{k+1} - x_k, F(x_{k+1}) - F(x_k)\rangle + r||x_{k+1} - x_k||^2}.
\end{aligned}$$

Como F é monótono $\langle x_{k+1} - x_k, F(x_{k+1}) - F(x_k)\rangle \geq 0$, disso segue:

$$\begin{aligned}
\theta_{k+1} &\leq \frac{||x_{k+1} - x_k||^2}{r||x_{k+1} - x_k||^2} \\
&\leq \frac{1}{r}.
\end{aligned}$$

Portanto $\frac{1}{\ell+r} \leq \theta_k \leq \frac{1}{r}$, para todo k. □

Com esses dois lemas temos o que é preciso para mostrar a convergência global:

Teorema 3.3.3. *Se **H1** e **H2** são satisfeitas e (x_k) é a sequência gerada pelo Algoritmo 3, então (x_k) converge para um $x^* \in \mathbb{R}^n$ tal que $F(x^*) = 0$.*

Demonstração. Primeiro vamos mostrar que a sequência $||x_{k+1} - x_k||$ converge para 0. De (31) temos:

$$\langle F(z_k), x_k - z_k \rangle = -\gamma_k \langle F(z_k), d_k \rangle \geq \sigma \gamma_k^2 ||d_k||^2 > 0. \tag{36}$$

Agora usando um argumento já visto no Teorema 3.2.2, seja x^* um ponto qualquer tal que $F(x^*) = 0$. Então, de (32), (36) e do Lema 3.2.1 segue que

$$||x_{k+1} - x^*||^2 \leq ||x_k - x^*||^2 - ||x_{k+1} - x_k||^2. \tag{37}$$

Portanto como já visto no Teorema 3.2.2 (x_k) é limitada e:

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

Veja que pelo Lema 3.3.2 θ_k é limitada, logo:

$$||d_k|| = \theta_k ||F(x_k)|| \leq \frac{1}{r} ||F(x_k)||.$$

Portanto (d_k) é limitada, consequentemente (z_k) também será, e com isso caímos numa situação análoga ao Teorema 3.2.2 e seguindo de forma similar a demonstração daquele teorema, após a equação (28), chegamos no resultado desejado. □

Neste capítulo estudamos três métodos que são "matrix-free", isto é, não demandam a resolução de sistemas lineares em cada iteração. O principal custo por iteração é o de uma ou mais avaliações do operador F, a depender do número de backtrackings (reduções do tamanho de passo) até que a condição da busca linear seja satisfeita. Infelizmente todos dependem de uma escolha externa de parâmetros, por exemplo, η_k no Algortimo 1 e r no Algoritmo 3, e tal escolha pode impactar a performance dos mesmos. Para uma investigação preliminar destas questões experimentos numéricos foram realizados e serão descritos no Capítulo 4.

Vale notar também que a teoria de convergência do Algoritmo 2 não pede a hipótese do operador ser Lipschitz contínuo (apenas contínuo), e por isso ele pode ser aplicado a uma classe mais ampla de equações monótonas.

Com isso concluímos o estudo teórico dos algoritmos para sistemas não lineares monótonos. No próximo capítulo iremos focar na implementação computacional, e fazer uma comparação dos algoritmos estudados com algoritmos genéricos para solução de sistemas não lineares – que não utilizam a monotonia do operador como ferramenta para chegar na solução. Este estudo numérico considera não apenas problemas teste da literatura como também duas aplicações em processamento de sinais, onde serão comparados com um algoritmo específico para este tipo de problema.

4 EXPERIMENTOS NUMÉRICOS

Neste capítulo serão apresentados experimentos numéricos a fim de analisar a eficiência dos algoritmos estudados. Além dos algoritmos DFSANME (Algoritmo 1) e o Algoritmo 3 que denominaremos por projexpec, que foram discutidos nos capítulos anteriores, também iremos considerar o algoritmo "Trust-region-dogleg" (MORÉ; SORENSEN, 1983) (implementado na rotina *fsolve* do Matlab/Octave para resolução de sistemas de equações não-lineares e problemas de quadrados mínimos não-lineares), o algoritmo FISTA (BECK; TEBOULLE, 2009) para problemas convexos e o algoritmo SGCS (XIAO; WANG; HU, 2011) muito semelhante ao Algoritmo 3, mudando apenas a escolha θ_{k+1} que será comentada no futuro.

Os experimentos foram divididos em três conjuntos de testes, o primeiro são problemas testes da literatura propostos em (LA CRUZ, 2017). O segundo consideramos um problema de processamento de sinais, mais especificamente na recuperação de sinais esparsos, e o terceiro trata do problema de deblurring de imagens, ambos problemas podem ser encontrados em (XIAO; WANG; HU, 2011). Vale ressaltar que para esses testes, não mostramos que a hipótese **H3** é satisfeita para aplicar o Algoritmo 1.

4.1 DETALHES DA IMPLEMENTAÇÃO

Todos os testes foram feitos no Octave e em um computador Intel Core i5 2.90GHz com 16GB de RAM, sistema operacional Windows10. Para todos os algoritmos definimos o número máximo de iterações como 10000. Os demais parâmetros algoritmicos são especificados nas seções a seguir.

4.2 PROBLEMAS TESTES DA LITERATURA

Nesta seção vamos considerar cinco problemas do tipo $F(x) = 0$ em que $F(x) = (f_1(x), ..., f_n(x))^T$ e $f_i : \mathbb{R}^n \to \mathbb{R}$ para $i = 1, \ ... \ , n$, com F monótono, extraídos de (LA CRUZ, 2017), para diferentes dimensões n e pontos iniciais x_0.

Foram usados os pontos iniciais $x_0^{(1)} = (10, ..., 10)^T$, $x_0^{(2)} = (1, ..., 1)^T$ e $x_0^{(3)} = (-1, ..., -1)^T$, e o critério de parada sendo $||F(x)|| \leq 10^{-4}$.

