Gabriel Cardoso de Sousa

Gradiente Conjugado Para Minimização de Quadráticas com Restrições Lineares

Florianópolis
2019

Gabriel Cardoso de Sousa

Gradiente Conjugado Para Minimização de Quadráticas com Restrições Lineares

Monografia desenvolvida para o Curso de Graduação em Matemática do Departamento de Matemática do Centro de Ciências Físicas e Matemáticas da Universidade Federal de Santa Catarina para a obtenção de grau Bacharelado em Matemática.

Universidade Federal de Santa Catarina
Curso de Graduação em Matemática do Departamento de
Matemática do Centro de Ciências
Físicas e Matemáticas

Orientador: Juliano de Bem Francisco

Florianópolis
2019

Gabriel Cardoso de Sousa

Gradiente Conjugado
Para Minimização de Quadráticas
com Restrições Lineares

Esta monografia foi julgada adequada como TRABALHO DE CONCLUSÃO DE CURSO no Curso de Matemática - Bacharelado e aprovada em sua forma final pela Banca Examinadora.

Prof. Dra. Silvia Martini de Holanda
Coordenadora do Curso

Banca Examinadora:

Prof. Dr. Juliano de Bem Francisco
Orientador

Prof. Dra. Melissa Weber Mendonça

Prof. Dr. Leonardo Silveira Borges

Resumo

O presente trabalho aborda a aplicação do método do gradiente conjugado na busca do mínimo de funções quadráticas com e sem restrições lineares. Começamos por métodos para a resolução de sistemas lineares irrestritos através de métodos diretos, estacionários e não-estacionários. Veremos os métodos de eliminação Gaussiana, Jacobi, Gauss-Seidel, gradiente e variações. Relatamos o comportamento dos métodos através da convergência e da aplicação dos métodos em sistemas lineares cuja matriz é esparsa, na forma de resultados numéricos.

Palavras-chave: sistemas lineares. gradiente. gradiente conjugado. quadráticas. esparsa.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O computador como conhecemos hoje, desenvolvido no século **XX** e aprimorado na transição para o século **XXI**, pode ser considerada uma das ferramentas mais poderosas utilizada pelo homem. Com ele fomos à lua, prevemos o tempo, conseguimos percorrer trajetórias entre dois pontos com o menor tempo possível, seja ela, terrestre, marítima ou aérea. Realizar transações bancárias através da internet com segurança. Aplicar investimentos onde o prejuízo seja mínimo e o lucro máximo, entre outras infinidades de aplicações.

Porém, estas aplicações tem um custo computacional atrelado, pois necessita que se armazene os dados, onde tais dados geralmente são massivos, de modo que é necessário um algoritmo eficiente e robusto para que após a análise, resulte em uma solução confiável e condizente com a realidade.

Quando a quantidade de entrada dos dados para estes problemas é massiva, classifica-se os dados de *BigData*. Em geral, esses dados são armazenados no computador através de matrizes e encontrar a solução do problema de otimização, consiste em resolver um, ou mais, sistemas lineares. A matriz deste sistema pode ter alguma estrutura que pode ser explorada, como por exemplo tridiagonal, simétrica, esparsa, definida positiva, etc.

Diversas maneiras podem ser empregadas para obter a solução de um sistema linear, das quais as principais são os métodos denominados de diretos, estacionários ou não-estacionários. O processo de implementação e aplicação dos métodos ocorreu entre os anos 60 e 70. Os métodos diretos foram aplicados inicialmente, devido a facilidade com que se consegue prever seu comportamento e pela sua robustez [16].

Próximo dos anos 70 ocorreu uma mudança. A utilização dos métodos iterativos estacionários e não-estacionários recebeu maior atenção, pois se percebeu que explorar o fato da matriz ser esparsa levava a uma economia no armazenamento e processamento. Em algumas situações é ainda necessário usar técnicas de pré-condicionamento para reduzir o número de iterações.

O trabalho está organizado da seguinte maneira: iniciamos o estudo no capítulo 2 através dos fundamentos da resolução de sistema lineares $\mathrm{A}\mathbf{x} = \mathbf{b}$, através dos métodos diretos, como a eliminação Gaussiana, eliminação Gaussiana com Pivoteamento e fatoração LU, pelas referências [4],[9],[10].

No capítulo 3 mencionamos os métodos estacionários como os métodos de Jacobi, Gauss-Seidel e SOR, além de mencionarmos algumas características de convergência, seguindo [4], [5], [9], [16]. No capítulo 4, introduzimos a definição e propriedades do gradiente, além de abordar o problema de minimizar a função quadrática e convexa em um conjunto irrestrito, de acordo com [3],[4], [12], [14].

No capítulo 5 desenvolvemos todas as nuances sobre o gradiente conjugado, desde as direções conjugadas até pré-condicionamento, encerrando com resultados numéricos obtidos através dos algoritmos implementados no programa MATLAB, pelas referências [7], [11], [13],[18]. Por fim, no capítulo 6 abordamos como minimizar a função quadrática e convexa sobre um conjunto definido por restrições lineares de igualdade através do gradiente conjugado, seguindo [13], [17].

2 Métodos Diretos Para Sistemas Lineares

Abordaremos inicialmente os sistemas lineares, mencionando operações elementares, sistemas equivalentes, e as classificações de um sistema linear com relação ao número de soluções.

Em seguida iremos tratar de fundamentos básicos na resolução do sistema linear $A\mathbf{x} = \mathbf{b}$ por métodos diretos. Neste caso, encontramos o vetor $\mathbf{x}$, a solução deste sistema, em uma quantidade finita de operações aritméticas.

A Eliminação Gaussiana com e sem pivoteamento e fatoração LU, são exemplos que entram nesta categoria e exporemos de forma breve sobre cada um.

2.1 Sistemas Lineares

Definição 2.1. Sejam a_{ij}, x_j e $b_i \in \mathbb{R}$, com $i \in \{1, ..., m\}$, $j \in \{1, ..., n\}$ e m,$n \in \mathbb{N}$. Um sistema linear é um sistema de equações lineares da forma,

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + & ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + & ... + a_{2n}x_n = b_2 \\ & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + & ... + a_{mn}x_n = b_m \end{cases}$$

□

Podemos interpretar o sistema acima de uma maneira matricial e é o modo como iremos adotar sempre que mencionarmos um sistema linear.

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \Leftrightarrow A\mathbf{x} = \mathbf{b}$$

com $\mathbf{x} = [x_1, x_2, ..., x_n]^T \in \mathbb{R}^n$ e $\mathbf{b} = [b_1, b_2, ..., b_m]^T \in \mathbb{R}^m$. Resolver o sistema linear $A\mathbf{x} = \mathbf{b}$ é encontrar um vetor $\mathbf{x}$ de modo a satisfazer as m linhas do sistema da **Definição 2.1**.

Exemplo 2.2. Sejam os sistemas

$$\begin{cases} 3x_1 + 2x_2 = 5 \\ 2x_1 + x_2 = 2 \end{cases} \qquad \begin{cases} 3x_1 + 2x_2 + x_3 = 5 \\ x_2 = 2 \end{cases} \qquad \begin{cases} 2x_1 + 3x_2 = 6 \\ 2x_1 + 3x_2 = 5 \end{cases}$$

As soluções são $\mathbf{x} = (-1, 4)^T$, $\mathbf{x} = ((1 - x_3)/3, 2, x_3)^T$ para qualquer $x_3 \in \mathbb{R}$ e o último não admite solução. Note que para o primeiro sistema temos uma única solução e para o segundo temos infinitas soluções. Segue a seguinte definição:

Definição 2.3. O sistema linear $A\mathbf{x} = \mathbf{b}$ é compatível e determinado se existe apenas uma única solução. É compatível e indeterminado se existem infinitas soluções e incompatível se não admite soluções. □

Uma maneira de obter a solução e que vai ser útil mais adiante são os sistemas equivalentes obtidos através de operações elementares sobre o sistema $A\mathbf{x} = \mathbf{b}$.

Definição 2.4. Realizamos operações elementares sobre o sistema $A\mathbf{x} = \mathbf{b}$ se efetuarmos uma ou mais operações do seguinte tipo:

a) A troca da i-ésima linha pela k-ésima linha denotaremos por $L_i \leftrightarrow L_k$

b) Multiplicar a i-ésima linha por uma constante $c \in \mathbb{R}\backslash\{0\}$, denotaremos por $L_i \leftarrow cL_i$.

c) Somar na i-ésima linha a k-ésima linha multiplicada por uma constante diferente de zero, que será denotada por $L_i \leftarrow L_i + cL_k$. □

Definição 2.5. O sistema $B\mathbf{x} = \mathbf{c}$ é dito ser equivalente ao sistema $A\mathbf{x} = \mathbf{b}$, se podemos obter o sistema $B\mathbf{x} = \mathbf{c}$ aplicando operações elementares às equações do sistema $A\mathbf{x} = \mathbf{b}$. □

Mostra-se que sistemas equivalentes tem o mesmo conjunto solução [1]. Uma ideia para resolver sistemas lineares é realizar operações elementares em um sistema linear de modo a obter um sistema de fácil resolução. Esta técnica é conhecida por eliminação gaussiana, ou método de Gauss.

2.2 Eliminação Gaussiana

Definição 2.6. Seja o sistema linear quadrado $A\mathbf{x} = \mathbf{b}$ com $A \in \mathbb{R}^{n\times n}$. A matriz aumentada com relação ao sistema linear é denotada por $[\ A \mid b\ ]$ é a matriz

$$[\ A \mid b\ ] = \left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & b_n \end{array}\right]$$

□

Desse modo considere $[\ A \mid b\ ]$ e suponha que $a_{11} \neq 0$. Para zerar a primeira coluna de A abaixo do coeficiente a_{11}, devemos fazer $L_2 \leftarrow L_2 + (-a_{21}/a_{11})L_1$. Logo, temos que a_{21} recebe $a_{21} + (-a_{21}/a_{11})a_{11} = 0$ e denotaremos por $a_{21} := a_{21} + (-a_{21}/a_{11})a_{11} = 0$.

Na terceira linha fazemos $L_3 \leftarrow L_3 + (-a_{31}/a_{11})L_1$, e assim $a_{31} = 0$. Procedendo dessa maneira $n-1$ vezes obtemos a matriz aumentada

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ 0 & \times & \vdots & \times & \times \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \times & \dots & \times & \times \end{array}\right]$$

onde $\times$ representa um elemento qualquer. Agora zeramos abaixo do elemento que está na posição da segunda coluna e segunda linha que denotaremos por $(2,2)$, chamado de pivô.

Caso o elemento seja zero, fazemos uma troca da linha por uma outra linha de modo a obter um elemento diferente de zero nesta posição. Note que $a_{22} := a_{22} + (-a_{21}/a_{11})a_{12}$ e $a_{32} := a_{32} + (-a_{31}/a_{11})a_{13}$.

Desse modo, utilizamos a linha L_2 para zerar abaixo do elemento da posição $(2,2)$, caso contrário perderíamos os zeros abaixo da posição $(1,1)$. Assim $L_3 \leftarrow L_3 + (-a_{32}/a_{22})L_2$. Após $n-2$ vezes obtemos a matriz

$$\left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ 0 & \times & \times & \vdots & \times & \times \\ \vdots & 0 & \times & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \times & \dots & \times & \times \end{array}\right]$$

Agora zeramos abaixo do elemento na posição $(3,3)$ de modo análogo como feito anteriormente. Continuando com estes passos iremos obter um sistema triangular superior da seguinte forma

$$\left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ & d_{22} & \times & \vdots & \times & \times \\ & & d_{33} & \vdots & \vdots & \vdots \\ & 0 & & \ddots & \times & \vdots \\ & & & & d_{nn} & c_n \end{array}\right] \tag{2.1}$$

Desse modo, se $d_{nn} \neq 0$ (caso contrário existem infinitas soluções ou nenhuma solução), pela última linha $d_{nn}x_n = c_n$. Ou seja, $x_n = c_n/d_{nn}$. Pela penúltima linha temos $b_{(n-1)(n-1)}x_{n-1} + d_{(n-1)n}x_n = c_{n-1}$.

Logo, por substituição $x_{n-1} = (c_{n-1} - d_{(n-1)n}(c_n/d_{nn}))/d_{(n-1)(n-1)}$. Dessa forma, conhecendo x_{n-1} e x_n encontramos x_{n-2} por substituição. Continuando o processo obtemos o vetor solução $\mathbf{x}$. Esta técnica é conhecida como substituição regressiva (ou *back-substitution*). Desse modo temos o seguinte algoritmo da eliminação Gaussiana, [9]

```
1  Algoritmo 2.1
2  Dados A, b
3  Para k=1:(n-1) faça
4       Para k=i:n faça
5            Se a_ki ≠ 0 então
6               p = k;
7            Senão se, k=n
8               Não existe solução única.
9            Fim
10      Fim
11      Se p ≠ i então
12         L_p ↔L_i
13      Fim
14      Para j=(i+1):n faça
15           m = a_ji/a_ii
16           L_j ←L_j − mL_i
17      Fim
18 Fim
19 x(n) = a_nn+1/a_nn
20 Para i=(n-1):1 faça
21      x(i) = ( a_in+1 − Σ_{j=i+1}^{n} a_ij x(j) ) /a_ii
22 Fim
```

Observe que para encontrar a solução do sistema, o processo computacional necessário tem custo de $\frac{2n^3}{3} + \frac{3n^2}{2} - \frac{7n}{6}$ operações ([9], p.356), onde n é a ordem da matriz A.

Porém existem sistemas onde o método apresenta a desvantagem de propagar erro de arredondamento e dessa forma a solução obtida é uma aproximação da solução exata.

Exemplo 2.7. Seja o sistema linear ([4],p.37)

$$\begin{bmatrix} 0.004 & 15.73 \\ 0.423 & -24.72 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 15.77 \\ -20.49 \end{bmatrix}$$

Aplicando eliminação Gaussiana utilizando apenas 4 casas decimais e arredondando a quinta casa, obtemos a solução $\mathbf{x} = [12.50, 0.9994]^T$.

Porém a solução exata é $\mathbf{x} = [10.0, 1.0]^T$. Para contornar este problema utilizamos a eliminação Gaussiana com pivoteamento.

2.3 Eliminação Gaussiana com Pivoteamento

Procedemos da mesma maneira que na eliminação Gaussiana adicionando o seguinte passo. Antes de zerar abaixo do pivô, percorremos a coluna do pivô em busca do máximo número em módulo.

Ou seja, dado o pivô da linha l e coluna k, a_{lk}, percorremos a coluna k realizando comparações em busca do máximo em módulo, $max\ |a_{rk}|$ para $l \leq r \leq n$. Encontrado o máximo, efetuamos a operação elementar de troca de linhas $L_l \leftrightarrow L_r$ e continuamos o processo da eliminação Gaussiana normalmente.

Exemplo 2.8. No mesmo sistema do **Exemplo 2.7** aplicando pivoteamento temos que $max\{|0.004|, |0.423|\} = |0.423|$. E aplicando a operação elementar $L_1 \leftrightarrow L_2$ o sistema fica

$$\left[\begin{array}{cc|c} 0.423 & -24.72 & -20.49 \\ 0.004 & 15.73 & 15.77 \end{array}\right]$$

Aplicando $L_2 \leftarrow L_2 + cL_1$ com $c = -0.9456 \times 10^{-2}$ ficamos com o sistema triangular abaixo e temos a solução exata.

$$\left[\begin{array}{cc|c} 0.423 & -24.72 & -20.49 \\ 0 & 15.9637 & 15.9637 \end{array}\right]$$

Desse modo o algoritmo fica:

```
1  Algoritmo 2.2
2  Dados A, b
3  Para k=1:(n-1) faça
4        v = |a_kk|
5        Para j=k:n faça
6              Se |a_jk| > v então
7                  v = |a_jk|
8                  p = j
9              Fim
10       Fim
11       L_i ↔ L_p
12       Para i=(k+1):n faça
13             m = a_ik/a_kk
14             b(i) = b(i) − mb(k)
15             Para j=(k+1):n faça
16                   a_ij = a_ij − ma_kj
17             Fim
18       Fim
19 Fim
20 x(n) = a_nn+1/a_nn
21 Para i=(n-1):1 faça
22       x(i) = ( a_in+1 − Σ_{j=i+1}^{n} a_ij x(j) ) /a_ii
23 Fim
```

2.4 Fatoração LU

Seja a matriz A quadrada de dimensão n. Uma outra possibilidade de resolver o sistema $A\mathbf{x} = \mathbf{b}$ é através da decomposição da matriz A.

Decompomos A em uma matriz triangular superior U, e em uma matriz triangular inferior L, de tal modo que $A = LU$. Esta fatoração é chamada de fatoração LU de A. A Figura 1 ilustra um exemplo da decomposição LU de uma matriz A, onde os pontos são elementos não nulos.

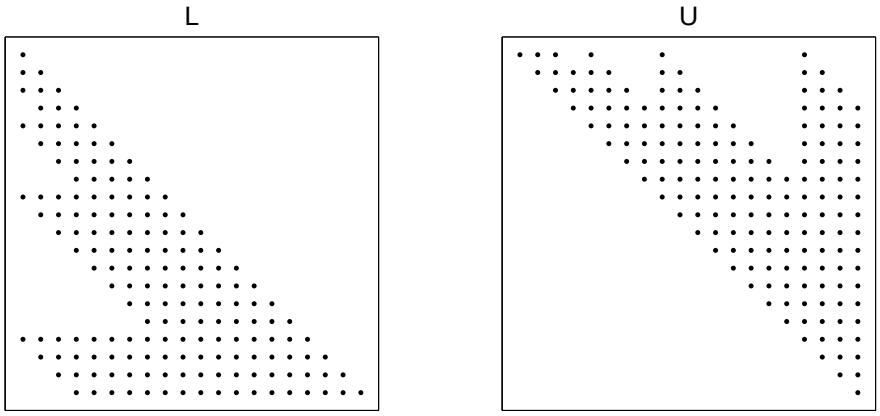

Figura 1 – Fatoração LU

Desse modo, resolver o sistema $A\mathbf{x} = \mathbf{b}$ é equivalente a resolver o sistema $LU\mathbf{x} = \mathbf{b}$ em duas etapas. Primeiro definimos $U\mathbf{x} = \mathbf{y}$ e resolvemos $L\mathbf{y} = \mathbf{b}$.

Encontrado o vetor solução $\mathbf{y}$, resolvemos $U\mathbf{x} = \mathbf{y}$. Para resolver estes dois sistemas basta realizar substituição pois se trata de sistemas triangulares.

Definição 2.9. A matriz $A \in \mathbb{R}^{n\times n}$ está na forma escalonada reduzida se satisfizer os três itens:

a) Os pivôs tem valor 1.

b) Considere uma linha i qualquer que não tem todos os elementos nulos. Então o número de zeros que precede o primeiro elemento não nulo na linha $i+1$ é maior que na linha i

c) Se existe alguma linha somente com zeros, ela está abaixo de qualquer outra linha que contenha elementos não nulos.

□

Exemplo 2.10. Exemplo de forma escalonada reduzida:

$$\begin{bmatrix} 1 & 5 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Chegamos neste formato realizando operações elementares como na eliminação Gaussiana. Uma outra forma de enxergar as operações elementares é através das matrizes elementares.

Definição 2.11. Uma matriz $E \in \mathbb{R}^{n\times n}$ é elementar se pode ser obtida após a realização de alguma das três operações elementares em uma matriz identidade. □

Exemplo 2.12. A matriz abaixo é elementar pois foi obtida da matriz identidade pela operação $L_2 \leftarrow L_2 + 2L_3$.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

O interessante é notar que realizar a operação elementar $L_2 \leftarrow L_2+2L_3$ em uma matriz A é equivalente a multiplicar a matriz E à esquerda de A, isto é EA. Com estas considerações temos o **Lema 2.13** a seguir, cuja prova pode ser encontrada em [2], e na sequência o **Teorema 2.14** que garante a existência dos fatores L e U da matriz A.

Lema 2.13. *Seja $E \in \mathbb{R}^{n\times n}$ uma matriz elementar qualquer e $L_1, L_2 \in \mathbb{R}^{n\times n}$ triangulares inferiores quaisquer. Então a matriz E é inversível, a multiplicação L_1L_2 é triangular inferior, e a L_1^{-1} é triangular inferior.* □

Teorema 2.14. *Seja a matriz $A \in \mathbb{R}^{n\times n}$. Se através de operações elementares conseguimos A na forma escalonada reduzida sem realizar troca de linhas, então existe $L \in \mathbb{R}^{n\times n}$ e $U \in \mathbb{R}^{n\times n}$ tal que $A = LU$.*

Dem. Note que pela eliminação Gaussiana vista anteriormente chegamos no sistema triangular superior (2.1). Ou seja, existem matrizes elementares $E_1, E_2, ..., E_k$ de modo que $E_kE_{k-1}...E_1A = U$. Pelo **Lema 2.13** existe a inversa de matrizes elementares, logo obtemos

$$A = E_1^{-1}E_2^{-1}...E_k^{-1}U.$$

Ou ainda $L = E_1^{-1}E_2^{-1}...E_k^{-1}$ e $U = E_kE_{k-1}...E_1A$. Vamos ver que L é triangular inferior e U é triangular superior. Note que se aplicar a eliminação Gaussiana em A, sem realizar troca de linhas, chegamos no sistema (2.1) e continuando o processo através da operação elementar $L_i \leftarrow cL_i$ para obter pivôs unitários, temos que U é triangular superior.

Por último, como as operações elementares realizadas são do tipo b) ou c), (ver **Definição 1.4**), ao aplicar o item c) a mudança é sempre abaixo da diagonal; logo temos que $E_1, ..., E_k$ são triangulares inferiores. Assim a inversa também será, pelo **Lema 2.13**. Agora, multiplicação de matrizes triangulares inferiores ainda é triangular inferior, através do **Lema 2.13**. Portanto, L é triangular inferior. □

Exemplo 2.15. Vamos calcular a decomposição LU da matriz

$$A = \begin{bmatrix} 2 & -2 & -2 \\ 0 & -2 & 2 \\ -1 & 5 & 2 \end{bmatrix}$$

Aplicando $L_3 \leftarrow L_3 + (1/2)L_1$ e armazenando as constantes sempre que realizarmos a operação elementar b) ou c) conforme o Teorema acima, temos $l_{31} = 1/2$ e

$$\begin{bmatrix} 2 & -2 & -2 \\ 0 & -2 & 2 \\ 0 & 4 & 1 \end{bmatrix}$$

Fazendo $L_3 \leftarrow L_3 + (2)L_2$ temos $l_{32} = 2$ com

$$U = \begin{bmatrix} 2 & -2 & -2 \\ 0 & -2 & 2 \\ 0 & 0 & 5 \end{bmatrix}$$

E a L fica:

$$\begin{bmatrix} 1 & 0 & 0 \\ -l_{21} & 1 & 0 \\ -l_{31} & -l_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1/2 & -2 & 1 \end{bmatrix} = L$$

Note que ocorre a troca do sinal pois L é a multiplicação de inversas de matrizes elementares. A fatoração LU de uma matriz A quadrada tem um custo computacional de $2n^2$ operações ([4], pg.40). Segue abaixo o algoritmo:

1 Algoritmo 2.3
2 Dados A, $L = I_n$ (identidade de ordem n).
3 Para j=2:n faça
4 $u_{1j} = a_{1j}/l_{11}$
5 $l_{j1} = a_{j1}/u_{11}$
6 Fim
7 Para i=2:(n-1) faça
8 Para j=(i+1):n faça
9 $u_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\right)$
10 $l_{ji} = (1/u_{ii})\left(a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{ki}\right)$
11 Fim
12 Fim

3 Métodos Estacionários

Os métodos iterativos são caracterizados por resolver o sistema $A\mathbf{x} = \mathbf{b}$ através de iterações que se aproximam da solução original a cada iteração feita, a partir de um ponto inicial. Isto é, o método gera uma sequência de números reais que pode ou não convergir dependendo da matriz do sistema.

