universidade estadual de campinas
Instituto de Matemática Estatística e Computação Científica
Departamento de Matemática Aplicada

Sistemas Ponto de Sela com uma aplicação à aceleração do Lagrangiano Aumentado

Viviana Analía Ramirez

Mestrado em Matemática Aplicada - Campinas - SP

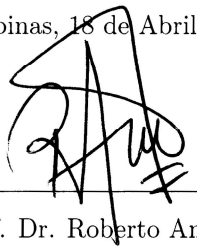
Orientador: Prof. Dr. Roberto Andreani

UNICAMP-IMECC

Sistemas Ponto de Sela com uma aplicação à aceleração do Lagrangiano Aumentado

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por **Viviana Analía Ramirez** e aprovada pela comissão julgadora.

Campinas, 18 de Abril de 2008



Prof. Dr. Roberto Andreani

Orientador

Banca Examinadora

1. Profa. Dra. María de los Angeles González Lima (CCE/USB/Venezuela)
2. Profa. Dra. Sandra Augusta Santos (IMECC/Unicamp/Brasil)
3. Prof. Dr. Roberto Andreani (IMECC/Unicamp/Brasil)

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, Unicamp, como requisito parcial para obtenção do Título de MESTRE em Matemática Aplicada.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Maria Júlia Milani Rodrigues

Ramirez, Viviana Analía
R145s Sistemas ponto de sela com uma aplicação à aceleração do Lagrangiano Aumentado / Viviana Analía Ramirez -- Campinas, [S.P. :s.n.], 2008.

Orientador : Roberto Andreani
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Sistemas ponto de sela. 2. Métodos numéricos. 3. Otimização matemática. I. Andreani, Roberto. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Título em inglês: Saddle point systems with an application to the acceleration of the Augmented Lagrangian.

Palavras-chave em inglês (Keywords): 1. Saddle point systems. 2. Numerical methods. 3. Mathematical optimization.

Área de concentração: Otimização

Titulação: Mestre em Matemática Aplicada

Banca examinadora:

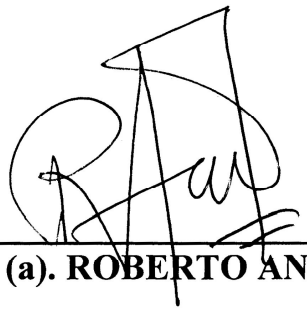
Prof. Dr. Roberto Andreani (IMECC-UNICAMP)
Profa. Dra. María de los Angeles González Lima (CCE/USB – Venezuela)
Profa. Dra. Sandra Augusta Santos (IMECC-UNICAMP)

Data da defesa: 18/04/2008

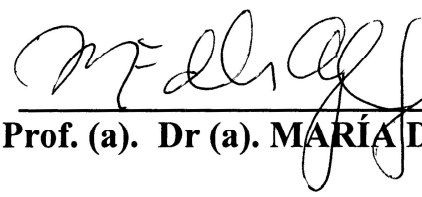
Programa de pós-graduação: Mestrado em Matemática Aplicada

Dissertação de Mestrado defendida em 18 de abril de 2008 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.



Prof. (a). Dr (a). ROBERTO ANDREANI



Prof. (a). Dr (a). MARÍA DE LOS ANGELES GONZÁLEZ LIMA

Prof. (a). Dr (a). SANDRA AUGUSTA SANTOS

"Não temas, porque eu sou contigo;
não te assombres, porque eu sou o teu Deus;
eu te fortaleço, e te ajudo,
e te sustento com a minha destra fiel".
Isaías 41:10

"Tudo posso naquele que me fortalece".
Filipenses 4:13

Agradecimentos

Primeramente agradeço a Deus por ser meu escudo e a minha força. Por ter-me dado a posibilidade de fazer este trabalho e ajudado para concluílo.

Agradeço a meu esposo Emilio, por seu amor, seu apoio e sua ajuda cada vez que precisei.

A meus pais Omar e Nidia e a meu irmão Gustavo, que por mais longe que estejam, tenho todo o seu apoio e afeto.

A Nino, por sua amizade, apoio, pela sua orientação e confiança que teve em mim para realizar este trabalho.

Agradeço às professoras Sandra e María de los Ángeles, pelas valiosas dicas para melhorar esta tese.

Agradeço as professoras Sandra e Cheti por terem me ajudado com a linguagem de FORTRAN.

Agradeço ao professor José Mario Martínez, por terme esclarecido dúvidas cada vez que eu solicitei.

Agradeço ao Professor Benar Fux Svaiter pelas sugestões para enriquecer este trabalho.

A Laura e Rodrigo, por terem me ajudado sempre que eu precisei.

A meus amigos Cleyson, Marcia, Elisa e Débora, por terem me apoiado ao longo desta jornada.

À CAPES pelo apoio financeiro.

Resumo

Os sistemas ponto de sela surgem em uma grande quantidade de áreas de investigação, como física, química, engenharia, reconstrução de imagens, etc. Portanto, são objeto de pesquisa, tanto as propriedades presentes neles como os métodos utilizados para a sua resolução. Diversos métodos foram desenvolvidos dependendo das caraterísticas do sistema, alguns deles com a propriedade de preservar a estrutura da matriz do sistema.

Neste trabalho utilizamos umo destes métodos para melhorar a precisão obtida pelo método ALGENCAN (Lagrangiano Aumentado usando GENCAN) em problemas de Programação Não Linear (PNL). Este método é muito robusto, ele obtém uma boa aproximação da solução com poucas iterações, mas perto da solução não consegue obter uma precisão muito exigente. Para melhorar esta precisão, aplicamos o método de Newton a um sistema KKT reduzido no ponto obtido por ALGENCAN, gerando um sistema ponto de sela. Para esta implementação utilizamos o método conhecido como fatoração LDL^T, escolhido por sua propriedade de preservar a estrutura esparsa do sistema.

Abstract

Saddle point systems arise in wide areas of research fields like physics, chemistry and engineering and images reconstructions, etc. Then, the properties of these systems and solving methods have been subjects of intense study in the last years. Depending upon the system properties, several methods were developed; some of these, exhibit the property of preserving the matrix structure system, like the sparsity. In this work, we have used one of these methods to improve the accuracy by using ALGECAN (Augmented Lagrangian using GENCAN) applied to Non-linear Programming (NLP) problems. This is a robust method which helps to get a good approximation to the solution. However, in several cases, it is not possible to get the desired accuracy. In order to improve the precision, we have applied Newton's method in a reduced KKT system, starting from a point given by ALGENCAN, which is a saddle point. We employ the so called LDL^T factorization in order to implement Newton's method, which give us better accuracy.

Sumário

Capítulo 1

Introdução

Em anos recentes, houve um aumento no interesse na resolução de sistemas de equações lineares de grande escala, na forma de *ponto de sela* (saddle point). Isto é devido a que tais problemas surgem em uma grande quantidade de aplicações científicas e técnicas. Alguns dos campos de aplicação são: dinâmica de fluidos (Glowinski 1984, Quarteroni e Valli 1994, Teman 1984, Tureck 1999, Wesseling 2001), estimação de quadrados mínimos ponderados e com restrições (Björk 1996, Golub e Van Loan 1996), otimização com restrições (Gill, Murray e Wright 1981, Wright 1992, Wright 1997), econômico (Arrow, Hurwicz e Uzawa 1958, Duchin e Szyld 1979, Leontief, Duchin e Szyld 1985), redes e circuitos elétricos (Bergen 1986, Chua, Desoer e Kuh 1987), electromagnetismo (Bossavit 1998, Perugia 1997, Perugia, Simoncini e Arioli 1999), reconstrução de imagens (Hall 1979), entre outras aplicações.

Nesta tese abordamos o estudo da resolução deste tipo de sistemas para acelerar os métodos Lagrangianos Aumentados, utilizados para a resolução de problemas de otimização. A metodología de trabalho foi estudar um artigo de Benzi, Golub e Liesen (2005), que contém uma análise completa sobre a resolução deste tipo de sistemas, estudando os aspectos teóricos presentes neles. Logo, mostrando a importância dos sistemas ponto de sela, os utilizamos para atingir uma precisão maior que a obtida pelos métodos Lagrangiano Aumentado.

Este trabalho está organizado da seguinte forma:

- **Capítulo 2**: Neste capítulo fazemos um estudo da teoria presente no artigo de M. Benzi, G. H. Golub e J. Liesen (2005).

  Começamos o capítulo definindo os sistemas ponto de sela, mostrando exemplos de problemas onde eles podem surgir. Em seguida fazemos um estudo da teoria deste tipo de sistemas e finalmente, estudamos os métodos utilizados para a sua resolução.

- **Capítulo 3**: Neste capítulo apresentamos o método do Langrangiano Aumentado com alguns propriedades. Definimos a função Lagrangiano Aumentado proposta por Power-Hestenes-Rockafeller (PHR) e algumas propriedades referentes à convergência do método.

- **Capítulo 4**: Este capítulo é o nexo entre os dois capítulos anteriores. Aqui apresentamos nossas propostas para acelerar o Lagrangiano Aumentado, utilizando um dos métodos estudados no Capítulo 2. Em seguida, mostramos os resultados numéricos que obtivemos em nossas execuções, e finalmente as conclusões.

- **Apêndice**: No apêndice estão alguns resultados teóricos que precisamos para certas demonstrações.

Capítulo 2

Sistemas Ponto de Sela

2.1 Introdução

Suponhamos que temos o seguinte problema clásico de **Programação Não Linear Convexa**:

$$\begin{array}{ll} min & \frac{1}{2}\, x^T A x - f^T x \\ suj.\, a & Bx = g \end{array} \tag{2.1}$$

com $A \in \mathbb{R}^{n\times n}$, $A = A^T$ (simétrica) e $A \geq 0$ (semi-definida positiva), $f, x \in \mathbb{R}^n$, $B \in \mathbb{R}^{m\times n}$ e $g \in \mathbb{R}^m$.

Das condições de otimalidade de primeira ordem temos o seguinte sistema linear

$$\begin{array}{lcl} Ax - f + B^T\lambda & = & 0 \\ Bx & = & g \end{array} \tag{2.2}$$

onde λ é o vetor de multiplicadores de Lagrange, ou matricialmente

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}. \tag{2.3}$$

Deste sistema, podemos destacar duas propriedades:

Propriedades:

1. Uma solução deste sistema é um ponto KKT do problema original, mais ainda, é solução dele.

2. Qualquer solução (x^*, λ^*) deste sistema é um *ponto de sela* da função Lagrangiana:

$$\mathcal{L}(x, \lambda) = \frac{1}{2}\, x^T A x - f^T x + (Bx - g)^T \lambda$$

ou seja

$$\mathcal{L}(x^*, \lambda) \;\leq\; \mathcal{L}(x^*, \lambda^*) \;\leq\; \mathcal{L}(x, \lambda^*) \quad \forall\; x \in \mathbb{R}^n, \lambda \in \mathbb{R}^m.$$

Demonstração: A primeira propriedade é uma conseqüência das condições de otimalidade de primeira ordem, assim só resta provar a segunda. Para esta, devemos mostrar que se $(x^*, \lambda^*) \in \mathbb{R}^{n+m}$ satisfaz (2.3) então

$$\mathcal{L}(x^*, \lambda) \ \leq \ \mathcal{L}(x^*, \lambda^*) \ \leq \ \mathcal{L}(x, \lambda^*) \quad \forall \ x \in \mathbb{R}^n, \lambda \in \mathbb{R}^m.$$

Seja (x^*, λ^*) um ponto que satisfaz (2.3), então x^* é um ponto KKT do problema (2.1) com λ^* vetor de multiplicadores de Lagrange associado. Como o problema (2.1) é um problema de programação convexa então (x^*, λ^*) é solução do problema dual de (2.1), para mais informação ver Luenberger (1989) (cap. 13) então

$$\mathcal{L}(x^*, \lambda^*) = \max_{\lambda} \mathcal{L}(x^*, \lambda) \ \geq \ \mathcal{L}(x^*, \lambda) \ \forall \ \lambda \in \ \mathbb{R}^m,$$

$$\therefore \qquad \mathcal{L}(x^*, \lambda^*) \ \geq \ \mathcal{L}(x^*, \lambda) \ \ \forall \lambda \in \mathbb{R}^m.$$

Para provar a outra desigualdade, consideremos a função Langrangiana no ponto (x, λ^*), que é uma função que depende da variável x somente:

$$\mathcal{L}(x, \lambda^*) = \frac{1}{2} \ x^T A x - f^T x + (Bx - g)^T \lambda^*.$$

Como (x^*, λ^*) satisfaz (2.3), então é um ponto estacionário da função $\mathcal{L}(x, \lambda^*)$. Por outro lado esta função é quadrática com $A = A^T$ e $A \geq 0$, então x^* é minimizador global dela (ver Proposição A.1). Logo

$$\mathcal{L}(x^*, \lambda^*) = \min_{x} \mathcal{L}(x, \lambda^*) \ \leq \ \mathcal{L}(x, \lambda^*) \ \forall \ x \in \ \mathbb{R}^n,$$

$$\therefore \qquad \mathcal{L}(x^*, \lambda^*) \ \leq \ \mathcal{L}(x, \lambda^*) \ \ \forall x \in \mathbb{R}^n.$$

Com isto finalizamos a prova. ■

Esta segunda propriedade dá origem ao nome de uma generalização deste tipo de sistemas, chamados *sistemas Ponto de Sela*.

2.2 Sistemas Ponto de Sela

Consideremos os sistemas lineares em blocos 2 x 2 com a seguinte estrutura:

$$\begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{2.4}$$

no qual

$$A \in \mathbb{R}^{n \times n}, \ B_1 \, e \, B_2 \in \mathbb{R}^{m \times n} \, e \, C \in \mathbb{R}^{m \times m}, \ \text{ com } \ m \leq n, \tag{2.5}$$

ou, chamando

$$\mathcal{A} = \begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix}, \ \mu = \begin{pmatrix} x \\ y \end{pmatrix} \ \text{ e } \ b = \begin{pmatrix} f \\ g \end{pmatrix}$$

o sistema anterior pode ser resumido como

$$\mathcal{A}\mu = b.$$

O caso em que A e/ou B_1 ou B_2 ou ambas sejam nulas é excluído.

Definição 2.1 *Um sistema linear na forma* (2.4) *e* (2.5) *descreve um problema* ***Ponto de Sela*** *ou* ***Saddle Point****, se os blocos* A, B_1, B_2 *e* C *satisfazem uma ou mais das seguintes condições:*

1. A *é simétrica:* $A = A^T$
2. *a parte simétrica de* A: $H = \frac{1}{2}(A + A^T)$ *é semi-definida positiva*
3. $B_1 = B_2 = B$
4. C *é simétrica* ($C = C^T$) *e semi-definida positiva*
5. $C = 0$.

A matriz $\mathcal{A}$ *é chamada* ***matriz ponto de sela****.*

Pode se observar que *5.* implica *4.*
O caso mais simples é quando todas as condições anteriores são satisfeitas. Neste caso A é simétrica, semi-definida positiva e sob estas condições o sistema (2.4) resulta em um *sistema linear simétrico* da forma

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{2.6}$$

que é justamente o sistema apresentado na introdução deste capítulo.

Um caso particular deste tipo de problemas é o seguinte. Apresentamos aqui um exemplo de uma situação onde surge um sistema *ponto de sela*. O problema desenvolvido refere-se a achar o melhor estimador linear despolarizado (*"best linear unbiased estimate"BLUE*) que é obtido encontrando a solução do seguinte problema de quadrados mínimos generalizado:

$$\min_x (f - Gx)^T W^{-1} (f - Gx)$$

onde $f \in \mathbb{R}^{\mathrm{n}}$, $G \in \mathbb{R}^{\mathrm{nxm}}$, $\mathrm{m} \leq \mathrm{n}$, $\mathrm{W} \in \mathbb{R}^{\mathrm{nxn}}$ simétrica e definida positiva. Seja $c(x) = (f - Gx)^T W^{-1}(f - Gx)$, como W é simétrica e definida positiva então W^{-1} também é simétrica e definida positiva, então existe sua fatoração de Cholesky. Seja $L \in \mathbb{R}^{\mathrm{nxn}}$ matriz triangular inferior tal que $W^{-1} = LL^T$ então

$$C(x) = (f - Gx)LL^T(f - Gx) = \| L^T(f - Gx) \|^2$$

Das condições de otimalidade de primeira ordem tem-se

$$\nabla C = 2\, G^T LL^T(f - Gx) = 0 \quad \Longleftrightarrow \quad G^T W^{-1}(f - Gx) = 0$$

Seja $\quad y = W^{-1}(f - Gx) \quad \Longrightarrow \quad f - Gx = Wy \quad \Longleftrightarrow \quad f = Wy + Gx$

Logo temos o seguinte sistema linear

$$\left\{ \begin{array}{lcl} Wy + Gx & = & f \\ G^T y & = & 0 \end{array} \right.$$

Matricialmente

$$\left(\begin{array}{cc} W & G \\ G^T & 0 \end{array} \right) \left(\begin{array}{c} y \\ x \end{array} \right) = \left(\begin{array}{c} f \\ 0 \end{array} \right)$$

Esta formulação também é conhecida como formulação do *"sistema aumentado"* e, como podemos verificar, corresponde a um sistema com a estrutura do sistema (2.6).

2.3 Fatoração das matrizes Ponto de Sela

A resolução de sistemas lineares cuja matriz do sistema é triangular é uma situação muito desejável. Uma maneira de obter sistemas com estruturas mais simples é fatorar a matriz do sistema. A fatoração consiste em decompor a matriz como produto de duas ou mais matrizes. É claro que não nos interessa qualquer fatoração, as que são de interesse são aquelas onde as matrizes que surgem na decomposição têm estruturas mais simples que a original, como por exemplo estrutura diagonal, triangular, ou ortogonal. Logo o sistema linear pode ser reescrito como a resolução de dois ou mais sistemas (dependendo da fatoração) de estruturas matricias mais simples.

Suponhamos A não singular, então a matriz $\mathcal{A}$ dada em (2.4) admite a seguinte fatoração em bloco:

$$\mathcal{A} = \left(\begin{array}{cc} A & B_1^T \\ B_2 & -C \end{array} \right) = \left(\begin{array}{cc} I & 0 \\ B_2 A^{-1} & I \end{array} \right) \left(\begin{array}{cc} A & 0 \\ 0 & S \end{array} \right) \left(\begin{array}{cc} I & A^{-1} B_1^T \\ 0 & I \end{array} \right) \tag{2.7}$$

onde $S = -(C + B_2 A^{-1} B_1^T)$ é o *complemento de Schur* de A em $\mathcal{A}$. Outras fatorações da matriz $\mathcal{A}$ são possíveis, porém, trabalharemos apenas com a decomposição (2.7) devido às propriedades que surgem dela, como veremos nas seguintes seções. Por outro lado, em geral as fatorações não preservam a esparsidade da matriz, enquanto que este tipo de fatoração (2.7) tem a propriedade de preservar esta estrutura.

Da fatoração anterior, podemos observar que se A é inversível, a matriz $\mathcal{A}$ é não singular se, e somente se, o complemento de Schur é inversível.

2.4 Condições de solubilidade

O conhecimento da existência ou não de soluções do problema que estamos interessados em resolver é uma tarefa muito importante, já que ela vai nos dar ferramentas para não trabalhar em vão, além de dar informação sobre o problema que este sistema representa. É

por isso que a análise da solubilidade é muito importante, para decidir se contiuamos com esse objetivo ou devemos mudá-lo. A seguir vamos estudar sob que condições os sistemas Ponto de Sela têm solução.

Caso simétrico

Considere o caso $A = A^T$, $A > 0$, $B_1 = B_2 = B$, $C = 0$, então $S = -BA^{-1}B^T$ é semi-definida negativa e se B tem posto completo (isto é, suas colunas formam um conjunto de vetores linearmente independentes) então S é inversível e portanto, $\mathcal{A}$ é inversível, logo o sistema (2.6) tem solução única.

Lema 2.1 *Sob as condições $A = A^T$, $A > 0$, $B_1 = B_2 = B$ com posto completo e C=0 , se (x^*, y^*) é a solução de* (2.6) *então x^* é a única solução de* (2.1).

Demonstração: Como (x^*, y^*) satisfaz (2.6) então $Bx^* = g$.
Sejam z e p tais que z é outra solução do sistema anterior, ou seja $Bz = g$ com $z \neq x^*$ e p satisfaz $p = x^* - z$ (note que $p \neq 0$).
Então

$$Bp = B(x^* - z) = Bx^* - Bz = 0 \implies p \in Nu(B).$$

Por outro lado,

$$J(z) = \frac{1}{2}\, z^T A z - f^T z,$$

mas como $z = x^* - p$ temos

$$J(x^* - p) = \frac{1}{2}\,(x^* - p)^T A(x^* - p) - f^T(x^* - p)$$

$$= \frac{1}{2}\,(x^*)^T A x^* - p^T A x^* + \frac{1}{2}\, p^T A p - f^T x^* + f^T p. \tag{2.8}$$

Para concluir esta prova é preciso fazer outra análise.
De (2.6)

$$Ax^* + By^* = f \implies Ax^* = f - By^*$$

$$\implies p^T A x^* = p^T(f - B^T y^*) = p^T f - p^T B^T y^* = p^T f - (Bp)^T y^*,$$

como $Bp = 0$, pois $p \in Nu(B)$, então tem-se

$$p^T A x^* = p^T f = f^T p.$$

Utilizando este resultado em (2.8) obtemos

$$J(x^* - p) = \frac{1}{2}\,(x^*)^T A x^* + \frac{1}{2}\, p^T A p - f^T x^* = J(x^*) + \frac{1}{2}\, p^T A p\ .$$

Como A é definida positiva e $p \neq 0$, então

$$J(x^*) + \frac{1}{2}\, p^T A p \; > J(x^*),$$

assim

$$J(x^* - p) > J(x^*) \iff J(z) > J(x^*).$$

Logo, como z é arbitrário, temos $J(z) > J(x^*)\ \forall z$ factível diferente de x^*.
Com isto fica provado que x^* é solução de (2.1).
Por outro lado, se w é solução de (2.1) então satisfaz (2.6), mas pelas hipóteses o sistema (2.6) tem uma única solução então $w = x^*$, logo a única solução de (2.1) é x^*. ■

Em seguida vamos a analisar o caso em que $A = A^T$ e definida positiva, $B_1 = B_2 = B$ e $C \neq 0$ semi-definida positiva. Sob estas condições $S = -(C + BA^{-1}B^T)$ é simétrica e semi-definida negativa. Sob estas hipóteses estabelecemos algumas proposições:

Proposição 2.1 *A matriz S é definida negativa se, e somente se,* $Nu(B^T) \cap Nu(C) = \{0\}$.

Demonstração: Suponhamos S definida negativa, isto é $x^T S x < 0\ \forall\, x \neq 0$ e $x^T S x = 0$ se, e somente se, $x = 0$.
Seja $x \in Nu(B^T) \cap Nu(C) \implies B^T x = 0$ e $Cx = 0$
Agora

$$x^T S x = -x^T (C + BA^{-1}B^T)\, x = -x^T C x - x^T B A^{-1} B^T x =$$
$$= -x^T C x - (B^T x)^T A^{-1} (B^T x) = 0$$

portanto

$$x^T S x = 0 \implies x = 0$$

Logo $Nu(B^T) \cap Nu(C) \subseteq \{0\}$. Mas como $0 \in Nu(B^T) \cap Nu(C)$ então $Nu(B^T) \cap Nu(C) = \{0\}$.

Reciprocamente, suponhamos que $\{0\} = Nu(B^T) \cap Nu(C)$. Como $C \geq 0$ e $A > 0$ então $S \leq 0$, logo somente temos que provar que $x^T S x = 0$ implica $x = 0$.
Seja x tal que

$$0 = x^T S x = -x^T (C + BA^{-1}B^T)\, x = -x^T C x - x^T B A^{-1} B^T x.$$

Como $C \geq 0$ e $A > 0$ e portanto A^{-1} também é definida positiva, temos o seguinte:
(a) $x^T C x = 0 \implies Cx = 0$ (ver A.2) $\implies x \in Nu(C)$.
(b) $(B^T x)^T A^{-1} (B^T x) = 0 \implies B^T x = 0 \implies x \in Nu(B^T)$.
Portanto $x \in Nu(B^T) \cap Nu(C) = \{0\} \implies x = 0$. Logo S é definida negativa. ■

Proposição 2.2 *Se* B^T *tem posto completo então* $\mathcal{A}$ *é inversível.*

Demonstração: Utilizando a fatoração (2.7), a matriz $\mathcal{A}$ é inversível se, e somente se, a matriz S é inversível e como esta última matriz é semi-definida negativa, basta provar que ela é definida negativa.
Seja x tal que $x^T Sx = 0 \implies x \in Nu(B^T) \cap Nu(C) \implies B^T x = 0$, mas como B^T tem posto completo temos que $x = 0$. Assim S é definida negativa. ■

Estas duas últimas proposições podem ser resumidas no seguinte teorema:

Teorema 2.1 *Seja A simétrica definida positiva, $B_1 = B_2 = B$ e C simétrica semi-definida positiva. $Nu(C) \cap Nu(B^T) = \{0\}$, se, e somente se, a matriz ponto de sela $\mathcal{A}$ é inversível. Em particular, se B tem posto completo, $\mathcal{A}$ é inversível.*

O seguinte teorema dá informação sobre que condições a matriz $\mathcal{A}$ é inversível, exigindo só que a matriz A seja simétrica e semi-definida positiva. A prova dele pode ser achada em Benzi, Golub e Liezen (2005, Teorema 3.3)

Teorema 2.2 *Seja A semi-definida positiva e simétrica, $B_1 = B_2 = B$ com posto completo e $C = 0$. Então uma condição necessária e suficiente para que a matriz ponto de sela $\mathcal{A}$ seja inversível é que $Nu(A) \cap Nu(B^T) = \{0\}$.*

O próximo teorema é uma generalização do teorema anterior para o caso em que $C = 0$. Ele fornece uma condição necessária para a não singularidade da matriz $\mathcal{A}$. Novamente a prova pode ser achada em Benzi, Golub e Liezen (2005, Teorema 3.3)

Teorema 2.3 *Se a matriz*

$$\mathcal{A} = \begin{pmatrix} A & B_1^T \\ B_2 & 0 \end{pmatrix}$$

é inversível, então posto$(B_1) = m$ e posto $\begin{pmatrix} A \\ B_2 \end{pmatrix} = n$.