Para esses problemas os parâmetros usados no Algoritmo 1 foram: $\gamma = 10^{-4}$, $\sigma = 0.5$, $\alpha_0 = 1$, $\alpha_{max} = 4.5 \times 10^5$ e $\eta_k = (0.999)^k(10^5 + ||F(x_0)||^2)$ para $k \geq 0$.

Para o Algoritmo 3 os parâmetros usados foram: $\theta_0 = 1$, $\beta = 0.5$, $\lambda = 0.01$, $r = 1$ e $r = 0.1$.

Comparamos esses algoritmos com o fsolve que teve como parâmetros o critério de parada e o máximo de iterações mencionados no começo deste capítulo, o algoritmo utlizado foi o padrão "Trust-region-dogleg", a Jacobiana do operador não foi passada ao fsolve, e com isso ele calculou as derivadas utilizando diferenças finitas com o tamanho de

		fsolve			Alg. 1			Alg. 3 ($r = 1$)			Alg. 3 ($r = 0.1$)		
n	x_0	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)
	$x_0^{(1)}$	626	313126	24.9	982	982	0.35	2979	11810	1.13	**	**	**
500	$x_0^{(2)}$	5	2005	0.15	504	504	0.18	1175	4113	0.41	1141	6478	0.57
	$x_0^{(3)}$	7	3007	0.22	994	994	0.36	1191	4153	0.42	1161	6485	0.56
	$x_0^{(1)}$	**	**	**	5063	11685	5.9	15143	60289	14.9	18355	100742	22.3
2500	$x_0^{(2)}$	4	7504	1.5	4670	10635	5.3	3307	12656	3.1	3804	21296	4.6
	$x_0^{(3)}$	11	25011	5.1	5063	12994	6.14	3312	12683	3.1	3854	21435	4.7

Tabela 1 – Resultados para o Problema 1

passo padrão da rotina.

Os resultados para cada problema foram organizados em forma de tabela, apresentando o número de variáveis n, o ponto inicial x_0, o número de iterações necessárias para atingir o critério de parada denotada por "IT" o número de avaliações da função representado por "F", e o tempo em segundos representado por "$t(s)$". Se em um determinado problema um algoritmo demandou mais que 90 segundos para chegar no critério de parada, consideramos uma falha, indicado nas tabelas por "**".

Abaixo listamos cada um dos problemas testes considerados, acompanhado por uma tabela com resultados e em seguida comentários.

Problema 1 As funções não-lineares que definem o problema são dadas por:

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

Pela Tabela 1 podemos perceber que para o ponto inicial $x_0^{(1)}$ o fsolve se mostra bem menos eficiente, enquanto para outros pontos o tempo é até menor que os outros. Isso nos mostra o quanto a escolha do ponto inicial no fsolve pode influenciar no desempenho do algoritmo. Isso provavelmente vem do fato que as derivadas do seno estão sendo calculadas por diferenças finitas e 10 é um numero longe de zero, onde se tem a melhor aproximação para a linearização.

Problema 2 Neste problema as funções não-lineares não são diferenciáveis em toda parte:

$$f_i(x) = 2x_i - sen(|x_i|) \quad i = 1, ..., n.$$

		fsolve			Alg. 1			Alg. 3 (r = 1)			Alg. 3 (r = 0.1)		
n	x_0	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)
	$x_0^{(1)}$	8	3508	0.21	7	7	0.005	21	44	0.005	10	23	0.002
500	$x_0^{(2)}$	6	2506	0.15	6	6	0.001	18	37	0.01	7	15	0.001
	$x_0^{(3)}$	3	1003	0.06	7	7	0.002	10	22	0.002	6	15	0.002
	$x_0^{(1)}$	8	17508	2.7	7	7	0.006	22	46	0.012	10	23	0.006
2500	$x_0^{(2)}$	6	12506	1.9	6	6	0.004	19	39	0.01	7	15	0.004
	$x_0^{(3)}$	2	2502	0.44	7	7	0.005	11	24	0.006	6	15	0.003
	$x_0^{(1)}$	8	35008	10.04	8	8	0.01	23	48	0.02	10	23	0.01
5000	$x_0^{(2)}$	6	25006	7.03	6	6	0.007	20	41	0.019	7	15	0.007
	$x_0^{(3)}$	2	5002	1.6	7	7	0.009	11	24	0.01	6	15	0.006

Tabela 2 – Resultados para o Problema 2

Da Tabela 2 podemos observar que mesmo com o aumento na dimensão n os Algoritmos 1 e 3 continuam bem eficientes, enquanto o fsolve começa a ficar mais lento, pois como a Jacobiana não é dada, uma consequência disso é o número de vezes que o algoritmo avalia o operador por iteração aumenta junto com o tamanho da dimensão, o tornando computacionalmente mais custoso e como consequência mais lento.

Problema 3 O operador é dado por:

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

Na Tabela 3, em termos de tempo, a diferença dos algoritmos em relação ao fsolve é bem aparente, e podemos observar também que o Algoritmo 3 demanda muito mais avaliações de função e iterações que o Algoritmo 1 o tornando mais custoso.

Problema 4 O operador é dado por:

$$f_i(x) = x_i - \frac{1}{n}x_i^2 + \frac{1}{n}\sum_{k=0}^{n} x_k + i, \quad i = 1, \ldots, n.$$

		fsolve			Alg. 1			Alg. 3 (r = 1)			Alg. 3 (r = 0.01)		
n	x_0	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)
	$x_0^{(1)}$	1561	781561	38.5	582	869	0.12	5705	12066	0.92	**	**	**
500	$x_0^{(2)}$	126	63127	3.1	99	234	0.12	4337	8951	0.69	**	**	**
	$x_0^{(3)}$	16	7516	0.3	20	27	0.004	768	1823	0.13	175	883	0.04
	$x_0^{(1)}$	**	**	**	166	293	0.08	6288	14584	2.01	**	**	**
2500	$x_0^{(2)}$	302	752802	79.15	669	1050	0.29	5191	11215	1.6	**	**	**
	$x_0^{(3)}$	16	37516	3.8	23	32	0.01	949	2711	0.35	399	2413	0.23
	$x_0^{(1)}$	**	**	**	349	510	0.25	6106	14980	3.3	**	**	**
5000	$x_0^{(2)}$	**	**	**	132	265	0.10	2811	7238	1.5	**	**	**
	$x_0^{(3)}$	16	75016	13.9	25	35	0.01	1137	3955	0.8	541	3411	0.52