Neste capítulo falaremos sobre Jacobi, Gauss-Seidel e SOR, além de mostrarmos alguns resultados sobre a convergência.

3.1 Método de Jacobi e Gauss-Seidel

Seja $A \in \mathbb{R}^{n\times n}$ em que cada elemento da diagonal seja não-nulo, isto é, $a_{ii} \neq 0, \forall i \in \{1, 2, .., n\}$. Considere o sistema linear $A\mathbf{x} = \mathbf{b}$, com $\mathbf{x}$,$\mathbf{b} \in \mathbb{R}^n$, na forma explícita:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + & ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + & ... + a_{2n}x_n = b_2 \\ & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + & ... + a_{nn}x_n = b_n \end{cases}$$

Podemos reescrever cada linha isolando x_i do coeficiente a_{ii}, no lado esquerdo da igualdade, ou seja,

$$\begin{aligned} x_1 &= \left(b_1 - \left(\sum_{j=2}^{n} a_{1j}x_j\right)\right)/a_{11} \\ x_2 &= \left(b_2 - \left(\sum_{\substack{j=1 \\ j\neq 2}}^{n} a_{2j}x_j\right)\right)/a_{22} \\ &\vdots \\ x_n &= \left(b_2 - \left(\sum_{j=1}^{n-1} a_{nj}x_j\right)\right)/a_{nn} \end{aligned} \tag{3.1}$$

Suponha agora que queiramos resolver o sistema $A\mathbf{x} = \mathbf{b}$ e ao invés de encontrar a solução por algum método direto, nos aproximamos da solução através da equação

(3.1). Considere um vetor inicial que chamaremos de $\mathbf{x}^{(0)} = [x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}]^T$ e aplique a equação (3.1) com o vetor inicial $\mathbf{x}^{(0)}$ e $\mathbf{b} = [b_1, b_2, ..., b_n]^T$.

Como resultado, teremos o vetor $\mathbf{x}^{(1)} = [x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)}]^T$. Fazendo novamente o processo temos $\mathbf{x}^{(2)} = [x_1^{(2)}, x_2^{(2)}, ..., x_n^{(2)}]^T$. O ponto central é que se continuarmos repetindo este método, obtemos sequências de vetores de modo que $\{x_1^{(i)}\} \to x_1$, $\{x_2^{(i)}\} \to x_2$, ..., $\{x_n^{(i)}\} \to x_n$, isto é, pode convergir para a solução do sistema $\mathbf{x}$.

Decomponha a matriz A nas matrizes L,D e U de modo que $A = L + D + U$. Neste caso,

$$L = \begin{bmatrix} 0 & \dots & \dots & 0 \\ a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn-1} & 0 \end{bmatrix} \quad D = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_{nn} \end{bmatrix} \quad U = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1n} \\ 0 & \dots & \dots & 0 \end{bmatrix} \tag{3.2}$$

Definição 3.1. Um método iterativo é estacionário se pode ser escrito na forma

$$\mathbf{x}^{(k+1)} = M\mathbf{x}^{(k)} + c.$$

Onde $\mathbf{c} \in \mathbb{R}^n$ e M é a matriz das iterações escrita em função de L,D ou U. □

Um método iterativo estacionário mais simples é o Método de Jacobi. Neste caso

$$M = -D^{-1}(L + U) \text{ e } \mathbf{c} = D^{-1}\mathbf{b}.$$

Isto é, $D\mathbf{x}^{(k+1)} = -(L + U)\mathbf{x}^{(k)} + \mathbf{b}$. A seguir apresentamos o pseudo-código ([4], pg.59):

```
1 Algoritmo 3.1
2 Dados A,b,x^(0), tol (aproximação da solução)
3 k = 0;
4 Enquanto ||Ax_i^(k) − b|| > tol faça
5          Para i=1:n faça
6                x_i^(k+1) = (1/a_ii) ( b_i − ( Σ_{j=1,j≠i}^{n} a_ij x_j^(k) ) ) /a_nn
7                k=k+1
8          Fim
9 Fim
```

Já no método de Gauss-Seidel aplicamos a equação (3.1) mas com a diferença de que utilizamos as componentes recém calculadas e não todo o vetor calculado na iteração

anterior. Por exemplo, com $n = 3$, iniciamos com a aproximação $\mathbf{x}^{(0)} = [x_1^{(0)}, x_2^{(0)}, x_3^{(0)}]^T$. Para a primeira iteração temos

$$\begin{aligned} x_1^{(1)} &= (b_1 - (a_{12}x_2^{(0)} + a_{13}x_3^{(0)}))/a_{11} \\ x_2^{(1)} &= (b_2 - (a_{21}x_1^{(1)} + a_{23}x_3^{(0)}))/a_{22} \\ x_3^{(1)} &= (b_3 - (a_{31}x_1^{(1)} + a_{32}x_2^{(1)}))/a_{33} \end{aligned}$$

Note que para calcularmos $x_2^{(1)}$ já utilizamos a aproximação mais recente disponível que é $x_1^{(1)}$. Ao contrário do método de Jacobi que utiliza todas as componentes de $\mathbf{x}^{(0)}$.

Assim, no método Gauss-Seidel $M = -(L + D)^{-1}U$ e $\mathbf{c} = (L + D)^{-1}\mathbf{b}$

$$\begin{aligned} A\mathbf{x} = \mathbf{b} &\Leftrightarrow (L + D + U)\mathbf{x} = \mathbf{b} \\ &\Leftrightarrow (L + D)\mathbf{x} = -U\mathbf{x} + \mathbf{b} \\ &\Leftrightarrow \mathbf{x}^{(k+1)} = -(L + D)^{-1}U\mathbf{x}^{(k)} + (L + D)^{-1}\mathbf{b} \end{aligned}$$

Segue abaixo o algoritmo [4]:

1 Algoritmo 3.2
2 Dados A,b,$x^{(0)}$, tol (aproximação da solução)
3 k = 0;
4 Enquanto $||Ax_i^{(k)} - b|| > tol$ faça
5 Para i=1:n faça
6 $$x_i^{(k+1)} = (1/a_{ii}) \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \right)$$
7 Fim
8 k=k+1
9 Fim

3.2 Método SOR

Antes de falarmos sobre o método SOR vamos ver algumas definições e resultados. Começaremos por autovalores com o intuito de definir raio espectral.

Definição 3.2. Seja a matriz $A \in \mathbb{R}^{n\times n}$. Os autovalores da matriz A são as raízes do polinômio $p(\lambda) = det(A - \lambda I_n)$, chamado polinômio característico da matriz A. □

Definição 3.3. O vetor não nulo $\mathbf{x}$ que satisfaz $(A - I_n\lambda)\mathbf{x} = 0$ é chamado de autovetor de A associado ao autovalor λ, onde I_n é a matriz identidade de ordem n. □

Definição 3.4. O raio espectral da matriz A é dado por $\rho(A) = max|\lambda|$ onde λ são os autovalores de A. □

Exemplo 3.5. Seja

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 3 \\ 0 & 1 & 1 \end{bmatrix}$$

O polinômio característico é $p(\lambda) = det(A - I_3\lambda) = (2 - \lambda)(1 - \lambda)(1 - \lambda)$, com autovalores $\lambda_1 = 2$, $\lambda_2 = \lambda_3 = 1$. Para encontrar os autovetores, resolvemos

$$(A - I_3\lambda_i)\mathbf{x} = \mathbf{0}, \quad i \in \{1, 2, 3\}.$$

E encontramos $\mathbf{x}_1 = [0, 1, 1]^T$, $\mathbf{x}_2 = [0, 0, 1]^T$ e $\mathbf{x}_3 = [1, 0, 0]^T$. Observe que a solução do sistema não é única. E ainda temos que $\rho(A) = max\{|2|, |1|\} = 2$.

Definiremos agora o que é uma matriz convergente com o propósito de mostrar a relação entre os autovalores da matriz de iteração M e a sequência gerada pelo método iterativo $\mathbf{x}^{(k+1)} = M\mathbf{x}^{(k)} + c$ convergir.

Definição 3.6. Dizemos que a matriz $A \in \mathbb{R}^{n\times n}$ é convergente se $\lim_{k\to\infty} A^k = 0$. □

Exemplo 3.7. Temos

$$A = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}, \qquad A^k = \begin{bmatrix} (1/2)^k & 0 \\ 0 & (1/2)^k \end{bmatrix}$$

Logo, como $\lim_{k\to\infty} (1/2)^k = 0$, segue que A é convergente.

Utilizaremos o seguinte resultado cuja prova pode ser encontrada em [8] e que nos garante o seguinte:

Teorema 3.8. *Dada a matriz A, temos as seguintes equivalências: $\rho(A) < 1 \Leftrightarrow \lim_{k\to\infty} A^k \boldsymbol{x} = 0$ para qualquer $\boldsymbol{x} \in \mathbb{R}^n \Leftrightarrow A$ é convergente.* □

Proposição 3.9. *Seja $M \in \mathbb{R}^{n\times n}$ tal que $\rho(M) < 1$. Então:*

a) $(I_n - M)^{-1}$ *existe;*

b) $(I_n - M)^{-1} = I_n + M + M^2 + ... = \sum_{k=0}^{\infty} M^k$.

Dem. Vamos ver o item *a*). Seja λ autovalor associado a $\mathbf{x}$ com relação à matriz M. Então $M\mathbf{x} = \lambda\mathbf{x}$. Ou, $-M\mathbf{x} = -\lambda\mathbf{x} \Leftrightarrow \mathbf{x} - M\mathbf{x} = \mathbf{x} - \lambda\mathbf{x} \Leftrightarrow (I_n - M)\mathbf{x} = (1-\lambda)\mathbf{x}$. Logo, λ ser autovalor de M é equivalente a $(1-\lambda)$ ser autovalor de $(I_n - M)$.

Por hipótese $\rho(M) < 1$, ou seja $|\lambda| \leq \rho(M) < 1$. Isto significa que $\lambda = 1$ não pode ser autovalor de M. Equivalentemente, 0 não pode ser autovalor de $I_n - M$. Portanto, existe $(I_n - M)^{-1}$.

Vamos ver agora o item *b*). Considere a soma finita $S_m = I_n + M + ... + M^m$. Logo $(I_n - M)S_m = (I_n - M)(I_n + M + ... + M^m) = (I_n + M + ... + M^m) - (M + M^2 + ... + M^{m+1}) = I_n - M^{m+1}$. Observe que pela hipótese temos $\rho(M) < 1$. Segue do **Teorema 3.8** que $\lim_{m\to\infty}(I_n - M)S_m = \lim_{m\to\infty}(I_n - M^{m+1}) = I_n$.

Por fim, $(I_n - M)^{-1}\left[\lim_{m\to\infty}(I_n - M)S_m = I_n\right] \Leftrightarrow (I_n - M)^{-1} = \lim_{m\to\infty} S_m = \sum_{m=0}^{\infty} M^m$ □.

Teorema 3.10. *Seja $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ qualquer. Considere a sequência $\{\boldsymbol{x}^k\}_{k\in N}$ gerada por $\boldsymbol{x}^{k+1} = M\boldsymbol{x}^k + c$. A sequência $\{\boldsymbol{x}^k\}_{k\in N}$ converge para uma única solução $\boldsymbol{x}$, se e somente se, $\rho(M) < 1$.*

Dem. Suponha $\rho(M) < 1$. Então para a iteração $(k-1)$ temos

$$\begin{aligned}
\mathbf{x}^k &= M\mathbf{x}^{k-1} + c \\
&= M(M\mathbf{x}^{k-2} + c) + c \\
&= M^2\mathbf{x}^{k-2} + (M + I_n)\mathbf{c} \\
&= M^2(M\mathbf{x}^{k-3} + \mathbf{c}) + (M + I_n)\mathbf{c} \\
&= M^3\mathbf{x}^{k-3} + M^2\mathbf{c} + (M + I_n)\mathbf{c} \\
&= M^3\mathbf{x}^{k-3} + (M^2 + M + I_n)\mathbf{c} \\
&\vdots \\
&= M^k\mathbf{x}^0 + (M^{k-1} + ... + I_n)\mathbf{c}.
\end{aligned}$$

Como M é convergente, pelo **Teorema 3.8**, temos $\lim_{k\to\infty} M^k\mathbf{x}^0 = 0$. Pela **Proposição 3.9**, temos que $\lim_{k\to\infty} \mathbf{x}^k = \lim_{k\to\infty} M^k\mathbf{x}^0 + \lim_{k\to\infty}(M^{k-1} + ... + I_n)\mathbf{c} = \lim_{k\to\infty} M^k\mathbf{x}^0 + \left[\sum_{k=0}^{\infty} M^k\right]\mathbf{c} = 0 + (I_n - M)^{-1}\mathbf{c}$. Isto é, $\{\mathbf{x}^k\}_{k\in N}$ converge para $(I_n - M)^{-1}\mathbf{c} = \mathbf{x}$. E a solução é única pela unicidade da convergência da sequência.

Suponha agora que $\lim_{k\to\infty} \mathbf{x}^k = \mathbf{x}$, ou seja, $\mathbf{x} = M\mathbf{x}+\mathbf{c}$. Vamos mostrar que $\lim_{n\to\infty} M^n\mathbf{z} = 0$ para qualquer $\mathbf{z}$ em $\mathbb{R}^n$. Seja $\mathbf{z} = \mathbf{x} - \mathbf{x}^{(0)}$, isto é, $\mathbf{x}^{(0)} = \mathbf{x} - \mathbf{z}$. Então

$$\begin{aligned}
\mathbf{x} - \mathbf{x}^k &= M\mathbf{x} + c - (M\mathbf{x}^{k-1} + \mathbf{c}) \\
&= M(\mathbf{x} - \mathbf{x}^{k-1}) \\
&= M(\mathbf{x} - M\mathbf{x}^{k-2} - \mathbf{c}) \\
&= M\mathbf{x} - M^2\mathbf{x}^{k-2} - M\mathbf{c} \\
&= M(\mathbf{x} - \mathbf{c}) - M^2\mathbf{x}^{k-2} \\
&= M(M\mathbf{x}) - M^2\mathbf{x}^{k-2} \\
&= M^2(\mathbf{x} - \mathbf{x}^{k-2}) \\
&\vdots \\
&= M^k(\mathbf{x} - \mathbf{x}^0) \\
&= M^k\mathbf{z}
\end{aligned}$$

Segue que, $\lim_{k\to\infty} M^k\mathbf{z} = \lim_{k\to\infty}(\mathbf{x} - \mathbf{x}^k) = 0$. Portanto pelo **Teorema 3.8** segue que $\rho(M) < 1$. □

Definição 3.11. A matriz $A \in \mathbb{R}^{n\times n}$ é definida positiva se para qualquer $\mathbf{x} \in \mathbb{R}^n\backslash\{\mathbf{0}\}$ temos que $\mathbf{x}^T A\mathbf{x} > 0$ e semi-definida positiva se $\mathbf{x}^T A\mathbf{x} \geq 0$, $\forall\ \mathbf{x} \in \mathbb{R}^n\backslash\{\mathbf{0}\}$. No caso em que $A \in \mathbb{C}^{n\times n}$, troca-se $\mathbf{x}^T$ por $\mathbf{x}^H = \overline{\mathbf{x}}^T$. □

Agora, vamos ver que dado o sistema $A\mathbf{x} = \mathbf{b}$, com $A \in \mathbb{R}^{n\times n}$ simétrica e definida positiva, o método de Gauss-Seidel converge para qualquer ponto inicial dado.

Teorema 3.12. *(Teorema de Householder-John) Sejam as matrizes A e B reais. Se A e $A - B - B^T$ forem simétricas e definida positiva, respectivamente, então a matriz $H = -(A - B)^{-1}B$ tem raio espectral menor que 1.*

Dem. Considere λ associado ao autovetor não nulo $\mathbf{w}$ de H. Ou seja, $H\mathbf{w} = \lambda\mathbf{w}$. Note que existe a possibilidade de $\mathbf{w}$ e λ serem complexos. De H temos,

$$
\begin{aligned}
H &= -(A-B)^{-1}B \\
(A-B)H &= -B \\
(A-B)H\mathbf{w} &= -B\mathbf{w} \\
(A-B)\lambda\mathbf{w} &= -B\mathbf{w} \\
A\lambda\mathbf{w} - \lambda B\mathbf{w} &= -B\mathbf{w} \\
B\mathbf{w}(-1+\lambda) &= \lambda A\mathbf{w}
\end{aligned}
$$

Note que se $\lambda = 1$ então temos $A\mathbf{w} = 0$. Ou seja, A é singular, isto é, não admite inversa. Contradição, pois A é definida positiva e isto implica que todos os autovalores são maiores que zero. Logo, $\lambda \neq 1$ e

$$
\overline{\mathbf{w}}^T B\mathbf{w} = \frac{\lambda}{-1+\lambda}\overline{\mathbf{w}}^T A\mathbf{w}
$$

Escreva $\mathbf{w} = \mathbf{u} + i\mathbf{v}$, onde $\mathbf{u}$ e $\mathbf{v}$ são vetores cuja componentes são reais. Então

$$
\begin{aligned}
\mathbf{w}^H A\mathbf{w} &= \overline{(\mathbf{u}+i\mathbf{v})}^T A(\mathbf{u}+i\mathbf{v}) \\
&= (\mathbf{u}^T - i\mathbf{v}^T)A(\mathbf{u}+i\mathbf{v}) \\
&= \mathbf{u}^T A\mathbf{u} + i\mathbf{u}^T A\mathbf{v} - i\mathbf{v}^T A\mathbf{u} + \mathbf{v}^T A\mathbf{v} \\
&= \mathbf{u}^T A\mathbf{u} + \mathbf{v}^T A\mathbf{v}
\end{aligned}
$$

Como A é definida positiva, $\overline{\mathbf{w}}^T A\mathbf{w} > 0$ e por hipótese $\overline{\mathbf{w}}^T(A - B - B^T)\mathbf{w} > 0$. Logo,

$$
\begin{aligned}
0 &< \overline{\mathbf{w}}^T(A - B - B^T)\mathbf{w} \\
&= \overline{\mathbf{w}}^T A\mathbf{w} - \overline{\mathbf{w}}^T B\mathbf{w} - \overline{\mathbf{w}}^T B^T\mathbf{w} \\
&= \overline{\mathbf{w}}^T A\mathbf{w} - \frac{\lambda}{-1+\lambda}\overline{\mathbf{w}}^T A\mathbf{w} - \frac{\overline{\lambda}}{-1+\overline{\lambda}}\overline{\mathbf{w}}^T A\mathbf{w} \\
&= \left(1 - \frac{\lambda}{-1+\lambda} - \frac{\overline{\lambda}}{-1+\overline{\lambda}}\right)\overline{\mathbf{w}}^T A\mathbf{w} \\
&= \frac{1-|\lambda|^2}{|-1+\lambda|^2}\overline{\mathbf{w}}^T A\mathbf{w}
\end{aligned}
$$

Como $\lambda \neq 1$, segue que o denominador é maior que zero. E como $\overline{\mathbf{w}}^T A\mathbf{w} > 0$ segue que $1 - |\lambda|^2 > 0$. Portanto $|\lambda| < 1$ e $\rho(H) < 1$. □

Corolário 3.13. *Se A é simétrica definida positiva então o método de Gauss-Seidel converge para a solução do sistema dado qualquer ponto inicial.*

Dem. Seja A simétrica definida positiva por hipótese. Defina B como sendo a matriz que recebe os elementos acima da diagonal de A. Então $A - B - B^T = D$, onde D é a diagonal de A.

Como A é definida positiva, seus elementos da diagonal são positivos. Logo, nenhum autovalor de D é negativo. Portanto D é simétrica definida positiva. E, aplicando **Teorema 3.12** temos que a matriz de iteração satisfaz $\rho(-(L+D)^{-1}U) < 1$. □

Apesar da caracterização de convergência, o método de Gauss-Seidel ainda pode convergir lentamente para a solução. Isto ocorre quando a matriz de iteração tem o raio espectral próximo de 1 [7].

Um exemplo é o seguinte, ([15],pg158). Seja o sistema $A\mathbf{x} = \mathbf{b}$, onde A tem dimensão 50×50 e é da seguinte forma:

$$A = \begin{bmatrix} 2.001 & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & 1 & \\ & & 1 & 2.001 & \end{bmatrix}$$

Definindo $\mathbf{b} = [1, .., 1]^T$ de tamanho 50×1, temos que o raio espectral da matriz de iteração de Gauss-Seidel é $0.9952 \approx 1$. Exigindo uma aproximação da solução exata de 10^{-5} o método de Gauss-Seidel converge em 757 iterações. Se modificarmos a diagonal para 3, o raio espectral fica 0.4428 e o método converge com 14 iterações.

O método SOR tenta corrigir esta desvantagem. Seja $\omega \in \mathbb{R}$ e realizando modificação no *Algoritmo 3.2* na linha 6, o método SOR é dado por:

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + (\omega/a_{ii})\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right)$$

E a sua forma geral é $\mathbf{x}^{(k+1)} = (D-\omega L)^{-1}[(1-\omega)D+\omega U]\mathbf{x}^{(k)} + \omega(D-\omega L)^{-1}\mathbf{b}$. Segue um resultado sobre a convergência do método SOR, que nos garante para quais $\omega \in \mathbb{R}$ o método converge.

Lema 3.14. *Sejam L,D e U como em (3.2). Então vale $det(D^{-1}) = det(D-\omega L)^{-1}$ e $det((1-\omega)I_n + \omega D^{-1}U) = det((1-\omega)I_n)$ com $\omega \in \mathbb{R}$.*

Dem. Note que $D - \omega L$ é da forma

$$\begin{bmatrix} d_{11} & 0 & \dots & 0 \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n1} & \dots & a_{nn-1} & d_{nn} \end{bmatrix}$$

Logo $det(D - \omega L) = d_{11}...d_{nn}$ e $det[(D - \omega L)^{-1}] = (det(D - \omega L))^{-1} = \dfrac{1}{d_{11}...d_{nn}} = det(D^{-1})$. Agora, note que $\omega D^{-1}U$ é da forma

$$\begin{bmatrix} 0 & \times & \dots & \times \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \times \\ 0 & \dots & \dots & 0 \end{bmatrix}$$

Portanto $det((1-\omega)I_n + \omega D^{-1}U) = det((1-\omega)I_n)$ □

Teorema 3.15. *Dado o sistema $A\boldsymbol{x} = \boldsymbol{b}$ com $A \in \mathbb{R}^{n\times n}$, se $a_{ii} \neq 0$ para $i = 1, ..., n$, então $\rho(M_{SOR}) = \rho((D - \omega L)^{-1}[(1-\omega)D + \omega U]) \geq |\omega - 1|$.*

Dem. Note que $det(M_{SOR}) = \prod_{i=1}^{n} \lambda_i$, onde λ_i são os autovalores de M_{SOR}. Segue que

$$\begin{aligned} det(M_{SOR}) &= det((D - \omega L)^{-1}[(1-\omega)D + \omega U]) \\ &= det((D - \omega L)^{-1})det((1-\omega)D + \omega U) \\ &= det(D^{-1})det((1-\omega)D + \omega U) \\ &= det((1-\omega)I_n + \omega D^{-1}U) \\ &= det((1-\omega)I_n) \\ &= (1-\omega)^n \end{aligned}$$

Agora, $\prod_{i=1}^{n} \lambda_i = (1-\omega)^n$. Porém, note que $\lambda_i \leq max\{|\lambda_i|, i = 1, .., n\} = |\lambda_k|$. Logo $\prod_{i=1}^{n} \lambda_i \leq |\lambda_k|^n$. Ou seja $(1-\omega)^n \leq |\lambda_k|^n \Leftrightarrow |1-\omega| \leq \rho(M_{SOR})$. □

Corolário 3.16. *O método SOR converge se, e somente se,* $0 < \omega < 2$.