O seguinte teorema extende as condições necessárias e suficientes do Teorema 2.2 para $\mathcal{A}$ ser inversível.

Teorema 2.4 *Seja H a parte simétrica de A, semi-definida positiva, $B_1 = B_2 = B$ com posto completo e C simétrica e semi-definida positiva.*

1. *Se $Nu(H) \cap Nu(B) = \{0\}$ então $\mathcal{A}$ é inversível.*
2. *Se $\mathcal{A}$ é inversível então $Nu(A) \cap Nu(B) = \{0\}$.*

Demonstração:
1. Para provar que $\mathcal{A}$ é inversível, vamos provar que o sistema

$$\begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0 \qquad ou \qquad \mathcal{A}\mu = 0 \tag{2.9}$$

tem como única solução $x = y = 0$.
Seja $\begin{pmatrix} x \\ y \end{pmatrix}$ tal que é solução do sistema (2.9), então

$$Ax + B^T y = 0, \tag{2.10}$$

$$Bx - Cy \ = 0. \tag{2.11}$$

Pré-multiplicando a equação (2.10) por x^T temos

$$x^T Ax + x^T B^T y = 0. \tag{2.12}$$

Pré-multiplicando a equação (2.11) por y^T temos

$$y^T Bx - y^T Cy = 0. \tag{2.13}$$

Observemos que $y^T Bx = (Bx)^T y = x^T B^T y$, substraindo (2.13) de (2.12) obtemos:

$$x^T Ax + y^T Cy = 0. \tag{2.14}$$

Por outro lado se $A = H + K$ onde H e K são a parte simétrica e anti-simétrica de A respectivamente, verifica-se que $x^T Ax = x^T Hx$, então substituindo este resultado em (2.14) tem-se:

$$x^T Hx + y^T Cy = 0. \tag{2.15}$$

Agora, como $H \geq 0$ e $C > 0$ e ambas são simétricas, verifica-se:

$$x^T Hx = 0 \quad \Longrightarrow \quad Hx = 0 \quad \Longrightarrow \quad x \in Nu(H) \tag{2.16}$$

$$y^T Cy = 0 \quad \Longrightarrow \quad y = 0. \tag{2.17}$$

Substituindo o valor de y em (2.11) tem-se:

$$Bx = 0 \qquad \Longrightarrow \qquad x \in Nu(B). \tag{2.18}$$

Logo, de (2.16) e (2.18) tem-se:

$$x \in Nu(H) \cap Nu(B) = \{0\} \qquad \Longrightarrow \qquad x = 0,$$

então, a única solução de (2.9) é a trivial. Logo $\mathcal{A}$ é inversível.

Agora, provemos o item 2. do Teorema. Seja $x \in Nu(A) \cap Nu(B)$ e seja $\mu = \begin{pmatrix} x \\ 0 \end{pmatrix}$, logo

$$\mathcal{A}\mu = \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} Ax \\ Bx \end{pmatrix}$$

como $x \in Nu(A) \cap Nu(B)$, então $\mathcal{A}\mu = 0$. Por outro lado $\mathcal{A}$ é inversível, assim verifica-se que $\mu = 0$ e portanto $x = 0$, com o qual fica provado o item 2. ■

Resumindo as condições de inversibilidade:

1. Sejam $A = A^T,\ B_1 = B_2 = B$ e $C = C^T,\ C \geq 0$.
   (a) $Nu(C) \cap Nu(B) = \{0\} \iff \mathcal{A}$ é inversível.
   (b) Se B^T tem posto completo então $\mathcal{A}$ é inversível.
2. Sejam $A = A^T,\ A \geq 0,\ B_1 = B_2 = B$ com posto completo e $C = 0$.
$$Nu(A) \cap Nu(B^T) = \{0\} \iff \mathcal{A} \text{ é inversível.}$$
3. Sejam $A = A^T$ e $C = 0$. Se $\mathcal{A}$ é inversível, então $posto(B_1) = m$ e $posto \begin{pmatrix} A \\ B_2 \end{pmatrix} = n$.
4. Seja H parte simétrica de A, $H \geq 0,\ B_1 = B_2 = B$ com posto completo e $C = C^T,\ C \geq 0$.
   (a) Se $Nu(H) \cap Nu(B^T) = \{0\} \implies \mathcal{A}$ é não singular.
   (b) $\mathcal{A}$ é inversiível $\implies Nu(A) \cap Nu(B) = \{0\}$.

2.5 A inversa das matrizes Ponto de Sela

Na seção 2.3, estudamos que sob a hipótese de A ser não singular, a matriz $\mathcal{A}$ é não singular se, e somente se, o *complemento de Schur* $S = -(C + B_2^T A^{-1} B_1^T)$ é não singular. Além disso obtivemos uma fatoração dela. Portanto assumamos que A é inversível e lembremos a descomposição da matriz $\mathcal{A}$, então temos

$$\mathcal{A} = \begin{pmatrix} I & O \\ B_2 A^{-1} & I \end{pmatrix} \begin{pmatrix} A & O \\ O & S \end{pmatrix} \begin{pmatrix} I & A^{-1} B_1^T \\ O & I \end{pmatrix}.$$

Vamos a usar esta descomposição para achar a inversa. Chamando $F_1,\ M$ e F_2 as três matrizes que aparecem na fatoração respetivamente, podemos escrever a seguinte igualdade

$$\mathcal{A} = F_1 M F_2,$$

logo

$$\mathcal{A}^{-1} = (F_2)^{-1} M^{-1} (F_1)^{-1}. \tag{2.19}$$

Calculando cada inversa de (2.19) separadamente, obtemos:

$$F_2^{-1} = \begin{pmatrix} I & -A^{-1} B_1^T \\ O & I \end{pmatrix} \quad M^{-1} = \begin{pmatrix} A^{-1} & O \\ O & S^{-1} \end{pmatrix} \quad F_1^{-1} = \begin{pmatrix} I & O \\ -B_2 A^{-1} & I \end{pmatrix}.$$

Finalmente fazendo as operações correspondentes, temos:

$$\mathcal{A}^{-1} = \begin{pmatrix} A^{-1} + A^{-1} B_1^T S^{-1} B_2 A^{-1} & -A^{-1} B_1^T S^{-1} \\ -S^{-1} B_2 A^{-1} & S^{-1} \end{pmatrix}$$

Se A for singular e C não singular, pode-se obter uma expressão análoga, assumindo que a matriz $A + B_1^T C^{-1} B_2$ (*complemento de Schur de* C em $\mathcal{A}$) é inversível.

Vamos a analisar o sistema ponto de sela para o caso particular em que A é simétrica definida positiva, $B = B_1 = B_2$, $C = 0$, $S = -BA^{-1}B^T$ não singular (portanto definida negativa) e $g = 0$. A solução (x^*, y^*) de (2.6) é:

$$\begin{pmatrix} x^* \\ y^* \end{pmatrix} = \mathcal{A}^{-1} \begin{pmatrix} f \\ 0 \end{pmatrix} = \begin{pmatrix} (I + A^{-1}B^T S^{-1}B)A^{-1}f \\ -S^{-1}BA^{-1}f \end{pmatrix}. \tag{2.20}$$

Seja $\pi = -A^{-1}B^T S^{-1}B$, por construção π é um operador linear, além disso satisfaz que $\pi^2 = \pi$ portanto π é projeção. Agora, dado $v \in \mathbb{R}^n$, já que $\pi^2 v = \pi v$ então $\pi v \in Im(A^{-1}B^T)$. Por outro lado, $v - \pi v \in Nu(\pi) \implies A^{-1}B^T S^{-1}B(v - \pi v) = 0 \implies B^T S^{-1}B(v - \pi v) = 0 \implies (v - \pi v)^T B^T S^{-1}B(v - \pi v) = 0$, mas como S^{-1} é definida negativa verifica-se que $B(v - \pi v) = 0 \implies v - \pi v \in Nu(B) \implies v - \pi v \perp ImB^T$. Portanto temos a seguinte relação:

$$\pi v \in Im(A^{-1}B^T) \quad \text{e} \quad v - \pi v \perp Im(B^T) \quad \forall v \in \mathbb{R}^n.$$

Da equação (2.20), a primeira componente pode ser escrita como

$$x^* = (I + A^{-1}B^T S^{-1}B)A^{-1}f = (I - \pi)\hat{x}$$

onde $\hat{x} = A^{-1}f$ é a solução do problema irrestrito com a função quadrática de (2.1). Da última expressão podem-se tirar duas observações: a primeira é que x^* é ortogonal a $Im(B^T)$. A segunda é que $\hat{x} = x^* + \pi\hat{x}$, significa que a solução do problema irrestrito está formada por duas partes, uma que está na $Im(A^{-1}B)$ e a outra que é ortogonal a B^T.
Voltando a nosso sistema *ponto de sela*, temos

$$\begin{aligned} Ax^* + B^T y^* = f &\implies Ax^* = f - B^T y^*, \\ Bx^* \qquad\quad = 0, & \end{aligned}$$

então

$$0 = Bx^* = BA^{-1}Ax^* = BA^{-1}(f - B^T y^*). \tag{2.21}$$

Como A é simétrica e definida positiva, sua inversa define o seguinte produto interno

$$\langle v, v \rangle_{A^{-1}} = v^T A^{-1} v,$$

e

$$\|v\|^2_{A^{-1}} = \langle v, v \rangle_{A^{-1}}.$$

Logo de (2.21), $f - B^T y^*$ é ortogonal a cada uma das colunas de B^T com relação a $\langle , \rangle_{A^{-1}}$, então $f - B^T y^* \perp_{\langle , \rangle_{A^{-1}}} Im(B^T)$ então $B^T y^*$ é a projeção ortogonal de f sobre $Im(B^T)$, portanto y^* é solução do seguinte problema de quadrados mínimos

$$\min_u \|f - B^T u\|_{A^{-1}}.$$

Assim, para resolver o problema *ponto de sela* (2.6) (com $g = 0$), pode-se resolver primeiro o problema de quadrados mínimos generalizados obtendo y^*, e em seguida calcular x^* de acordo com

$$x^* = A^{-1}(f - B^T y^*).$$

2.6 Propriedades Espectrais

O estudo das propriedades espectrais é muito importante para o desenvolvimento de métodos algorítmicos, como também na análise da convergência deles. Na otimização, o conhecimento do signal dos autovalores da matriz Hessiana permite nos determinar se o ponto estacionário obtido é em um minimizador ou maximizador local, ou, um ponto de sela. Este fato é o que motiva esta seção.
Antes de tratar algumas propriedades de autovalores, vamos estudar outras propriedades matricias que serão úteis. Trabalharemos com matrizes reais, mas as propriedades, definições e teoremas desenvolvidos também se estendem naturalmente ao caso complexo.

Definição 2.2 *Duas matrizes A e B se dizem "congruentes", se existe uma matriz inversível P tal que $P^T AP = B$.*

Exemplo: Consideremos as seguintes matrizes

$$A = \begin{pmatrix} -2 & 0 \\ 1 & -1 \end{pmatrix} \qquad \text{e} \qquad B = \begin{pmatrix} -1 & 1 \\ 0 & -2 \end{pmatrix}$$

Elas são *congruentes*, pois escolhendo

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

verifica-se que $P^T AP = B$.

A seguir vamos a estudar alguns resultados referentes à congruência de matrizes, começando com a relação de equivalência.

Lema 2.2 *A "congruência"é uma relação de equivalência. Isto é:*

1. *A é congruente a A.*
2. *Se A é congruente a B então B é congruente a A.*
3. *Se A é congruente a B e B é congruente a C, então A é congruente a C.*

Demonstração:
1. Basta escolher $P = I$.
2. Se $A = P^T BP$, com P inversível, então $B = (P^T)^{-1} AP^{-1}$.
3. Se $A = P^T BP$ e $B = S^T CS$ então $A = (SP)^T C(SP)$. ■

Propriedades:

1. *Duas matrizes congruentes têm o mesmo posto.*

2. *Sejam A e B matrizes congruentes, então A é definida positiva (negativa) se, e somente se, B é definida positiva (negativa).*

Demonstração:
1. Sejam A e B matrizes congruentes $\implies$ $\exists\, P$ inversível tal que $P^T AP = B$ ou $A = P^{-T} BP^{-1}$.
Agora

$$posto(B) = posto(P^T AP) \leq posto(P^T A) \leq posto(A)$$

$$\implies \quad posto(B) \leq posto(A).$$

Com um argumento similar, utilizando $A = P^{-T} BP^{-1}$ podemos obter

$$posto(A) \leq posto(B).$$

Juntando ambos resultados tem-se

$$posto(A) = posto(B).$$

Para a prova da segunda propriedade analisamos o caso em que $A > 0$, queremos provar que $B > 0$, ou seja $P^T AP > 0$.
Pela hipótese,

$$x^T P^T APx > 0 \quad \forall\ Px \neq 0.$$

Agora $Px = 0 \iff x = 0$ pois P é inversível $\implies x^T P^T APx > 0\ \forall x \neq 0$. Logo B é definida positiva.
Para a recíproca, basta trocar B por A na prova anterior. Assim fica demonstrado 2. ■

Corolário 2.1 *Sejam A e B duas matrizes congruentes. Então A é indefinida $\iff$ B é indefinida.*

Outro conceito de uma matriz que adquire importância no momento de obter informação acerca dos autovalores dela é o conceito de Inércia da matriz que é definido a seguir.

Definição 2.3 *Seja $A \in \mathbb{R}^{nxn}$ uma matriz simétrica. Chamamos "Inércia de A " à terna ordenada:*

$$i(A) = (i_+(A), i_-(A), i_0(A))$$

onde $i_+(A)$ é a quantidade de autovalores positivos de A, $i_-(A)$ é a quantidade de autovalores negativos de A, e $i_0(A)$ é a quantidade de autovalores nulos de A, contando multiplicidade. Note que posto(A) é igual a $i_+(A) + i_-(A)$.

Observação: Seja $A \in \mathbb{R}^{nxn}$ matriz simétrica, então $\exists\ U \in \mathbb{R}^{nxn}$ matriz ortogonal tal que $A = U\Lambda U^T$ com $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$, λ_i i-ésimo autovalor de A e $\lambda_i \in \mathbb{R}$ pois A é simétrica. Suponhamos que A tem r autovalores positivos, k negativos e $n-k-r$ nulos. Sem perda de generalidade, vamos supor que estão ordenados tais que $\lambda_1, \ldots, \lambda_r$ são positivos, $\lambda_{r+1}, \ldots, \lambda_{r+k}$ são negativos e $\lambda_{r+k+1}, \ldots, \lambda_{n-r-k}$ nulos.
Definindo $D = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}, \sqrt{-\lambda_{r+1}}, \ldots, \sqrt{-\lambda_{r+k}}, 1, \ldots 1)$ temos que D é não singular e $\Lambda = DI(A)D$, onde $I(A) = diag(1, \ldots 1, -1, \ldots -1, 0, \ldots 0)$. Os primeiros r elementos da diagonal são 1's, os $r+k$ seguintes tem o valor -1 e os $n-k-r$ restantes são 0, ou, também podemos dizer que $I(A)$ tem $i_+(A)$ termos que são "+1", $i_-(A)$ termos que são "-1"e $i_0(A)$ termos que são "0".
Matricialmente

$$\Lambda = D \begin{pmatrix} 1 & & & & & \\ & 1 & & & 0 & \\ & & -1 & & & \\ & & & & & \\ & & & -1 & & \\ & 0 & & & 0 & \\ & & & & & 0 \end{pmatrix} D = DI(A)D. \tag{2.22}$$

A matriz $I(A)$ é chamada *matriz inércia de A*.
Agora a matriz A pode ser escrita como

$$A = UDI(A)DU^T = NI(A)N^T, \tag{2.23}$$

onde $N = DU^T$. Logo A é congruente a $I(A)$. Assim toda matriz simétrica é congruente à sua matriz de inércia (ver Horn e Johnson (1985), pág. 223).

Teorema 2.5 *Lei da Inércia de Sylvester:*
Duas matrizes simétricas são congruentes, se e somente têm a mesma inércia, isto é, a mesma quantidade de autovalores positivos, negativos e nulos.

(ver Horn e Johnson (1985), pág. 224).

Demonstração: Suponhamos que A e B têm a mesma inércia, então cada uma pode ser expressa da forma (2.23), com a mesma matriz de inércia, mas provavelmente, com diferentes N . Agora, pela parte 3 do *lema* 2.2, a congruência é uma relação transitiva e neste caso, temos que A e B são congruentes à mesma matriz de inércia, portanto B é congruente a A.

Recíprocamente, suponhamos A congruente a B e seja P matriz inversível tal que $A = P^T BP$. Como A e B têm o mesmo posto (pela *propriedade* 1.), então tem a mesma quantidade de autovalores nulos. Assim só precisamos provar que elas tem a mesma quantidade de autovalores positivos.

Suponhamos que A tem r autovalores positivos e sejam $v_1, \ldots, v_r$ autovetores ortonormais de A correspondentes aos autovalores positivos $\lambda_1, \ldots, \lambda_r$. Seja $T = span\{v_1, \ldots, v_r\}$ então a dimensão de T é r, e se $x \in T$ com $x \neq 0$ então

$$
\begin{aligned}
& x = \alpha_1 v_1 + \ldots + \alpha_r v_r \\
\Longrightarrow \qquad & Ax = \lambda_1 \alpha_1 v_1 + \ldots + \lambda_r \alpha_r v_r \\
\Longrightarrow \qquad & x^T Ax = \sum_{i=1}^{r} (\alpha_i v_i)^T \sum_{j=1}^{r} \alpha_j \lambda_j v_j = \sum_{i=1}^{r} \alpha_i^2 \lambda_i > 0 \\
\Longrightarrow \qquad & x^T P^T BPx > 0 \\
\Longrightarrow \qquad & (Px)^T B(Px) > 0
\end{aligned}
$$

então $y^T By > 0 \ \ \forall y \in M = span\{Pv_1, \ldots, Pv_r\}$ com $y \neq 0$. Como P é inversível e $v_1, \ldots, v_r$ são ortonormais então $Pv_1, \ldots Pv_r$ são linearmente independentes, portanto M tem dimensão r, então B tem no mínimo r autovalores positivos (ver corolario A.1). Assim $I_+(B) \geq I_+(A)$. Agora como os papeis entre A e B neste argumento podem ser trocados, podemos concluir que $I_+(A) = I_+(B)$. ■

Este teorema é muito importante, dado que no caso das condições de otimalidade em problemas de otimização, podemos obter informação valiosa acerca dos pontos estacionários do problema.

Agora sim, já estamos em condições de estudar algumas propriedades referentes aos autovalores das matrizes presentes nos sistemas *ponto de sela*.

Assumamos que A é simétrica e definida positiva, $B_1 = B_2 = B$ com suas colunas linearmente independentes, seja C simétrica semi-definida positiva. Então de (2.7) temos

$$
\begin{pmatrix} I & 0 \\ -BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} I & -A^{-1}B^T \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}
$$

com $S = -(C - BA^{-1}B^T)$ simétrica definida negativa.

Logo $\mathcal{A}$ é congruente à matriz em bloco

$$
\begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}.
$$

Como esta última matriz é indefinida, pois tem n autovalores positivos, que são os autovalores da matriz A e tem m autovalores negativos correspondentes à matriz S, então $\mathcal{A}$ é indefinida com n autovalores positivos e m autovalores negativos. O mesmo acontece se B tiver posto incompleto, ou suas colunas não forem linearmente independentes, pois S continua sendo definida negativa. Se S tiver posto deficiente, por exemplo $m-r$, então S teria $m-r$ autovalores negativos e r nulos. Logo $\mathcal{A}$ teria n autovalores positivos, $m-r$ negativos e r nulos.

O seguinte resultado estabelece um limitante para os autovalores de um tipo das matrizes $\mathcal{A}$, que correspondem ao tipo de matrizes que são de nosso interesse.