Tabela 3 – Resultados para o Problema 3

		fsolve			Alg. 1			Alg. 3 (r= 1)			Alg. 3 (r = 0.1)		
n	x_0	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)
	$x_0^{(1)}$	9	4009	0.26	22	23	0.007	71	174	0.01	67	248	0.01
500	$x_0^{(2)}$	16	7516	0.48	22	23	0.005	81	197	0.01	64	229	0.01
	$x_0^{(3)}$	16	7516	0.50	22	23	0.005	80	195	0.01	71	262	0.01
	$x_0^{(1)}$	13	30013	4.42	27	28	0.009	103	251	0.03	80	291	0.02
2500	$x_0^{(2)}$	20	47520	6.96	27	28	0.009	101	248	0.03	80	294	0.02
	$x_0^{(3)}$	20	47520	6.73	27	28	0.009	99	243	0.03	85	312	0.03
	$x_0^{(1)}$	15	70015	20.3	27	28	0.01	110	269	0.04	86	313	0.04
5000	$x_0^{(2)}$	21	100021	29.1	27	28	0.01	112	273	0.04	84	312	0.04
	$x_0^{(3)}$	21	100021	29.7	27	28	0.01	113	275	0.04	83	307	0.04

Tabela 4 – Resultados para o Problema 4

Das Tabelas 4 e 5 podemos concluir o mesmo que nos problemas anteriores.

Problema 5 O operador é linear $F(x) = Ax + b$, tal que $b = (-1, -1, ..., -1)^T$, e

		fsolve			Alg. 1			Alg. 3 (r = 1)			Alg. 3 (r = 0.1)		
n	x_0	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)	IT	F	t(s)
	$x_0^{(1)}$	2	502	0.03	20	21	0.005	15	36	0.003	60	217	0.01
500	$x_0^{(2)}$	2	502	0.03	16	16	0.003	11	26	0.002	23	71	0.005
	$x_0^{(3)}$	3	1003	0.06	17	18	0.004	15	39	0.004	42	154	0.01
	$x_0^{(1)}$	2	2502	0.36	20	21	0.008	14	33	0.004	67	252	0.02
2500	$x_0^{(2)}$	2	2502	0.36	14	15	0.006	11	26	0.003	37	134	0.01
	$x_0^{(3)}$	3	5003	0.74	17	18	0.007	13	32	0.004	54	210	0.02
	$x_0^{(1)}$	2	5002	1.4	20	21	0.01	16	39	0.008	74	276	0.04
5000	$x_0^{(2)}$	2	5002	1.4	14	15	0.009	11	26	0.006	31	112	0.02
	$x_0^{(3)}$	3	10003	2.9	17	18	0.01	13	31	0.007	58	221	0.03

Tabela 5 – Resultados para o Problema 5

$$A = \begin{pmatrix} \frac{5}{2} & 1 & & & & & & & \\ 1 & \frac{5}{2} & 1 & & & & & & \\ & & \cdot & \cdot & \cdot & & & & \\ & & & \cdot & \cdot & \cdot & & & \\ & & & & \cdot & \cdot & \cdot & & \\ & & & & & 1 & \frac{5}{2} & 1 & \\ & & & & & & 1 & \frac{5}{2} \end{pmatrix}.$$

Das tabelas e dos comentários feitos podemos tirar uma conclusão bem aparente: a rotina fsolve demanda a avaliação da função $F(x)$ muito mais vezes (a cada iteração, em n pontos da forma $x + he_i$), que consequentemente o torna mais lento e mais custoso que os outros dois algoritmos que não usam (e nem tentam aproximar) a derivada. Mesmo em problemas nos quais o fsolve precisou de menos iterações que os outros, por calcular a função muitas vezes ele ainda é mais lento. Portanto, apenas das tabelas, conseguimos deduzir que o fsolve é o algoritmo menos eficiente neste conjunto de problemas teste – supondo que a Jacobiana não é fornecida.

A fim de fornecer um panorama geral dos algoritmos considerando todos os exemplares de problemas considerados, também traçamos o perfil de desempenho (DOLAN; MORÉ, 2002), como apresentado na Figura 1.

Tal gráfico é organizado da seguinte forma: no eixo horizontal temos τ, que é calculado para cada problema e dimensão, tomando o tempo t_1, t_2, t_3 que os três algoritmos levaram para chegar na solução, e dividindo pelo menor tempo $t_0 = min\{t_1, t_2, t_3\}$ entre os tempos obtidos, ou seja:

