UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

MAYCON PEREIRA DE SOUZA

Precondicionador Separador Duplo: Uma nova classe de Precondicionadores

Campinas

2024

Maycon Pereira de Souza

Precondicionador Separador Duplo: Uma nova classe de Precondicionadores

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Matemática Aplicada.

Orientador: Aurelio Ribeiro Leite de Oliveira

Este trabalho corresponde à versão final da Tese defendida pelo aluno Maycon Pereira de Souza e orientada pelo Prof. Dr. Aurelio Ribeiro Leite de Oliveira.

Campinas

2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

So89p Souza, Maycon Pereira de, 1985-
Precondicionador separador duplo : uma nova classe de precondicionadores / Maycon Pereira de Souza. – Campinas, SP : [s.n.], 2024.

Orientador: Aurelio Ribeiro Leite de Oliveira.
Tese (doutorado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Matemática, Estatística e Computação Científica.

1. Pré-condicionadores. 2. Programação linear. 3. Métodos de pontos interiores. I. Oliveira, Aurelio Ribeiro Leite de, 1962-. II. Universidade Estadual de Campinas (UNICAMP). Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações Complementares

Título em outro idioma: Double splitting preconditioner : a new class of preconditioners
Palavras-chave em inglês:
Preconditioners
Linear programming
Interior-point methods
Área de concentração: Matemática Aplicada
Titulação: Doutor em Matemática Aplicada
Banca examinadora:
Aurelio Ribeiro Leite de Oliveira [Orientador]
Francisco de Assis Magalhães Gomes Neto
Luciana Casacio
Cecilia Orellana Castro
Manolo Rodriguez Heredia
Data de defesa: 24-06-2024
Programa de Pós-Graduação: Matemática Aplicada

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0001-7991-9794
- Currículo Lattes do autor: http://lattes.cnpq.br/8205872091845759

Tese de Doutorado defendida em 24 de junho de 2024 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). AURELIO RIBEIRO LEITE DE OLIVEIRA

Prof(a). Dr(a). FRANCISCO DE ASSIS MAGALHÃES GOMES NETO

Prof(a). Dr(a). LUCIANA CASACIO

Prof(a). Dr(a). CECILIA ORELLANA CASTRO

Prof(a). Dr(a). MANOLO RODRIGUEZ HEREDIA

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Aos meus pais, à minha esposa Cleidiane e aos meus filhos Davi, Estêvão, Teresa e Catarina. E sobretudo a Deus, que me deu a vida e a oportunidade de estudar e de sempre querer ir além.

"Somos na realidade o que somos diante de Deus."

São Francisco de Assis

Agradecimentos

À minha esposa Cleidiane, a quem amo muito, por toda compreensão e força dados durante esse estudo.

Aos meus filhos, Davi, Estêvão, Teresa e Catarina, que são as joias da minha vida.

A meus pais e irmãos, a quem amo muito.

Ao meu orientador Aurelio Ribeiro Leite de Oliveira, por gentilmente me aceitar como aluno e por todo o apoio e suporte dados ao longo desta tese.

Aos meus colegas de Doutorado, em especial ao Samuel Francisco.

À Catalina Jaramillo.

À banca, pelas valiosas sugestões e pelo trabalho dedicado à avaliação do presente estudo.

Agradeço ao Instituto Federal de Goiás por me conceder a oportunidade de aprimoramento por meio do programa de capacitação de servidores.

E, sobretudo, a Deus por Sua misericórdia infinita, que está sempre sobre todos nós.

Resumo

Neste trabalho propomos uma generalização da ideia dada em (OLIVEIRA, 1997). O autor citado apresenta uma nova abordagem de se precondicionar o sistema aumentado, sistema oriundo da aplicação dos métodos de pontos interiores a um problema de Programação Linear, denominado Precondicionador Separador, em que é feita uma separação nos elementos da matriz do primeiro bloco do sistema aumentado, formando dois conjuntos distintos.

Foi mostrado também que o Precondicionador Separador funciona muito bem próximo a uma solução do problema de programação linear, característica que é muito bem vista, pois próximo à solução o sistema tende a ser extremamente mal condicionado.

Com base na ideia apresentada por (OLIVEIRA, 1997), propomos um precondicionador duplo para o sistema aumentado, acrescentando alguns parâmetros. Dessa forma, surgem novos precondicionadores, que chamamos de Precondicionador Separador Duplo. Conseguimos mostrar que o precondicionador assim construído é uma generalização do Precondicionador Separador proposto por (OLIVEIRA, 1997). Além disso, mostramos que temos uma variedade de escolhas para alguns blocos que compõem o Precondicionador Separador Duplo e que levam à mesma matriz K (primeiro bloco da matriz precondicionada duplamente) que é a chave principal para a eficiência do método proposto. Através dos parâmetros que fazem parte do Precondicionador Separador Duplo, podemos fazer com que a matriz em questão não tenha autovalores suficientemente próximos de zero, o que pode prejudicar o desempenho de métodos iterativos. Além do mais, através de escolhas adequadas desses parâmetros, podemos fazer com que a matriz seja muito bem condicionada nas iterações finais do método proposto.

Conseguimos encontrar funções desses parâmetros que fornecem o número de condição na norma-2, assim pudemos minimizar essas funções e encontrar toda uma família de soluções que produzem o mesmo número de condição entre o Precondicionador Separador Duplo e o Precondicionador Separador. Provamos assim que o Precondicionador Separador é ótimo nesse sentido. Também é demonstrado que o Precondicionador Separador é ótimo para a norma-1 e norma-∞ dada a melhor opção de parâmetros para minimizar a norma-2. Conseguimos empiricamente, com a utilização da norma-1 e norma-∞, verificar que o Precondicionador Separador Duplo pode obter um número de condição menor do que o Precondicionador Separador.

Palavras-chave: Precondicionador Separador Duplo; Sistema Aumentado; Método de Pontos Interiores.

Abstract

In this work we propose a generalization of the idea given in (OLIVEIRA, 1997). The aforementioned author brings a new approach to precondition the augmented system, a system that comes from the application of methods of interiors points in a Linear Programming problem, called Splitting Preconditioner, in which a separation is made in the elements of the matrix of the first block of the augmented system, forming two distinct sets.

It was also demonstrated that the Splitting Preconditioner works very well close to a solution to the linear programming problem, a characteristic that is very well regarded, as close to the solution the system tends to be extremely poorly conditioned.

Based on the idea presented by (OLIVEIRA, 1997), we propose a double preconditioner for the augmented system, adding some parameters. In this way, it allowed us a greater range of possibilities for new preconditioners, which we call Double Splitting Preconditioner. We were able to show that the preconditioner thus constructed is a generalization of the Splitting Preconditioner proposed by (OLIVEIRA, 1997). Furthermore, we show that we have a variety of choices for some blocks that make up the Double Splitting Preconditioner and that lead to the same matrix K (first block of the doubly preconditioned matrix) which is the main key to the efficiency of the method proposed. Through the parameters that are part of the Double Splitting Preconditioner, we can ensure that the matrix in question does not have eigenvalues close enough to zero, which can harm the performance of the iterative methods. Furthermore, through appropriate choices of these parameters, we can ensure that the matrix is very well conditioned in the final iterations of the proposed method.

We were able to find functions of these parameters that give the condition number in the 2-norm, so we were able to minimize these functions and find a whole family of solutions that produce the same condition number between the Double Splitting Preconditioner and the Splitting Preconditioner. We thus prove that the Splitting Preconditioner is optimal in this sense. It is also shown that the Splitting Preconditioner is optimal for 1-norm and ∞-norm given the best choice of parameters to minimize 2-norm. We were able to empirically, using the 1-norm and ∞-norm, verify that the Double Splitting Preconditioner can obtain a lower condition number than the Splitting Preconditioner.

Keywords: Double Splitting Preconditioner; Augmented System; Interior Point Method.

Lista de ilustrações

Figura 1 – Caso $n > 2m$ Fonte: Elaborado pelo autor. 59
Figura 2 – Caso $n = 2m$ Fonte: Elaborado pelo autor. 59
Figura 3 – Caso $n < 2m$ Fonte: Elaborado pelo autor. 61
Figura 4 – Região de autovalores, caso λ_{max} esteja a direita da origem Fonte: Elaborado pelo autor. 62
Figura 5 – Região de autovalores, caso $-\lambda_{max}$ esteja a esquerda da origem Fonte: Elaborado pelo autor. 63
Figura 6 – Região de autovalores, caso λ_{max} esteja a direita da origem Fonte: Elaborado pelo autor. 63
Figura 7 – Região de autovalores, caso $-\lambda_{max}$ esteja a esquerda da origem Fonte: Elaborado pelo autor. 64
Figura 8 – Região de autovalores, caso λ_{max} esteja a direita da origem Fonte: Elaborado pelo autor. 64
Figura 9 – Região de autovalores, caso $-\lambda_{max}$ esteja a esquerda da origem Fonte: Elaborado pelo autor. 65
Figura 10 – Decaimento no número de condição da matriz K, caso $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ Fonte: Elaborado pelo autor. 67
Figura 11 – Decaimento no número de condição da matriz K, caso $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$ Fonte: Elaborado pelo autor trocar esse figura. 67
Figura 12 – Caso $n > 2m$ e W matriz de posto incompleto Fonte: Elaborado pelo autor. 71
Figura 13 – Caso $n = 2m$ e W matriz de posto incompleto Fonte: Elaborado pelo autor. 71
Figura 14 – Caso $n < 2m$ e W matriz de posto incompleto Fonte: Elaborado pelo autor. 73
Figura 15 – SAPD x Separador (Parâmetros $\alpha = 1$, $\beta = 0,5$ e $\gamma = 0,0001$ para SAPD.) Fonte: Elaborado pelo autor. 89
Figura 16 – SAPD x Separador: Problemas em que há convergência para a solução ótima. Fonte: Elaborado pelo autor. 90
Figura 17 – SAPD x Separador (Parâmetros $\alpha = 1$, $\beta = 1$ e $\gamma = 0,01$ para SAPD.) Fonte: Elaborado pelo autor. 92
Figura 18 – SAPD x Separador: Problemas em que há convergência para a solução ótima. Fonte: Elaborado pelo autor. 92
Figura 19 – SAPD x Separador com a utilização da função **fmincon**. 97

Lista de símbolos

$\mathbb{R}_+^*$	Conjunto dos números reais positivos não nulos.
$\mathbb{R}^{m\times n}$	Espaço vetorial das matrizes de ordem $m \times n$.
$\mathbf{I}_n$	Matriz identidade de ordem n.
$A \in \mathbb{R}^{m\times n}$	Matriz retangular de ordem $m \times n$.
x^t	Transposto do vetor x.
$x^t y$	Produto interno entre os vetores x e y.
$\|x\|_1$	Norma-1 de vetores reais ou matrizes reais.
$\|x\|_2$	Norma-2 de vetores reais ou matrizes reais (Norma Euclidiana).
$\|A\|_F$	Norma de Frobenius da Matriz A.
$\|x\|$	Módulo de um número real.
e	Vetor de uns, ou seja, $e = (1, 1, \cdots, 1)$.
$diag(x)$	Matriz diagonal com o vetor x na diagonal.
$min(a, b)$	Mínimo entre os valores a e b.
0	Significa tanto o escalar zero quanto a matriz nula de ordem apropriada.
(a, b)	Intervalo aberto em a e b.
[a, b)	Intervalo fechado em a e aberto em b.
$\kappa(A)$	Número de condição da matriz A segundo a norma-2.

Sumário

1 INTRODUÇÃO 14

2 PROGRAMAÇÃO LINEAR 17
2.1 Preâmbulo 17
2.2 Método de Newton 19
2.3 Método de pontos interiores Primal-Dual 19
2.4 Método Preditor-Corretor 22
2.5 Sistema Aumentado *versus* Equações Normais 25
2.6 Variáveis canalizadas 27
2.7 Precondicionamento 29
2.8 Método dos Gradientes Conjugados 31
2.9 Programação Quadrática Convexa 33

3 PRECONDICIONADORES 36
3.1 Fatoração controlada de Cholesky 36
3.2 Precondicionador Separador 38
3.3 Precondicionador Híbrido 42

4 PRECONDICIONADOR SEPARADOR DUPLO 45
4.1 Construção do Precondicionador Duplo 45
4.2 Escolha dos blocos do Sistema Precondicionado Duplamente 46
4.3 Escolha dos blocos *F, U e V* 48
4.4 Análise da matriz do 1º Bloco 49
4.5 Aplicação do Precondicionador Separador Duplo na resolução do Sistema Aumentado 52
4.6 Algumas considerações sobre as matrizes S e K 55
4.7 Considerações sobre o número de condição da matriz K na norma-2 61
4.8 Um pouco sobre a matriz W 65
4.8.1 Diferença entre os precondicionadores 67
4.9 Matriz W de posto Incompleto 68
4.10 Número de condição da matriz K revisitado 74
4.10.1 Valores Ótimos 79
4.11 Número de condição da matriz K na norma-∞ ou norma-1 80
4.12 Comentário dos resultados do Programa Mathematica 83
4.13 Solução do tipo $\gamma = -2\alpha\beta$ na norma infinita 85
4.14 Combinando as normas 86

5 EXPERIMENTOS COMPUTACIONAIS 87
5.1 Testes em problemas utilizando a norma-∞ e norma-1 87

6 CONCLUSÕES 99
6.1 Perspectivas Futuras 100

REFERÊNCIAS 101

1 Introdução

De acordo com (BAZARAA; JARVIS; SHERALI, 2010), a programação linear trata o problema de minimizar ou maximizar uma função linear sujeita a restrições lineares de igualdade e/ou desigualdade. Em 1947 houve o desenvolvimento do método Simplex por George B. Dantzig, e desde então a programação linear tem sido utilizada nos mais variados campos, como por exemplo, governamental, militar, industrial e planejamento urbano, entre outros. A popularidade da programação linear pode ser atribuída, por exemplo, ao fato de se resolver problemas grandes e complexos em um período de tempo razoável com a utilização de algoritmos eficientes e computadores modernos.

Ainda em consonância com (BAZARAA; JARVIS; SHERALI, 2010), o método Simplex possui grande aceitação por causa da sua capacidade de produzir soluções em um período de tempo razoável.

De acordo com (GONDZIO, 2012), desde 1940 a programação linear tem sido o centro de atenção em otimização. Por décadas o algoritmo Simplex foi o único método disponível para resolver essa importante classe de problemas em otimização. Embora o método Simplex provavelmente tenha complexidade não polinomial (no pior caso pode levar muitas iterações que dependem exponencialmente da dimensão do problema, ou seja, quanto maior a dimensão do problema, mais rápido aumenta o tempo computacional para encontrar uma solução ótima), na prática tem-se mostrado um método muito eficiente e confiável. O método Simplex visita os vértices do politopo definido pelas restrições do problema, e como esse número pode ser elevadíssimo, o método corre o risco de ter que visitar muitos deles antes de encontrar um ótimo.

Ainda em concordância com (GONDZIO, 2012), o primeiro algoritmo polinomial para programação linear foi desenvolvido por Khachiyan, (KHACHIYAN, 1979). Seu algoritmo elipsoide constrói uma série de elipsoides inscritos no conjunto factível. Os centros desses elipsoides formam uma sequência de pontos que convergem para uma solução ótima de um problema de programação linear. E como citado em (GOLDFARB; TODD, 1988), temos que o método dos elipsoides é de pouca importância para a computação prática, embora seu significado teórico para otimização linear, não-linear e combinatória seja de grande importância. Ademais, abriu-se assim uma nova vertente para a programação linear, incentivando a busca de outros métodos de programação linear, culminando nos métodos de pontos interiores.

Em (VANDERBEI, 2014), é feita uma comparação entre uma variação do método Simplex e uma variação dos métodos de pontos interiores, a qual nos mostra que para muitos problemas de pequeno e médio porte, o método Simplex é mais rápido. Já

para problemas maiores os métodos de pontos interiores tendem a ser melhores, embora esse resultado dependa da classe dos problemas em questão.

Em (GONDZIO, 2012) é dito que tanto o método Simplex quanto os métodos de pontos interiores são amplamente usados hoje em dia e continuam a competir entre si. Embora problemas de grande porte geralmente parecem favorecer os métodos de pontos interiores, nem sempre é possível predizer um vencedor em uma determinada classe de problemas, pois pode depender da estrutura do problema e sua esparsidade.

Dado uma sistema linear, às vezes esse sistema não possui boas características, como por exemplo, não possuindo um bom número de condição. Desse modo, uma técnica muito utilizada para melhorar o número de condição dos coeficientes de uma matriz é a técnica de precondicionamento, pois podemos fazer com que a matriz dos coeficientes tenha melhores propriedades espectrais, como por exemplo, número de condição próximo a 1 ou autovalores agrupados. Utilizamos os precondicionadores para resolver o sistema linear através de algum método iterativo.

A partir da ideia apresentada por (OLIVEIRA, 1997), propomos precondicionar duas vezes o sistema aumentado, incrementando alguns parâmetros. Desta forma, surgem novos precondicionadores, que chamamos de Precondicionador Separador Duplo. Conseguimos mostrar que o precondicionador assim construído é uma generalização do Precondicionador Separador proposto por (OLIVEIRA, 1997), isto é, o sistema aumentado oriundo do Precondicionador Separador Duplo é mais geral do que o Precondicionador Separador.

Além disso, obtemos um importante resultado que mostra que temos uma variedade de escolhas para alguns blocos que compõem o Precondicionador Separador Duplo, e que conduzem a mesma matriz K (primeiro bloco da matriz precondicionada duplamente), que é chave principal para eficiência do método proposto. Através dos parâmetros que fazem parte do Precondicionador Separador Duplo, podemos fazer com que a matriz em questão não tenha autovalores suficientemente próximos de zero, o que pode prejudicar o desempenho de métodos iterativos. Além do mais, através de escolhas adequadas desses parâmetros, podemos fazer com que a matriz seja muito bem condicionada nas iterações finais do método proposto. E de certo modo, temos a priori, um conhecimento sobre qual será o número de condição da matriz nas iterações finais.

Portanto, um dos objetivos desta tese é proporcionar um precondicionador que melhore o número de condição da matriz precondicionada.

Esta tese é constituída de mais cinco capítulos. No Capítulo 2 fazemos uma revisão de alguns conceitos importantes de programação linear, bem como do método preditor-corretor, também fizemos uma contraposição entre o sistema aumentado e o sistema de equações normais. Ainda neste capítulo, é apresentado o método de gradientes

conjugados e comentamos sobre programação quadrática convexa.

No Capítulo 3 falamos especificamente sobre três importantes tipos de precondicionadores, a Fatoração Controlada de Cholesky (CAMPOS, 1995), o Precondicionador Separador (OLIVEIRA, 1997) e o Precondicionador Híbrido (BOCANEGRA, 2005).

Já no Capítulo 4, discorremos sobre o Precondicionador Separador Duplo, como se deu a construção do mesmo e mostramos alguns teoremas que nos ajudam a compreender o raio espectral da matriz K. Também discorremos sobre a aplicação do Precondicionador Separador Duplo na resolução do sistema aumentado, e ainda fizemos algumas considerações sobre o número de condição da matriz K. Mostramos que o Precondicionador Separador é ótimo para a norma-2. Mostramos também que dados os parâmetros ótimos para a norma-2, que o número de condição ótimo para a norma-1 e norma-∞, é o mesmo número do Precondicionador Separador.

No Capítulo 5 fizemos testes computacionais para quantificarmos a eficácia do precondicionador. Encontramos situações onde o Precondicionador Separador Duplo é superior.

O Capítulo 6 mostra a conclusão sobre a nova abordagem proposta e perspectivas para pesquisa futura.

2 Programação Linear

2.1 Preâmbulo

Discorreremos neste capítulo sobre alguns conceitos básicos e definições, que serão necessários para o entendimento da teoria que se segue. Algumas das ideias apresentadas aqui podem ser encontradas em (WRIGHT, 1997).

Um problema de programação linear pode ser escrito na forma padrão como abaixo

$$\begin{aligned} \min\ & f(x) = c^t x \\ \text{s.a}\ & Ax = b \\ & x \geqslant 0, \end{aligned} \tag{2.1}$$

em que $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ e $x \in \mathbb{R}^n$.

Chamamos A de matriz dos coeficientes, c é o vetor de custos, b é o vetor dos termos independentes ou de recursos e x é o vetor das variáveis ou incógnitas, conforme (ARENALES; ARMENTANO, 2006). Sem perda de generalidade considera-se que a matriz A possui posto completo m, ou seja, não possui linhas redundantes.

Note que em (2.1), a variável x é não negativa, quando ocorrer em um dado problema que alguma variável seja livre, ou seja, sem sinal definido, podemos reescrever esta variável como diferença entre duas variáveis não negativas.

Definição 2.1. *Um ponto x que satisfaz as restrições do problema (2.1) é chamado de factível.*

Definição 2.2. *Chamamos de conjunto factível ao conjunto de todos os pontos factíveis.*

Definição 2.3. *Um ponto factível x^* que minimiza a função objetivo é chamado de solução ótima.*

Associado a (2.1), chamado de problema primal, temos o seguinte problema dual

$$\begin{aligned} \max\ & g(y,z) = b^t y \\ \text{s.a}\ & A^t y + z = c \\ & y \text{ livre},\ z \geqslant 0, \end{aligned} \tag{2.2}$$

em que $y \in \mathbb{R}^m$ é o vetor das variáveis duais livres e $z \in \mathbb{R}^n$ é o vetor das variáveis de folga duais.

Definição 2.4. *Dizemos que $x^0 \in \mathbb{R}^n$ é ponto factível primal se satisfazer $Ax^0 = b$ e $x^0 \geqslant 0$, e se tivermos x^0 estritamente maior que zero, dizemos que o mesmo é um ponto interior factível primal.*

Definição 2.5. *Dizemos que $y^0 \in \mathbb{R}^m$ e $z^0 \in \mathbb{R}^n$ é ponto factível dual se satisfazer $A^t y^0 + z^0 = c$, com $z^0 \geqslant 0$ e y^0 livre, e se tivermos z^0 estritamente maior que zero, dizemos que o mesmo é um ponto interior factível dual.*

Chamamos de par primal-dual os problemas (2.1) e (2.2). Para um ponto x e (y, z) factível de (2.1) e (2.2), respectivamente, temos que $g(y, z) = b^t y \leqslant c^t x = f(x)$, ou seja, a função objetivo do problema de maximização dual é um limitante inferior para o problema de minimização primal, já a função objetivo do problema primal é um limitante superior para o dual. E as duas funções objetivo coincidem no valor ótimo, quando este existe, ou seja, $g(y^*, z^*) = f(x^*)$ para x^* solução de (2.1) e (y^*, z^*) solução de (2.2).

Observação 1: Note que na Definição 2.5 não temos restrição alguma sobre o vetor $y \in \mathbb{R}^m$, ou seja, o mesmo pode tanto ser positivo ou negativo. O *gap* do par primal-dual é dado pela diferença entre o valor da função objetivo primal e dual, isto é, $\gamma = c^t x - b^t y$. E para x e (y, z) pontos factíveis, temos que $\gamma = z^t x$.

Definição 2.6. *Dizemos que x^* é minimizador global de (2.1), se $f(x^*) \leqslant f(x)$, $\forall\, x$ factível.*

Discorreremos agora sobre quais aspectos nos dizem que um determinado ponto pode ou não possuir um valor ótimo nos problemas (2.1) e (2.2), isto é, para que um ponto x e (y, z) seja minimizador global do problema (2.1) e maximizador global de (2.2), respectivamente, ele deve satisfazer as condições dadas em (2.3).

Condições de otimalidade

Para que um determinado ponto (x, y, z) seja ótimo para os problemas (2.1) e (2.2), ele deve satisfazer as seguintes condições

$$\begin{aligned} Ax = b, \quad x \geqslant 0 \quad & \text{(Factibilidade Primal)} \\ A^t y + z = c, \quad z \geqslant 0 \quad & \text{(Factibilidade Dual)} \\ XZe = 0, \quad & \text{(Condição de Complementaridade)} \end{aligned} \tag{2.3}$$

onde $X = diag(x)$, $Z = diag(z)$ e e é vetor de uns. Isto nos diz que para um ponto (x, y, z) ser ótimo do par primal-dual, ele deve satisfazer as duas primeiras condições lineares de igualdades em (2.3) e uma condição não linear.

Antes de descrevermos o método de pontos interiores primal-dual, vamos apresentar o método de Newton, que é de extrema relevância para compreendermos bem o método em questão.

2.2 Método de Newton

De acordo (RIBEIRO; KARAS, 2014), o método de Newton é uma das ferramentas matemáticas mais importantes em otimização. Desse modo, aplicaremos o método no sistema não linear (2.3) sem as restrições de não negatividade.

Considere uma função vetorial $F : \mathbb{R}^n \to \mathbb{R}^n$ de classe C^1. Queremos resolver o problema

$$F(x) = 0.$$

Para isto, tomamos a aproximação de Taylor de primeira ordem de F em torno do ponto x^k, que é dada por $F(x) \approx F(x^k) + J(x^k)(x - x^k)$, onde $J(x^k) = \nabla F(x^k)^t$ é o jacobiano da função F. Logo uma aproximação para o problema é $F(x^k)+J(x^k)(x-x^k) = 0$. Tomando $d^k = (x - x^k)$, e caso o jacobiano da função seja não singular, temos $d^k = -J(x^k)^{-1}F(x^k)$. Seja t_k o tamanho de passo a se tomar. Logo, o próximo ponto é

$$x^{k+1} = x^k + t_k d^k. \tag{2.4}$$

Note que se tomarmos $t_k = 1$ em (2.4), temos o método de Newton puro. Perceba ainda que, de fato, não invertemos a matriz jacobiana, pois esta é uma operação muito cara computacionalmente, mas resolvemos o sistema linear $J(x^k)d^k = -F(x^k)$.

2.3 Método de pontos interiores Primal-Dual

Os métodos primal-dual, como o nome sugere, resolvem simultaneamente os problemas primal e dual.

De acordo com (WRIGHT, 1997) os métodos do tipo primal-dual encontram uma solução (x^*, y^*, z^*) pela aplicação de variações do método de Newton às condições de otimalidade dada em (2.3), e assim, modificam-se as direções de busca e tamanho de passo, de tal modo que as desigualdades $(x, z) \geqslant 0$ sejam satisfeitas estritamente a cada iteração.

Considere a função vetorial $F : \mathbb{R}^{2n+m} \to \mathbb{R}^{2n+m}$ dada por

$$F(x, y, z) = \begin{bmatrix} Ax - b \\ A^t y + z - c \\ XZe \end{bmatrix} = 0, \tag{2.5}$$

para $(x, z) \geqslant 0$. Todos os métodos do tipo primal-dual geram iterandos (x^k, y^k, z^k) que satisfazem $(x, z) > 0$. E é justamente essa propriedade que origina o termo ponto interior.

Como a maioria dos algoritmos iterativos em otimização, os métodos de pontos interiores do tipo primal-dual têm dois elementos básicos: um procedimento para determinar

a direção que devemos caminhar e uma medida do quanto se deseja andar nesta direção, ou seja, o tamanho do passo a se considerar. Vamos aplicar o método de Newton em (2.5) para descobrirmos a direção (dx, dy, dz) que devemos tomar. Temos

$$J(x, y, z) \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = -F(x, y, z),$$

ou seja,

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^t & \mathbf{I}_n \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \begin{bmatrix} b - Ax \\ c - z - A^t y \\ -XZe \end{bmatrix}. \tag{2.6}$$

Se o ponto corrente for um ponto factível, então temos $b - Ax = 0$ e $c - z - A^t y = 0$, assim, o sistema (2.6) pode ser escrito como

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^t & \mathbf{I}_n \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -XZe \end{bmatrix}.$$

Um passo completo nessa direção geralmente não é permitido, pois isso violaria o limite $(x, z) > 0$. Para evitar essa dificuldade, realizamos uma busca linear ao longo da direção de Newton, assim a nova iteração é $(x, y, z) + \alpha(dx, dy, dz)$, para um parâmetro $\alpha \in (0, 1]$. Normalmente podemos andar apenas uma fração de passo $(\alpha < 1)$ antes de violarmos a condição $(x, z) > 0$. Portanto, um passo de Newton completo não é feito sempre.

A maioria dos algoritmos do tipo primal-dual consideram caminhar ao longo de direções do ortante (x, z) estritamente positivo, e que de certa forma estejam longe de zero, ou seja, pontos factíveis (x, z) que sejam interiores. Esses pontos produzem um caminho conhecido como trajetória central, e que se aproxima da fronteira dessa região, formada pelas restrições do problema, apenas no limite. Para isso, fazemos que os pares (x, z) convirjam a zero na mesma razão, dessa maneira é possível dar passos ao longo da trajetória central antes de violar a condição de não negatividade.

Desta forma, introduzimos um parâmetro de centralidade $\sigma \in [0, 1]$, onde esse valor é comumente utilizado para ajustar a centralização da trajetória e uma medida μ de dualidade definida por

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i z_i = \frac{x^t z}{n},$$

onde μ mede o valor médio dos produtos dos pares $x_i z_i$.

Com esses conceitos, podemos formular uma estrutura básica do algoritmo para o método primal-dual.

Algoritmo 1: Primal-Dual

Entrada: Dados (x^0, z^0) ponto inicial estritamente factível e y^0 factível.

Saída: Solução do método primal-dual.

início

Para $k = 0, 1, 2, \cdots$

Resolva

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^t & \mathbf{I}_n \\ Z^k & 0 & X^k \end{bmatrix} \begin{bmatrix} dx^k \\ dy^k \\ dz^k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -X^k Z^k e + \sigma_k \mu_k e \end{bmatrix},$$

onde $\sigma_k \in [0, 1]$ e $\mu_k = \dfrac{(x^k)^t z^k}{n}$;

faça $(x^{k+1}, y^{k+1}, z^{k+1}) \leftarrow (x^k, y^k, z^k) + \alpha_k (dx^k, dy^k, dz^k)$;

escolha α_k tal que $(x^{k+1}, z^{k+1}) > 0$

faça $k = k + 1$;

fim

A maioria das implementações dos métodos do tipo primal-dual, são um caso particular do Algoritmo 1. Note que é exigido que o ponto inicial seja estritamente factível, ou seja, estamos exigindo que $Ax^k = b$ e $A^t y^k + z^k = c$ para qualquer iteração k, em particular para $k = 0$. Contudo, para muitos problemas, essa imposição de se encontrar um ponto inicial estritamente factível se torna uma tarefa cara de se realizar. Assim, uma pequena modificação é feita, e exigindo não mais que o ponto seja factível, mas apenas interior, ou seja, queremos apenas que $(x^0, z^0) > 0$. Desta forma, a busca na direção de Newton é dada por

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^t & \mathbf{I}_n \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \begin{bmatrix} r_p \\ r_d \\ -XZe + \sigma\mu e \end{bmatrix}, \tag{2.7}$$

onde chamamos $r_p = b - Ax$ de resíduo primal e $r_d = c - z - A^t y$ resíduo dual. Quando em (2.7) tomamos $\sigma = 0$, temos a direção de Newton puro, conhecida também por direção afim-escala.

Neste trabalho não abordaremos sobre os métodos primal-dual afim-escala e seguidor de caminho, visto que a grande maioria dos *Solvers* não utilizam esses métodos. No entanto, para estudo e verificação dos algoritmos, os mesmos podem ser encontrados nos trabalhos de (SUÑAGUA, 2014) e (CASACIO, 2015).

A seguir, falaremos sobre uma variante do método Preditor-Corretor, proposto

por Sanjay Mehrotra (MEHROTRA, 1992).

2.4 Método Preditor-Corretor

Segundo (WRIGHT, 1997), a maioria dos métodos de pontos interiores existentes para problemas de programação linear são baseados no método preditor-corretor proposto em (MEHROTRA, 1992).

O método preditor-corretor gera uma sequência de pontos infactíveis (x^k, y^k, z^k) para o qual temos $(x^k, z^k) > 0$. E a direção de busca para cada iteração consiste em três componentes descritas a seguir:

- Direção afim-escala, chamada de "preditora", calculada através da direção de Newton puro para a função $F(x, y, z)$ definida em (2.5);
- Parâmetro de centralidade σ escolhido adaptativamente;
- Direção "corretora", que tenta compensar algumas das não linearidades na direção afim-escala.