Teorema 2.6 *Sejam A simétrica definida positiva, $B_1 = B_2 = B$ com posto completo e $C = 0$, μ_1 e μ_n o maior e o menor autovalores de A respectivamente, σ_1 e σ_m o maior e o menor valores singulares de B e $\sigma(\mathcal{A})$ o espectro de $\mathcal{A}$. Então*

$$\sigma(\mathcal{A}) \subset I^- \cup I^+,$$

onde

$$I^- = [\ \frac{1}{2}(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\ ,\ \frac{1}{2}(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2}\ )\ ]$$

e

$$I^+ = [\ \mu_n\ ,\ \frac{1}{2}(\mu_1 + \sqrt{\mu_1^2 + 4\sigma_1^2}\ )\ ].$$

(ver Rusten e Winter (1992))

Demonstração: Seja λ um autovalor de $\mathcal{A}$ com autovetor associado $\begin{pmatrix} x \\ y \end{pmatrix}$, isto é

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix},$$

então

$$Ax + B^T y = \lambda x, \tag{2.24}$$

$$Bx = \lambda y. \tag{2.25}$$

Se $x = 0$, de (2.24) temos $B^T y = 0$, logo, $y = 0$ pois B tem posto completo, mas isto não pode acontecer, já que $\begin{pmatrix} x \\ y \end{pmatrix}$ é autovetor, portanto $x \neq 0$.
Pré-multiplicando (2.24) por x temos

$$x^T Ax + x^T B^T y = \lambda \|x\|^2, \tag{2.26}$$

e pré-multiplicando (2.25) por y temos

$$y^T Bx = \lambda \|y\|^2. \tag{2.27}$$

Agora, substituindo (2.27) em (2.26) obtemos

$$x^T Ax + \lambda \|y\|^2 = \lambda \|x\|^2. \tag{2.28}$$

Como $\mathcal{A}$ tem autovalores positivos e negativos, vamos dividir a demonstração em duas partes, uma considerando os autovalores positivos e a outra considerando os negativos.
Analisemos o primeiro caso. Sabe-se que $x^TAx \geq \mu_n\|x\|^2$ (ver Proposição A.3), então substituindo em (2.28)

$$\begin{aligned} & x^TAx + \lambda\|y\|^2 \geq \mu_n\|x\|^2 + \lambda\|y\|^2 \\ \Longrightarrow \quad & \lambda\|x\|^2 \geq \mu_n\|x\|^2 + \lambda\|y\|^2 \\ \Longrightarrow \quad & -\lambda\|y\|^2 \geq (\mu_n - \lambda)\|x\|^2 \\ \Longrightarrow \quad & 0 \geq \mu_n - \lambda \\ \Longrightarrow \quad & \lambda \geq \mu_n. \end{aligned} \tag{2.29}$$

Por outro lado, de (2.25) temos

$$Bx = \lambda y \implies \|Bx\| = \lambda\|y\| \implies \frac{1}{\lambda}\|Bx\| = \|y\|.$$

Então substituindo em (2.28), segue que

$$x^TAx + \frac{1}{\lambda}\|Bx\|^2 = \lambda\|x\|^2. \tag{2.30}$$

Considerando a desigualdade (ver Proposição A.3 e Proposição A.4)

$$x^TAx + \frac{1}{\lambda}\|Bx\|^2 \leq \mu_1\|x\|^2 + \frac{1}{\lambda}\,\sigma_1^2\|x\|^2,$$

logo, (2.30) fica

$$\begin{aligned} & \lambda^2\|x\|^2 \leq (\mu_1\lambda + \sigma_1^2)\|x\|^2 \\ \Longrightarrow \quad & 0 \leq (-\lambda^2 + \mu_1\lambda + \sigma_1^2)\|x\|^2 \\ \Longrightarrow \quad & 0 \geq (\lambda^2 - \mu_1\lambda - \sigma_1^2) \\ \Longleftrightarrow \quad & 0 \geq (\lambda - \lambda_1)(\lambda - \lambda_2), \end{aligned} \tag{2.31}$$

onde

$$\lambda_1 = \frac{1}{2}(\mu_1 + \sqrt{\mu_1^2 + 4\sigma_1^2}\ ),$$

$$\lambda_2 = \frac{1}{2}(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_1^2}\ ).$$

Como os autovalores e os valores singulares são positivos, das duas últimas expressões temos que $\lambda_1 > 0$ e $\lambda_2 < 0$. Agora da desigualdade (2.31) tem-se duas opções

1. $\lambda \geq \lambda_1$ e $\lambda \leq \lambda_2$
2. $\lambda \leq \lambda_1$ e $\lambda \geq \lambda_2$

1. não pode acontecer, pois se $\lambda \leq \lambda_2$ então $\lambda < 0$ que é absurdo. Portanto

$$0 < \lambda \leq \lambda_1 \implies \lambda < \frac{1}{2}(\mu_1 + \sqrt{\mu_1^2 + 4\sigma_1^2}\ ). \tag{2.32}$$

Logo, juntando (2.29) com (2.32) temos

$$\lambda \in [\mu_n\ ,\ \frac{1}{2}(\mu_1 + \sqrt{\mu_1^2 + 4\sigma_1^2}\ )\ ].$$

Agora analisemos o caso dos autovalores negativos. De (2.25) temos

$$\lambda y = Bx \implies |\lambda|\|y\| = \|Bx\| \implies \|y\| = \frac{1}{|\lambda|}\|Bx\|$$

substituindo em (2.28)

$$x^T Ax + \frac{1}{\lambda}\|Bx\|^2 = \lambda\|x\|^2, \tag{2.33}$$

e lembrando que

$$x^T Ax \geq \mu_n\|x\|^2,$$

e

$$\frac{1}{\lambda}\|Bx\|^2 \geq \frac{\sigma_1^2}{\lambda}\|x\|^2,$$

(2.33) fica

$$\begin{aligned} \lambda\|x\|^2 \ &\geq\ \mu_n\|x\|^2 + \frac{1}{\lambda}\,\sigma_1^2\|x\|^2 \\ \implies \qquad (\lambda^2 - \mu_n\lambda - \sigma_1^2)\|x\|^2 &\leq 0 \\ \implies \qquad (\lambda - \lambda_1)(\lambda - \lambda_2) &\leq 0, \end{aligned} \tag{2.34}$$

onde

$$\lambda_1 = \frac{1}{2}(\mu_n + \sqrt{\mu_n^2 + 4\sigma_1^2}\ ),$$

e

$$\lambda_2 = \frac{1}{2}(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\ ).$$

Observemos que $\lambda_1 > 0$ e $\lambda_2 < 0$. Assim como antes, da desigualdade (2.34) temos duas opções

1. $\lambda \geq \lambda_1$ e $\lambda \leq \lambda_2$
2. $\lambda \leq \lambda_1$ e $\lambda \geq \lambda_2$

Se acontecer 1. então $\lambda > 0$ o qual contradiz nossa hipóteses. Portanto

$$\lambda_2 \leq \lambda < 0 \implies \lambda \geq \frac{1}{2}(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\ ). \tag{2.35}$$

Agora, vamos a procurar o limitante superior. Sabemos que $x \in \mathbb{R}^n$, e $\mathbb{R}^n$ é soma direta de $Nu(B)$ e $Nu(B)^\perp \implies \exists\, v \in Nu(B)$ e $w \in Nu(B)^\perp$ tais que $x = v + w$, então fazendo (2.24) produto interno com v temos

$$\langle Ax, v\rangle + \langle B^T y, v\rangle = \langle \lambda x, v\rangle$$

$$\implies \langle Av, v\rangle + \langle Aw, v\rangle = \lambda\langle v, v\rangle + \lambda\langle w, v\rangle - \langle B^T y, v\rangle.$$

Como $\langle w, v\rangle = 0$ temos

$$\implies \langle Aw, v\rangle = \lambda\|v\|^2 - \langle Av, v\rangle - \langle B^T y, v\rangle. \tag{2.36}$$

Agora

$$\mu_n\|v\|^2 \leq \langle Av, v\rangle \leq \mu_1\|v\|^2 \implies -\mu_n\|v\|^2 \geq -\langle Av, v\rangle \geq -\mu_1\|v\|^2. \tag{2.37}$$

Também temos $\langle B^T y, v\rangle = y^T Bv$, mas como $Bx = \lambda y \implies y = \frac{1}{\lambda}Bx = \frac{1}{\lambda}B(v + w) = \frac{1}{\lambda}Bv \implies y = \frac{1}{\lambda}Bv$, logo

$$\langle B^T y, v\rangle = \frac{1}{\lambda}v^T B^T Bv = \frac{1}{\lambda}\|Bv\|^2 \tag{2.38}$$

e

$$-\frac{1}{\lambda}\sigma_m\|v\| \leq -\frac{1}{\lambda}\|Bv\| \leq -\frac{1}{\lambda}\sigma_1\|v\|. \tag{2.39}$$

Substituindo (2.37)-(2.39) em (2.36) temos

$$\langle Aw, v\rangle = \lambda\|v\|^2 - \langle Av, v\rangle - \langle B^T y, v\rangle \geq \lambda\|v\|^2 - \mu_1\|v\|^2 - \frac{1}{\lambda}\sigma_m^2\|v\|^2,$$

portanto

$$\langle Aw, v\rangle \geq \lambda\|v\|^2 - \mu_1\|v\|^2 - \frac{1}{\lambda}\sigma_m^2\|v\|^2. \tag{2.40}$$

Por outro lado, fazendo o produto interno de (2.24) com w obtemos

$$\langle Ax, w\rangle + \langle B^T y, w\rangle = \langle \lambda x, w\rangle.$$

$$\implies \langle Av, w\rangle + \langle Aw, w\rangle = \lambda\langle v, w\rangle + \lambda\langle w, w\rangle - \langle B^T y, w\rangle.$$

Como $\langle v, w\rangle = 0$ e $\langle B^T y, w\rangle = y^T Bw = 0$, temos

$$\langle Av, w\rangle = \lambda\|w\|^2 - \langle Aw, w\rangle \leq \lambda\|w\|^2 - \mu_n\|w\|^2 = (\lambda - \mu_n)\|w\|^2.$$

Assim,

$$\langle Av, w\rangle \leq (\lambda - \mu_n)\|w\|^2. \tag{2.41}$$

Como $A = A^T$ se cumpre que $\langle Av, w\rangle = \langle Aw, v\rangle$, logo podemos juntar (2.40) e (2.41) obtendo

$$\lambda\|v\|^2 - \mu_1\|v\|^2 - \frac{1}{\lambda}\sigma_m^2\|v\|^2 \leq (\lambda - \mu_n)\|w\|^2.$$

Lembrando que $\lambda < 0$ e $\mu_n > 0$, o lado direito desta desigualdade fica negativo, então verifica-se que

$$\lambda\|v\|^2 - \mu_1\|v\|^2 - \frac{1}{\lambda}\sigma_m^2\|v\|^2 \leq 0$$

$$\implies \qquad (\lambda^2 - \lambda\mu_1 - \sigma_m^2)\|v\|^2 \geq 0.$$

Assim pode acontecer uma das seguintes opções:

1. $\lambda^2 - \lambda\mu_1 - \sigma_m^2 \geq 0$
2. $\|v\|^2 = 0$

Se acontecer 2. $\|v\|^2 = 0 \implies v = 0 \implies x = w$, então (2.24) e (2.25) ficam

$$Aw + B^T y = \lambda w, \tag{2.42}$$

$$Bw = \lambda y. \tag{2.43}$$

Como $w \in Nu(B)^\perp$ de (2.43) temos que $y = 0$, logo (2.42) fica $Aw = \lambda w \implies \lambda$ é um autovalor de A, mas isto não pode acontecer, como A é simétrica definida positiva, tem todos os seus autovalores positivos. Assim

$$\lambda^2 - \lambda\mu_1 - \sigma_m^2 \geq 0,$$

que fatorando resulta:

$$(\lambda - \lambda_1)(\lambda - \lambda_2) \geq 0, \tag{2.44}$$

onde

$$\lambda_1 = \frac{1}{2}(\mu_1 + \sqrt{\mu_1^2 + 4\sigma_m^2}\ )$$

e

$$\lambda_2 = \frac{1}{2}(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2}\ ).$$

Notemos que $\lambda_1 > 0$ e $\lambda_2 < 0$. Da desigualdade (2.44) temos duas opções

1. $\lambda \geq \lambda_1$ e $\lambda \geq \lambda_2$
2. $\lambda \leq \lambda_1$ e $\lambda \leq \lambda_2$

Se acontecer 1. temos $\lambda > 0$, o qual contradiz a hipóteses. Portanto temos

$$\lambda \leq \lambda_1 \quad \text{e} \quad \lambda \leq \lambda_2 \implies \lambda \leq \lambda_2.$$

Assim

$$\lambda \leq \frac{1}{2}(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2}\ ). \tag{2.45}$$

Finalmente, juntando (2.35) com (2.45) obtemos

$$\lambda \in [\frac{1}{2}(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\ ), \frac{1}{2}(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2}\ )]$$

o que conclui nossa prova. ■

Este último resultado pode ser estendido para o caso em que $C \neq 0$ é semi-definida positiva, mas nossa análise fica por aqui.

Estes limitantes para os autovalores são usados para medir a rapidez de convergência de alguns métodos iterativos aplicados ao sistema *ponto de sela*.

2.7 Métodos de resolução de sistemas Ponto de Sela

A solução algorítmica para problemas *ponto de sela* pode ser dividida em duas categorias: *métodos segregados* e *métodos acoplados*. Os métodos segregados calculam dois vetores desconhecidos x e y separadamente, dependendo do caso será determinado x ou y primeiro. Esta aproximação envolve a solução de dois sistemas lineares de tamanho menor que $n+m$, chamados *sistemas reduzidos*. Estes métodos podem ser diretos ou iterativos, ou envolver uma combinação dos dois. Os representantes mais importantes destes métodos são: o *método da redução do complemento de Schur*, que se baseia na fatoração LU de $\mathcal{A}$ e o *método do espaço nulo* que depende de uma base para o espaço nulo das restrições.

Os métodos acoplados acham x e y (ou aproximações deles) simultaneamente, sem fazer uso explícito do sistema reduzido. Estes métodos incluem resoluções diretas, baseadas na fatoração triangular da matriz $\mathcal{A}$ e resoluções iterativas como o método do subespaço de Krylov.

Após este breve resumo vamos apresentar os métodos mencionados anteriormente e alguns outros.

2.7.1 Redução ao Complemento de Schur

Do sistema (2.4) temos as seguintes equações

$$\begin{array}{rcrcl} Ax & + & B_1^T y & = & f, \\ B_2 x & + & (-C)y & = & g. \end{array} \tag{2.46}$$

Supomos A e $\mathcal{A}$ não singulares, e lembrando a fatoração de $\mathcal{A}$ temos

$$\begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B_2A^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} I & A^{-1}B_1^T \\ 0 & I \end{pmatrix}, \tag{2.47}$$

com $S = -(C + B_2A^{-1}B_1^T)$ o *complemento de Schur* de A em $\mathcal{A}$. Desta igualdade podemos inferir que S é inversível. Fazendo a multiplicação matricial com as duas últimas matrizes que aparecem no lado direito da igualdade temos

$$\begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B_2A^{-1} & I \end{pmatrix} \begin{pmatrix} A & B_1^T \\ 0 & S \end{pmatrix}. \tag{2.48}$$

Esta fatoração de $\mathcal{A}$ corresponde à sua fatoração LU, onde

$$L = \begin{pmatrix} I & 0 \\ B_2A^{-1} & I \end{pmatrix} \quad \text{e} \quad U = \begin{pmatrix} A & B_1^T \\ 0 & S \end{pmatrix}.$$

Voltemos ao sistema (2.46). Pré-multiplicando a primeira equação por B_2A^{-1} obtemos

$$B_2x + B_2A^{-1}B_1^Ty = B_2A^{-1}f$$

Como $B_2x = g + Cy$, a equação anterior fica

$$\begin{aligned} & Cy + B_2A^{-1}B_1^Ty = B_2A^{-1}f - g \\ \iff \qquad & (C + B_2A^{-1}B_1^T)y = B_2A^{-1}f - g. \end{aligned} \tag{2.49}$$

No lado direito deve-se resolver o sistema $Av = f$, a menos que f seja nula. É importante destacar que o sistema (2.49) tem uma única solução, pois a matriz $(C + B_2A^{-1}B_1^T) = -S$, sendo portanto inversível.
Uma vez encontrada a solução y^* de (2.49), x^* pode ser obtido resolvendo

$$Ax = f - B_1^Ty^*,$$

que é o sistema reduzido de ordem n. Estes dois sistemas para x e y podem ser obtidos usando a fatoração LU de $\mathcal{A}$. Como

$$\mathcal{A} = LU,$$

temos que

$$\begin{aligned} & \mathcal{A}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \implies L^{-1}\mathcal{A}\begin{pmatrix} x \\ y \end{pmatrix} = L^{-1}\begin{pmatrix} f \\ g \end{pmatrix} \\ \implies & U\begin{pmatrix} x \\ y \end{pmatrix} = L^{-1}\begin{pmatrix} f \\ g \end{pmatrix} \\ \implies & \begin{pmatrix} A & B_1^T \\ 0 & S \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} I & 0 \\ -B_2A^{-1} & I \end{pmatrix}\begin{pmatrix} f \\ g \end{pmatrix} \end{aligned}$$

$$\Longrightarrow \qquad \begin{pmatrix} A & B_1^T \\ 0 & S \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ -B_2 A^{-1} f + g \end{pmatrix}$$

Este é um sistema triangular superior por blocos, do qual podemos obter as equações apresentadas anteriormente e cuja resolução é mas simples que a do sistema original.
A solução destes dois sistemas pode ser achada por métodos diretos ou iterativos. Um caso especial surge quando A e $-S$ são simétricas definidas positivas, desta forma os métodos que podem ser implementados são a fatoração de *Cholesky* ou *Gradientes Conjugados*.
O método desenvolvido nesta seção é conhecido com diferentes nomes dependendo da área de pesquisa. Em otimização é conhecido como *método do espaço imagem* e é recomendável quando o ordem m do sistema reduzido é pequena e o sistema linear com a matriz A pode ser resolvido eficientemente.
As principais desvantagens do método são a necessidade de A ser inversível, e o complemento de Schur poderia ser completamente cheio e muito custoso de fatorar. Também podemos notar que ao formar S pode surgir instabilidade numérica, especialmente se A é mal condicionada.

2.7.2 Método do espaço nulo

Suponhamos $B_1 = B_2 = B$ com posto completo, $C = 0$ e $Nu(H) \cap Nu(B) = \{0\}$, onde H é a parte simétrica de A. Então o sistema ponto de sela fica

$$Ax + B^T y = f, \tag{2.50}$$

$$Bx \qquad = g. \tag{2.51}$$

Este método assume que são conhecidas:

1. uma solução particular x^* de (2.51);
2. a matriz $Z \in \mathbb{R}^{\mathrm{nx(n-m)}}$ tal que $Im(Z) = Nu(B)$, isto é $BZ = 0$.

Podemos observar que as colunas de Z geram $Nu(B)$, então sob estas hipóteses, Z tem posto coluna completo, ou seja, suas colunas são linearmente independentes.
Com estas duas hipóteses, o conjunto solução do sistema (2.51) é dado por $x = Zv + x^*$, com $v \in \mathbb{R}^{\mathrm{nx(n-m)}}$.
Substituindo esta expressão em (2.50) obtemos

$$\begin{array}{rcccl} & A(Zv + x^*) & + & B^T y & = & f \\ \Longrightarrow & AZv & + & Ax^* & = & f - B^T y, \end{array}$$

pré-multiplicando por Z^T, segue que

$$Z^T AZv + Z^T Ax^* = Z^T f - Z^T B^T y = Z^T f - (BZ)^T y = Z^T f.$$

Portanto

$$Z^T AZv = Z^T f - Z^T Ax^*,$$

ou seja

$$Z^T AZv = Z^T(f - Ax^*). \tag{2.52}$$

Proposição 2.3 *Sob as hipóteses estabelecidas, a matriz* $Z^T AZ$ *é inversível.*

Demonstração: Seja x solução do sistema

$$Z^T AZx = 0. \tag{2.53}$$

Como H e K representam a parte simétrica e anti-simétrica de A respectivamente, então $A = H + K$ e substituindo em (2.53) temos

$$Z^T(H+K)Zx = 0 \implies x^T Z^T(H+K)Zx = 0 \implies x^T Z^T HZx + x^T Z^T KZx = 0.$$

Lembrando que como K é anti-simétrica, então $y^T Ky = 0 \ \forall\, y$, assim a última igualdade fica

$$x^T Z^T HZx = 0 \implies Zx = 0 \implies x = 0.$$

Portanto a matriz $Z^T AZ$ é inversível. ■
Assim o sistema (2.52) tem uma única solução v^* . Uma vez que encontramos v^*, podemos calcular $x_{sol} = Zv^* + x^*$, e finalmente y_{sol} pode ser obtido substituindo x_{sol} em (2.50)

$$B^T y = f - Ax_{sol},$$

pré-multiplicando por B

$$BB^T y = B(f - Ax_{sol}).$$

Este sistema corresponde às equações normais do sistema sobredeterminado

$$B^T y = f - Ax_{sol},$$

que é obtido a partir do problema

$$\min_y \|(f - Ax_{sol}) - B^T y\|.$$

Uma vantagem deste método é não precisar calcular A^{-1}. De fato, o método só precisa da condição $Nu(H) \cap Nu(B) = \{0\}$.
A principal dificuldade deste método é achar uma matriz Z cujas colunas sejam uma base para $Nu(B)$.

Uma escolha desta matriz está baseada em encontrar uma base fundamental. Como estamos interesados em reduzir o custo computacional, esta estratégia utiliza a informação que é conhecida. A idéia é utilizar as colunas l.i.'s da matriz B para construir a matriz Z.

O processo é o seguinte:
Seja P a matriz de permutação tal que $BP = [B_b \;\; B_N]$, onde $B_b \in \mathbb{R}^{m\times m}$ é inversível e $B_N \in \mathbb{R}^{m\times n}$. Então a escolha para a matriz Z é:

$$Z = P \begin{pmatrix} -B_b^{-1}B_N \\ I \end{pmatrix}$$

Este tipo de base é chamada *"base fundamental"*. Só resta-nos achar uma solução para o sistema (2.51).
Como B_b é inversível, então o sistema $B_b x = g$ tem uma única solucão, seja x_b a solução e definamos

$$x^* = P \begin{pmatrix} x_b \\ 0 \end{pmatrix}.$$

Então

$$BPPx^* = \begin{pmatrix} B_b & B_N \end{pmatrix} \begin{pmatrix} x_b \\ 0 \end{pmatrix} = B_b x_b = g.$$

Portanto x^* é solução do sistema (2.51). Logo, x^* pode ser a solução que precisamos para aplicar o método.

Pode acontecer que esta escolha para Z seja mal condicionada, então outra forma de obtê-la é utilizando a fatoração QR de B^T, ou seja, fazer

$$B^T = Q\tilde{R}\ ,$$

com $Q \in \mathbb{R}^{n\times n}$ ortogonal, $Q = [Q_1 \;\; Q_2]$, $Q_1 \in \mathbb{R}^{n\times m}$, $Q_2 \in \mathbb{R}^{n\times(n-m)}$ e $\tilde{R} \in \mathbb{R}^{n\times m}$, $\tilde{R} = \begin{pmatrix} R \\ 0 \end{pmatrix}$, $R \in \mathbb{R}^{n\times n}$ inversível. Fazendo as substituções temos

$$B = \tilde{R}Q^T = \begin{pmatrix} R^T & 0 \end{pmatrix} \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} = R^T Q_1^T$$

Logo,

$$Bx = g \iff R^T Q_1^T x = g \implies Q_1^T x = R^{-T} g \implies x = Q_1 R^{-T} g$$

Assim

$$x^* = Q_1 R^{-T} g,$$
$$Z = Q_2.$$

2.7.3 Métodos diretos acoplados

Consideremos o caso onde $A = A^T$, $B_1 = B_2 = B$ e $C = C^T$ (possivelmente nula). Há várias maneiras de desenvolver a eliminação Gaussiana de uma matriz simétrica, possivelmente indefinida, utilizando e preservando a simetria.
A fatoração da forma

$$\mathcal{A} = \mathcal{Q}^{\mathcal{T}} \widetilde{\mathcal{L}} \mathcal{D} \widetilde{\mathcal{L}}^{\mathcal{T}} \mathcal{Q},$$

onde $\mathcal{Q}$ é matriz de permutação, $\widetilde{\mathcal{L}}$ triangular inferior unitária e $\mathcal{D}$ uma matriz diagonal em blocos, com blocos de dimensão 1 e 2, é geralmente conhecida como uma fatoração LDL^T. A idéia foi desenvolvida por Bunch e Parlett (1971), obtendo um algoritmo mais estável para a fatoração de matrizes simétricas indefinidas a um custo comparável à fatoração de Cholesky.

Para exemplificar, consideremos o caso em que A é simétrica definida positiva, B com posto completo e $C = 0$. Vamos a ver que sob estas hipóteses $\mathcal{A}$ admite uma fatoração LDL^T com $\mathcal{D}$ diagonal e $\mathcal{Q} = I$.
Como $A > 0$ e $A = A^T$, podemos escrever $A = L_A D_A L_A^T$, com L_A matriz triangular inferior unitária, D_A matriz diagonal e definida positiva. Por outro lado o complemento de Schur $S = -(C + BA^{-1}B^T)$ é simétrico definido negativo, então admite a fatoração $S = -L_S D_S L_S^T$, portanto,

$$\mathcal{A} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} I & A^{-1}B^T \\ 0 & I \end{pmatrix} =$$

$$\begin{pmatrix} I & 0 \\ BL_A^{-T}D_A^{-1}L_A^{-1} & I \end{pmatrix} \begin{pmatrix} L_A D_A L_A^T & 0 \\ 0 & L_S D_S L_S^T \end{pmatrix} \begin{pmatrix} I & L_A^{-T}D_A^{-1}L_A^{-1}B^T \\ 0 & I \end{pmatrix} =$$

$$\begin{pmatrix} L_A & 0 \\ BL_A^{-T}D_A^{-1} & L_S \end{pmatrix} \begin{pmatrix} D_A & 0 \\ 0 & D_S \end{pmatrix} \begin{pmatrix} L_A^T & D_A^{-1}L_A^{-1}B^T \\ 0 & L_S^T \end{pmatrix} = \widetilde{\mathcal{L}}\mathcal{D}\widetilde{\mathcal{L}}^{\mathcal{T}}$$

Logo $\mathcal{A} = \widetilde{\mathcal{L}}\mathcal{D}\widetilde{\mathcal{L}}^{\mathcal{T}}$ onde $\widetilde{\mathcal{L}}$ é triangular inferior e $\mathcal{D}$ é diagonal em blocos.
Este tipo de fatoração é amplamente escolhido para trabalhar com matrizes indefinidas densas e simétricas. Nos casos em que a matriz é esparsa, diversos códigos foram desenvolvidos para preservar a esparsidade na matriz $\widetilde{\mathcal{L}}$, os quais fazem parte da biblioteca *Harwell Subroutine Library* (HSL). Um desses códigos é o MA57, que foi utilizado neste trabalho.

