Estudo de um método baseado em autovalores generalizados para o subproblema de região de confiança[1]

Jean Carlos A. Medeiros
Profa. Dra. Sandra Augusta Santos
Departamento de Matemática Aplicada - IMECC/UNICAMP

21 de fevereiro de 2019

[1]Este trabalho contou com o apoio do CNPq: 137035/2017-9.

Sumário

Introdução

Nosso problema de interesse é a minimização irrestrita:

$$\min_{x \in \mathbb{R}^n} f(x) \tag{PI}$$

onde $f : \mathbb{R}^n \mapsto \mathbb{R}$ e f de classe C^2. Para resolver numericamente problemas do tipo PI são utilizados métodos iterativos, neste trabalho nos concentraremos nos métodos de região de confiança [2, 5].

Os métodos de região de confiança que estudaremos aqui baseiam-se em estipular um modelo quadrático da função objetivo no iterado corrente e uma região em torno desse ponto onde espera-se que o modelo represente bem a função objetivo, essa região é chamada região de confiança; então encontramos uma aproximação $\overline{x}$ para o minimizador do modelo. Se $\overline{x}$ proporcionar uma redução satisfatória, então atualizamos o ponto corrente para $\overline{x}$ e repetimos o processo, se a redução não for satisfatória, isso implica que o modelo não está representando muito bem a função, então reduzimos o raio da região de confiança e repetimos o processo para encontrar um novo minimizador $\overline{x}$.

Diferentemente dos métodos de busca linear [3], que primeiro calculam uma direção de descida e depois encontram um tamanho de passo para andar nessa direção, os métodos de região de confiança vão estabelecer primeiro o tamanho máximo de passo a ser andado, ou seja, o raio da região de confiança e depois vão calcular a direção, resolvendo o subproblema de minimizar o modelo sujeito a região de confiança.

Este trabalho foi dividido em dois capítulos, no primeiro capítulo vamos estudar os métodos de região de confiança mais clássicos, analisando desde as abordagens mais utilizadas na construção do modelo até a resolução do subproblema (de forma exata ou aproximada), de modo a obter a direção que gerará o novo iterado, nossos estudos foram feitos com base em [8, 9]. No segundo capítulo, nosso objetivo é analisar, com uma abordagem geométrica, um método [1] proposto recentemente para resolução do subproblema de região de confiança, que se baseia em autovalores generalizados.

Capítulo 1

Métodos de região de confiança para minimização irrestrita

1.1 Motivação

Antes de começarmos o nosso estudo sobre os métodos de região de confiança, vamos tentar compreender primeiro qual é a motivação por trás desses métodos, pois dessa forma ao nos depararmos com um problema de minimização irrestrita teremos mais elementos para escolher entre um método de busca linear ou um método de região de confiança.

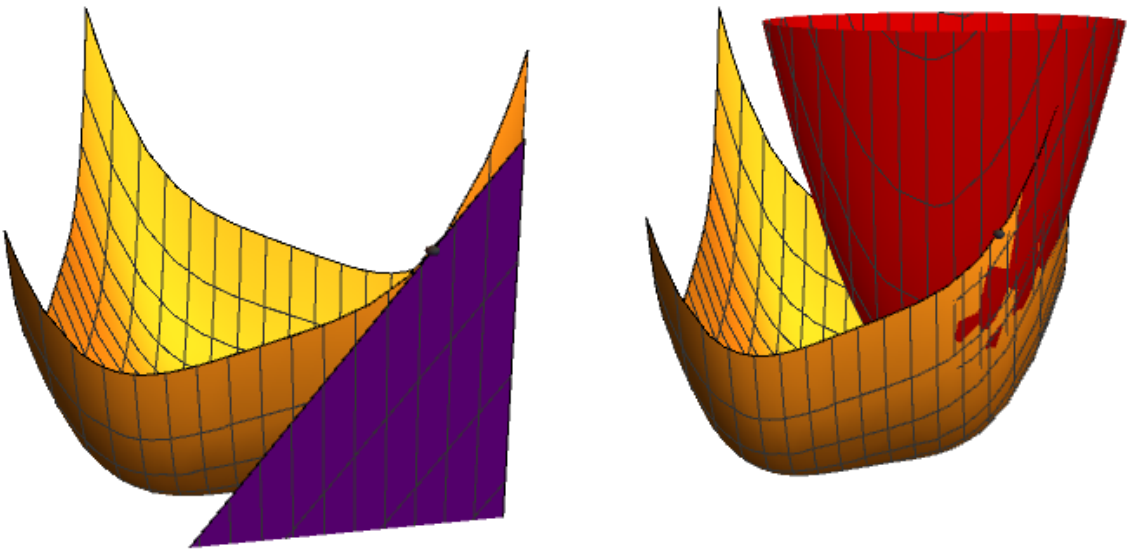
Os métodos mais clássicos de minimização irrestrita são os métodos de busca linear, onde dado um ponto inicial, calculamos uma direção de descida e um tamanho de passo, como por exemplo o Método de Máxima Descida, Método de Newton e Métodos Quasi-Newton. Já os métodos de região de confiança têm uma abordagem diferente, pois a cada iteração é construído um modelo para a função objetivo, geralmente quadrático, que é minimizado, sujeito a uma região em que acredita-se que o modelo seja uma boa aproximação para a função.