Desta forma, para uma iteração k qualquer, a direção preditora (dx_1^k, dy_1^k, dz_1^k), relativa a primeira componente acima, é encontrada resolvendo o sistema linear (2.8) a seguir,

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^t & \mathbf{I}_n \\ Z^k & 0 & X^k \end{bmatrix} \begin{bmatrix} dx_1^k \\ dy_1^k \\ dz_1^k \end{bmatrix} = \begin{bmatrix} r_p^k \\ r_d^k \\ r_a^k \end{bmatrix}, \tag{2.8}$$

isto é,

$$\begin{cases} Adx_1^k = r_p^k \\ A^t dy_1^k + dz_1^k = r_d^k \\ Z^k dx_1^k + X^k dz_1^k = r_a^k, \end{cases} \tag{2.9}$$

onde $r_p^k = b - Ax^k$, $r_d^k = c - z^k - A^t y^k$ e $r_a^k = -X^k Z^k e$ é o resíduo afim.

Considerando a direção de Newton puro, segue-se os seguintes resíduos. Para o próximo resíduo primal, temos $r_p^{k+1} = b - Ax^{k+1} = b - A(x^k + dx_1^k) = b - Ax^k - Adx_1^k = r_p^k - r_p^k = 0$, e o próximo resíduo dual é $r_d^{k+1} = c - z^{k+1} - A^t y^{k+1} = c - (z^k + dz_1^k) - A^t(y^k + dy_1^k) = c - z^k - A^t y^k - (A^t dy_1^k + dz_1^k) = r_d^k - r_d^k = 0$, podemos assim compreender que se utilizarmos a direção de Newton puro, em uma iteração, temos um ponto primal e dual factível. Entretanto, nas condições de complementaridade temos

$r_a^{k+1} = -X^{k+1}Z^{k+1}e = -(X^k + DX_1^k)(Z^k + DZ_1^k)e = -X^kZ^ke - Z^k(DX_1^k)e - X^k(DZ_1^k)e - (DX_1^k)(DZ_1^k)e = -X^kZ^ke - Z^kdx_1^k - X^kdz_1^k - (DX_1^k)(DZ_1^k)e = r_a - r_a - (DX_1^k)(DZ_1^k)e = -(DX_1^k)(DZ_1^k)e$, logo o próximo resíduo do termo não linear é $r_a^{k+1} = -(DX_1^k)(DZ_1^k)e$, onde $DX_1^k = diag(dx_1^k)$ e $DZ_1^k = diag(dz_1^k)$.

Desta forma, a direção corretora, relacionado aos outros dois itens, é dada calculando a correção no ponto $(x^{k+1}, y^{k+1}, z^{k+1})$ e introduzindo a perturbação μ no termo não linear, isto é, resolvemos um segundo sistema linear com a mesma matriz de coeficientes[1], alternando apenas o lado direito de (2.8), tomando $r_p^k = 0$, $r_d^k = 0$ e substituindo r_a^k por $r_c^k = \mu e - (DX_1^k)(DZ_1^k)e$, ou seja, para encontrarmos a direção corretora (dx_2^k, dy_2^k, dz_2^k), resolvemos o sistema

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^t & \mathbf{I}_n \\ Z^k & 0 & X^k \end{bmatrix} \begin{bmatrix} dx_2^k \\ dy_2^k \\ dz_2^k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ r_c^k \end{bmatrix}. \tag{2.10}$$

Perceba que foram utilizadas algumas das componentes da direção afim-escala no cálculo da resolução do sistema (2.10).

Desta forma, a direção final (dx^k, dy^k, dz^k), pode ser obtida pela soma das direções do sistema (2.8) e (2.10), ou seja,

$$dx^k = dx_1^k + dx_2^k,\ dy^k = dy_1^k + dy_2^k \text{ e } dz^k = dz_1^k + dz_2^k.$$

Entretanto, podemos encontrar diretamente a direção final resolvendo o sistema

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^t & \mathbf{I}_n \\ Z^k & 0 & X^k \end{bmatrix} \begin{bmatrix} dx^k \\ dy^k \\ dz^k \end{bmatrix} = \begin{bmatrix} r_p^k \\ r_d^k \\ r_a^k + r_c^k \end{bmatrix}.$$

Observação 2: Se a direção afim-escala produz uma boa redução na medida μ de dualidade enquanto permanece dentro do ortante positivo definido por $(x, z) > 0$, concluímos que pouca centralização é necessária nessa iteração, portanto atribuímos um valor pequeno à σ. Se, por outro lado, pudermos mover apenas uma pequena distância ao longo da direção afim-escala antes de violarmos as restrições $(x, z) > 0$, concluímos que é necessária uma quantidade significativa de centralização, assim, tomamos σ próximo de 1. Para medirmos a eficiência da direção afim-escala, tomamos

$$\tilde{\gamma} = (x + \alpha_p dx_1)^t (z + \alpha_d dz_1); \tag{2.11}$$

$$\sigma = \left(\frac{\tilde{\gamma}}{\gamma}\right)^p, \text{ onde } p = 2 \text{ ou } 3; \tag{2.12}$$

[1] Se na expressão (2.8) tivéssemos $X^{k+1} = X^k + DX_1^k$ e $Z^{k+1} = Z^k + DZ_1^k$ na matriz de coeficientes, este seria o passo de Newton da próxima iteração, dado que tomamos o passo completo na iteração atual, entretanto, aproximamos X^{k+1} e Z^{k+1} por X^k e Z^k.

$$\mu = \sigma\left(\frac{\gamma}{n}\right). \tag{2.13}$$

Os tamanhos de passos em (2.11) que utilizamos para que o próximo iterando $(x^{k+1}, z^{k+1}) = (x^k + \alpha_p dx^k, z^k + \alpha_d dz^k)$, também precisa ser interior, isto é, $(x^{k+1}, z^{k+1}) > 0$ é dado por

$$\alpha_p = \min\left\{1, \tau\min\left\{-\frac{x_j}{dx_j} \ / dx_j < 0\right\}\right\} \tag{2.14}$$

e

$$\alpha_d = \min\left\{1, \tau\min\left\{-\frac{z_j}{dz_j} \ / dz_j < 0\right\}\right\}, \tag{2.15}$$

onde consideramos $\tau = 0.99995$, veja (LUSTIG; MARSTEN; SHANNO, 1994). Na prática utilizamos tamanhos de passos diferentes, ou seja, $\alpha_p \neq \alpha_d$.

Perceba ainda que no método preditor-corretor, é preciso resolver dois sistemas lineares a cada iteração, que partilham a mesma matriz de coeficientes, porém a resolução do segundo sistema torna-se relativamente fácil, já que podemos aproveitar a fatoração obtida no primeiro sistema, se estiver sendo usado métodos diretos.

É bom salientarmos ainda que o método continua sendo base para muitas implementações de métodos de pontos interiores, por exemplo, (OLIVEIRA, 1997), (OLIVEIRA; SORENSEN, 2005), (BOCANEGRA, 2005) e (SUÑAGUA, 2014).

Ponto inicial

O ponto inicial (x^0, y^0, z^0) no método preditor-corretor não necessita ser factível, apenas que seja interior, isto é, $(x^0, z^0) > 0$. Para isto, tomamos o mesmo como segue

$$\begin{aligned}
\tilde{x} &= A^t(AA^t)^{-1}b, \\
\epsilon_1 &= \max\left\{\text{-min } \{\tilde{x}_j\}; \frac{\|b\|_1}{\epsilon_2\|A\|_1}; \epsilon_2\right\}, \text{ onde } \epsilon_2 = 100, \\
x_j^0 &= \max\{\tilde{x_j}, \epsilon_1\}, \\
y^0 &= 0,
\end{aligned}$$

$$z_j^0 = \begin{cases} c_j + \epsilon_3, & c_j \geqslant 0 \\ -c_j, & c_j \leqslant -\epsilon_3 \\ \epsilon_3, & -\epsilon_3 \leqslant c_j \leqslant 0, \text{ onde } \epsilon_3 = \|c\|_1 + 1. \end{cases}$$

O ponto inicial é semelhante ao dado em (LUSTIG; MARSTEN; SHANNO, 1992). Desta forma temos que o ponto satisfaz $(x^0, z^0) > 0$ e o par (x_j^0, z_j^0) não é pequeno, ou seja, o mesmo é iniciado com valores relativamente grandes.

Critério de Parada

Os métodos de pontos interiores terminam quando as condições de otimalidade descritas em (2.3) são satisfeitas. Essas condições podem ser medidas calculando-se a factibilidade primal, factibilidade dual e o *gap* do par primal-dual. Assim temos

$$\frac{\|r_p^k\|_2}{1+\|b\|_2} = \frac{\|b - Ax^k\|_2}{1+\|b\|_2} \leqslant \epsilon \tag{2.16}$$

$$\frac{\|r_d^k\|_2}{1+\|c\|_2} = \frac{\|c - A^t y^k - z^k\|_2}{1+\|c\|_2} \leqslant \epsilon \tag{2.17}$$

$$\frac{|c^t x^k - b^t y^k|}{1+|c^t x^k|} \leqslant \epsilon, \tag{2.18}$$

onde $\epsilon > 0$, é uma medida de erro aceitável, como por exemplo, $\epsilon = 10^{-8}$, veja (WRIGHT, 1997).

2.5 Sistema Aumentado *versus* Equações Normais

Por questão de simplicidade e a fim de facilitar a leitura do texto, o índice k referente a iteração será suprimido nesta parte do presente trabalho.

Vejamos como podemos calcular a direção de Newton (dx_1, dy_1, dz_1). O sistema (2.9) a ser resolvido a cada iteração dos métodos de pontos interiores é escrito como

$$\begin{cases} Adx_1 = r_p \\ A^t dy_1 + dz_1 = r_d \\ Zdx_1 + Xdz_1 = r_a, \end{cases} \tag{2.19}$$

onde $r_p = b - Ax$, $r_d = c - z - A^t y$ e $r_a = -XZe$. Para (x, z) pontos estritamente positivos, as matrizes $X = diag(x)$ e $Z = diag(z)$ são invertíveis, assim, podemos isolar dz_1 da terceira equação de (2.19), isto é, $dz_1 = X^{-1}(r_a - Zdx_1)$, e substituindo esta expressão na segunda equação, e chamando $D^{-1} = X^{-1}Z$ temos

$$\begin{cases} Adx_1 = r_p \\ -D^{-1}dx_1 + A^t dy_1 = r_d - X^{-1}r_a, \end{cases}$$

ou seja

$$\begin{bmatrix} -D^{-1} & A^t \\ A & 0 \end{bmatrix} \begin{bmatrix} dx_1 \\ dy_1 \end{bmatrix} = \begin{bmatrix} r_d - X^{-1}r_a \\ r_p \end{bmatrix}. \tag{2.20}$$

Ao sistema (2.20) damos o nome de sistema aumentado. Como a matriz D é invertível, podemos multiplicar a primeira equação do sistema aumentado por AD.

Somando-se essa a segunda equação de (2.20), podemos eliminar a variável dx_1, chegando assim nas equações normais, a saber,

$$ADA^t dy_1 = r_p + AD(r_d - X^{-1}r_a), \tag{2.21}$$

onde as variáveis dx_1 e dz_1 são dadas por

$$dx_1 = D(A^t dy_1 - r_d + X^{-1}r_a), \tag{2.22}$$

$$dz_1 = X^{-1}(r_a - Zdx_1). \tag{2.23}$$

A matriz A tem posto completo, logo a matriz ADA^t é simétrica e definida positiva. A equação (2.21) pode ser resolvida através, por exemplo, da fatoração de Cholesky ou pelo método iterativo dos gradientes conjugados, a escolha de um ou outro método depende da dimensão do problema e/ou esparsidade da matriz de restrições.

Perceba ainda, que a presença de apenas uma coluna densa na matriz A já é suficiente para fazer com que a matriz ADA^t seja muito mais densa, de fato, isto pode ser comprovado escrevendo as equações normais como soma de produtos externos de colunas da matriz A com pesos $\frac{x_i}{z_i}$, ou seja,

$$ADA^t = \sum_{i=1}^{n} \frac{x_i}{z_i} A_{*i} A_{*i}^t,$$

onde A_{*i}, denota a coluna i da matriz A.

Em (OLIVEIRA, 1997), foi mostrado que a matriz ADA^t se torna muito mal condicionada nas iterações finais dos métodos de pontos interiores, isto é, perto de uma solução, assim é difícil o uso de métodos iterativos, como gradientes conjugados. Neste caso, quando a matriz é mal condicionada, uma alternativa para resolver esse problema consiste em precondicionar o sistema, de tal modo a encontrar um outro sistema equivalente, porém este possuindo melhores características, como por exemplo, um número de condição melhor do que o sistema original. Vale a pena citar que, normalmente, o fator de Cholesky da matriz ADA^t possui muito mais entradas não nulas[2] do que a matriz não fatorada, isto é, ao fatorar a matriz perde-se esparsidade, veja (OLIVEIRA; SORENSEN, 2005), o que é uma característica não desejada.

Como citado em (OLIVEIRA, 1997) uma boa justificativa para escolha do sistema aumentado e não as equações normais, é que as equações normais possuem a desvantagem de mudar a esparsidade do problema original, o que não acontece quando o sistema aumentado é usado, ou seja, podemos tirar proveito da estrutura do sistema aumentado, ao contrário das equações normais. Outra desvantagem é que normalmente o sistema de equações normais (2.21) é mais mal condicionado que o sistema aumentado

[2] Quando isso ocorre, dizemos que houve fill-in "preenchimento" do fator de Cholesky.

(2.20). Não obstante as equações normais sendo importantes, pois as mesmas são simétricas, definidas positivas e reduzem a dimensão do problema, de ordem $(m+n)$ para ordem m. Apesar disso, escolhemos o sistema aumentado, e um forte argumento para isso, é devido ao lema 4.1 em (OLIVEIRA, 1997) ou lema 3.1 em (OLIVEIRA; SORENSEN, 2005), onde foi mostrado que todo precondicionador para equações normais produz um precondicionador equivalente para o sistema aumentado, mas o contrário não é verdadeiro. Além disso, optamos pelo sistema aumentado devido à maior oportunidade de encontrar um precondicionador eficaz, onde evitamos lidar com as equações normais.

De acordo com (GONDZIO, 2012), o sistema que dá origem ao sistema aumentado, por exemplo, o sistema (2.8), é quadrado de dimensão $(2n+m)$ e por causa da presença de vários blocos zero e matrizes diagonais, ele exibe um padrão de esparsidade interessante e explorável. Além disso, se um problema de otimização for de grande porte, espera-se que a própria matriz A seja esparsa.

Note também que o sistema (2.20) é simétrico e indefinido, no entanto, preserva a esparsidade do problema original quanto a matriz A. Assim, seguiremos a mesma abordagem apresentada em (OLIVEIRA, 1997) e (OLIVEIRA; SORENSEN, 2005), onde os autores precondicionaram o sistema aumentado (2.20), resultando no Precondicionador Separador.

2.6 Variáveis canalizadas

Veremos que o problema com variáveis canalizadas, conduz ao mesmo sistema aumentado (2.20), porém com ligeira mudança na matriz D.

Considere o problema:

$$\begin{aligned} \min\ f(x) = c^t x \\ \text{s.a } Ax = b \\ Ex \leqslant u \\ x \geqslant 0 \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \min\ f(x) = c^t x \\ \text{s.a } Ax = b \\ Ex + s = u \\ (x, s) \geqslant 0, \end{aligned} \tag{2.24}$$

onde u é um vetor de limitante superior das variáveis canalizadas e $E \in \mathbb{R}^{q \times n}$ é uma matriz que possui q elementos de x canalizados e é formada por q colunas da matriz identidade, correspondendo as variáveis canalizadas e as demais colunas nulas, isto é, $E = [\mathbf{I}_q\ 0]$, assim se, $E = \mathbf{I}_n$, todas as coordenadas de x são canalizadas, ou seja, possuem limite superior. Sem perda de generalidade, assumimos que as q primeiras variáveis estão canalizadas.

O problema canalizado (2.24) possui o seguinte dual

$$\begin{aligned} \max\ & g(y,w) = b^t y - u^t w \\ \text{s.a}\ & A^t y - E^t w + z = c \\ & (w,z) \geqslant 0. \end{aligned} \tag{2.25}$$

Observe que s e w são vetores de dimensão igual ao número de variáveis canalizadas. As condições de otimalidade associados aos problemas (2.24) e (2.25), para $(x,s) \geqslant 0$ e $(w,z) \geqslant 0$ são

$$\begin{aligned} Ax &= b, && \text{(Factibilidade Primal)} \\ Ex + s &= u, && \text{(Factibilidade Primal)} \\ A^t y - E^t w + z &= c, && \text{(Factibilidade Dual)} \\ XZe &= 0, && \text{(Condição de Complementaridade)} \\ SWe &= 0, && \text{(Condição de Complementaridade)} \end{aligned} \tag{2.26}$$

onde $S = diag(s)$ e $W = diag(w)$. Assim, para determinar a direção (dx, ds, dy, dz, dw), aplicamos o método de Newton nas condições de otimalidade (2.26). Logo temos

$$J(x,s,y,z,w) \begin{bmatrix} dx \\ ds \\ dy \\ dz \\ dw \end{bmatrix} = -F(x,s,y,z,w),$$

ou seja,

$$\begin{bmatrix} A & 0 & 0 & 0 & 0 \\ E & \mathbf{I}_q & 0 & 0 & 0 \\ 0 & 0 & A^t & \mathbf{I}_n & -E^t \\ Z & 0 & 0 & X & 0 \\ 0 & W & 0 & 0 & S \end{bmatrix} \begin{bmatrix} dx \\ ds \\ dy \\ dz \\ dw \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{bmatrix}, \tag{2.27}$$

onde $r_1 = b - Ax$, $r_2 = u - s - Ex$, $r_3 = c - A^t y + E^t w - z$, $r_4 = -XZe$ e $r_5 = -SWe$. Assim temos o seguinte sistema

$$\begin{cases} Adx = r_1 \\ Edx + ds = r_2 \\ A^t dy + dz - E^t dw = r_3 \\ Zdx + Xdz = r_4 \\ Wds + Sdw = r_5. \end{cases} \tag{2.28}$$

Assumindo x e s estritamente positivos, podemos isolar a quarta e a quinta equações de (2.28), $dz = X^{-1}(r_4 - Zdx)$ e $dw = S^{-1}(r_5 - Wds)$, e substituindo na terceira equação da mesma, chegando no sistema

$$\begin{cases} Adx = r_1 \\ Edx + ds = r_2 \\ A^t dy + X^{-1}r_4 - X^{-1}Zdx - E^t S^{-1}r_5 + E^t S^{-1}Wds = r_3, \end{cases} \tag{2.29}$$

isolando a segunda equação de (2.29), isto é, $ds = r_2 - Edx$, e substituindo na terceira equação, e fazendo as devidas manipulações algébricas, chegamos em

$$\begin{cases} Adx = r_1 \\ (-X^{-1}Z - E^t S^{-1}WE)dx + A^t dy = r_3 - X^{-1}r_4 + E^t S^{-1}r_5 - E^t S^{-1}Wr_2, \end{cases}$$

que pode ser escrito ainda como

$$\begin{bmatrix} -D^{-1} & A^t \\ A & 0 \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} r_6 \\ r_1 \end{bmatrix}, \tag{2.30}$$

onde $D = (X^{-1}Z + E^t S^{-1}WE)^{-1}$ e $r_6 = r_3 - X^{-1}r_4 + E^t S^{-1}r_5 - E^t S^{-1}Wr_2$.

Veja que o sistema (2.30) é semelhante a (2.20). Para esse caso, temos as equações normais que são

$$ADA^t dy = r_1 + ADr_6.$$

Já as demais direções são dadas por

$$\begin{aligned} dx &= D(A^t dy - r_6), \\ ds &= r_2 - Edx, \\ dz &= X^{-1}(r_4 - Zdx), \\ dw &= S^{-1}(r_5 - Wds). \end{aligned}$$

Note que a matriz dos coeficientes do sistema (2.27) é altamente esparsa, com exceção apenas dos blocos das matrizes A e A^t, todos os outros blocos são matrizes nulas e/ou diagonais.

A abordagem desse trabalho, que é descrita em detalhes no Capítulo 4, pode ser implementada para o problema de variáveis canalizadas.

2.7 Precondicionamento

Seja o sistema linear

$$Ax = b, \tag{2.31}$$

onde $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ e $x \in \mathbb{R}^n$. Precondicionar o sistema (2.31), significa encontrar um outro que possui o mesmo conjunto solução. No entanto, o sistema precondicionado apresenta melhores propriedades espectrais, como por exemplo, menor número de condição ou autovalores agrupados. Segundo (SAAD, 2003), tanto a robustez quanto a eficiência de métodos iterativos podem ser melhorados com a utilização de precondicionadores. De acordo com (TREFETHEN; III, 1997), a aplicação de precondicionadores é essencial para a maioria das implementações de métodos iterativos, pois os mesmos podem melhorar drasticamente as propriedades da matriz do sistema. Assim, ao resolver o sistema precondicionado, e com a utilização de um precondicionador adequado, pode-se encontrar a solução muito mais rápido do que se tentasse resolver o sistema sem precondicionamento.

Podemos precondicionar o sistema (2.31) apenas pela direita, ou seja, multiplicamos a direita pela matriz não singular M, fazendo $Ax = b \Leftrightarrow M^{-1}Ax = M^{-1}b$, onde dizemos que $M^{-1}A$ é a matriz precondicionada. Um exemplo de precondicionador dessa classe é o precondicionador escala diagonal, onde a matriz M é formada pelos elementos da diagonal da matriz A, ou seja, $M = diag(A)$, o precondicionador assim obtido é simples e fácil de ser implementado, e pode inclusive obter bons resultados para certos problemas, veja (TREFETHEN; III, 1997), página 316. Note que a matriz precondicionada desta forma possui diagonal unitária.

Podemos também precondicionar o sistema apenas pela esquerda, ou seja, fazemos $Ax = b \Leftrightarrow AM^{-1}Mx = b$, e chamando $\tilde{x} = Mx$, temos $AM^{-1}\tilde{x} = b$, onde aqui AM^{-1} é a matriz precondicionada.

Porém, podemos precondicionar o sistema tanto à direita quanto à esquerda, fazendo $Ax = b \Leftrightarrow M^{-1}Ax = M^{-1}b \Leftrightarrow M^{-1}AN^{-1}Nx = M^{-1}b$, e chamando $\tilde{x} = Nx$ e $\tilde{b} = M^{-1}b$, temos $M^{-1}AN^{-1}\tilde{x} = \tilde{b}$, onde M e N são matrizes não singulares. E se quisermos que a matriz precondicionada seja simétrica, caso A seja simétrica, tomamos $N^{-1} = M^{-t}$, assim temos $M^{-1}AM^{-t}\tilde{x} = \tilde{b}$, onde $M^{-1}AM^{-t}$ é a matriz precondicionada.

Um excelente precondicionador, que pertence a essa classe, é o fator de Cholesky de uma matriz. De fato, assuma que a matriz A é simétrica e definida positiva, logo possui decomposição de Cholesky, isto é, $A = MM^t$. Assim temos

$$M^{-1}AM^{-t} = M^{-1}MM^tM^{-t} = \mathbf{I}_m.$$

Nesse caso, o número de condição da matriz precondicionada $M^{-1}AM^{-t}$ é o mesmo da matriz identidade, ou seja, a matriz precondicionada é extremamente bem condicionada. No entanto, nem sempre é fácil encontrar o fator de Cholesky de uma dada matriz, sobretudo para matrizes de grande porte e que não sejam esparsas, que nesse caso seria usado diretamente para resolver o sistema linear. Assim, faz-se necessário tentarmos encontrar outros precondicionadores, que não sejam computacionalmente caros de se calcular, mas que tenham uma boa eficiência.

Os precondicionadores podem ser utilizados para resolver sistemas lineares de grande porte mediante métodos iterativos que normalmente não envolvem calcular explicitamente o produto de matrizes, ou seja, não é calculado de fato $(M^{-1}AM^{-t})v$, mas sim $(M^{-1}(A(M^{-t}v)))$, isto é, calcula-se apenas produto matriz-vetor.

2.8 Método dos Gradientes Conjugados

Discorreremos agora brevemente sobre o método dos gradientes conjugados. Para mais detalhes veja por exemplo (SHEWCHUK, 1994) ou (NOCEDAL; WRIGHT, 2006).

O método dos gradientes conjugados pode ser visto como um método de se minimizar irrestritamente a função quadrática estritamente convexa dada em (2.32). Chamamos de função quadrática estritamente convexa a toda função $f : \mathbb{R}^n \to \mathbb{R}$ escrita na forma

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c, \tag{2.32}$$

com $A \in \mathbb{R}^{n\times n}$, simétrica e definida positiva, $b \in \mathbb{R}^n$ e $c \in \mathbb{R}$.

Dessa forma, derivando (2.32) e igualando a zero, temos:

$$\nabla f(x) = Ax - b = 0 \Rightarrow Ax = b. \tag{2.33}$$

Assim, pela teoria de otimização convexa, esse problema possui um único minimizador, que é global, logo resolvermos o sistema linear $Ax = b$ é equivalente a minimizarmos a função dada em (2.32). Desta forma, o método sugere que dada uma aproximação inicial x_0 para a solução de (2.33), constrói-se um conjunto de direções A-conjugadas, onde a primeira direção é a de Máxima Descida, ou seja, oposta do gradiente, e todas as outras são A-conjugadas entre si. Dizemos que dois vetores não nulos d_i e d_j são A-conjugados, se $d_i^t A d_j = 0$, para $i \neq j$. Desse modo, em no máximo a quantidade de autovalores distintos da matriz A, é possível atingir o minimizador global x^* de (2.32), conforme (NOCEDAL; WRIGHT, 2006), página 115. No Algoritmo 2 resumimos o método

dos gradientes conjugados conforme (NOCEDAL; WRIGHT, 2006).

Algoritmo 2: Gradientes Conjugados

Entrada: Dado x_0 ponto inicial.

Saída: Solução do sistema $Ax = b$.

início

  Tome $r_0 = Ax_0 - b$, $p_0 = -r_0$, $k = 0$;

  enquanto $r_k \neq 0$. **faça**

    $\alpha_k = \dfrac{r_k^t r_k}{p_k^t A p_k}$; (Tamanho do passo)

    $x_{k+1} = x_k + \alpha_k p_k$; (Próximo ponto)

    $r_{k+1} = r_k + \alpha_k A p_k$; (Próximo resíduo)

    $\beta_{k+1} = \dfrac{r_{k+1}^t r_{k+1}}{r_k^t r_k}$; (Responsável para que as direções sejam A-conjugadas)

    $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$; (Conjunto de direções A-conjugadas)

    $k = k + 1$; (Atualização)

  fim

fim

Note que no método dos gradientes conjugados, Algoritmo 2, o maior esforço computacional é o cálculo de produto matriz-vetor.

Para completar as informações e contextualizar a convergência do método dos gradientes conjugados, o número de condição da matriz A, considerando a norma-2, denotado por $cond_2(A)$ ou $\kappa(A)$ é dado por $\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \dfrac{\lambda_{max}}{\lambda_{min}}$.

Desta forma, uma estimativa para se encontrar a solução do sistema via método dos gradientes conjugados é dada por

$$\|x_k - x^*\|_A \leqslant 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \|x_0 - x^*\|_A, \tag{2.34}$$

onde $\|x\|_A = \sqrt{x^t A x}$ é a norma ponderada. A expressão (2.34) nos diz que quanto mais bem condicionada for a matriz A, menor será o número de iterações efetuadas para se encontrar a solução x^*. Por outro lado, se a matriz não for bem condicionada, então teremos que efetuar muitas iterações para encontrarmos a solução. Por exemplo, se tivermos $\kappa(A) = 1$, o método dos gradientes conjugados converge em uma iteração. Isto mostra que é extremamente relevante a procura de bons precondicionadores que venham a diminuir o número de condição do sistema precondicionado.

Apresentamos a seguir o método dos gradientes conjugados precondicionado, Algoritmo 3.

Se tomarmos o precondicionador como sendo a matriz identidade, isto é, $M = \mathbf{I}_m$, temos que o Algoritmo 3 é o mesmo que o Algoritmo 2. E conforme (NOCEDAL;

Algoritmo 3: Gradientes Conjugados Precondicionado

Entrada: Dado um ponto inicial x_0 e um precondicionador M.
Saída: Solução do sistema precondicionado.

início
  Tome $r_0 = Ax_0 - b$;
  Resolva $My_0 = r_0$ para y_0;
  Tome $p_0 = -y_0,\ k = 0$;
  enquanto $r_k \neq 0.$ **faça**
    $\alpha_k = \dfrac{r_k^t y_k}{p_k^t A p_k}$; (Tamanho do passo)
    $x_{k+1} = x_k + \alpha_k p_k$; (Próximo ponto)
    $r_{k+1} = r_k + \alpha_k A p_k$; (Próximo resíduo)
    Resolva $My_{k+1} = r_{k+1}$;
    $\beta_{k+1} = \dfrac{r_{k+1}^t y_{k+1}}{r_k^t y_k}$; (Responsável para que as direções sejam A-conjugadas)
    $p_{k+1} = -y_{k+1} + \beta_{k+1} p_k$; (Conjunto de direções A-conjugadas)
    $k = k + 1$; (Atualização)
  fim
fim

WRIGHT, 2006) em termos de esforço computacional, a principal diferença entre os dois algoritmos está na resolução de um sistema linear do tipo $My = r$ no Algoritmo 3.

Observação 3: De acordo com (NOCEDAL; WRIGHT, 2006), no Algoritmo 3 foi utilizado o precondicionador simétrico e definido positivo $M = C^t C$ para o sistema precondicionado $(C^{-t}AC^{-1})\hat{x} = C^{-t}b$, onde $\hat{x} = Cx$. Note ainda que poderíamos ter aplicado diretamente o Algoritmo 2 ao sistema $(C^{-t}AC^{-1})\hat{x} = C^{-t}b$ e resolver o sistema adicional $x = C^{-1}\hat{x}$ para encontrarmos a solução desejada x.

A grande dificuldade de se encontrar um bom precondicionador é que desejamos que o mesmo seja fácil de construir e de aplicar e ao mesmo tempo acelere a convergência do sistema, ou seja, queremos um precondicionador que seja aplicável e computacionalmente barato de construir.

Conforme (MEYER, 2000), construir bons precondicionadores e provar que eles realmente funcionam, continua sendo uma área ativa de pesquisa no campo da análise numérica.

2.9 Programação Quadrática Convexa

Abordaremos agora a programação quadrática convexa, pois como veremos, a matriz do sistema aumentado não difere quanto a sua estrutura. Deste modo, podemos também aplicar a nossa proposta, que será vista no Capítulo 4, a esta importante classe de problemas. Uma apresentação deste tópico pode ser encontrada em (GONDZIO, 2012).

Considere o seguinte problema

$$\begin{aligned} \min\ f(x) &= c^t x + \frac{1}{2} x^t Q x \\ \text{s.a } Ax &= b \\ x &\geqslant 0, \end{aligned} \tag{2.35}$$

em que $A \in \mathbb{R}^{m\times n}$ é uma matriz de posto completo, $m \leqslant n$, $Q \in \mathbb{R}^{n\times n}$ é uma matriz semidefinida positiva, $x, c \in \mathbb{R}^n$ e $b \in \mathbb{R}^m$.

Associado ao problema (2.35), chamado de problema primal, temos o seguinte dual

$$\begin{aligned} \max\ g(y, x, z) &= b^t y - \frac{1}{2} x^t Q x \\ \text{s.a } A^t y + z - Qx &= c \\ y \text{ livre}, (x, z) &\geqslant 0, \end{aligned} \tag{2.36}$$

onde $y \in \mathbb{R}^m$ é o vetor de variáveis dual livre e $z \in \mathbb{R}^n$ é o vetor de variáveis de folga dual.

Note que se tivermos $Q = 0$ em (2.35) e (2.36) recaímos no par primal-dual (2.1)-(2.2).