2.7.4 Iterações estacionárias

Dado o sistema ponto de sela

$$\begin{array}{rcrcl} Ax & + & B_1^T y & = & f, \\ B_2 x & - & Cy & = & g, \end{array}$$

estes métodos consistem de iterações simultâneas de x e y . Elas podem ser obtidas de equações que surgem de separar a matriz do sistema. Por eliminação de uma das incógnitas eles podem ser interpretados como iterações de um sistema reduzido. Assim obtemos algoritmos de resoluções acoplados ou separados.
Dois métodos que se destacam aqui são os métodos de *Arrow-Hurwicz* e *Uzawa*. Detalharemos as suas propostas a seguir:

Método de Uzawa

Por simplicidade vamos assumir que A é inversível, $B_1 = B_2 = B$ e $C = 0$. Começando com um valor inicial x_0, y_0 este método consiste na seguinte iteração acoplada

$$Ax_{k+1} = f - B^T y_k, \tag{2.54}$$

$$y_{k+1} = y_k + w(Bx_{k+1} - g), \tag{2.55}$$

onde $w > 0$ é um parâmetro de relaxação. O processo iterativo pode ser obtido em termos da matriz $\mathcal{A}$ através de iterações de ponto fixo da seguinte maneira:

Dada $\mathcal{A}$, dividimos esta matriz como $\mathcal{A} = P - Q$ com P inversível e consideramos o sistema

$$\mathcal{A}\mu = b,$$

substituindo temos

$$(P - Q)\mu = b$$
$$\implies \quad P\mu = b + Qx \quad \implies \quad \mu = P^{-1}(b + Q\mu) = P^{-1}b + P^{-1}Q\mu$$

logo usando a iteração de ponto fixo temos:

$$\mu_{n+1} = P^{-1}b + P^{-1}Q\mu_n$$

ou

$$P\mu_{n+1} = b + Q\mu_n.$$

Escolhendo

$$P = \begin{pmatrix} A & 0 \\ B & -\frac{1}{w}I \end{pmatrix}, \quad Q = \begin{pmatrix} 0 & -B^T \\ 0 & -\frac{1}{w}I \end{pmatrix}, \quad \mu_k = \begin{pmatrix} x_k \\ y_k \end{pmatrix}, \quad \text{e} \quad b = \begin{pmatrix} f \\ g \end{pmatrix}$$

e fazendo as operações correspondentes obtemos o sistema (2.54)-(2.55).
Por outro lado, se usarmos a equação (2.54) para isolar x_{k+1} obtemos

$$x_{k+1} = A^{-1}(f - B^T y_k)$$

e substituindo na equação (2.55) temos que

$$y_{k+1} = y_k - w[g - BA^{-1}(f - B^T y_k)]$$

o que é equivalente a aplicar a iteração de Richardson

$$z_{n+1} = z_n + \alpha(d - Mz_n)$$

com $\alpha \geq 0$, ao sistema do *complemento de Schur*

$$BA^{-1}B^T y = BA^{-1}f - g$$

onde $z_j = y_j$, $\alpha = w$, $M = BA^{-1}B^T$ e $d = BA^{-1}f - g$.

Um resultado interessante que podemos obter, em relação à convergência do método, surge de considerar A é simétrica. Sob esta hipótese matriz $BA^{-1}B^T$ também resulta simétrica. Chamando λ_{max} e λ_{min} os autovalores máximo e mínimo respetivamente de $BA^{-1}B^T$, a iteração de Richardson converge se $0 < w < \frac{2}{\lambda_{max}}$ (ver Proposição A.5), o que nos diz que o método de Uzawa converge sob estas hipóteses.

Método de Arrow-Hurwicz

Este método pode ser considerado como uma alternativa menos custosa computacionalmente que o método de Uzawa.
A solução da equação (2.54) é o minimizador da função objetivo

$$\phi(x) = \frac{1}{2}\, x^T\, Ax - x^T(f - B^T y_k).$$

Agora para conseguir essa economia se propõe a dar um passo na direção do gradiente negativo de $\phi(x)$, com um tamanho α, ou seja, dar um passo na direção do método da máxima descida. Logo a iteração do método fica

$$x_{k+1} = x_k + \alpha(f - B^T y_k - Ax_k),$$

$$y_{k+1} = y_k + w(Bx_{k+1} - g).$$

De modo similar ao método de Uzawa, este método pode ser deduzido a partir de uma iteração de ponto fixo, separando a matriz $\mathcal{A}$ como

$$\mathcal{A} = P - Q$$

onde

$$P = \begin{pmatrix} \frac{1}{\alpha} & 0 \\ B & -\frac{1}{w}I \end{pmatrix}, \quad Q = \begin{pmatrix} \frac{1}{\alpha}I - A & -B^T \\ 0 & -\frac{1}{w}I \end{pmatrix}.$$

A convergência deste método depende dos parâmetros de relaxação α e w.

2.7.5 Método de penalidade

Assumamos $A = A^T$ semi-definida positiva, $B_1 = B_2 = B$ com posto completo, $C = 0$ e $Nu(A) \cap Nu(B) = \{0\}$, portanto, o sistema (2.4) tem solução única. Como já foi apresentado, encontrar a solução do sistema ponto de sela é equivalente a minimizar o seguinte problema com restrições:

$$\min \quad J(x) = \frac{1}{2}\, x^T Ax - f^T x, \tag{2.56}$$

$$\text{suj. a} \qquad Bx = g. \tag{2.57}$$

Um método para resolver este problema é o *"método de penalização quadrática"*. Este método consiste em substituir o problema original por uma seqüência de subproblemas sem restrições. A função a ser minimizada em cada subproblema, chamada *função de penalidade*, é formada por duas partes: a função original mais um termo positivo que penaliza o fato de um ponto não ser factível. Assim cada subproblema a ser resolvido é

$$\min \hat{J}(x) = \min \;\; J(x) + \frac{\gamma}{2}\|Bx - g\|^2.$$

com $\gamma > 0$. Como $Nu(A) \cap Nu(B) = \{0\}$, a função $\hat{J}(x)$ é estritamente convexa, portanto tem uma única solução $x(\gamma)$. Logo, é gerada uma seqüência de minimizadores de $\hat{J}(x(\gamma))$ cujo ponto limite é solução do problema original (Nocedal (1999)), isto é

$$\lim_{\gamma \to \infty} x(\gamma) = x^*,$$

onde x^* é solução do problema original. O minizador $x(\gamma)$ pode ser obtido igualando o gradiente de $\hat{J}(x)$ a zero, ficando o sistema linear

$$(A + \gamma B^T B)x = f + \gamma B^T g. \tag{2.58}$$

Chamando $y(\gamma) = f + \gamma B^T g$, e considerando $x(\gamma)$ solução de (2.58) cumpre-se a seguinte proposição:

Proposição 2.4

$$\|x^* - x(\gamma)\| = O(\gamma^{-1}) \qquad \text{e} \qquad \|y^* - y(\gamma)\| = O(\gamma^{-1}) \quad \text{para} \quad \gamma \longrightarrow \infty$$

(Glowinski (1984), pág. 21-22)

Demonstração:

$(A + \gamma B^T B)x = f + \gamma B^T g \quad \Longrightarrow \quad Ax + \gamma B^T(Bx - g) = f$. Chamando $y = \gamma(Bx - g)$, temos que:

$$Ax + B^T y = f. \tag{2.59}$$

Por outro lado $y = \gamma(Bx - g) \quad \Longrightarrow \quad \gamma^{-1} y = Bx - g$, portanto

$$Bx - \gamma^{-1} y = g. \tag{2.60}$$

Logo juntando (2.59) e (2.60) e chamando $\epsilon = \gamma^{-1}$, obtemos o sistema linear ponto de sela

$$\begin{pmatrix} A & B^T \\ B & -\epsilon \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}. \tag{2.61}$$

Seja

$$F(x, y, \epsilon) = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \begin{pmatrix} Ax + B^T y - f \\ Bx - \epsilon y - g \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

O ponto $\begin{pmatrix} x \\ y \end{pmatrix}$ que satisfaz (2.61) é solução de

$$F(x, y, \epsilon) = 0.$$

Derivando implicitamente com respeito a ϵ temos

$$\frac{\partial F_1}{\partial \epsilon} \;=\; Ax'(\epsilon) \;+\; B^T y'(\epsilon) \qquad\qquad =\; 0, \tag{2.62}$$

$$\frac{\partial F_2}{\partial \epsilon} \;=\; Bx'(\epsilon) \;-\; \epsilon y'(\epsilon) \;-\; y(\epsilon) = \; 0. \tag{2.63}$$

Observe que se $(x^*,\ y^*)$ é solução do sistema ponto de sela com $x(\epsilon) \longrightarrow x^*$ e $y(\epsilon) \longrightarrow y^*$ quando $\epsilon \longrightarrow 0$, então $(x^*, y^*, 0)$ é solução de $F(x, y, \epsilon) = 0$, e também $x(0) = x^*$ e $y(0) = y^*$.
Fazendo o desenvolvimento de Taylor de $x(\epsilon)$ e $y(\epsilon)$ em torno de $\epsilon = 0$ temos

$$x(\epsilon) = x^* + x'(0)\epsilon + O(\epsilon^2) \implies x(\epsilon) - x^* = x'(0)\epsilon + O(\epsilon^2),$$

$$y(\epsilon) = y^* + y'(0)\epsilon + O(\epsilon^2) \implies y(\epsilon) - y^* = y'(0)\epsilon + O(\epsilon^2).$$

Voltando às equaçoẽs (2.62) e (2.63) escolhendo $\epsilon = 0$ temos

$$Ax'(0) + B^T y'(0) = 0,$$

$$Bx'(0) - y(0) = 0,$$

ou

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x'(0) \\ y'(0) \end{pmatrix} = \begin{pmatrix} 0 \\ y(0) \end{pmatrix}. \tag{2.64}$$

Como $Nu(A) \cap Nu(B) = \{0\}$, o sistema ponto de sela (2.64) tem solução única e $x'(0) = 0 = y'(0)$ se, e somente se, $y(0) = 0$. Logo

$$\|x(\epsilon) - x^*\| = O(\epsilon), \qquad \|y(\epsilon) - y^*\| = O(\epsilon)$$

e fica provada a afirmação. ■

Uma desvantagem deste método é que o sistema linear (2.58) é mal condicionado. O número de condição da matriz $A + \gamma B^T B$ aumenta com γ. Uma maneira de solucionar este problema é trocar o método de penalidade pelo *Método dos Multiplicadores*, obtendo o método conhecido como o *Método do Lagrangiano Aumentado*. Os detalhes deste método serão desenvolvidos no próximo capítulo.

2.7.6 Método do Subespaço de Krylov

Um *método geral de projeção* para resolver sistemas lineares do tipo

$$Ax = b$$

obtém uma solução aproximada de um subespaço de $\mathbb{R}^n$, que denotaremos por $\mathcal{K}$. Se a sua dimensão for m, então precisamos de m restrições para obter uma única solução nesse subespaço. Uma maneira tipica para obter essas restrições é impor m condições de ortogonalidade. Mais especificamente, impomos que o vetor resíduo $b - Ax$ seja ortogonal a m vetores linearmente independentes. Estes vetores definem outro subespaço de dimensão m, chamado *subespaço das restrições*.

Resumindo, estes métodos obtêm uma solução aproximada x_m do sistema $Ax = b$, de um subespaço afim $x_0 + \mathcal{K}_m$ de dimensão k, impondo a condição conhecida como condição de Petrov-Galerkin

$$b - Ax_m \perp \mathcal{C}_m,$$

onde $\mathcal{C}_m$ é outro subespaço de dimensão m, x_0 é um ponto inicial arbitrário que aproxima a solução e $b - Ax_m$ é o resíduo no passo m que denotaremos por r_m. Em cada passo atualizam-se $\mathcal{K}_m$, $\mathcal{C}_m$ e x_0 é o último ponto x_m achado.
O método do subespaço de Krylov, é um método que está baseado neste tipo de idéia. O subespaço $\mathcal{K}_m$ é o *subespaço de Krylov*

$$\mathcal{K}_m(A, r_0) = span\{r_0, Ar_0, A^2r_0, \ldots, A^{m-1}r_0\}.$$

Analisemos como seria este método. Lembrando, o método geral de projeção gera uma seqüência tal que o k-ésimo iterando satizfaz

$$u_k \in u_{k-1} + \mathcal{K}_k \tag{2.65}$$

então,

$$u_k = a_k u_{k-1} + \sum_{i=0}^{k-1} \alpha_i^{(k-1)} A^i r_0, \tag{2.66}$$

com a_k e $\alpha_i^{(k-1)}$ constantes reais, e por outro lado

$$u_{k-1} \in u_{k-2} + \mathcal{K}_{k-1} \implies u_{k-1} = a_{k-1}u_{k-2} + \sum_{i=0}^{k-2} \alpha_i^{(k-2)} A^i r_0.$$

Substituindo em (2.66) e agrupando convenientemente, obtemos:

$$u_k = a_k a_{k-1} u_{k-2} + \sum_{i=0}^{k-1} \beta_i^{(k-1)} A^i r_0,$$

e continuando com este processo, no passo k teremos

$$u_k = c\, u_0 + \sum_{i=0}^{k-1} \gamma_i A^i r_0,$$

onde c e $\gamma_i \ \ \forall\, i$ são constantes reais, obtidas fazendo as correspondentes operações de soma e multiplicação. Logo

$$u_k \in u_0 + span\{r_0, Ar_0, \ldots, A^{k-1}r_0\}.$$

***Resumindo*:**
Suponhamos que u_0 é uma solução aproximada do sistema $\mathcal{A}u = b$ e seja $r_0 = b - \mathcal{A}u_0$ o

resíduo inicial. Os métodos do subspaço do Krylov são métodos iterativos tais que o k-ésimo iterando satisfaz

$$u_k \in u_0 + \mathcal{K}_k(\mathcal{A}, r_0), \qquad \text{com} \quad k = 1, 2, \ldots.$$

Observar:
$\mathcal{K}_j(A, r_0) \subseteq \mathcal{K}_{j+1}(A, r_0)\ \forall\ m \in \mathbb{N}$. Assim, cada subespaço de Krylov contém os subespaços obtidos nas iterações anteriores.

Proposição 2.5 *Se dim* $\mathcal{K}_{n+m}(A, r_0) = d \leq n + m$, *então* $r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{d-1}r_0$ *são linearmente independentes.*

Demonstração: Suponhamos $r_0 \neq 0$. Se $r_0 = 0$, a tese cumpre-se trivialmente. Seja $k \geq 1$ tal que $E = \{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{k-1}r_0\}$ seja um conjunto linearmente independente, $E \neq \emptyset$ pois $E = \{r_0\}$ é um conjunto linearmente independente.
Agora podem acontecer duas opções:

1. $k = d$,
2. $k \leq d - 1$.

Se acontecer a primeira opção, não precisamos provar nada.
Analisemos o segundo caso. Vamos provar que $r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^k r_0$, são linearmente independentes usando indução sobre k.
Com $k = 1$, o conjunto formado por r_0 é linearmente independente. Pela hipótese indutiva suponhamos que $r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{k-1}r_0$ formam um conjunto linearmente independente. Para provar a tese indutiva supomos que $r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^k r_0$ são linearmente dependentes, então

$$\mathcal{A}^k r_0 = \sum_{i=0}^{k-1} \alpha_i \mathcal{A}^i r_0 \implies \mathcal{A}^k r_0 \in span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{k-1}r_0\}.$$

Multiplicando por $\mathcal{A}$ ambos lados da igualdade anterior temos

$$\mathcal{A}^{k+1} r_0 = \sum_{i=0}^{k-1} \alpha_i \mathcal{A}^{i+1} r_0 = \sum_{i=0}^{k-2} \alpha_i \mathcal{A}^{i+1} r_0 + \alpha_{k-1} \mathcal{A}^k r_0$$

substituindo $\mathcal{A}^k r_0$, segue que

$$\mathcal{A}^{k+1} r_0 = \sum_{i=0}^{k-2} \alpha_i \mathcal{A}^{i+1} r_0 + \alpha_{k-1} (\sum_{i=0}^{k-1} \alpha_i \mathcal{A}^i r_0)$$

e agrupando convenientemente, podemos escrever

$$\mathcal{A}^{k+1} r_0 = \sum_{i=0}^{k-1} \beta_i \mathcal{A}^i r_0,$$

onde $\beta_i \in \mathbb{R}$ $\forall$ i, obtidas fazendo as correspondentes operações de soma e multiplicação. Portanto,

$$\mathcal{A}^{k+1}r_0 \in span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{k-1}r_0\}.$$

Continuando com este processo teremos

$$\mathcal{A}^{\ell}r_0 = \sum_{i=0}^{k-1} \gamma_i \mathcal{A}^i r_0 \quad \forall\ \ell \quad \Longrightarrow \quad \mathcal{A}^{\ell}r_0 \in span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{k-1}r_0\} \quad \forall\ \ell.$$

Portanto

$$span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{n+m}r_0\} = span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{k-1}r_0.\}$$

Como $span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{k-1}r_0\}$ é um conjunto linearmente independente pela hipótese indutiva, este conjunto tem dimensão k. Logo

$$\Longrightarrow \quad dim\ \mathcal{K}_{n+m}(\mathcal{A}, r_0) = k < d.$$

Mas isto contradiz a hipótese. Assim o conjunto $span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^k r_0\}$ é linearmente independente $\forall\ k \leq d-1$, em particular verifica-se para $k = d-1$, com o qual fica provada a proposição. ■

Logo $span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{d-1}r_0\} = span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{n+m}r_0\}$. Então,

$$\mathcal{K}_1(\mathcal{A}, r_0) \subseteq \ldots. \subseteq \mathcal{K}_d(\mathcal{A}, r_0) = \ldots. = \mathcal{K}_{n+m}(\mathcal{A}, r_0).$$

Para cada $\ell \leq d$ verifica-se que $dim\ \mathcal{K}_\ell(\mathcal{A}, r_0) = \ell$. Logo precisam-se ℓ restrições para fazer u_ℓ único. Nos métodos do subespaço de Krylov isto é obtido pedindo que o k-ésimo resíduo

$$r_k = b - \mathcal{A}u_k$$

seja ortogonal a um subespaço C_k de dimensão k, chamado *subespaço de restrições*. Logo

$$r_k = b - \mathcal{A}u_k \in r_0 + \mathcal{A}\mathcal{K}_k(\mathcal{A}, r_0), \quad r_k \perp \mathcal{C}_k,$$

onde a ortogonalidade é medida com respeito ao produto interno euclidiano.

A seguir faremos uma breve análise. Sejam

$$\{v_1, \ldots, v_k\} \quad \text{uma base para } \mathcal{K}_k, \text{ e definamos } V = [v_1\ \ldots\ v_k], \quad V \in \mathbb{R}^{(n+m)\times k}$$

e

$$\{w_1, \ldots, w_k\} \quad \text{uma base para } \mathcal{C}_k, \text{ e definamos } W = [w_1\ \ldots\ w_k], \quad W \in \mathbb{R}^{(n+m)\times k}$$

onde w_i são as colunas da matriz W.
Sabemos que se $u_k \in u_0 + \mathcal{K}_k \implies \exists\, y \in \mathbb{R}^k$ tal que $u_k = u_0 + Vy$. Multiplicando por $\mathcal{A}$ ambos lados desta útima igualdade obtemos

$$\mathcal{A}u_k = \mathcal{A}u_0 + \mathcal{A}Vy.$$

Multiplicando por (-1) ambos os lados e somando b obtemos

$$b - \mathcal{A}u_k = b - \mathcal{A}u_0 - \mathcal{A}Vy.$$

Então

$$r_k = r_0 - \mathcal{A}Vy. \tag{2.67}$$

Por outro lado $r_k \perp \mathcal{C}_k \implies r_k \perp w_i \ \forall\, i = 1, \ldots, k \implies w_i^T r_k = 0 \,\forall\, i \implies W^T r_k = 0.$ Logo, multiplicando por W ambos os lados de (2.67) obtemos

$$W^T r_k = W^T r_0 - W^T \mathcal{A}Vy,$$

ou

$$0 = W^T r_0 - W^T \mathcal{A}Vy$$

implicando que

$$W^T \mathcal{A}Vy = W^T r_0.$$

Se $W^T\mathcal{A}V$ é inversível, então este último sistema linear tem solução única e, neste caso, u_k ficaria univocamente determinado (Saad (2003), pág. 131). ■

Teorema 2.7 *Assumamos que o subespaço de Krylov $\mathcal{K}_k(\mathcal{A}, r_0)$ tem dimensão k. Se*

1. *$\mathcal{A}$ é simétrica definida positiva e $\mathcal{C}_k = \mathcal{K}_k(\mathcal{A}, r_0)$ ou*
2. *$\mathcal{A}$ é não singular e $\mathcal{C}_k = \mathcal{A}\mathcal{K}_k(\mathcal{A}, r_0)$*

então existe um único u_k tal que $u_k \in u_0 + \mathcal{K}_k(\mathcal{A}, r_0) \;\; k = 1, 2, \ldots$ para o qual $r_k = b - \mathcal{A}u_k$ é ortogonal a $\mathcal{C}_k$.

(Saad (2003), pág. 132)

Demonstração: Aqui vamos a provar que $W^T\mathcal{A}V$ é inversível para quaisquer W e V bases de $\mathcal{K}_k$ e $\mathcal{C}_k$ respectivamente, e pela análise prévia temos garantido que u_k definido pelo teorema existe e é único.
Suponhamos que verifica-se (*1.*) $\mathcal{C}_k = \mathcal{K}_k(\mathcal{A}, r_0)$. Do mesmo jeito que na análise prévia, seja $\{v_1, \ldots, v_k\}$ alguma base de $\mathcal{K}_k$ e $V = [v_1 \ \ldots \ v_k]$ matriz cujas colunas são os vetores v_i e seja $\{w_1, \ldots, w_k\}$ alguma base de $\mathcal{C}_k$ e $W = [w_1 \ \ldots \ w_k]$ matriz cujas colunas são os vetores w_i. Como $\mathcal{C}_k = \mathcal{K}_k(\mathcal{A}, r_0) \implies \{w_i\}$ é base de $\mathcal{K}_k \;\exists\; G \in \mathbb{R}^{k\times k}$ inversível tal que $W = VG$. Logo $W^T\mathcal{A}V = G^T V^T \mathcal{A}V$.
Agora $\mathcal{A}$ é definida positiva, então $x^T V^T \mathcal{A}Vx > 0 \;\forall\; Vx \neq 0$. Como V tem posto completo, se $Vx = 0$ então $x = 0$. Portanto $x^T V^T \mathcal{A}Vx > 0 \;\forall\; x \neq 0$, e assim $V^T\mathcal{A}V$ é inversível.
Por outro lado G é inversível, então $G^T V^T \mathcal{A}V$ é inversível, portanto $W^T\mathcal{A}V$ é inversível e como $\{v_i\}$ e $\{w_i\}$ são arbitrárias, verificam-se as hipóteses.
Suponhamos que a condição (*2.*) seja satisfeita: $\mathcal{C}_k = \mathcal{A}\mathcal{K}_k(\mathcal{A}, r_0)$. Sejam V e W como no caso anterior,

$$V = [v_1, \ldots, v_k] \implies \mathcal{A}V = [\mathcal{A}v_1, \ldots, \mathcal{A}v_k]$$

e

$$\mathcal{A}v_i \in \mathcal{A}\mathcal{K}_k \ \forall\, i \implies \mathcal{A}v_i \in \mathcal{C}_k \ \forall i.$$

Como $\mathcal{A}$ é inversível e $\{v_1, \dots v_k\}$ é linearmente independente então $\mathcal{A}v_1, \dots, \mathcal{A}v_k$ é linearmente independente. Assim $\{\mathcal{A}v_i\}_{i=1}^k$ são k vetores linearmente independentes em $\mathcal{C}_k$, portanto formam uma base para $\mathcal{C}_k$, então $\exists\ G \in \mathbb{R}^{k\times k}$ inversível tal que $W = \mathcal{A}VG$. Logo $W^T\mathcal{A}V = G^TV^T\mathcal{A}^T\mathcal{A}V$.
Agora seja x tal que

$$V^T\mathcal{A}^T\mathcal{A}Vx = 0 \implies x^TV^T\mathcal{A}^T\mathcal{A}Vx = 0 \iff \|\mathcal{A}Vx\|^2 = 0$$

$$\iff \mathcal{A}Vx = 0 \iff Vx = 0 \implies x = 0$$

portanto $V^T\mathcal{A}^T\mathcal{A}V$ é inversível, $\implies$ $G^TV^T\mathcal{A}^T\mathcal{A}V$ é inversível $\implies$ $W^T\mathcal{A}V$ é inversível e como $\{v_i\}$ e $\{w_i\}$ são arbitrárias, as hipóteses são satisfeitas. ■

O item *1.* no teorema anterior, caracteriza o *método dos Gradientes Conjugados* e o item *2.* caracteriza o *método de Resíduos Mínimos* (MINRES).

Observe que na demostração anterior não usamos o fato da matriz A ser simétrica. Portanto, o *Teorema* 2.7 pode ser generalizado para o caso de matrizes não simétricas, com demonstração semelhante à que fizemos.

Análise de convergência

Em aritmética exata, os métodos definidos por *1.* e *2.* do *Teorema* 2.7 para matrizes $\mathcal{A}$ simétricas não singulares encontram a solução no máximo no passo $d = dim\ \mathcal{K}_{n+m}(\mathcal{A}, r_0)$, ou seja $u_d = u^*$, com

$$\mathcal{K}_{n+m}(\mathcal{A}, r_0) = span\{r_0, \mathcal{A}r_0, \dots, \mathcal{A}^{d-1}r_0\}$$

observado na seção anterior.
Mais ainda, o resultado para o método dos Gradientes Conjugados, pode ser generalizado para matrizes não simétricas. Neste caso temos:

$$\mathcal{K}_k(\mathcal{A}, r_0) = \{r_0, \mathcal{A}r_0, \dots, \mathcal{A}^{k-1}r_0\}, \qquad \mathcal{C}_k = \mathcal{K}_k,$$

$$r_k \in r_0 + \mathcal{A}\mathcal{K}_k(\mathcal{A}, r_0) = r_0 + span\{\mathcal{A}r_0, \dots, \mathcal{A}^kr_0\} \subseteq span\{r_0, \mathcal{A}r_0, \dots, \mathcal{A}^kr_0\}$$
$$\implies r_k \in span\{r_0, \mathcal{A}r_0, \dots, \mathcal{A}^kr_0\}.$$

Por outro lado temos que

$$r_k \perp \mathcal{C}_k \quad \text{ou} \quad r_k \perp span\{r_0, \mathcal{A}r_0, \dots, \mathcal{A}^{k-1}r_0\}.$$

Agora, suponhamos $r_k \neq 0\ \ \forall k = 1, \dots, d,$ no passo d temos:

$$r_d \perp span\{r_0, \mathcal{A}r_0, \dots, \mathcal{A}^{d-1}r_0\} \tag{2.68}$$

$$r_d \in span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^d r_0\} = span\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{d-1} r_0\} \tag{2.69}$$

logo juntando (2.68) e (2.69) temos que $r_d = 0$. Assim o método converge no máximo em $d = dim\ \mathcal{K}_{n+m}(\mathcal{A}, r_o)$ passos (ver Luenberger).