Dem. Pelo **Teorema 3.15**, $\rho(M_{SOR}) < 1$. Logo $|w - 1| < 1 \Leftrightarrow 0 < w < 2$. □

4 Métodos Não-Estacionários

Nesta parte vamos estabelecer algumas definições e considerações com o propósito de falar sobre métodos baseados no gradiente. Veremos condições sobre a existência do mínimo local, assim como minimizar $f : \mathbb{R}^n \to \mathbb{R}$ continuamente diferenciável em $\mathbb{R}^n$. Seja $f : \mathbb{R}^n \to \mathbb{R}$ continuamente diferenciável em $\mathbb{R}^n$, isto é, suas derivadas parciais de primeira ordem são contínuas. O gradiente de f no ponto $\mathbf{x} = [x_1, ..., x_n]^T \in \mathbb{R}^n$, denotado por $\nabla f(\mathbf{x})$, é dado por,

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, ..., \frac{\partial f(\mathbf{x})}{\partial x_n}\right]^T.$$

A seguir são apresentadas algumas propriedades do gradiente. Seja $c \in \mathbb{R}$ e defina o conjunto $N_c = \{x \in \mathbb{R}^n : f(\mathbf{x}) = c\}$, chamado de conjunto das curvas de nível de f. Geometricamente se $n = 2$, N_c pode ser interpretado como sendo a intersecção do plano $z = \mathbf{c}$ com o gráfico de $f(\mathbf{x})$. Note que dado o ponto $\mathbf{x}_0 \in N_c$, o vetor $\nabla f(\mathbf{x})$ é ortogonal ao vetor tangente à curva de nível em $\mathbf{x}_0$.

Com efeito, seja a curva γ que pertence à N_c e parametrizada pela função continuamente diferenciável $g : \mathbb{R} \to \mathbb{R}^n$, onde $g(t_0) = \mathbf{x}_0$, e a derivada de g, que denotaremos por $Dg(t)$, satisfaça $Dg(t_0) \neq 0$. Seja $h(t) = f(g(t))$ e então $h'(t_0) = Df(g(t_0))Dg(t_0)$. Como $\gamma \in N_c$, temos que $f(g(t)) = c_1$. Logo $h'(t_0) = 0$, ou seja, $Df(g(t_0))Dg(t_0) = 0 \Leftrightarrow \nabla f(\mathbf{x}_0)^T Dg(t_0) = 0$, onde $Dg(t_0)$ é o vetor tangente em $\mathbf{x}_0$.

Vamos ver um outro resultado que nos mostra que a direção em que o vetor $\nabla f(\mathbf{x}_0)$ aponta é a de maior crescimento da função f, na vizinhança de $\mathbf{x}_0$.

Teorema 4.1. *Dada $f : \mathbb{R}^n \to \mathbb{R}$ diferenciável em $\boldsymbol{x}_0$, e um vetor unitário $\boldsymbol{u}$ qualquer, a direção de maior crescimento de f é na direção de $\nabla f(\boldsymbol{x}_0)$.*

Dem. Considere os vetores $\nabla f(\mathbf{x}_0)$ e $\mathbf{u}$. Pela lei dos cossenos,

$$\langle \nabla f(\mathbf{x}_0), \mathbf{u} \rangle = ||\nabla f(\mathbf{x}_0)||||u||cos(\theta).$$

Onde θ é o ângulo formado entre os vetores $\nabla f(\mathbf{x}_0)$ e $\mathbf{u}$. Segue que,

$$\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{u}} = Df(\mathbf{x}_0)(\mathbf{u}) = \nabla f(\mathbf{x}_0)\mathbf{u} = \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle.$$

Note que a taxa de variação é máxima na direção de $\mathbf{u}$ quando $\theta = 0$. Logo $\nabla f(\mathbf{x}_0)$ e $\mathbf{u}$ são colineares e como $\mathbf{u}$ é unitário, $\mathbf{u} = \dfrac{\nabla f(\mathbf{x}_0)}{||\nabla f(\mathbf{x}_0)||}$. □

Para identificar um minímo local precisamos estudar o gradiente e o que chamamos de Hessiana denotada por $\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{n\times n}$. Utilizaremos para esse estudo o Teorema de Taylor cuja demonstração pode ser encontrada em [14].

Definição 4.2. Seja $f : \mathbb{R}^n \to \mathbb{R}$ e considere $x^* \in \mathbb{R}^n$. O ponto $\mathbf{x}^*$ é ponto de mínimo local se existir uma bola aberta com $r > 0$ centrada em $\mathbf{x}^*$, denotado por, $B(r, \mathbf{x}^*)$, de modo que $f(\mathbf{x}^*) \leq f(\mathbf{x})$, $\forall\ \mathbf{x} \in B(r, \mathbf{x}^*)$. □

Teorema 4.3. *(Teorema de Taylor) Seja $f : \mathbb{R}^n \to \mathbb{R}$ continuamente diferenciável e $\boldsymbol{u} \in \mathbb{R}^n$. Então, temos que $\exists\ \alpha \in (0, 1)$ tal que:*

a) $f(\boldsymbol{x} + \boldsymbol{u}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x} + \alpha\boldsymbol{u})^T \boldsymbol{u}$.
b) Supondo f duas vezes continuamente diferenciável temos:

$$f(\boldsymbol{x} + \boldsymbol{u}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T \boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^T \nabla^2 f(\boldsymbol{x} + \alpha\boldsymbol{u})\boldsymbol{u}.$$

□

Teorema 4.4. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ continuamente diferenciável e $\boldsymbol{x}^*$ um ponto de mínimo local de f em $B(r, \boldsymbol{x}^*) \subset \mathbb{R}^n$. Então $\nabla f(\boldsymbol{x}^*) = 0$.*

Dem. Suponha que $\nabla f(\mathbf{x}^*) \neq 0$, e considere a direção $\mathbf{u} = -\nabla f(\mathbf{x}^*)$. Observe que

$$\begin{aligned}
\nabla f(\mathbf{x}^*)^T \mathbf{u} &= -\nabla f(\mathbf{x}^*)^T \nabla f(\mathbf{x}^*) \\
&= -\langle \nabla f(\mathbf{x}^*), \nabla f(\mathbf{x}^*) \rangle \\
&= -||\nabla f(\mathbf{x}^*)||^2 \\
&< 0
\end{aligned}$$

Por hipótese $\nabla f(\mathbf{x}^*)$ é contínua em $B(r, \mathbf{x}^*) \subset \mathbb{R}^n$. Logo existe $T > 0$ de modo que $\nabla f(\mathbf{x}^* + \alpha\mathbf{u})^T \mathbf{u} < 0$ com $\alpha \in [0, T]$.

Pelo **Teorema 4.3.**, segue que $f(\mathbf{x}^* + t\mathbf{u}) = f(\mathbf{x}^*) + t\nabla f(\mathbf{x}^* + \alpha\mathbf{u})^T \mathbf{u}$ com $0 < \alpha \leq t \leq T$. Ou seja, $f(\mathbf{x}^* + t\mathbf{u}) < f(\mathbf{x}^*)$, $t \in (0, T] \subset B(r, \mathbf{x}^*)$. Contradição, pois $\mathbf{x}^*$ é mínimo local em $B(r, \mathbf{x}^*)$. □

Definição 4.5. Dada $f : \mathbb{R}^n \to \mathbb{R}$, se $\mathbf{x}^*$ satisfizer a condição $\nabla f(\mathbf{x}^*) = 0$, dizemos que $\mathbf{x}^*$ é um ponto estacionário. □

Com o teorema acima temos que dado um ponto qualquer no domínio de f, se o gradiente for diferente de zero então ele não pode ser mínimo local.

Segue agora um resultado que não basta verificar se o gradiente se anula no ponto para garantir que este ponto seja mínimo. Denotaremos por B a vizinhança $B(r, \mathbf{x}^*)$.

Teorema 4.6. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ duas vezes continuamente diferenciável com $\boldsymbol{x}^*$ mínimo local de f e $\nabla^2 f(\boldsymbol{x})$ contínua em B. Então $\nabla f(\boldsymbol{x}^*) = 0$ e $\nabla^2 f(\boldsymbol{x}^*)$ é semi-definida positiva.*

Dem. O resultado $\nabla f(\mathbf{x}^*) = 0$ segue direto do **Teorema 4.4**. Suponha agora que $\nabla^2 f(\mathbf{x}^*)$ não seja semi-definida positiva. Pela **Definição 3.11**, existe $\mathbf{u} \neq \mathbf{0}$ de modo que $\mathbf{u}^T \nabla^2 f(\mathbf{x}^*)\mathbf{u} < 0$.

Como $\nabla^2 f(\mathbf{x})$ é contínua em B, existe $T > 0$ de modo que $\mathbf{u}^T \nabla^2 f(\mathbf{x}^* + t\mathbf{u})\mathbf{u} < 0$ com $0 \leq t \leq T$. Pelo **Teorema 4.3**, em B, existe $\alpha \in (0, T]$ e $t \in (0, \alpha)$ de modo que

$$f(\mathbf{x}^* + \alpha\mathbf{u}) = f(\mathbf{x}^*) + \alpha \nabla f(\mathbf{x}^*)^T \mathbf{u} + \alpha^2 \frac{1}{2} \mathbf{u}^T \nabla^2 f(\mathbf{x}^* + t\mathbf{u})\mathbf{u}.$$

Segue, $f(\mathbf{x}^* + \alpha\mathbf{u}) < f(\mathbf{x}^*)$ em B. O que é uma contradição. □

Com os dois resultados anteriores conseguimos caracterizar quando um ponto é um minimizador local.

Teorema 4.7. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ duas vezes continuamente diferenciável de modo que $\nabla f(\boldsymbol{x}^*) = 0$ em B e $\nabla^2 f(\boldsymbol{x})$ contínua em B com $\nabla^2 f(\boldsymbol{x}^*)$ definida positiva. Então o ponto $\boldsymbol{x}^*$ é mínimo local em B.*

Dem. Como $\nabla^2 f(\mathbf{x})$ é contínua em B, $\nabla^2 f(\mathbf{x})$ é definida positiva em $B(\frac{r}{2}, \mathbf{x}^*)$. Isto é, para qualquer $\mathbf{v} \in B(\frac{r}{2}, \mathbf{x}^*)$ temos que $\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v} > 0$.

Logo, considerando $\mathbf{u} \in B(\frac{r}{2}, \mathbf{x}^*)$ segue $\mathbf{x}^* + \mathbf{u} \in B(\frac{r}{2}, \mathbf{x}^*)$ e pelo **Teorema 4.3**, temos $f(\mathbf{x}^* + \mathbf{u}) = f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T \mathbf{u} + \frac{1}{2} \mathbf{u}^T \nabla^2 f(\mathbf{x}^* + \alpha\mathbf{u})\mathbf{u}$ com $\alpha \in (0, 1)$.

Como $\nabla f(\mathbf{x}^*) = 0$, segue $f(\mathbf{x}^* + \mathbf{u}) = f(\mathbf{x}^*) + \frac{1}{2} \mathbf{u}^T \nabla^2 f(\mathbf{x}^* + \alpha\mathbf{u})\mathbf{u}$. Porém, $\mathbf{u}^T \nabla^2 f(\mathbf{x}^* + \alpha\mathbf{u})\mathbf{u} > 0$ em $B(\frac{r}{2}, \mathbf{x}^*)$ pois $||\mathbf{x}^* - \alpha\mathbf{u} - \mathbf{x}^*|| = \alpha||\mathbf{u}|| \leq \frac{r}{2}$. Portanto $f(\mathbf{x}^* + \mathbf{u}) > f(\mathbf{x}^*)$. □

Colocados estes resultados, podemos falar sobre o método do Gradiente com f quadrática e convexa com a matriz representante de f simétrica definida positiva (SDP), isto é, $Q \in \mathbb{R}^{n \times n}$ é a matriz SDP, $\mathbf{q} \in \mathbb{R}^n$ e $c \in \mathbb{R}$.

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T Q \mathbf{x} - \mathbf{q}^T \mathbf{x} + c \tag{4.1}$$

Note que $\nabla f(\mathbf{x}) = Q\mathbf{x} - \mathbf{q}$, logo, resolver o sistema $Q\mathbf{x} = \mathbf{q}$ é equivalente a minimizar $f(\mathbf{x})$, pois $\nabla f(\mathbf{x}) = 0 \Leftrightarrow Q\mathbf{x} = \mathbf{q}$. Desse modo, dado $\mathbf{x}^{(0)}$ ponto inicial, considere

a direção $-\nabla f(\mathbf{x}^{(k)})$ e a sequência $\{\mathbf{x}^{(k)}\}$ dada por $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \nabla f(\mathbf{x}^{(k)})$, com $\alpha_k \in argmin_{\alpha \geq 0} f(\mathbf{x}^{(k)} + \alpha \nabla f(\mathbf{x}^{(k)}))$, onde $argmin$ é o conjunto dos minimizadores da função $\phi_k(\alpha) = f(\mathbf{x}^{(k)} + \alpha \nabla f(\mathbf{x}^{(k)}))$.

O parâmetro $\alpha_k \in [0, +\infty)$ informa em cada iteração o maior decrescimento de $f(\mathbf{x})$ na direção do $\nabla f(\mathbf{x}^{(k)})$, e o calculamos da seguinte forma. Seja $\mathbf{u} \in \mathbb{R}^n$ e defina $\phi_k(\mathbf{u}) = f(\mathbf{x}^{(k)} + \alpha \mathbf{u})$. Resolvendo $\phi_k'(\alpha) = 0$ temos $\nabla f(\mathbf{x}^{(k)} + \alpha \mathbf{u})^T \mathbf{u} = 0 \Leftrightarrow (Q(\mathbf{x}^{(k)} + \alpha \mathbf{u}) - \mathbf{q})^T \mathbf{u} = 0 \Leftrightarrow (\nabla f(\mathbf{x}^k) + \alpha Q \mathbf{u})^T \mathbf{u} = 0 \Leftrightarrow \alpha_k = -\frac{\nabla f(\mathbf{x}^k)^T \mathbf{u}}{\mathbf{u}^T Q \mathbf{u}}$. Denominando de resíduo a igualdade $\nabla f(\mathbf{x}^{(k)}) = Q\mathbf{x} - \mathbf{q} = \mathbf{r}_k$ e definindo $\mathbf{u} = \mathbf{r}_k$, temos o passo exato

$$\alpha_k = -\frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_k^T Q \mathbf{r}_k}.$$

Desse modo, para cada iteração escolhemos α_k que nos fornece o maior decrescimento na direção do gradiente. O algoritmo do gradiente com a busca linear exata é dado por,

```
Algoritmo 4.1
Dados Q, q, x^(0) e ε
r_0 = Qx^(0) − q
k = 0
Enquanto ||r_k|| ≥ ε faça
        α_k = −(r_k^T r_k)/(r_k^T Q r_k)
        x^(k+1) = x^(k) + α_k r_k
        k = k + 1
        r_k = Qx^(k) − q
Fim
```

Seguem dois resultados sobre a sequência gerada pelo método do Gradiente com busca linear exata.

Proposição 4.8. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ continuamente diferenciável. A sequência gerada pelo Algoritmo 4.1 satisfaz a propriedade de que $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$ é ortogonal a $\boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)}$ para todo $k \in \{0, 1, 2, ...\}$.*

Dem. Queremos mostrar que $\langle (\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}), (\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}) \rangle = 0$. Como $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = \alpha_k \nabla f(\mathbf{x}^{(k)})$ e $\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)} = \alpha_{k+1} \nabla f(\mathbf{x}^{(k+1)})$. Segue que,

$$\langle (\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}), (\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}) \rangle = \alpha_k \alpha_{k+1} \langle \nabla f(\mathbf{x}^{(k)}), \nabla f(\mathbf{x}^{(k+1)}) \rangle.$$

Logo vamos ver que $\langle \nabla f(\mathbf{x}^{(k)}), \nabla f(\mathbf{x}^{(k+1)}) \rangle = 0$. Seja α_k minimizador de $\phi_k(\alpha)$. Então

$$\begin{aligned}
0 &= \frac{d\phi_k(\alpha)}{d\alpha} \\
&= \frac{D}{d\alpha}[f(\mathbf{x}^{(k)} + \alpha\nabla f(\mathbf{x}^{(k)}))] \\
&= (\nabla f(\mathbf{x}^{(k)} + \alpha\nabla f(\mathbf{x}^{(k)}))^T \nabla f(\mathbf{x}^{(k)}) \\
&= \langle \nabla f(\mathbf{x}^{(k)} + \alpha\nabla f(\mathbf{x}^{(k)})), \nabla f(\mathbf{x}^{(k)}) \rangle \\
&= \langle \nabla f(\mathbf{x}^{(k+1)}), \nabla f(\mathbf{x}^{(k)}) \rangle
\end{aligned}$$

□

Proposição 4.9. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ continuamente diferenciável. Se $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$ então $f(\boldsymbol{x}^{(k+1)}) \leq f(\boldsymbol{x}^{(k)})$.*

Dem. Seja $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \nabla f(\mathbf{x}^{(k)})$. Considere $\phi_k(\alpha) = f(\mathbf{x}^{(k)} - \alpha\nabla f(\mathbf{x}^{(k)}))$ com $\alpha_k \geq 0$ o mínimo de $\phi_k(\alpha)$ com $\alpha \in [0, +\infty)$. Logo $\phi_k(\alpha_k) \leq \phi_k(\alpha)$, para qualquer α em $[0, +\infty)$.

Como $\dfrac{d\phi_k(0)}{d\alpha} = -||\nabla f(\mathbf{x}^{(k)})||^2$, segue que $\dfrac{d\phi_k(0)}{d\alpha} < 0$, logo, $\phi_k(0)$ é decrescente em $(0, t]$ com $t > 0$, ou seja, $\phi_k(0) > \phi_k(\alpha)$ para todo $\alpha \in (0, t]$. Portanto $f(\mathbf{x}^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(t) < \phi_k(0) = f(\mathbf{x}^{(k)})$. □

A seguir veremos que quando temos a função (4.1) e existe um mínimo local podemos afirmar que este ponto é um mínimo global.

Definição 4.10. A função $f : \mathbb{R}^n \to \mathbb{R}$ é convexa em $R \subseteq \mathbb{R}^n$ se satisfaz a desigualdade,

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) \tag{4.2}$$

para quaisquer $\mathbf{x}, \mathbf{y} \in R$ e $\alpha \in [0, 1]$. No caso em que temos a desigualdade estrita ($<$) dizemos que f é estritamente convexa. □

Proposição 4.11. *Sejam funções $g(\boldsymbol{x})$ e $h(\boldsymbol{x})$ de $\mathbb{R}^n$ em $\mathbb{R}$ estritamente convexa e convexa, respectivamente. Então $g(\boldsymbol{x}) + h(\boldsymbol{x})$ é estritamente convexa.*

Dem. Note que g e h satisfazem (4.2) por hipótese. Queremos mostrar que $(g+h)(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < \alpha(g+h)(\mathbf{x}) + (1-\alpha)(g+h)(\mathbf{y})$ em R. Segue que $(g+h)(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) = g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) + h(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < \alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y}) + \alpha h(\mathbf{x}) + (1-\alpha)h(\mathbf{y}) = \alpha(g+h)(\mathbf{x}) + (1-\alpha)(g+h)(\mathbf{x})$ como queríamos. □

Proposição 4.12. *Seja $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{q}^T\boldsymbol{x} + c$ com Q SDP. Então $f(\boldsymbol{x})$ é estritamente convexa em $\mathbb{R}^n$.*

Dem. Note que $g(\mathbf{x}) = -\mathbf{q}^T\mathbf{x}$ é convexa e o mesmo ocorre com $h(\mathbf{x}) = c$. Como $g(\mathbf{x}) + h(\mathbf{x})$ é convexa, basta verificar que $s(\mathbf{x}) = \mathbf{x}^T Q\mathbf{x}$ é estritamente convexa, pois a soma de

função convexa com estritamente convexa é estritamente convexa pela **Proposição 4.11**. Vamos ver que $\alpha s(\mathbf{x}) + (1-\alpha)s(\mathbf{y}) - s(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) > 0$. Desenvolvendo temos,

$$\alpha(1-\alpha)\mathbf{x}^T Q\mathbf{x} + \alpha(1-\alpha)\mathbf{y}^T Q\mathbf{y} - \alpha(1-\alpha)\mathbf{x}^T Q\mathbf{y} - \alpha(1-\alpha)\mathbf{y}^T Q\mathbf{x} > 0$$

Como $\alpha \in (0,1)$ podemos dividir por $\frac{1}{\alpha(1-\alpha)}$. E obtemos $\mathbf{x}^T Q\mathbf{x} + \mathbf{y}^T Q\mathbf{y} - \mathbf{x}^T Q\mathbf{y} - \mathbf{y}^T Q\mathbf{x} > 0 \Leftrightarrow (\mathbf{x}-\mathbf{y})^T Q(\mathbf{x}-\mathbf{y}) > 0$. Note que Q é SDP, logo $(\mathbf{x}-\mathbf{y})^T Q(\mathbf{x}-\mathbf{y}) > 0$. Portanto $f(\mathbf{x})$ é estritamente convexa. □

Teorema 4.13. *Se $f : \mathbb{R}^n \to \mathbb{R}$ é continuamente diferenciável e estritamente convexa com $\boldsymbol{x}^*$ um ponto estacionário qualquer, então $\boldsymbol{x}^*$ é mínimo global de f em $\mathbb{R}^n$.*

Dem. Suponha que $\mathbf{x}^*$ não seja mínimo global, isto é, é um mínimo local. Logo existe $\mathbf{z} \in \mathbb{R}^n$ com $f(\mathbf{z}) < f(\mathbf{x}^*)$ em $\mathbb{R}^n$. Considere o vetor não nulo $\mathbf{u} = \mathbf{z} - \mathbf{x}^*$. Note, utilizando o fato de f ser estritamente convexa, que

$$\begin{aligned}
\nabla f(\mathbf{x}^*)(\mathbf{z}-\mathbf{x}^*) &= Df(\mathbf{x}^*)(\mathbf{z}-\mathbf{x}^*) \\
&= \frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{u}} \\
&= \lim_{h\to 0} \frac{f(\mathbf{x}^* + h(\mathbf{z}-\mathbf{x}^*)) - f(\mathbf{x}^*)}{h} \\
&< \lim_{h\to 0} \frac{hf(\mathbf{z}) + (1-h)f(\mathbf{x}^*) - f(\mathbf{x}^*)}{h} \\
&= f(\mathbf{z}) - f(\mathbf{x}^*) \\
&< 0
\end{aligned}$$

Segue que $\nabla f(\mathbf{x}^*) \neq 0$ e $\mathbf{x}^*$ não é um ponto estacionário, o que é uma contradição. □

4.1 Método de Barzilai-Borwein

Antes de mostrar um exemplo, existe uma variação do método do Gradiente, que é o método de Barzilai-Borwein [3]. Defina $\mathbf{s}^{(k-1)} = \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$ e $\Delta\mathbf{r}_k = \mathbf{r}_k - \mathbf{r}_{k-1}$. O método calcula o passo α_k de modo que

$$\alpha_k = argmin_{\alpha>0} \frac{1}{2}||\mathbf{s}^{(k-1)}\alpha - \Delta\mathbf{r}_k||^2 \Rightarrow \alpha_k = \frac{(\mathbf{s}^{(k-1)})^T \Delta\mathbf{r}_k}{(\mathbf{s}^{(k-1)})^T \mathbf{s}^{(k-1)}} \tag{4.3}$$

e temos $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \dfrac{1}{\alpha_k}\mathbf{r}_k$. Ou ainda,

$$\alpha_k = argmin_{\alpha>0}\frac{1}{2}||\mathbf{s}^{(k-1)} - \Delta\mathbf{r}_k\alpha||^2 \Rightarrow \alpha_k = \frac{(\mathbf{s}^{(k-1)})^T\Delta\mathbf{r}_k}{(\Delta\mathbf{r}_k)^T\Delta\mathbf{r}_k} \tag{4.4}$$

com $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{r}_k$. O algoritmo pode ser escrito da seguinte forma, quando k é ímpar aplicamos o passo (4.3) e quando k é par aplicamos o passo (4.4).