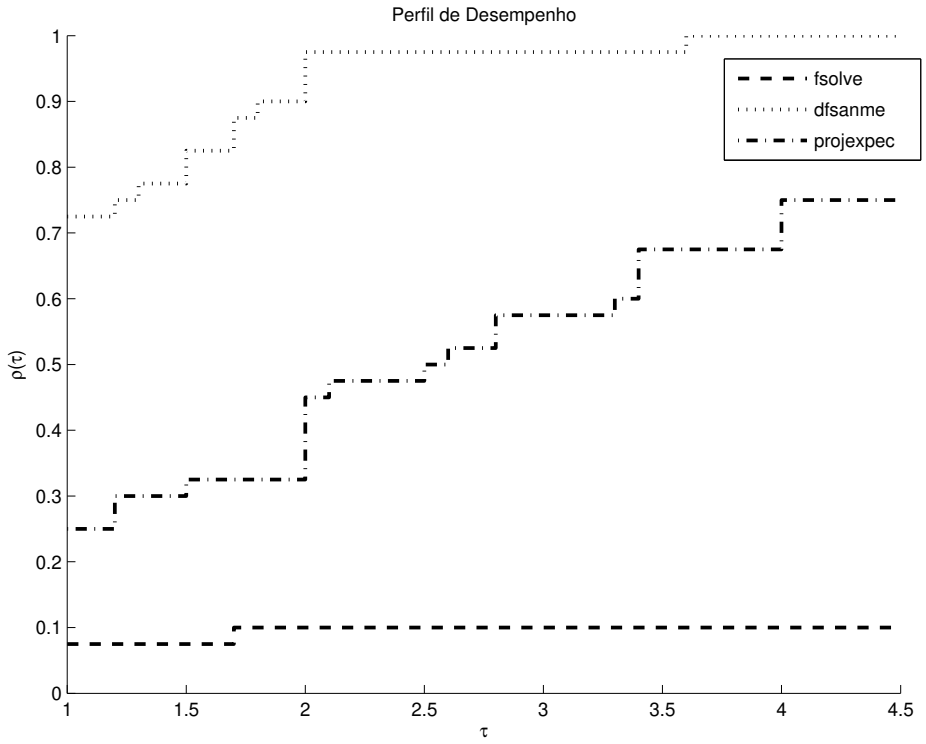


Figura 1 – Perfil de desempenho.

$$\tau_i = \frac{t_i}{t_0} \quad \text{tal que } \; i = 1, 2, 3.$$

No eixo vertical temos $\rho(\tau)$ que para cada τ nos diz a porcentagem de problemas que o algoritmo resolveu em uma proporção τ comparado ao menor tempo, ou seja, em $\tau = 1$, $\rho(\tau)$ nos diz a porcentagem de problemas que o algoritmo resolveu no menor tempo entre os três – dizemos que o algoritmo com maior valor $\rho(1)$ é o mais eficiente. Além disso, aquele que atinge $\rho(\tau) = 1$ para o menor valor de τ é dito o mais robusto.

Do gráfico da Figura 1 é fácil de ver que o fsolve realmente é o que tem a pior performance, mas apenas das tabelas estava difícil de concluir qual dos dois algoritmos restantes é mais eficiente. Com esse gráfico em mãos agora é fácil perceber que o dfsanme é o mais eficiente, e também o mais robusto, entre os três com uma boa margem (para este conjunto de problemas teste).

4.3 RECUPERAÇÃO DE SINAL ESPARSO

Nesta seção apresentaremos um modelo que transforma um problema de processamento de sinais, conhecido como recuperação de sinal esparso, em um sistema não linear monótono como proposto no artigo (XIAO; WANG; HU, 2011).

4.3.1 Redefinindo o problema

No problema de recuperação de sinal uma abordagem comum é buscar a solução mais esparsa (maior número possível de componentes nulas) compatível com um certo conjunto de medições (observações). Este modelo é baseado em uma medição (observação) linear, na qual é medido (observado) um vetor $b \in \mathbb{R}^m$ tal que $b = A\bar{x}$ com $\bar{x}$ sendo o sinal original, e A é uma "matriz de medida" $m \times n$ ($m \leq n$) que representa um operador.

Para isso, poderíamos minimizar a $\|x\|_0$ (número de componentes não-nulas de x) sujeito a $Ax = b$. Mas esse problema possui complexidade computacional NP-difícil (NATARAJAN, 1995). Com isso um modelo alternativo é utilizar a norma-1 (DONOHO, D. L., 2006) e resolver o problema de "Basis-Persuit (BP)" (CHEN; DONOHO, D. L.; SAUNDERS, 2001):

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \|x\|_1 \\ \text{s.a} \quad & Ax = b. \end{aligned} \tag{38}$$

A equação $Ax = b$ é muitas vezes relaxada, sobretudo quando as medições b contêm ruído, usando uma função de penalidade quadrática $h(x) = ||Ax - b||_2^2$. Com isso podemos considerar o problema:

$$\min_{x \in \mathbb{R}^n} \quad \tau\|x\|_1 + \frac{1}{2}||Ax - b||_2^2 \tag{39}$$

o qual é equivalente a (38) para um valor adequado de $\tau > 0$ (detalhes em (BRUCKSTEIN; DONOHO, D.; ELAD, 2009)).

Agora, escrevendo $x = u - v$ tal que $u, v \in \mathbb{R}^n$, $u \geq 0, v \geq 0$ e $u_i = (x_i)_+$, $v_i = (-x_i)_+$ para todo $i = 1, ..., n$ com $(\cdot)_+ = max\{0, \cdot\}$, podemos expressar a ℓ_1-norma como $||x||_1 = e_n^T u + e_n^T v$ onde e_n é o vetor com todas entradas iguais a 1. Com isso podemos reescrever o problema novamente como:

$$\begin{aligned} \min_{u,v \in \mathbb{R}^n} \quad & \frac{1}{2}||b - A(u - v)||^2 + \tau e_n^T u + \tau e_n^T v \\ \text{s.a} \quad & u \geq 0 \\ & v \geq 0. \end{aligned} \tag{40}$$

Definindo

$$z = \begin{bmatrix} u \\ v \end{bmatrix}, \; y = A^T b, \quad c = \tau e_{2n} + \begin{bmatrix} -y \\ y \end{bmatrix}, \; H = \begin{bmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{bmatrix}$$

podemos mostrar que (40) equivale a

$$\begin{aligned} \min_{z \in \mathbb{R}^{2n}} \quad & \frac{1}{2} z^T H z + c^T z \\ \text{s.a} \quad & z \geq 0. \end{aligned} \tag{41}$$

Com efeito, primeiro note que

$$\begin{aligned} z^T Hz &= \begin{bmatrix} u^T & v^T \end{bmatrix} \begin{bmatrix} A^TA & -A^TA \\ A^TA & A^TA \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \\ &= \begin{bmatrix} u^T & v^T \end{bmatrix} \begin{bmatrix} A^TAu - A^TAv \\ A^TAv - A^TAu \end{bmatrix} \\ &= u^TA^TAu - 2u^TA^TAv + v^TA^TAv \\ &= u^TA^TAu - 2u^TA^TAv + v^TA^TAv \\ &= ||Au - Av||_2^2 \\ &= ||A(u-v)||_2^2 \geq 0. \end{aligned} \tag{42}$$