No próximo capítulo faremos uma descrição do método do Lagrangiano Aumentado aplicado a problemas de Programação Não Linear (PNL).

Capítulo 3

Lagrangiano Aumentado

3.1 Introdução

Um problema de Programação Não Linear (PNL) é definido como:

$$\min_{x\in\Omega} \quad f(x) \tag{3.1}$$

$$suj.\ a\ \ h(x) = 0 \tag{3.2}$$

$$g(x) \leq 0 \tag{3.3}$$

com $\Omega \subseteq \mathbb{R}^n$, $f : \Omega \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$, $h : \Omega \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^m$ e $g : \Omega \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^p$, funções contínuas e deriváveis.

O método desenvolvido neste capítulo faz parte da grande classe de métodos desenvolvidos para achar a solução deste tipo de problemas. Em anos recentes estes métodos perderam popularidade, pois surgiram métodos de convergência mais rápida. No entanto, eles tem propriedades muito boas que fazem que sua implementação esteja vigente. Como por exemplo, a convergência a minimizadores globais não precisa da diferenciabilidade das funções que definem o problema de Programação Não Linear. Outra característica importante é que devido à esparsidade do sistema KKT, pode ser muito difícil a fatoração da matriz do sistema. Este fato dificulta a resolução dos problemas mediante métodos baseados em Newton, mas isto não apresenta uma dificuldade para alguns métodos tipo Lagrangiano Aumentado. Assim como as mencionadas, outras vantagens muito importantes podem ser encontradas na literatura referentes a este método (ver R. Andreani, et al. (2007)).

Os métodos de Lagrangiano Aumentado consistem em substituir o problema original pela resolução de uma seqüência de subproblemas mais simples. Em cada subproblema são fixados o parâmetro de penalidade e os multiplicadores de Lagrange, sendo atualizados em cada iteração externa.

Nestes processos, são combinados os métodos de penalidade com os multiplicadores de Lagrange. Uma vantagem sobre os métodos de penalidade, é que o parâmetro de penalidade não precisa ser aumentado em cada iteração externa, fato que reduz o mal condicionamento dos subproblemas.

Muitos métodos tipo Lagrangiano Aumentado surgem, dependendo da função de penalização utilizada. O algoritmo com melhor desempenho para minimização com restrições de desigualdade é o método Powell-Hestenes-Rockafellar (PHR) (E. G. Birgin et al. (2005)).

Dado $\rho > 0$, parâmetro de penalidade, a função Lagrangiano Aumentado PHR (Power-Hestenes-Rockafellar) define-se como:

$$\mathcal{L}_\rho(x,\lambda,\mu) = f(x) + \frac{\rho}{2}\left\{\sum_{i=1}^{m}\left[h_i(x) + \frac{\lambda_i}{2}\right]^2 + \sum_{i=1}^{p}\max\left(0, g_i(x) + \frac{\mu_i}{\rho}\right)^2\right\}. \tag{3.4}$$

3.2 Algoritmo - Lagrangiano Aumentado

Algoritmo 3.1 ALGENCAN

Sejam $\lambda_{min} < \lambda_{max}$, $\mu_{max} > 0$, $\tau > 1$, $0 < \rho < 1$, $\{\epsilon_k\}$ *uma seqüência de números positivos tal que* $\lim_{k\to\in\infty}\epsilon_k = 0$, $\lambda_i^1 \in [\lambda_{min}, \lambda_{max}]$ $i = 1,..,m$, $\mu_i^1 \in [0, \mu_{max}]$ $i = 1,..,p$, $\rho_1 > 0$ *e* $x_0 \in \Omega = \{x \in \mathbb{R}^n \ / \ l \leq x \leq u\}$ *um ponto inicial arbitrário .*
Fazer $k \longleftarrow 1$.

Passo 1 *Achar o iterando* x_k *que é solução do subproblema*

$$\min_{x\in\Omega} \ \mathcal{L}_{\rho_k}(x, \lambda_k, \mu_k)$$

x_k *deve satisfazer:*

$$\|P_\Omega(x_k - \nabla_x\mathcal{L}_{\rho_k}(x, \lambda_k, \mu_k)) - x_k\|_\infty \leq \epsilon_k,$$

onde P_Ω *é a projeção Euclidiana sobre* Ω.

Passo 2 *Definir*

$$V_i^k = \max\left\{g_i(x_k), -\frac{\mu_i^k}{\rho_i^k}\right\} \ i = 1,..,p.$$

Se $k = 1$ *ou*

$$\max\left\{\|h(x_k)\|_\infty, \ \|V^k\|_\infty\right\} \leq \tau\max\left\{\|h(x_{k-1})\|_\infty, \ \|V^{k-1}\|_\infty\right\}$$

fazer $\rho_{k+1} = \rho_k$. *Caso contrário, fazer* $\rho_{k+1} = \tau\rho_k$.

Passo 3 *Calcular* $\lambda_i^{k+1} \in [\lambda_{min},\ \lambda_{max}]\ i = 1, .., m$ *e* $\mu_i^{k+1} \in [0,\ \mu_{k+1}]\ i = 1, .., p$ *(Geralmente,* $\lambda_i^{k+1} = \min\{\max\{\lambda_{min},\ \lambda_i^k + \rho_k h_i(x_k)\},\ \lambda_{max}\}$ *e* $\mu_i^{k+1} = \min\{\max\{0,\ \mu_i^k + \rho_k g_i(x_k)\},\ \mu_{max}\}\ i = 1, .., p.$*)*
Fazer $k \longleftarrow k + 1$ *e voltar ao passo 1.*

A rotina ALGENCAN (Augmented Lagrangian algoritm using GENCAN) é o método de Lagrangiano Aumentado do Projeto TANGO, disponível nos endereços
`www.ime.usp.br/~egbirgin/tango` ou `www.ime.unicamp.br/~martinez/software`.

Os fundamentos teóricos deste método podem ser encontrados em J. M. Martínez (2006). No presente trabalho, apresentamos três resultados desse estudo que achamos relevantes, para garantir a convergência do método no caso em que os subproblemas são resolvidos por aproximações. De modo geral, os três resultados podem ser resumidos como:

1. Os pontos limites da seqüência gerada pelo Algoritmo 3.1 são pontos admissíveis do problema original, ou são pontos KKT da soma dos quadrados das infactibilidades, ou não sastisfazem a condição CPLD (Dependência Linear Positiva Constante) com respeito às restrições de Ω.

2. Se x^* é um ponto limite admissível que satisfaz a CPLD com respeito ao problema original, então ele é um ponto estacionário ou KKT do problema original.

3. Sob certas hipóteses, a seqüência de parâmetros de penalidade é limitada.

Antes de enunciar os teoremas vamos definir quando um ponto satisfaz a condição CPLD:

Definição 3.1 *Sejam* $h(x) = 0,\ g(x) \leq 0,\ I_A = \{i \in \{1, .., p\} / \ g_i(x) = 0\}$*, dizemos que* x *satisfaz a Condição de Dependência Linear Positiva Constante (CPLD) se a existência de* $I_h \subset \{1, .., m\},\ I_g \subset \{1, .., p\},\ \lambda_i \in \mathbb{R}\ \forall\ \mathrm{i} \in \mathrm{I_h},\ \mu_\mathrm{i} > 0\ \ \forall\ \mathrm{i} \in \mathrm{I_g}$ *tais que*

$$\sum_{i \in I_h} \lambda_i \nabla h_i(x) + \sum_{i \in I_g} \mu_i \nabla g_i(x) = 0$$

com $\sum_{i \in I_h} |\lambda_i| + \sum_{i \in I_g} \mu_i > 0$*, implica que existe* $\delta > 0$ *tal que os gradientes*

$$\{\nabla h_i(z)\}_{i \in I_h},\ \{\nabla g_i(z)\}_{i \in I_g}$$

são linearmente dependentes para todo $z \in \mathbb{R}^\mathrm{n}$ *tal que* $\|z - x\| \leq \delta$.

A prova do teorema seguinte pode ser encontrada em R. Andreani et al.(2008), Teorema 4.1, ou em J. M. Martínez (2006), Teorema 7.1.

Teorema 3.1 *Seja* $\{x_k\}$ *uma seqüência infinita gerada pelo Algoritmo 3.1. Seja* x^* *um ponto limite de* $\{x_k\}$*. Então, se a seqüência de parâmetros de penalidade* $\{\rho_k\}$ *é limitada,* x^* *é admissível. Caso contrário, verifica-se pelo menos uma das seguintes possibilidades:*

1. *O ponto x^* é um ponto KKT do problema*

$$\min_{x\in\Omega}\left[\sum_{i=1}^{m} h_i(x)^2 + \sum_{i=1}^{p}\max\{0,\ g_i(x)\}^2\right]$$

2. *x^* não satisfaz a condição CPLD associada a Ω.*

Em J. M. Martínez (2006), Teorema 7.2, pode ser achada a prova do próximo Teorema.

Teorema 3.2 *Seja $\{x_k\}$ uma seqüência infinita gerada pelo Algoritmo 3.1, com ponto limite x^*. Suponhamos que x^* é um ponto admissível que satisfaz a CPLD com respeito a todas as restrições do problema original. Então, x^* é ponto KKT do problema original.*

Para enunciar o último teorema vamos assumir as seguintes hipóteses:

Hipóteses

1. A seqüência $\{x_k\}$ gerada pelo Algoritmo 3.1 converge a x^*.
2. O ponto x^* é admissível (ou seja satisfaz $h(x^*) = 0,\ \text{e}\ \ g(x^*) \leq 0$).
3. O conjunto formado pelos gradientes das restrições de igualdade em x^*, junto com os gradientes das restrições de desigualdade ativas em tal ponto ($g(x^*) = 0$), é linearmente independente.
4. $\lambda_i^* \in (\lambda_{min},\ \lambda_{max})\ i = 1, .., m\ \ \text{e}\ \mu_i^* \in [0,\ \mu_{max})\ i = 1, .., p.$
5. As funções $f,\ h\ \ \text{e}\ \ g$ possuem derivadas segundas contínuas em uma vizinhança de x^*.
6. Para todo $\ i = 1, \cdots, p\ $ tais que $g_i(x^*) = 0,\ $ temos $\ \mu_i^* > 0.$
7. Seja $Z \in \mathbb{R}^{\mathrm{n}\times(\mathrm{n}-\mathrm{m})}$ uma matriz cujas colunas formam uma base do núcleo de $\nabla h(x^*)^T$. Então, $Z^T\nabla^2\mathcal{L}_0(x^*, \lambda^*)Z$ é definida positiva.

A demonstração do teorema enunciado a seguir pode ser achada em J. M. Martínez (2006), Teorema 8.2.

Teorema 3.3 *Suponhamos que verificam-se as hipóteses enunciadas anteriormente e que existe uma seqüência $\{\eta_k\}$ tal que*

$$\lim_{k\to\infty}\eta_k = 0$$

e

$$\epsilon_k \leq \eta_k \max\{\|h(x_k)\|_\infty,\ \|V^k\|_\infty\}\ \ \forall\ k \in \mathbb{N}.$$

Então a seqüência dos parâmetros de penalidade $\{\rho_k\}$ é limitada.

O leitor interessado em fazer um estudo mais profundo do Lagrangiano Aumentado, pode consultar, por exemplo, o trabalho de Martínez (2006).

Capítulo 4

Aceleração do Lagrangiano Aumentado

4.1 Introdução

Consideremos o seguinte problema de Programação Não Linear (PNL)

$$\begin{array}{lll} \min & f(x) & \\ \text{suj. a} & h(x) & = \ 0, \\ & g(x) & \leq \ 0. \end{array}$$

com $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ e $g : \mathbb{R}^n \longrightarrow \mathbb{R}^p$, funções de clase $C^1(\mathbb{R}^n)$.

Como vimos no capítulo anterior, este problema pode ser resolvido mediante métodos de tipo Lagrangiano Aumentado. Lembrando o que apresentamos, estes métodos utilizam parâmetros de penalidade que podem ser aumentados ou permanecer constantes de uma iteração para outra, dependendo da medida da factibilidade. Como acontece com as estratégias que utilizam parâmetros de penalidade, os métodos tipo Lagrangiano Aumentado tornam-se mal condicionados quando os parâmetros de penalidade aumentam seu valor e, os problemas associados tornam-se difíceis de resolver. Por outro lado, um crescimento lento dos parâmetros produz uma convergência lenta do algoritmo.

Neste trabalho foi utilizado o Lagrangiano Aumentado baseado na fórmula PHR. Este método é tal que consegue chegar perto da solução muito rápido, mas, perto da solução o método torna-se lento, ou seja, não consegue atingir uma precisão muito exigente, como 10^{-14}. Este fato motivou-nos a procurar por métodos de convergência local rápida perto da solução, escolhendo como ponto de partida destes métodos o ponto obtido do Lagrangiano Aumentado PHR.

Trabalhos com este objetivo já foram desenvolvidos. Em L. F. Mendonça et al. (2006), procurou-se acelerar o método do Lagrangiano Aumentado com uma estratégia *quasi Newton*. Em outro estudo recente (ver E. G. Birgin, J. M. Matínez (2008)), no qual nos baseamos

para este trabalho, foram implementados dois processos, um onde o Lagrangiano Aumentado é combinado com o método de pontos interiores e o outro onde ele é combinado com o método de Newton.

Nossa proposta de pequisa, é combinar o algoritmo ALGENCAN com o método de Newton aplicado às condições de Karush-Kuhn-Tucker (condições KKT), as quais determinam um sistema não linear. Uma maneira de encontrar a solução deste tipo de sistema é utilizar o método de Newton, pois sob certas hipóteses, perto da solução, ele tem convergência quadrática. Para a utilização do método de Newton nos baseamos em um dos métodos desenvolvidos no Capítulo 2.

Antes de adentrarnos na nossa proposta de trabalho, vamos realizar um breve resumo do método de Newton.

Método de Newton

Neste método procuramos determinar a solução da equação não linear

$$F(x) = 0,$$

com $F : \mathbb{R}^n \to \mathbb{R}^n$ e $F \in C^1(\mathbb{R}^n)$. A idéia é a seguinte: dado $z \in \mathbb{R}^n$, fazemos uma aproximação linear de $F(x)$ ao redor do ponto z

$$F(x) \approx F(z) + J_F(z)(x - z).$$

Igualando a zero o lado direito da expressão anterior e isolando x temos:

$$x = z - J_F(z)^{-1}F(z).$$

Fazendo

$$d = x - z,$$

devemos resolver o seguinte sistema não linear:

$$J_F(z)d = -F(z)$$

para, em seguida, atualizar z, ou seja, $z \leftarrow z + d$.

Agora, já estamos em condições de desenvolver nossa proposta.

4.2 Implementações propostas

Nesta seção apresentamos as três implementações que foram realizadas. Primeiramente as listamos e logo em seguida expomos cada uma com mais detalhes.

- ALGENCAN - Newton 1.
- ALGENCAN - Newton 2.
- ALGENCAN - ALGENCAN aliviado.

Cada uma delas é dividida em duas etapas, como veremos a seguir.

4.2.1 ALGENCAN - Newton 1

Na primeira etapa desta proposta, o algoritmo ALGENCAN é executado até que ele satisfaça os critérios de parada próprios do processo. Seja x_k o ponto obtido nesta etapa com λ_k e μ_k os vetores correspondentes aos multiplicadores de Lagrange associados. Estes dados serão utilizados como valores inicias para a segunda etapa. Como x_k é o valor obtido por ALGENCAN, então ele satisfaz

- $\|h(x_k)\|_\infty \leq \epsilon$
- $\|g(x_k)_+\|_\infty \leq \epsilon$, onde $g(x_k)_+ = max\{0,\ g(x_k)\}$ e $\epsilon > 0$.
- $\mu_i = 0$ quando $g_i(x_k) \leq -\epsilon$.

Para a segunda etapa define-se $I_A = \{i \mid g_i(x_k) \geq -\epsilon\}$ o conjunto dos índices das restrições candidatas a serem restrições ativas. Em seguida, o seguinte sistema não linear é resolvido:

$$\begin{aligned} \nabla f(x) + \textstyle\sum_{i=1}^{m} \lambda_i \nabla h_i(x) + \sum_{i \in I_A} \mu_i \nabla g_i(x) &= 0 \quad , \\ h(x) &= 0 \quad , \\ g_i(x) + \tfrac{z_i}{2}^2 &= 0 \quad \forall\, i \in I_A, \\ \mu_i z_i &= 0 \quad \forall\, i \in I_A, \end{aligned} \tag{4.1}$$

que representa um sistema KKT.

Detalhes da implementação

Nesta implementação obtemos uma aproximação inicial da solução do problema usando o código ALGENCAN. Em seguida, testamos quais restrições pertencem ao conjunto I_A. Para esta proposta, nos baseamos no seguinte fato: o ponto que obtemos de ALGENCAN está perto da solução, então aquelas restrições que estão longe de ser ativas, pela continuidade, vão continuar sendo inativas no ponto solução do problema. Assim podemos definir um novo problema, onde a função objetivo continua sendo a mesma, mas o conjunto das restrições está formado somente pelas restrições de igualdade e aquelas restrições de desigualdade que pertençam ao conjunto I_A. Uma tentativa natural que surge é resolver o seguinte problema:

$$\begin{aligned} \min \quad & f(x) \\ \text{suj. a} \quad & h(x) = 0, \\ & g_i(x) = 0 \quad \forall\, i \in I_A. \end{aligned}$$

A dificuldade nesta formulação está no fato que não temos certeza se todas as restrições do conjunto I_A são ativas na solução de problema original. Portanto, uma solução do problema anterior pode não ser a solução que estamos procurando. Assim, o mais conveniente na formulação para o novo problema é considerar as restrições do conjunto I_A como desigualdades. Logo, o problema a ser resolvido pode ser escrito como:

$$\begin{aligned} \min \quad & f(x) && (4.2)\\ \text{suj. a} \quad & h(x) = 0, && (4.3)\\ & g_i(x) \leq 0 \ \ \forall\, i \in I_A. && (4.4) \end{aligned}$$

Para facilitar e compreensão, doravante utilizaremos as seguintes notações:

- $r = dim\{I_A\}$
- $\tilde{g}(x) : \mathbb{R}^n \longrightarrow \mathbb{R}^k$ a função vetorial cujas coordenadas correspondem às restrições de desigualdade que pertencem a I_A.
- Fazendo um abuso de notação, denotamos por μ o vetor em $\mathbb{R}^k$ cujas componentes correspondem aos multiplicadores associados a $\tilde{g}_i(x)$ respectivamente.

Assim, o problema (4.2)-(4.4) pode ser reescrito como

$$\begin{aligned} \min \quad & f(x) && (4.5)\\ \text{suj. a} \quad & h(x) = 0, && (4.6)\\ & \tilde{g}_i(x) \leq 0 \ \ \forall\, i = 1, ..., r. && (4.7) \end{aligned}$$

Agora, que relação existe entre este problema e o sistema não linear (4.1)? Como veremos a seguir, um ponto (x^*, λ^*, μ^*) com $\mu_i \geq 0$ que é solução do sistema KKT (4.1) pode ser interpretado como um ponto estacionário do problema (4.5)-(4.7).

Quando temos que resolver problemas de PNL com restrições de igualdade e desigualdade, é bastante tentador querer transformar o problema original em outro equivalente, onde somente apareçam restrições de igualdade. Na procura de lograr este objetivo, vamos fazer uma breve discussão que fundamente nossa reformulação do problema original. Uma primeira proposta consiste em somar constantes positivas a cada uma das restrições de desigualdade, de modo tal de transformar as desigualdades em igualdades, isto é, propor-se resolver o seguinte problema

$$\begin{aligned} \min \quad & f(x) && && \\ \text{suj. a} \quad & h(x) && = \ 0, && \\ & \tilde{g}_i(x) + z_i && = \ 0 && \forall\, i = 1, ..., r,\\ & z_i && \geq \ 0 && \forall\, i = 1, ..., r. \end{aligned}$$

Mas, continuamos tendo restrições de desigualdade. Para contornar essa dificuldade, consideremos o seguinte problema:

$$\begin{aligned} \min \quad & f(x) && (4.8)\\ \text{suj. a} \quad & h(x) = 0, && (4.9)\\ & \tilde{g}_i(x) + \frac{{z_i}^2}{2} = 0 \ \ \forall\, i = 1, ..., r. && (4.10) \end{aligned}$$

com $x \in \mathbb{R}^n$ e $z \in \mathbb{R}^k$. A pergunta natural que surge é: se obtivermos pontos estacionários neste problema, podemos s achar os pontos estacionários do problema original? A resposta é afirmativa, sob certas hipóteses, como veremos a seguir.

Proposição 4.1 *Se x^* é um ponto estacionário do problema definido pelas equações* (4.5) − (4.7) *então $\exists\, z^*$ tal que (x^*, z^*) é um ponto estacionário do problema definido pelas equações* (4.8) − (4.10)*.*

Demonstração: Suponhamos que x^* é ponto estacionário do problema definido por (4.5)-(4.7). Então x^* satisfaz as condições KKT desse problema, ou seja

$$\begin{aligned} \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{r} \mu_i \nabla \tilde{g}_i(x^*) &= 0, && (4.11)\\ h(x^*) &= 0, && (4.12)\\ \tilde{g}_i(x^*) &\leq 0 \ \ \forall\, i = 1, ..., r, && (4.13)\\ \tilde{g}_i(x^*)\mu_i &= 0 \ \ \forall\, i = 1, ..., r, && (4.14)\\ \mu_i &\geq 0 \ \ \forall\, i = 1, ..., r, && (4.15) \end{aligned}$$

com $\lambda \in \mathbb{R}^m$ e $\mu \in \mathbb{R}^r$ vetores dos multiplicadores de Lagrange associados. Das condições de otimalidade de primeira ordem do problema definido por (4.8)-(4.10) temos

$$\begin{aligned} \nabla f(x) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla h_i(x) + \sum_{i=1}^{r} \tilde{\mu}_i \nabla \tilde{g}_i(x) &= 0, && (4.16)\\ h(x) &= 0, && (4.17)\\ \tilde{g}_i(x) + \frac{{z_i}^2}{2} &= 0 \ \ \forall\, i = 1, ..., r, && (4.18)\\ \tilde{\mu}_i z_i &= 0 \ \ \forall\, i = 1, ..., r. && (4.19) \end{aligned}$$

Agora, escolhendo $x = x^*$, $\lambda = \tilde{\lambda}$ e $\mu = \tilde{\mu}$, temos que as equações (4.16) e (4.17) são satisfeitas. Para completar a prova falta achar $z \in R^r$ tal que as equações (4.18) e (4.19)

sejam satisfeitas. Seja z^* tal que $z_i^* = \sqrt{-2\, g_i(x^*)}$, z^* está bem definido já que de (4.13) temos $g_i(x^*) \leq 0 \ \ \forall\, i = 1, ..., r$. Com esta escolha a equação (4.18) é satisfeita. De (4.14) e (4.15) temos que

$$\left\{ \begin{array}{lllllllllll} se & \mu_i > 0 & \Longrightarrow & \tilde{g}_i(x^*) & = & 0 & \Longrightarrow & z_i^* & = & 0 & \Longrightarrow \ \mu_i z_i^* = 0, \\ se & \tilde{g}_i(x^*) < 0 & \Longrightarrow & \mu_i & = & 0 & \Longrightarrow & \mu_i z_i^* & = & 0. & \end{array} \right.$$

Logo, temos que a equação (4.19) é cumprida. Assim fica provada esta proposição. ■

A recíproca desta proposição não é verdade, pois não podemos garantir que os multiplicadores associados às restrições de desigualdade ativas sejam não negativos. Mas a recíproca é cumprida no caso em que os multiplicadores associados as restrições onde aparecem variáveis de folga sejam não negativos. Isto é o que nos diz a seguinte proposição:

Proposição 4.2 *Se (x^*, z^*) é um ponto estacionário do problema definido pelas equações $(4.8) - (4.10)$, com multiplicadores de Lagrange associados às restrições onde aparece alguma componente da variável z não negativa, então x^* é um ponto estacionário do problema definido pelas equações $(4.5) - (4.7)$.*

Demonstração: Suponhamos que (x^*, z^*) é ponto estacionário do problema definido por (4.8)-(4.10). Então x^* satisfaz as condições KKT do problema, ou seja:

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{i=1}^{r} \mu_i \nabla \tilde{g}_i(x^*) = 0, \tag{4.20}$$

$$h(x^*) = 0, \tag{4.21}$$

$$\tilde{g}_i(x^*) + \frac{{z_i^*}^2}{2} = 0 \ \ \forall\, i = 1, ..., r, \tag{4.22}$$

$$\mu_i z_i^* = 0 \ \ \forall\, i = 1, ..., r, \tag{4.23}$$

com λ e μ vetores dos multiplicadores de Lagrange associados tais que $\mu_i \geq 0 \ \ \forall\, i = 1, ..., r$. Devemos provar que existem $\tilde{\lambda}$ e $\tilde{\mu}$ tais que x^* satisfaz:

$$\nabla f(x^*) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla h_i(x^*) + \sum_{i=1}^{r} \tilde{\mu}_i \nabla \tilde{g}_i(x^*) = 0, \tag{4.24}$$

$$h(x^*) = 0, \tag{4.25}$$

$$\tilde{g}_i(x^*) \leq 0 \ \ \forall\, i = 1, ..., r, \tag{4.26}$$

$$g_i(x^*)\tilde{\mu}_i = 0 \ \ \forall\, i = 1, ..., r, \tag{4.27}$$

$$\tilde{\mu}_i \geq 0 \ \ \forall\, i = 1, ..., r. \tag{4.28}$$

Se escolhermos $\tilde{\lambda} = \lambda$ e $\tilde{\mu} = \mu$, as equações (4.24) e (4.25) são cumpridas. De (4.22) temos que $\tilde{g}_i(x^*) \leq 0$ e por hipótese (4.28) é satisfeita. Agora, de (4.22) e (4.23) temos

$$\begin{cases} se \quad \mu_i > 0 \implies z_i^* = 0 \implies \tilde{g}_i(x^*) = 0 \implies \mu_i \tilde{g}_i(x^*) = 0, \\ se \quad z_i = 0 \implies \tilde{g}_i(x^*) = 0 \implies \mu_i \tilde{g}_i(x^*) = 0. \end{cases}$$

Portanto (4.27) também é cumprida. Com isto, proposição fica provada. ■

Assim, com base nesta última proposição, para encontrar pontos estacionários de problemas de PNL, vamos nos concentrar em achar os pontos estacionários dos problemas definidos como em (4.8)-(4.10), onde as variáveis z_i são chamadas *variáveis de folga*. Como podemos notar, as condições de otimalidade deste último problema correspondem às equações do sistema KKT que devemos resolver na nossa proposta.