Para não tratarmos com as restrições de desigualdades em (2.35) fazemos uso da função barreira logarítmica, dada por

$$\begin{aligned} \min\ f(x) &= c^t x + \frac{1}{2} x^t Q x - \mu \sum_{j=1}^{n} ln(x_j) \\ \text{s.a } Ax &= b. \end{aligned} \tag{2.37}$$

Tomemos a função lagrangeana em (2.37) que é

$$L(x, y) = c^t x + \frac{1}{2} x^t Q x - \mu \sum_{j=1}^{n} ln(x_j) + y^t (b - Ax).$$

Uma condição necessária para um ponto ótimo é termos $\nabla L(x, y) = 0$. Assim segue

$$\nabla_x L(x, y) = c + Qx - \mu X^{-1} e - A^t y = 0 \text{ e } \nabla_y L(x, y) = b - Ax = 0.$$

Seja $z = \mu X^{-1} e \Rightarrow XZe = \mu e$, logo para $(x, z) \geqslant 0$, temos as seguintes condições de otimalidade para programação quadrática convexa

$$\begin{aligned} Ax &= b, \quad \text{(Factibilidade Primal)} \\ A^t y + z - Qx &= c, \quad \text{(Factibilidade Dual)} \\ XZe &= \mu e. \quad \text{(Condição de Complementaridade)} \end{aligned} \tag{2.38}$$

Aplicando o método de Newton em (2.38), temos

$$\begin{bmatrix} A & 0 & 0 \\ -Q & A^t & \mathbf{I}_n \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} dx_1 \\ dy_1 \\ dz_1 \end{bmatrix} = \begin{bmatrix} r_p \\ r_d \\ r_c \end{bmatrix},$$

onde $r_p = b - Ax$, $r_d = c + Qx - A^t y - z$ e $r_c = \mu e - XZe$.

Fazendo algumas eliminações de variáveis, temos o seguinte sistema aumentado para programação quadrática convexa.

$$\begin{bmatrix} -D^{-1} & A^t \\ A & 0 \end{bmatrix} \begin{bmatrix} dx_1 \\ dy_1 \end{bmatrix} = \begin{bmatrix} r_d - X^{-1} r_c \\ r_p \end{bmatrix}, \tag{2.39}$$

onde em (2.39) temos $D^{-1} = Q + X^{-1}Z$. Quando Q for uma matriz diagonal teremos então um problema com variáveis separáveis. E o sistema de equações normais resultante demanda o mesmo esforço computacional que em problemas lineares.

No próximo capítulo trataremos mais especificamente três importantes tipos de precondicionadores usados iterativamente em métodos de pontos interiores.

3 Precondicionadores

3.1 Fatoração controlada de Cholesky

Nesta Seção trataremos da Fatoração Controlada de Cholesky (FCC) (CAMPOS, 1995), a qual é um tipo de fatoração incompleta de Cholesky.

Uma boa apresentação sobre a fatoração controlada de Cholesky pode ser encontrada, por exemplo, nos trabalhos (BOCANEGRA, 2005) e (CASACIO, 2015).

Fatorações incompletas podem ser utilizadas como precondicionadores. Sabemos que dada uma matriz A simétrica e definida positiva, existe uma matriz triangular inferior L com elementos na diagonal positivos, de tal forma que temos a decomposição $A = LL^t$, onde, L é chamado de fator de Cholesky, e dizemos que temos uma fatoração completa. Observe que neste caso, quando utilizamos o fator L, temos a matriz precondicionada mais bem condicionada possível. Porém, normalmente o fator de Cholesky obtido pela fatoração, possui mais elementos não nulos do que a matriz original A, ou seja, houve preenchimento (fill-in) no fator produzido. Entretanto, pelo custo de processamento e de memória, isso é uma característica não desejada. Uma alternativa para se evitar esse preenchimento é aplicar não o fator completo da matriz fatorada, mas sim, com um fator incompleto, denotado por $\tilde{L}$, ou seja, com uma aproximação, e com esse fator incompleto, temos um certo controle sobre o preenchimento da matriz.

Campos propôs a chamada fatoração controlada de Cholesky (CAMPOS, 1995) e (CAMPOS; BIRKETT, 1998), onde pode-se permitir que o fator tenha mais ou menos elementos não nulos do que a matriz original, isto de acordo com a quantidade de memória disponível.

Seja L o fator de Cholesky da matriz A quando a fatoração for completa e $\tilde{L}$ quando a fatoração for incompleta. Temos:

$$A = LL^t = \tilde{L}\tilde{L}^t + R,$$

onde R é a matriz resto.

Usando o fator incompleto $\tilde{L}$ como precondicionador para a matriz A, temos

$$\tilde{L}^{-1}A\tilde{L}^{-t} = \tilde{L}^{-1}(LL^t)\tilde{L}^{-t} = (\tilde{L}^{-1}L)(\tilde{L}^{-1}L)^t. \tag{3.1}$$

Definindo $E = L - \tilde{L}$, e substituindo em (3.1), temos

$$\tilde{L}^{-1}A\tilde{L}^{-t} = (\mathbf{I} + \tilde{L}^{-1}E)(\mathbf{I} + \tilde{L}^{-1}E)^t.$$

Desse modo, quando $\tilde{L} \to L \Rightarrow E \to 0$ e $\tilde{L}^{-1}A\tilde{L}^{-t} \to \mathbf{I}$.

Note que a matriz resto é dada por

$$\begin{aligned} R &= LL^t - \tilde{L}\tilde{L}^t \\ &= LL^t - L\tilde{L}^t + L\tilde{L}^t - \tilde{L}\tilde{L}^t \\ &= L(L^t - \tilde{L}^t) + (L - \tilde{L})\tilde{L}^t \\ &= LE^t + E\tilde{L}^t. \end{aligned}$$

Portanto, quando $E \to 0 \Rightarrow R \to 0$.

Desse modo a fatoração controlada de Cholesky é baseada na minimização da norma de Frobenius da matriz E.

$$min\|E\|_F^2 = \sum_{j=1}^{m} c_j,$$

com $c_j = \sum_{i=1}^{m} |l_{ij} - \tilde{l}_{ij}|^2$, onde m é a ordem da matriz quadrada A.

Sem perda de generalidade, assuma que os elementos da coluna j estão ordenados, em módulo, por ordem decrescente.

Separando c_j em dois somatórios temos

$$c_j = \sum_{i=1}^{t_j+\eta} |l_{ij} - \tilde{l}_{ij}|^2 + \sum_{i=t_j+\eta+1}^{m} |l_{ij}|^2, \tag{3.2}$$

onde t_j é o número de elementos não negativos abaixo da diagonal da j-ésima coluna de A e η é o número extra de elementos permitidos por coluna.

Note que no primeiro somatório de (3.2) estão os $t_j + \eta$ elementos não nulos da j-ésima coluna de $\tilde{L}$. O segundo, contém apenas os elementos do fator completo de L que não estão em $\tilde{L}$. Baseado nestes fatos, Campos propôs a seguinte heurística para minimizar E:

• Aumentar o valor de η, desta forma se permite um maior preenchimento. Portanto c_j decresce, consequentemente $\|E\|_F$ também, pois o primeiro somatório contém mais termos que o segundo. Uma regra que se usa para o incremento de η é aumentá-lo em 10 unidades (BOCANEGRA, 2005).

• Para um η fixo, escolher os maiores $t_j + \eta$ elementos de $\tilde{L}$, em valor absoluto. Desta forma os maiores elementos estão no primeiro somatório e os menores estão no segundo, produzindo um fator $\tilde{L}$ ótimo para uma determinada quantidade de armazenamento.

Como citado em (BOCANEGRA, 2005), a fatoração controlada de Cholesky, não considera o padrão de esparsidade da matriz A, assim, os elementos que são armazenados no precondicionador podem ou não estar nas posições correspondentes aos da matriz A, pois a escolha é baseada levando em consideração o valor absoluto e não a posição do elemento.

Em consonância com (BOCANEGRA; CAMPOS; OLIVEIRA, 2007) temos que na (FCC), se $\eta = -m$, temos o Precondicionador Escala Diagonal. Caso $\eta = m$, temos a fatoração completa de Cholesky, onde η pertence ao intervalo $[-m, m]$. Assim, se $\eta > 0$ temos preenchimento (fill-in) e caso $\eta < 0$ temos retirada de elementos (drop-out). Caso $\eta = 0$ o armazenamento é exatamente o mesmo que da matriz A. Desta forma, podemos ter um armazenamento de memória previsível. Assim, o precondicionador é muito versátil, pois dependendo da quantidade de memória disponível podemos fazer com que o precondicionador seja mais ou menos eficiente.

Perceba que o parâmetro de incremento η é responsável por permitir um maior preenchimento na matriz precondicionadora, desta forma, dependendo do problema, é possível permitir uma maior quantidade de elementos não nulos por coluna, isto se for possível por causa de limitação de memória computacional. Tipicamente, quando η aumenta, o número de iterações do método dos gradientes conjugados diminui, já que o precondicionador é melhorado. Por outro lado, o tempo necessário para construir tal precondicionador é maior (BOCANEGRA, 2005).

Deve ser mencionado que fatorações incompletas podem ter a desvantagem de possuir pivôs muito pequenos ou negativos. Desta forma, para tentar contornar essa característica, na (FCC) foi utilizado um incremento exponencial na diagonal da matriz $\tilde{L}$.

Um exemplo de aplicação da (FCC) é o trabalho de (SILVA L. M. DA; OLIVEIRA, 2021), onde foram apresentadas duas novas propostas de modificações na fatoração controlada de Cholesky, para reduzir o tempo computacional necessário para a resolução de problemas de programação linear. Tais propostas basearam-se no fato de que a (FCC) é suscetível à falhas durante o processo de fatoração, por causa de elementos não positivos e/ou muito pequenos na diagonal. Desta forma, quando ocorre uma falha na diagonal da matriz das equações normais, é incrementado um número α e a fatoração da nova matriz é reiniciada. As modificações propostas consideram duas situações: Uma buscou reduzir o número de vezes que a (FCC) pode ser reiniciada até concluir a construção do precondicionador, já a outra considerou evitar tais reinícios.

3.2 Precondicionador Separador

O Precondicionador Separador foi proposto por (OLIVEIRA, 1997) como um precondicionador para o sistema aumentado (2.20). O principal objetivo dele é evitar a

presença das equações normais $AD^{-1}A^t$. Para essa finalidade, foi empregado o precondicionador geral dado em blocos

$$M^{-1} = \begin{bmatrix} F & G \\ H & J \end{bmatrix},$$

e precondicionando a matriz dos coeficientes[3]

$$S = \begin{bmatrix} -D & A^t \\ A & 0 \end{bmatrix},$$

temos,

$$M^{-1}SM^{-t} = \begin{bmatrix} -FDF^t + FA^tG^t + GAF^t & -FDH^t + FA^tJ^t + GAH^t \\ -HDF^t + HA^tG^t + JAF^t & -HDH^t + HA^tJ^t + JAH^t \end{bmatrix}.$$

Tomando $J = 0$ evita-se a presença das equações normais, ou seja, as mesmas não aparecem na matriz precondicionada. Assim segue

$$M^{-1}SM^{-t} = \begin{bmatrix} -FDF^t + FA^tG^t + GAF^t & -FDH^t + GAH^t \\ -HDF^t + HA^tG^t & -HDH^t \end{bmatrix}.$$

Seja agora, $H = [\mathbf{I}_m \quad 0]P$, com $P \in \mathbb{R}^{n\times n}$ uma matriz de permutação tal que HA^t seja não singular, assim, o bloco $-HDH^t$ é equivalente a uma matriz diagonal denotada por $-D_B$, sendo

$$PDP^t = \begin{bmatrix} D_B & 0 \\ 0 & D_N \end{bmatrix}.$$

Para que a matriz precondicionada se torne bloco diagonal, tome $G^t = (HA^t)^{-1}HDF^t$, ficando como

$$M^{-1}SM^{-t} = \begin{bmatrix} -FDF^t + FA^tG^t + GAF^t & 0 \\ 0 & -D_B \end{bmatrix}.$$

Por fim, tome $F = D^{-\frac{1}{2}}$.

Deste modo a matriz precondicionada é

$$M^{-1}\begin{bmatrix} -D & A^t \\ A & 0 \end{bmatrix}M^{-t} = \begin{bmatrix} -\mathbf{I}_n + D^{-\frac{1}{2}}A^tG^t + GAD^{-\frac{1}{2}} & 0 \\ 0 & -D_B \end{bmatrix}, \tag{3.3}$$

onde

$$M^{-1} = \begin{bmatrix} D^{-\frac{1}{2}} & G \\ H & 0 \end{bmatrix},$$

com $G = H^tD_B^{\frac{1}{2}}B^{-1}$, $HP^t = [\mathbf{I}_m \quad 0]$ e $AP^t = [B \quad N]$.

[3] Aqui semelhante a (OLIVEIRA, 1997) foi tomado $D = X^{-1}Z$.

Note que podemos considerar que o Precondicionador Separador reduz a dimensão do sistema aumentado $(m+n)$ para n, já que a parte inferior do sistema (3.3) é diagonal, sendo facilmente resolvido.

Observe que o sistema (3.3) possui no primeiro bloco a matriz indefinida

$$-\mathbf{I}_n + D^{-\frac{1}{2}} A^t G^t + GAD^{-\frac{1}{2}},$$

onde, através do Complemento de Schur, a mesma pode ser reduzida às seguintes matrizes simétricas e definidas positivas

$$\mathbf{I}_m + D_B^{\frac{1}{2}} B^{-1} N D_N^{-1} N^t B^{-t} D_B^{\frac{1}{2}} \tag{3.4}$$

ou

$$\mathbf{I}_{n-m} + D_N^{-\frac{1}{2}} N^t B^{-t} D_B B^{-1} N D_N^{-\frac{1}{2}}. \tag{3.5}$$

A matriz (3.4) possui dimensão m, já a matriz (3.5) possui dimensão $(n-m)$.

Chamando $W = D_N^{-\frac{1}{2}} N^t B^{-t} D_B^{\frac{1}{2}}$, as matrizes (3.4) e (3.5) podem ser escritos como $\mathbf{I}_m + W^t W$ e $\mathbf{I}_{n-m} + WW^t$, respectivamente.

Essas matrizes possuem uma característica desejável em métodos iterativos, a de possuir todos os seus autovalores, em módulo, maiores ou iguais a 1, visto que autovalores próximos de zero podem prejudicar o desempenho de métodos iterativos na solução de sistemas lineares.

No trabalho de (CASTRO, 2017), foi mostrado que as matrizes $I_m + WW^t$ e $I_{n-m} + W^t W$ possuem o mesmo espectro, ou seja, possuem o mesmo conjunto de autovalores, não contando multiplicidade, mesmo não sendo de mesma dimensão. Além disso, foi mostrado que essas matrizes em questão possuem o mesmo número de condição, isto é, $\kappa(I_m + WW^t) = \kappa(I_{n-m} + W^t W)$.

Equivalência as equações normais

Vejamos que a matriz de (3.4) pode ser obtida através das equações normais. Seja $A = [B \;\; N]P$, com $P \in \mathbb{R}^{n\times n}$ uma matriz de permutação tal que B seja não singular, logo temos:

$$AD^{-1}A^t = [B \;\; N] P D^{-1} P^t \begin{bmatrix} B^t \\ N^t \end{bmatrix} = [B \;\; N] \begin{bmatrix} D_B^{-1} & 0 \\ 0 & D_N^{-1} \end{bmatrix} \begin{bmatrix} B^t \\ N^t \end{bmatrix} \Rightarrow$$

$$AD^{-1}A^t = BD_B^{-1}B^t + ND_N^{-1}N^t. \tag{3.6}$$

Utilizando $D_B^{\frac{1}{2}} B^{-1}$ como precondicionador em (3.6), temos

$$\begin{aligned} T &= D_B^{\frac{1}{2}} B^{-1} (AD^{-1}A^t) B^{-t} D_B^{\frac{1}{2}} = D_B^{\frac{1}{2}} B^{-1} (BD_B^{-1}B^t + ND_N^{-1}N^t) B^{-t} D_B^{\frac{1}{2}} \\ &= \mathbf{I}_m + D_B^{\frac{1}{2}} B^{-1} N D_N^{-1} N^t B^{-t} D_B^{\frac{1}{2}} = \mathbf{I}_m + W^t W. \end{aligned}$$

Observe que em $D = X^{-1}Z$, ao menos $(n - m)$ elementos são grandes, desta forma uma escolha adequada das colunas da matriz B, faz com que D_B e D_N^{-1} sejam muito pequenos próximo de uma solução do sistema linear, assim W tende a matriz nula, logo T se aproxima da matriz identidade e portanto temos que o número de condição tende a 1. Ou seja, perto de uma solução do sistema linear temos que o Precondicionador Separador torna a matriz precondicionada extremamente bem condicionada.

Explicando melhor, a matriz $D = (X^{-1})Z$ possui elementos da forma

$$d_{ii} = \frac{z_i}{x_i},$$

$\forall\ i = \{1, 2, 3, \cdots, n\}$, assim, à medida que o método de pontos interiores se aproxima de uma solução ótima, as variáveis primais e duais convergem para seus respectivos valores ótimos, de modo que o conjunto de índices $\{1, 2, 3, \cdots, n\}$, se particiona em dois distintos subconjuntos, B relacionado as variáveis básicas e N, relacionado as variáveis não-básicas.

Sabemos que próximo a uma solução do problema, os valores correspondendo a matriz não-básica tendem a zero, isto é, $X_N \Rightarrow 0$, assim pela complementaridade devemos ter $Z_N \Rightarrow \hat{z} \neq 0$ sem degenerescência, assim $D_N = (X_N^{-1})Z_N \Rightarrow \infty$ para os índices em N, consequentemente, $D_N^{-1} = (Z_N^{-1})X_N \Rightarrow 0$. Para os índices em B devemos ter $X_B \Rightarrow \hat{x} \neq 0$ e também pela complementaridade temos $Z_B \Rightarrow 0$, assim $D_B = (X_B^{-1})Z_B \Rightarrow 0$, logo a matriz $W = D_N^{-\frac{1}{2}} N^t B^{-t} D_B^{\frac{1}{2}}$ tende a matriz nula.

Com uma boa escolha de colunas de A para formar a matriz B, o Precondicionador Separador tem bom desempenho perto de uma solução do problema de programação linear, onde o sistema se torna extremamente mal condicionado.

O preço pago por se evitar o sistema de equações normais é determinar quais colunas de A devem ser escolhidas para se formar B. No entanto, a fatoração $QB = LU$ é tipicamente mais fácil de calcular do que a fatoração de Cholesky de $AD^{-1}A^t$, (BOCANEGRA; CAMPOS; OLIVEIRA, 2007).

Uma estratégia para formar a matriz B é minimizar $\|D_B^{\frac{1}{2}} B^{-1} N D_N^{-\frac{1}{2}}\|$. Este problema é difícil de se resolver, mas pode ser aproximado com uma heurística barata computacionalmente. Para isso, seleciona-se as m primeiras colunas linearmente independentes de AD^{-1}, com menor norma-1. Porém, foi mostrado em (VELAZCO; OLIVEIRA; CAMPOS, 2011) que a utilização da norma-2 produz melhores resultados que a norma-1, havendo assim uma diminuição no número de iterações no método dos gradientes conjugados na solução do sistema linear.

Uma boa característica do Precondicionador Separador é que ele pode utilizar a mesma matriz B por algumas iterações, deste modo, o tempo de solução é reduzido, já que o maior esforço computacional está em escolher as colunas de A para formar B e calcular a fatoração LU dessa matriz. Como consequência, nestas iterações o precondicionador é muito barato computacionalmente. No entanto, mesmo aproveitando a matriz B por

algumas iterações, isto não significa que o precondicionador seja idêntico, pois o mesmo depende da matriz D que muda de iteração para iteração.

No trabalho de (SUÑAGUA; OLIVEIRA, 2017), foi proposta uma nova abordagem para se encontrar a base para formar a matriz B, que é de extrema importância na utilização do Precondicionador Separador. Foi utilizada uma fatoração LU retangular da matriz transposta das restrições escalada. Mostrou-se que essa nova abordagem, na maioria da vezes, produz uma matriz mais bem condicionada que a base utilizada por (OLIVEIRA, 1997).

Em (CASACIO, 2015), a autora traz uma nova abordagem para a construção do Precondicionador Separador proposto por (OLIVEIRA, 1997). A proposta melhora as características de esparsidade dos sistemas de equações, através de um novo critério de ordenamento das colunas da matriz precondicionadora, a fim de que a mesma tenha o número de condição limitado e seja o mais esparsa possível.

Como o Precondicionador Separador foi projetado para as iterações finais do método de pontos interiores, foi utilizado o Precondicionador Escala Diagonal nas primeiras iterações do método iterativo.

E como citado em (OLIVEIRA; SORENSEN, 2005), para que o Precondicionador Separador possa ser competitivo contra abordagens diretas, temos que o mesmo deve ter uma cuidadosa implementação.

Atualmente, a abordagem iterativa híbrida é a mais eficiente que existe para resolução de problemas lineares através dos métodos de pontos interiores.

3.3 Precondicionador Híbrido

O precondicionador Híbrido (BOCANEGRA, 2005) e (BOCANEGRA; CAMPOS; OLIVEIRA, 2007), foi proposto para resolução dos sistemas de equações normais pelo método dos gradientes conjugados. O método assume a existência de duas fases na resolução do problema através dos métodos de pontos interiores. Na Fase I, o precondicionador é construído pela fatoração controlada de Cholesky, onde normalmente a matriz dos coeficientes ainda não é mal condicionada, sendo assim eficiente neste cenário. Já na Fase II o sistema utiliza o Precondicionador Separador, onde em via de regra, o sistema se torna muito mal condicionado e no qual o Precondicionador Separador apresenta melhores resultados. Deve-se ressaltar que mesmo existindo duas fases, não significa que em todos os problemas sejam sempre utilizadas as duas fases, pois pode acontecer que apenas com a utilização da Fase I o método dos gradientes conjugados seja eficiente, atingindo assim a solução do problema.

Foi mostrado que a abordagem híbrida possui desempenho superior na solução

de algumas classes de problemas quando comparada a abordagem direta. Tipicamente, esse fato ocorre quando o fator de Cholesky tem um grande número de elementos não nulos. Inclusive com essa perspectiva, foi possível determinar a solução ótima de problemas nos quais a abordagem direta não atingiu solução (BOCANEGRA, 2005).

Na abordagem inicial, proposta em (BOCANEGRA, 2005), foi utilizada a seguinte heurística para a troca do precondicionador, isto é, tem-se a mudança quando a (FCC) começa perder eficiência, tornando assim lenta a convergência pelo método dos gradientes conjugados.

A troca de fase proposta por (BOCANEGRA, 2005) é identificada se uma das seguintes condições ocorrer:

1) O *gap* inicial é reduzido por um fator de 10^6 e o número de iterações dos gradientes conjugados é maior que $m/4$, sendo m a ordem do sistema.

2) Foram feitas mais de 10 correções na diagonal para evitar falhas na (FCC) e o número de colunas do problema é inferior a 16000.

3) O número de iterações do método dos gradientes conjugados é igual a m e o parâmetro η já foi incrementado três ou mais vezes.

Assim, quando a primeira condição é atendida, isto indica que a (FCC) já não está apresentando um bom desempenho e o Precondicionador Separador poderá obter melhores resultados, já que pode estar próximo da solução do problema. A segunda situação prevê casos em que a (FCC) apresenta dificuldades em construir o precondicionador, pois a mesma utiliza um incremento exponencial, para evitar que a diagonal da matriz tenha elementos muito pequenos ou negativos. Já a terceira situação sugere que não foi possível encontrar um bom parâmetro η, deste modo não está sendo atingida a convergência pelo método dos gradientes conjugados. Quando isto acontece, foi sugerido que o parâmetro η seja incrementado em 10 unidades.

No trabalho de (VELAZCO; OLIVEIRA; CAMPOS, 2011), foi sugerido uma nova mudança para troca do Precondicionador Híbrido, pois foi mostrado que utilizando a fatoração controlada de Cholesky por um maior tempo possível, permitiu a obtenção de melhores resultados do que a abordagem original feita por (BOCANEGRA, 2005).

Deste modo, a troca do precondicionador se dá quando:

• Se o número de iterações do método dos gradientes conjugados é maior ou igual a $m/6$ então é verificado se η atingiu o maior valor. Se sim, então a troca de fase é realizada. Caso contrário, η é incrementado em 10 unidades, e o processo continua com o precondicionador obtido pela fatoração controlada de Cholesky. Note que η é utilizado apenas na primeira fase do Precondicionador Híbrido.

Observe que esta abordagem tenta utilizar o precondicionador (FCC) por um maior tempo possível, já que o mesmo é mais barato computacionalmente e requer menos memória do que o Precondicionador Separador. Entretanto, para alguns problemas essa abordagem falha, pois muito perto de uma solução a (FCC) não tem bom desempenho, conforme (VELAZCO; OLIVEIRA; CAMPOS, 2011).

No Capítulo 4, falaremos sobre uma generalização do precondicionador proposto por (OLIVEIRA, 1997), chamado de Precondicionador Separador Duplo, onde precondicionamos o sistema aumentado (2.20) duas vezes.

4 Precondicionador Separador Duplo

Dado o sistema linear $Ax = b$, onde $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ e $x \in \mathbb{R}^n$, podemos precondicionar este sistema, a fim de se obter um outro sistema com o mesmo conjunto solução, porém mais fácil de se resolver por algum método iterativo, ou seja, podemos tomar os seguintes sistemas equivalentes:

$Ax = b \Leftrightarrow M^{-1}Ax = M^{-1}b \Leftrightarrow M^{-1}AN^{-1}Nx = M^{-1}b$, e chamando $\tilde{x} = Nx$ e $\tilde{b} = M^{-1}b$, temos $M^{-1}AN^{-1}\tilde{x} = \tilde{b}$, onde M e N são matrizes não singulares. E se quisermos que a matriz precondicionadora seja simétrica, caso A seja simétrica, tomamos $N^{-1} = M^{-t}$, assim temos

$$M^{-1}AM^{-t}\tilde{x} = \tilde{b}. \tag{4.1}$$

A ideia que agora propomos, é precondicionar novamente o sistema (4.1)

$$\begin{aligned} M^{-1}AM^{-t}\tilde{x} &= \tilde{b} \\ Q^{-1}M^{-1}AM^{-t}\tilde{x} &= Q^{-1}\tilde{b} \\ Q^{-1}M^{-1}AM^{-t}R^{-1}R\tilde{x} &= Q^{-1}\tilde{b}, \end{aligned}$$

onde Q e R são matrizes não singulares, e chamando $\overline{\overline{x}} = R\tilde{x}$ e $\overline{\overline{b}} = Q^{-1}\tilde{b}$, temos

$$Q^{-1}M^{-1}AM^{-t}R^{-1}\overline{\overline{x}} = \overline{\overline{b}}.$$

E se quisermos que o precondicionador seja simétrico, se $A = A^t$, tomamos $R^{-1} = Q^{-t}$, donde temos

$$Q^{-1}M^{-1}AM^{-t}Q^{-t}\ \overline{\overline{x}} = \overline{\overline{b}}, \tag{4.2}$$

onde $\overline{\overline{x}} = Q^tM^tx$, $\overline{\overline{b}} = Q^{-1}M^{-1}b$ e $Q^{-1}M^{-1}AM^{-t}Q^{-t}$ é a matriz precondicionada.

Dessa forma, vamos aplicar o precondicionador duplo na matriz do sistema aumentando (2.20).

4.1 Construção do Precondicionador Duplo

Seja

$$M^{-1} = \begin{bmatrix} F & G \\ H & J \end{bmatrix}, \tag{4.3}$$

o primeiro precondicionador, onde $F \in \mathbb{R}^{n\times n}$, $G \in \mathbb{R}^{n\times m}$, $H \in \mathbb{R}^{m\times n}$ e $J \in \mathbb{R}^{m\times m}$. Aplicando o precondicionador (4.3) à matriz do sistema aumentado (2.20) que é

$$S = \begin{bmatrix} -D^{-1} & A^t \\ A & 0 \end{bmatrix},$$

temos

$$M^{-1}SM^{-t} = \begin{bmatrix} -FD^{-1}F^t + FA^tG^t + GAF^t & -FD^{-1}H^t + FA^tJ^t + GAH^t \\ -HD^{-1}F^t + HA^tG^t + JAF^t & -HD^{-1}H^t + HA^tJ^t + JAH^t \end{bmatrix}. \quad (4.4)$$

Considerando agora o precondicionador,

$$Q^{-1} = \begin{bmatrix} U & V \\ X & Z \end{bmatrix}, \quad (4.5)$$

onde $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{n\times m}$, $X \in \mathbb{R}^{m\times n}$ e $Z \in \mathbb{R}^{m\times m}$, e se utilizarmos (4.5) para precondicionarmos o sistema (4.4), temos

$$Q^{-1}M^{-1}SM^{-t}Q^{-t} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \quad (4.6)$$

onde S_{11}, S_{12}, S_{21} e S_{22} são dados a seguir

$$S_{11} = -UFD^{-1}F^tU^t + UFA^tG^tU^t + UGAF^tU^t - UFD^{-1}H^tV^t + UFA^tJ^tV^t + UGAH^tV^t - VHD^{-1}F^tU^t + VHA^tG^tU^t + VJAF^tU^t - VHD^{-1}H^tV^t + VHA^tJ^tV^t + VJAH^tV^t.$$

$$S_{12} = -UFD^{-1}F^tX^t + UFA^tG^tX^t + UGAF^tX^t - UFD^{-1}H^tZ^t + UFA^tJ^tZ^t + UGAH^tZ^t - VHD^{-1}F^tX^t + VHA^tG^tX^t + VJAF^tX^t - VHD^{-1}H^tZ^t + VHA^tJ^tZ^t + VJAH^tZ^t.$$

$$S_{21} = -XFD^{-1}F^tU^t + XFA^tG^tU^t + XGAF^tU^t - XFD^{-1}H^tV^t + XFA^tJ^tV^t + XGAH^tV^t - ZHD^{-1}F^tU^t + ZHA^tG^tU^t + ZJAF^tU^t - ZHD^{-1}H^tV^t + ZHA^tJ^tV^t + ZJAH^tV^t.$$

$$S_{22} = -XFD^{-1}F^tX^t + XFA^tG^tX^t + XGAF^tX^t - XFD^{-1}H^tZ^t + XFA^tJ^tZ^t + XGAH^tZ^t - ZHD^{-1}F^tX^t + ZHA^tG^tX^t + ZJAF^tX^t - ZHD^{-1}H^tZ^t + ZHA^tJ^tZ^t + ZJAH^tZ^t.$$

Observação 4: Como S é matriz simétrica, logo o sistema precondicionado (4.6) é simétrico, isto é, $S_{12} = (S_{21})^t$.

4.2 Escolha dos blocos do Sistema Precondicionado Duplamente

Para escolha dos blocos das matrizes M^{-1} e Q^{-1}, temos que ter alguns objetivos.

i) Evitar o surgimento das equações normais.

ii) Fazer com que o sistema precondicionado duplamente (4.6) seja relativamente fácil de se resolver, como por exemplo, um sistema formado por matrizes diagonais.

Assim sendo, tome

$$M^{-1} = \begin{bmatrix} F & G \\ H & 0 \end{bmatrix} \text{ e } Q^{-1} = \begin{bmatrix} U & V \\ 0 & \mathbf{I}_m \end{bmatrix},$$

dessa forma temos:

$$S_{22} = \cancelto{0}{-XFD^{-1}F^tX^t} + \cancelto{0}{XFA^tG^tX^t} + \cancelto{0}{XGAF^tX^t} - \cancelto{0}{XFD^{-1}H^tZ^t} + \cancelto{0}{XFA^tJ^tZ^t} + \cancelto{0}{XGAH^tZ^t} - \cancelto{0}{ZHD^{-1}F^tX^t} + \cancelto{0}{ZHA^tG^tX^t} + \cancelto{0}{ZJAF^tX^t} - HD^{-1}H^t + \cancelto{0}{ZHA^tJ^tZ^t} + \cancelto{0}{ZJAH^tZ^t}.$$

Portanto, $S_{22} = -HD^{-1}H^t$.