1 Algoritmo 4.2
2 Dados $x^{(0)}$, Q, q, ε
3 $r_0 = Qx^{(0)} - q$
4 k = 0
5 Enquanto $||r_k|| > \varepsilon$ faça
6 Se k = 0 então
7 $\alpha_k = -\dfrac{(r_k)^T r_k}{(r_k)^T Q r_k}$
8 Senão, se k é ímpar
9 $\alpha_k = \dfrac{(s^{(k-1)})^T s^{(k-1)}}{(s^{(k-1)})^T \Delta r_k}$
10 Senão
11 $\alpha_k = \dfrac{(s^{(k-1)})^T \Delta r_k}{(\Delta r_k)^T \Delta r_k}$
12 Fim
13 $x^{(k+1)} = x^{(k)} + \alpha_k r_k$
14 k=k+1
15 $r_k = Qx^{(k)} - q$
16 Fim

Exemplo 4.14. Vamos resolver dois sistemas $Q_1\mathbf{x} = \mathbf{q}_1$ e $Q_2\mathbf{x} = \mathbf{q}_2$ com $\mathbf{q}_1 = Q_1\mathbf{u}^T$ e $\mathbf{q}_2 = Q_2\mathbf{u}^T$ onde $\mathbf{u} = [1, 1, .., 1]^T \in \mathbb{R}^{200}$ e com as matrizes Q_1 e Q_2 de tamanho 200×200, simétricas e ambas definidas positivas, com entradas entre (0,1). Aplicando Gradiente e Gradiente-BB no sistema $Q_1\mathbf{x} = \mathbf{q}_1$, com a tolerância de 10^{-3} (uma aproximação de 3 casas decimais da solução exata), precisaram de 4 iterações, ambos os métodos.

Já para o sistema $Q_2\mathbf{x} = \mathbf{q}_2$ com a mesma tolerância, ocorreu a convergência em 435 iterações para o método do Gradiente e 31 iterações para o método Gradiente-BB. Graficamente, definindo o eixo x como o número de iterações e o eixo y na escala logarítmica a $||\mathbf{r}_k||$, temos o seguinte gráfico.

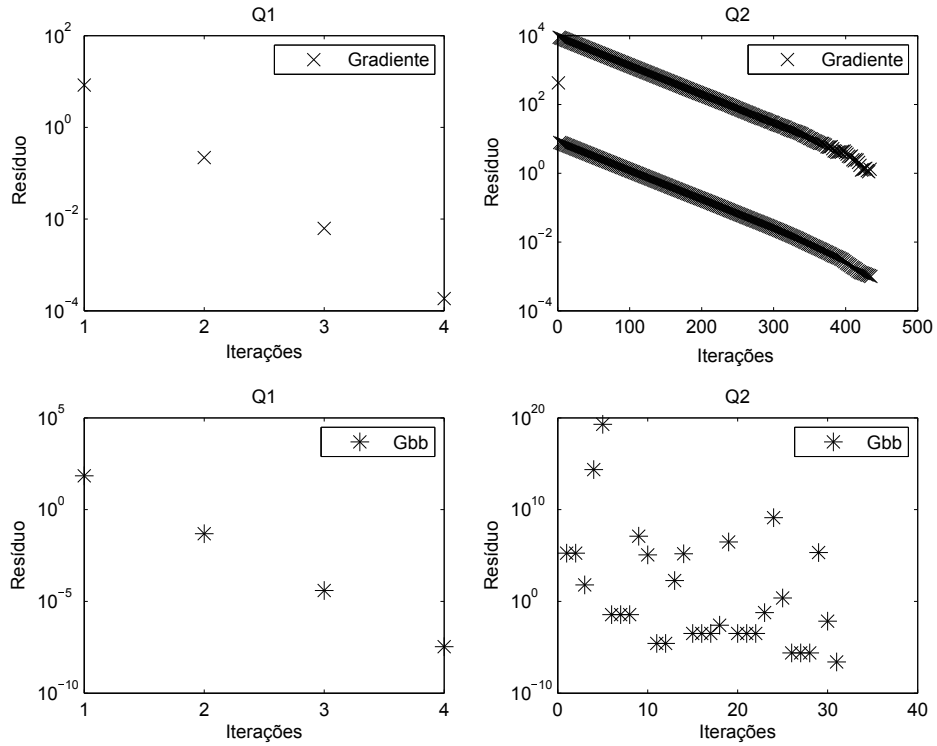

Figura 2 – Comparação entre Gradiente e Gradiente-BB

Note que para o sistema $Q_1\mathbf{x} = \mathbf{q}_1$ tanto o Gradiente como Gradiente-BB tiveram um bom desempenho. Porém, ambos os métodos tiverem um aumento significativo de iterações ao resolver o sistema $Q_2\mathbf{x} = \mathbf{q}_2$. A razão desse aumento tem relação exclusivamente sobre a matriz do sistema e pode ser explicado através do chamado número de condição da matriz.

Definição 4.15. Seja Q uma matriz simétrica definida positiva. O número de condição da matriz Q é definido por $\kappa(Q) = \dfrac{\lambda_{max}(Q)}{\lambda_{min}(Q)}$, onde $\lambda_{max}(Q)$ e $\lambda_{min}(Q)$ são os autovalores de máximo e mínimo de Q respectivamente. □

No **Exemplo 4.14**, temos que $\kappa(Q_1) \approx 2$ e $\kappa(Q_2) \approx 3 \times 10^7$ e isso explica o aumento no número de iterações. Como Q é simétrica definida positiva, as curvas de níveis da quadrática tem o formato de elipses, quanto maior o número de condição da matriz do sistema, mais lenta é a convergência, pois as elipses serão alongadas, além de estarem rotacionadas e transladadas com relação aos eixos coordenados.

Esta característica alongada esta associado ao fato que o autovalor λ_{max} indica a distância do eixo maior e λ_{min} indica a distância do eixo menor da elipse, conforme ilustrado na Figura 3.

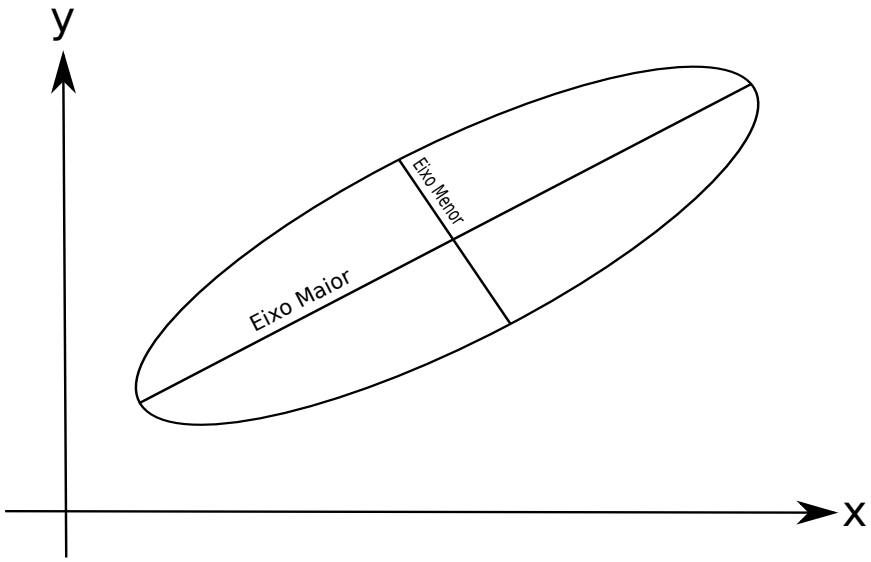

Figura 3 – Elipse

Desse modo a convergência do método do Gradiente é na forma de "zigue-zague". Como a Figura 4, onde $c1$,$c2$,.. são as curvas de níveis de $f(\mathbf{x})$ e $\mathbf{x}^*$ é o mínimo. Um modo de melhorar a convergência nesse caso é utilizando o método do Gradiente com direções conjugadas. O próximo capítulo é dedicado a este assunto.

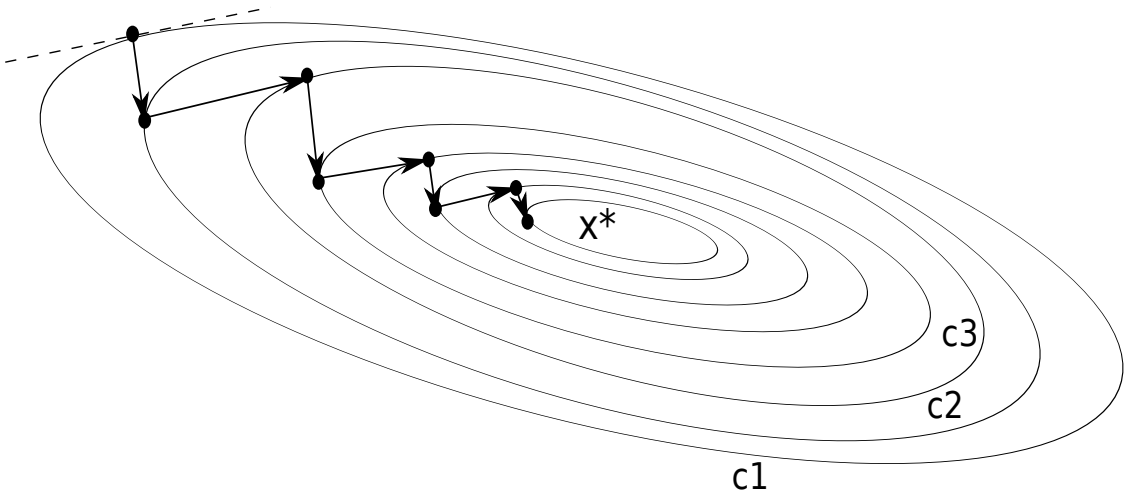

Figura 4 – Método do gradiente

5 Gradiente Conjugado

Neste capítulo exporemos sobre o gradiente conjugado, estudando a convergência e um modo de acelerar esta convergência através do que chamamos de pré-condicionamento, além de resultados numéricos dos métodos estacionários e não-estacionários.

Utilizaremos nos experimentos matrizes esparsas. Neste caso, as matrizes têm um grande número de entradas nulas. As matrizes que faremos os testes são do tipo simétrica definida positiva e esparsa, retiradas de uma coleção de matrizes chamadas de *The SuiteSparse Matrix Collection*[1] , provenientes de aplicações reais.

Os métodos serão aplicados em 20 matrizes de dimensões que variam de 19×19 até 9604×9604, que vieram de problemas reais do tipo combinatória, estrutural, grafos, eletromagnetismo e de energia em um circuito.

Vamos iniciar definindo o que são vetores conjugados, mostrar maneiras de construir este conjunto e ver algumas propriedades.

Definição 5.1. Seja Q simétrica e definida positiva. Os vetores não nulos $\{\mathbf{v}_0, ..., \mathbf{v}_{n-1}\}$, $\mathbf{v}_i \in \mathbb{R}^n$, são conjugados com relação a matriz Q ou Q-conjugados se $\mathbf{v}_s^T Q \mathbf{v}_t = 0$ para qualquer $s, t \in \{0, 1, 2, ..., n-1\}$ com $s \neq t$. □

Exemplo 5.2. Vamos ver que os autovetores de Q são Q-conjugados. Primeiro note que dado os autovetores $\mathbf{v}_i$ e $\mathbf{v}_j$, onde $i, j \in \{1, ..., n\}$, $i \neq j$, com os autovalores λ_i e λ_j respectivamente, temos $\lambda_i \mathbf{v}_i^T \mathbf{v}_j = \mathbf{v}_i^T \lambda_i \mathbf{v}_j = \mathbf{v}_i^T Q \mathbf{v}_j = \mathbf{v}_j^T Q \mathbf{v}_i = \lambda_j \mathbf{v}_j^T \mathbf{v}_i = \lambda_j \mathbf{v}_i^T \mathbf{v}_j \Leftrightarrow (\lambda_i - \lambda_j) \mathbf{v}_i^T \mathbf{v}_j = 0$. Como os autovalores são distintos segue que $\mathbf{v}_i^T \mathbf{v}_j = 0$.

Desse modo, $\mathbf{v}_i^T Q \mathbf{v}_j = \mathbf{v}_i \lambda_j \mathbf{v}_j = \lambda_j \mathbf{v}_i^T \mathbf{v}_j = 0$. Portanto o conjunto formado pelos autovetores de Q são Q-conjugados.

Exemplo 5.3. Outro modo de obter vetores Q-conjugados é por meio do processo de Gram-Schmidt. Dado um conjunto de vetores linearmente independentes em $\mathbb{R}^n$, $\{\mathbf{u}_0, \mathbf{u}_1, ..., \mathbf{u}_{n-1}\}$, defina $\mathbf{v}_0 = \mathbf{u}_0$ e considere para cada $\mathbf{v}_{k+1}$ com $k \in \{0, ..., n-2\}$ o seguinte:

$$\mathbf{v}_{k+1} = \mathbf{u}_{k+1} - \sum_{i=0}^{k} \frac{\mathbf{u}_{k+1}^T Q \mathbf{v}_i}{\mathbf{v}_i^T Q \mathbf{v}_i} \mathbf{v}_i.$$

Os vetores $\{\mathbf{v}_0, ..., \mathbf{v}_{n-1}\}$ são Q-conjugados.

[1] https://sparse.tamu.edu/. Acesso em: 30 nov. 2019.

Proposição 5.4. *O conjunto de vetores não nulos* $\{\boldsymbol{v}_0, ..., \boldsymbol{v}_{n-1}\}$ *que são Q-conjugados são linearmente independentes.*

Dem. Suponha que o conjunto não seja linearmente independente. Então é possível escrever $\mathbf{v}_i = c_0\mathbf{v}_0 + ... + c_{i-1}\mathbf{v}_{i-1} + c_{i+1}\mathbf{v}_{i+1} + ... + c_{n-1}\mathbf{v}_{n-1}$. Multiplicando por $Q\mathbf{v}_i$ segue

$$\mathbf{v}_i^T Q\mathbf{v}_i = c_0\mathbf{v}_0^T Q\mathbf{v}_i + ... + c_{i-1}\mathbf{v}_{i-1}^T Q\mathbf{v}_i + c_{i+1}\mathbf{v}_{i+1}^T Q\mathbf{v}_i + ... + c_{n-1}\mathbf{v}_{n-1}^T Q\mathbf{v}_i.$$

Como no lado direito da igualdade estamos realizando multiplicação de vetores Q-conjugados, segue $\mathbf{v}_i^T Q\mathbf{v}_i = 0 \Rightarrow \mathbf{v}_i = 0$. O que é contradição. □

Observe que os dois exemplos anteriores são maneiras de obter vetores Q-conjugados, porém, existe um custo alto atrelado a esses dois procedimentos. Por hora, vamos supor que conhecemos as direções conjugadas $\{\mathbf{v}_0, ..., \mathbf{v}_{n-1}\}$ e ver alguns resultados.

Similar ao *Algoritmo 4.1*, consideramos a cada iteração a direção $\mathbf{v}_k$ ao invés do gradiente. Também conseguimos de modo explícito definir α_k realizando os mesmos cálculos, isto é, minimizando $\phi_k(\alpha)$ na direção de $\mathbf{v}_k$, neste caso $\alpha_k = -\dfrac{(\mathbf{r}_k)^T\mathbf{v}_k}{(\mathbf{v}_k)^T Q\mathbf{v}_k}$.

```
1 Algoritmo 5.1
2 Dados x^(0), v_0, ..., v_{n-1}
3 Para k=0,1,2,..,n-1 faça
4       α_k = -(r_k)^T v_k / (v_k)^T Q v_k
5       x^(k+1) = x^(k) + α_k v_k
6 Fim
```

O interessante é que conseguimos minimizar $f(\mathbf{x})$ em, no máximo, n passos. Conforme mostramos no próximo teorema.

Teorema 5.5. *Dado um conjunto de vetores Q-conjugados* $\{\boldsymbol{v}_0, ..., \boldsymbol{v}_{n-1}\}$*, é possível minimizar f(*$\boldsymbol{x}$*) em, no máximo, n iterações.*

Dem. Seja $\mathbf{x}^{(0)}$ um vetor inicial qualquer e $\mathbf{x}^*$ solução de $Q\mathbf{x} = \mathbf{q}$. Queremos mostrar que a sequência $\{\mathbf{x}^{(k)}\}$ converge para $\mathbf{x}^*$ em, no máximo, n iterações.

Como os vetores Q-conjugados formam uma base para $\mathbb{R}^n$ pela **Proposição 5.4**, podemos escrever $\mathbf{x}^* - \mathbf{x}^{(0)} = c_0\mathbf{v}_0 + ... + c_{n-1}\mathbf{v}_{n-1}$. Agora,

$$(\mathbf{v}_k)^T Q(\mathbf{x}^* - \mathbf{x}^{(0)}) = (\mathbf{v}_k)^T Qc_0\mathbf{v}_0 + (\mathbf{v}_k)^T Qc_1\mathbf{v}_1 + ... + (\mathbf{v}_k)^T Qc_{n-1}\mathbf{v}_{n-1},$$

para algum $k \in \{0, .., n-1\}$. Por definição de vetores conjugados $(\mathbf{v}_k)^T Q(\mathbf{x}^* - \mathbf{x}^{(0)}) = c_k(\mathbf{v}_k)^T Q(\mathbf{v}_k)$. Portanto $c_k = \dfrac{(\mathbf{v}_k)^T Q(\mathbf{x}^* - \mathbf{x}^{(0)})}{(\mathbf{v}_k)^T Q\mathbf{v}_k}$.

Primeiro observe que para qualquer iteração j, pelo *Algoritmo 5.1* linha 5, podemos escrever $\mathbf{x}^{(j)} = \mathbf{x}^{(0)} + \alpha_0\mathbf{v}_0 + \alpha_1\mathbf{v}_1 + ... + \alpha_{j-1}\mathbf{v}_{j-1}$. Veremos que $\alpha_k = c_k$ onde $k \in \{0, ..., n-1\}$. Pela observação inicial $\mathbf{x}^{(k)} - \mathbf{x}^{(0)} = c_0\mathbf{v}_0 + ... + c_{k-1}\mathbf{v}_{k-1}$ (1).

Utilizando a hipótese, $(\mathbf{v}_k)^T Q(\mathbf{x}^{(k)} - \mathbf{x}^{(0)}) = 0$. Portanto

$$c_k = \frac{(\mathbf{v}_k)^T Q(\mathbf{x}^* - \mathbf{x}^{(0)})}{(\mathbf{v}_k)^T Q\mathbf{v}_k} = \frac{\mathbf{v}_k^T\mathbf{q} - \mathbf{v}_k^T Q\mathbf{x}^{(0)}}{(\mathbf{v}_k)^T Q\mathbf{v}_k} = \frac{(\mathbf{v}_k)^T\mathbf{r}_k}{(\mathbf{v}_k)^T Q\mathbf{v}_k} = \alpha_k,$$

como queríamos, pois multiplicando por $\mathbf{v}_k Q$ a equação (1) e utilizando a definição de vetores conjugados, chegamos em $\mathbf{v}_k^T Q\mathbf{x}^{(k)} - \mathbf{v}_k^T Q\mathbf{x}^{(0)} = 0$. □

Definição 5.6. Considere o ponto $\mathbf{x}_0$ e as direções conjugadas $\mathbf{v}_0, ..., \mathbf{v}_{k-1}$ em $\mathbb{R}^n$, defina o conjunto

$$\mathfrak{L} = \{\mathbf{x} : \mathbf{x} = \mathbf{x}_0 + span\{\mathbf{v}_0, ..., \mathbf{v}_{k-1}\}\}.$$

□

O resultado seguinte nos garante mais duas propriedades com relação a vetores Q-conjugados. Seja a sequência de vetores $\{\mathbf{x}^{(k)}\}$ gerado pelo *Algoritmo 5.1.* Então, para cada iteração k, $\mathbf{x}^{(k)}$ minimiza $f(\mathbf{x})$ no conjunto $\mathfrak{L}$. Isto é, como os vetores $\{\mathbf{v}_0, ..., \mathbf{v}_{k-1}\}$ são linearmente independentes pela **Proposição5.4**, a medida que k cresce, terá um instante em que teremos uma base para $\mathbb{R}^n$ e o mínimo $\mathbf{x}^*$ de $f(\mathbf{x})$ estará nesta expansão.

A outra propriedade é que $\mathbf{r}_k^T\mathbf{v}_i = 0$ com $i \in \{0, .., k-1\}$, onde podemos escrever o resíduo da forma $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}Q\mathbf{v}_{k-1}$, pois $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k Q\mathbf{v}_k$, assim, $Q\mathbf{x}^{(k+1)} = Q\mathbf{x}^{(k)} + \alpha_k Q\mathbf{v}_k$. Ou ainda, $Q\mathbf{x}^{(k+1)} - \mathbf{q} = Q\mathbf{x}^{(k)} - \mathbf{q} + \alpha_k Q\mathbf{v}_k$. Isto é, $\mathbf{r}_k = \mathbf{r}_{k-1} + \alpha_{k-1}Q\mathbf{v}_{k-1}$. Logo, o resíduo na iteração k é ortogonal a todos os vetores $\mathbf{v}_i$ das iterações anteriores à k.