Consideremos o seguinte problema de PNL

$$\begin{array}{lll} \min & f(x) & \\ \text{suj. a} & h(x) = & 0, \\ & g(x) \leq & 0. \end{array}$$

Definimos a função *Lagrangiano* deste problema como:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + h(x)^T \lambda + g(x)^T \mu.$$

Um ponto KKT ou ponto estacionário de um problema de PNL, também pode ser interpretado como um ponto estacionário da função Lagrangiana definida para o problema de PNL. Provemos isto:

A função Lagrangiano em problemas da forma (4.8)-(4.10), é

$$\mathcal{L}(x, \lambda, \mu) = f(x) + h(x)^T \lambda + \left(\tilde{g}(x) + \frac{z^2}{2} \right)^T \mu,$$

onde $\frac{z}{2}^2$ representa um vetor no qual cada componente é multiplicada por ela mesma vezes $\frac{1}{2}$, e z representa o vetor das variáveis de folga. As condições de otimalidade de primeira ordem desta função são:

$$\begin{array}{lll} \nabla_x \mathcal{L} = & \nabla f(x) + J_h(x)^T \lambda + J_{\tilde{g}}(x)^T \mu = & 0, \\ \nabla_\lambda \mathcal{L} = & h(x) = & 0, \\ \nabla_\mu \mathcal{L} = & \tilde{g}(x) + \frac{z}{2}^2 = & 0, \\ \nabla_z \mathcal{L} = & ZUe = & 0, \end{array}$$

onde

$$Z = diag(z_1, ..., z_r),$$

$$U = diag(\mu_1, ..., \mu_r),$$

$$e = (1, ..., 1)^T, \quad e \in \mathbb{R}^r.$$

Estas equações representam as condições de otimalidade de primeira ordem do problema definido por (4.8)-(4.10). Assim fica provada nossa interpretação. ■

Em nossa proposta, para achar a solução do sistema KKT devemos resolver um sistema não linear, que como já foi dito, é feito usando o método de Newton. Para isto, devemos determinar duas coisas: primeiro, qual é a função que representa a equação não linear e segundo, a matriz Jacobiana desta função. A função não linear é representada pelas equações do sistema não linear, ou seja

$$F(x, \lambda, \mu, z) = \begin{pmatrix} \nabla_x \mathcal{L} \\ \nabla_\lambda \mathcal{L} \\ \nabla_\mu \mathcal{L} \\ \nabla_z \mathcal{L} \end{pmatrix} = \begin{pmatrix} \nabla f(x) + J_h(x)^T \lambda + J_{\tilde{g}}(x)^T \mu \\ h(x) \\ \tilde{g}(x) + \frac{z}{2}^2 \\ ZUe \end{pmatrix},$$

e sua Jacobiana é

$$J_F(x, \lambda, \mu, z) = \begin{pmatrix} M & J_h(x)^T & J_{\tilde{g}}(x)^T & 0 \\ J_h(x) & 0 & 0 & 0 \\ J_{\tilde{g}}(x) & 0 & 0 & Z \\ 0 & 0 & Z & U \end{pmatrix},$$

onde

$$M = \nabla^2 f(x) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x) + \sum_{i=1}^{r} \mu_i \nabla^2 \tilde{g}(x).$$

Observemos que a matriz Jacobiana de $F(x, \lambda, \mu, z)$ pode ser escrita como uma matriz em blocos 2×2 da forma

$$\begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}$$

escolhendo:

$$A = \begin{pmatrix} M & J_h(x)^T \\ J_h(x) & 0 \end{pmatrix}, \quad B = \begin{pmatrix} J_{\tilde{g}(x)} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{e} \quad C = -\begin{pmatrix} 0 & Z \\ Z & U \end{pmatrix}.$$

Como as matrizes A e C são simétricas, a matriz J_F é uma matriz ponto de sela. Portanto, da Definição 2.1, o sistema que ela define é um sistema ponto de sela. Assim, a solução desse sistema linear pode ser encontrada usando algum dos métodos desenvolvidos no capítulo 2.

Finalmente, podemos resumir esta proposta no seguinte algoritmo:

Algoritmo 4.1

Etapa 1 *Executar ALGENCAN. Sejam* $\tilde{x}$, $\tilde{\lambda}$ *e* $\tilde{\mu}$ *a aproximação obtida nesta etapa para a solução do problema e os multiplicadores de Lagrange respectivamente. Usar esta solução como ponto inicial da próxima etapa.*

Etapa 2 *Definir* $I_A = \{i \mid g_i(\tilde{x}) \geq -\epsilon\}$, *acrescentar as variáveis de folga às restrições de* I_A *e definir* $z_i = \sqrt{2\,max\{0, -g_i(\tilde{x})\}} \quad \forall i \in I_A$ *como vetor inicial dessas variáveis. Executar o método de Newton para resolver o sistema*

$$J_F(x, \lambda, \mu, z)d = -F((x, \lambda, \mu, z),$$

atualizar os parâmetros. Declarar convergência de Newton quando

$$\|\nabla_x \mathcal{L}(x, \lambda, \mu, z)\|_\infty \leq \ tol,$$

onde tol é uma tolerância positiva. Se a tolerância não é atingida, parar o algoritmo quando o número máximo de iterações for atingido.

Observe que no critério de parada, não é considerada a medida da infactibilidade, pois ALGENCAN consegui atingir ela.

4.2.2 ALGENCAN - Newton 2

Esta proposta é uma variante da proposta anterior. Como fizemos em ALGENCAN-Newton 1, executamos ALGENCAN obtendo o ponto inicial para a próxima etapa. Na segunda etapa, são determinados dois conjuntos de índices: o conjunto das restrições candidatas a serem restrições ativas $I_A = \{i \mid g_i(x) \geq -\epsilon\}$ e o conjunto $I_{sf} = \{i \in I_A \mid \mu_i \geq \beta\}$, restrições "sem folga", onde β é uma constante positiva e μ_i é o multiplicador de Lagrange associado à restrição $g_i(x)$. Este conjunto representa as restrições de I_A cujo multiplicador associado tem valor "grande". Logo, como na proposta anterior, estabelecemos o sistema KKT para ser resolvido. A mudança que é feita em relação ao sistema (4.1) aparece nas últimas duas equações que são substituídas pelas seguintes:

$$\begin{aligned} g_i(x) &= 0 \quad \forall i \ \in I_{sf}, \\ g_i(x) + \tfrac{z_i}{2}^2 &= 0 \quad \forall \ i \in I_A - I_{sf}, \\ \mu_i z_i &= 0 \quad \forall \ i \in I_A - I_{sf}. \end{aligned}$$

Assim o sistema linear a ser resolvido nesta etapa fica:

$$\begin{aligned} \nabla f(x) + \textstyle\sum_{i=1}^{m} \lambda_i \nabla h_i(x) + \sum_{i \in I_A} \mu_i \nabla g_i(x) &= 0 \quad , \\ h(x) &= 0 \quad , \\ g_i(x) &= 0 \quad \forall i \ \in I_{sf}, \\ g_i(x) + \tfrac{z_i}{2}^2 &= 0 \quad \forall \ i \in I_A - I_{sf}, \\ \mu_i z_i &= 0 \quad \forall \ i \in I_A - I_{sf}. \end{aligned} \tag{4.29}$$

O sistema não linear (4.29) pode ser interpretado como as condições de otimalidade de primeira ordem do seguinte problema

$$\begin{array}{lrcll} \min & f(x) & & & \\ \text{suj. a} & h(x) & = & 0, & \\ & g_i(x) & = & 0 & \forall\, i \in I_{sf}, \\ & g_i(x) + \frac{z_i}{2}^2 & = & 0 & \forall\, i \in I_A - I_{sf}, \end{array} \tag{4.30}$$

que sob a hipótese que os multiplicadores de Lagrange, correspondentes às restrições que pertencem a I_A tenham valores positivos, é equivalente a resolver o problema de programação não linear

$$\begin{array}{lrcll} \min & f(x) & & & \\ \text{suj. a} & h(x) & = & 0, & \\ & g_i(x) & = & 0 & \forall\, i \in I_{sf}, \\ & g_i(x) & \leq & 0 & \forall\, i \in I_A - I_{sf}. \end{array} \tag{4.31}$$

A prova desta equivalência é similar à feita nas proposições 4.1 e 4.2 da subseção 4.2.1.

Detalhes da implementação

Esta abordagem pode ser interpretada como uma segunda versão da abordagem anterior, pois a primeira etapa não muda, mas na segunda etapa realizamos uma modificação.
Para esta abordagem nos baseamos na suposição que como o ponto obtido por ALGENCAN está perto da solução, consideramos que as estimativas dos multiplicadores poderiam nos dar informação acerca de quais restrições do conjunto I_A são verdadeiramente ativas, ou seja, para quais cumpre-se $g_i(x) = 0$.
Agora como usarmos os multiplicadores para determinar as restrições que consideramos "verdadeiramente ativas"?, para isto consideremos somente as seguintes condições de otimalidade KKT:

$$\begin{array}{rcll} \mu_i g_i(x) & = & 0, & \\ g_i(x) & \leq & 0 & \forall\, i = 1, ..., p, \\ \mu_i & \geq & 0 & \forall\, i = 1, ..., p. \end{array} \tag{4.32}$$

A análise que pode ser feita destas relações é que quanto mais aumenta a variável μ_i o valor de $g_i(x)$ deve estar mais perto de zero para que a primeira condição continue se cumprindo. Assim, consideramos aquelas restrições de desigualdade cujo multiplicador associado tem valor grande (maior que um certo parâmetro positivo), como restrições ativas. Agora, o ponto para o qual vamos determinar o conjunto I_{sf}, é o ponto obtido na primeira etapa, sendo, portanto, uma boa aproximação da solução do problema. Então por continuidade, supomos que aquelas restrições continuarão pertencendo ao conjunto I_{sf} na solução do problema.

Novamente para achar a solução do sistema não linear (4.29) vamos a aplicar o método de Newton. A função da equação não linear

$$F(y) = 0$$

é a seguinte

$$F(x,\lambda,\mu,z) = \begin{pmatrix} \nabla f(x) + \sum_i \tilde{\lambda}_i \nabla \tilde{h}_i(x) + \sum_i \mu_i \nabla g_{cf}(x) \\ \tilde{h}(x) \\ g(x)_{cf} + \frac{z_{cf}}{2}^2 \\ \mu_{cf} z_{cf}. \end{pmatrix},$$

onde

$$\tilde{h}(x) = \begin{pmatrix} h(x) \\ g_{sf}(x) \end{pmatrix},$$

cujas compontentes estão formadas pelas restrições de igualdade $h(x)$ e as restrições de desigualdade de I_A que não precisam variáveis de folga, representadas por $g_{sf}(x)$,

$$\tilde{\lambda} = \begin{pmatrix} \lambda \\ \mu_{sf} \end{pmatrix},$$

no qual a componente λ corresponde a um vetor contendo os multiplicadores de Lagrange associados às restrições de igualdade e a componente μ_{sf} é o vetor formado pelos multiplicadores de Lagrange associados às restrições de desigualdade que não precisam de variáveis de folga,

$g_{cf}(x)$ denota o vetor das restrições de desigualdades de I_A que precisam de variáveis de folga,

μ_{cf} corresponde ao vetor dos multiplicadores de Lagrange associados as componentes do vetor $g_{cf}(x)$,

z_{cf} é o vetor das variáveis de folga, onde a notação $\frac{z_{cf}}{2}^2$ representa cada componente deste vetor ao quadrado, vezes o fator $\frac{1}{2}$.
Logo a matriz Jacobiana de F é

$$\begin{pmatrix} M & J_{\tilde{h}}^T(x) & J_{g_{cf}}^T(x) & 0 \\ J_{\tilde{h}}(x) & 0 & 0 & 0 \\ J_{g_{cf}}(x) & 0 & 0 & Z_{cf} \\ 0 & 0 & Z_{cf} & U_{cf} \end{pmatrix}, \tag{4.33}$$

onde, supondo $k = p - dim I_{sf}$, temos:

- $M = \nabla^2 f(x) + \sum_{i=1}^{m+dim I_{sf}} \tilde{\lambda}_i \nabla^2 \tilde{h}_i(x) + \sum_{i=1}^{k} \mu_i \nabla^2 g_{cf}(x)$,
- $Z_{cf} = diag\{z_{cf}^1 \dots z_{cf}^k\}$,
- $U_{cf} = diag\{\mu_{cf}^1 \dots \mu_{cf}^k\}$.

Novamente temos uma matriz ponto de sela, portanto, o sistema que ela define é um sistema ponto de sela. A sua resolução pode ser determinada por alguns dos métodos desenvolvidos no Capítulo 2.

Assim, o algoritmo para esta proposta pode ser escrito como:

Algoritmo 4.2 Etapa 1 *Executar ALGENCAN. Sejam* $\tilde{x}$, $\tilde{\lambda}$ *e* $\tilde{\mu}$ *a aproximação obtida nesta etapa para a solução do problema e os multiplicadores de Lagrange, respectivamente. Usar esta solução como ponto inicial da próxima etapa.*

Etapa 2 *Definir* $I_A = \{i \mid g_i(\tilde{x}) \geq -\epsilon\}$ *e* $I_{sf} = \{i \in I_A \mid \mu_i \geq \beta\}$. *Acrescentar variáveis de folga às restrições candidatas a serem ativas* $g_i(\tilde{x})$ *tais que* $i \in I_A - I_{sf}$ *e, definir* $z_i = \sqrt{2\,max\{0, -g_i(\tilde{x})\}}$ $\forall i \in I_A - I_{sf}$ *como vetor inicial dessas variáveis. Executar Newton para resolver o sistema*

$$J_F(x, \lambda, \mu, z)d = -F((x, \lambda, \mu, z),$$

atualizar os parâmetros. Declarar convergência de Newton quando

$$\|\nabla_x \mathcal{L}(x, \lambda, \mu, z)\|_\infty \leq \ tol, \quad (tol > 0).$$

Se a tolerância não é atingida, parar o algoritmo quando o número máximo de iterações for atingido.

A função $\mathcal{L}((x, \lambda, \mu, z)$ *representa a função Lagrangiano definida para o problema (4.30).*

Novamente observe que no critério de parada, não é considerada a medida da infactibilidade, pois ALGENCAN consegui atingir ela.

4.2.3 ALGENCAN - ALGENCAN aliviado

Na primeira etapa desta implementação executamos ALGENCAN no problema original até que o processo termine por cumprir algum dos critérios de parada do algoritmo. Na segunda etapa redefinimos o problema: a função a ser minimizada não muda com respeito ao problema original mas sim mudam as restrições, ficando restrito às restrições de igualdade do problema original e àquelas restrições de desigualdade tais que no ponto obtido por ALGENCAN satisfaçam $g_i(x) \geq -\epsilon$, (restrições candidatas a serem *restrições ativas*). Logo executa-se ALGENCAN neste novo problema, escolhendo como ponto inicial o ponto obtido na etapa anterior.

Detalhes da implementação

Para sermos justos com ALGENCAN, assim como fizemos nas propostas anteriores de reduzir o conjunto de restrições, nosso objetivo aqui foi verificar o que acontecia com o código ALGENCAN, se apenas considerássemos as restrições ativas perto da solução do problema.

Assim o algoritmo sugerido neste caso é o seguinte:

Algoritmo 4.3

Etapa 1 *Executar ALGENCAN. Sejam* $\tilde{x}$, $\tilde{\lambda}$ *e* $\tilde{\mu}$ *a aproximação obtida nesta etapa para a solução do problema e os multiplicadores de Lagrange respectivamente, que serão utilizados como ponto inicial para a segunda etapa.*

Etapa 2 *Definir* $I_A = \{i \mid g_i(\tilde{x}) \geq -\epsilon\}$ *e resolver, usando ALGENCAN, o seguinte problema*

$$\begin{aligned} \min \quad & f(x) \\ suj.\ a\ \ & h(x) = 0, \\ & g_i(x) \leq 0 \ \ \forall\, i \in I_A. \end{aligned}$$

4.3 Subrotina utilizada para a resolução dos sistemas Ponto de Sela

Nas três abordagens consideradas, as resoluções dos sistemas ponto de sela foram realizadas com o auxilio da subrotina MA57, que é parte integrante da biblioteca da HSL (*Harwell Subroutine Library*). Esta subrotina resolve sistemas lineares da forma

$$Ax = b,$$

onde a matriz A é simétrica, esparsa e não necessariamente definida. Este é um método direto, baseado em uma variante da eliminação Gaussiana. A resolução é feita utilizando a fatoração LDL^T desenvolvida no Capítulo 2. A idéia do algoritmo é a seguinte: inicialmente os elementos da matriz A são ordenados por uma técnica do tipo mínimo grau para preservar esparsidade (ver Duff et al (1986)). Em seguida, é realizada a fatoração LDL^T mesmo que A tenha posto incompleto e, finalmente, é resolvido o sistema linear.

4.4 Experimentos Computacionais

Em nossos experimentos computacionais, consideramos problemas de pequeno e grande porte, lembrando que essa clasificação é feita de acordo com o número de variáveis do problema.

4.4.1 Problemas de pequeno porte

Os problemas apresentados a seguir, foram retirados de W. Hock, K. Schittkowski (1981) e do *Projeto Tango*.

Problema 1:

$$\begin{aligned} \min \quad & x_1 - x_2 \\ \text{suj. a} \quad & -3x_1^2 + 2x_1x_2^2 + 1 \geq 0 \end{aligned}$$

Problema 2:

$$\begin{aligned} \min \quad & (x_1 - 2)^2 + (x_2 - 1)^2 \\ \text{suj. a} \quad & -x_1^2 + x_2 \geq 0 \\ & -x_1 - x_2 + 2 \geq 0 \end{aligned}$$

Problema 3:

$$\begin{aligned} \min \quad & -x_1x_2x_3 \\ \text{suj. a} \quad & -x_1^2 - 2x_2^2 - 4x_3^2 + 48 \geq 0 \end{aligned}$$

Problema 4:

$$\begin{aligned} \min \quad & x_1^2 + x_2^2 + 2x_3^2 + x_4^2 - 5x_1 - 5x_2 - 21x_3 + 7x_4 \\ \text{suj. a} \quad & 8 - x_1^2 - x_2^2 - x_3^2 - x_4^2 - x_1 + x_2 - x_3 + x_4 \geq 0 \\ & 10 - x_1^2 - 2x_2^2 - x_3^2 - 2x_4^2 + x_1 + x_4 \geq 0 \\ & 5 - 2x_1^2 - x_2^2 - x_3^2 - 2x_1 + x_2 + x_4 \geq 0 \end{aligned}$$

Problema 5:

$$\begin{aligned} \min \quad & x_1^2 + x_2^2 \\ \text{suj. a} \quad & x_1 + x_2 + 1 \leq 0 \\ & x_1 - x_2 - 1 \leq 0 \end{aligned}$$

Problema 6:

$$\begin{aligned} \min \quad & (x_1 - 10)^2 + 5(x_2 - 12)^2 + x_3^4 + 3(x_4 - 11)^2 + 10x_5^6 + 7x_6^2 \\ & + x_7^4 - 4x_6x_7 - 10x_6 - 8x_7 \\ \text{suj. a} \quad & 127 - 2x_1^2 - 3x_2^4 - x_3 - 4x_4^2 - 5x_5 \geq 0 \\ & 282 - 7x_1 - 3x_2 - 10x_3^2 - x_4 + x_5 \geq 0 \\ & 196 - 23x_1 - x_2^2 - 6x_6^2 + 8x_7 \geq 0 \\ & -4x_1^2 - x_2^2 + 3x_1x_2 - 2x_3^2 - 5x_6 + 11x_7 \geq 0 \end{aligned}$$

Problema 7:

$$\begin{aligned}
\min \quad & x_1^2 + x_2^2 + x_1x_2 - 14x_1 - 16x_2 + (x_3 - 10)^2 + 4(x_4 - 5)^2 \\
& + (x_5 - 3)^2 + 2(x_6 - 1)^2 + 5x_7^2 + 7(x_8 - 11)^2 \\
& + 2(x_9 - 10)^2 + (x_{10} - 7)^2 + 45
\end{aligned}$$

$$\begin{aligned}
\text{suj. a} \quad & 105 - 4x_1 - 5x_2 + 3x_7 - 9x_8 && \geq 0 \\
& -10x_1 + 8x_2 + 17x_7 - 2x_8 && \geq 0 \\
& 8x_1 - 2x_2 - 5x_9 + 2x_{10} + 12 && \geq 0 \\
& -3(x_1 - 2)^2 - 4(x_2 - 3)^2 - 2x_3^2 + 7x_4 + 120 && \geq 0 \\
& -5x_1^2 - 8x_2 - (x_3 - 6)^2 + 2x_4 + 40 && \geq 0 \\
& -0.5(x_1 - 8)^2 - 2(x_2 - 4)^2 - 3x_5^2 + x_6 + 30 && \geq 0 \\
& -x_1^2 - 2(x_2 - 2)^2 + 2x_1x_2 - 14x_5 + 6x_6 && \geq 0 \\
& 3x_1 - 6x_2 - 12(x_9 - 8)^2 + 7x_{10} && \geq 0
\end{aligned}$$

Problema 8:

$$\begin{aligned}
\min \quad & 5.04x_1 + 0.035x_2 + 10x_3 + 3.36x_5 - 0.063x_4x_7 \\
\text{suj. a} \quad & g_1(x) = 735.82 - 0.222x_{10} - bx_9 && \geq 0 \\
& g_2(x) = -133 + 3x_7 + 3x_7 - ax_{10} && \geq 0 \\
& g_3(x) = -g_1(x) + x_9(1/b - b) && \geq 0 \\
& g_4(x) = -g_2(x) + (1/a - a) && \geq 0 \\
& g_5(x) = 1.12x_1 + 0.13167x_1x_8 - 0.00667x_1x_8^2 - ax_4 && \geq 0 \\
& g_6(x) = 57.425 + 1.098x_8 - 0.038x_8^2 + 0.325x_6 - ax_7 && \geq 0 \\
& g_7(x) = -g_5(x) + (1/a - a)x_4 && \geq 0 \\
& g_8(x) = -g_6(x) + (1/a - a)x_7 && \geq 0 \\
& g_9(x) = 1.22x_4 - x_1 - x_5 && = 0 \\
& g_{10}(x) = 98000x_3/(x_4x_9 + 1000x_3) - x_6 && = 0 \\
& g_{11}(x) = (x_2 + x_5)/x_1 - x_8 && = 0 \\
& a = 0.99 \\
& b = 0.9
\end{aligned}$$

$$
\begin{array}{ccccc}
0.00001 & \leq & x_1 & \leq & 2000 \\
0.00001 & \leq & x_2 & \leq & 16000 \\
0.00001 & \leq & x_3 & \leq & 120 \\
0.00001 & \leq & x_4 & \leq & 5000 \\
0.00001 & \leq & x_5 & \leq & 2000 \\
85 & \leq & x_6 & \leq & 93 \\
90 & \leq & x_7 & \leq & 95 \\
3 & \leq & x_8 & \leq & 12 \\
1.2 & \leq & x_9 & \leq & 4 \\
145 & \leq & x_{10} & \leq & 162
\end{array}
$$