Considere $H = [\mathbf{I}_m \;\; 0]P$, onde $P \in \mathbb{R}^{n\times n}$ é uma matriz de permutação tal que AH^t seja não singular, assim, o bloco $S_{22} = -HD^{-1}H^t$ é equivalente a uma matriz diagonal denotada por $-D_B^{-1}$, sendo

$$PD^{-1}P^t = \begin{bmatrix} D_B^{-1} & 0 \\ 0 & {D_N}^{-1} \end{bmatrix}.$$

De fato, pois

$$-HD^{-1}H^t = -\begin{bmatrix} \mathbf{I}_m & 0 \end{bmatrix} PD^{-1}P^t \begin{bmatrix} \mathbf{I}_m \\ 0 \end{bmatrix} = -\begin{bmatrix} \mathbf{I}_m & 0 \end{bmatrix}\begin{bmatrix} D_B^{-1} & 0 \\ 0 & {D_N}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{I}_m \\ 0 \end{bmatrix} = -D_B^{-1}.$$

Logo $-HD^{-1}H^t = -D_B^{-1}$. Portanto $S_{22} = -D_B^{-1}$.

Seja $AP^t = [B \;\; N]$, uma partição da matriz A escolhida adequadamente, assim temos

$$H^t = P^t \begin{bmatrix} \mathbf{I}_m \\ 0 \end{bmatrix} \Rightarrow AH^t = AP^t \begin{bmatrix} \mathbf{I}_m \\ 0 \end{bmatrix} = [B \;\; N] \begin{bmatrix} \mathbf{I}_m \\ 0 \end{bmatrix} = B,$$

daí $AH^t = B$.

Para determinar os demais blocos da matriz precondicionada (4.6), vamos escolher as submatrizes de M^{-1} e Q^{-1}, de tal modo que $S_{12} = 0$. Com as escolhas de $J = 0$, $X = 0$ e $Z = \mathbf{I}_m$, temos:

$$S_{12} = -UFD^{-1}H^t + UGAH^t - VHD^{-1}H^t = -UFD^{-1}H^t + UGB - VD_B^{-1} = 0,$$

daí $UG = UFD^{-1}H^tB^{-1} + VD_B^{-1}B^{-1} = (UFD^{-1}H^t + VD_B^{-1})B^{-1} = LB^{-1}$. Portanto, se tomarmos $UG = LB^{-1}$, onde $L = UFD^{-1}H^t + VD_B^{-1}$, teremos $S_{12} = 0$, e consequentemente $S_{21} = 0$, pois o precondicionador é simétrico. Note que podemos explicitar a matriz G, visto que a matriz U deve ser de posto completo, para que Q^{-1} seja não singular, assim temos $G = U^{-1}LB^{-1}$.

Agora, voltemos nossa atenção para o bloco S_{11} de (4.6). Já com as escolhas de $J = 0$, $X = 0$ e $Z = \mathbf{I}_m$, temos

$S_{11} = -UFD^{-1}F^tU^t + UFA^tG^tU^t + UGAF^tU^t - UFD^{-1}H^tV^t + UGAH^tV^t - VHD^{-1}F^tU^t + VHA^tG^tU^t - VHD^{-1}H^tV^t$, e considerando $AH^t = B$, $HD^{-1}H^t = D_B^{-1}$ e $L = UFD^{-1}H^t + VD_B^{-1}$, chegamos em

$$S_{11} = -UFD^{-1}F^tU^t + UFA^tB^{-t}L^t + LB^{-1}AF^tU^t + VD_B^{-1}V^t. \tag{4.7}$$

Portanto, o precondicionador duplo (4.6) pode ser escrito como

$$Q^{-1}M^{-1}SM^{-t}Q^{-t} = \begin{bmatrix} S_{11} & 0 \\ 0 & -D_B^{-1} \end{bmatrix}. \tag{4.8}$$

4.3 Escolha dos blocos *F, U e V*

A fim de que apareça na diagonal do bloco S_{11} a matriz identidade ou um múltiplo seu, tome $F = \alpha D^{-\frac{1}{2}}$ e $U = \beta D$, com α e $\beta \in \mathbb{R}^*$. Assim, a matriz (4.7) é dada por

$$S_{11} = -(\alpha\beta)^2\mathbf{I}_n + \alpha\beta D^{\frac{1}{2}}A^tB^{-t}L^t + \alpha\beta LB^{-1}AD^{\frac{1}{2}} + VD_B^{-1}V^t \tag{4.9}$$

e $L = \alpha\beta D^{-\frac{1}{2}}H^t + VD_B^{-1}$.

E as matrizes precondicionadoras são

$$M^{-1} = \begin{bmatrix} \alpha D^{-\frac{1}{2}} & G \\ H & 0 \end{bmatrix} \text{ e } Q^{-1} = \begin{bmatrix} \beta D & V \\ 0 & \mathbf{I}_m \end{bmatrix},$$

com $H = [\mathbf{I}_m \;\; 0]P$ e $G = \beta^{-1}D^{-1}LB^{-1}$. E finalmente definimos

$$V = \gamma P^t \begin{bmatrix} {D_B}^{\frac{1}{2}} \\ 0 \end{bmatrix}, \text{ com } \gamma \in \mathbb{R}.$$

Note que na matriz acima, γ pode inclusive ser nulo. Perceba ainda que V está recebendo a ação da matriz de permutação.

Observação 5:

Considerando $L = \alpha\beta D^{-\frac{1}{2}}H^t + VD_B^{-1}$, $H = [\mathbf{I}_m \;\; 0]P$ e $V = \gamma P^t \begin{bmatrix} {D_B}^{\frac{1}{2}} \\ 0 \end{bmatrix}$,

então L pode ser escrito na forma

$$L = (\alpha\beta + \gamma)P^t \begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix}.$$

De fato, pois $L = \alpha\beta D^{-\frac{1}{2}}H^t + VD_B^{-1} = \alpha\beta D^{-\frac{1}{2}}P^t \begin{bmatrix} \mathbf{I}_m \\ 0 \end{bmatrix} + \gamma P^t \begin{bmatrix} {D_B}^{\frac{1}{2}} \\ 0 \end{bmatrix} D_B^{-1} \Rightarrow$

$$PL = \alpha\beta PD^{-\frac{1}{2}}P^t \begin{bmatrix} \mathbf{I}_m \\ 0 \end{bmatrix} + \gamma \begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix} \text{ daí}$$

$$PL = \alpha\beta \begin{bmatrix} {D_B}^{-\frac{1}{2}} & 0 \\ 0 & {D_N}^{-\frac{1}{2}} \end{bmatrix}\begin{bmatrix} \mathbf{I}_m \\ 0 \end{bmatrix} + \gamma \begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix} \Rightarrow PL = (\alpha\beta+\gamma)\begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix}.$$

Portanto

$$L = (\alpha\beta + \gamma)P^t \begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix}.$$

4.4 Análise da matriz do 1º Bloco

Vejamos algumas características da matriz do primeiro bloco do sistema precondicionado duplamente (4.8).

$$\text{Seja } \hat{K} = S_{11} = -(\alpha\beta)^2\mathbf{I}_n + \alpha\beta D^{\frac{1}{2}}A^tB^{-t}L^t + \alpha\beta LB^{-1}AD^{\frac{1}{2}} + VD_B^{-1}V^t. \tag{4.10}$$

Perceba que a matriz $\hat{K}$ é simétrica, pois:

$$\begin{aligned}\hat{K}^t &= (-(\alpha\beta)^2\mathbf{I}_n + \alpha\beta D^{\frac{1}{2}}A^tB^{-t}L^t + \alpha\beta LB^{-1}AD^{\frac{1}{2}} + VD_B^{-1}V^t)^t \\ &= -(\alpha\beta)^2\mathbf{I}_n + \alpha\beta LB^{-1}AD^{\frac{1}{2}} + \alpha\beta D^{\frac{1}{2}}A^tB^{-t}L^t + VD_B^{-1}V^t = \hat{K}.\end{aligned}$$

Seja agora $C = LB^{-1}AD^{\frac{1}{2}}$, então temos

$$\hat{K} = -(\alpha\beta)^2\mathbf{I}_n + \alpha\beta C^t + \alpha\beta C + VD_B^{-1}V^t \tag{4.11}$$

e considerando

$$L = (\alpha\beta+\gamma)P^t\begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix},\ V = \gamma P^t\begin{bmatrix} {D_B}^{\frac{1}{2}} \\ 0 \end{bmatrix},$$

e $A = [B \quad N]P$, então temos a seguinte relação:

$$C = LB^{-1}AD^{\frac{1}{2}} = (\alpha\beta+\gamma)P^t\begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix}B^{-1}[B \quad N]PD^{\frac{1}{2}} \Rightarrow$$

$$PCP^t = (\alpha\beta+\gamma)\begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix}[\mathbf{I}_m \quad B^{-1}N]PD^{\frac{1}{2}}P^t \Rightarrow$$

$$PCP^t = (\alpha\beta+\gamma)\begin{bmatrix} D_B^{-\frac{1}{2}} & D_B^{-\frac{1}{2}}B^{-1}N \\ 0 & 0 \end{bmatrix}\begin{bmatrix} D_B^{\frac{1}{2}} & 0 \\ 0 & {D_N}^{\frac{1}{2}} \end{bmatrix} \Rightarrow$$

$$PCP^t = (\alpha\beta+\gamma)\begin{bmatrix} \mathbf{I}_m & D_B^{-\frac{1}{2}}B^{-1}ND_N^{\frac{1}{2}} \\ 0 & 0 \end{bmatrix}.$$

Logo,

$$C = (\alpha\beta+\gamma)P^t\begin{bmatrix} \mathbf{I}_m & D_B^{-\frac{1}{2}}B^{-1}ND_N^{\frac{1}{2}} \\ 0 & 0 \end{bmatrix}P.$$

Portanto, temos

$$C = P^t\begin{bmatrix} (\alpha\beta+\gamma)\mathbf{I}_m & (\alpha\beta+\gamma)D_B^{-\frac{1}{2}}B^{-1}ND_N^{\frac{1}{2}} \\ 0 & 0 \end{bmatrix}P.$$

E assim segue que

$$C^t = P^t \begin{bmatrix} (\alpha\beta+\gamma)\mathbf{I}_m & 0 \\ (\alpha\beta+\gamma)D_N^{\frac{1}{2}}N^tB^{-t}D_B^{-\frac{1}{2}} & 0 \end{bmatrix} P.$$

Note que

$$-(\alpha\beta)^2\mathbf{I}_n = P^t \begin{bmatrix} -(\alpha\beta)^2\mathbf{I}_m & 0 \\ 0 & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} P$$

e $VD_B^{-1}V^t = \gamma P^t \begin{bmatrix} D_B{}^{\frac{1}{2}} \\ 0 \end{bmatrix} D_B^{-1}\gamma[D_B^{\frac{1}{2}} \;\; 0]P = P^t \begin{bmatrix} \gamma^2\mathbf{I}_m & 0 \\ 0 & 0 \end{bmatrix} P.$

Desta forma, podemos escrever (4.11) como

$$\hat{K} = P^t \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)D_B^{-\frac{1}{2}}B^{-1}ND_N^{\frac{1}{2}} \\ \alpha\beta(\alpha\beta+\gamma)D_N^{\frac{1}{2}}N^tB^{-t}D_B^{-\frac{1}{2}} & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} P. \tag{4.12}$$

Note que para o precondicionador funcionar, devemos ter $\gamma \neq -\alpha\beta$.

Observe, ainda, temos uma variedade de escolhas para as matrizes que compõem o Precondicionador Separador Duplo e que todas conduzem à mesma matriz (4.12). Isso ocorre devido às escolhas dos blocos que compõem o Precondicionador. Como a matriz S_{11}, definida em (4.9), envolve apenas produtos na forma UF, é possível combinar essas duas matrizes e considerar $UF = \alpha\beta D^{\frac{1}{2}}$.

Na sequência, definiremos o complemento de Schur da Matriz (4.12), veja por exemplo, (BOYD; BOYD; VANDENBERGHE, 2004).

O complemento de Schur em relação à matriz $(-(\alpha\beta)^2\mathbf{I}_{n-m})$ é dado por

$$(\alpha\beta+\gamma)^2\mathbf{I}_m + (\alpha\beta+\gamma)^2D_B^{-\frac{1}{2}}B^{-1}ND_NN^tB^{-t}D_B^{-\frac{1}{2}}. \tag{4.13}$$

Já o complemento de Schur em relação à matriz $(\alpha\beta+\gamma)^2\mathbf{I}_m$ é dado por

$$-(\alpha\beta)^2\mathbf{I}_{n-m} - (\alpha\beta)^2D_N^{\frac{1}{2}}N^tB^{-t}D_B^{-1}B^{-1}ND_N^{\frac{1}{2}}. \tag{4.14}$$

Note que podemos encontrar o complemento de Schur em (4.13) através das equações normais. De fato, seja $A = [B \;\; N]P$, com $P \in \mathbb{R}^{n\times n}$ uma matriz de permutação tal que B seja não singular, logo temos:

$$ADA^t = [B \;\; N]PDP^t \begin{bmatrix} B^t \\ N^t \end{bmatrix} = [B \;\; N] \begin{bmatrix} D_B & 0 \\ 0 & D_N \end{bmatrix} \begin{bmatrix} B^t \\ N^t \end{bmatrix} \Rightarrow$$

$$ADA^t = BD_BB^t + ND_NN^t. \tag{4.15}$$

Assim, utilizando $(\alpha\beta+\gamma)D_B^{-\frac{1}{2}}B^{-1}$ como precondicionador em (4.15), temos

$$\begin{aligned}&(\alpha\beta+\gamma)D_B^{-\frac{1}{2}}B^{-1}ADA^t(\alpha\beta+\gamma)B^{-t}D_B^{-\frac{1}{2}}\\ =\ &(\alpha\beta+\gamma)D_B^{-\frac{1}{2}}B^{-1}(BD_BB^t+ND_NN^t)(\alpha\beta+\gamma)B^{-t}D_B^{-\frac{1}{2}}\\ =\ &(\alpha\beta+\gamma)^2\mathbf{I}_m+(\alpha\beta+\gamma)^2D_B^{-\frac{1}{2}}B^{-1}ND_NN^tB^{-t}D_B^{-\frac{1}{2}}.\end{aligned}$$

Seja $W = D_B^{-\frac{1}{2}}B^{-1}ND_N^{\frac{1}{2}}$, onde $W \in \mathbb{R}^{m\times(n-m)}$, então (4.13) pode ser escrito como

$$(\alpha\beta+\gamma)^2\mathbf{I}_m+(\alpha\beta+\gamma)^2WW^t \tag{4.16}$$

e a expressão (4.14) pode ser escrita como

$$-(\alpha\beta)^2\mathbf{I}_{n-m}-(\alpha\beta)^2W^tW. \tag{4.17}$$

Note que a matriz $\hat{K}$ dada em (4.12) pode ser escrita como

$$\hat{K} = P^t\begin{bmatrix}(\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W\\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m}\end{bmatrix}P. \tag{4.18}$$

Observe que a expressão em (4.13) é equivalente a dada em (OLIVEIRA, 1997), página 56, com exceção do fator multiplicativo $(\alpha\beta+\gamma)^2$ e a forma que foi definida a matriz D, ou seja, as equações normais para os dois precondicionadores são equivalentes, porém, como mostraremos adiante, os sistemas aumentados não são equivalentes.

Os Teoremas 4.1 e 4.2 apresentados a seguir, podem nos ajudar a compreender melhor os autovalores da matriz $\hat{K}$ em (4.18).

Teorema 4.1. *A matriz $(\alpha\beta+\gamma)^2\mathbf{I}_m+(\alpha\beta+\gamma)^2WW^t$ é definida positiva e possui todos os autovalores maiores do que $(\alpha\beta+\gamma)^2$.*

Demonstração. Seja $0 \neq v \in \mathbb{R}^m$ um autovetor normalizado da matriz

$$(\alpha\beta+\gamma)^2\mathbf{I}_m+(\alpha\beta+\gamma)^2WW^t$$

associado ao autovalor λ, isto é, $((\alpha\beta+\gamma)^2\mathbf{I}_m+(\alpha\beta+\gamma)^2WW^t)v = \lambda v$. Logo, $(\alpha\beta+\gamma)^2v+(\alpha\beta+\gamma)^2WW^tv = \lambda v$ e multiplicando pelo transposto do vetor v, temos $(\alpha\beta+\gamma)^2+(\alpha\beta+\gamma)^2v^tWW^tv = \lambda$, como $(W^tv)^t(W^tv) = \|W^tv\|_2^2 > 0\ \forall\ v \neq 0$ e $(\alpha\beta+\gamma)^2 > 0$, segue que $\lambda > (\alpha\beta+\gamma)^2$. □

O Teorema 4.1 nos mostra que a matriz (4.16) possui m autovalores positivos e maiores do que $(\alpha\beta+\gamma)^2$.

Teorema 4.2. *A matriz $-(\alpha\beta)^2\mathbf{I}_{n-m}-(\alpha\beta)^2W^tW$ é definida negativa e possui todos os autovalores menores ou iguais a $-(\alpha\beta)^2$.*

Demonstração. Seja $0 \neq v \in \mathbb{R}^{n-m}$ um autovetor normalizado da matriz

$$-(\alpha\beta)^2\mathbf{I}_{n-m} - (\alpha\beta)^2 W^t W$$

associado ao autovalor λ, ou seja, $(-(\alpha\beta)^2\mathbf{I}_{n-m} - (\alpha\beta)^2 W^t W)v = \lambda v$. Logo, $-(\alpha\beta)^2 v - (\alpha\beta)^2 W^t W v = \lambda v$ e multiplicando pelo transposto do vetor v, temos $-(\alpha\beta)^2 - (\alpha\beta)^2 v^t W^t W v = \lambda$, como $(Wv)^t(Wv) = \|Wv\|_2^2 \geqslant 0 \; \forall \; v \neq 0$ e $-(\alpha\beta)^2\|Wv\|_2^2 \leqslant 0$, segue que $\lambda \leqslant -(\alpha\beta)^2$. □

Assim, o Teorema 4.2 nos mostra que a matriz (4.17) possui $(n-m)$ autovalores menores ou iguais a $-(\alpha\beta)^2$.

4.5 Aplicação do Precondicionador Separador Duplo na resolução do Sistema Aumentado

Sabemos que o sistema aumentado é dado por

$$\underbrace{\begin{bmatrix} -D^{-1} & A^t \\ A & 0 \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} dx \\ dy \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} r_1 \\ r_2 \end{bmatrix}}_{b}, \tag{4.19}$$

onde $r_1 = r_d - X^{-1}r_a$ e $r_2 = r_p$. Considere o processo[6] em uma iteração k qualquer.

Vamos agora resolver o sistema (4.19) usando o Precondicionador Separador Duplo. Note que sempre devemos ter $UF = \alpha\beta D^{\frac{1}{2}}$.

Desta forma, temos os seguintes precondicionadores

$$M^{-1} = \begin{bmatrix} F & G \\ H & 0 \end{bmatrix} \text{ e } Q^{-1} = \begin{bmatrix} U & V \\ 0 & \mathbf{I}_m \end{bmatrix}, \text{ com } V = \gamma P^t \begin{bmatrix} {D_B}^{\frac{1}{2}} \\ 0 \end{bmatrix},$$

$H = [\mathbf{I}_m \;\; 0]P$, $G = U^{-1}LB^{-1}$ e $L = (\alpha\beta + \gamma)P^t \begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix}$, aplicados em (4.19), temos

$$Q^{-1}M^{-1}SM^{-t}Q^{-t}\,\overline{\overline{x}} = \overline{\overline{b}}, \tag{4.20}$$

onde $\overline{\overline{x}} = Q^t M^t x$ e $\overline{\overline{b}} = Q^{-1}M^{-1}b$, assim segue

$$\overline{\overline{b}} = \begin{bmatrix} U & V \\ 0 & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} F & G \\ H & 0 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} UFr_1 + VHr_1 + UGr_2 \\ Hr_1 \end{bmatrix}.$$

E da expressão (4.7) e (4.8) temos

$$Q^{-1}M^{-1}SM^{-t}Q^{-t} = \begin{bmatrix} -(\alpha\beta)^2\mathbf{I}_n + \alpha\beta D^{\frac{1}{2}}A^t B^{-t}L^t + \alpha\beta LB^{-1}AD^{\frac{1}{2}} + VD_B^{-1}V^t & 0 \\ 0 & -D_B^{-1} \end{bmatrix}.$$

[6] Aqui consideramos apenas o cálculo da direção afim escala, que estamos chamando simplesmente de (dx, dy, dz).

Assim, podemos escrever (4.20) como se segue:

$$\begin{bmatrix} -(\alpha\beta)^2\mathbf{I}_n + \alpha\beta D^{\frac{1}{2}}A^tB^{-t}L^t + \alpha\beta LB^{-1}AD^{\frac{1}{2}} + VD_B^{-1}V^t & 0 \\ 0 & -D_B^{-1} \end{bmatrix} \begin{bmatrix} \overline{\overline{x_1}} \\ \overline{\overline{x_2}} \end{bmatrix} =$$

$$\begin{bmatrix} \alpha\beta D^{\frac{1}{2}}r_1 + \gamma P^t \begin{bmatrix} D_B^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} Pr_1 + (\alpha\beta+\gamma)P^t \begin{bmatrix} D_B^{-\frac{1}{2}} \\ 0 \end{bmatrix} B^{-1}r_2 \\ [\mathbf{I}_m \;\; 0]Pr_1 \end{bmatrix}, \tag{4.21}$$

onde $\overline{\overline{x}} = \begin{bmatrix} \overline{\overline{x_1}} \\ \overline{\overline{x_2}} \end{bmatrix}$, com $\overline{\overline{x_1}} \in \mathbb{R}^n$ e $\overline{\overline{x_2}} \in \mathbb{R}^m$, daí segue

$-D_B^{-1}\overline{\overline{x_2}} = [\mathbf{I}_m \;\; 0]Pr_1 \Rightarrow \overline{\overline{x_2}} = [-D_B \;\; 0]Pr_1$. Desde modo, podemos compreender que o Precondicionador Separador Duplo reduz a dimensão do sistema (4.19), que é $n+m$, para n, já que $\overline{\overline{x_2}}$ é facilmente calculado.

Note ainda que colocamos os elementos que compõem o sistema (4.21) todos em função dos componentes primordiais da matriz A e da matriz de permutação P.

Voltemos nossa atenção para a parte superior do sistema (4.21). Temos que $(-(\alpha\beta)^2\mathbf{I}_n + \alpha\beta D^{\frac{1}{2}}A^tB^{-t}L^t + \alpha\beta LB^{-1}AD^{\frac{1}{2}} + VD_B^{-1}V^t)\overline{\overline{x_1}} =$

$$\alpha\beta D^{\frac{1}{2}}r_1 + \gamma P^t \begin{bmatrix} D_B^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} Pr_1 + (\alpha\beta+\gamma)P^t \begin{bmatrix} D_B^{-\frac{1}{2}} \\ 0 \end{bmatrix} B^{-1}r_2.$$

Utilizando a expressão (4.10) e dado

$$\hat{K} = P^t \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} P,$$

considere

$$\overline{\overline{x_1}} = \begin{bmatrix} \overline{\overline{y_1}} \\ \overline{\overline{y_2}} \end{bmatrix}, \text{ com } \overline{\overline{y_1}} \in \mathbb{R}^m \text{ e } \overline{\overline{y_2}} \in \mathbb{R}^{n-m}$$

$$\text{e } \alpha\beta D^{\frac{1}{2}}r_1 + \gamma P^t \begin{bmatrix} D_B^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} Pr_1 + (\alpha\beta+\gamma)P^t \begin{bmatrix} D_B^{-\frac{1}{2}} \\ 0 \end{bmatrix} B^{-1}r_2 = \begin{bmatrix} r_3 \\ r_4 \end{bmatrix},$$

onde $r_3 \in \mathbb{R}^m$ e $r_4 \in \mathbb{R}^{n-m}$, assim temos

$$P^t \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} P \begin{bmatrix} \overline{\overline{y_1}} \\ \overline{\overline{y_2}} \end{bmatrix} = \begin{bmatrix} r_3 \\ r_4 \end{bmatrix} \Rightarrow$$

$$\begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} P \begin{bmatrix} \overline{\overline{y_1}} \\ \overline{\overline{y_2}} \end{bmatrix} = P \begin{bmatrix} r_3 \\ r_4 \end{bmatrix}.$$

Definindo $P \begin{bmatrix} \overline{\overline{y_1}} \\ \overline{\overline{y_2}} \end{bmatrix} = \begin{bmatrix} \overline{\overline{z_1}} \\ \overline{\overline{z_2}} \end{bmatrix}$ e $P \begin{bmatrix} r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} r_5 \\ r_6 \end{bmatrix}$, com $r_5 \in \mathbb{R}^m$ e $r_6 \in \mathbb{R}^{n-m}$.

Assim segue

$$\begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} \begin{bmatrix} \overline{\overline{z_1}} \\ \overline{\overline{z_2}} \end{bmatrix} = \begin{bmatrix} r_5 \\ r_6 \end{bmatrix}. \tag{4.22}$$

Desta forma, temos que

$$\overline{\overline{x_1}} = P^t \begin{bmatrix} \overline{\overline{z_1}} \\ \overline{\overline{z_2}} \end{bmatrix}, \text{ pois } \begin{bmatrix} \overline{\overline{y_1}} \\ \overline{\overline{y_2}} \end{bmatrix} = P^t \begin{bmatrix} \overline{\overline{z_1}} \\ \overline{\overline{z_2}} \end{bmatrix}.$$

Portanto, temos $\overline{\overline{x}} = \begin{bmatrix} \overline{\overline{x_1}} \\ \overline{\overline{x_2}} \end{bmatrix}$, e assim $x = M^{-t}Q^{-t}\overline{\overline{x}}$, onde $x = \begin{bmatrix} dx \\ dy \end{bmatrix}$ é a direção de Newton do primeiro sistema linear do método preditor-corretor.

Observe que na prática resolvemos o sistema linear dado em (4.22), por um método iterativo, o método Minres, por exemplo. Note ainda que a componente dz dada em (2.23) não é alterada, pois a mesma não é utilizada no sistema aumentado, a saber, $dz = X^{-1}(r_a - Zdx)$.

Temos no Algoritmo 4 o método preditor-corretor, no qual as direções de Newton foram calculadas através da resolução do sistema aumentado precondicionado duplamente, chamado simplesmente de (SAPD).

Algoritmo 4: Algoritmo SAPD

Entrada: Dado (x^0, y^0, z^0) ponto inicial, α, β e γ aceitáveis.

Saída: Solução do sistema precondicionado duplamente.

início
- Para $k = 0, 1, 2, \cdots$
- **enquanto** $k \leqslant maxiter.$ **faça**
  - Calcule o critério de parada conforme (2.16), (2.17) e (2.18).
  - Resolva o sistema (4.19), utilizando o Precondicionador Separador Duplo, encontrando assim (dx_1^k, dy_1^k, dz_1^k). (Direção preditora)
  - Calcule os parâmetros de (2.11) - (2.13).
  - Através do Precondicionador Separador Duplo, resolva (4.19), substituindo o lado direito por $r_1^k = -(X^k)^{-1}r_c^k$ e $r_2^k = 0$, obtendo (dx_2^k, dy_2^k, dz_2^k). (Direção preditora)
  - Calcule os tamanhos dos passos por (2.14) e (2.15).
  - Faça $x^{k+1} \leftarrow x^k + \alpha_p dx^k$ e $(y^{k+1}, z^{k+1}) \leftarrow (y^k, z^k) + \alpha_d(dy^k, dz^k)$.
  - $k = k + 1$; (Atualização)
- **fim**

fim

4.6 Algumas considerações sobre as matrizes S e K

Mostraremos agora que a matriz

$$S = \begin{bmatrix} -D^{-1} & A^t \\ A & 0 \end{bmatrix},$$

onde $S \in \mathbb{R}^{(n+m)\times(n+m)}$, D^{-1} matriz diagonal com $d_{ii} > 0 \ \forall i = 1, \cdots, n$, e A sendo matriz de posto completo, é não singular e indefinida, possuindo n autovalores negativos e m autovalores positivos.

Para isso utilizaremos a relação abaixo e a Lei de Inércia de Sylvester (GOLUB; LOAN, 2012).

$$S = \begin{bmatrix} -D^{-1} & A^t \\ A & 0 \end{bmatrix} = \begin{bmatrix} -D^{-1} & 0 \\ A & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} -D & 0 \\ 0 & ADA^t \end{bmatrix} \begin{bmatrix} -D^{-1} & A^t \\ 0 & \mathbf{I}_m \end{bmatrix}.$$

Assim, $det(S) = det(-D^{-1})det(-D)det(ADA^t)det(-D^{-1}) = det(-D^{-1})det(ADA^t) \neq 0$.

Logo, S é não singular, isto é, não possui autovalor nulo. E temos que o número de autovalores positivos de S é m, pois ADA^t é matriz definida positiva, e o número de autovalores negativos de S é n, pois $-D$ é matriz definida negativa, assim, S é matriz indefinida.

Note, no entanto, que S pode ter autovalores próximos de zero, o que pode prejudicar o desempenho de métodos iterativos.

Além disso, perceba que a matriz

$$Q^{-1}M^{-1}SM^{-t}Q^{-t} = \begin{bmatrix} K & 0 \\ 0 & -D_B^{-1} \end{bmatrix},$$

possui a mesma quantidade de autovalores positivos e negativos que a matriz S, conforme a escolha de algumas matrizes, apenas estão mais bem distribuídos como veremos a frente.

Em relação a matriz $\hat{K}$ dada em (4.18), por simplicidade, sem considerar a matriz P, a mesma será denotada por K e pode ser escrita como

$$K = \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_m & -(\alpha\beta)^{-1}(\alpha\beta+\gamma)W \\ 0 & \mathbf{I}_{n-m} \end{bmatrix}$$

$$\begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m + (\alpha\beta+\gamma)^2WW^t & 0 \\ 0 & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} \begin{bmatrix} \mathbf{I}_m & 0 \\ -(\alpha\beta)^{-1}(\alpha\beta+\gamma)W^t & \mathbf{I}_{n-m} \end{bmatrix}. \tag{4.23}$$

Assim, $det(K) = det(-(\alpha\beta)^2\mathbf{I}_{n-m})det((\alpha\beta+\gamma)^2\mathbf{I}_m + (\alpha\beta+\gamma)^2WW^t) = (-(\alpha\beta)^2)^{(n-m)}det((\alpha\beta+\gamma)^2\mathbf{I}_m + (\alpha\beta+\gamma)^2WW^t) \neq 0$.

Logo, K é não singular, e também pela Lei de Inércia de Sylvester, K possui m autovalores positivos, pois $((\alpha\beta+\gamma)^2\mathbf{I}_m + (\alpha\beta+\gamma)^2 WW^t)$ é matriz definida positiva e $(n-m)$ autovalores negativos. Note que temos uma relação semelhante a (4.23) se considerarmos agora o complemento de Schur dado em (4.18).

Perceba ainda que na expressão dada em (4.23) não utilizamos a matriz de permutação, porém, a mesma não altera os autovalores da matriz K.

De fato, seja $0 \neq v \in \mathbb{R}^n$ um autovetor da matriz K associado ao autovalor λ, assim temos $Kv = \lambda v \Rightarrow P^tKv = \lambda P^tv \Rightarrow P^tKPP^tv = \lambda P^tv$. Seja $y = P^tv$, assim temos, $P^tKPy = \lambda y$, e, portanto, λ é também um autovalor da matriz P^tKP, porém houve uma reordenação nos elementos do vetor v.

Desta forma, vamos calcular de fato os autovalores da matriz K. Para isso, vamos dividir em três casos, conforme o número de linhas e colunas da matriz W. Assim, nos três casos, temos que $K \in \mathbb{R}^{n\times n}$ e $W \in \mathbb{R}^{m\times(n-m)}$, porém o posto da matriz W mudará de acordo com a situação. Em todos os casos a seguir W é matriz de posto completo.