Com isso temos que H é semidefinida positiva. Agora calculando c^Tz temos

$$\begin{aligned} \begin{bmatrix} \tau e_n^T - y^T & \tau e_n^T + y^T \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} &= \tau e_n^T u - y^T u + \tau e_n^T v + y^T v \\ &= b^TAv - b^TAu + \tau e_n^T u + \tau e_n^T v \\ &= b^T(-A(u-v)) + \tau e_n^T u + \tau e_n^T v \end{aligned} \tag{43}$$

e assim

$$\frac{1}{2}z^THz + c^Tz = \frac{1}{2}||A(u-v)||_2^2 + b^T(-A(u-v)) + \tau e_n^T u + \tau e_n^T v.$$

Como

$$\frac{1}{2}||b - A(u-v)||^2 = \frac{1}{2}\|b\|^2 - b^TA(u-v) + \frac{1}{2}\|A(u-v)\|^2,$$

e o termo $\frac{1}{2}\|b\|^2$ independe de u, v, provamos a equivalência desejada.

Note que se $z \geq 0$ é solução de (41) temos que $\forall w \in \mathbb{R}^{2n}$ as seguintes desigualdes são satisfeitas:

$$\begin{aligned} \frac{1}{2}w^THw + c^Tw &\geq \frac{1}{2}z^THz + c^Tz \\ w^THw + 2c^Tw &\geq z^THz + 2c^Tz \\ w^THw + 2c^Tw - z^THz - 2c^Tz &\geq 0 \\ \langle Hw, w\rangle + \langle c, 2w\rangle - \langle Hz, z\rangle - \langle c, 2z\rangle &\geq 0 \\ \langle Hz + c, 2w\rangle + \langle Hz + c, -2z\rangle &\geq 0 \\ \langle Hz + c, w\rangle + \langle Hz + c, -z\rangle &\geq 0 \\ \langle Hz + c, w - z\rangle &\geq 0 \end{aligned} \tag{44}$$

Logo encontrar $z \in \mathbb{R}^{2n}$ que é solução de (41) é o mesmo que encontrar $z \in \mathbb{R}^{2n}$ tal que $\langle Hz + c, w - z\rangle \geq 0 \ \forall w \geq 0$.

Como temos $\langle Hz + c, w\rangle - \langle Hz + c, z\rangle \geq 0$, e a região viável de $z \in \mathbb{R}^{2n}$ é $z \geq 0$, temos que se $Hz + c \geq 0$, $\langle Hz + c, z\rangle \geq 0$ e como w pode ser o vetor nulo, o problema (44) é equivalente a encontrar $z \in \mathbb{R}^{2n}$ tal que

$$z \geq 0, \quad Hz + c \geq 0 \quad \text{e} \quad \langle Hz + c, z\rangle = 0. \tag{45}$$

Portanto z é solução de (45) se e somente se z satifaz

$$F(z) = \min\{Hz + c, z\} = 0. \tag{46}$$

Em que F é um operador que calcula o mínimo em cada componente entre os dois vetores. Agora basta mostrar que o operador é monótono e Lipschitz para podermos aplicar os algoritmos estudados nesse tipo de problema.

4.3.2 O operador é monótono e Lipschitz contínuo

Lema 4.3.1. *O operador $F(z) = \min\{Hz + c, z\}$ em que $z \in \mathbb{R}^{2n}$ é Lipschitz.*

Demonstração. Sejam $p, q \in \mathbb{R}^{2n}$ e $p, q \geq 0$. Logo

$$||F(p) - F(q)|| = ||\min\{p, Hp + c\} - \min\{q, Hq + c\}||.$$

Divindo em casos para cada $i \in \{1, ..., 2n\}$ temos:
Caso(i): $\min\{p_i, (Hp + c)_i\} = p_i$ e $\min\{q_i, (Hq + c)_i\} = q_i$:

$$\begin{aligned} |\min\{p_i, (Hp + c)_i\} - \min\{q_i, (Hq + c)_i\}| &= |p_i - q_i| \\ &\leq ||p - q||_\infty. \end{aligned}$$

Caso(ii): $\min\{p_i, (Hp + c)_i\} = (Hp + c)_i$ e $\min\{q_i, (Hq + c)_i\} = (Hq + c)_i$

$$\begin{aligned} |\min\{p_i, (Hp + c)_i\} - \min\{q_i, (Hq + c)_i\}| &= |(Hp + c)_i - (Hq + c)_i| \\ &= |(Hp)_i - (Hq)_i| \\ &= |H(p - q)_i| \\ &\leq ||H||_\infty \, ||p - q||_\infty. \end{aligned}$$

Caso(iii): $\min\{p_i, (Hp + c)_i\} = p_i$ e $\min\{q_i, (Hq + c)_i\} = (Hq + c)_i$

$$|\min\{p_i, (Hp + c)_i\} - \min\{q_i, (Hq + c)_i\}| = |p_i - (Hq + c)_i|.$$

Para esse caso vamos primeiro considerar os índices i tais que $(Hq - c)_i > p_i$. Com isso obtemos

$$\begin{aligned} |p_i - (Hq + c)_i| = |(Hq + c)_i - p_i| &= (Hq + c)_i - p_i \\ &< q_i - p_i \\ &\leq |q_i - p_i| \\ &\leq ||p - q||_\infty. \end{aligned}$$

Agora para os índices em que $(Hq + c)_i < p_i$ temos

$$\begin{aligned} |p_i - (Hq + c)_i| &= p_i - (Hq + c)_i \\ &< (Hp + c)_i - (Hq + c)_i \\ &\leq |(Hp + c)_i - (Hq + c)_i| \\ &= |H(p - q)_i| \\ &\leq ||H||_\infty ||p - q||_\infty. \end{aligned}$$

Daí segue que

$$|p_i - (Hq + c)_i| \leq \lambda ||p - q||_\infty,$$

em que $\lambda = max\{||H||_\infty, 1\}$.