Problema 9:

$$
\begin{array}{lll}
\min & x_{11} + x_{12} + x_{13} & \\
\text{suj. a} & x_3 - x_2 & \geq 0 \\
& x_2 - x_1 & \geq 0 \\
& 1 - 0.002x_7 + 0.002x_8 & \geq 0 \\
& x_{11} + x_{12} + x_{13} & \geq 50 \\
& 250 - x_{11} - x_{12} - x_{13} & \geq 0 \\
& x_{13} - 1.262626x_{10} + 1.231059x_3x_{10} & \geq 0 \\
& x_5 - 0.03475x_2 - 0.975x_2x_5 + 0.00975x_2^2 & \geq 0 \\
& x_6 - 0.03475x_3 - 0.975x_3x_6 + 0.00975x_3^2 & \geq 0 \\
& x_5x_7 - x_1x_8 - x_4x_7 + x_4x_8 & \geq 0 \\
& 1 - 0.002(x_2x_9 + x_5x_8 - x_1x_8 - x_6x_9) - x_5 - x_6 & \geq 0 \\
& x_2x_9 - x_3x_{10} - x_6x_9 - 500x_2 + 500x_6 + x_2x_{10} & \geq 0 \\
& x_2 - 0.9 - 0.002(x_2x_{10} - x_3x_{10}) & \geq 0 \\
& x_4 - 0.03475x_1 - 0.975x_1x_4 + 0.00975x_1^2 & \geq 0 \\
& x_{11} - 1.262626x_8 + 1.231059x_1x_8 & \geq 0 \\
& x_{12} - 1.262626x_9 + 1.231059x_2x_9 & \geq 0
\end{array}
$$

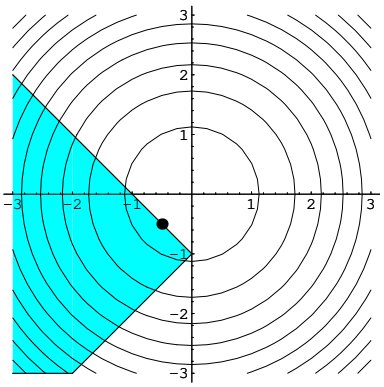
$$
\begin{array}{ccccc}
0.1 & \leq & x_1 & \leq & 1 \\
0.1 & \leq & x_2 & \leq & 1 \\
0.1 & \leq & x_3 & \leq & 1 \\
0.0001 & \leq & x_4 & \leq & 0.1 \\
0.1 & \leq & x_5 & \leq & 0.9 \\
0.1 & \leq & x_6 & \leq & 0.9 \\
0.1 & \leq & x_7 & \leq & 1000 \\
0.1 & \leq & x_8 & \leq & 1000 \\
500 & \leq & x_9 & \leq & 1000 \\
0.1 & \leq & x_{10} & \leq & 500 \\
1 & \leq & x_{11} & \leq & 150 \\
0.0001 & \leq & x_{12} & \leq & 150 \\
0.0001 & \leq & x_{13} & \leq & 150
\end{array}
$$

Problemas 10, 11 e 12:

Eles formam parte dos problemas presentes no ***projeto Tango***. O desenvolvimento deles será realizado para o caso de problemas de grande porte. Para a execução destes problemas dois parâmetros devem ser definidos: np que representa quantidade de pontos e $ndim$, que representa a dimensão à qual pertencem tais pontos. Na seguinte tabela, são apresentadas as execuções realizadas mudando np e $ndim$, junto com a quantidade de variáveis do problema e a quantidade de restrições representadas por n e m respectivamente.

	np	$ndim$	n	m
hardspheres	5	2	11	15
kissing	6	3	18	21
kissing	15	5	75	120

Vamos destacar que o problema 5 não foi retirado da bibliografía antes mencionada. Ele foi o escolhido por sua simplicidade para desenhar o conjunto das restrições e ter idéia geometricamente da solução do problema, como podemos observar no seguinte gráfico:



4.4.2 Problemas de grande porte

Os problemas que utilizamos em nossos experimentos computacionais foram retirados do *projeto Tango*. Eles são:

Problema do passo da montanha

Estes problemas consistem em resolver a seguinte situação: Dada uma superfície $S(x, y)$ e $p_i,\ p_f \in \mathbb{R}^2$ pontos inicial e final respectivamente, procura-se encontrar uma poligonal $p_i, p_1, p_2, ..., p_N, p_f,$ tal que

$$\max_{1 \leq k \leq N} S(p_k)$$

seja o menor possível, ou seja, uma poligonal φ tal que

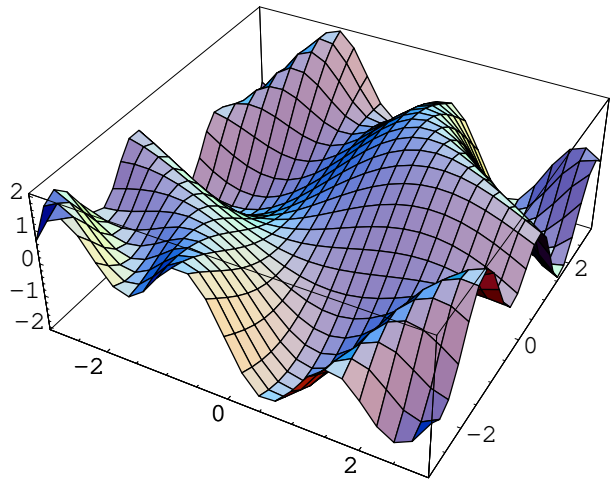
$$\varphi = \min_{\psi \in \Gamma} \max_{p_k \in \psi} S(p_k),$$

onde Γ representa a família de todos os caminhos que unem p_i com p_f. Além disso, a distância entre pontos consecutivos da poligonal deve ser menor ou igual a uma tolerância estabelecida. O problema possui $n = 2N + 1$ variáveis e $m = 2N + 1$ restrições, onde N corresponde à quantidade de pontos entre p_i e p_f.

Logo, podemos formular o problema como o seguinte problema de Programação Não Linear

$$\begin{array}{lrcll} \min & z & & & \\ \text{suj. a} & \|p_i - p_1\|^2 & \leq & d_{max}, & \\ & \|p_i - p_1\|^2 & \leq & d_{max}, & \\ & \|p_k - p_{k+1}\|^2 & \leq & d_{max} & \forall\ k = 1, ..., N-1, \\ & \|p_N - p_f\|^2 & \leq & d_{max}, & \\ & S(pk) & \leq & z & \forall\ k = 1, 2, ..., N. \end{array}$$

A superfície considerada neste trabalho é:



$$S(x, y) = sin(xy) + sin(x + y).$$

Nos experimentos numéricos, este problema é chamado a *"mountain"*.

Problema de Esferas Rígidas

Estes problemas consistem em: Dado um conjunto de np pontos $p_1, ..., p_{np}$ em uma esfera unitária pertencente a um espaço de dimensão nd, procura-se maximizar a distância mínima entre dois pontos quaisquer.

Formulando o problema como um problema de Programação Não Linear temos:

$$\begin{aligned} \max \ \min_{i \neq j} \quad & \|p_i - p_j, \| \\ \text{suj. a} \qquad & \|p_i\| = 1 \quad \forall\, i = 1, ..., np. \end{aligned}$$

Reformulando o problema, podemos reescrevê-lo

$$\begin{aligned} \min \qquad & z \\ \text{suj. a} \quad & \langle p_i, p_j \rangle \leq z \qquad \forall\, i \neq j, \\ & \|p_i\| = 1 \quad \forall\, i = 1, ..., np. \end{aligned}$$

onde z é uma variável real.
Nos experimentos numéricos, este problema é chamado *"hardspheres"*

Problemas kissing

Neste problema, devemos determinar np pontos $p_1, p_2, ..., p_{np}$ em uma esfera unitária de um espaço de dimensão nd tal que

$$\text{dist}(p_i, p_j) \geq 1, \quad \forall\, i < j.$$

Reformulando o problema como o um problema de Programação Não Linear temos

$$\begin{aligned} \min \quad & 0 \\ \text{suj. a} \quad & \text{dist}(p_i, p_j) \geq 1 \qquad \forall\, i < j, \\ & \|p_i\| = 1 \quad \forall\, i = 1, ..., np. \end{aligned}$$

Nos testes numéricos, este problema é chamado *"kissing"*.

Na tabela apresentada a seguir, são detalhados cada um dos problemas anteriores com os distintos parâmetros testados junto com o número de variáveis do problema e o número de restrições, representados por n e m respectivamente:

	np	$ndim$	n	m
mountain	50		101	101
	60		121	121
	75		151	151
	90		181	181
	100		201	201
hardspheres	15	20	301	120
	20	22	441	210
	10	22	221	55
	10	30	301	55
	15	40	601	120
	40	6	241	820
kissing	17	27	459	153
	27	26	702	378
	26	19	494	351
	20	20	400	210
	28	20	560	406
	35	10	350	630

4.4.3 Resultados Numéricos

ALGENCAN-Newton 1 e ALGENCAN Newton 2

Os resultados dos experimentos computacionais, são apresentados em diferentes tabelas. Elas são diferenciadas em problemas de pequeno porte e grande porte. Os valores obtidos em relação ao tempo, correspondem a uma média obtida de testar 100 vezes cada problema. A precisão proposta em nossos testes foi 10^{-14}, a precisão exigida para ALGENCAN foi 10^{-8}. Destacamos que algumos problemas não atingiram a precisão de 10^{-14} devido a que, o tamanho do passo para o método de Newton resultou inferior a este valor, assim o passo tornou-se desprezível.

Problemas de pequenos porte

Descrição dos problemas

		ALGENCAN		ALG.-Newton 1		ALG.-Newton 2	
		n	m	n	m	n	m
1	Problema 1	2	1	4	1	3	1
2	Problema 2	2	2	6	2	4	2
3	Problema 3	3	1	5	1	4	1
4	Problema 4	4	3	8	2	6	2
5	Problema 5	2	2	4	1	3	1
6	Problema 6	7	4	11	2	9	2
7	Problema 7	10	8	22	6	17	6
8	Problema 8	10	31	25	9	19	9
9	Problema 9	13	41	43	15	30	15
10	hardspheres	11	15	26	10	21	10
11	kissing	18	21	38	13	38	13
12	kissing	75	120	106	23	106	23

n: representa quantidade de variáveis do problema presentes em cada proposta.

m: representa quantidade de restrições do problema presentes em cada proposta.

Tabela 1

Precisão

Para medir a precisão, calculamos a medida da otimalidade e a medida da infactibilidade, e denotamos-as como $\|\nabla\mathcal{L}\|_\infty$ e $\|g\|_\infty$ respetivamente.

	ALGENCAN		ALGENCAN-Newton 1		ALGECAN-Newton 2	
	$\|\nabla\mathcal{L}\|_\infty$	$\|g\|_\infty$	$\|\nabla\mathcal{L}\|_\infty$	$\|g\|_\infty$	$\|\nabla\mathcal{L}\|_\infty$	$\|g\|_\infty$
1	6.44 E-12	3.13E-07	0	4.477E-17	0	4.477E-17
2	7.520E-09	8.483E-09	1.110E-16	2.427E-09	0	0
3	3.485E-09	3.545E-10	3.557E-15	3.545E-10	1.776E-15	2.185E-15
4	9.439E-09	5.095E-09	1.776E-15	3.333E-09	8.881E-16	1.033E-16
5	4.44E-16	9.455E-9	4.44E-16	9.455E-9	4.44E-16	9.455E-9
6	2.908E-09	2.404E-09	2.664E-15	1.823E-10	1.420E-14	0
7	8.260E-09	9.960E-09	7.105E-15	7.327E-15	7.105E-15	7.327E-15
8	7.214E-10	4.911E-09	2.387E-12	6.039E-14	2.387E-12	6.039E-14
9	1.698E-06	7.430E-09	5.684E-14	2.614E-14	5.684E-13	1.772E-12
10	6.238E-09	4.608E-10	5.551E-17	1.317E-10	3.122E-17	1.296E-16
11	1.847E-09	1.8614E-10	1.57567E-19	8.2019E-17	1.57567E-19	8.2019E-17
12	5.740E-09	1.725E-09	1.158E-16	1.578E-16	1.158E-16	1.578E-16

Tabela 2

Problemas de grande porte

Descrição dos problemas

		ALGENCAN		ALG.-Newton 1		ALG.-Newton 2	
		n	m	n	m	n	m
1	mountain	101	101	121	10	118	10
2		121	121	127	3	124	3
3		151	151	279	21	236	21
4		181	181	225	22	216	22
5		201	201	239	19	236	19
6	hardspheres	301	120	526	120	526	120
7		441	210	841	210	841	210
8		221	55	321	55	321	55
9		301	55	401	55	401	55
10		601	120	826	120	826	120
11		241	820	833	316	833	316
12	kissing	459	153	512	35	512	35
13		702	378	733	29	733	29
14		494	351	546	39	546	39
15		400	210	434	27	434	27
16		560	406	624	46	624	46
17		350	630	417	51	417	51

n: representa quantidade de variáveis do problema presentes em cada proposta.

m: representa quantidade de restrições do problema presentes em cada proposta.

Tabela 3

Precisão e Infactibilidade

	ALGENCAN		ALGENCAN-Newton 1		ALGECAN-Newton 2	
	$\|\nabla\mathcal{L}\|_\infty$	$\|g\|_\infty$	$\|\nabla\mathcal{L}\|_\infty$	$\|g\|_\infty$	$\|\nabla\mathcal{L}\|_\infty$	$\|g\|_\infty$
1	6.392E-09	3.683E-09	6.106E-16	2.521E-09	9.436E-16	4.773E-16
2	6.731E-09	1.518E-09	1.11E-16	1.303E-09	3.885E-16	3.159E-16
3	4.4205E-09	1.875E-09	2.220E-16	5.908E-10	1.110E-16	1.328E-16
4	7.671E-07	2.493E-10	1.332E-14	1.745E-15	2.220E-15	2.063E-11
5	7.554E-10	3.700E-09	3.330E-16	3.700E-09	2.220E-16	4.105E-16
6	1.981E-09	3.393E-09	4.440E-16	2.423E-09	4.440E-16	2.423E-09
7	6.717E-09	1.800E-09	4.440E-16	1.756E-09	4.440E-16	1.756E-09
8	7.773E-10	9.742E-10	2.2201E-16	9.718E-10	2.2201E-16	9.718E-10
9	1.966E-09	7.142E-09	1.110E-16	5.929E-09	1.110E-16	5.929E-09
10	6.863E-09	1.861E-09	2.027E-15	1.049E-09	2.027E-15	1.049E-09
11	8.690E-09	2.916E-09	7.000E-15	1.019E-13	7.000E-15	1.019E-13
12	1.9641E-09	5.038E-09	6.149E-18	3.015E-16	6.149E-18	3.015E-16
13	2.884E-10	2.962E-09	2.091E-19	2.775E-16	2.091E-19	2.775E-16
14	4.170E-09	2.996E-09	1.689E-17	1.167E-15	1.689E-17	1.167E-15
15	2.327E-09	9.439E-09	9.592E-17	6.581E-15	9.592E-17	6.581E-15
16	2.089E-09	4.668E-09	1.352E-14	4.029E-16	1.352E-14	4.029E-16
17	2.777E-09	8.847E-09	4.552E-16	1.141E-13	4.552E-16	1.141E-13

Tabela 4

Tempo total de execução

	Problemas de pequeno porte (seg.)	
	ALG.-Newton 1	ALG.-Newton 2
1	0.00071	0.00055
2	0.00053	0.00041
3	0.00054	0.00051
4	0.00065	0.00066
5	0.00017	0.00017
6	0.00096	0.00075
7	0.0067	0.00051
8	0.181	0.135
9	0.014	0.021
10	0.00094	0.00086
11	0.0023	0.0024
12	0.0080	0.0086

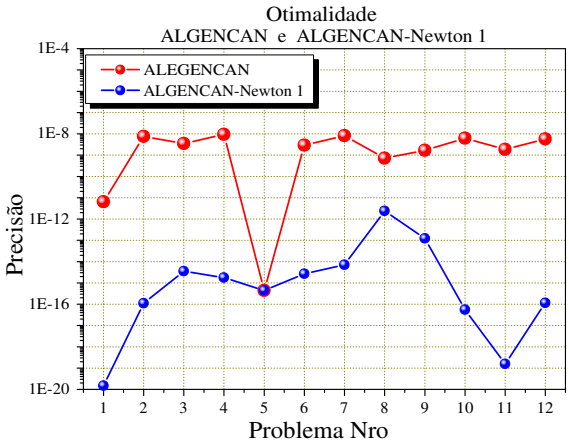
Tabela 5

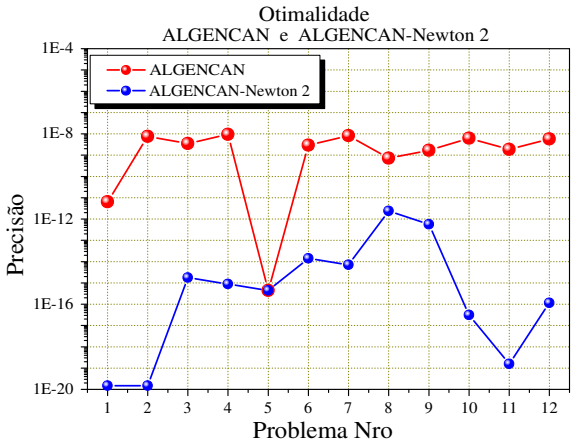
	Problemas de grande porte (seg.)	
	ALG.-Newton 1	ALG.-Newton 2
1	0.0069	0.0069
2	0.0062	0.0051
3	0.0371	0.0341
4	0.0531	0.0265
5	0.054	0.049
6	1.04	1.07
7	6.17	6.01
8	0.269	0.267
9	0.357	0.356
10	4.629	5.011
11	3.360	3.351
12	0.308	0.274
13	0.587	0.576
14	0.346	0.355
15	0.169	0.151
16	0.837	0.826
17	0.212	0.212

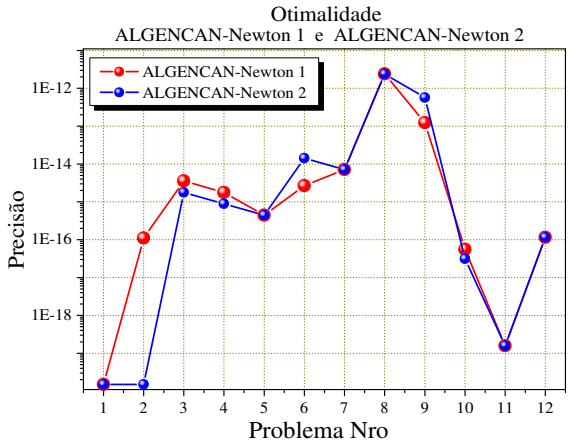
Tabela 6

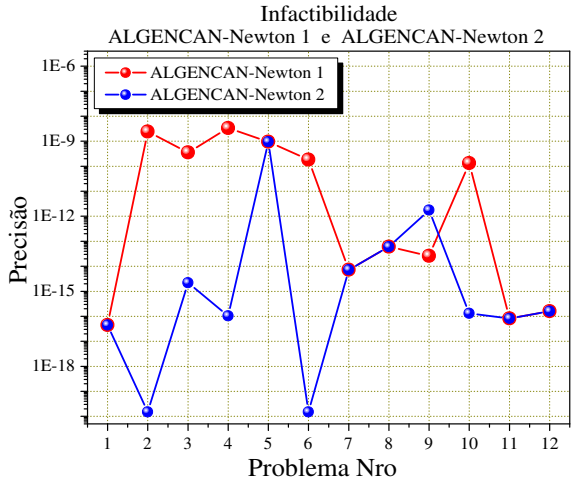
Em seguida, apresentaremos graficamente os resultados obtidos.

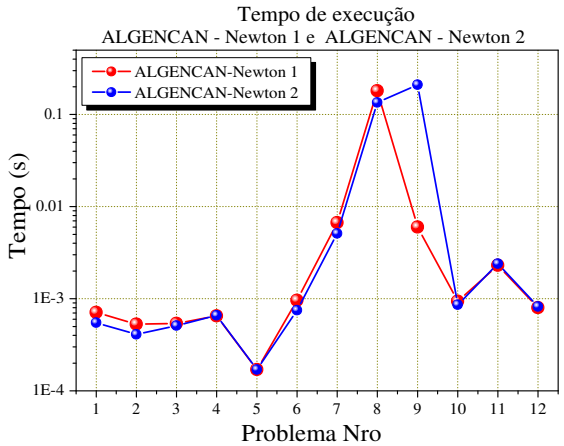
Problemas de pequeno porte



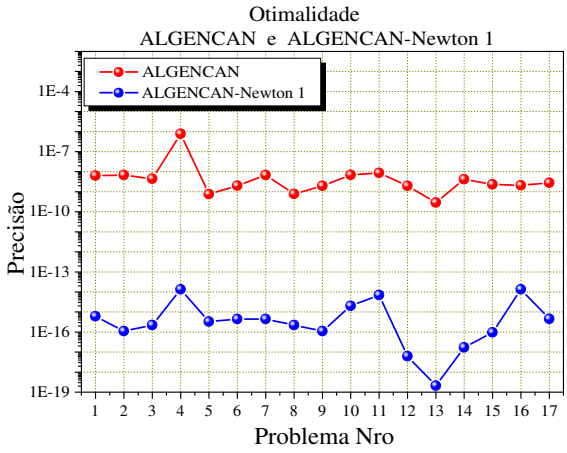


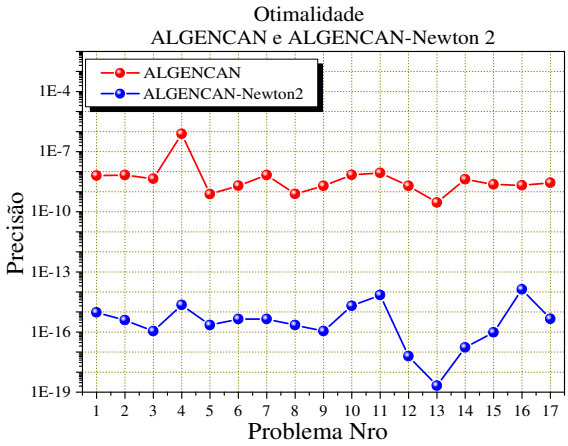


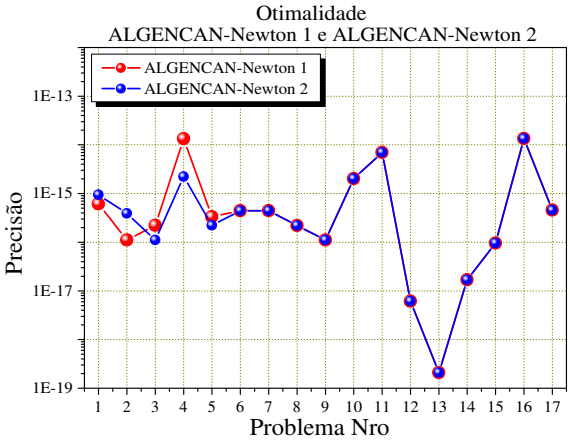


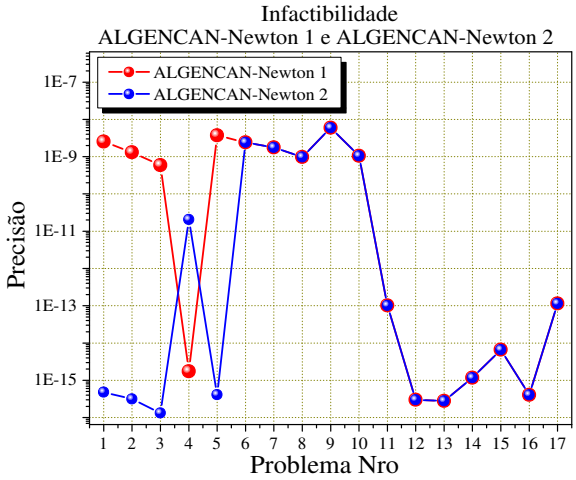


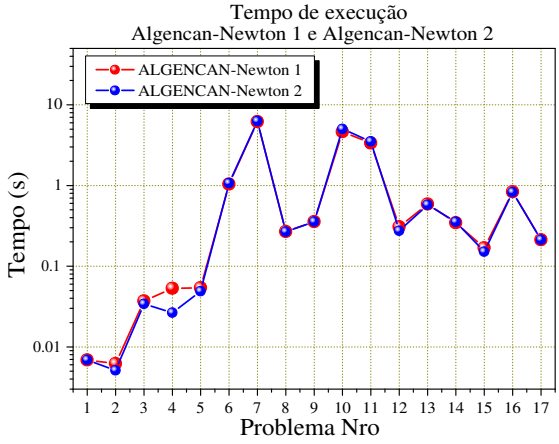
Problemas de grande porte











ALGENCAN-ALGENCAN aliviado

Nesta proposta não obtivemos o resultado desejado. Em nehum dos casos testados, o processo finalizou por cumprir-se o critério da convergência. Ele finalizou por atingir o número máximo de iterações externas, internas ou por ter o parâmetro de penalidade muito alto, o que torna o problema inestável. Em algumos testes, o seqüência foi para outro ponto.