Segue que a matriz $\hat{K}$ sem a matriz de permutação é

$$K = \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix}. \tag{4.24}$$

Como já observado, devemos ter $\alpha, \beta \in \mathbb{R}^*,\ \gamma \in \mathbb{R}$ e $\gamma \neq -\alpha\beta$.

i) Caso n > 2m

Teorema 4.3. *Considerando a matriz dada em (4.24), posto*$(W) = m$*, $n > 2m$, então a matriz K possui m autovalores positivos e maiores do que $(\alpha\beta+\gamma)^2$ e $(n-m)$ autovalores negativos, dos quais m são menores do que $-(\alpha\beta)^2$ e exatamente $(n-2m)$ são iguais a $-(\alpha\beta)^2$.*

Demonstração. Sabemos que a matriz K é simétrica, logo todos os seus n autovalores são reais, veja por exemplo (MEYER, 2000). Seja

$$0 \neq v = \begin{bmatrix} x \\ y \end{bmatrix}$$

um autovetor $\in \mathbb{R}^n$ de K associado ao autovalor λ, com $x \in \mathbb{R}^m$ e $y \in \mathbb{R}^{n-m}$, isto é, $Kv = \lambda v$, então temos

$$\begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow$$

$$\begin{cases} ((\alpha\beta+\gamma)^2\mathbf{I}_m - \lambda\mathbf{I}_m)x + \alpha\beta(\alpha\beta+\gamma)Wy = 0 \\ \alpha\beta(\alpha\beta+\gamma)W^tx - ((\alpha\beta)^2\mathbf{I}_{n-m} + \lambda\mathbf{I}_{n-m})y = 0. \end{cases} \tag{4.25}$$

Vamos mostrar primeiramente que $\lambda \neq (\alpha\beta+\gamma)^2$.

Suponha por absurdo que $\lambda = (\alpha\beta+\gamma)^2$. Assim de (4.25) temos

$$\begin{cases} \alpha\beta(\alpha\beta+\gamma)Wy = 0 \\ \alpha\beta(\alpha\beta+\gamma)W^tx - ((\alpha\beta)^2 + (\alpha\beta+\gamma)^2)y = 0 \end{cases}$$

e isolando y na segunda equação e substituindo na primeira equação, chegamos em $((\alpha\beta)^2 + (\alpha\beta+\gamma)^2)^{-1}(\alpha\beta)^2(\alpha\beta+\gamma)^2WW^tx = 0 \Rightarrow WW^tx = 0$, mas $WW^t \in \mathbb{R}^{m\times m}$ e como $posto(W) = m$, então WW^t é matriz definida positiva, logo não singular, porém isso implica $x = 0$ e assim $y = 0$, o que não pode acontecer, pois

$$v = \begin{bmatrix} x \\ y \end{bmatrix} \neq 0.$$

Portanto $\lambda \neq (\alpha\beta+\gamma)^2$. Assim, a matriz $((\alpha\beta+\gamma)^2\mathbf{I}_m - \lambda\mathbf{I}_m)$ é não singular, daí temos $x = \dfrac{-\alpha\beta(\alpha\beta+\gamma)Wy}{(\alpha\beta+\gamma)^2 - \lambda}$ e substituindo na segunda equação de (4.25) e fazendo as operações algébricas necessárias, chegamos em

$$W^tWy = \left(\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda - 1\right)y. \tag{4.26}$$

Temos que $posto(W) = m$ e $2m < n$. E sabemos que W^tW é matriz semidefinida positiva com $posto(W^tW) = m$.

Desse modo, a matriz W^tW em (4.26) possui m autovalores positivos que vamos denotar por $\delta_1, \cdots, \delta_m$, e um autovalor $\varphi = 0$, com multiplicidade algébrica $p = n - 2m$. Desse modo, considere (δ_i, y_i) um autopar da matriz W^tW.

Assim, $\forall\ i = 1, \cdots, m$, temos

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 = \delta_i \Rightarrow$$

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 - \delta_i = 0. \tag{4.27}$$

Desta forma, aplicando a fórmula de Bhaskara em (4.27), obtemos:

$$\lambda_i = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma + 2(\alpha\beta)^2 + \gamma^2)^2 + 4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}. \tag{4.28}$$

Deste modo, para cada autovalor δ_i de W^tW, temos dois autovalores associados a matriz K, dados por

$$\lambda_i^+ = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}$$

e

$$\lambda_i^- = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}.$$

Como $\delta_i > 0 \ \forall \ i = 1, \cdots, m$, logo

$$\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i} > (2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2).$$

Assim temos $\lambda_i^+ > \frac{(2\alpha\beta\gamma+\gamma^2)}{2}+\frac{1}{2}(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2) = 2\alpha\beta\gamma+(\alpha\beta)^2+\gamma^2 = (\alpha\beta+\gamma)^2$ e $\lambda_i^- < \frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2) = -(\alpha\beta)^2$.

Desta forma, temos que K possui, por enquanto, m autovalores positivos maiores do que $(\alpha\beta+\gamma)^2$ e m autovalores negativos menores do que $-(\alpha\beta)^2$.

Para o autovalor $\varphi = 0$ da matriz W^tW, com multiplicidade algébrica $p = n-2m$, temos

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 = 0, \forall \ i = m+1, \cdots, n-m.$$

E novamente utilizando a fórmula Bhaskara obtemos:

$$\lambda_i = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2} = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2).$$

Assim, temos $\lambda_i^+ = 2\alpha\beta\gamma+(\alpha\beta)^2+\gamma^2 = (\alpha\beta+\gamma)^2$ e $\lambda_i^- = -(\alpha\beta)^2$.

Porém, já sabemos que $\lambda = (\alpha\beta+\gamma)^2$ não pode ser autovalor da matriz K, logo, temos que todos os demais $n-2m$ autovalores da matriz K são iguais a $-(\alpha\beta)^2$.

Portanto, a matriz K é indefinida e possui m autovalores positivos e maiores do que $(\alpha\beta+\gamma)^2$, m autovalores negativos menores do que $-(\alpha\beta)^2$ e exatamente $(n-2m)$ autovalores iguais a $-(\alpha\beta)^2$, para o caso onde $n > 2m$. □

A Figura 1 ilustra a região dos autovalores para este caso.

Dessa forma, podemos escolher convenientemente os escalares α, β e γ de tal forma que a matriz K não tenha autovalores tão próximos de zero, visto que isso pode prejudicar o desempenho de métodos iterativos.

ii) Caso n = 2m

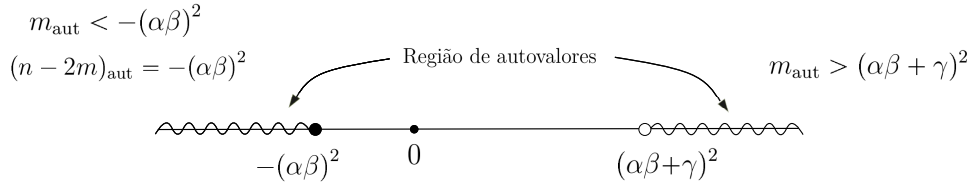

Figura 1 – Caso $n > 2m$
Fonte: Elaborado pelo autor.

Teorema 4.4. *Considerando a matriz dada em (4.24), posto*$(W) = m$*,* $n = 2m$*, então a matriz* K *possui* m *autovalores positivos e maiores do que* $(\alpha\beta+\gamma)^2$ *e* m *autovalores negativos e menores do que* $-(\alpha\beta)^2$*.*

Demonstração. Para o caso onde $n = 2m$, temos que a matriz W, de ordem $m \times (2m-m) = m \times m$, é matriz definida positiva. Assim, seguindo os mesmos passos feitos para o caso $n > 2m$, apenas com exceção do autovalor nulo, segue que a matriz K possui m autovalores positivos e maiores do que $(\alpha\beta+\gamma)^2$ e m autovalores negativos e menores do que $-(\alpha\beta)^2$. □

A Figura 2 ilustra a região dos autovalores para este caso.

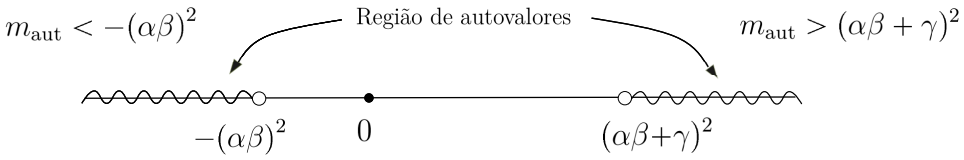

Figura 2 – Caso $n = 2m$
Fonte: Elaborado pelo autor.

iii) Caso n < 2m

Teorema 4.5. *Considerando a matriz dada em (4.24), posto*$(W) = n-m$*,* $n < 2m$*, então a matriz* K *possui* m *autovalores positivos, dos quais* $(n-m)$ *são maiores do que* $(\alpha\beta+\gamma)^2$ *e exatamente* $(2m-n)$ *iguais a* $(\alpha\beta+\gamma)^2$ *e* $(n-m)$ *autovalores negativos e menores do que* $-(\alpha\beta)^2$*.*

Demonstração. Para o caso onde $n < 2m$, temos que a matriz $W \in \mathbb{R}^{m\times(n-m)}$ possui posto $(n-m)$, pois $n-m < m$, daí $W^tW \in \mathbb{R}^{(n-m)\times(n-m)}$ é definida positiva. Seja

$$0 \neq v = \begin{bmatrix} x \\ y \end{bmatrix}$$

um autovetor $\in \mathbb{R}^n$ de K associado ao autovalor λ, com $x \in \mathbb{R}^m$ e $y \in \mathbb{R}^{n-m}$, isto é, $Kv = \lambda v$, então temos

$$\begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow$$

$$\begin{cases} ((\alpha\beta+\gamma)^2\mathbf{I}_m - \lambda\mathbf{I}_m)x + \alpha\beta(\alpha\beta+\gamma)Wy = 0 \\ \alpha\beta(\alpha\beta+\gamma)W^tx - ((\alpha\beta)^2\mathbf{I}_{n-m} + \lambda\mathbf{I}_{n-m})y = 0. \end{cases} \tag{4.29}$$

Mostraremos que $\lambda \neq -(\alpha\beta)^2$.

Suponha por absurdo que $\lambda = -(\alpha\beta)^2$, assim temos que:

$$\begin{cases} ((\alpha\beta+\gamma)^2\mathbf{I}_m + (\alpha\beta)^2\mathbf{I}_m)x + \alpha\beta(\alpha\beta+\gamma)Wy = 0 \\ \alpha\beta(\alpha\beta+\gamma)W^tx = 0. \end{cases} \tag{4.30}$$

Isolando x na primeira equação temos $x = \dfrac{-\alpha\beta(\alpha\beta+\gamma)Wy}{(\alpha\beta+\gamma)^2 + (\alpha\beta)^2}$, e substituindo na segunda equação de (4.30) chegamos em $W^tWy = 0 \Rightarrow y = 0$, logo $x = 0$, o que é uma contradição, pois $v \neq 0$, assim, $\lambda \neq -(\alpha\beta)^2$.

Como $\lambda \neq -(\alpha\beta)^2$, logo a matriz $((\alpha\beta)^2\mathbf{I}_{n-m} + \lambda\mathbf{I}_{n-m})$ é não singular, assim, podemos isolar y na segunda equação de (4.29), isto é, $y = \dfrac{\alpha\beta(\alpha\beta+\gamma)W^tx}{(\alpha\beta)^2 + \lambda}$, e substituindo na primeira equação da mesma e fazendo as operações algébricas necessárias, chegamos em

$$WW^tx = \left(\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda - 1\right)x.$$

Temos que $posto(W) = n - m$ e $n < 2m$. Sabemos que WW^t é matriz semidefinida positiva com $posto(WW^t) = n - m$.

Assim, a matriz WW^t possui $n-m$ autovalores positivos, que são $\delta_1, \cdots, \delta_{n-m}$ e um autovalor $\varphi = 0$, com multiplicidade algébrica $p = 2m - n$.

Desse modo, considere (δ_i, x_i) um autopar da matriz WW^t.

Assim, $\forall\ i = 1, \cdots, n-m$, temos

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 = \delta_i \Rightarrow$$

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 - \delta_i = 0. \tag{4.31}$$

Note que na expressão (4.31) são exatamente os mesmos coeficientes da equação do segundo grau dado em (4.27), logo temos $\lambda_i^+ > (\alpha\beta+\gamma)^2$ e $\lambda_i^- < -(\alpha\beta)^2$, $\forall\ i = 1, \cdots, n-m$.

Para o autovalor $\varphi = 0$ da matriz WW^t com multiplicidade algébrica $p = 2m-n$, chegamos em $\lambda_i^+ = (\alpha\beta+\gamma)^2$ e $\lambda_i^- = -(\alpha\beta)^2$, $\forall\ i = n-m+1, \cdots, m$.

Entretanto, sabemos que para $n < 2m$, $\lambda = -(\alpha\beta)^2$ não pode ser autovalor da matriz K, assim temos que $(2m-n)$ autovalores da matriz K são iguais a $(\alpha\beta+\gamma)^2$.

Portanto, para o caso onde $n < 2m$, a matriz K possui m autovalores positivos, dos quais $(n-m)$ são maiores do que $(\alpha\beta+\gamma)^2$ e exatamente $(2m-n)$ iguais a $(\alpha\beta+\gamma)^2$ e $(n-m)$ autovalores negativos e menores do que $-(\alpha\beta)^2$. □

A Figura 3 mostra a região dos autovalores para este caso.

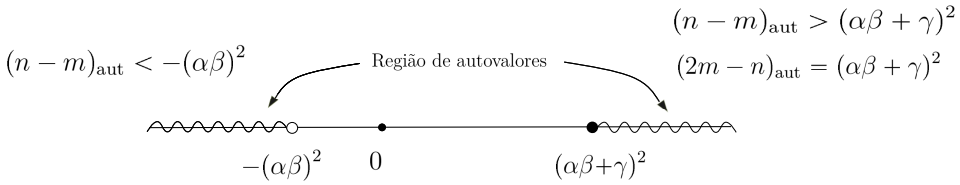

Figura 3 – Caso $n < 2m$
Fonte: Elaborado pelo autor.

Na próxima seção falaremos sobre o número de condição da matriz K, e como esse valor pode ser extremamente útil na resolução de um sistema linear.

4.7 Considerações sobre o número de condição da matriz K na norma-2

Definição 4.1. *O raio espectral de uma matriz $A \in \mathbb{R}^{n\times n}$ é dado por $\rho(A) = max\ \{|\lambda_1|, \cdots, |\lambda_n|\}$, onde $\lambda_1, \cdots, \lambda_n$ são os autovalores da matriz A.*

Para uma matriz real simétrica, logo hermitiana, temos que a norma-2 pode ser calculada como $\|A\|_2 = \rho(A)$, isto é, a norma-2 é o maior dos autovalores, em módulo, da matriz A, veja (SAAD, 2003).

Desta forma, calcularemos o número de condição da matriz K, que é dado por $cond_2(K) = \|K\|_2\|K^{-1}\|_2$, onde

$$K = \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix}.$$

Consideraremos três casos, conforme o número de linhas e colunas da matriz W, e para cada um destes as duas possibilidades, no caso $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ ou $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$.

i) n > 2m

Já vimos, pelo Teorema 4.3 que todos os autovalores da matriz K, que é simétrica, estão no intervalo $(-\infty, -(\alpha\beta)^2] \bigcup ((\alpha\beta+\gamma)^2, +\infty)$, isto é, existem autovalores

que são maiores do que $(\alpha\beta+\gamma)^2$, e existem autovalores, em sua grande maioria, que são menores ou iguais a $-(\alpha\beta)^2$.

a) Consideremos que todos os autovalores de K, estão em $[-\lambda_0, -(\alpha\beta)^2]\bigcup((\alpha\beta+\gamma)^2, \lambda_{max}]$, com λ_{max} maior autovalor positivo, $-\lambda_0$ o menor autovalor negativo e $\lambda_{max} \geqslant -\lambda_0$, isto é, o maior autovalor de K, está do lado positivo.

A Figura 4 ilustra a região dos autovalores para este caso.

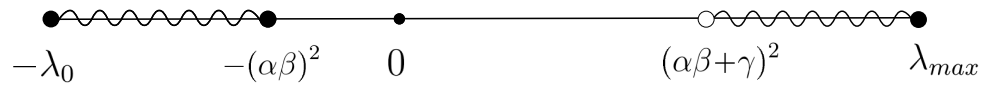

Figura 4 – Região de autovalores, caso λ_{max} esteja a direita da origem
Fonte: Elaborado pelo autor.

Assim, temos que $\|K\|_2 = \lambda_{max}$.

Por outro lado, sabemos que para uma matriz A inversível, se $\lambda \neq 0$ é autovalor de A, então λ^{-1} é autovalor de A^{-1}, de fato, pois temos $Av = \lambda v \Rightarrow A^{-1}Av = \lambda A^{-1}v \Rightarrow A^{-1}v = \lambda^{-1}v$, e como os autovalores de K, estão em $[-\lambda_0, -(\alpha\beta)^2]\bigcup((\alpha\beta+\gamma)^2, \lambda_{max}]$, logo os autovalores da matriz inversa, K^{-1}, estão no intervalo $\left[\frac{-1}{(\alpha\beta)^2}, \frac{-1}{\lambda_0}\right]\bigcup\left[\frac{1}{\lambda_{max}}, \frac{1}{(\alpha\beta+\gamma)^2}\right)$.

Note que podem ocorrer duas alternativas:

I) $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$.

II) $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$.

Caso $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, temos que $|-(\alpha\beta)^2| = (\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ e $\lambda_{max} \geqslant \lambda_0$, e assim segue que $\forall\ \alpha, \beta$ e γ aceitáveis que:

$$\left|\frac{-1}{(\alpha\beta)^2}\right| > \left|\frac{-1}{\lambda_0}\right| > \frac{1}{(\alpha\beta+\gamma)^2} > \frac{1}{\lambda_{max}} \text{ ou } \left|\frac{-1}{(\alpha\beta)^2}\right| \geqslant \frac{1}{(\alpha\beta+\gamma)^2} > \left|\frac{-1}{\lambda_0}\right| \geqslant \frac{1}{\lambda_{max}}.$$

Assim, o maior autovalor, em módulo, da matriz K^{-1}, é $(\alpha\beta)^{-2}$.

Caso $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, temos que $|-(\alpha\beta)^2| = (\alpha\beta)^2 > (\alpha\beta+\gamma)^2$ e $\lambda_{max} \geqslant |-\lambda_0|$, e assim, segue que:

$$\frac{1}{(\alpha\beta+\gamma)^2} > \left|\frac{-1}{\lambda_0}\right| > \frac{1}{(\alpha\beta)^2} > \frac{1}{\lambda_{max}} \text{ ou } \frac{1}{(\alpha\beta+\gamma)^2} > \frac{1}{(\alpha\beta)^2} > \left|\frac{-1}{\lambda_0}\right| \geqslant \frac{1}{\lambda_{max}}.$$

Logo, o maior autovalor da matriz K^{-1}, é $(\alpha\beta+\gamma)^{-2}$, e portanto, $\|K^{-1}\|_2 < (\alpha\beta+\gamma)^{-2}$. Perceba que neste caso, tomamos uma desigualdade na norma-2 da matriz K^{-1}, pois a mesma não está definida em $(\alpha\beta+\gamma)^2$.

Portanto, segue que $cond_2(K) = \|K\|_2\|K^{-1}\|_2 = (\alpha\beta)^{-2}\lambda_{max}$, se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, ou $cond_2(K) = \|K\|_2\|K^{-1}\|_2 < (\alpha\beta+\gamma)^{-2}\lambda_{max}$, se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$.

Assim, de um modo geral, podemos dizer que

$$cond_2(K) \leqslant \min((\alpha\beta)^{-2}, (\alpha\beta+\gamma)^{-2})\lambda_{max},$$

onde λ_{max} é o maior autovalor da matriz K.

b) Considerando a mesma faixa para os autovalores de K, porém, admita $-\lambda_{max}$ do lado esquerdo da origem e $|-\lambda_{max}| > \lambda_0$.

Logo o intervalo é $[-\lambda_{max}, -(\alpha\beta)^2] \bigcup ((\alpha\beta+\gamma)^2, \lambda_0]$. Veja Figura 5.

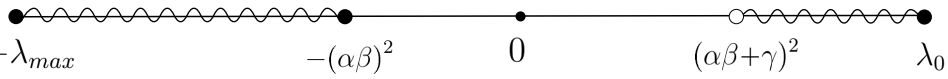

Figura 5 – Região de autovalores, caso $-\lambda_{max}$ esteja a esquerda da origem
Fonte: Elaborado pelo autor.

Assim, segue que $\|K\|_2 = |-\lambda_{max}| = \lambda_{max}$, e os autovalores da matriz K^{-1} estão em $\left[\frac{-1}{(\alpha\beta)^2}, \frac{-1}{\lambda_{max}},\right] \bigcup \left[\frac{1}{\lambda_0}, \frac{1}{(\alpha\beta+\gamma)^2}\right)$.

Se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ temos que $\|K^{-1}\|_2 = (\alpha\beta)^{-2}$, logo, $cond_2(K) = (\alpha\beta)^{-2}\lambda_{max}$, porém se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, temos que $cond_2(K) < (\alpha\beta+\gamma)^{-2}\lambda_{max}$.

Portanto, ao considerarmos $n > 2m$, e independente do sinal de λ_{max} ser positivo ou negativo, temos $cond_2(K) \leqslant \min((\alpha\beta)^{-2}, (\alpha\beta+\gamma)^{-2})\lambda_{max}$, onde λ_{max} é o maior autovalor, em módulo, da matriz K.

O resultado $cond_2(K) \leqslant \min((\alpha\beta)^{-2}, (\alpha\beta+\gamma)^{-2})\lambda_{max}$, mostra que a matriz ser mal condicionada ou não, pode estar intimamente relacionado com um múltiplo do maior autovalor, em módulo, da matriz K. Além disso, o número de condição da matriz não cresce descontroladamente e está limitado ao maior autovalor, em módulo, da matriz K.

ii) n = 2m

De acordo com o Teorema 4.5 temos que todos os autovalores da matriz K estão no intervalo $(-\infty, -(\alpha\beta)^2) \bigcup ((\alpha\beta+\gamma)^2, +\infty)$.

a) Considere primeiro que o maior autovalor λ_{max} esteja a direita da origem, e seja $-\lambda_0$, o menor autovalor negativo, assim os autovalores estão em $[-\lambda_0, -(\alpha\beta)^2) \bigcup ((\alpha\beta+\gamma)^2, \lambda_{max}]$, com $\lambda_{max} \geqslant \lambda_0$, conforme ilustra a Figura 6.

Figura 6 – Região de autovalores, caso λ_{max} esteja a direita da origem
Fonte: Elaborado pelo autor.

Assim temos $\|K\|_2 = \lambda_{max}$.

Já os autovalores da matriz K^{-1} estão em $\left(\frac{-1}{(\alpha\beta)^2}, \frac{-1}{\lambda_0}\right] \bigcup \left[\frac{1}{\lambda_{max}}, \frac{1}{(\alpha\beta+\gamma)^2}\right)$.

Da mesma maneira, pode ocorrer $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ então $\left|\frac{-1}{(\alpha\beta)^2}\right| = \left|\frac{1}{(\alpha\beta)^2}\right| \geqslant \frac{1}{(\alpha\beta+\gamma)^2}$, daí temos $\|K^{-1}\|_2 < (\alpha\beta)^{-2}$, portanto $cond_2(K) < (\alpha\beta)^{-2}\lambda_{max}$.

Pode ocorrer também $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, logo $\left|\frac{-1}{(\alpha\beta)^2}\right| = \left|\frac{1}{(\alpha\beta)^2}\right| < \frac{1}{(\alpha\beta+\gamma)^2}$ assim, $\|K^{-1}\|_2 < (\alpha\beta+\gamma)^{-2}$. Note que para esta situação $n = 2m$, o intervalo é aberto em $-(\alpha\beta)^2$ e $(\alpha\beta+\gamma)^2$.

b) Agora $-\lambda_{max}$ está a esquerda da origem, e considere que os autovalores estão em $[-\lambda_{max}, -(\alpha\beta)^2) \bigcup ((\alpha\beta+\gamma)^2, \lambda_0]$, com $|-\lambda_{max}| > \lambda_0$, conforme mostra a Figura 7.

Figura 7 – Região de autovalores, caso $-\lambda_{max}$ esteja a esquerda da origem
Fonte: Elaborado pelo autor.

Assim temos que $\|K\|_2 = |-\lambda_{max}| = \lambda_{max}$. Quanto a matriz K^{-1}, temos que os autovalores estão em $\left(\frac{-1}{(\alpha\beta)^2}, \frac{-1}{\lambda_{max}}\right] \bigcup \left[\frac{1}{\lambda_0}, \frac{1}{(\alpha\beta+\gamma)^2}\right)$.

Da mesma maneira temos $\|K^{-1}\|_2 < (\alpha\beta)^{-2}$ e $cond_2(K) < (\alpha\beta)^{-2}\lambda_{max}$, caso $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ e $\|K^{-1}\|_2 < (\alpha\beta+\gamma)^{-2}$ e $cond_2(K) < (\alpha\beta+\gamma)^{-2}\lambda_{max}$, caso $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$.

iii) n < 2m

Conforme o Teorema 4.5, sabemos que todos os autovalores da matriz K estão no intervalo $(-\infty, -(\alpha\beta)^2) \bigcup [(\alpha\beta+\gamma)^2, +\infty)$.

a) Considere que o maior autovalor λ_{max} esteja a direita da origem, e seja $-\lambda_0$, o menor autovalor negativo, deste modo temos que os autovalores de K estão em $[-\lambda_0, -(\alpha\beta)^2) \bigcup [(\alpha\beta+\gamma)^2, \lambda_{max}]$, com $\lambda_{max} \geqslant \lambda_0$. Assim segue que $\|K\|_2 = \lambda_{max}$. A Figura 8 ilustra esse caso.

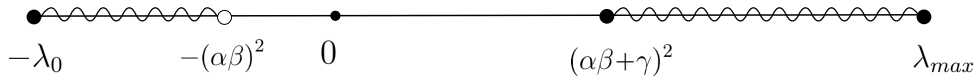

Figura 8 – Região de autovalores, caso λ_{max} esteja a direita da origem
Fonte: Elaborado pelo autor.

Os autovalores da matriz K^{-1} estão em $\left(\frac{-1}{(\alpha\beta)^2}, \frac{-1}{\lambda_0}\right] \bigcup \left[\frac{1}{\lambda_{max}}, \frac{1}{(\alpha\beta+\gamma)^2}\right]$.

Do mesmo modo, caso $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, temos que $\|K^{-1}\|_2 < (\alpha\beta)^{-2}$. Logo, $cond_2(K) < (\alpha\beta)^{-2}\lambda_{max}$, $\|K^{-1}\|_2 = (\alpha\beta+\gamma)^{-2}$ e $cond_2(K) = (\alpha\beta+\gamma)^{-2}\lambda_{max}$, caso $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$. Note que a matriz K não está definida em $-(\alpha\beta)^2$.

b) Neste momento, para o último caso, considere que $-\lambda_{max}$ esteja a esquerda da origem, e que os autovalores estejam em $[-\lambda_{max}, -(\alpha\beta)^2)\bigcup[(\alpha\beta+\gamma)^2, \lambda_0]$, onde $|-\lambda_{max}| > \lambda_0$, conforme Figura 9.

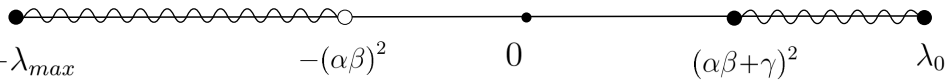

Figura 9 – Região de autovalores, caso $-\lambda_{max}$ esteja a esquerda da origem
Fonte: Elaborado pelo autor.

Também temos $\|K\|_2 = |-\lambda_{max}| = \lambda_{max}$. Em relação a matriz K^{-1}, temos que os autovalores estão em $\left(\frac{-1}{(\alpha\beta)^2}, \frac{-1}{\lambda_{max}}\right]\bigcup\left[\frac{1}{\lambda_0}, \frac{1}{(\alpha\beta+\gamma)^2}\right]$.

Da mesma maneira temos $\|K^{-1}\|_2 < (\alpha\beta)^{-2}$ e $cond_2(K) < (\alpha\beta)^{-2}\lambda_{max}$, se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ e $\|K^{-1}\|_2 = (\alpha\beta+\gamma)^{-2}$ e $cond_2(K) = (\alpha\beta+\gamma)^{-2}\lambda_{max}$, se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$.

Assim, para os três casos, temos

$$cond_2(K) \leqslant min((\alpha\beta)^{-2}, (\alpha\beta+\gamma)^{-2})\lambda_{max}, \tag{4.32}$$

onde λ_{max} é o maior autovalor, em módulo, de K.

Perceba que a expressão dada em (4.32) nos dá uma estimativa do número de condição de uma matriz dependendo apenas do maior autovalor, em módulo, da matriz K e dos parâmetros α, β e γ. Desta forma, se empregarmos um bom método para encontrarmos este maior autovalor, o número de condição pode ser muito bem estimado por esta expressão.

Veremos a seguir, como utilizar o Precondicionador Separador Duplo na resolução do sistema aumentado.

4.8 Um pouco sobre a matriz W

Note que a matriz K não está fixa, a matriz $W = D_B^{-\frac{1}{2}}B^{-1}ND_N^{\frac{1}{2}}$, possui elementos que mudam enquanto vão sendo efetuadas iterações do método de pontos interiores. Veja que próximo a uma solução do sistema, ao menos $(n-m)$ elementos em D^{-1} são grandes.

Desta forma, uma escolha adequada das colunas da matriz A, faz com que $D_N^{\frac{1}{2}}$ e $D_B^{-\frac{1}{2}}$ sejam muito pequenos próximos a uma solução do método de pontos interiores,

logo a matriz $W = D_B^{-\frac{1}{2}} B^{-1} N D_N^{\frac{1}{2}}$ tende a matriz nula, assim temos que

$$\begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix}, \text{ tende a } \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & 0 \\ 0 & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix}.$$

Explicando melhor, a matriz $D^k = (Z^{-1})^k X^k$ possui elementos da forma

$$d_{ii}^k = \frac{x_i^k}{z_i^k},$$

$\forall\ i = \{1, 2, 3, \cdots, n\}$, assim, à medida que o método de pontos interiores se aproxima de uma solução ótima, as variáveis primais e duais convergem para seus respectivos valores ótimos, de modo que o conjunto de índices $\{1, 2, 3, \cdots, n\}$, se particiona em dois distintos subconjuntos, B relacionado as variáveis básicas e N, relacionado as variáveis não-básicas.

Sabemos que próximo a uma solução do problema, os valores correspondendo a matriz não-básica tendem a zero, isto é, $X_N \Rightarrow 0$, assim pela complementaridade devemos ter $Z_N \Rightarrow \hat{z} \neq 0$ sem degenerescência, assim $D_N^k = (Z_N^{-1})^k X_N^k \Rightarrow 0$ para os índices em N. Para os índices em B devemos ter $X_B \Rightarrow \hat{x} \neq 0$ e também pela complementaridade temos $Z_B \Rightarrow 0$, assim $D_B^k = (Z_B^{-1})^k X_B^k \Rightarrow \infty$, logo a matriz $(D_B)^{-1} \Rightarrow 0$. Além disso, esse fato faz com que o sistema aumentado (2.20) ou as equações normais (2.21) se tornem mal condicionadas à medida que o método se aproxima de uma solução ótima. Vale ressaltar que as equações normais se tornam muito mais mal condicionadas do que o sistema aumentado, veja (OLIVEIRA, 1997), sendo essa outra vantagem de se trabalhar com o sistema aumentado.

Desse modo, temos que $W = D_B^{-\frac{1}{2}} B^{-1} N D_N^{\frac{1}{2}}$ tende a matriz nula, e consequentemente K tende a uma matriz diagonal, possuindo assim um número de condição fácil de calcular. Perceba que deste modo, os autovalores positivos da matriz K tendem a $(\alpha\beta+\gamma)^2$, já os autovalores negativos tendem a $-(\alpha\beta)^2$.

Desta forma, próximo a solução do sistema, temos que o número de condição da matriz K é dado por

$$cond_2(K) \approx \frac{(\alpha\beta+\gamma)^2}{(\alpha\beta)^2} = \frac{(\alpha\beta)^2 + 2\alpha\beta\gamma + \gamma^2}{(\alpha\beta)^2} = \left(1 + \frac{\gamma}{\alpha\beta}\right)^2. \tag{4.33}$$

Em (4.33) levamos em consideração que $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$.