Caso(iv): análogo ao caso (iii).

Com isso obtemos limitantes para todo $i \in \{1, ..., 2n\}$ de $|(F(p) - F(q))_i|$. Logo

$$||F(p) - F(q)||_\infty \leq \lambda ||p - q||_\infty \leq \lambda ||p - q||.$$

Mas como em dimensão finita as normas são equivalentes, segue que

$$||F(p) - F(q)|| \leq \sqrt{n}||F(p) - F(q)||_\infty \leq \sqrt{n}\lambda ||p - q||$$

Portanto F é Lipschitz contínuo, com a constante de Lipschitz $\ell = max\{\sqrt{n}||H||_\infty, \sqrt{n}\}$. □

Lema 4.3.2. *Seja $z \in \mathbb{R}^{2n}$ operador $F(z) = \min\{Hz + c, z\}$. Se existe $q \in \mathbb{R}^{2n}$ tal que $F(q) = 0$, então F é monótono.*

Demonstração. Seja $p \in \mathbb{R}^{2n}$ e $q \in \mathbb{R}^{2n}$ tal que $F(q) = 0$. Tomando o m definido no Lema 2.3.2 temos que para $i = 1, ..., 2n$

$$\begin{aligned} min(p_i, (Hp + c)_i) - min(q_i, (Hq + c)_i) &= (1 - m_i)(p_i - q_i) \\ &\quad + m_i((Hp + c)_i - (Hq + c)_i). \end{aligned} \tag{47}$$

Note que fazendo $M \in \mathbb{R}^{2n \times 2n}$ uma matriz diagonal em que sua diagonal é $(m_1, ..., m_{2n})$ e $I \in \mathbb{R}^{2n \times 2n}$ a matriz identidade, podemos escrever a equação (47) da seguinte forma

$$\begin{aligned} F(p) - F(q) &= (I - M)(p - q) + M(Hp + c - Hq - c) \\ F(p) &= (I - M)(p - q) + MH(p - q) \\ F(p) &= (I - M + MH)(p - q). \end{aligned} \tag{48}$$

Logo, a equação (48) nos diz que para qualquer $w \in \mathbb{R}^{2n}$

$$F(w) = (I - M + MH)(w - q). \tag{49}$$

Finalmente

$$\begin{aligned}\langle F(p) - F(w), p - w\rangle &= \langle (I - M + MH)(p - q) \\ &\quad - (I - M + MH)(w - q), p - w\rangle \\ &= \langle (I - M + MH)(p - w), p - w\rangle\end{aligned}$$

e como $I - M$ é uma matriz diagonal com os elementos da diagonal maiores ou iguais a zero e H é uma matriz semidefinida positiva, a matriz MH também será. Logo $(I - M + MH)$ é semidefinida positiva, assim temos $\langle (I - M + MH)(p - w), p - w\rangle \geq 0$, como p e w são vetores quaisquer de $\mathbb{R}^{2n}$, segue que F é monótono. □

Tendo em mente, como dito no começo deste capítulo, que não estaremos mostrando a hipótese **H3** para os testes com o Algoritmo 1, podemos aplicar os algoritmos estudados para o problema de processamento de sinais, como apresentado a seguir.

4.3.3 Parâmetros utilizados no problema de processamento de sinais

Para o problema de processamento de sinais foi usado o ponto inicial $x_0 = (1, ..., 1)^T$. Para o critério de parada desse problema consideramos o utilizado anteriormente nos problemas teste da literatura e também o proposto no artigo (XIAO; WANG; HU, 2011):

$$\frac{|f_k - f_{k-1}|}{|f_{k-1}|} < 10^{-5}$$

em que

$$f_k = f(x_k) = \tau||x_k||_1 + \frac{1}{2}||Ax_k - b||_2^2.$$

Para esse problema os parâmetros usados no Algoritmo 1 foram: $\gamma = 10^{-4}$, $\sigma = 0.5$, $\alpha_0 = 1$, $\alpha_{max} = 4.5 \times 10^5$ e $\eta_k = (0.999)^k(10^5 + ||F(x_0)||^2)$ para $k \geq 0$.

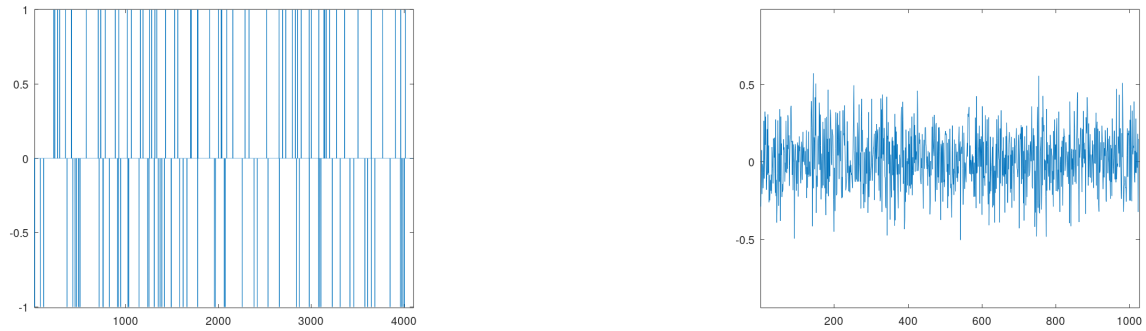
Para o Algoritmo 3 fizemos uma pequena mudança no passo 5, para ficar igual ao algoritmo SGCS proposto em (XIAO; WANG; HU, 2011), mudando apenas a escolha de $\theta_{k+1} = \frac{\langle s_k, s_k\rangle}{\langle s_k, y_k\rangle}$ por $\theta_{k+1} = \max\{\theta_{min}, \min(\hat{\theta}_k, \theta_{max})\}$, com $\hat{\theta}_k = \frac{\langle s_k, s_k\rangle}{\langle s_k, y_k\rangle}$. Os parâmetros foram: $\theta_0 = 1$, $\theta_{min} = 10^{-4}$, $\theta_{max} = 10^{10}$, $\beta = 0.1$, $\lambda = 1.2$, $r = 10^{-4}$.