4.4.4 Conclusões dos experimentos computacionais

Dos problemas testados, obtivemos:

Propostas: ALGENCAN - Newton 1 e ALGENCAN - Newton 2

- Sobre um total de 29 problemas testados, ambas propostas conseguiram obter uma aproximação da solução, ou seja os algoritmos convergeram em 100% dos problemas testados.
- Em ALGENCAN - Newton 1 e ALGENCAN - Newton 2 a precisão exigida foi atingida em 25 dos problemas (86,20%).
- Em ambas propostas, 26 dos problemas precisaram de no máximo 5 iterações do método de Newton para a convergência (89.65%).
- Em ALGENCAN - Newton 1, a medida da infactibilidade decresceu em 27 dos problemas (93,10%), enquanto que em ALGENCAN - Newton 2 decresceu em 28 dos problemas (96,55%).
- A medida da infactibilidade obtida com ALGENCAN - Newton 2 em relação à medida obtida por ALGENCAN - Newton 1 foi: inferior em 9 problemas, igual em 18 problemas, e maior em 2 problemas.
- Nos experimentos onde o número de variáveis da proposta ALGENCAN-Newton 2 foi menor ao da proposta ALGENCAN-Newton 1, o tempo de execução da primeira proposta foi inferior ao da segunda. No caso onde o número de variáveis é o mesmo, ambas propostas não têm diferença.

Destacamos que quando no método ALGENCAN exigimos a precisão de 10^{-14}, a medida da infactibilidade atingiu esse valor, enquanto que a medida da otimalidade não (com exceção nos problemas kissing, onde ALGENCAN convergiu em todos os casos testados). No entanto, a proposta ALGENCAN-Newton convergiu em 21 dos problemas (72,4%), dos quais 11 deles obtiveram melhor tempo de execução comparado com as tabelas apresentadas anteriormente.

Proposta: ALGENCAN - ALGENCAN aliviado

Como já expusemos anteriormente, nesta proposta os problemas testados não atingiram a precisão desejada. Mas este resultado era de se esperar, pois o Lagrangiano Aumentado é um método de primeira ordem e os multiplicadores de Lagrange não dão informação muito precisa sobre o problema dual, eles cumprem a função de manter o parâmetro de penalidade controlado no sentido de não crecer muito rapidamente. Assim era, de certa forma, previsível o resultado obtido em nossos testes.

4.5 Conclusões finais

Como podemos observar nas tabelas, as propostas ALGENCAN-Newton 1 e ALGENCAN-Newton 2 tiveram quase o mesmo desempenho em relação à medida de otimalidade, enquanto que com respeito à medida da infactibilidade, a segunda proposta foi melhor. Com ambas propostas conseguimos melhorar a precisão no ponto obtido por ALGENCAN. No entanto ALGENCAN - ALGENCAN aliviado, não logrou essa melhora. A diferença desta última proposta em relação às duas primeiras é que ALGENCAN - ALGENCAN aliviado é um método de primeira ordem, enquanto que ALGENCAN-Newton 1 e ALGENCAN-Newton 2 são métodos de segunda ordem o que os torna métodos de convergência local rápida.

Destacamos que ALGENCAN é um método robusto e obtém uma boa aproximação da solução do problema de PNL com poucas iterações, não entanto, se exigimos maior precisão na solução, como ele é um método de primeira ordem, tem dificuldade para atingir tal precisão. Para lograr melhorar esta medida, aplicamos o método de Newton a um sistema KKT reduzido no ponto obtido por ALGENCAN. Para esta implementação utilizamos um dos métodos estudados no capítulo 1 para resolver sistemas ponto de sela. O método escolhido foi a fatoração LDL^T por suas propriedades de preservar a estrutura esparsa do sistema. Junto com esta melhora, a medida da infactibilidade também decresceu. Assim, conseguimos melhorar a precisão na medida da otimalidade e na medida da infactibilidade. Por outro lado, a variante de eliminar folgas naquelas restrições que têm o multiplicador grande também foi vantajosa pois o processo converge em menos iterações e portanto, o tempo de execução é menor. Finalmente podemos concluir que nossos objetivos foram atingidos, ficando como trabalho futuro a implementação de outros métodos para resolver sistemas ponto de sela.

Apêndice A

Conceitos básicos

Proposição A.1 *Seja* $J(x) = \frac{1}{2}\, x^T H x - f^T x + g$. *Suponhamos* $H = H^T$ *e* $H \geq 0$. *Se* x^* *é um ponto estacionário de* H, *então é minimizador global de* H.

Demonstração: x^* é ponto estacionário de $J(x)$, então

$$\nabla J(x^*) = 0 \iff Hx^* = f$$

$$J(x) = \frac{1}{2}\, x^T H x - f^T x + g =$$

$$\frac{1}{2}\, x^T H x + \frac{1}{2}\, (x^*)^T H x^* - \frac{1}{2}\, (x^*)^T H x - \frac{1}{2}\, x^T H x^* - \frac{1}{2}\, (x^*)^T H x^* +$$

$$+\frac{1}{2}\, (x^*)^T H x + \frac{1}{2}\, x^T H x^* - f^T + g =$$

$$(x^* - x)^T H (x^* - x) - \frac{1}{2}\, (x^*)^T H x^* + (x^*)^T H x - f^T x + g =$$

Agora $(x^*)^T H x^* = (Hx^*)^T x = f^T x$ então

$$(x^* - x)^T H (x^* - x) - \frac{1}{2}\, (x^*)^T H x^* + f^T x - f^T x + g \geq -\frac{1}{2}\, (x^*)^T H x^* + g =$$

$$(x^*)^T H x^* - \frac{1}{2}\, (x^*)^T H x^* - (x^*)^T H x^* + g =$$

$$(x^*)^T H x^* - f^T x^* + g = J(x^*)$$

finalmente

$$J(x) \geq J(x^*) \ \ \forall x. \blacksquare$$

Proposição A.2 *Seja* $A \in \mathbb{R}^{nxn}$ *semidefinida positiva. Se* $x^T A x = 0$ *então* $Ax = 0$

Demonstração: Seja $p(t) = (x+ty)^T A(x+ty)$ com $y \in \mathbb{R}^n$ e x tal que $x^TAx = 0$. Como $A \geq 0$ então $p(t) \geq 0 \ \forall t$.
Agora $p(0) = x^TAx = 0 \implies t = 0$ é um ponto de mínimo de $p(t) \implies p'(0) = 0$.
Derivando $p(t)$ temos

$$\frac{dp}{dt} = 2y^TA(x+ty),$$

$$0 = p'(0) = 2y^TAx \implies y^TAx = 0.$$

Como y é arbitrário temos que $y^TAx = 0 \ \forall y \in \mathbb{R}^n$, em particular escolhendo $y = Ax$

$$(Ax)^TAx = 0 \iff ||Ax||^2 = 0 \iff Ax = 0. \blacksquare$$

Teorema A.1 *Sejam $A \in \mathbb{R}^{nxn}$ simétrica, k um inteiro tal que $1 \leq k \leq n$, $\lambda_1 \leq \ldots \leq \lambda_n$ autovalores de A e $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$. Seja S_k um subespaço k-dimensional de $\mathbb{R}^n$.*

1. Se existe uma constante c_2 tal que $x^TAx \geq c_2x^Tx \quad \forall x \in S_k$, então $\lambda_n \geq \lambda_{n-1} \geq \ldots \geq \lambda_{n-k+1} \geq c_2$.

2. Se existe uma constante c_1 tal que $x^TAx \leq c_1x^Tx \quad \forall x \in S_k$, então $c_1 \geq \lambda_k \geq \ldots \geq \lambda_1$

Demonstração:
1. Por hipótese c_2 satisfaz

$$c_2 \leq \frac{x^TAx}{x^Tx} \ \forall x \in S_k,$$

então

$$c_2 \leq \min_{x \neq 0, \ x \in S_k} \frac{x^TAx}{x^Tx} \leq \max_{V \subseteq \mathbb{R}^n, \ \dim V = k} \ \min_{x \neq 0, \ x \in V} \frac{x^TAx}{x^Tx} = \lambda_{n-k+1},$$

esta última desigualdade é obtida usando o teorema de Courant-Fisher (Horn e Johnson (1985), pág. 179).
Portanto $c_2 \leq \lambda_{n-k+1} \leq .. \leq \lambda_n$, logo fica provada a primeira parte do teorema.
Provamos agora a segunda parte do teorema. Por hipótese c_1 satisfaz

$$c_1 \geq \frac{x^TAx}{x^Tx} \ \forall x \in S_k,$$

então

$$c_1 \geq \max_{x \neq 0, \ x \in S_k} \frac{x^TAx}{x^Tx} \geq \min_{V \subseteq \mathbb{R}^n, \ \dim V = k} \ \max_{x \neq 0, \ x \in V} \frac{x^TAx}{x^Tx} = \lambda_k,$$

novamente última desigualdade é obtida usando o teorema de Courant-Fisher.
Assim $c_1 \geq \lambda_k \geq .. \geq \lambda_1$, o que finaliza a prova do teorema. $\blacksquare$

Corolário A.1 *Sejam A matriz simétrica e S um subespaço de* $\mathbb{R}^n$ *de dimensão k.*
1. Se $x^TAx \geq 0 \ \ \forall x \in S$ *então A tem pelo menos k autovalores não negativos.*
2. Se $x^TAx > 0 \ \ \forall x \in S$, *com x não nulo então A tem pelo menos k autovalores positivos.*

Demonstração:
1. $x^TAx \geq 0 \ \ \forall x \in S$, com $dim(S) = k \implies \lambda_n \geq \ldots \geq \lambda_{n-k+1} \geq 0$ pelo teorema anterior, logo se verifica a primeira afirmação do corolário.
2. Pela primeira parte do corolário, sabemos que A tem pelo menos k autovalores não negativos. Falta provar que eles não podem ser zero. Suponhamos que $\lambda_{n-k+1} = 0$, então

$$0 = \lambda_{n-k+1} = \max_{V \subseteq \mathbb{R}^n,\ \dim V = k} \ \min_{x \neq 0,\ x \in V} \frac{x^TAx}{x^Tx} \geq \min_{x \neq 0,\ x \in S} \frac{x^TAx}{x^Tx} \geq 0,$$

então

$$0 = \min_{x \neq 0,\ x \in S} \frac{x^TAx}{x^Tx} = \min_{||x||_2 = 1,\ x \in S} x^TAx.$$

Agora S é de dimensão finita, $D = \{x \in S \ / \ x^Tx = 1\}$ é compacto (observar que $0 \notin D$), e a função x^TAx é contínua em D, então atinge seu mínimo, quer dizer $\exists\, x_0 \in D$ tal que $x_0^TAx_0 = 0$, mas como $x_0 \neq 0$ isto contradiz a hipótese que $x^TAx > 0 \ \ \forall x \in S$ e $x \neq 0$. Portanto A tem pelo menos k autovalores positivos. ■

Proposição A.3 *Sejam* $M \in \mathbb{R}^{nxn}$ *matriz simétrica definida positiva, e* μ_1 *e* μ_n *o maior e o menor autovalor de M respectivamente. Então*

$$\mu_n\|x\|^2 \leq \langle Mx, x\rangle \leq \mu_1\|x\|^2.$$

Demonstração: Seja λ autovalor de M e x autovetor associado então

$$Mx = \lambda x$$

$$\implies \langle Mx, x\rangle = \lambda\langle x, x\rangle = \lambda\|x\|^2 \leq \mu_1\|x\|^2$$

$$\implies \langle Mx, x\rangle \leq \mu_1\|x\|^2.$$

Por outro lado

$$\lambda\|x\|^2 \geq \mu_n\|x\|^2$$

$$\implies \langle Mx, x\rangle \geq \mu_n\|x\|^2.$$

Portanto, dado que x autovetor de M é arbitrário, nossa proposição se verifica para todo autovetor de M. Agora, como M é simétrica, seus autovetores formam uma base ortonormal para $\mathbb{R}^n$, então $v = \sum_{i=1}^n \alpha_i x_i$ com $\alpha_i \in \mathbb{R} \ \forall\, i = 1, \ldots, n$ e $\{x_1, \ldots, x_n\}$ base ortonormal de autovetores de M. Assim

$$\langle Mv, v\rangle = \sum_{j=1}^n \lambda_j\alpha_j \sum_{i=1}^n \alpha_i\langle x_j, x_i\rangle = \sum_{j=1}^n \lambda_j\alpha_j^2 \leq \mu_1 \sum_{j=1}^n \alpha_j 2 = \mu_1\|v\|^2$$

Portanto temos

$$\langle Mv, v\rangle \leq \mu_1 \|v\|^2.$$

Com um argumento similar, só que limitando por baixo obtemos:

$$\langle Mv, v\rangle \geq \mu_n \|v\|^2$$

e assim fica provada a afirmação. ■

Proposição A.4 *Sejam σ_1 e σ_m o maior e o menor valor singular de B respectivamente, $B \in \mathbb{R}^{mxn}$ com linhas linearmente independentes, então*

$$\sigma_m \leq \frac{\|By\|}{\|y\|} \leq \sigma_1 \ \ \forall y.$$

Demonstração: Seja $y \in \mathbb{R}^n$, pela hipótese sabemos que σ_1 e σ_m correspondem ao maior e menor valor singular de B, então σ_1^2 e σ_m^2 correspondem ao maior e ao menor autovalor da matriz B^TB, logo pela proposição A.3 temos que

$$\sigma_m^2 \|y\|^2 \leq \langle B^TBy, y\rangle \leq \sigma_1^2 \|y\|^2.$$

Por outro lado

$$\langle B^TBy, y\rangle = (B^TBy)^Ty = y^TB^TBy = \|By\|^2.$$

Logo juntanto estos dois resultados obtemos

$$\sigma_m^2 \|y\|^2 \leq \|By\|^2 \leq \sigma_1^2 \|y\|^2,$$

como y é um vetor arbitrario de $\mathbb{R}^n$, fica provada a afirmação. ■

Proposição A.5 *Seja $A \in \mathbb{R}^{n\times n}$ simétrica e definida positiva. Sejam λ_{min} e λ_{max} o menor e o maior autovalor dela. Seja o seguinte método iterativo para resolver o sistema linear $Ax = b$ conhecido como o método de Richardson*

$$w_{n+1} = w_n + \alpha(b - Aw_n),$$

então ele converge se

$$0 < \alpha < \frac{2}{\lambda_{max}},$$

e se reescrevemos a iteração como

$$w_{n+1} = (I - \alpha A)w_n + b,$$

obtemos que $G_\alpha = I - \alpha A$ é a matriz de iteração do método, logo o α que minimiza o raio espectral de G_α ($\rho(G_\alpha)$) é

$$\alpha^* = \frac{2}{\lambda_{max} + \lambda_{min}}$$

(Saad (2003), pág. 114)

Demonstração: O método iterativo converge se e somente se

$$\rho(I - \alpha A) < 1,$$

então

$$|1 - \alpha\lambda_i| < 1 \quad \forall\ \lambda_i \in \rho(A),$$

isto é

$$-1 < 1 - \alpha\lambda_i < 1 \ \ \forall\ \lambda_i \in \rho(A).$$

Portanto,

1. $1 - \alpha\lambda_i < 1 \implies \alpha\lambda_i > 0 \implies \alpha > 0,$
2. $1 - \alpha\lambda_i > -1 \implies 2 > \alpha\lambda_i \implies \frac{2}{\lambda_i} > \alpha \ \ \forall\ \lambda_i \in \rho(A) \implies \frac{2}{\lambda_{max}} > \alpha.$

Logo

$$0 < \alpha < \frac{2}{\lambda_{max}}\ .$$

Agora qual é o α que minimiza $\rho(G_\alpha)$?

$$\rho(G_\alpha) = \max_i\{\ |1 - \alpha\lambda_i|\ \} = max\{|1 - \alpha\lambda_{min}|\,, |1 - \alpha\lambda_{max}|\},$$

então, o problema pode se reescreer como

$$\min_\alpha\ \max\{|1 - \alpha\lambda_{min}|\,, |1 - \alpha\lambda_{max}|\}$$

o mínimo é atingido quando

$$|1 - \alpha\lambda_{min}| = |1 - \alpha\lambda_{max}|\,,$$

assim

1. se $1 - \alpha\lambda_{min} > 0$
   (a) $1 - \alpha\lambda_{min} = \ \ 1 - \alpha\lambda_{max}, \qquad$ se $1 - \alpha\lambda_{max} > 0,$
   (b) $1 - \alpha\lambda_{min} = -1 + \alpha\lambda_{max}, \qquad$ se $1 - \alpha\lambda_{max} < 0,$
2. se $1 - \alpha\lambda_{min} < 0$
   (a) $-1 + \alpha\lambda_{min} = \ \ 1 - \alpha\lambda_{max}, \qquad$ se $1 - \alpha\lambda_{max} > 0,$
   (b) $-1 + \alpha\lambda_{min} = -1 + \alpha\lambda_{max}, \qquad$ se $1 - \alpha\lambda_{max} < 0.$

Analisando os diferentes casos temos:

1.(a) $1 - \alpha\lambda_{min} = 1 - \alpha\lambda_{max} \implies \lambda_{max} = \lambda_{max}$,
então,

$$\min_{\alpha} \rho(G_\alpha) = \min_{\alpha} |1 - \alpha\lambda_{min}| = 0 \implies \alpha = \frac{1}{\lambda_{min}} = \frac{2}{\lambda_{max} + \lambda_{min}},$$

1.(b)

$$1 - \alpha\lambda_{min} = -1 + \alpha\lambda_{max} \implies \alpha = \frac{2}{\lambda_{max} + \lambda_{min}}.$$

Com argumentos similares aos anteriores, de 2.(a) e 2.(b) obtemos que o

$$\alpha = \frac{2}{\lambda_{max} + \lambda_{min}} .$$

Portanto o α que minimiza $\rho(G_\alpha)$ é

$$\alpha^* = \frac{2}{\lambda_{max} + \lambda_{min}} . \blacksquare$$

Referências Bibliográficas

[1] R. Andreani, E. G. Birgin, J. M. Martínez, M. L. Schuverdt (2008), "Augmented Lagrangian methods under the Constant Positive Linear Dependence constraint qualification", Mathematical Programming **111**, pp. 5-32.

[2] R. Andreani, E. G. Birgin, J. M. Martínez, M. L. Schuverdt (2007), "On Augmented Lagrangian methods with general lower-level constraints", SIAM Journal on Optimization **18**, pp. 1286-1309.

[3] K. J. Arrow, L. Hurwicz, H. Uzawa (1958), *Studies in Linear and Nonlinear Programming*, Standford University Press, Stanford, CA.

[4] M. Benzi, G. H Golub, J. Liesen (2005), "Numerical Solution of saddle point problems", Cambridge University Press.

[5] A. R. Bergen (1986), *Power Systems Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

[6] E. G. Birgin, R. Castillo and J. M. Martínez (2005), "Numerical comparison of Augmented Lagrangian algoritms for nonconvex problems", Computational Optimization and Applications **31**, pp. 31-56.

[7] E. G. Birgin, J. M. Martínez (2008), "Improving ultimate convergence of an Augmented Lagrangian method", Optimization Methods and Software **23**, pp. 177-195.

[8] A. Björk (1996), *Numerical Methods for Least Squares Problems* , SIAM, Philadelphia, PA.

[9] A. Bossavit (1998), "Mixed systems of algebric equations in computational electromagnetics", COMPEL **17**, pp. 59-63.

[10] L. O. Chua, C. A. Desoer, E. S. Kuh (1987), *Linear and Nonlinear Circuits*, McGraw-Hill, New York.

[11] F. Duchin, D. B. Szyld (1979), "Application of sparse matrix techniques to inter-regional input-output analysis", Economics of Planning **15**, pp. 142-167.

[12] I. S. Duff, A. M. Erisman, J. K. Reid (1986), *Direct Methods for Sparse Matrices*, Monographs on Numerical Analysis, Oxford University Press, Oxford.

[13] R. Fletcher (1987), *Practical Methods of Optimization*, 2nd ed, Wiley, Chichester, UK.

[14] P. E. Gill, W. Murray, M. H. Wright (1981), *Practical Optimization*, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London.

[15] R. Glowinski (1984), *Numerical Methods for Nonlinear Variational Problems*, Springer Series in Computational Physics, Springer, New York.

[16] G. H. Golub, C. Grief (2003), "On solving block-structured indefinite linear systems", SIAM J. Sci. Comput. **24**, pp. 2076-2092.

[17] G. H. Golub, M. L. Overton (1988), "The convergence of inexact Chebyshev and Richardson iterative methods for solving linear sistems", Numer. Math. **53**, pp. 571-593.

[18] G. H. Golub, C. F. Van Loan (1996), *Matriz Computations*, Johns Hopkins Studies in the Mathematical Sciences, 3rd ed, Johns Hopkins University Press, Baltimore, MD.

[19] E. L. Hall (1979), *Computer Image Processing and Recognition*, Academic Press, New York.

[20] W. Hock, K. Schittkowski (1981), "Test examples for nonlinear programming codes", Lecture Notes in Economics and Mathematical Systems Vol. 187, Spring-Verlag, Berlin-New York, 1981.

[21] A. Horn, C. R. Johnson (1985), *Matrix Analysis*, Cambridge University Press, Cambridge.

[22] W. Leontief, F. Duchin, D. B. Szyld (1985), "New approaches in economic analysis", Science **228**, pp. 419-422.

[23] G. Luenberger (1989), *Linear and Nonlinear Programming*, Addison Wesley, Reading, MA.

[24] J. M. Martínez, S. A. Santos (1995), *Métodos Computacionais de Otimização*, Departamento de Matemática Aplicada, IMECC.

[25] J. M. Martínez (2006), *Otimização Prática Usando o Langrangeano Aumentado*, Departamento de Matemática Aplicada, IMECC, UNICAMP.

[26] L. F. de Mendonça, V. L da Rocha Lopez e J. M. Martínez (2006), "Aceleração de métodos do tipo Langrangeano Aumentado para resolver problemas de otimização com restrições de desigualdade", Departamento de Matemática Aplicada, IMECC, UNICAMP.

[27] J. Nocedal , S. J. Wright (1999), *Numerical Optimization*, Springer Series in Operations Research, Springer, Berlin.

[28] I. Perugia (1997), "A field-based mixed formulation for two-dimensional magnetostatic problem", SIAM J. Numer. Anal. **34**, pp. 2382-2391.

[29] I. Perugia, V, Simoncini, M. Arioli (1999), "Linear algebra methods in a mixed approximation of magnetostatic problems", SIAM J. Sci. Comput. **21**, pp. 1085-1101.

[30] A. Quarteroni , A. Valli (1994), *Numerical Approximation of Partial Differential Equations*, Vol. 23, Springer Series in Computational Mathematics, Springer, Berlin.

[31] T. Rusten , R. Winther (1992), *A preconditioned iterative methods for saddle point problems*, SIAM J. Matrix Anal. Appl. **13**, pp. 887-904.

[32] Y. Saad (2003), *Iterative Methods for Sparse Linear Systems*, 2nd ed, SIAM, Philadelphia, PA.

[33] D. B. Szyld (1981), *Using sparse matrix techniques to solve a model of the world economy, in Sparse Matrices and Their Uses*, (I. S. Duff, ed), Academic Press, pp. 357-365.

[34] R. Termam (1984), *Navier-Stokes Equations: Theory and Numerical Analysis*, Vol. 2, Studies in Mathematics and its Applications, 3rd ed, North-Holland, Amsterdam.

[35] S. Turek (1999), *Efficient Solvers for Incomprenssible Flow Problems: An Algorithmic and Computational Approach*, Vol. 6, Lecture Notes in Computational Science and Engineering, Springer, Berlin.

[36] P. Wesseling (2001), *Principles of Computational Fuid Dynamics*, Vol. 29, Springer Series in Computational Mathemathics, Springer, Berlin.

[37] M. H. Wright (1992), "Interior methods for constrained optimization", Acta Numérica, Vol. 1, Cambridge University Press, pp. 341-407.

[38] S. J. Wright (1997), *Primal-Dual Interior Point Methods*, SIAM, Philadelphia, PA.