Lema 5.7. *O vetor $\overline{\boldsymbol{x}}$ minimiza $f(\boldsymbol{x})$ em $\mathfrak{L}$ se, e somente se, $\nabla f(\overline{\boldsymbol{x}})^T\boldsymbol{v}_i = 0$ com $i \in \{0, 1, ..., k-1\}$.*

Dem. Suponha que $\overline{\mathbf{x}}$ minimiza $f(\mathbf{x})$ em $\mathfrak{L}$. Veremos que $\nabla f(\overline{\mathbf{x}})^T\mathbf{v}_i = 0$. Considere a função $\phi(\mathbf{c}) = f(\mathbf{x}_0 + c_0\mathbf{v}_0 +, ..., +c_{k-1}\mathbf{v}_{k-1})$, onde $\mathbf{c} = [c_0, c_1, ..., c_{k-1}]^T \in \mathbb{R}^k$. Observe que, por definição f é estritamente convexa, logo $\phi(\mathbf{c})$ é estritamente convexa. Segue que existe um único ponto $\mathbf{c}^*$ que minimiza $\phi(\mathbf{c})$. Ou seja, fazendo $\mathbf{c}^* = [c_0^*, c_1^*, ..., c_{k-1}^*]^T$, temos

$$\frac{\partial\phi(\mathbf{c})}{\partial\mathbf{c}_i}\Big|_{\mathbf{c}^*} = 0, \quad i = 0, ..., k-1$$

Segue que $0 = \nabla\phi(\mathbf{x}_0 + c_0^*\mathbf{v}_0 +, ..., +c_{k-1}^*\mathbf{v}_{k-1})^T\mathbf{v}_i = \nabla f(\overline{\mathbf{x}})^T\mathbf{v}_i$. Por outro lado, se $\nabla f(\overline{\mathbf{x}})^T\mathbf{v}_i = 0$ então $\nabla f(\overline{\mathbf{x}})^T = 0$ pois $\mathbf{v}_i$ são não nulos. Portanto, $\overline{\mathbf{x}}$ minimiza $f(\mathbf{x})$.

□

Teorema 5.8. *Seja $\boldsymbol{x}^{(0)}$ um vetor inicial qualquer e a sequência gerada pela linha 5 do Algoritmo 5.1, isto é, sequência gerada pelas direções dos vetores conjugados. Então $(\boldsymbol{r}_k)^T \boldsymbol{v}_i = 0$ com $i \in \{0, 1, ..., k-1\}$ e além disso $\boldsymbol{x}^{(k)}$ minimiza $f(\boldsymbol{x})$ no conjunto $\mathfrak{L}$.*

Dem. Veremos que $(\mathbf{r}_k)^T \mathbf{v}_i = 0$ por indução. Para $k = 1$ temos $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha_0 \mathbf{v}_0$. Pelo **Lema 5.7**, $\mathbf{x}^{(1)}$ minimiza f na direção de $\mathbf{v}_0$, ou seja, $\mathbf{r}_1^T \mathbf{v}_0 = 0$. Suponha que nossa hipótese de indução seja $\mathbf{r}_{k-1}^T \mathbf{v}_i = 0$ com $i \in \{0, 1, ..., k-2\}$. Como $\mathbf{r}_k = \mathbf{r}_{k-1} + \alpha_{k-1} Q \mathbf{v}_{k-1}$, segue que

$$\begin{aligned}
(\mathbf{v}_{k-1})^T \mathbf{r}_k &= (\mathbf{v}_{k-1})^T \mathbf{r}_{k-1} + \alpha_{k-1} (\mathbf{v}_{k-1})^T Q \mathbf{v}_{k-1} \\
&= (\mathbf{v}_{k-1})^T \mathbf{r}_{k-1} + \frac{(\mathbf{r}_{k-1})^T \mathbf{v}_{k-1}}{(\mathbf{v}_{k-1})^T Q \mathbf{v}_{k-1}} (\mathbf{v}_{k-1})^T Q \mathbf{v}_{k-1} \\
&= 0.
\end{aligned}$$

Agora, para $i \in \{0, 1, ..., k-2\}$, temos $(\mathbf{v}_i)^T \mathbf{r}_k = (\mathbf{v}_i)^T \mathbf{r}_{k-1} + \alpha_{k-1} (\mathbf{v}_i)^T Q \mathbf{v}_{k-1} = 0$ pela hipótese de indução e pelo motivo de que $\mathbf{v}_i$ são vetores Q-conjugados. Portanto $(\mathbf{r}_k)^T \mathbf{v}_i = 0$ para $i \in \{0, 1, ..., k-1\}$ como queríamos. □

Note que o desenvolvimento realizado no **Teorema 5.8** não impomos nenhuma condição além de serem vetores conjugados. Também estávamos supondo que conhecemos os vetores Q-conjugados.

Voltando ao problema de como obter as direções conjugadas, podemos considerar a seguinte sequência de vetores $\{\mathbf{v}_n\}_{n=0}^{k-1}$, onde cada vetor é a combinação linear da direção contrária do gradiente, isto é, $\mathbf{r}_k = Q\mathbf{x}^{(k)} - \mathbf{q}$ mais a direção do vetor anterior $\mathbf{v}_{k-1}$. Ou seja, $\mathbf{v}_{k+1} = -\mathbf{r}_{k+1} + \beta_{k+1} \mathbf{v}_k$, onde $\beta_k \in \mathbb{R}$ é obtido impondo o fato de que $\mathbf{v}_{k-1}$ e $\mathbf{v}_k$ sejam Q-conjugados.

Podemos pensar nesse β_k como sendo o tamanho do passo de modo a obter o próximo vetor conjugado da sequência. Conseguimos calculá-lo explicitamente, da seguinte forma. Multiplicando $\mathbf{v}_{k+1} = -\mathbf{r}_{k+1} + \beta_{k+1} \mathbf{v}_k$ por $\mathbf{v}_k^T Q$, temos $\mathbf{v}_k^T Q \mathbf{v}_{k+1} = -\mathbf{v}_k^T Q \mathbf{r}_k + \beta_{k+1} \mathbf{v}_k^T Q \mathbf{v}_k$. Como $\mathbf{v}_k^T Q \mathbf{v}_{k+1} = 0$, segue $\beta_{k+1} = \dfrac{(\mathbf{v}_k)^T Q \mathbf{r}_{k+1}}{(\mathbf{v}_k)^T Q \mathbf{v}_k}$. Note que através dessa relação, precisamos saber apenas o vetor da iteração anterior para construir o próximo. Dessa maneira temos o algoritmo:

1 `Algoritmo 5.2`
2 `Dados` $x^{(0)}$, $v_0 = -r_0$, $k = 0$
3 `Equanto` $||r_k|| \neq 0$
4 $\quad \alpha_k = -\dfrac{(r_k)^T v_k}{(v_k)^T Q v_k}$
5 $\quad x^{(k+1)} = x^{(k)} + \alpha_k v_k$

6 $r_{k+1} = Qx^{(k+1)} - q$
7 $\beta_{k+1} = \frac{(r_{k+1})^T Q v_k}{(v_k)^T Q v_k}$
8 $v_{k+1} = -r_{k+1} + \beta_{k+1} v_k$
9 k=k+1
10 Fim

Com o objetivo de mostrar que os vetores $\mathbf{v}_k$ construídos dessa maneira são vetores Q-conjugados, vamos precisar definir subespaço de Krylov.

Definição 5.9. Sejam $\mathbf{u} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ e $k \in \mathbb{N}$. O subespaço de Krylov de dimensão k com relação à A e $\mathbf{u}$ é definido por $K(\mathbf{u}, k) = span\{\mathbf{u}, A\mathbf{u}, ..., A^k\mathbf{u}\}$.

Considerando o subespaço de Krylov, $\mathbf{u}$ como sendo $\mathbf{r}_0$ e $A = Q$, o teorema a seguir mostra que os resíduos são ortogonais e que $span\{\mathbf{r}_0, \mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_k\} = span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^k\mathbf{r}_0\}$.

Teorema 5.10. *Caso o método do gradiente conjugado (Algoritmo 5.2) não encontre a solução em até k iterações, então:*

(i) $(\boldsymbol{r}_k)^T \boldsymbol{r}_i = 0$ *para* $i \in \{0, ..., k-1\}$.
(ii) $span\{\boldsymbol{r}_0, \boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_k\} = K(\boldsymbol{r}_0, k)$.
(iii) $span\{\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_k\} = K(\boldsymbol{r}_0, k)$.
(iv) $(\boldsymbol{v}_k)^T Q \boldsymbol{v}_i = 0$ *para* $i \in \{1, ..., k-1\}$.

Ainda, a sequência gerada pelo Algoritmo 5.2 converge para a solução em, no máximo, n iterações.

Dem. Vamos mostrar *(ii)* e *(iii)* por indução sobre k. Para $k = 0$ temos $span\{\mathbf{r}_0\} = span\{\mathbf{r}_0\}$ e $span\{\mathbf{v}_0\} = span\{\mathbf{r}_0\}$. Suponha que seja verdade para k, isto é,

$$span\{\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_k\} = span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^k\mathbf{r}_0\},$$
$$span\{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_k\} = span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^k\mathbf{r}_0\}.$$

Vamos ver que $span\{\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_{k+1}\} \subset K(\mathbf{r}_0, k+1)$. Pela hipótese de indução $\mathbf{r}_k \in K(\mathbf{r}_0, k)$ e $\mathbf{v}_k \in K(\mathbf{r}_0, k)$. Logo $Q\mathbf{v}_k \in span\{Q\mathbf{r}_0, ..., Q^{k+1}\mathbf{r}_0\}$. Como $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1} Q\mathbf{v}_{k-1}$, ou ainda $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k Q\mathbf{v}_k$ e segue que $\mathbf{r}_{k+1} \in span\{\mathbf{r}_0, ..., Q^{k+1}\mathbf{r}_0\}$. Utilizando a hipótese de indução $span\{\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_k\} = span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^k\mathbf{r}_0\}$, temos o resultado.

Por outro lado, pela hipótese de indução $span\{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_k\} = span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^k\mathbf{r}_0\}$, temos $Q^{k+1}\mathbf{r}_0 \in span\{Q\mathbf{v}_0, ..., Q\mathbf{v}_k\}$. Como $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k Q\mathbf{v}_k$, temos $Q\mathbf{v}_j = \frac{(\mathbf{r}_j - \mathbf{r}_{j+1})}{\alpha_j}$ para $j \in \{0, .., k\}$. Segue que $Q^{k+1}\mathbf{r}_0 \in span\{\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_{k+1}\}$. Pela hipótese de indução, $span\{\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_k\} = span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^k\mathbf{r}_0\}$. Portanto,

$$span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^{k+1}\mathbf{r}_0\} \subset span\{\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_{k+1}\}.$$

Verificaremos o item (iii). Pela linha 8 do *Algoritmo 5.2* podemos escrever $span\{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_{k+1}\} = span\{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_k, \mathbf{r}_{k+1}\}$. Pela hipótese de indução e pelo item (ii),

$$\begin{aligned} span\{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_k, \mathbf{r}_{k+1}\} = \\ span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^k\mathbf{r}_0, \mathbf{r}_{k+1}\} = \\ span\{\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_k, \mathbf{r}_{k+1}\} = K(\mathbf{r}_0, k+1). \end{aligned}$$

(iv) Note que se multiplicarmos a linha 8 do *Algoritmo 5.2* por $Q\mathbf{v}_i$ pela esquerda, onde $i \in \{0, ..., k\}$ e substituir por β_{k+1} obtemos que $(\mathbf{v}_{k+1})^T Q\mathbf{v}_i = 0$ para $k = i$. No caso $i \leq k-1$ note que pela hipótese de indução os vetores $\mathbf{v}_i$ são conjugados e pelo **Teorema 5.8**, $(\mathbf{r}_{k+1})^T\mathbf{v}_i = 0$ para $i \in \{0, 1, ..., k\}$. Aplicando o item (iii), temos

$$\begin{aligned} Q\mathbf{v}_i \in Qspan\{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_i\} &= span\{Q\mathbf{r}_0, Q^2\mathbf{r}_0, ..., Q^{i+1}\mathbf{r}_0\} \\ &\subset span\{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_k\}, \end{aligned}$$

para $i \in \{0, 1, ..., k-1\}$, onde $Qspan\{\mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_i\} = \alpha_0 Q\mathbf{v}_0 + ... + \alpha_i Q\mathbf{v}_i$, com $\alpha_t \in \mathbb{R}$.

Segue que $(\mathbf{r}_{k+1})^T Q\mathbf{v}_i = 0$ e da hipótese de indução $(\mathbf{v}_{k+1})^T Q\mathbf{v}_i = 0$ com $i \in \{0, 1, .., k\}$. Com isso, mostramos que a construção realizada no *Algoritmo 5.2* nos fornecem $\mathbf{v}_i$ que são vetores Q-conjugado e, pelo **Teorema 5.5**, temos a convergência em, no máximo, n iterações.

(i) Pelo **Teorema 5.8**, temos $(\mathbf{r}_k)^T\mathbf{v}_i = 0$ com $i \in \{1, ..., k-1\}$ e $k \in \{1, ..., n-1\}$. Pela linha 8 do *Algoritmo 5.2*, segue que $\mathbf{v}_i = -\mathbf{r}_i + \beta_i\mathbf{v}_{i-1}$. Desta forma, $\mathbf{r}_i \in span\{\mathbf{v}_i, \mathbf{v}_{i-1}\}$. Ou seja $(\mathbf{r}_k)^T\mathbf{r}_i = (\mathbf{r}_k)^T(-\mathbf{v}_i + \beta_i\mathbf{v}_{i-1}) = -(\mathbf{r}_k)^T\mathbf{v}_i + \beta_i(\mathbf{r}_k)^T\mathbf{v}_{i-1} = 0$ para $i \in \{1, .., k-1\}$, como queríamos. □

Com estes resultados é possível aperfeiçoar o *Algoritmo 5.2* no sentido computacional. Utilizando o **Teorema 5.8** e a linha 8 do *Algoritmo 5.2*, onde temos $\mathbf{v}_{k+1} = -\mathbf{r}_{k+1} + \beta_{k+1}\mathbf{v}_k$, podemos recalcular α_k da linha 4. Temos,

$$\alpha_k = -\frac{(\mathbf{r}_k)^T\mathbf{v}_k}{(\mathbf{v}_k)^T Q\mathbf{v}_k} = -\frac{(\mathbf{r}_k)^T(-\mathbf{r}_k + \beta_k\mathbf{v}_{k-1})}{(\mathbf{v}_k)^T Q\mathbf{v}_k} = \frac{-\mathbf{r}_k^T\mathbf{r}_k + \beta_k\mathbf{r}_k^T\mathbf{v}_{k-1}}{(\mathbf{v}_k)^T Q\mathbf{v}_k} = \frac{(\mathbf{r}_k)^T\mathbf{r}_k}{(\mathbf{v}_k)^T Q\mathbf{v}_k}.$$

Note que agora precisamos acessar o $\mathbf{v}_k$ apenas uma vez para realizar o produto interno $\langle \mathbf{v}_k, Q\mathbf{v}_k \rangle$. Similarmente, conseguimos um aperfeiçoamento em β_k, utilizando **Teorema 5.10** e **Teorema 5.8**. Note que $\mathbf{r}_k = \mathbf{r}_{k-1} + \alpha_{k-1}Q\mathbf{v}_{k-1}$. Logo, para a iteração $k+1$ temos $\mathbf{r}_{k+1} = \mathbf{r}_k + \alpha_k Q\mathbf{v}_k$. Ainda, $\mathbf{r}_{k+1} - \mathbf{r}_k = \alpha_k Q\mathbf{v}_k$. Segue que,

$$\beta_{k+1} = \frac{(\mathbf{r}_{k+1})^T Q\mathbf{v}_k}{(\mathbf{v}_k)^T Q\mathbf{v}_k} = \frac{\alpha(\mathbf{r}_{k+1})^T Q\mathbf{v}_k}{\alpha(\mathbf{v}_k)^T Q\mathbf{v}_k} = \frac{(\mathbf{r}_{k+1})^T(\mathbf{r}_{k+1} - \mathbf{r}_k)}{(\mathbf{v}_k)^T(\mathbf{r}_{k+1} - \mathbf{r}_k)} = \frac{(\mathbf{r}_{k+1})^T(\mathbf{r}_{k+1}) - (\mathbf{r}_{k+1})^T\mathbf{r}_k}{(\mathbf{v}_k)^T(\mathbf{r}_{k+1} - \mathbf{r}_k)}$$

$$= \frac{(\mathbf{r}_{k+1})^T\mathbf{r}_{k+1}}{(\mathbf{v}_k)^T(\mathbf{r}_{k+1} - \mathbf{r}_k)}. \text{ Como } \mathbf{v}_k^T = -\mathbf{r}_k^T + \beta_k\mathbf{v}_{k-1}^T, \text{ segue}$$

$$\frac{(\mathbf{r}_{k+1})^T\mathbf{r}_{k+1}}{(-\mathbf{r}_k^T + \beta_k\mathbf{v}_{k-1}^T)(\mathbf{r}_{k+1} - \mathbf{r}_k)} = \frac{(\mathbf{r}_{k+1})^T\mathbf{r}_{k+1}}{-\mathbf{r}_k^T\mathbf{r}_{k+1} + \mathbf{r}_k^T\mathbf{r}_k + \beta_k\mathbf{v}_{k-1}^T\mathbf{r}_k - \beta_k\mathbf{v}_{k-1}^T\mathbf{r}_k}.$$

Note que $-\mathbf{r}_k^T\mathbf{r}_{k+1} = 0$ e $(\mathbf{r}_k)^T\mathbf{v}_i = 0$ com $i \in \{0, 1, ..., k-1\}$, pelo **Teorema 5.8**. Portanto $\beta_{k+1} = \frac{(\mathbf{r}_{k+1})^T\mathbf{r}_{k+1}}{(\mathbf{r}_k)^T\mathbf{r}_k}$. Desse modo, temos a versão final do algoritmo gradiente conjugado:

1 Algoritmo 5.3
2 Dados $x^{(0)}, v_0 = -r_0$ e $k = 0$
3 Equanto $||r_k|| \neq 0$
4 $\alpha_k = \frac{(r_k)^T r_k}{(v_k)^T Q v_k}$
5 $x^{(k+1)} = x^{(k)} + \alpha_k v_k$
6 $r_{k+1} = r_k + \alpha_k Q v_k$
7 $\beta_{k+1} = \frac{(r_{k+1})^T r_{k+1}}{(r_k)^T r_k}$
8 $v_{k+1} = -r_{k+1} + \beta_{k+1} v_k$
9 k=k+1
10 Fim

5.1 Convergência

Vamos ver que existe uma relação entre os autovalores da matriz Q do sistema $Q\mathbf{x} = \mathbf{q}$ e a velocidade de convergência do *Algoritmo 5.3*. Para este estudo, vamos definir uma norma com relação à matriz Q.

Definição 5.11. Seja $Q \in \mathbb{R}^{n\times n}$ simétrica definida positiva. Dado $\mathbf{x}$ em $\mathbb{R}^n$, a norma de $\mathbf{x}$ com relação à Q é $||\mathbf{x}||_Q = \sqrt{\mathbf{x}^T Q\mathbf{x}}$. □

Seja $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{v}_k$. Como $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \alpha_{k-1}\mathbf{v}_{k-1}$, temos

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k-1)} + \alpha_{k-1}\mathbf{v}_{k-1} + \alpha_k\mathbf{v}_k.$$

Reescrevendo dessa forma até $\mathbf{x}^{(0)}$, temos que $x^{(k+1)} = \mathbf{x}^{(0)} + \alpha_0\mathbf{v}_0 + ... + \alpha_k\mathbf{v}_k$. Aplicando o **Teorema 5.10**(iii) ficamos com

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{x}^{(0)} + \eta_0\mathbf{r}_0 + \eta_1 Q\mathbf{r}_0 + ... + \eta_k Q^k\mathbf{r}_0 \\ &= \mathbf{x}^{(0)} + (\eta_0 + \eta_1 Q + ... + \eta_k Q^k)\mathbf{r}_0\end{aligned}$$

onde $\eta_i \in \mathbb{R}$. Desse modo, para cada iteração $(k+1)$ consideramos o polinômio $\overline{P_k}(Q) = \eta_0 I_n + \eta_1 Q + ... + \eta_k Q^k$, e podemos escrever $\mathbf{x}^{(k+1)} = \mathbf{x}^{(0)} + \overline{P_k}(Q)\mathbf{r}_0$. Utilizaremos esta igualdade nos desenvolvimentos seguintes, que nos auxiliará em dois resultados sobre a convergência.

Lema 5.12. *Seja* $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} - \boldsymbol{q}^T\boldsymbol{x} + c$ *e* $\boldsymbol{x}^*$ *o minimizador de* $f(\boldsymbol{x})$*. Então*

$$\frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}^*||_Q^2 = f(\boldsymbol{x}) - f(\boldsymbol{x}^*).$$

Dem. Note que $\frac{1}{2}||\mathbf{x} - \mathbf{x}^*||_Q^2 = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T Q(\mathbf{x} - \mathbf{x}^*) = \frac{1}{2}(\mathbf{x}^T Q - (\mathbf{x}^*)^T Q)(\mathbf{x} - \mathbf{x}^*) = \frac{1}{2}(\mathbf{x}^T Q\mathbf{x} - \mathbf{x}^T Q\mathbf{x}^* - (\mathbf{x}^*)^T Q\mathbf{x} + (\mathbf{x}^*)^T Q\mathbf{x}^*) = \frac{1}{2}(\mathbf{x}^T Q\mathbf{x} - \mathbf{x}^T\mathbf{q} - (\mathbf{x}^*)^T Q\mathbf{x} + (\mathbf{x}^*)^T\mathbf{q}) = f(\mathbf{x}) - f(\mathbf{x}^*)$. □

Desse modo, pelo **Lema 5.12**, minimizar $f(\mathbf{x})$ no domínio de f é equivalente a minimizar $\frac{1}{2}||\mathbf{x} - \mathbf{x}^*||_Q^2$, pois $f(\mathbf{x}^*)$ é constante. Pelo **Teorema 5.8**, $\mathbf{x}^{(k+1)}$ minimiza $f(\mathbf{x})$ em $\mathfrak{L}$. Logo, $\mathbf{x}^{(k+1)} = argmin\frac{1}{2}||\mathbf{x} - \mathbf{x}^*||_Q^2$ em $\mathfrak{L}$. Pelo **Teorema 5.10** *(iii)*, temos que $\mathfrak{L} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{x}_0 + span\{\mathbf{r}_0, Q\mathbf{r}_0, ..., Q^k\mathbf{r}_0\}\}$. Ou seja, $\overline{P_k}(Q)$ minimiza

$$\frac{1}{2}||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_Q^2 = \frac{1}{2}||\mathbf{x}^{(0)} + \overline{P_k}(Q)\mathbf{r}_0 - \mathbf{x}^*||_Q^2.$$

Assim, para $\mathbf{x}^{(j+1)}$ existe $\overline{P_j}(Q) \in P = \{\beta_0 + \beta_1 Q + ... + \beta_j Q^j : \beta_i \in \mathbb{R}\}$ que minimiza $\frac{1}{2}||\mathbf{x}^{(0)} + P_j(Q)\mathbf{r}_0 - \mathbf{x}^*||_Q^2$. Note que como Q é simétrica definida positiva, dado os autovalores $0 < \lambda_1 \leq ... \leq \lambda_n$ e os autovetores correspondentes $\mathbf{u}_1, ..., \mathbf{u}_n$, os autovetores $\mathbf{u}_i$ e $\mathbf{u}_j$ são ortogonais para $i \neq j$. Este fato é similar ao **Exemplo 5.2**.