Se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, temos que o número de condição da matriz K é

$$cond_2(K) \approx \frac{(\alpha\beta)^2}{(\alpha\beta+\gamma)^2} = \left(\frac{\alpha\beta}{\alpha\beta+\gamma}\right)^2. \tag{4.34}$$

Note em (4.33) ou (4.34) que o número de condição da matriz K, perto da solução tende a depender apenas dos escalares α, β e γ, ou seja, independe inclusive da

matriz B. Assim, para escolhas convenientes desses escalares, podemos fazer com que o número de condição da matriz K seja próximo a unidade, por exemplo, tomando $\alpha = 10$, $\beta = 3$ e $\gamma = 0.001$ em (4.33), temos $cond_2(K) = 1.0000666667$. Dessa maneira, podemos fazer que próximo a uma solução do método de pontos interiores a matriz do sistema fique bem condicionada, fato extremamente desejado na resolução de sistemas lineares.

Assim, nas primeiras iterações o número de condição da matriz K é sempre menor ou igual a um múltiplo do maior autovalor da matriz K. Já próximo a solução do sistema, a matriz dos coeficientes tende a não depender mais do maior autovalor, mas apenas dos parâmetros. Além disso, para escolhas adequadas desses escalares, a matriz tende a ficar bem condicionada, e há um decaimento no número de condição de K , indo de

$$(\alpha\beta)^{-2}\lambda_{max} \text{ até } (\alpha\beta)^{-2}(\alpha\beta+\gamma)^2 = \left(1+\frac{\gamma}{\alpha\beta}\right)^2.$$

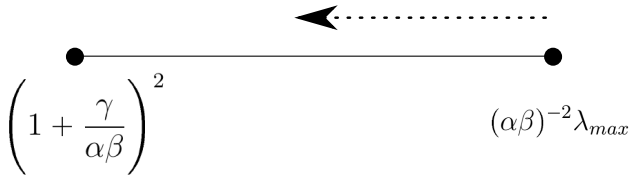

Figura 10 – Decaimento no número de condição da matriz K, caso $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$
Fonte: Elaborado pelo autor.

Ou

$$(\alpha\beta+\gamma)^{-2}\lambda_{max} \text{ até } (\alpha\beta+\gamma)^{-2}(\alpha\beta)^2 = \left(\frac{\alpha\beta}{\alpha\beta+\gamma}\right)^2.$$

Figura 11 – Decaimento no número de condição da matriz K, caso $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$
Fonte: Elaborado pelo autor trocar esse figura.

Observe, quanto mais próximo γ estiver de zero (mantendo α e β controlados), mais bem condicionada a matriz se torna, visto que o número de condição da matriz K, nas iterações finais, depende apenas de α, β e γ.

4.8.1 Diferença entre os precondicionadores

Vejamos agora qual a diferença entre os sistemas aumentados para o Precondicionador Separador Duplo e o proposto em (OLIVEIRA, 1997). Recordemos os elementos

que compõem o Precondicionador Separador Duplo:

$$M^{-1} = \begin{bmatrix} F & G \\ H & 0 \end{bmatrix} \text{ e } Q^{-1} = \begin{bmatrix} U & V \\ 0 & \mathbf{I}_m \end{bmatrix}, \text{ com } V = \gamma P^t \begin{bmatrix} {D_B}^{\frac{1}{2}} \\ 0 \end{bmatrix},$$

$$H = [\mathbf{I}_m \;\; 0]P, \; G = U^{-1}LB^{-1} \text{ e } L = (\alpha\beta + \gamma)P^t \begin{bmatrix} {D_B}^{-\frac{1}{2}} \\ 0 \end{bmatrix}, \text{ assim temos}$$

$$Q^{-1}M^{-1} = \begin{bmatrix} U & V \\ 0 & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} F & G \\ H & 0 \end{bmatrix} = \begin{bmatrix} UF + VH & UG \\ H & 0 \end{bmatrix} =$$

$$\begin{bmatrix} \alpha\beta D^{\frac{1}{2}} + \gamma P^t \begin{bmatrix} D_B^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} P & (\alpha\beta + \gamma)P^t \begin{bmatrix} D_B^{-\frac{1}{2}} \\ 0 \end{bmatrix} B^{-1} \\ H & 0 \end{bmatrix}. \tag{4.35}$$

Observe na matriz (4.35) que devemos ter $UF = \alpha\beta D^{\frac{1}{2}}$. Já o precondicionador proposto por (OLIVEIRA, 1997) é dado por

$$\begin{bmatrix} D^{-\frac{1}{2}} & P^t \begin{bmatrix} D_B^{\frac{1}{2}} \\ 0 \end{bmatrix} B^{-1} \\ H & 0 \end{bmatrix}. \tag{4.36}$$

Perceba que a expressão dada em (4.35) é uma generalização de (4.36), para vermos isso, basta tomarmos $\alpha\beta = 1$ e $\gamma = 0$. Repare que esses valores são equivalentes ao Precondicionador Separador. Na verdade, foi mostrado na Subseção 4.10.1 que toda solução da forma $\gamma = -2\alpha\beta$ produz o mesmo número de condição entre os dois precondicionadores. Note ainda que em (OLIVEIRA, 1997) foi tomado $D = X^{-1}Z$, contrário da maneira que expusemos no presente texto. Dessa forma, o Precondicionador Separador pode ser visto como um caso particular do Precondicionador Separador Duplo.

Portanto, uma boa justificativa da proposta do Precondicionador Separador Duplo, se dá em tentarmos saber qual deverá ser, se de fato há, uma melhor escolha para os parâmetros α, β e γ.

4.9 Matriz W de posto Incompleto

Trataremos o caso onde $W = D_B^{-\frac{1}{2}} B^{-1} N D_N^{\frac{1}{2}}$ é matriz de posto incompleto. Recorde a matriz K é dada por:

$$K = \begin{bmatrix} (\alpha\beta + \gamma)^2 \mathbf{I}_m & \alpha\beta(\alpha\beta + \gamma)W \\ \alpha\beta(\alpha\beta + \gamma)W^t & -(\alpha\beta)^2 \mathbf{I}_{n-m} \end{bmatrix}. \tag{4.37}$$

Vamos dividir em três casos, conforme o número de linhas e colunas da matriz W, porém para todos esses casos, W é matriz de posto incompleto.

i) Caso $n > 2m$

Teorema 4.6. *Considerando a matriz dada em (4.37), posto$(W) = r$, $r < m$ e $n > 2m$, então a matriz K possui m autovalores positivos, dos quais r são maiores do que $(\alpha\beta+\gamma)^2$ e exatamente $(m-r)$ iguais a $(\alpha\beta+\gamma)^2$ e $(n-m)$ autovalores negativos, dos quais r são menores do que $-(\alpha\beta)^2$ e exatamente $(n-m-r)$ são iguais a $-(\alpha\beta)^2$.*

Demonstração. Seja

$$0 \neq v = \begin{bmatrix} x \\ y \end{bmatrix}$$

um autovetor $\in \mathbb{R}^n$ de K associado ao autovalor λ, com $x \in \mathbb{R}^m$ e $y \in \mathbb{R}^{n-m}$, isto é, $Kv = \lambda v$, então temos

$$\begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow$$

$$\begin{cases} ((\alpha\beta+\gamma)^2\mathbf{I}_m - \lambda\mathbf{I}_m)x + \alpha\beta(\alpha\beta+\gamma)Wy = 0 \\ \alpha\beta(\alpha\beta+\gamma)W^t x - ((\alpha\beta)^2\mathbf{I}_{n-m} + \lambda\mathbf{I}_{n-m})y = 0. \end{cases}$$

O que é equivalente a

$$\begin{cases} Wy = \left(\dfrac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} - \dfrac{(\alpha\beta+\gamma)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)x \\ W^t x = \left(\dfrac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} + \dfrac{(\alpha\beta)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)y. \end{cases} \tag{4.38}$$

Pré multiplicando a segunda equação de (4.38) por W, temos

$$WW^t x = \left(\frac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} + \frac{(\alpha\beta)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)Wy$$

e usando a primeira equação da mesma, segue

$$WW^t x = \left(\frac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} + \frac{(\alpha\beta)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)\left(\frac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} - \frac{(\alpha\beta+\gamma)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)x$$

$$WW^t x = \left(\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda - 1\right)x.$$

Temos que $posto(WW^t) = r < m < n-m$. Desse modo, a matriz WW^t possui r autovalores positivos, denotados por $\delta_1, \cdots, \delta_r$ e um autovalor $\varphi = 0$, com multiplicidade algébrica $p = m - r$. Considere (δ_i, x_i) um autopar da matriz WW^t.

Assim, $\forall\ i = 1, \cdots, r$, temos

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 = \delta_i \Rightarrow$$

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 - \delta_i = 0. \tag{4.39}$$

Perceba que em (4.39) temos a mesma expressão que em (4.28). Assim segue:

$$\lambda_i = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma + 2(\alpha\beta)^2+\gamma^2)^2 + 4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}.$$

Deste modo, para cada autovalor δ_i de WW^t, temos dois autovalores associados a matriz K, dados por

$$\lambda_i^+ = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma + 2(\alpha\beta)^2+\gamma^2)^2 + 4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}$$

e

$$\lambda_i^- = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma + 2(\alpha\beta)^2+\gamma^2)^2 + 4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}.$$

Como $\delta_i > 0,\ \forall\ i = 1, \cdots, r$, logo

$$\sqrt{(2\alpha\beta\gamma + 2(\alpha\beta)^2+\gamma^2)^2 + 4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i} > (2\alpha\beta\gamma + 2(\alpha\beta)^2+\gamma^2).$$

Assim temos $\lambda_i^+ > \frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2) = 2\alpha\beta\gamma+(\alpha\beta)^2+\gamma^2 = (\alpha\beta+\gamma)^2$ e $\lambda_i^- < \frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}(2\alpha\beta\gamma + 2(\alpha\beta)^2+\gamma^2) = -(\alpha\beta)^2$.

Desta forma, temos que K possui, por enquanto, r autovalores positivos maiores do que $(\alpha\beta+\gamma)^2$ e r autovalores negativos menores do que $-(\alpha\beta)^2$.

Para o autovalor $\varphi = 0$ da matriz WW^t, com multiplicidade algébrica $p = m-r$, temos

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 = 0, \forall\ i = r+1, \cdots, m.$$

E novamente utilizando a fórmula de Bhaskara obtemos

$$\lambda_i = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma + 2(\alpha\beta)^2+\gamma^2)^2} = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}(2\alpha\beta\gamma + 2(\alpha\beta)^2+\gamma^2).$$

Assim, temos $\lambda_i^+ = 2\alpha\beta\gamma + (\alpha\beta)^2+\gamma^2 = (\alpha\beta+\gamma)^2$ e $\lambda_i^- = -(\alpha\beta)^2$.

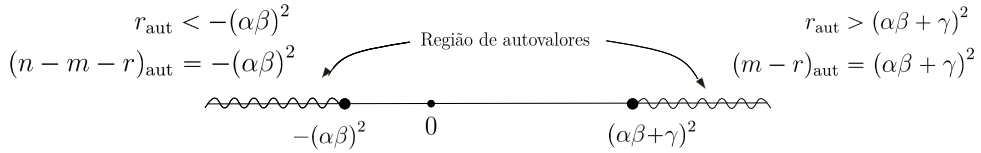

Figura 12 – Caso $n > 2m$ e W matriz de posto incompleto
Fonte: Elaborado pelo autor.

Portanto, a matriz K é indefinida e possui m autovalores positivos, dos quais r são maiores do que $(\alpha\beta+\gamma)^2$ e exatamente $(m-r)$ iguais a $(\alpha\beta+\gamma)^2$ e $(n-m)$ autovalores negativos, dos quais r são menores do que $-(\alpha\beta)^2$ e exatamente $(n-m-r)$ são iguais a $-(\alpha\beta)^2$. A Figura 12 apresenta a a região dos autovalores para este caso. □

ii) Caso n = 2m

Teorema 4.7. *Considerando a matriz dada em (4.37), posto*$(W) = r < m$, $n = 2m$, *então a matriz* K *possui* m *autovalores positivos, dos quais* r *são maiores do que* $(\alpha\beta+\gamma)^2$ *e exatamente* $(m-r)$ *são iguais a* $(\alpha\beta+\gamma)^2$ *e* m *autovalores negativos, dos quais* r *são menores do que* $-(\alpha\beta)^2$ *e exatamente* $(m-r)$ *são iguais a* $-(\alpha\beta)^2$.

Demonstração. Para o caso onde $n = 2m$, temos que a matriz W, de ordem $m \times (2m-m) = m \times m$, com *posto*$(W) = r$ e $r < m$ é matriz semidefinida positiva. Assim, seguindo os mesmos passos feitos para o caso $n > 2m$, apenas com exceção do autovalor nulo, segue que a matriz K possui m autovalores positivos, dos quais r são maiores do que $(\alpha\beta+\gamma)^2$ e exatamente $(m-r)$ são iguais a $(\alpha\beta+\gamma)^2$ e m autovalores negativos, dos quais r são menores do que $-(\alpha\beta)^2$ e exatamente $(m-r)$ são iguais a $-(\alpha\beta)^2$, conforme Figura 13. □

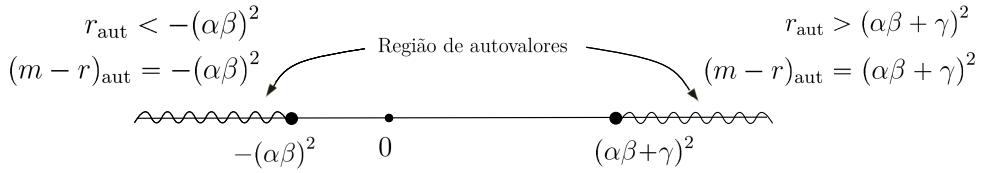

Figura 13 – Caso $n = 2m$ e W matriz de posto incompleto
Fonte: Elaborado pelo autor.

iii) Caso n < 2m

Teorema 4.8. *Considerando a matriz dada em (4.37), posto*$(W) = r$, $r < n-m$ *e* $n < 2m$, *então a matriz* K *possui* m *autovalores positivos, dos quais* r *são maiores do que* $(\alpha\beta+\gamma)^2$ *e exatamente* $(m-r)$ *iguais a* $(\alpha\beta+\gamma)^2$ *e* $(n-m)$ *autovalores negativos, dos quais* r *são menores do que* $-(\alpha\beta)^2$ *e exatamente* $(n-m-r)$ *são iguais a* $-(\alpha\beta)^2$.

Demonstração. Seja

$$0 \neq v = \begin{bmatrix} x \\ y \end{bmatrix}$$

um autovetor $\in \mathbb{R}^n$ de K associado ao autovalor λ, com $x \in \mathbb{R}^m$ e $y \in \mathbb{R}^{n-m}$, isto é, $Kv = \lambda v$, então temos

$$\begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow$$

$$\begin{cases} ((\alpha\beta+\gamma)^2\mathbf{I}_m - \lambda\mathbf{I}_m)x + \alpha\beta(\alpha\beta+\gamma)Wy = 0 \\ \alpha\beta(\alpha\beta+\gamma)W^t x - ((\alpha\beta)^2\mathbf{I}_{n-m} + \lambda\mathbf{I}_{n-m})y = 0. \end{cases}$$

$$\begin{cases} Wy = \left(\dfrac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} - \dfrac{(\alpha\beta+\gamma)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)x \\ W^t x = \left(\dfrac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} + \dfrac{(\alpha\beta)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)y. \end{cases} \tag{4.40}$$

Pré multiplicando a primeira equação de (4.40) por W^t, temos

$$W^tWy = \left(\frac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} - \frac{(\alpha\beta+\gamma)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)W^t x$$

e usando a segunda equação da mesma, segue

$$W^tWy = \left(\frac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} - \frac{(\alpha\beta+\gamma)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)\left(\frac{\lambda}{\alpha\beta(\alpha\beta+\gamma)} + \frac{(\alpha\beta)^2}{\alpha\beta(\alpha\beta+\gamma)}\right)y$$

$$W^tWy = \left(\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda - 1\right)y.$$

Temos que $posto(W^tW) = r < n-m < m$. Desse modo, a matriz W^tW possui r autovalores positivos, denotados por $\delta_1, \cdots, \delta_r$ e um autovalor $\varphi = 0$, com multiplicidade algébrica $p = n-m-r$. Considere (δ_i, y_i) um autopar da matriz W^tW.

Assim, $\forall\ i = 1, \cdots, r$, temos

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 = \delta_i \Rightarrow$$

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 - \delta_i = 0. \tag{4.41}$$

Perceba que em (4.41) temos novamente a mesma expressão que em (4.28). Assim segue:

$$\lambda_i = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma + 2(\alpha\beta)^2 + \gamma^2)^2 + 4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}.$$

Deste modo, para cada autovalor δ_i de W^tW, temos dois autovalores associados a matriz K, dados por

$$\lambda_i^+ = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}$$

e

$$\lambda_i^- = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}.$$

Como $\delta_i > 0,\ \forall\ i = 1, \cdots, r$, logo

$$\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i} > (2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2).$$

Assim temos $\lambda_i^+ > \frac{(2\alpha\beta\gamma+\gamma^2)}{2}+\frac{1}{2}(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2) = 2\alpha\beta\gamma+(\alpha\beta)^2+\gamma^2 = (\alpha\beta+\gamma)^2$ e $\lambda_i^- < \frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2) = -(\alpha\beta)^2$.

Desta forma, temos que K possui, por enquanto, r autovalores positivos maiores do que $(\alpha\beta+\gamma)^2$ e r autovalores negativos menores do que $-(\alpha\beta)^2$.

Para o autovalor $\varphi = 0$ da matriz W^tW, com multiplicidade algébrica $p = n-m-r$, temos

$$\frac{1}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i^2 - \frac{(2\alpha\beta\gamma+\gamma^2)}{(\alpha\beta)^2(\alpha\beta+\gamma)^2}\lambda_i - 1 = 0, \forall\ i = r+1, \cdots, n-m.$$

E novamente utilizando a fórmula Bhaskara obtemos:

$$\lambda_i = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2} = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2).$$

Assim, temos $\lambda_i^+ = 2\alpha\beta\gamma+(\alpha\beta)^2+\gamma^2 = (\alpha\beta+\gamma)^2$ e $\lambda_i^- = -(\alpha\beta)^2$.

Portanto, a matriz K é indefinida e possui m autovalores positivos, dos quais r são maiores do que $(\alpha\beta+\gamma)^2$ e exatamente $(m-r)$ iguais a $(\alpha\beta+\gamma)^2$ e $(n-m)$ autovalores negativos, dos quais r são menores do que $-(\alpha\beta)^2$ e exatamente $(n-m-r)$ são iguais a $-(\alpha\beta)^2$, conforme mostra a Figura 14. □

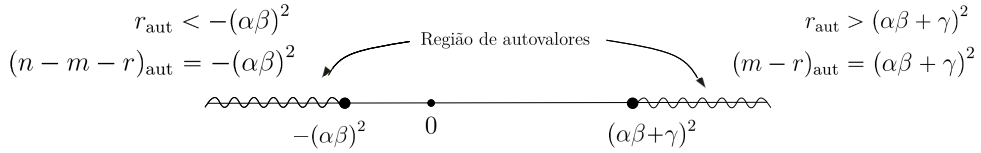

Figura 14 – Caso $n < 2m$ e W matriz de posto incompleto
Fonte: Elaborado pelo autor.

Perceba que em todos os três casos possíveis, quando W é matriz de posto incompleto, os menores autovalores em módulos são atingidos, no caso, $(\alpha\beta+\gamma)^2$ para os autovalores positivos e $-(\alpha\beta)^2$ para os autovaloers negativos. Note ainda que se tomarmos $r = m$ nos Teoremas (4.6) ou (4.7), temos os Teoremas (4.3) ou (4.4). E se tomarmos $r = n-m$ no Teorema (4.8), temos o Teorema (4.5).

4.10 Número de condição da matriz K revisitado

Faremos uma análise mais detalhada sobre o número de condição, na norma-2, da matriz K dada abaixo:

$$K = \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix}.$$

Vamos dividir em três casos, conforme o número de linhas e colunas da matriz W. Assim, nos três casos, temos que $K \in \mathbb{R}^{n\times n}$ e $W \in \mathbb{R}^{m\times(n-m)}$, porém o posto da matriz W muda de acordo com a situação em que se encontra.

De acordo com (4.28) todos os autovalores da matriz K são do tipo:

$$\lambda_i = \frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_i}.$$

Vamos utilizar o seguinte resultado.

Se A é simétrica, então $cond_2(A) = \dfrac{max\{|\lambda|, \text{onde } \lambda \text{ é autovalor de A}\}}{min\{|\lambda|, \text{onde } \lambda \text{ é autovalor de A}\}}$.

i) Caso n > 2m

a) Matriz W de posto completo

Sabemos que W^tW é matriz semidefinida positiva. Seja $\delta_{max}(W^tW)$ o maior autovalor de W^tW, e $\delta_{min}(W^tW) = 0$ o menor autovalor. Assim segue que o número de condição da matriz K é dado por:

$$cond_2(K) = \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{min}(W^tW)}\right|}$$

$$= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2}\right|}$$

$$= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}|(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)|\right|}$$

$$= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|-(\alpha\beta)^2,(\alpha\beta+\gamma)^2\right|} \tag{4.42}$$

Note que nesta situação o valor $(\alpha\beta+\gamma)^2$ não é atingido, se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, então $2\alpha\beta\gamma+\gamma^2 \geqslant 0$, e que podemos retirar o módulo do numerador de (4.42), assim segue

$$cond_2(K) = \frac{\frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}}{(\alpha\beta)^2}. \tag{4.43}$$

Porém, se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, logo $(2\alpha\beta\gamma+\gamma^2) < 0$, e assim em (4.42), conseguiremos apenas um limitante superior.

$$cond_2(K) < \frac{\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{(\alpha\beta+\gamma)^2}. \tag{4.44}$$

b) Matriz W de posto incompleto

Para o caso onde W é matriz de posto incompleto, então sempre o menor autovalor é atingido, sendo em $|-(\alpha\beta)^2|$ ou $(\alpha\beta+\gamma)^2$, segue

$$\begin{aligned}
cond_2(K) &= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2}\right|} \\
&= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}|(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)|\right|} \\
&= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|-(\alpha\beta)^2,(\alpha\beta+\gamma)^2\right|}.
\end{aligned} \tag{4.45}$$

E do mesmo modo, segue que se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, então por (4.45) segue

$$cond_2(K) = \frac{\frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}}{(\alpha\beta)^2}. \tag{4.46}$$

E se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, então

$$cond_2(K) = \frac{\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{(\alpha\beta+\gamma)^2}. \tag{4.47}$$

ii) Caso n = 2m

a) Matriz W de posto completo

Nesta situação tanto $-(\alpha\beta)^2$ quanto $(\alpha\beta+\gamma)^2$ são atingidos. Assim conseguimos apenas um limitante superior para duas possibilidades:

$(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ ou $(\alpha\beta)^2 > (\alpha\beta\gamma)^2$. Assim segue:

$$\begin{aligned}
cond_2(K) &= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{min}(W^tW)}\right|} \\
&< \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2}\right|} \\
&< \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}|(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)|\right|} \\
&< \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|-(\alpha\beta)^2,(\alpha\beta+\gamma)^2\right|}.
\end{aligned} \tag{4.48}$$

Se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, logo $\gamma^2+2\alpha\beta\gamma \geqslant 0$, assim por (4.48) segue

$$cond_2(K) < \frac{\frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}}{(\alpha\beta)^2}. \tag{4.49}$$

E se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$ segue

$$cond_2(K) < \frac{\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{(\alpha\beta+\gamma)^2}. \tag{4.50}$$

b) Matriz W de posto incompleto

Para o caso $n = 2m$ e W matriz de posto incompleto, o menor autovalor é atingido, sendo em $|-(\alpha\beta)^2|$ ou $(\alpha\beta+\gamma)^2$, assim segue

$$\begin{aligned} cond_2(K) &= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2}\right|} \\ &= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm \frac{1}{2}|(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)|\right|} \\ &= \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{min\left|-(\alpha\beta)^2,(\alpha\beta+\gamma)^2\right|}. \end{aligned} \tag{4.51}$$

Se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ então por (4.51) temos

$$cond_2(K) = \frac{\frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}}{(\alpha\beta)^2}. \tag{4.52}$$

E se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$ então

$$cond_2(K) = \frac{\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(W^tW)}\right|}{(\alpha\beta+\gamma)^2}. \tag{4.53}$$

iii) Caso n < 2m

a) Matriz W de posto completo

Para este caso, temos que a matriz WW^t é semidefinida positiva. Seja $\delta_{max}(WW^t)$ o maior autovalor de WW^t, e seja $\delta_{min}(WW^t) = 0$ o menor autovalor. Assim segue que o número de condição da matriz K é dado por

$$cond_2(K) = \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}\right|}{min\left|-(\alpha\beta)^2,(\alpha\beta+\gamma)^2\right|}. \tag{4.54}$$

Relembre que nesta situação o valor $-(\alpha\beta)^2$ não é atingido.

Desse modo, se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, logo $\gamma^2+2\alpha\beta\gamma \geqslant 0$. Assim em (4.54), temos

$$cond_2(K) < \frac{\frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}}{(\alpha\beta)^2}. \tag{4.55}$$

Porém, se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, segue

$$cond_2(K) = \frac{\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}\right|}{(\alpha\beta+\gamma)^2}. \tag{4.56}$$

b) Matriz W de posto incompleto

Para o caso onde W é matriz de posto incompleto, então sempre o menor autovalor é atingido, sendo em $|-(\alpha\beta)^2|$ ou $(\alpha\beta+\gamma)^2$, segue

$$cond_2(K) = \frac{max\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} \pm\frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}\right|}{min\left|-(\alpha\beta)^2,(\alpha\beta+\gamma)^2\right|}. \tag{4.57}$$

E do mesmo modo, se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, logo $\gamma^2+2\alpha\beta\gamma \geqslant 0$. Assim por (4.57) temos

$$cond_2(K) = \frac{\frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}}{(\alpha\beta)^2}. \tag{4.58}$$

E se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, segue

$$cond_2(K) = \frac{\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}\right|}{(\alpha\beta+\gamma)^2}. \tag{4.59}$$

Vale destacar no caso onde W é matriz de posto incompleto, e para os três casos, tanto $n > 2m$, $n = 2m$ e $n < 2m$, quanto $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$ ou $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, temos sempre a igualdade na fórmula que nos fornece o número de condição.

Se $(\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2$, e de acordo com (4.43), (4.49) e (4.55), segue que:

$$cond_2(K) \leqslant \frac{\frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}}{(\alpha\beta)^2}. \tag{4.60}$$

Porém, se $(\alpha\beta)^2 > (\alpha\beta+\gamma)^2$, e de acordo com (4.44), (4.50) e (4.56), segue que:

$$cond_2(K) \leqslant \frac{\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}\right|}{(\alpha\beta+\gamma)^2}. \tag{4.61}$$

4.10.1 Valores Ótimos

Para encontrarmos os valores ótimos de (4.60) e (4.61), devemos resolver os seguintes problemas de minimização com restrição:

$$\min f(\alpha,\beta,\gamma) = \frac{\frac{(2\alpha\beta\gamma+\gamma^2)}{2} + \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}}{(\alpha\beta)^2} \tag{4.62}$$

$$\text{s.a: } (\alpha\beta)^2 \leqslant (\alpha\beta+\gamma)^2.$$

E

$$\min f(\alpha,\beta,\gamma) = \frac{\left|\frac{(2\alpha\beta\gamma+\gamma^2)}{2} - \frac{1}{2}\sqrt{(2\alpha\beta\gamma+2(\alpha\beta)^2+\gamma^2)^2+4(\alpha\beta)^2(\alpha\beta+\gamma)^2\delta_{max}(WW^t)}\right|}{(\alpha\beta+\gamma)^2} \tag{4.63}$$

$$\text{s.a: } (\alpha\beta)^2 > (\alpha\beta+\gamma)^2.$$

Para ambos os problemas, (4.62) e (4.63), foram encontrados os seguintes pontos mínimos:

a) $\gamma = -2\alpha\beta$, com $\alpha\beta > 0$[7].

b) $\gamma = 0$ e $\forall\ \alpha,\beta \in \mathbb{R}^*$, em particular $\alpha\beta = 1$ e $\gamma = 0$.

Perceba em **b)**, que o ponto da forma $\alpha\beta = 1$ e $\gamma = 0$ é equivalente ao Precondicionador Separador. Já o ponto **a)** nos diz que encontramos toda uma família de pontos ótimos que produzem o mesmo número de condição que o Precondicionador Separador. Vale ressaltar que o ponto ótimo do Separador não é um caso particular da família $\gamma = -2\alpha\beta$, ou seja, encontramos todos esses infinitos pontos de uma maneira totalmente distinta do Precondicionador Separador. A seguir, mostramos que de fato esses dois pontos produzem o mesmo número de condição.

Substituindo o ponto **a)** na função que nos dá o número de condição (4.62) ou (4.63) resulta em $\sqrt{1+\delta_{max}(WW^t)}$, que é exatamente o mesmo valor quando substituímos o ponto **b)** em (4.62) ou (4.63). Note ainda que o número de condição da matriz K depende do maior autovalor da matriz W.

[7] essa solução de (4.63) foi encontrada através do programa Mathematica para $(\alpha\beta)^2 \geqslant (\alpha\beta+\gamma)^2$.

Portanto, para $\gamma = -2\alpha\beta$, temos que o número de condição do Precondicionador Separador Duplo é exatamente igual ao número de condição do Precondicionador Separador. Ou seja, o número de condição do Precondicionador Separador Duplo no ponto ótimo, é igual ao número de condição do Precondicionador Separador na norma-2. Nesse sentido, o Precondicionador Separador é ótimo.

Entretanto, como veremos nos testes no Capítulo 5, se utilizarmos outra norma, como a norma-∞ ou norma-1, o número de condição do Precondicionador Separador Duplo foi melhor do que em vários casos do que o Precondicionador Separador.

4.11 Número de condição da matriz K na norma-∞ ou norma-1

Vamos calcular o número de condição da matriz K, dada a seguir, na norma-∞.

$$K = \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix}.$$

O número de condição da matriz K na norma infinita é $cond_\infty(K)=\|K\|_\infty\|K^{-1}\|_\infty$.

Observe que é indiferente trabalharmos com a norma-∞ ou norma-1, pois a matriz K é simétrica.

A norma infinita da matriz K é dada por

$$\|K\|_\infty = max((\alpha\beta+\gamma)^2 + |\alpha\beta(\alpha\beta+\gamma)|\|W\|_\infty, (\alpha\beta)^2 + |\alpha\beta(\alpha\beta+\gamma)|\|W^t\|_\infty).$$

A inversa da matriz K é

$$K^{-1} = \begin{bmatrix} (\alpha\beta+\gamma)^{-2}(\mathbf{I}_m + WW^t)^{-1} & (\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}W(\mathbf{I}_{n-m} + W^tW)^{-1} \\ (\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}(\mathbf{I}_{n-m} + W^tW)^{-t}W^t & -(\alpha\beta)^{-2}(\mathbf{I}_{n-m} + W^tW)^{-1} \end{bmatrix} \tag{4.64}$$

Para calcular a norma infinita da matriz K^{-1} em (4.64), devemos dividir em duas matrizes, a primeira contendo o primeiro e o segundo bloco. E a segunda matriz, contendo o terceiro e o quarto bloco, como a seguir:

$$K^{-1} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix}.$$

Onde

$$K_1 = [(\alpha\beta+\gamma)^{-2}(\mathbf{I}_m + WW^t)^{-1} \quad (\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}W(\mathbf{I}_{n-m} + W^tW)^{-1}]$$

e

$$K_2 = [(\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}(\mathbf{I}_{n-m} + W^tW)^{-t}W^t \quad -(\alpha\beta)^{-2}(\mathbf{I}_{n-m} + W^tW)^{-1}].$$

Desse modo, segue que a norma da matriz K^{-1} é dada por

$$\|K^{-1}\|_\infty = max(\|K_1\|_\infty, \|K_2\|_\infty).$$

Portanto, o número de condição da matriz K na norma infinita é dada por

$$cond_\infty(K)=(max((\alpha\beta+\gamma)^2+|\alpha\beta(\alpha\beta+\gamma)|\|W\|_\infty,(\alpha\beta)^2+|\alpha\beta(\alpha\beta+\gamma)|\|W^t\|_\infty))(max(\|K_1\|_\infty,\|K_2\|_\infty)). \tag{4.65}$$

Seja $A_1 = (\mathbf{I}_m + WW^t)^{-1}$, $A_2 = (\mathbf{I}_{n-m} + W^tW)^{-1}$ e $A_3 = WA_2$.