Nestes experimentos consideramos também o algoritmo FISTA (BECK; TEBOULLE, 2009), usando um valor $L \geq ||A||$ como uma estimativa para a constante de Lipschitz associada ao gradiente da função $\frac{1}{2}||Ax - b||_2^2$.

Para os testes consideramos $\tau = 10^{-2}$, e b o sinal contaminado com ruído da forma

$$b = A\bar{x} + w$$

em que w é um vetor de ruído Gaussiano, cujas entradas vem da distribuição $N(0, \sigma^2 I)$ com $\sigma = 10^{-4}$ e $\bar{x}$ é o vetor que representa o sinal original.



(a) sinal original (b) medição com ruído

Figura 2 – Sinal original e medição com ruído

	dfsanme			SGCS			FISTA		
x_0	IT	MSE	t(s)	IT	MSE	t(s)	IT	MSE	t(s)
x_0^1	457	1.7e-04	3.2	1564	2.3e-04	16.08	104	3.2e-04	1.31
x_0^2	134	1.7e-04	0.94	940	2.3e-04	10.7	73	2.5e-04	1.7

Tabela 6 – Resultados para reconstrução do sinal esparso

4.4 SINAIS ESPARSOS

Primeiro consideramos o problema de reconstruir um sinal esparso de tamanho n a partir de m observações. Testamos com um sinal de tamanho pequeno com $n = 2^{12}$ e $m = 2^{10}$, com o sinal original possuindo 2^7 elementos aleatórios não nulos. A matriz A é a matriz Gaussiana, em que seus elementos são gerados através de uma distribuição normal $N(0, 1)$ ($randn(m, n)$ no Matlab). Nesse problema além de utilizar o ponto inicial $x_0^1 = (1, ..., 1)^T$ também testamos com o ponto $x_0^2 = (0, ..., 0)^T$.

Para medir a qualidade da restauração da imagem utilizamos o MSE (mean squared error que pode ser estudado em (STATHAKI, 2008, Capítulo 5)) definido como

$$MSE(x^*; \bar{x}) = \frac{1}{n}||\bar{x} - x^*||^2$$

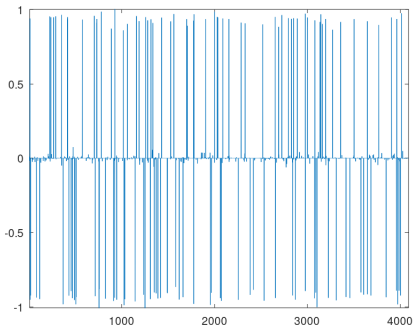
em que x^* é o sinal restaurado e $\bar{x}$ é o sinal original.

Organizamos os resultados em forma de tabela, comparando para atingir algum critério de parada o número de iterações "IT", o tempo "$t(s)$" e o "MSE" do iterado final.

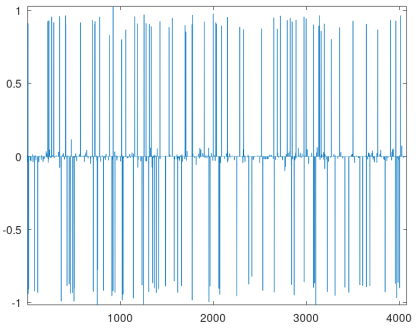
Da Tabela 6 podemos concluir que o FISTA é o algoritmo que demanda menos iterações, e o SGCS demanda um tempo muito maior que os outros. Das Figuras 2, 3 e do valor do MSE percebemos que todos os três reconstroem um sinal semelhante com uma diferença de MSE na ordem de 10^{-5}.

4.5 RECUPERAÇÃO DE IMAGENS BORRADAS

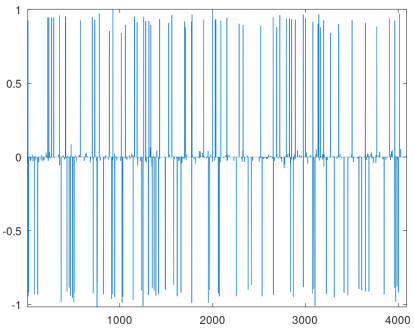
Para o segundo experimento consideramos o problema de reconstrução de imagem, que utiliza o modelo visto na Seção 4.3.1, em que A é um operador de blur Gaussiano e b



(a) DFSANME



(b) FISTA



(c) SGCS

Figura 3 – Sinais recuperados

	dfsanme			SGCS			FISTA		
imagem $n \times n$	IT	SNR	t(s)	IT	SNR	t(s)	IT	SNR	t(s)
cameraman 256×256	15	23.49	0.37	531	24.81	22.12	1263	24.95	14.09
lena 256×256	34	23.54	0.84	445	23.33	18.77	1514	22.8	17.04

Tabela 7 – Resultados para reconstrução da imagem

é a imagem borrada (blured) e com ruído.

Para medir a qualidade da restauração da imagem utilizamos o SNR (signal-to-noise ratio que pode ser estudado em (STATHAKI, 2008, Capítulo 19)) definido como

$$SNR(x^*, \bar{x}) = 20 \log_{10} \left(\frac{||\bar{x}||}{||x * - \bar{x}||} \right)$$

em que x^* é o sinal restaurado e $\bar{x}$ é o sinal original.

Além de comparar as imagems entre si visualmente, os resultados foram organizados em forma de tabela, apresentando a imagem e a sua dimensão $n \times n$, o número de iterações necessárias para atingir o critério de parada denotado por "IT", o "SNR", e o tempo em segundos representado por "$t(s)$".

Neste teste, como a matriz A é desconhecida no problema de processamento de imagem, por ser um operador de blur, utilizamos uma constante grande o suficiente $L = 1000$.