Considerando $\mathbf{u}_i$ unitário, podemos aplicar a decomposição espectral da matriz Q [2], e escrevemos $Q = UDU^T$ ou ainda, $Q = \lambda_1\mathbf{u}_1\mathbf{u}_1^T + \lambda_2\mathbf{u}_2\mathbf{u}_2^T + ... + \lambda_n\mathbf{u}_n\mathbf{u}_n^T$. O próximo lema nos diz que se $\mathbf{s}$ é um autovetor de Q, então $\mathbf{s}$ também será um autovetor de $P_k(Q)$.

Lema 5.13. *Seja* λ *autovalor de* Q *com* $\boldsymbol{s}$ *autovetor correspondente. Então* $\boldsymbol{s}$ *é autovetor de* $P_k(Q)$ *com autovalor* $P_k(\lambda)$*, isto é,* $P_k(Q)\boldsymbol{s} = P_k(\lambda)\boldsymbol{s}$*, com* $P_k(Q) \in P$*.*

Dem. Observe que como $Q^j\mathbf{s} = \lambda^j\mathbf{s}$, segue

$$\begin{aligned} P_k(Q)\mathbf{s} &= (\eta_0 I_n + \eta_1 Q + \eta_2 Q^2 + ... + \eta_k Q^k)\mathbf{s} \\ &= \eta_0 I_n\mathbf{s} + \eta_1 Q\mathbf{s} + \eta_2 Q^2\mathbf{s} + ... + \eta_k Q^k\mathbf{s} \\ &= \eta_0\mathbf{s} + \eta_1\lambda\mathbf{s} + \eta_2\lambda^2\mathbf{s} + ... + \eta_k\lambda^k\mathbf{s} \\ &= (\eta_0 + \eta_1\lambda + \eta_2\lambda^2 + ... + \eta_k\lambda^k)\mathbf{s} \\ &= P_k(\lambda)\mathbf{s}. \end{aligned}$$

□

Os autovetores $\mathbf{u}_i$ formam uma base para $\mathbb{R}^n$, pois Q é simétrica definida positiva, e podemos escrever $\mathbf{x}^{(0)} - \mathbf{x}^* = c_1\mathbf{u}_1 + ... + c_n\mathbf{u}_n$, $c_i \in \mathbb{R}$.

Proposição 5.14. *Temos que* $||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||_Q^2 = \lambda_1\boldsymbol{c}_1^2 + ... + \lambda_n\boldsymbol{c}_n^2$

Dem. Observe que

$$\begin{aligned} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2 &= (\mathbf{x}^{(0)} - \mathbf{x}^*)^T Q(\mathbf{x}^{(0)} - \mathbf{x}^*) \\ &= (c_1\mathbf{u}_1 + ... + c_n\mathbf{u}_n)^T Q(c_1\mathbf{u}_1 + ... + c_n\mathbf{u}_n) \\ &= (c_1\mathbf{u}_1^T + ... + c_n\mathbf{u}_n^T)Q(c_1\mathbf{u}_1 + ... + c_n\mathbf{u}_n) \\ &= c_1^2\mathbf{u}_1^T Q\mathbf{u}_1 + ... + c_n^2\mathbf{u}_n^T Q\mathbf{u}_n \end{aligned}$$

Utilizando o fato na última igualdade de que $c_ic_j\mathbf{u}_i^T Q\mathbf{u}_j = 0$ pois $Q\mathbf{u}_j = \lambda_i\mathbf{u}_j$, onde $\mathbf{u}_i$ e $\mathbf{u}_j$ são ortogonais.

$$\begin{aligned} c_1^2\mathbf{u}_1^T Q\mathbf{u}_1 + ... + c_n^2\mathbf{u}_n^T Q\mathbf{u}_n &= c_1^2\mathbf{u}_1^T\lambda_1\mathbf{u}_1 + ... + c_n^2\mathbf{u}_n^T\lambda_n\mathbf{u}_n \\ &= \lambda_1\mathbf{c}_1^2 + ... + \lambda_n\mathbf{c}_n^2 \end{aligned}$$

Pois u_i são unitários. □

O último resultado auxiliar nos fornece uma maneira de calcular a norma $||.||_Q^2$ de um vetor qualquer em $\mathbb{R}^n$.

Lema 5.15. *Dado* $\boldsymbol{y} \in \mathbb{R}^n$ *então* $||\boldsymbol{y}||_Q^2 = \sum_{i=1}^{n} \lambda_i(\boldsymbol{u}_i^T\boldsymbol{y})^2$.

Dem. Observe que $||\mathbf{y}||_Q^2 = \mathbf{y}^T Q\mathbf{y} = \mathbf{y}^T\left(\sum_{i=1}^{n}\lambda_i\mathbf{u}_i\mathbf{u}_i^T\right)\mathbf{y} = \sum_{i=1}^{n}\lambda_i\mathbf{y}^T\mathbf{u}_i\mathbf{u}_i^T\mathbf{y} =$

$= \sum_{i=1}^{n}\lambda_i(\mathbf{u}_i^T\mathbf{y})^2$ como queríamos. □

Note que $\mathbf{r}_0 = Q\mathbf{x}^{(0)} - \mathbf{q} = Q\mathbf{x}^{(0)} - Q\mathbf{x}^* = Q(\mathbf{x}^{(0)} - \mathbf{x}^*)$. Então,

$$\begin{aligned}\mathbf{x}^{(k+1)} = \mathbf{x}^{(0)} + \overline{P_k}(Q)\mathbf{r}_0 \Leftrightarrow \mathbf{x}^{(k+1)} - \mathbf{x}^* &= \mathbf{x}^{(0)} + \overline{P_k}(Q)\mathbf{r}_0 - \mathbf{x}^* \\ &= \mathbf{x}^{(0)} + \overline{P_k}(Q)Q(\mathbf{x}^{(0)} - \mathbf{x}^*) - \mathbf{x}^* \\ &= (\mathbf{x}^{(0)} - \mathbf{x}^*) + \overline{P_k}(Q)Q(\mathbf{x}^{(0)} - \mathbf{x}^*) \\ &= (I_n + \overline{P_k}(Q)Q)(\mathbf{x}^{(0)} - \mathbf{x}^*)\end{aligned}$$

Como $\mathbf{x}^{(0)} - \mathbf{x}^* = \sum_{i=1}^{n} c_i\mathbf{u}_i$, segue

$$\begin{aligned}(I_n + \overline{P_k}(Q)Q)(\mathbf{x}^{(0)} - \mathbf{x}^*) &= (I_n + \overline{P_k}(Q)Q)(\sum_{i=1}^{n} c_i\mathbf{u}_i) \\ &= \sum_{i=1}^{n}(I_n c_i\mathbf{u}_i + c_i\overline{P_k}(Q)Q\mathbf{u}_i) \\ &= \sum_{i=1}^{n}(I_n c_i\mathbf{u}_i + c_i\overline{P_k}(Q)\lambda_i\mathbf{u}_i) \\ &= \sum_{i=1}^{n}(I_n c_i\mathbf{u}_i + c_i\overline{P_k}(\lambda_i)\lambda_i\mathbf{u}_i) \\ &= \sum_{i=1}^{n}(1 + \lambda_i\overline{P_k}(\lambda_i))c_i\mathbf{u}_i.\end{aligned}$$

Portanto $\mathbf{x}^{(k+1)} - \mathbf{x}^* = \sum_{i=1}^{n}(1 + \lambda_i\overline{P_k}(\lambda_i))c_i\mathbf{u}_i$. Pelo **Lema 5.15**,

$$\begin{aligned}||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_Q^2 &= \sum_{i=1}^{n}\lambda_i(\mathbf{u}_i^T(1 + \lambda_i\overline{P_k}(\lambda_i))c_i\mathbf{u}_i)^2 \\ &= \sum_{i=1}^{n}\lambda_i(\mathbf{u}_i^T c_i\mathbf{u}_i + \lambda_i\overline{P_k}(\lambda_i)c_i\mathbf{u}_i^T\mathbf{u}_i)^2 \\ &= \sum_{i=1}^{n}\lambda_i(c_i + \lambda_i\overline{P_k}(\lambda_i)c_i)^2 \\ &= \sum_{i=1}^{n}\lambda_i((1 + \lambda_i\overline{P_k}(\lambda_i))c_i)^2 \\ &= \sum_{i=1}^{n}\lambda_i(1 + \lambda_i\overline{P_k}(\lambda_i))^2(c_i)^2\end{aligned}$$

Como $\overline{P_k}(\lambda)$ minimiza $||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_Q^2$, temos que

$$||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_Q^2 = \min_{P_k \in P}\sum_{i=1}^{n}\lambda_i(1 + \lambda_i P_k(\lambda_i))^2(c_i)^2$$

Com $P = \{\beta_0 + \beta_1 x + ... + \beta_k x^k : \beta_i \in \mathbb{R}\}$. Considerando o

$$\max\{(1 + \lambda_1 P_k(\lambda_1))^2, ..., (1 + \lambda_n P_k(\lambda_n))^2\} = L.$$

Segue que

$$\min_{P_k \in P} \sum_{i=1}^{n} \lambda_i (1 + \lambda_i P_k(\lambda_i))^2 (c_i)^2 \leq \min_{P_k \in P} \sum_{i=1}^{n} \lambda_i L c_i^2 \leq \min_{P_k \in P} L \sum_{i=1}^{n} \lambda_i c_i^2.$$

Pelo **Lema 5.14**,

$$\min_{P_k \in P} L \sum_{i=1}^{n} \lambda_i c_i^2 = \min_{P_k \in P} L ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2.$$

Portanto

$$||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_Q^2 \leq \min_{P_k \in P} \max_{1 \leq i \leq n} \{(1 + \lambda_i P_k(\lambda_i))^2\} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2 \tag{5.1}$$

Note que $T(\lambda) = 1 + \lambda P_k(\lambda)$ é um polinômio de grau $k+1$ onde $T(0) = 1$, isto é, os polinômios que resolvem (5.1) quando aplicados em zero, resultam em um. O próximo resultado nos diz que é possível o *Algoritmo 5.3* convergir para a solução $\mathbf{x}^*$ em r iterações onde $r < n$, com $Q \in \mathbb{R}^{n \times n}$.

Teorema 5.16. *Seja a matriz $Q \in \mathbb{R}^{n \times n}$, simétrica definida positiva com os autovalores distintos $0 < \lambda_1 < \lambda_2 < ... < \lambda_r$. Então o Algoritmo 5.3 converge, no máximo, em r iterações.*

Dem. Considere o polinômio $K_r(\lambda) = \dfrac{(-1)^r}{\lambda_1 \lambda_2 ... \lambda_r}(\lambda - \lambda_1)(\lambda - \lambda_2)...(\lambda - \lambda_r)$. Observe que $K_r(\lambda_j) = 0$ com $j \in \{1, ..., n\}$, pois $\lambda_j \in \{\lambda_1, ..., \lambda_r\}$. Temos também que $K_r(0) = 1$, pois $K_r(0) = \dfrac{(-1)^r}{\lambda_1 \lambda_2 ... \lambda_r}(-\lambda_1)(-\lambda_2)...(-\lambda_r) = \dfrac{(-1)^{(2r)} \lambda_1 \lambda_2 ... \lambda_r}{\lambda_1 \lambda_2 ... \lambda_r} = 1.$

Seja $V_r(\lambda) = K_r(\lambda) - 1$. Note que V é um polinômio de grau r com raiz 0, pois $V_r(0) = K_r(0) - 1 = 0$. Defina $S_{r-1}(\lambda) = \dfrac{V_r(\lambda)}{(\lambda - 0)}$. Como estamos dividindo $V_r(\lambda)$ por uma raiz, $S_{r-1}(\lambda)$ tem grau $r - 1$.

Desse modo, pela desigualdade (5.1), temos com $P = \{\beta_0 + \beta_1 x + ... + \beta_{r-1} x^{r-1} : \beta_i \in \mathbb{R}\}$ que

$$\begin{aligned} ||\mathbf{x}^{(r)} - \mathbf{x}^*||_Q^2 &\leq \min_{P_{r-1} \in P} \max_{1 \leq i \leq n} \{(1 + \lambda_i P_{r-1}(\lambda_i))^2\} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2 \\ &\leq \max_{1 \leq i \leq n} \{(1 + \lambda_i S_{r-1}(\lambda_i))^2\} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2 \\ &= \max_{1 \leq i \leq n} (K_r(\lambda_i))^2 ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2 \\ &= 0 \end{aligned}$$

Portanto $\mathbf{x}^{(r)} = \mathbf{x}^*$ como queríamos. □

Vamos mostrar um lema que será utilizado em outro teorema, que nos fornece uma outra caracterização da convergência do *Algoritmo 5.3*.

Lema 5.17. *Seja $p(x)$ um polinômio de grau s. Então o polinômio $p'(x)$ tem suas $s-1$ raízes entre as raízes de $p(x)$.*

Dem. Seja x_i uma raiz de $p'(x)$, onde $i \in \{1, ..., s-1\}$ e y_k a raiz de $p(x)$ com $k \in \{1, ..., s\}$. Pelo Teorema do Valor Médio [12], existe $c \in (y_k, y_{k+1})$ de modo que $p(y_{k+1}) - p(y_k) = p'(c)(y_{k+1} - y_k) \Leftrightarrow p'(c) = 0$. Logo $c = x_i$ para algum i e x_i está entre as raízes de $p(x)$.

□

Teorema 5.18. *Sejam $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ os autovalores da matriz Q de modo que para $m \in \{1, ..., n-1\}$ os $n-m$ autovalores pertençam ao intervalo $[a, b]$, com $a > 0$ e os m autovalores restantes sejam maiores do que b. Então*

$$||\boldsymbol{x}^{(m+1)} - \boldsymbol{x}^*||_Q^2 \leq \left(\frac{\lambda_{n-m} - \lambda_1}{\lambda_{n-m} + \lambda_1}\right)^2 ||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||_Q^2. \tag{5.2}$$

Dem. Considere a desigualdade (5.1). Defina o polinômio $q(\lambda) = 1 + \lambda P_m(\lambda)$, onde $P_m \in P = \{\beta_0 + \beta_1 x + ... + \beta_m x^m : \beta_i \in \mathbb{R}\}$, de modo que $\dfrac{a+b}{2}$ e os m autovalores $\lambda_{n-m+1} < .... < \lambda_n$ sejam raízes de $q(\lambda)$. Desse modo, para $P_m(\lambda)$ com $\lambda_i \in [a, b]$, temos,

$$||\mathbf{x}^{(m+1)} - \mathbf{x}^*||_Q^2 \leq \min_{P_m \in P} \max_{1 \leq i \leq n} \{(1 + \lambda_i P_m(\lambda_i))^2\} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2.$$

Note que $q(\lambda)$ tem grau $m+1$, como $q'(\lambda)$ irá ter grau m, pelo **Lema 5.17** suas raízes estão entre as raízes de $q(\lambda)$. Pelo mesmo motivo, $q''(\lambda)$ com grau $m-1$, tem suas raízes entre as raízes de $q'(\lambda)$.

Por construção $q(\lambda)$ não tem raiz no intervalo $(-\infty, \frac{a+b}{2})$, logo $q''(\lambda)$ não muda de sinal neste intervalo. Como $q''(\lambda) > 0$, pois se $q''(\lambda) < 0$ em $(-\infty, \frac{a+b}{2})$ então $q'(\lambda)$ é decrescente em $(-\infty, \frac{a+b}{2})$, logo $q'(\lambda) > 0$ em $(-\infty, \frac{a+b}{2})$. Logo $q(\lambda)$ é crescente em $(-\infty, \frac{a+b}{2})$. Contradição. Segue que $q(\lambda)$ é convexa em $(-\infty, \frac{a+b}{2})$ e em particular $[0, \frac{a+b}{2}]$, isto é

$$q\left(t0 + (1-t)\frac{a+b}{2}\right) \leq tq(0) + (1-t)q\left(\frac{a+b}{2}\right), \quad t \in [0, 1]$$
$$q\left((1-t)\frac{a+b}{2}\right) \leq t$$

Fazendo $\lambda = (1-t)\frac{a+b}{2}$ obtemos $t = 1 - \frac{2\lambda}{a+b}$. Desse modo, $q(\lambda) \leq 1 - \frac{2\lambda}{a+b}$ em $[0, \frac{a+b}{2}]$. Note que $q(\lambda) \geq 1 - \frac{2\lambda}{a+b}$ em $[\frac{a+b}{2}, b]$, pois $q(\lambda)$ esta abaixo da reta $1 - \frac{a+b}{2}$ em $[0, \frac{a+b}{2}]$ onde $\frac{a+b}{2}$ é raiz de $q(\lambda)$. Como $q(\lambda)$ não tem raiz no intervalo $(-\infty, \frac{a+b}{2})$ e existem m

autovalores restantes que são raízes de $q(\lambda)$ maiores que b, segue que $q(\lambda)$ é crescente em $[\frac{a+b}{2}, b]$, logo maior que $1 - \frac{2\lambda}{a+b}$. Seque que, $|1 + \lambda P_m(\lambda)| \leq |1 - \frac{2\lambda}{a+b}|$ em $[a, b]$. Ou seja,

$$\begin{aligned}
||\mathbf{x}^{(m+1)} - \mathbf{x}^*||_Q^2 &\leq \min_{P_m \in P} \max_{1 \leq i \leq n} \{(1 + \lambda_i P_m(\lambda_i))^2\} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2. \\
&\leq |1 - \frac{2\lambda_i}{a+b}|^2 ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2, \quad \lambda_i \in [a, b] \\
&\leq |1 - \frac{2a}{a+b}|^2 ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2 \\
&\leq \left(\frac{b-a}{a+b}\right)^2 ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2
\end{aligned}$$

Com $a = \lambda_1$ e $b = \lambda_{n-m}$, temos (5.2).

□

O **Teorema 5.18** nos da uma informação importante e ao mesmo tempo interessante. Primeiro note que,

$$\left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right) = \left(\frac{\lambda_1(\frac{\lambda_{n-k}}{\lambda_1} - 1)}{\lambda_1(\frac{\lambda_{n-k}}{\lambda_1} + 1)}\right) = \left(\frac{\frac{\lambda_{n-k}}{\lambda_1} - 1}{\frac{\lambda_{n-k}}{\lambda_1} + 1}\right).$$

Desse modo, se λ_{n-k} está próximo de λ_1, então $\dfrac{\lambda_{n-k}}{\lambda_1} \approx 1$. Isto é, $\dfrac{\lambda_{n-k}}{\lambda_1} - 1 \approx 0$. E portanto $\mathbf{x}^{(k+1)} \approx \mathbf{x}^*$. Ou ainda, podemos dizer que a proximidade dos autovalores com relação a λ_1 é levado em consideração na convergência do *Algoritmo 5.3*.

Quanto maior é a acumulação entorno de λ_1 dos autovalores $\lambda_2, \lambda_3, ..., \lambda_n$, maior é a velocidade de convergência. Pode ocorrer também que tenhamos autovalores acumulados mas que não estão próximos de λ_1. Porém ainda assim, o algoritmo converge rápido.

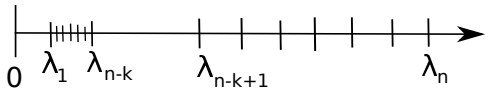

Figura 5 – Distribuição dos Autovalores

Exemplo 5.19. Vamos aplicar o *Algoritmo 5.3* em duas matrizes em que os autovalores estão agrupados ou acumulados e verificar se a desigualdade (5.2) corrobora com a prática.

Em cada matriz simétrica definida positiva, agrupamos os autovalores em três intervalos, $\left(\frac{1}{2}, \frac{3}{2}\right)$, $(19, 21)$ e $(29, 31)$, onde variamos a quantidade em cada um deles, indicado com % na Tabela 1 e Tabela 2.

As tabelas mostram a quantidade de autovalores (ordem da matriz Q), o número de iterações necessárias para uma aproximação da ordem de 10^{-5} da solução exata, isto é $||\mathbf{r}_k|| < 10^{-5}$, e a % de autovalores distribuídos em cada intervalo, com o menor autovalor $\lambda_1 \in \left(\frac{1}{2}, \frac{3}{2}\right)$.

n	20	20	20	20	20	20
$\left(\frac{1}{2}, \frac{3}{2}\right)$	45% (9)	80% (16)	20% (4)	100%	0%	0%
(19,21)	0%	0%	40% (8)	0%	100%	0%
(29,31)	55% (11)	20% (4)	40% (8)	0%	0%	100%
Iterações	14	11	13	8	3	3

Tabela 1 – Iterações realizadas para cada matriz de ordem 20 e a % dos autovalores em cada intervalo

n	2500	2500	2500	2500	2500	2500
$\left(\frac{1}{2}, \frac{3}{2}\right)$	45% (1125)	96% (2400)	20% (500)	100%	0%	0%
(19,21)	0%	0%	40% (1000)	0%	100%	0%
(29,31)	55% (1375)	4% (100)	40% (1000)	0%	0%	100%
Iterações	21	21	26	9	3	3

Tabela 2 – Iterações realizadas para cada matriz de ordem 2500 e a % dos autovalores em cada intervalo

Note que em todos os casos, o número de iterações necessárias foi menor do que a dimensão da matriz Q correspondente, como esperado pelo **Teorema 5.16**.

Note também que quando temos todos os autovalores próximos de λ_1 como na Tabela 1 ou 2, onde temos 100% em cada um dos intervalos, o algoritmo converge mais rápido do que quando temos os autovalores distribuídos de maneira uniforme em dois ou três intervalos. O que já era esperado pelo **Teorema 5.18**.

É possível chegar em uma caracterização da convergência do *Algoritmo 5.3* similar à desigualdade (5.2) utilizando o número de condição da matriz Q, $\kappa(Q)$.

Teorema 5.20. *Seja a matriz $Q \in \mathbb{R}^{n\times n}$, simétrica definida positiva e seja $\kappa(Q)$ o número de condição da matriz Q. Então a sequência $\{\boldsymbol{x}^{(k)}\}$ gerada pelo Algoritmo 5.3 satisfaz,*

$$||\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*||_Q^2 \leq 2\left(\frac{\sqrt{\kappa(Q)}-1}{\sqrt{\kappa(Q)}+1}\right)^k ||\boldsymbol{x}^{(0)} - \boldsymbol{x}^*||_Q^2.$$

Dem. Primeiro vamos observar que o polinômio que resolve o problema

$$\min_{P_k \in P} \max_{1\leq i\leq n} \{(1+\lambda_i P_k(\lambda_i))^2\}$$

são os polinômios de Chebyshev [18]. Tais polinômios são definidos, onde t é o grau, como

$$T_t(x) = \frac{1}{2}\left((x+\sqrt{x^2-1})^t + (x-\sqrt{x^2-1})^t\right). \tag{5.3}$$

Uma propriedade desses polinômios é que $T(0) = 1$ e que $|T_t(x)| \leq 1$ com $x \in [-1, 1]$. Desse modo defina o polinômio,

$$P_k(\lambda) = \frac{T_k\left(\dfrac{\lambda_{max}+\lambda_{min}-2\lambda}{\lambda_{max}-\lambda_{min}}\right)}{T_k\left(\dfrac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)}.$$

O $P_k(\lambda)$ definido dessa forma resolve o problema $\min\limits_{P_k \in P} \max\limits_{1\leq i\leq n}\{(1+\lambda_i P_k(\lambda_i))^2\}$ [18].