Desse modo, a norma infinita de K_1 e K_2 podem ser escritas e estimadas por:

$K_1 = [(\alpha\beta+\gamma)^{-2}A_1 \quad (\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}A_3] \Rightarrow \|K_1\|_\infty \leqslant (\alpha\beta+\gamma)^{-2}\|A_1\|_\infty + |(\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}|\|A_3\|_\infty$.

$K_2 = [(\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}{A_3}^t \quad -(\alpha\beta)^{-2}A_2] \Rightarrow \|K_2\|_\infty \leqslant |(\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}|\|{A_3}^t\|_\infty + (\alpha\beta)^{-2}\|A_2\|_\infty$.

Em (4.65) temos 4 possibilidades a se considerar. A fim de facilitar a leitura, vamos chamar $u = \alpha\beta$, $F_1 = (u+\gamma)^2 + |u(u+\gamma)|\|W\|_\infty$, $F_2 = u^2 + |u(u+\gamma)|\|W^t\|_\infty$, $F_3 = (u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|\|A_3\|_\infty$ e $F_4 = |(u(u+\gamma))^{-1}|\|A_3^t\|_\infty + u^{-2}\|A_2\|_\infty$.

Primeira possibilidade

$$\begin{aligned} &\textbf{min } F_1F_3 \\ &\textbf{s.a}: F_1 - F_2 \geqslant 0 \\ &\quad F_3 - F_4 \geqslant 0. \end{aligned} \tag{4.66}$$

Portanto, temos que resolver o seguinte problema de minimização com restrições

$$\begin{aligned} &\textbf{min } f(u,\gamma) = ((u+\gamma)^2 + |u(u+\gamma)|\|W\|_\infty)((u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|\|A_3\|_\infty) \\ &\textbf{s.a}: (u+\gamma)^2 + |u(u+\gamma)|\|W\|_\infty - u^2 - |u(u+\gamma)|\|W^t\|_\infty \geqslant 0 \\ &\quad (u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|\|A_3\|_\infty - |(u(u+\gamma))^{-1}|\|{A_3}^t\|_\infty - u^{-2}\|A_2\|_\infty \geqslant 0. \end{aligned} \tag{4.67}$$

Simplificando o conjunto de restrições (4.67), temos

$$\begin{aligned} &\textbf{min } f(u,\gamma) = ((u+\gamma)^2 + |u(u+\gamma)|\|W\|_\infty)((u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|\|A_3\|_\infty) \\ &\textbf{s.a}: 2u\gamma + \gamma^2 + |u(u+\gamma)|(\|W\|_\infty - \|W^t\|_\infty) \geqslant 0 \\ &\quad (u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|(\|A_3\|_\infty - \|A_3^t\|_\infty) - u^{-2}\|A_2\|_\infty \geqslant 0. \end{aligned}$$

Segunda possibilidade

$$\begin{aligned} &\textbf{min } F_1F_4 \\ &\textbf{s.a}: F_1 - F_2 \geqslant 0 \\ &\quad F_4 - F_3 \geqslant 0. \end{aligned} \tag{4.68}$$

Logo, temos que resolver o seguinte problema de minimização com restrições

$$\begin{aligned} &\textbf{min } f(u,\gamma) = ((u+\gamma)^2 + |u(u+\gamma)|\|W\|_\infty)(|(u(u+\gamma))^{-1}|\|A_3^t\|_\infty + u^{-2}\|A_2\|_\infty) \\ &\textbf{s.a}: (u+\gamma)^2 + |u(u+\gamma)|\|W\|_\infty - u^2 - |u(u+\gamma)|\|W^t\|_\infty \geqslant 0 \\ &\quad |(u(u+\gamma))^{-1}|\|A_3^t\|_\infty + u^{-2}\|A_2\|_\infty - (u+\gamma)^{-2}\|A_1\|_\infty - |(u(u+\gamma))^{-1}|\|A_3\|_\infty \geqslant 0. \end{aligned} \tag{4.69}$$

Simplificando o conjunto de restrições (4.69), temos

$$\begin{aligned} &\textbf{min } f(u,\gamma) = ((u+\gamma)^2 + |u(u+\gamma)|\|W\|_\infty)(|(u(u+\gamma))^{-1}|\|A_3^t\|_\infty + u^{-2}\|A_2\|_\infty) \\ &\textbf{s.a}: \ 2u\gamma + \gamma^2 + |u(u+\gamma)|(\|W\|_\infty - \|W^t\|_\infty) \geqslant 0 \\ &\quad |(u(u+\gamma))^{-1}|(\|A_3^t\|_\infty - \|A_3\|_\infty) + u^{-2}\|A_2\|_\infty - (u+\gamma)^{-2}\|A_1\|_\infty \geqslant 0. \end{aligned}$$

Terceira possibilidade

$$\begin{aligned} &\textbf{min } F_2F_3 \\ &\textbf{s.a}: F_2 - F_1 \geqslant 0 \\ &\quad F_3 - F_4 \geqslant 0. \end{aligned} \tag{4.70}$$

Assim, temos que resolver o seguinte problema de minimização com restrições

$$\begin{aligned} &\textbf{min } f(u,\gamma) = (u^2 + |u(u+\gamma)|\|W^t\|_\infty)((u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|\|A_3\|_\infty) \\ &\textbf{s.a}: u^2 + |u(u+\gamma)|\|W^t\|_\infty - (u+\gamma)^2 - |u(u+\gamma)|\|W\|_\infty \geqslant 0 \\ &\quad (u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|\|A_3\|_\infty - |(u(u+\gamma))^{-1}|\|A_3^t\|_\infty - u^{-2}\|A_2\|_\infty \geqslant 0. \end{aligned} \tag{4.71}$$

Simplificando o conjunto de restrições (4.71), temos

$$\begin{aligned} &\textbf{min } f(u,\gamma) = (u^2 + |u(u+\gamma)|\|W^t\|_\infty)((u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|\|A_3\|_\infty) \\ &\textbf{s.a}: \ -2u\gamma - \gamma^2 + |u(u+\gamma)|(|W^t\|_\infty - |W\|_\infty) \geqslant 0 \\ &\quad (u+\gamma)^{-2}\|A_1\|_\infty + |(u(u+\gamma))^{-1}|(\|A_3\|_\infty - \|A_3^t\|_\infty) - u^{-2}\|A_2\|_\infty \geqslant 0. \end{aligned}$$

Quarta possibilidade

$$\begin{aligned} &\textbf{min } F_2F_4 \\ &\textbf{s.a}: F_2 - F_1 \geqslant 0 \\ &\quad F_4 - F_3 \geqslant 0. \end{aligned} \tag{4.72}$$

Portanto, temos que resolver o seguinte problema de minimização com restrições

$$\begin{aligned}
&\textbf{min } f(u,\gamma) = (u^2 + |u(u+\gamma)|\|W^t\|_\infty)(|(u(u+\gamma))^{-1}|\|A_3^t\|_\infty + u^{-2}\|A_2\|_\infty) \\
&\textbf{s.a}\text{: } u^2 + |u(u+\gamma)|\|W^t\|_\infty - (u+\gamma)^2 - |u(u+\gamma)|\|W\|_\infty \geqslant 0 \\
&\quad |(u(u+\gamma))^{-1}|\|A_3^t\|_\infty + u^{-2}\|A_2\|_\infty - (u+\gamma)^{-2}\|A_1\|_\infty - |(u(u+\gamma))^{-1}|\|A_3\|_\infty \geqslant 0.
\end{aligned} \tag{4.73}$$

Simplificando (4.73) segue

$$\begin{aligned}
&\textbf{min } f(u,\gamma) = (u^2 + |u(u+\gamma)|\|W^t\|_\infty)(|(u(u+\gamma))^{-1}|\|A_3^t\|_\infty + u^{-2}\|A_2\|_\infty) \\
&\textbf{s.a}\text{: } -2u\gamma - \gamma^2 + |u(u+\gamma)|(\|W^t\|_\infty - \|W\|_\infty) \geqslant 0 \\
&\quad |(u(u+\gamma))^{-1}|(\|A_3^t\|_\infty - \|A_3\|_\infty) + u^{-2}\|A_2\|_\infty - (u+\gamma)^{-2}\|A_1\|_\infty \geqslant 0.
\end{aligned}$$

4.12 Comentário dos resultados do Programa Mathematica

Todos os problemas foram abordados com o programa Mathematica 13.1. Chame $C_1 = \|W\|_\infty$, $C_2 = \|A_1\|_\infty$, $C_3 = \|A_3\|_\infty$, $C_4 = \|W^t\|_\infty$, $C_5 = \|A_2\|_\infty$ e $C_6 = \|A_3{}^t\|_\infty$

Para o problema (4.66) não conseguimos encontrar solução. Porém, se tratarmos apenas o problema **min** F_1F_3, sem as restrições, conseguimos como resposta:

$\gamma = -1$ e $u = -\dfrac{C_3}{-C_3 + \sqrt{C_1C_2C_3}}$, se $C_1C_2 > C_3$.

$\gamma = -1$ e $u = \dfrac{C_3}{C_3 + \sqrt{C_1C_2C_3}}$, se $C_1C_2 \leqslant C_3$.

Desse modo, vemos que devemos ter $\gamma = -1$. Assim, conseguimos minimizar a função objetivo com a primeira restrição, ou seja, conseguimos:

$$\begin{aligned}
&\textbf{min } F_1F_3 \\
&\textbf{s.a}\text{: } F_1 - F_2 \geqslant 0.
\end{aligned} \tag{4.74}$$

Para o problema (4.74), já com $\gamma = -1$, tivemos como resposta:

$u = -\dfrac{C_3}{-C_3 + \sqrt{C_1C_2C_3}}$ se $(C_3 < C_1C_2$ & $(C_1 = C_4 || (C_3(2 + C_1^2 + \sqrt{4 + (C_1 - C_4)^2}$ $(C_1 - C_4) + C_4^2) > 2C_1(C_2 + C_3C_4)$ &$C_1 \geqslant C_4)))||(C_3(2 + C_1^2 + C_4^2 + \sqrt{4 + (C_1 - C_4)^2}|C_1 - C_4|) < 2C_1(C_2 + C_3C_4)$ &$C_1 \neq C_4)$.

$u = \dfrac{C_3}{C_3 + \sqrt{C_1C_2C_3}}$, se $C_3 \geqslant C_1C_2$ &$C_1 > C_4$ &$C_3\sqrt{4 + (C_1 - C_4)^2}(C_1 - C_4) + 2C_1(C_2 + C_3C_4) > C_3(2 + C_1^2 + C_4^2)$.

$u = \dfrac{1}{2}$, se $C_1 = C_4$ &$C3 \geqslant C1C2$.

$u = \dfrac{-2 + C_1 + \sqrt{4 + (C_1 - C_4)^2} - C_4}{2(C_1 - C_4)}$, se $C_1 > C_4$&$(C_1^2C_3 + 2C_1(C_2 - C_3C_4) + C_3(2 + C_4^2) \leqslant C_3\sqrt{4 + (C1 - C4)^2}(C_1 - C_4)||C_3\sqrt{4 + (C_1 - C_4)^2}(C_1 - C_4) + 2C_1(C_2 + C_3C_4) \leqslant C_3(2 + C_1^2 + C_4^2))$.

$u = -\dfrac{2 - C_1 + \sqrt{4 + (C_1 - C_4)^2} + C_4}{2(C_1 - C_4)}$, se $C_3(2 + C_1^2 + \sqrt{4 + (C_1 - C_4)^2}(C_1 - C_4) + C_4^2) = 2C_1(C_2 + C_3C_4)\&C_1 > C_4$.

$u = \dfrac{2 + C_1 + \sqrt{4 + (C_1 - C_4)^2} - C_4}{2(C_1 - C_4)}$, caso contrário.

Entretanto, quando tentamos minimizar todo o problema (4.66), o programa Mathematica não consegue efetuar as contas.

Para o problema (4.68) também não encontramos resposta. Porém, ao resolvermos o problema sem restrição **min** F_1F_4, tivemos como resposta $u = 1$ e $\gamma = -1$, o que não faz sentido para o nosso problema.

Também para o problema (4.70) não encontramos solução. Ao se resolver o problema sem restrição **min** F_2F_3, tivemos como resposta $u = 0$ e $\gamma = -1$, também o que não faz sentido para o nosso caso.

Para o problema (4.72), o programa Mathematica não conseguiu resolver o problema todo, ou seja, com as duas restrições $F_2 - F_1 \geqslant 0$ e $F_4 - F_3 \geqslant 0$. Mas ao resolver **min** F_2F_4, tivemos como resposta $u = -1$ e $\gamma = 1 - \dfrac{C_6}{\sqrt{C_4C_5C_6}}$.

Assim, vemos que devemos ter $u = -1$.

Desse modo, conseguimos minimizar a função objetivo com a primeira restrição, ou seja, conseguimos:

$$\begin{aligned} &\textbf{min } F_2F_4 \\ &\textbf{s.a}: F_2 - F_1 \geqslant 0. \end{aligned} \tag{4.75}$$

Para o problema (4.75), já com $u = -1$, tivemos como resposta:

$\gamma = \dfrac{1}{2}(2 + C_1 - \sqrt{4 + (C_1 - C_4)^2} - C4)$, se $C_1 \neq C_4\&(\sqrt{4 + (C_1 - C_4)^2}C_4C_6 \geqslant 2C_4C_5 + (2 + (C_1 - C_4)^2)C_6 + C_1\sqrt{4 + (C_1 - C_4)^2}C6||2C_4C_5 + \sqrt{4 + (C_1 - C_4)^2}C_4C_6 \leqslant (2 + (C_1 - C_4)^2)C_6 + C_1\sqrt{4 + (C_1 - C_4)^2}C_6)$.

$\gamma = 1 - \dfrac{C_6}{\sqrt{C_4C_5C_6}}$, se $(2C_4C_5 + \sqrt{4 + (C_1 - C_4)^2}C_4C_6 > (2 + (C_1 - C_4)^2)C_6 + C_1\sqrt{4 + (C_1 - C_4)^2}C_6\&((C_1 < C_4\&C_4C_5 < C_6)||C_1 > C_4))||(C_1 \leqslant C_4\&C_4C_5 > C_6)$.

$\gamma = 0$, se $(C_1 = C_4\&C_4C_5 \leqslant C_6)||(C_1 < C_4\&C_4C_5 = C_6)$.

E γ indeterminado, caso contrário aos listados.

Como pode-se observar, não foi possível minimizar completamente nenhum dos casos possíveis, que são (4.66), (4.68), (4.70) e (4.72). No entanto, como pode ser verificado no Capítulo 5, encontramos empiricamente valores de escalares que resultam num número de condição melhor do que o Precondicionador Separador com a utilização da norma-∞. Desse modo, pode-se verificar que o Precondicionador Separador Duplo teve um melhor

desempenho do que o Precondicionador Separador para determinados valores de escalares. Ainda, testes com a função **fmincon** do Matlab obtiveram alguns bons resultados.

4.13 Solução do tipo $\gamma = -2\alpha\beta$ na norma infinita

Perceba, fazendo $\gamma = -2\alpha\beta$ na matriz K dada em (4.24) e na sua inversa em (4.64), temos que o número de condição na norma-∞ entre os dois precondicionadores será o mesmo. De fato, temos:

$$K = \begin{bmatrix} (\alpha\beta+\gamma)^2\mathbf{I}_m & \alpha\beta(\alpha\beta+\gamma)W \\ \alpha\beta(\alpha\beta+\gamma)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix}.$$

Assim, segue

$$\begin{aligned} K &= \begin{bmatrix} (\alpha\beta-2\alpha\beta)^2\mathbf{I}_m & \alpha\beta(\alpha\beta-2\alpha\beta)W \\ \alpha\beta(\alpha\beta-2\alpha\beta)W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} = \begin{bmatrix} (-\alpha\beta)^2\mathbf{I}_m & -(\alpha\beta)^2W \\ -(\alpha\beta)^2W^t & -(\alpha\beta)^2\mathbf{I}_{n-m} \end{bmatrix} \\ &= (\alpha\beta)^2 \begin{bmatrix} \mathbf{I}_m & -W \\ -W^t & -\mathbf{I}_{n-m} \end{bmatrix}. \end{aligned}$$

Logo

$$\|K\|_\infty = \left\| (\alpha\beta)^2 \begin{bmatrix} \mathbf{I}_m & -W \\ -W^t & -\mathbf{I}_{n-m} \end{bmatrix} \right\|_\infty = (\alpha\beta)^2 \left\| \begin{bmatrix} \mathbf{I}_m & -W \\ -W^t & -\mathbf{I}_{n-m} \end{bmatrix} \right\|_\infty.$$

Na matriz inversa K^{-1}, dada a seguir:

$$K^{-1} = \begin{bmatrix} (\alpha\beta+\gamma)^{-2}(\mathbf{I}_m + WW^t)^{-1} & (\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}W(\mathbf{I}_{n-m}+W^tW)^{-1} \\ (\alpha\beta)^{-1}(\alpha\beta+\gamma)^{-1}(\mathbf{I}_{n-m}+W^tW)^{-t}W^t & -(\alpha\beta)^{-2}(\mathbf{I}_{n-m}+W^tW)^{-1} \end{bmatrix}.$$

Temos

$$\begin{aligned} K^{-1} &= \begin{bmatrix} (\alpha\beta-2\alpha\beta)^{-2}(\mathbf{I}_m + WW^t)^{-1} & (\alpha\beta)^{-1}(\alpha\beta-2\alpha\beta)^{-1}W(\mathbf{I}_{n-m}+W^tW)^{-1} \\ (\alpha\beta)^{-1}(\alpha\beta-2\alpha\beta)^{-1}(\mathbf{I}_{n-m}+W^tW)^{-t}W^t & -(\alpha\beta)^{-2}(\mathbf{I}_{n-m}+W^tW)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} (-\alpha\beta)^{-2}(\mathbf{I}_m + WW^t)^{-1} & -(\alpha\beta)^{-2}W(\mathbf{I}_{n-m}+W^tW)^{-1} \\ -(\alpha\beta)^{-2}(\mathbf{I}_{n-m}+W^tW)^{-t}W^t & -(\alpha\beta)^{-2}(\mathbf{I}_{n-m}+W^tW)^{-1} \end{bmatrix} \\ &= (\alpha\beta)^{-2} \begin{bmatrix} (\mathbf{I}_m + WW^t)^{-1} & -W(\mathbf{I}_{n-m}+W^tW)^{-1} \\ -(\mathbf{I}_{n-m}+W^tW)^{-t}W^t & -(\mathbf{I}_{n-m}+W^tW)^{-1} \end{bmatrix}. \end{aligned}$$

Assim

$$\|K^{-1}\|_\infty = (\alpha\beta)^{-2} \left\| \begin{bmatrix} (\mathbf{I}_m + WW^t)^{-1} & -W(\mathbf{I}_{n-m}+W^tW)^{-1} \\ -(\mathbf{I}_{n-m}+W^tW)^{-t}W^t & -(\mathbf{I}_{n-m}+W^tW)^{-1} \end{bmatrix} \right\|_\infty.$$

Portanto, temos

$$cond_\infty(K) = (\alpha\beta)^2 \left\| \begin{bmatrix} \mathbf{I}_m & -W \\ -W^t & -\mathbf{I}_{n-m} \end{bmatrix} \right\|_\infty (\alpha\beta)^{-2} \left\| \begin{bmatrix} (\mathbf{I}_m + WW^t)^{-1} & -W(\mathbf{I}_{n-m} + W^tW)^{-1} \\ -(\mathbf{I}_{n-m} + W^tW)^{-t}W^t & -(\mathbf{I}_{n-m} + W^tW)^{-1} \end{bmatrix} \right\|_\infty$$

$$= \left\| \begin{bmatrix} \mathbf{I}_m & -W \\ -W^t & -\mathbf{I}_{n-m} \end{bmatrix} \right\|_\infty \left\| \begin{bmatrix} (\mathbf{I}_m + WW^t)^{-1} & -W(\mathbf{I}_{n-m} + W^tW)^{-1} \\ -(\mathbf{I}_{n-m} + W^tW)^{-t}W^t & -(\mathbf{I}_{n-m} + W^tW)^{-1} \end{bmatrix} \right\|_\infty$$

$$= cond_\infty(\text{Separador}).$$

Note ainda, para $\gamma = 0$ e $\forall\ \alpha, \beta \in \mathbb{R}^*$, tanto na norma infinita quanto em uma norma qualquer, temos que o número de condição entre os dois precondicionadores é o mesmo. De fato, tomando $\gamma = 0$ e $\forall\ \alpha, \beta \in \mathbb{R}^*$, em (4.24) e (4.64), estamos fazendo apenas um escalamento na matriz e assim o número de condição não se altera.

4.14 Combinando as normas

Ao substituir a solução ótima para a norma-2, $\gamma = -2u$ nos problemas (4.66), (4.68), (4.70) e (4.72) para as norma-1 ou norma-∞, encontramos em todos como resposta o valor de $u = -1$, assim o ponto que minimiza cada uma dessas funções é dado por $u = -1$ e $\gamma = 2$. Ou seja, a solução ótima para a norma-1 e norma-∞ dada a solução ótima na norma-2 é $\alpha\beta = -1$ e $\gamma = 2$. Note que esses valores levam à mesma matriz K do Precondicionador Separador Duplo. Porém, vale ressaltar que não encontramos um minimizador geral para o problema na norma-1 ou norma-∞.

5 Experimentos Computacionais

Neste capítulo, mostraremos alguns experimentos numéricos feitos com a abordagem proposta com a utilização da norma-∞.

Os experimentos numéricos foram feitos em um Notebook Lenovo ideapad S145 i7 - 8565U, memória RAM 8G e/ou Desktop OptiPlex 3080, intel core I3, memória RAM 16 GB e 256 GB de SSD. A resolução dos problemas foi feita por meio de uma implementação do método no programa livre Octave, versão 8.4.0 e também uma implementação no Octave, juntamente com Matlab 2020 utilizando a função **fmincon**.

Foram feitos testes em 56 problemas da coleção Netlib com 68 combinações diferentes de parâmetros. Nestes testes, foi feita uma comparação entre o número de condição do Precondicionador Separador Duplo *versus* Precondicionador Separador. Com esta finalidade, foi utilizada a fatoração de Cholesky para resolver os sistemas lineares nos métodos de pontos interiores com a abordagem proposta e assim extrair as matrizes para investigar o número de condição a cada iteração.

É bom salientarmos que não foram feitos testes com valores do tipo $\gamma = -2\alpha\beta$ ou $\gamma = 0$ e $\forall\ \alpha, \beta \in \mathbb{R}^*$, já que estes valores produzem o mesmo número de condição entre o Precondicionador Separador Duplo e o Precondicionador Separador.

5.1 Testes em problemas utilizando a norma-∞ e norma-1

Nos problemas das Tabelas 1, 2, 3, 4, 5 e 6, fizemos testes computacionais para quantificarmos o número de vezes onde o número de condição do Precondicionador Separador Duplo foi melhor do que o Precondicionador Separador.

Na primeira coluna da Tabela 1 e 2, temos o nome do problema, na segunda coluna o número de linhas e na terceira o de colunas. Na quarta coluna temos o número de iterações necessárias para a convergência do método de pontos interiores. A quinta coluna nos mostra o número de iterações onde o número de condição do Precondicionador Separador Duplo (SAPD) foi melhor. Já na sexta coluna o número de iterações onde o número de condição utilizando o Precondicionador Separador foi melhor. E na sétima coluna o resultado para se resolver o problema, note que temos ainda os seguintes símbolos nesta coluna, que são: $\star$ significa que o problema atingiu o número máximo de iterações, que no caso definimos como 100; já o símbolo $\bullet$ significa que o programa Octave excedeu o limite de 40 minutos na resolução do problema. O resultado (SAPD) significa que a abordagem proposta pelo Precondicionador Separador Duplo foi melhor do que o Precondicionador Separador, ou seja, teve mais iterações nas quais teve menor número de condição do que o

Precondicionador Separador, já o resultado (Separador) indica o contrário, e o resultado (Empate), significa que as duas abordagens tiveram a mesma quantidade de vezes de ganho do número de condição.

Para todos os problemas da Tabela 1, foram utilizados como parâmetros para o Precondicionador Separador Duplo, $\alpha = 1$, $\beta = 0,5$ e $\gamma = 0,0001$. Já no Precondicionador Separador, temos $\alpha = 1$, $\beta = 1$ e $\gamma = 0$.

Tabela 1 – Precondicionador Separador Duplo *versus* Precondicionador Separador $\alpha = 1$, $\beta = 0,5$ e $\gamma = 0,0001$.

Problema	**Nº linhas**	**Nº colunas**	**Nº Iteração**	**SAPD**	**Separador**	**Resultado**
adlittle	56	138	12	7	5	SAPD
afiro	27	51	8	5	3	SAPD
agg	488	615	16	11	5	SAPD
agg2	516	758	16	7	9	Separador
bandm	305	472	18	9	9	Empate
beaconfd	173	295	11	8	3	SAPD
blend	74	114	12	8	4	SAPD
bnl1	643	1586	34	23	11	SAPD •
capri	271	482	16	5	11	Separador
cycle	1903	3371	5	2	3	Separador •
etamacro	400	816	87	28	59	Separador
fffff800	524	1028	23	20	3	SAPD
finnis	497	1064	100	60	40	SAPD ⋆
fit1d	24	1049	70	58	12	SAPD
fit1p	627	1677	14	5	9	Separador
gfrd_pnc	616	1160	77	5	72	Separador
greenbea	2392	5598	2	1	1	Empate •
grow7	140	301	71	40	31	SAPD
grow15	300	645	82	58	24	SAPD
grow22	440	946	82	69	13	SAPD
israel	174	316	21	7	14	Separador
kb2	43	68	74	63	11	SAPD
lotfi	153	366	15	8	7	SAPD
maros	846	1966	17	12	5	SAPD •
modszk1	687	1620	55	55	0	SAPD •
perold	625	1506	37	20	17	SAPD •
pilot4	410	1123	87	52	35	SAPD •
pilotnov	975	2446	8	3	5	Separador •
recipe	91	204	57	45	12	SAPD
sc50a	50	78	8	5	3	SAPD
sc50b	50	78	7	7	0	SAPD
sc105	50	78	9	6	3	SAPD
sc205	205	317	10	7	3	SAPD
scagr7	129	185	12	7	5	SAPD
scagr25	471	671	15	4	11	Separador
scfxm1	330	600	100	51	49	SAPD ⋆
scfxm2	660	1200	82	56	26	SAPD •
scfxm3	990	1800	26	17	9	SAPD •
scrs8	490	1275	27	17	10	SAPD
scsd1	77	760	9	4	5	Separador

Tabela 1 - Continuação.

Problema	**Nº linhas**	**Nº colunas**	**Nº Iteração**	**SAPD**	**Separador**	**Resultado**
scsd6	147	1350	11	8	3	SAPD
scsd8	397	2750	6	5	1	SAPD •
sctap1	300	660	17	13	4	SAPD
sctap2	1090	2500	13	3	10	Separador •
sctap3	1480	3340	6	5	1	SAPD •
share1b	117	253	21	9	12	Separador
share2b	96	162	11	7	4	SAPD
shell	536	1777	10	10	0	SAPD
sierra	1227	2735	7	4	3	SAPD •
stair	356	614	25	17	8	SAPD
standata	359	1274	17	15	2	SAPD
standgub	361	1383	68	68	0	SAPD •
standmps	467	1274	20	18	2	SAPD
stocfor1	117	165	15	9	6	SAPD
stocfor2	2157	3045	9	4	5	Separador •
vtp_base	198	346	18	10	8	SAPD

Foram testados 56 problemas nos quais podemos fazer uma análise mais abrangente. Desses 56 problemas, 41 tiveram resultado favorável ao Precondicionador Separador Duplo (SAPD), 2 empates e apenas 13 foram desfavoráveis (Separador). Ou seja, aproximadamente 73% dos problemas tiveram um número de condição menor utilizando o (SAPD). Em apenas 23,21,% dos problemas o número de condição do Precondicionador Separador teve um número de condição menor do que o (SAPD). E tivemos em apenas 2 problemas empate entre os dois precondicionadores, veja Figura 15.

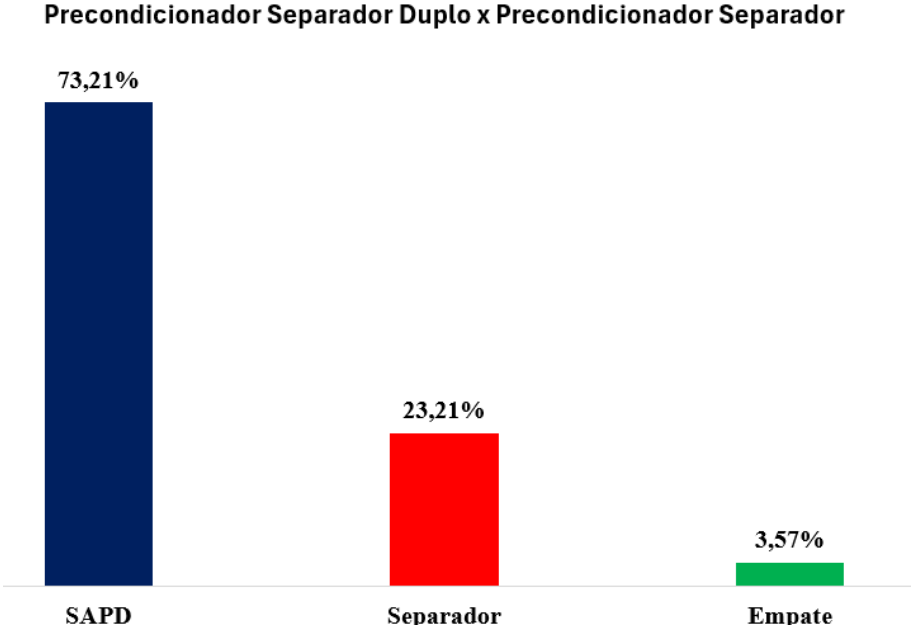

Figura 15 – SAPD x Separador
(Parâmetros $\alpha = 1$, $\beta = 0,5$ e $\gamma = 0,0001$ para SAPD.)
Fonte: Elaborado pelo autor.

Podemos ainda observar na Tabela 2 que os problemas **modszk1, sc50b, shell** e **standgub**, o Precondicionador Separador Duplo teve um número condição menor em todas as iterações.

Se considerarmos apenas os problemas onde se pode encontrar a solução, que foram 38 problemas, tivemos um ganho de quase de 74% para o Precondicionador Separador Duplo, contra menos do que 26% para o Precondicionador Separador, veja Figura 16.

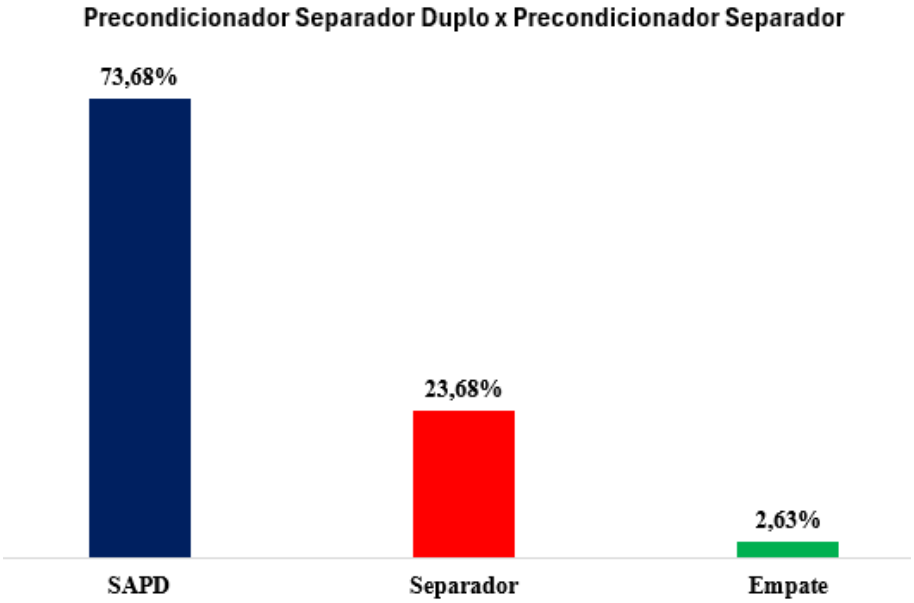

Figura 16 – SAPD x Separador: Problemas em que há convergência para a solução ótima. Fonte: Elaborado pelo autor.