Dos resultados numéricos da Tabela 7, observa-se que o dfsanme é o mais eficiente

(a) Imagem original

(b) Imagem com ruído

Figura 4 – Imagem original e borrada

(a) DFSANME

(b) FISTA

(c) SGCS

Figura 5 – Imagens restauradas

(a) imagem original

(b) imagem com ruído

Figura 6 – Imagem original e borrada

(a) DFSANME

(b) FISTA

(c) SGCS

Figura 7 – Imagem restaurada

por alcançar o critério com menos iterações e tempo, apesar de reconstruir uma imagem com um SNR um pouco menor. Comparando os outros dois algoritmos vemos que SGCS atinge o critério de parada em menos iterações, embora precise de mais tempo que o FISTA. O fato de FISTA demandar um número maior de iterações pode ser consequência de não conhecermos a constante de Lipschitz do operador, e por isso termos utilizado uma constante muito grande, o que leva a um tamanho de passo muito pequeno.

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho realizamos um estudo sobre métodos iterativos para solução de sistema não lineares monótonos. Uma vantagem dos métodos propostos é que eles não dependem do uso de derivadas, o que pode ser benéfico já que em algumas aplicações tal informação não está disponível (seja porque F não é diferenciável, ou porque $F'(x)$ é inacessível). Por isso consideramos a direção $d_k = -tF(x_k)$, em que t é um escalar positivo adequado.

Neste contexto, estudamos o algoritmo DFSANME proposto por La Cruz (LA CRUZ, 2017), que emprega uma busca não-monótona juntamente com o uso de um parâmetro espectral.

Consideramos também um algoritmo alternativo, de Solodov e Svaiter (SOLODOV; SVAITER, 1999), baseado em projeções sobre semi-espaços que contem o conjunto solução. Por fim, analisamos um algoritmo proposto em (YU *et al.*, 2009), que combina as duas estratégias mencionadas anteriormente.

Apresentamos de forma detalhada a boa definição e análise de convergência de todos os algoritmos.

Realizamos experimentos numéricos/computacionais a fim de comparar o desempenho dos algoritmos estudados e outros algoritmos mais genéricos. Para isso, resolvemos alguns problemas testes da literatura, e consideramos algumas aplicações em processamento de sinais. Em particular, descrevemos como o problema de recuperação de sinal esparso pode ser reescrito como um sistema não linear monótono. Com base nos resultados numéricos, o DFSANME (Algoritmo 1) demonstrou um desempenho superior aos demais algoritmos na maioria dos problemas considerados.

Como trabalhos futuros, consideramos fazer um estudo mais profundo sobre a motivação e teoria do parâmetro espectral, conceito utilizado em dois dos algoritmos estudados e estudar outros problemas envolvendo operadores monótonos como o problema de inclusão monótona.

REFERÊNCIAS

BAUSCHKE, H.; COMBETTES, P. **Convex Analysis and Monotone Operator Theory in Hilbert Spaces**. 2. ed. New York: Springer Publishing Company, Incorporated, 2017.

BECK, A.; TEBOULLE, M. **A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems**. Tel-Aviv University: SIAM Journal on Imaging Sciences, v. 2, n. 1, p. 183-202, 2009.

BERTSEKAS, D. **Nonlinear Programming**. 2. ed. Michigan: Athena Scientific, 2003.

BRUCKSTEIN, A.; DONOHO, D.; ELAD, M. **From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images**. [*S.l.*]: SIAM Review, v. 51, p.34-81, 2009.

CHEN, S. S.; DONOHO, D. L.; SAUNDERS, M. A. **Atomic Decomposition by Basis Pursuit**. Stanford University: SIAM Journal on Scientific Computing, v. 43, n. 1, p. 129-159, 2001.

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

DOLAN, E. D.; MORÉ, J. J. **Benchmarking optimization software with performance profiles**. [*S.l.*]: Mathematical Programming, v. 91, p. 201-213, 2002.

DONOHO, D. L. **For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution**. Stanford University: Communications on Pure e Applied Mathematics, v. 59, p. 797-829, 2006.

KHALIL, H.K. **Nonlinear Systems**. 3. ed. Upper Saddle River, New Jersey: Prentice Hall, 2002.

LA CRUZ, W. **A spectral algorithm for large-scale systems of nonlinear monotone equations**. Central University of Venezuela: Numerical Algorithms, v. 76, p. 1109-1130, 2017.

MORÉ, J. J.; SORENSEN, D. C. **Computing a Trust Region Step**. [*S.l.*]: SIAM Journal on Scientific e Statistical Computing, v. 4, p. 553-572, 1983.

NATARAJAN, B. K. **Sparse Approximate Solutions to Linear Systems**. Cornell University: SIAM Journal on Computing, v. 24, n. 2, p. 227-234, 1995.

ROCKAFELLAR, R. T.; WETS, R. J.B. **Variational Analysis**. 1. ed. Heidelberg, New York: Springer Verlag, 1998.

SMITH, S. W. **The scientist and engineer's guide to digital signal processing**. 1. ed. San Diego: California Technical Publishing, 1997.

SOLODOV, M. V.; SVAITER, B. F. A Globally Convergent Inexact Newton Method for Systems of Monotone Equations. *In*: REFORMULATION: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods. Boston, MA: Springer US, 1999. p. 355–369.

STATHAKI, Tania. **Image Fusion: Algorithms and Applications**. 1. ed. Oxford: Academic Press, Inc., 2008.

XIAO, Y.; WANG, Q.; HU, Q. **Non-smooth equations based method for -norm problems with applications to compressed sensing**. Henan University: Nonlinear Analysis, v. 74, p. 3570-3577, 2011.

YU, Z.; LIN, J.; SUN, J.; XIAO, Y.; LIU, L.; LI, Z. **Spectral gradient projection method for monotone nonlinear equations with convex constraints**. [*S.l.*]: Applied Numerical Mathematics, v. 59, p. 2416-2423, 2009.