Como $-1 \leq T_k\left(\dfrac{\lambda_{max}+\lambda_{min}-2\lambda}{\lambda_{max}-\lambda_{min}}\right) \leq 1$ com $\left(\dfrac{\lambda_{max}+\lambda_{min}-2\lambda}{\lambda_{max}-\lambda_{min}}\right) \in [-1, 1]$. Segue que,

$$\begin{aligned}
||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_Q^2 &\leq \min_{P_k \in P} \max_{1\leq i\leq n}\{(1+\lambda_i P_k(\lambda_i))^2\}||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2 \\
&= P_k(\lambda_i)||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2, \quad \text{para algum } i \in \{1, ..., n\} \\
&\leq T_k\left(\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)^{-1} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2 \\
&= T_k\left(\frac{\kappa(Q)+1}{\kappa(Q)-1}\right)^{-1} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2. \qquad (5.4)
\end{aligned}$$

Aplicando (5.3), temos que (5.4) é igual à,

$$2\left[\left(\frac{\sqrt{\kappa(Q)}+1}{\sqrt{\kappa(Q)}-1}\right)^k + \left(\frac{\sqrt{\kappa(Q)}-1}{\sqrt{\kappa(Q)}+1}\right)^k\right]^{-1} ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2.$$

Como $\lim\limits_{k\to\infty}\left(\dfrac{\sqrt{\kappa(Q)}-1}{\sqrt{\kappa(Q)}+1}\right)^k = 0$, segue a desigualdade desejada,

$$||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_Q^2 \leq 2\left(\frac{\sqrt{\kappa(Q)}-1}{\sqrt{\kappa(Q)}+1}\right)^k ||\mathbf{x}^{(0)} - \mathbf{x}^*||_Q^2. \tag{5.5}$$

□

5.2 Técnicas de Pré-condicionamento

Iremos estudar nesta seção maneiras de construir um pré-condicionador para o sistema $Q\mathbf{x} = \mathbf{q}$. Isto é, queremos uma matriz $P = P_1P_2$, de modo que o sistema $P_1^{-1}QP_2^{-1}\mathbf{z} = P_1^{-1}\mathbf{q}$ seja mais fácil de resolver que $Q\mathbf{x} = \mathbf{q}$, onde $\mathbf{z} = P_2\mathbf{x}$. Existem várias maneiras de construir um precondicionador [6]. Nesta seção, falaremos de técnicas baseadas na fatoração LU incompleta (ILU).

Seja $Q \in \mathbb{R}^{n\times n}$ uma matriz esparsa e considere $K \subset \{(i,j) \mid 1 \leq i,j \leq n\}$, com $i \neq j$. A fatoração LU incompleta (ILU) consiste em encontrar matrizes esparsas L triangular inferior e U triangular superior de modo que $Q = LU + R$, onde o padrão de esparsidade de L e U depende do subconjunto K, que chamaremos de conjunto padrão zero. O precondicionador P é então dado por $P = LU$, onde $R = Q - P$ é o resíduo da aproximação.

Obtemos as matrizes L e U de modo que os elementos q_{ij} cujo índice $(i,j) \in P$ não são processados na fatoração. Dessa forma, temos o algoritmo de fatoração ILU geral.

```
1 Algoritmo 5.4
2 Para i = 2, ..., n faça
3       Para k = 1, ..., i − 1 e para (i, j) ∉ P faça
4             q_ik = q_ik / q_kk
5             Para j = k + 1, ..., n e (i, j) ∉ P faça
6                   q_ij = q_ij − q_ik q_kj
7             Fim
8       Fim
9 Fim
```

Se definimos o conjunto $K = \{(i,j) \mid a_{ij} = 0\}$, temos a fatoração LU incompleta de nível 0 ($ILU(0)$), onde os fatores L e U possuem o mesmo padrão de esparsidade das partes triangular inferior e triangular superior da matriz Q respectivamente. Observe que o número de elementos não nulos do produto LU é maior que o da matriz Q.

Existem problemas onde a fatoração incompleta $ILU(0)$ não produz um bom precondicionador [6]. Para melhorar a precisão da fatoração introduziremos implementações que se diferem da $ILU(0)$, permitindo a inserção de alguns elementos na estrutura original da matriz, assim os fatores L e U terão mais elementos não nulos do que as partes triangular inferior e superior da matriz Q respectivamente. Para explicar essas implementações, vamos introduzir o conceito de níveis de preenchimento que é atribuído a cada elemento que é processado pela eliminação gaussiana.

Definição 5.21. Seja Q uma matriz esparsa, o nível de preenchimento de um elemento

q_{ij} é definido por

$$niv_{ij} = \left\{ \begin{array}{cc} 0 & se\ q_{ij} \neq 0,\ ou\ i = j \\ \infty & \text{caso contrário} \end{array} \right\}$$

Como a cada iteração este elemento é modificado na linha 6 do *Algoritmo 5.4*, temos que atualizar niv_{ij} por

$$niv(q_{ij}) = niv_{ij} = min\{niv_{ij}, niv_{ik} + niv_{kj} + 1\}. \tag{5.6}$$

Com a definição acima, podemos obter o seguinte conjunto

$$K_m = \{(i, j) \mid niv_{ij} > m\}$$

onde niv_{ij} é o nível de preenchimento depois de todas as atualizações (5.6). Nesse caso, os elementos cujo nível de preenchimento não excede m são mantidos. Com esse conjunto, podemos implementar a fatoração LU incompleta de nível m ($ILU(m)$).

```
1  Algoritmo 5.5
2  Defina niv_ij = 0 onde q_ij ≠ 0
3  Para i = 2,...,n faça
4       Para k = 1,...,i-1 e niv_ij ≤ m faça
5            q_ik = q_ik / q_kk
6            Para j = k+1,...,n Faça
7                 q_ij = q_ij - q_ik q_kj
8                 Atualize niv_ij usando (5.6)
9            Fim
10      Fim
11      Zere os elementos da linha i cujo niv_ij > m
12 Fim
```

Sabemos que a matriz Q é simétrica definida positiva, logo, pode-se adaptar a fatoração Cholesky no *Algoritmo 5.5* e assim obter a implementação que é denominada Cholesky Incompleta de nível m. A Figura 6 mostra a matriz Q, a quantidade de elementos não nulos, indicado por nz, os fatores da decomposição de Cholesky Incompleta de nível zero de Q, definidas como L e L^T, e os fatores da fatoração de Cholesky Incompleta de nível $m = n$, onde n é a ordem da matriz Q, definidas como $L2$ e $L2^T$.

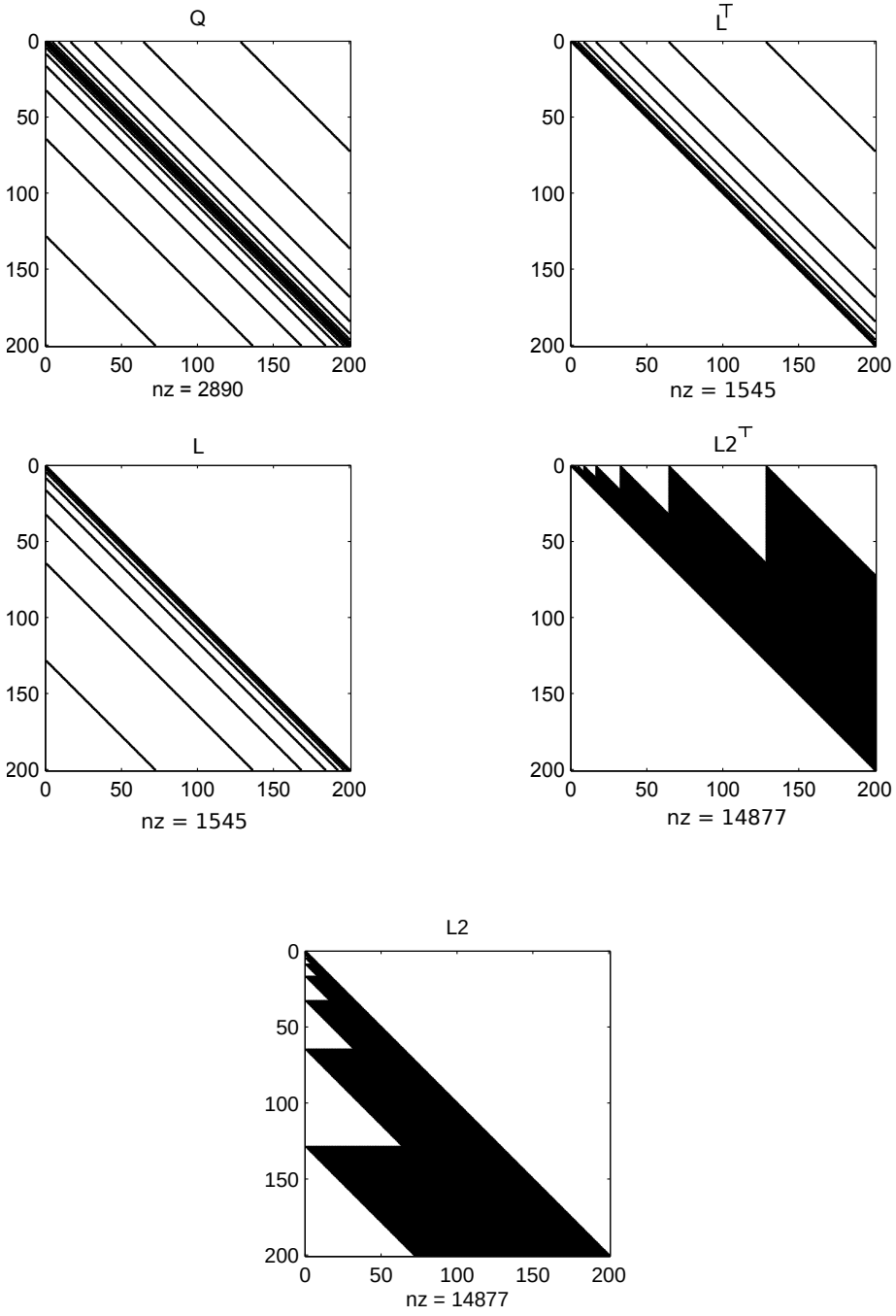

Figura 6 – Fatoração de Cholesky e Cholesky Incompleta da matriz Q

5.3 Pré-Condicionamento

O pré-condicionamento tem como objetivo, através de uma mudança de variável na equação (4.1), aumentar a velocidade de convergência do *Algoritmo 5.2*, fazendo uma redistribuição dos autovalores. Buscamos uma redistribuição que faça com que em (5.2), o número $\dfrac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}$, seja o menor possível, ou em (5.5), de modo que a constante

$\frac{\sqrt{\kappa(Q)}-1}{\sqrt{\kappa(Q)}+1}$, seja a menor possível. Isto é, no primeiro caso, fazer com que os autovalores fiquem agrupados e no segundo caso diminuir o número de condicionamento da matriz obtida através da mudança de variável, com relação ao número de condicionamento da matriz Q.

Seja a quadrática (4.1). Fazendo a mudança $\mathbf{y} = V\mathbf{x}$, onde V é uma matriz inversível em $\mathbb{R}^{n\times n}$. Temos que,

$$\begin{aligned} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{q}^T\mathbf{x} + c \Leftrightarrow f(\mathbf{y}) &= \frac{1}{2}(V^{-1}\mathbf{y})^T Q(V^{-1}\mathbf{y}) - \mathbf{q}^T(V^{-1}\mathbf{y}) + c \\ &= \frac{1}{2}(\mathbf{y}^T V^{-T})Q(V^{-1}\mathbf{y}) - (V^{-T}\mathbf{q})^T\mathbf{y} + c \\ &= \frac{1}{2}\mathbf{y}^T(V^{-T}QV^{-1})\mathbf{y} - (V^{-T}\mathbf{q})^T\mathbf{y} + c \end{aligned}$$

Logo, como feito no Capítulo 4, queremos resolver o sistema $(V^{-T}QV^{-1})\mathbf{y} = V^{-T}\mathbf{q}$. Note que este sistema é equivalente a $Q\mathbf{x} = \mathbf{q}$. De fato, $(V^{-T}QV^{-1})\mathbf{y} = V^{-T}\mathbf{q} \Leftrightarrow V^{-T}Q\mathbf{x} = V^{-T}\mathbf{q} \Leftrightarrow Q\mathbf{x} = \mathbf{q}$.

Seja P a matriz pré-condicionadora. Queremos P de modo que as características descritas no início desta seção ocorram. Uma maneira de obter a P é usar a fatoração de Cholesky Incompleta que abordamos na seção anterior. Note primeiro que a fatoração de Cholesky decompõe Q da forma $Q = LL^T$, onde L é triangular inferior. A fatoração de Cholesky Incompleta nos fornece uma aproximação de L através da matriz S, onde S é esparsa com $S \approx L$, ou seja, $Q \approx SS^T$.

Note que $S^{-1}QS^{-T} \approx S^{-1}SS^TS^{-T} \approx I_n$. Isto é atrativo já que faz uma redistribuição dos autovalores de modo a acelerar a convergência do *Algoritmo 5.3*. Acrescentando a fatoração de Cholesky Incompleta no *Algoritmo 5.3* temos o Gradiente Conjugado com Pré-Condicionamento.

1 Algoritmo 5.6
2 Dados $x^{(0)}$, Q, S e q
3 $P = SS^T$
4 $r_0 = Qx^{(0)} - q$
5 Resolva o sistema $SS^Tt_0 = r_0$
6 $v^{(0)} = -t_0$
7 $k = 0$
8 Enquanto $r_k \neq 0$
9 $\quad \alpha_k = \frac{(r_k)^T t_k}{(v_k)^T Q v_k}$
10 $\quad x^{(k+1)} = x^{(k)} + \alpha_k v_k$
11 $\quad r_{k+1} = r_k + \alpha_k Q v_k$
12 $\quad$ Resolva o sistema $SS^Tt_{k+1} = r_{k+1}$
13 $\quad \beta_{k+1} = \frac{(r_{k+1})^T t_{k+1}}{(r_k)^T t_k}$

```
14          v_{k+1} = -t_{k+1} + β_{k+1} v_k
15          k = k + 1
16  Fim
```

Exemplo 5.22. Vamos aplicar o *Algoritmo 5.3* e o *Algoritmo 5.6* em uma matriz Q esparsa com dimensão 200×200, simétrica definida positiva, onde $\mathbf{q}$ é um vetor arbitrário com entradas em $(0, 1)$.

A Figura 7 mostra nos itens (a) e (b) a convergência do método do gradiente conjugado e do gradiente conjugando com pré-condicionamento respectivamente, com aproximação de 10^{-3} da solução exata, isto é, $||\mathbf{r}_k|| < 10^{-3}$.

Nos itens (c) e (d) temos os 200 autovalores, antes do pré-condicionamento e após, respectivamente. Note que no gradiente conjugado, os autovalores estão entre 0 e 1300 e a convergência ocorreu em 100 iterações.

Após o pré-condicionamento com a fatoração incompleta de Cholesky os autovalores assumiram valores entre 0 e $3,5$, isto é, fez com que os autovalores ficassem agrupados, o que é ideal para acelerar a convergência conforme discutido anteriormente. De fato, com o pré-condicionamento obtemos a aproximação da solução em 5 iterações.

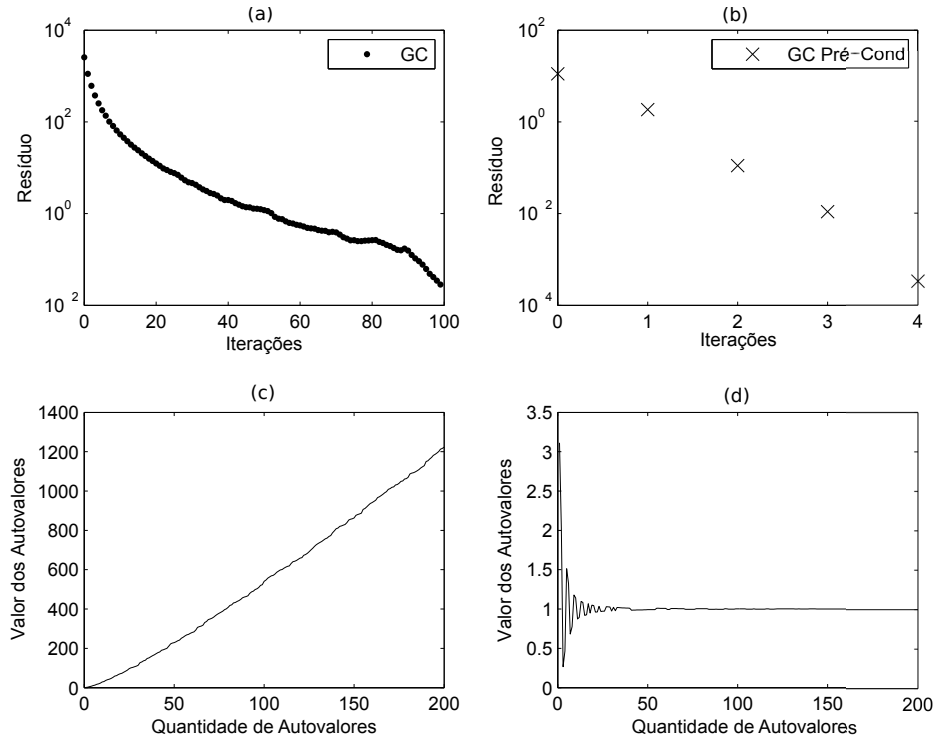

Figura 7 – Gradiente Conjugado e Gradiente Conjugado com Pré-Condicionamento

5.4 Resultados Numéricos

Os resultados numéricos foram obtidos ao resolver o sistema $Q\mathbf{x} = \mathbf{q}$, com Q simétrica definida positiva e esparsa, retiradas de *The SuiteSparse Matrix Collection*. O vetor $\mathbf{q}$ foi considerado como sendo $A\mathbf{x}^*$, com $\mathbf{x}^* = [1, 1, ..., 1]^T$ e o ponto inicial considerado foi o vetor nulo.

Na Tabela 3 esta reunido os seguintes itens. As abreviações J, G.S, Gbb, Gc e Gcp são, respectivamente, o método de Jacobi, Gauss-Seidel, Gradiente Barzilai-Borwein, Gradiente Conjugado e Gradiente Conjugado com Pré-condicionamento. Cada método foi aplicado com um limite de 20000 iterações e um tempo máximo de 1 hora, medido através da função interna *tic* do MATLAB. Os experimentos foram realizados em um computador pessoal com 4Gb de memória e CPU B980 de 2.40Ghz.

Os campos abaixo de cada método são os números de iterações necessárias para uma aproximação da solução exata na ordem de 10^{-3}, isto é, $||\mathbf{r}_k|| < 10^{-3}$, e o tempo necessário. O campo em que não aparece o tempo, a convergência ocorre em menos de 1 segundo.

O campo que aparece N.c significa que não houve convergência com o número máximo de iterações e N.c.t devido ao tempo ter excedido. Os campos $\kappa(Q)$ e Nz, significam o condicionamento, obtido através do comando *condest* do MATLAB e o número de elementos não-nulos de Q, onde a escrita 2.06e+08 significa 2.06×10^8, respectivamente.

As figuras abaixo mostram os gráficos dos métodos do gradiente conjugado com e sem Pré-Condicionamento, SOR com parâmetros 0,5, 1,2 e 1,9, Jacobi, Gauss-Seidel e GBB aplicados à matriz $Trefethen\_200$.

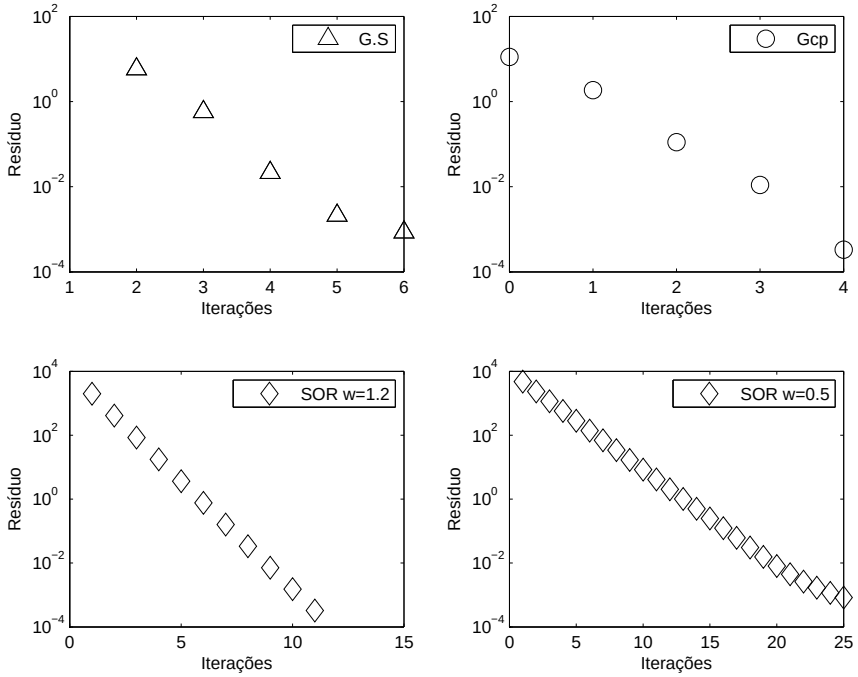

Figura 8 – Gráficos com os algoritmos aplicados à matriz $Trefethen\_200$

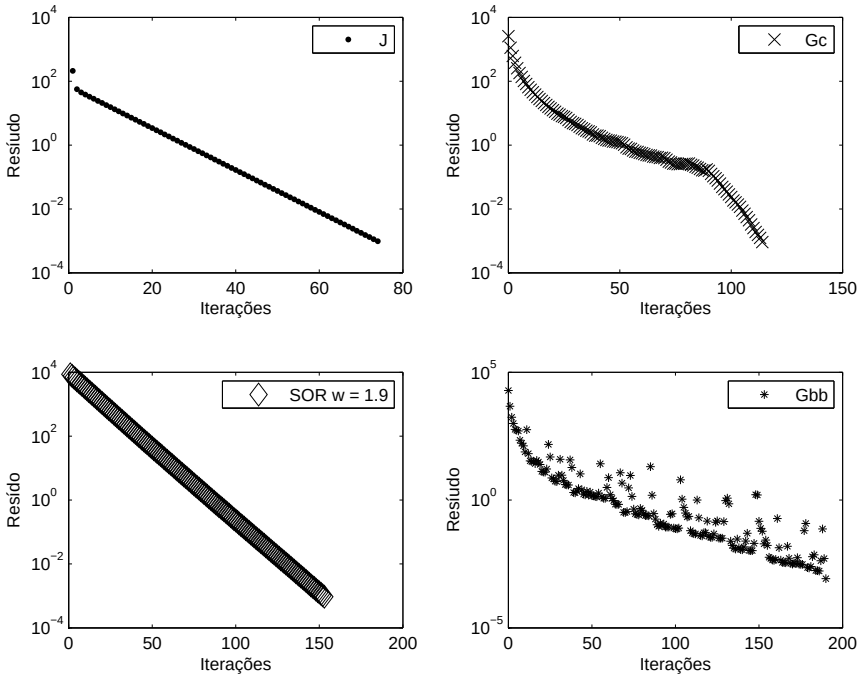

Figura 9 – Gráficos com os algoritmos aplicados à matriz $Trefethen\_200$

Na Figura 8, os gráficos correspondentes aos métodos G.S, Gcp, SOR com parâmetro 1,2 e 0,5 convergiram com 6, 5, 11 e 26 iterações respectivamente. Os gráficos dos métodos J, Gc, SOR com parâmetro 1,9 e Gbb apresentados na Figura 9, convergiram com 75, 115, 153 e 191 respectivamente, desse modo, segue a Tabela 3 com os resultados.