Na Tabela 2, fizemos testes com os valores de $\alpha = 1$, $\beta = 1$ e $\gamma = 0,01$. Já para o Precondicionador Separador, $\alpha = 1$, $\beta = 1$ e $\gamma = 0$.

Tabela 2 – Precondicionador Separador Duplo *versus* Precondicionador Separador $\alpha = 1$, $\beta = 1$ e $\gamma = 0,01$.

Problema	**Nº linhas**	**Nº colunas**	**Nº Iteração**	**SAPD**	**Separador**	**Resultado**
adlittle	56	138	12	6	6	Empate
afiro	27	51	8	5	3	SAPD
agg	488	615	16	11	5	SAPD
agg2	516	758	16	7	9	Separador
bandm	305	472	18	9	9	Empate
beaconfd	173	295	11	7	4	SAPD
blend	74	114	12	8	4	SAPD
bnl1	643	1586	34	23	11	SAPD •
capri	271	482	16	4	12	Separador
cycle	1903	3371	5	2	3	Separador •
etamacro	400	816	87	24	63	Separador
fffff800	524	1028	23	20	3	SAPD
finnis	497	1064	100	59	41	SAPD ⋆

Tabela 2 - Continuação.

Problema	**Nº linhas**	**Nº colunas**	**Nº Iteração**	**SAPD**	**Separador**	**Resultado**
fit1d	24	1049	70	58	12	SAPD
fit1p	627	1677	14	5	9	Separador
gfrd_pnc	616	1160	77	5	72	Separador
greenbea	2392	5598	2	0	2	Separador •
grow7	140	301	71	38	33	SAPD
grow15	300	645	82	57	25	SAPD
grow22	440	946	82	67	15	SAPD
israel	174	316	21	6	15	Separador
kb2	43	68	74	62	12	SAPD
lotfi	153	366	15	8	7	SAPD
maros	846	1966	17	12	5	SAPD •
modszk1	687	1620	55	55	0	SAPD •
perold	625	1506	37	19	18	SAPD •
pilot4	410	1123	87	50	37	SAPD •
pilotnov	975	2446	8	3	5	Separador •
recipe	91	204	57	45	12	SAPD
sc50a	50	78	8	5	3	SAPD
sc50b	50	78	7	7	0	SAPD
sc105	50	78	9	5	4	SAPD
sc205	205	317	10	7	3	SAPD
scagr7	129	185	12	6	6	Empate
scagr25	471	671	15	4	11	Separador
scfxm1	330	600	100	43	57	Separador ⋆
scfxm2	660	1200	82	46	36	SAPD •
scfxm3	990	1800	26	17	9	SAPD •
scrs8	490	1275	27	16	11	SAPD
scsd1	77	760	9	2	7	Separador
scsd6	147	1350	11	8	3	SAPD
scsd8	397	2750	6	5	1	SAPD •
sctap1	300	660	17	13	4	SAPD
sctap2	1090	2500	13	3	10	Separador •
sctap3	1480	3340	6	5	1	SAPD •
share1b	117	253	21	9	12	Separador
share2b	96	162	11	7	4	SAPD
shell	536	1777	10	10	0	SAPD
sierra	1227	2735	7	4	3	SAPD •
stair	356	614	25	17	8	SAPD
standata	359	1274	17	15	2	SAPD
standgub	361	1383	68	68	0	SAPD •
standmps	467	1274	20	18	2	SAPD
stocfor1	117	165	15	9	6	SAPD
stocfor2	2157	3045	9	4	5	Separador •
vtp_base	198	346	18	10	8	SAPD

Dos 56 problemas, 38 tiveram resultado favorável ao Precondicionador Separador Duplo, 3 empates e apenas 15 foram desfavoráveis. Ou seja, mais de 67% dos problemas tiveram um número de condição menor utilizando o (SAPD). Em menos de 27% dos problemas o número de condição do Precondicionador Separador teve um número de condição menor do que o (SAPD). E tivemos empate em cerca de 5% dos problemas, veja Figura 17.

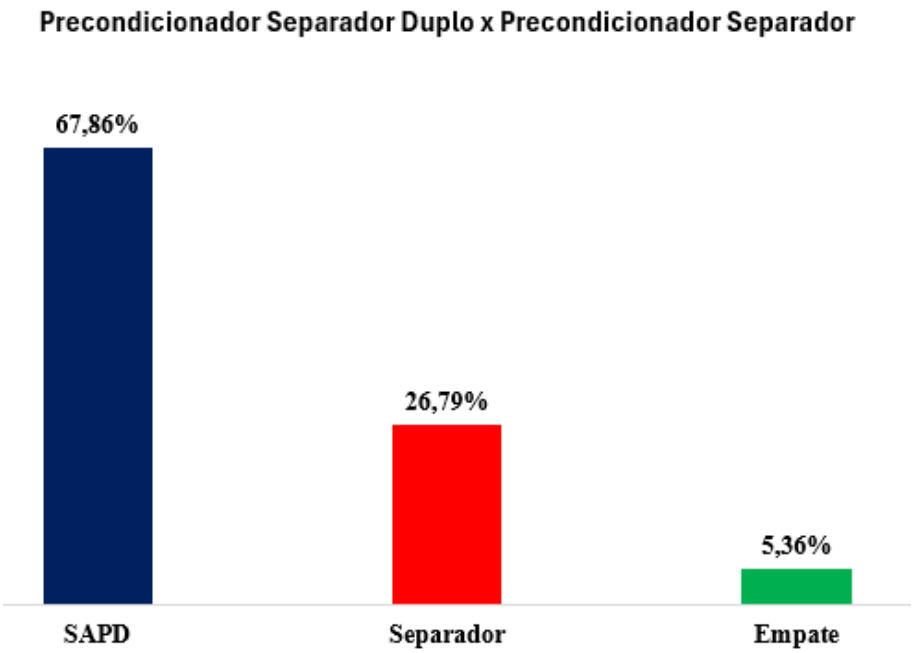

Figura 17 – SAPD x Separador
(Parâmetros $\alpha = 1$, $\beta = 1$ e $\gamma = 0,01$ para SAPD.)
Fonte: Elaborado pelo autor.

Se considerarmos apenas os problemas onde se pode encontrar a solução, que no caso foram 38 problemas, tivemos um ganho de quase 68% para o Precondicionador Separador Duplo, contra menos 24% para o Precondicionador Separador, conforme Figura 18.

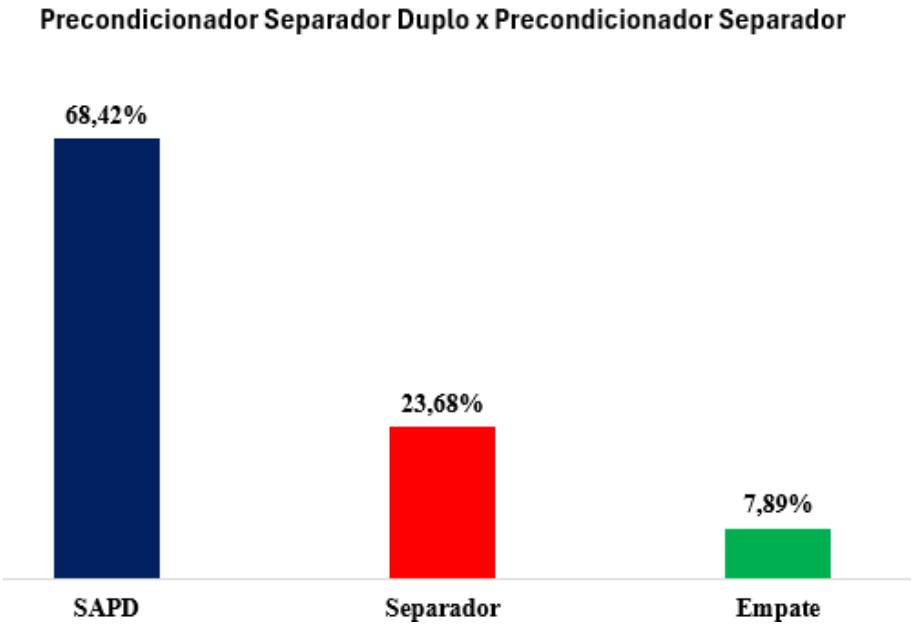

Figura 18 – SAPD x Separador: Problemas em que há convergência para a solução ótima.
Fonte: Elaborado pelo autor.

Experimentos adicionais

Fizemos diversos outros testes com os seguintes valores, sempre fixando $\alpha = 1$ e variando β e γ. O conjunto de valores foi:
$\beta \in \{-2; -1; -0,5; 0,001; 0,01; 0,1; 1; 0,5; 0,9; 1; 2; 5; 10; 50; 100\}$ e $\gamma \in \{-5; -4; -3; -2; -1; -0,01; -0,0009; 0,0001; 0,0009; 0,001; 0,009; 0,01; 0,09; 0,1; 0,9; 1; 2; 3; 4; 5\}$.

Apresentamos nas Tabelas 3 e 4 os resultados com diversos valores de β e γ.

Tabela 3 – Precondicionador Separador Duplo *versus* Precondicionador Separador para diversos parâmetros, $|\gamma| \geqslant 1$.

Parâmetros	
$\alpha = 1$ $\beta = -2$	$\gamma = 1$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -2$	$\gamma = 3$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -2$	$\gamma = 5$
SAPD	2
Empate	2
Separador	52

Parâmetros	
$\alpha = 1$ $\beta = -2$	$\gamma = -1$
SAPD	2
Empate	3
Separador	51

Parâmetros	
$\alpha = 1$ $\beta = -2$	$\gamma = -2$
SAPD	1
Empate	0
Separador	55

Parâmetros	
$\alpha = 1$ $\beta = -2$	$\gamma = -3$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -2$	$\gamma = -4$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -2$	$\gamma = -5$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -1$	$\gamma = 3$
SAPD	1
Empate	0
Separador	55

Parâmetros	
$\alpha = 1$ $\beta = -1$	$\gamma = 4$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -1$	$\gamma = 5$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -1$	$\gamma = -1$
SAPD	1
Empate	0
Separador	55

Parâmetros	
$\alpha = 1$ $\beta = -1$	$\gamma = -2$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -1$	$\gamma = -3$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -1$	$\gamma = -4$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -1$	$\gamma = -5$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = 1$	$\gamma = 1$
SAPD	1
Empate	0
Separador	55

Parâmetros	
$\alpha = 1$ $\beta = 1$	$\gamma = 2$
SAPD	0
Empate	0
Separador	56

Tabela 3 - Continuação.

Parâmetros	
$\alpha = 1 \quad \beta = 2 \quad \gamma = 1$	
SAPD	2
Empate	2
Separador	52

Parâmetros	
$\alpha = 1 \quad \beta = 2 \quad \gamma = 2$	
SAPD	1
Empate	1
Separador	54

Parâmetros	
$\alpha = 1 \quad \beta = 2 \quad \gamma = 3$	
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1 \quad \beta = 2 \quad \gamma = 4$	
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1 \quad \beta = 2 \quad \gamma = 5$	
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1 \quad \beta = 2 \quad \gamma = -1$	
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1 \quad \beta = 2 \quad \gamma = -3$	
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1 \quad \beta = 2 \quad \gamma = -5$	
SAPD	2
Empate	3
Separador	51

Pode-se observar na Tabela 3 que não houve melhora no número de condição do Precondicionador Separador Duplo com os escalares no seguinte intervalo:
$|\gamma| \geqslant 1$ e $|\beta| \geqslant 1$.

Tabela 4 – Precondicionador Separador Duplo *versus* Precondicionador Separador para diversos parâmetros, $|\gamma| < 1$.

Parâmetros	
$\alpha = 1 \quad \beta = 1 \quad \gamma = 0,9$	
SAPD	1
Empate	0
Separador	55

Parâmetros	
$\alpha = 1 \quad \beta = 1 \quad \gamma = 0,1$	
SAPD	20
Empate	2
Separador	34

Parâmetros	
$\alpha = 1 \quad \beta = 1 \quad \gamma = 0,09$	
SAPD	22
Empate	2
Separador	32

Parâmetros	
$\alpha = 1 \quad \beta = 1 \quad \gamma = 0,01$	
SAPD	38
Empate	3
Separador	15

Parâmetros	
$\alpha = 1 \quad \beta = 1 \quad \gamma = 0,009$	
SAPD	39
Empate	2
Separador	15

Parâmetros	
$\alpha = 1 \quad \beta = 1 \quad \gamma = 0,001$	
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha = 1 \quad \beta = 1 \quad \gamma = 0,0009$	
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha = 1 \quad \beta = 1 \quad \gamma = 0,0001$	
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha = 1 \quad \beta = 0,5 \quad \gamma = 0,9$	
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1 \quad \beta = 0,5 \quad \gamma = 0,1$	
SAPD	10
Empate	1
Separador	45

Parâmetros	
$\alpha = 1 \quad \beta = 0,5 \quad \gamma = 0,09$	
SAPD	12
Empate	1
Separador	43

Parâmetros	
$\alpha = 1 \quad \beta = 0,5 \quad \gamma = 0,01$	
SAPD	39
Empate	2
Separador	15

Tabela 4 - Continuação.

Parâmetros	
$\alpha=1$ $\beta=0,5$	$\gamma=0,009$
SAPD	36
Empate	3
Separador	17

Parâmetros	
$\alpha=1$ $\beta=0,5$	$\gamma=0,001$
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha=1$ $\beta=0,5$	$\gamma=0,0009$
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha=1$ $\beta=0,5$	$\gamma=0,0001$
SAPD	41
Empate	2
Separador	13

Parâmetros	
$\alpha=1$ $\beta=0,1$	$\gamma=0,009$
SAPD	21
Empate	3
Separador	32

Parâmetros	
$\alpha=1$ $\beta=0,1$	$\gamma=0,0009$
SAPD	39
Empate	2
Separador	15

Parâmetros	
$\alpha=1$ $\beta=0,9$	$\gamma=0,0009$
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha=1$ $\beta=0,01$	$\gamma=0,0009$
SAPD	21
Empate	3
Separador	32

Parâmetros	
$\alpha=1$ $\beta=0,001$	$\gamma=0,0009$
SAPD	1
Empate	0
Separador	55

Parâmetros	
$\alpha=1$ $\beta=2$	$\gamma=0,0009$
SAPD	41
Empate	2
Separador	13

Parâmetros	
$\alpha=1$ $\beta=5$	$\gamma=0,0009$
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha=1$ $\beta=10$	$\gamma=0,0009$
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha=1$ $\beta=50$	$\gamma=0,0009$
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha=1$ $\beta=100$	$\gamma=0,0009$
SAPD	40
Empate	2
Separador	14

Parâmetros	
$\alpha=1$ $\beta=-2$	$\gamma=0,0009$
SAPD	14
Empate	2
Separador	40

Parâmetros	
$\alpha=1$ $\beta=2$	$\gamma=-0,0009$
SAPD	14
Empate	2
Separador	40

Parâmetros	
$\alpha=1$ $\beta=-0,5$	$\gamma=0,01$
SAPD	12
Empate	3
Separador	41

Parâmetros	
$\alpha=1$ $\beta=0,5$	$\gamma=-0,01$
SAPD	12
Empate	3
Separador	41

Na Tabela 4 pode-se observar, em geral, a medida que diminuímos o valor de γ, aumentamos a quantidade de vezes onde o número de condição do Precondicionador Separador Duplo foi melhor. Note que para $\beta = 1$ e $\gamma = 0,09$ temos 22 resultados favoráveis, agora se considerarmos o mesmo valor de β e $\gamma = 0,009$ já temos 39 resultados favoráveis. Destaque ainda, que o número máximo que conseguimos foi de 41 resultados favoráveis na abordagem proposta, contra apenas 13 favorável ao Precondicionador Separador, que foram em $\beta = 0,5$ e $\gamma = 0,0001$ e em $\beta = 2$ e $\gamma = 0,0009$. Assim, podemos inferir que para $\gamma \in [10^{-4}; 10^{-2}]$ e $\beta \notin [10^{-2}, 10^{-1}]$ o método produziu bons resultados.

Note ainda na mesma Tabela 4, que houve diversos valores de β e γ que tiveram 40 resultados favoráveis para o Precondicionador Separador Duplo, ou seja, para vários valores de parâmetros tivemos um ganho de 71,42% dos problemas com o método sugerido.

Tabela 5 – Precondicionador Separador Duplo *versus* Precondicionador Separador para diversos parâmetros, $|\beta| < 1$ e $\gamma > 1$.

Parâmetros	
$\alpha = 1$ $\beta = 0,1$	$\gamma = 2$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = 0,01$	$\gamma = 2$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = 0,001$	$\gamma = 2$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = 0,1$	$\gamma = 10$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = 0,01$	$\gamma = 10$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = 0,001$	$\gamma = 10$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -0,1$	$\gamma = 2$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -0,01$	$\gamma = 2$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -0,001$	$\gamma = 2$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -0,1$	$\gamma = 10$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -0,01$	$\gamma = 10$
SAPD	0
Empate	0
Separador	56

Parâmetros	
$\alpha = 1$ $\beta = -0,001$	$\gamma = 10$
SAPD	0
Empate	0
Separador	56

Na Tabela 5 podemos perceber que para valores de $|\beta| < 1$ não houve ganho no número de condição do Precondicionador Separador Duplo em relação ao Precondicionador Separador.

Na Tabela 6, apresentamos uma comparação do número de condição entre os dois precondicionadores para todas as iterações no problema **agg**. Pode-se observar que, em 11 das 16 iterações (em negrito), o Precondicionador Separador Duplo apresentou um menor número de condição do que o Precondicionador Separador. A quinta coluna, Dif. absoluta (Diferença absoluta), mostra a distância entre os números de condição dos dois precondicionadores. Já a última coluna, Dif. rel (Diferença relativa), apresenta a diferença relativa entre eles. Note ainda que, pelo fato da diferença relativa ser pequena, não podemos afirmar se isso alterará o número de iterações no método de pontos interiores.

Tabela 6 – Comparativo entre os números de condições no problema **agg**, com $\alpha = 1$, $\beta = 1$ e $\gamma = 0,0009$.

N° de iterações	SAPD	Separador	Resultado	Dif. absoluta	Dif. rel (%)
1	**11315,17**	**11323,65**	**SAPD**	**8,47**	**0,07**
2	**50527466,34**	**50571187,45**	**SAPD**	**43721,01**	**0,08**
3	**77932998,44**	**77991506,48**	**SAPD**	**58508,04**	**0,07**
4	**34997683,89**	**35023065,23**	**SAPD**	**25381,33**	**0,07**
5	3396540,53	3395530,28	Separador	1010,25	0,02
6	**10552349,97**	**10560559,64**	**SAPD**	**8209,66**	**0,07**
7	**16522077,28**	**16533954,59**	**SAPD**	**11877,31**	**0,07**
8	3867678,00	3866192,52	Separador	1485,47	0,03
9	**375378,30**	**375638,88**	**SAPD**	**260,58**	**0,06**
10	**734701,69**	**735188,54**	**SAPD**	**486,85**	**0,06**
11	**338197,42**	**338438,36**	**SAPD**	**240,94**	**0,07**
12	**38908636,48**	**38929247,51**	**SAPD**	**20611,03**	0,05
13	2904377,44	2902186,55	Separador	2190,89	0,07
14	**199124,35**	**199232,31**	**SAPD**	**107,95**	**0,05**
15	522980,00	522671,37	Separador	308,63	0,05
16	3395767,24	3393392,04	Separador	2375,20	0,06

Fonte: Elaborado pelo autor.

Experimentos Computacionais com a função fmincon

Com o objetivo de obter os parâmetros ótimos para o Precondicionador Separador Duplo, resolvemos os problemas de otimização (4.66), (4.68), (4.70) e (4.72), usando a função **fmincon** do Matlab. Foram realizados testes em 56 problemas; desses, 24 foram favoráveis ao Precondicionador Separador Duplo, 27 ao Precondicionador Separador e 5 resultaram em empates, veja Figura 19.

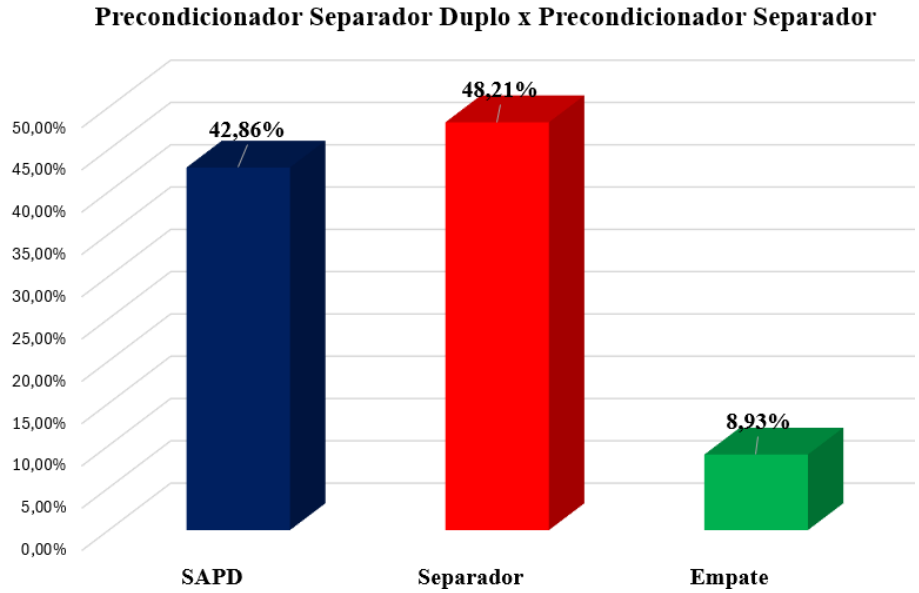

Figura 19 – SAPD x Separador com a utilização da função **fmincon**.

Vale ressaltar que o objetivo da função **fmincon** é encontrar os melhores valores de β e γ, de modo que o número de condição seja o menor possível.

O fato do precondicionador Separador Duplo não ter tido um desempenho melhor do que o Precondicionador Separador, com a utilização da função **fmincon**, pode ser pelo fato de estarmos minimizando funções que foram obtidas por meio de uma desigualdade, pois ao calcularmos a inversa da matriz K, em (4.64) utilizamos a desigualdade triangular para calcularmos a norma de cada parte da inversa de K. Não foi também possível usar uma tolerância muito pequena. Adicionalmente, esse problema deve ter mínimos locais.

Desse modo, para determinados parâmetros, podemos concluir que o Precondicionador Separador Duplo obteve uma diminuição no número de condição na norma-∞ em relação ao Precondicionador Separador.

6 Conclusões

Neste trabalho desenvolvemos uma generalização do Precondicionador Separador (OLIVEIRA, 1997), denominado Precondicionador Separador Duplo. Mostramos que o sistema aumentado oriundo do Precondicionador Separador Duplo é mais geral do que o Precondicionador Separador. Além disso, mostramos que temos uma variedade de escolhas para alguns blocos que compõem o Precondicionador Separador Duplo, e que conduzem a mesma matriz K. Desse modo, temos uma nova classe de precondicionadores.

Através dos parâmetros que fazem parte do precondicionador, podemos fazer com que a matriz precondicionada K, não tenha autovalores suficientemente próximos de zero, o que pode prejudicar o desempenho de métodos iterativos. Além do mais, através de escolhas adequadas desses parâmetros, podemos fazer com que a matriz seja muito bem condicionada nas iterações finais do método proposto. E de certo modo, temos a priori, um conhecimento sobre qual será o número de condição da matriz nas iterações finais.

Ainda, mostramos como é a aplicação dessa abordagem na resolução do sistema linear. Também foi mostrados os valores ótimos desses parâmetros que minimizam a função que fornece o número de condição na norma-2.

Conseguimos demonstrar que a utilização ótima dos parâmetros na norma-2 produz o mesmo número de condição do Precondicionador Separador. Com efeito, encontramos toda uma família de soluções, a saber, $\gamma = -2\alpha\beta$, que obtém o mesmo número de condição que o Precondicionador Separador, que portanto é ótimo para a norma-2 nesse sentido. Vale destacar, que o Precondicionador Separador, $\alpha\beta = 1$ e $\gamma = 0$, não é um caso particular da família $\gamma = -2\alpha\beta$, ou seja, encontramos toda uma família de soluções ótimas de uma maneira totalmente distinta do Precondicionador Separador.

Conseguimos demonstrar ainda que, dados os parâmetros ótimos para a norma-2, as soluções que são ótimas para a norma-∞ e norma-1; são de precondicionadores equivalente ao Separador.

Entretanto, conseguimos empiricamente, utilizando a norma-∞, verificar que o Precondicionador Separador Duplo obtém um número de condição menor do que o Precondicionador Separador. Verificou-se que, para diversas escolhas de parâmetros, aproximadamente 71% dos problemas testados apresentaram um número de condição melhor do que o obtido com o Precondicionador Separador. Essas situações tendem a ocorrer com valores de γ entre 10^{-2} e 10^{-4}.

6.1 Perspectivas Futuras

Uma perspectiva futura pode ser a realização de testes computacionais em problemas de grande porte, utilizando o Precondicionador Separador Duplo incorporado ao solver PCx e fazendo uma comparação com o Precondicionador Híbrido descrito em (BOCANEGRA, 2005). Desse modo, pode-se verificar o desempenho, principalmente em relação ao número de iterações e ao tempo de processamento para diversos valores dos parâmetros. Para essa comparação, pode-se utilizar o Perfil de Desempenho dado em (DOLAN; MORÉ, 2002).

Outra perspectiva pode ser a de tentar ver qual é a relação dos parâmetros com o tamanho do problema de Programação Linear. Ou seja, para uma determinada classe e/ou tamanho de problema, verificar se existem outros valores de parâmetros que também produzem bons resultados.

Ainda, pode-se utilizar o Precondicionador Separador Duplo no software Matlab mais recente com a função Minres. Desse modo, resolveremos efetivamente o problema com a matriz K e, assim, veremos a influência dos parâmetros na resolução do mesmo.

Também, pode-se fazer uma comparação entre o Precondicionador Separador Duplo e a abordagem direta (Cholesky), para resolução de um sistema linear oriundo dos métodos de pontos interiores.

Uma outra perspectiva é encontrar entre as soluções ótimas do Precondicionador Separador Duplo, da forma $\gamma = -2\alpha\beta$, ou seja, solução ótima na norma-2, aquela onde seus autovalores estejam melhor agrupados. Desse modo, possivelmente teremos um número menor de iterações com a utilização do método Minres, e consequentemente, um menor tempo de processamento para solução do problema.

Finalmente, pode se tentar resolver de forma analítica as quatro funções ótimas para os problemas da norma-∞ e norma-1, talvez usando o Mathematica em máquinas mais sofisticadas.

Referências

ARENALES, M.; ARMENTANO, V. *Pesquisa operacional*. [S.l.]: Elsevier Brasil, 2006.

BAZARAA, M. S.; JARVIS, J. J.; SHERALI, H. D. *Linear programming and network flows*. 4nd. ed. [S.l.: s.n.], 2010.

BOCANEGRA, S. Algoritmos de Newton-Krylov precondicionados para métodos de pontos interiores. Tese de doutorado. Universidade Federal de Minas Gerais, 2005.

BOCANEGRA, S.; CAMPOS, F. F.; OLIVEIRA, A. R. Using a hybrid preconditioner for solving large-scale linear systems arising from interior point methods. *Computational Optimization and Applications*, Springer, v. 36, n. 2-3, p. 149–164, 2007.

BOYD, S.; BOYD, S. P.; VANDENBERGHE, L. *Convex optimization*. [S.l.]: Cambridge university press, 2004.

CAMPOS, F. F. *Analysis of conjugate gradients-type methods for solving linear equations.* Tese (Doutorado) — University of Oxford, 1995.

CAMPOS, F. F.; BIRKETT, N. R. An efficient solver for multi–right-hand-side linear systems based on the cccg (η) method with applications to implicit time-dependent partial differential equations. *SIAM Journal on Scientific Computing*, SIAM, v. 19, n. 1, p. 126–138, 1998.

CASACIO, L. Aperfeiçoamento de precondicionadores para solução de sistemas lineares dos métodos de pontos interiores. Tese de doutorado, 2015.

CASTRO, C. O. Estudo do número de condição para aprimorar o cálculo da base do precondicionador separador no método de pontos interiores. Tese de doutorado. Universidade Estadual de Campinas, 2017.

DOLAN, E. D.; MORÉ, J. J. Benchmarking optimization software with performance profiles. *Mathematical programming*, Springer, v. 91, n. 2, p. 201–213, 2002.

GOLDFARB, D.; TODD, M. *Linear programming*. [S.l.], 1988.

GOLUB, G. H.; LOAN, C. F. V. *Matrix computations*. [S.l.]: JHU press, 2012.

GONDZIO, J. Interior point methods 25 years later. *European Journal of Operational Research*, Elsevier, v. 218, n. 3, p. 587–601, 2012.

KHACHIYAN, L. G. A polynomial algorithm in linear programming. In: RUSSIAN ACADEMY OF SCIENCES. *Doklady Akademii Nauk*. [S.l.], 1979. v. 244, n. 5, p. 1093–1096.

LUSTIG, I. J.; MARSTEN, R. E.; SHANNO, D. F. On implementing mehrotra's predictor–corrector interior-point method for linear programming. *SIAM Journal on Optimization*, SIAM, v. 2, n. 3, p. 435–449, 1992.

_____. Interior point methods for linear programming: Computational state of the art. *ORSA Journal on Computing*, INFORMS, v. 6, n. 1, p. 1–14, 1994.

MEHROTRA, S. On the implementation of a primal-dual interior point method. *SIAM Journal on optimization*, SIAM, v. 2, n. 4, p. 575–601, 1992.

MEYER, C. D. *Matrix analysis and applied linear algebra*. [S.l.]: Siam, 2000. v. 71.

NOCEDAL, J.; WRIGHT, S. J. Numerical optimization 2ed. Springer. New York, 2006.

OLIVEIRA, A. R. A new class of preconditioners for large-scale linear systems from interior point methods for linear programming. 1997. PhD thesis, Department of Computational and Applied Mathematics, Rice University, 1997. http://hdl.handle.net/1911/19149 .

OLIVEIRA, A. R.; SORENSEN, D. C. A new class of preconditioners for large-scale linear systems from interior point methods for linear programming. *Linear Algebra and its applications*, Elsevier, v. 394, p. 1–24, 2005.

RIBEIRO, A. A.; KARAS, E. W. *Otimização Contínua: aspectos teóricos e computacionais*. São Paulo: Cengage Learning, 2014.

SAAD, Y. *Iterative methods for sparse linear systems*. [S.l.]: SIAM, 2003.

SHEWCHUK, J. R. *An introduction to the conjugate gradient method without the agonizing pain*. Pittsburgh (PA): Carnegie-Mellon University. Department of Computer Science, 1994.

SILVA L. M. DA; OLIVEIRA, A. Modified controlled cholesky factorization for preconditioning linear systems from the interior-point method. *Computational and Applied Mathematics*, v. 40, p. 154, 2021.

SUÑAGUA, P. Uma nova abordagem para encontrar uma base do precondicionador separador para sistemas lineares no método de pontos interiores. 2014. Tese de doutorado.

SUÑAGUA, P.; OLIVEIRA, A. R. A new approach for finding a basis for the splitting preconditioner for linear systems from interior point methods. *Computational Optimization and Applications*, Springer, v. 67, p. 111–127, 2017.

TREFETHEN, L. N.; III, D. B. *Numerical linear algebra*. [S.l.]: Siam, 1997.

VANDERBEI, R. J. *Linear programming Foundations and Extensions*. 4nd. ed. [S.l.: s.n.], 2014.

VELAZCO, M. I. V.; OLIVEIRA, A. R. L. d.; CAMPOS, F. F. Heuristics for implementation of a hybrid preconditioner for interior-point methods. *Pesquisa Operacional*, SciELO Brasil, v. 31, n. 3, p. 579–591, 2011.

WRIGHT, S. J. *Primal-dual interior-point methods*. [S.l.]: SIAM, 1997.