Matriz	n	$\kappa(Q)$	Nz	J	G.S	SOR $\omega = 0.5$	SOR $\omega = 1.2$	SOR $\omega = 1.9$	Gbb	Gc	Gcp
$Trefethen\_20b$	19	44	147 (40%)	21	5	17	9	115	34	19	4
$Trefethen\_20$	20	95	158(39%)	46	4	12	7	83	28	20	5
$mesh1e1$	48	8	306(13%)	30	5	12	7	72	9	7	4
$bcsstm02$	66	9	66(1.5%)	1	1	5	3	31	6	4	1
$bcsstm05$	153	12	153(0.65%)	1	1	8	4	48	12	8	1
$Trefethen\_200$	200	1590	2890(7%)	75	6	26	11	153	191	115	5
$bcsstm07$	420	1.32e+04	7252(4%)	N.c	46	144 4.5s	28	159 4.8s	511	153	9
494_bus	494	3.9e+06	1666(0.68%)	N.c.	N.c	N.c	N.c	6727 3min	N.c	692	74
662_bus	662	8.2e+05	2474(0.56%)	N.c	N.c	N.c	19137 15min	2858 2min	3396	366	47
$msc00726$	726	1.7e+06	34e+03(6.45%)	N.c	3414 7min	10219 21min	2279 5min	1047 2min	N.c	1289	44
$bcsstm08$	1074	8.2e+06	1074(0.09%)	1	1	27 2s	12	172 12s	15255	90	1
$mhd1280b$	1280	5.9e+12	22e+03(1.3%)	N.c	13	26 2.5s	10	106 15s	975	386	3
$bcsstm26$	1922	2.6e+05	1922(0.05%)	1	1	9	4	56	27	47	1
$mhd3200b$	3200	2.0e+13	18e+03(0.17%)	2410 23min	5 3s	7 4s	4 2s	47 37s	50	27	3
$bcsstm24$	3562	1.8e+13	3562(0.02%)	1	1	28 17s	12 8s	182 2min	8682	1794	1
$msc04515$	4515	5.4e+06	97e+03(0.47%)	522 22min	N.c.t	N.c.t	N.c.t	N.c.t	N.c	5190 6s	4060 7s
$crystm01$	4875	421	10e+04(0.42%)	1 2s	1 2s	1 2s	1 2s	1 2s	1	1	1
Muu	7102	155	17e+04(0.33%)	6 35s	3 19s	6 38s	4 25s	34 3min	4	4	1
$aft01$	8205	9.3e+18	12e+049(0.17%)	N.c	N.c.t	N.c.t	N.c.t	N.c.t	1614	4781 14s	71 5s
$fv1$	9604	12	85e+03(0.09%)	47 5min	25 2min	75 8min	17 2min	89 11min	14	11	5

Tabela 3 – Resultados Numéricos

Observe que os métodos não-estacionários (Gbb, Gc e Gcp) tiveram um desempenho superior se comparado com os métodos estacionários (J, G.S e SOR), tanto no quesito de número de iterações, como no de tempo. O motivo é que os métodos estacionários não exploram o fato da matriz Q ser simétrica definida positiva, por outro lado, os métodos não-estacionários exploram este fato, como por exemplo, no cálculo do passo α_k.

Entre todos os métodos, o gradiente conjugado com pré-condicionamento foi o método que teve melhor desempenho. Isto se deve ao fato da redistribuição dos autovalores quando aplica-se o pré-condicionamento, conforme discutido nas seções anteriores.

Nos casos em que o gradiente conjugado precisou de mais iterações que a ordem da matriz do sistema, indica que existe uma perda de vetores Q-conjugados, isto é, $\mathbf{v}_i Q \mathbf{v}_j \neq 0$, $i \neq j$, o que implica a necessidade de mais iterações para ter a igualdade.

6 Minimização de Quadráticas Convexas com Restrições Lineares

Neste capítulo iremos resolver o problema abaixo,

$$\min f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{q}^T\mathbf{x} + c \tag{6.1}$$

$$\text{sujeito a} \qquad A^T\mathbf{x} = \mathbf{b} \tag{6.2}$$

em que $Q \in \mathbb{R}^{n\times n}$ é simétrica definida positiva, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{q} \in \mathbb{R}^n$, $c \in \mathbb{R}$ e $A \in \mathbb{R}^{n\times m}$, com colunas linearmente independentes e $\mathbf{b} \in \mathbb{R}^m$, com $m < n$. O objetivo é transformar o problema restrito (6.1) no irrestrito (4.1), pois este, nós sabemos resolver. Iremos utilizar conceitos como base, espaço nulo e espaço coluna, que podem ser encontrados em [2].

Note que para qualquer vetor $\mathbf{v} \in \mathbb{R}^n$, é possível decompô-lo como uma soma da forma $\mathbf{x} = Z\mathbf{v}_1 + W\mathbf{v}_2$ onde W é a matriz cujas colunas formam uma base para o espaço coluna de A, isto é, $R(A)$. E Z é matriz cujas colunas formam uma base para o espaço nulo de A^T, isto é, $null(A^T)$. Onde $\mathbf{v}_1 \in \mathbb{R}^{n-m}$, $Z \in \mathbb{R}^{n\times(n-m)}$, $W \in \mathbb{R}^{n\times m}$ e $\mathbf{v}_2 \in \mathbb{R}^m$.

Segue que $A^T\mathbf{x} = A^T(Z\mathbf{v}_1 + W\mathbf{v}_2) = A^TZ\mathbf{v}_1 + A^TW\mathbf{v}_2$. Como Z é formado por colunas que são base do espaço nulo de A^T, temos $A^TZ = 0$. Logo $A^T\mathbf{x} = A^TW\mathbf{v}_2 = \mathbf{b}$. Como A^T e W possuem colunas linearmente independentes, o sistema $A^TW\mathbf{v}_2 = \mathbf{b}$ possui única solução $\mathbf{v}_2^*$.

Desse modo $\mathbf{x} = Z\mathbf{v}_1 + W\mathbf{v}_2^*$ satisfaz (6.2) e temos $f(\mathbf{x}) = f(Z\mathbf{v}_1 + W\mathbf{v}_2^*) = g(\mathbf{v}_1)$, de modo que $\mathbf{x}$ elimina as restrições lineares (6.2) quando consideramos a $g(\mathbf{v}_1)$. Isto é, queremos minimizar a função irrestrita

$$\begin{aligned}
g(\mathbf{v}_1) &= \frac{1}{2}(Z\mathbf{v}_1 + W\mathbf{v}_2^*)^TQ(Z\mathbf{v}_1 + W\mathbf{v}_2^*) - \mathbf{q}^T(Z\mathbf{v}_1 + W\mathbf{v}_2^*) + c \\
&= \frac{1}{2}(\mathbf{v}_1^TZ^T + (\mathbf{v}_2^*)^TW^T)Q(Z\mathbf{v}_1 + W\mathbf{v}_2^*) - \mathbf{q}^T(Z\mathbf{v}_1 + W\mathbf{v}_2^*) + c \\
&= \frac{1}{2}(\mathbf{v}_1^TZ^TQZ\mathbf{v}_1 + \mathbf{v}_1^TZ^TQW\mathbf{v}_2^* + (\mathbf{v}_2^*)^TW^TQZ\mathbf{v}_1 + (\mathbf{v}_2^*)^TW^TQW\mathbf{v}_2^*) \\
&- \mathbf{q}^T(Z\mathbf{v}_1 + W\mathbf{v}_2^*) + c \\
&= \frac{1}{2}\mathbf{v}_1^TZ^TQZ\mathbf{v}_1 + \mathbf{v}_1^TZ^T(QW\mathbf{v}_2^* - \mathbf{q}) + (\mathbf{v}_2^*)^TW^TQW\mathbf{v}_2^* - \mathbf{q}^TW\mathbf{v}_2^* + c,
\end{aligned}$$

onde $k = (\mathbf{v}_2^*)^TW^TQW\mathbf{v}_2^* - \mathbf{q}^TW\mathbf{v}_2^* + c$ é constante. Desse modo, fazendo $\nabla(g(\mathbf{v}_1)) = 0$, temos o sistema equivalente,

$$\begin{aligned}\nabla(g(\mathbf{v}_1)) = 0 &\Leftrightarrow \nabla\left(\frac{1}{2}\mathbf{v}_1^T Z^T Q Z \mathbf{v}_1 + \mathbf{v}_1^T Z^T (QW\mathbf{v}_2^* - \mathbf{q}) + k\right) = 0 \\ &\Leftrightarrow \nabla(\frac{1}{2}\mathbf{v}_1^T Z^T Q Z \mathbf{v}_1) + \nabla(\mathbf{v}_1^T Z^T (QW\mathbf{v}_2^* - \mathbf{q})) = 0 \\ &\Leftrightarrow Z^T Q Z \mathbf{v}_1 + Z^T Q W \mathbf{v}_2^* - Z^T \mathbf{q} = 0 \\ &\Leftrightarrow Z^T Q Z \mathbf{v}_1 = -Z^T (QW\mathbf{v}_2^* - \mathbf{q}) \qquad (6.3)\end{aligned}$$

No caso em que a matriz Z^TQZ seja simétrica definida positiva, podemos aplicar o *Algoritmo 5.3*, ou ainda o *Algoritmo 5.6* para encontrar $\mathbf{v}_1$. Portanto, conhecidos $\mathbf{v}_1$ e $\mathbf{v}_2^*$, temos $\mathbf{x}$ que satisfaz (6.1) e (6.2).

Sabemos que o *Algoritmo 5.6* aplica o pré-condicionamento para acelerar a convergência. Desse modo, iremos aplicar o pré-condicionamento em (6.1) com a fatoração de Cholesky Incompleta, de modo análogo à Seção 5.3, onde $Q \approx SS^T$ e $V = S^T$. Realizando a mudança de variável, $\mathbf{y} = V\mathbf{x} = S^T\mathbf{x}$, segue que queremos resolver,

$$\min f(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T(S^{-1}QS^{-T})\mathbf{y} - (S^{-1}\mathbf{q})^T\mathbf{y} + c \qquad (6.4)$$

$$\text{sujeito à} \qquad (A^TS^{-T})\mathbf{y} = \mathbf{b} \qquad (6.5)$$

Escrevendo $\mathbf{y} = Z\mathbf{w}_1 + W\mathbf{w}_2^*$ onde as colunas de Z formam uma base para o espaço nulo de A^TS^{-T} e as colunas da matriz W formam uma base para o espaço coluna de A^TS^{-T}, chegamos no sistema $Z^TS^{-1}QS^{-T}Z\mathbf{w}_1 = Z^T(S^{-1}\mathbf{q} - S^{-1}QS^{-T}W\mathbf{w}_2^*)$. Uma forma de evitar o cálculo explícito da Z é utilizando a matriz de projeção $P = I_n - S^{-1}A(A^TS^{-T}S^{-1}A)^{-1}A^TS^{-T}$ [13]. Dessa forma, o algoritmo pré-condicionado fica,

1 Algoritmo 6.1
2 Dados Q, A^TS^{-T}, b, q
3 Resolva o sistema $A^TS^{-T}y = b$
4 $r_0 = Qy + q$;
5 Resolva o sistema $g_0 = Pr_0$
6 $d_0 = -g_0$
7 $k = 0$
8 Enquanto $||(r_k)^Tg_k|| \neq 0$
9 $\quad \alpha_k = \dfrac{(r_k)^Tg_k}{(d_k)^TQd_k}$
10 $\quad y^{(k+1)} = y^{(k)} + \alpha_k d_k$
11 $\quad r^{(k+1)} = r^{(k)} + \alpha_k Q d_k$
12 $\quad$ Resolva o sistema $g_{k+1} = Pr_{k+1}$
13 $\quad \beta_{k+1} = \dfrac{(r_{k+1})^Tg_{k+1}}{(r_k)^Tg_k}$
14 $\quad d_{k+1} = -g_{k+1} + \beta_{k+1}d_k$

15 $r_k = r_{k+1}$
16 $g_k = g_{k+1}$
17 k=k+1
18 Fim

Note que para resolver a linha 12 temos que,

$$\begin{aligned} P\mathbf{r}_{k+1} &= (I_n - S^{-1}A(A^TS^{-T}S^{-1}A)^{-1}A^TS^{-T})\mathbf{r}_{k+1} \\ &= \mathbf{r}_{k+1} - S^{-1}A(A^TS^{-T}S^{-1}A)^{-1}A^TS^{-T}\mathbf{r}_{k+1} \\ &= \mathbf{r}_{k+1} - S^{-1}A\mathbf{u}_{k+1} \end{aligned}$$

Assim, $(A^TS^{-T}S^{-1}A)^{-1}A^TS^{-T}\mathbf{r}_{k+1} = \mathbf{u}_{k+1}$. Portanto, podemos resolver o sistema $A^TS^{-T}\mathbf{r}_{k+1} = (A^TS^{-T}S^{-1}A)\mathbf{u}_{k+1}$. Desse modo, conhecido $\mathbf{y}$ e resolvendo $V\mathbf{x} = \mathbf{y}$, temos o $\mathbf{x}$ que satisfaz (6.1) e (6.2).

Exemplo 6.1. Na Tabela 4 temos os resultados obtidos pelo *Algoritmo 6.1* com e sem pré-condicionamento para os mesmos sistemas da Tabela 3, com dimensões entre 19 e 494, e restrições lineares $A^T\mathbf{x} = \mathbf{b}$, com ponto inicial o vetor nulo e a matriz A^T de ordem $\frac{n}{2} \times n$ com entradas reais entre $(0, 1)$. Temos os campos com o tempo necessário para realizar a Fatoração de Cholesky Incompleta, identificado como $Ichol$, além das iterações e o tempo realizado com pré-condicionamento (Pre) e sem (Spr).

Matriz	n	*Iterações*		*Tempo*		*Tempo*
		Pre	*Spr*	*Pre*	*Spr*	*Ichol*
Trefethen_20b	19	7	8	6ms	2ms	310 μs
Trefethen_20	20	7	9	7ms	39ms	458 μs
mesh1e1	48	6	7	13ms	2ms	301 μs
bcsstm02	66	3	4	7ms	6ms	297μs
bcsstm05	153	6	7	21ms	39ms	32 μs
Trefethen_200	200	16	19	249ms	27ms	336 μs
bcsstm07	420	98	150	5.9s	617ms	600 μs
494_bus	494	89	119	11ms	736ms	115 μs

Tabela 4 – *Algoritmo 6.1* com e sem pré-condicionamento para problemas esparsos

Segue abaixo o gráfico ao aplicarmos o *Algoritmo 6.1* na matriz *bcsstm07* com uma aproximação de 10^{-3} da solução exata, em que o eixo x representa a iteração e o eixo y na escala logarítmica o resíduo.

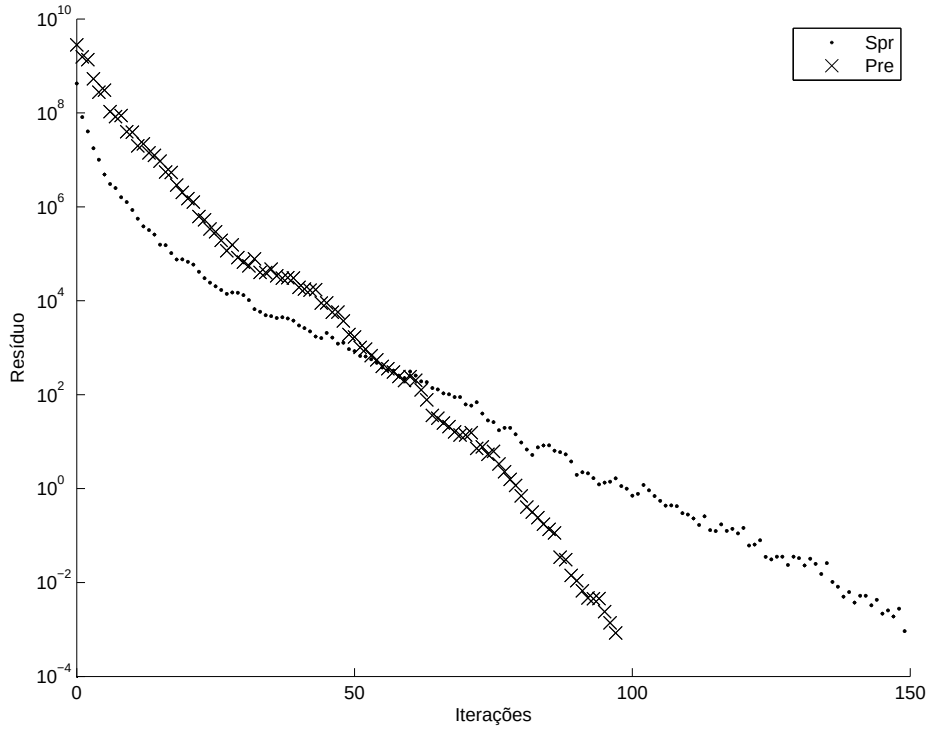

Figura 10 – Gráfico do Gradiente Conjugado com Restrições Lineares

Note que para os sistemas cuja matriz está entre as dimensões 19 e 153, não ocorre diferença de desempenho entre aplicar o *Algoritmo 6.1* com e sem pré-condicionamento. A partir da matriz do sistema com dimensão 420, o desempenho do pré-condicionamento se mostra superior.

Observe que realizar a fatoração de Cholesky Incompleta levou menos de um segundo em todos os casos. Também é importante notar a aplicação da matriz de projeção na linha 12 do *Algoritmo 6.1*, pois se evita calcular a matriz Z, economizando tempo e armazenamento.

7 Conclusão

Vimos neste trabalho diversos modelos para a resolução de sistemas lineares. Iniciamos pelas maneiras clássicas de resolver que são os métodos diretos. Seguindo pelos modelos iterativos estacionários, vimos que o método SOR acelera a convergência do modelo Gauss-Seidel, que depende do parâmetro ω. Neste caso, só há convergência se $0 < \omega < 2$.

Nos métodos não-estacionários vimos algumas propriedades do gradiente e mencionamos também o gradiente proposto por Barzilai-Borwein. No método do gradiente, observamos que a convergência é lenta ao resolvermos o problema de minimizar quadráticas convexas cujas curvas de níveis são elipses com formato alongado. Isto é, a convergência depende do raio espectral da matriz Q.

Nesta situação, abordamos o problema através das direções conjugadas, que apresenta duas características. No caso em que os autovalores estejam distribuídos de maneira uniforme a convergência é lenta, e quando os autovalores estão acumulados na reta e próximos do menor autovalor, a convergência é rápida. Vimos também uma caracterização da convergência do gradiente conjugado pelos polinômios de Chebyshev. No caso em que temos distribuição uniforme dos autovalores, podemos aplicar o pré-condicionamento.

O pré-condicionamento faz uma redistribuição dos autovalores de modo que fiquem acumulados, com o objetivo de acelerar a convergência. A matriz pré-condicionadora utilizada foi a fatoração de Cholesky Incompleta, que faz a aproximação esparsa de um dos fatores da fatoração de Cholesky de Q. Realizamos experimentos numéricos com os métodos estacionários e não-estacionários. Podemos concluir que entre estes dois conjuntos de métodos, o método do gradiente com pré-condicionamento teve o melhor desempenho, seguido do gradiente conjugado.

Por fim, vimos como minimizar quadráticas convexas com restrições lineares. A ideia principal é eliminar a restrição da quadrática através da decomposição da variável em uma soma, que considera a base do espaço nulo e a base do espaço coluna da matriz da restrição linear. Desse modo, finalizamos os experimentos numéricos através do gradiente conjugado com e sem pré-condicionamento para quadráticas convexas com restrições lineares

O desenvolvimento futuro deste trabalho seria resolver o problema de minimizar a função quadrática convexa sujeito a restrições lineares com variáveis canalizadas, isto é, cada componente do vetor que satisfaz a restrição linear esta limitada inferiormente e superiormente por uma constante.

Referências

[1] CALLIOLI, Carlos A. et al. **Álgebra linear e aplicações**. 4. ed. rev. São Paulo: Atual, 1983.

[2] ANTON, Howard; BUSBY, Robert C. **Álgebra linear contemporânea**. Porto Alagre: Bookman, 2006.

[3] BARZILAI, Jonathan; BORWEIN, Jonathan M. Two-Point Step Size Gradient Methods. **IMA Journal of Numerical Analysis**, v. 8, p. 141-148, 1 jan. 1988.

[4] CUNHA, Maria C. **Métodos Numéricos**. 1. ed. Campinas: Ed. da Unicamp, 1993.

[5] SCHÖNLIEB, Carola-Bibiane. **Mathematical Tripos Part II Michaelmas Term 2015**: Numerical Analysis Lecture 17, 2015.

[6] DONGARRA, Jack J. et al. **Numerical Linear Algebra for High-Performance Computers**. Philadelphia: Society for Industrial and Applied Mathematics, 1998.

[7] GOLUB, Gene H.; VAN LOAN, Charles F. **Matrix computations**. 3. ed. Baltimore: Johns Hopkins University Press, 1996.

[8] ISAACSON, Eugene; KELLER, Herbert B. **Analysis of numerical methods**. New York: Dover Publications, 1994

[9] FAIRES, John D.; BURDEN, Richard L. **Numerical analysis**. 8. ed. rev. Belmont, CA: Thomson Brooks/Cole, 2005.

[10] LEON, Steven J. **Linear algebra with applications**. 4. ed. Englewood Cliffs: Prentice-Hall, 1994.

[11] LUENBERGER, David G. **Linear and nonlinear programming**. 2. ed. Boston: Addison-Wesley, 1984.

[12] MARSDEN, Jerrold E.; HOFFMAN, Michael J. **Elementary classical analysis**. 2. ed. New York: W. H. Freeman, 1993.

[13] NOCEDAL, Jorge; WRIGHT, Stephen J. **Numerical optimization**. 2nd ed. New York: Springer, 2006.

[14] CHONG, Edwin K. P.; ZAK, Stanislaw H. **An introduction to optimization**. New York: Wiley Interscience, 1996

[15] QUARTERONI, Alfio; SALERI, Fausto. **Scientific Computing with MATLAB and Octave**. 2. ed. Berlin, Heidelberg: Springer, 2006.

[16] SAAD, Y. **Iterative methods for sparse linear systems**. Boston: PWS, 1996.

[17] SHARIFF, M. H. B. M. A Constrained Conjugate Gradient Method and the Solution of Linear Equations. **Computers Math. Applic.**, v. 30, p. 25-37, 21 dez. 1995.

[18] SHEWCHUK, Jonathan R. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. **School of Computer Science Carnegie Mellon University**, Pittsburgh, p. 41-62, 4 ago. 1994.

[19] LIMA, Elon L. **Geometria analítica e álgebra linear**. 2. ed. Rio de Janeiro: IMPA, 2006.