À primeira a vista pode parecer algo contraproducente, já que estamos trocando um problema de minimização irrestrita por uma sequência de problemas restritos, mas se analisarmos com cautela veremos que nós temos a vantagem de que essa sequência de problemas restritos tratam sempre de minimizar uma função quadrática, ou seja, a função, o seu gradiente e sua hessiana em todas iterações vão ser sempre da mesma forma.

Note que a facilidade em calcular o vetor gradiente e a matriz hessiana vão depender da função objetivo, quanto mais complicada for a função, mais penoso será calcular o seu vetor gradiente e matriz hessiana. Vale a pena destacar que os subproblemas de região de confiança possuem uma estrutura muito especial, particularmente quando a região é definida com a norma euclidiana, a minimização do modelo quadrático sujeito à bola de confiança pode ser convertido em um problema escalar (unidimensional), o que é bastante favorável. De qualquer forma, para casos de dimensão grande e/ou funções difíceis pode ser vantajoso o uso de métodos de região de confiança.

1.2 Construindo o modelo

No método de região de confiança definimos um modelo para função objetivo a partir de um ponto corrente x^k e estabelecemos uma bola fechada centrada em x^k e com raio Δ_k. Essa vizinhança em torno de x^k é chamada de *região de confiança*, pois nessa região vamos confiar que o modelo gera uma boa aproximação para a função objetivo. O modelo mais simples possível é o linear, que geometricamente falando é o hiperplano tangente à superfície gráfico da função objetivo no ponto corrente. Porém, esse não é o modelo mais eficiente, uma vez que o hiperplano não consegue "capturar" a curvatura da função, ou seja, o modelo só consegue representar bem a função numa vizinhança muito próxima do ponto corrente. Então, apesar de ser um modelo mais fácil de computar, seria necessário um número de iterações maior para conseguir minimizar a função objetivo. Em geral, o modelo quadrático é o mais utilizado, pois é o modelo mais simples capaz de aproximar a curvatura da função objetivo. A figura 1.1 ilustra o gráfico de uma função de duas variáveis e os seus modelos linear e quadrático em torno de um ponto destacado.



(a) Modelo linear (b) Modelo quadrático

Figura 1.1: *Exemplo em $\mathbb{R}^2$. A superfície em laranja representa o gráfico da função objetivo, o plano roxo representa o gráfico do modelo linear e o parabolóide elíptico em vermelho representa o gráfico do modelo quadrático, ambos em torno do mesmo ponto x^k.*

Então seja um ponto corrente $x^k \in \mathbb{R}^n$ e uma matriz $A_k \in \mathbb{R}^{n\times n}$ simétrica e com norma uniformemente limitada, definimos o modelo quadrático da f em torno do ponto x^k da seguinte maneira:

$$m_k(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T A_k d$$

onde $d = x - x^k$, ou seja, $\|d\|$ é distância do ponto corrente x^k até o ponto x.

Os métodos que adotam $A_k = \nabla^2 f(x^k)$ são chamados de Métodos Newtonianos de Região de Confiança. Podemos observar que nesse caso o modelo é o Polinômio de Taylor de ordem 2 da função objetivo em torno de x^k, então sabemos que o erro de aproximação é $O(\|d\|^3)$, ou seja, quanto menor for a distância $\|d\|$, menor será o erro de aproximação. Isso implica que numa vizinhança muito próxima de x^k o modelo é uma boa representação para a função objetivo. Mais ainda, fixada uma região de confiança para os modelos linear e quadrático, a aproximação do modelo quadrático é de fato melhor do que a do modelo linear, pois enquanto

o erro de aproximação do modelo quadrático é $O(\|d\|^3)$, o do modelo linear é $O(\|d\|^2)$. Teoricamente, se conseguirmos garantir a diferenciabilidade além da segunda ordem é possível construir modelos que aproximem a função objetivo ainda melhor; se a função objetivo for de classe C^3, por exemplo, é possível construirmos um modelo cúbico, conforme exemplifica a figura 1.2:

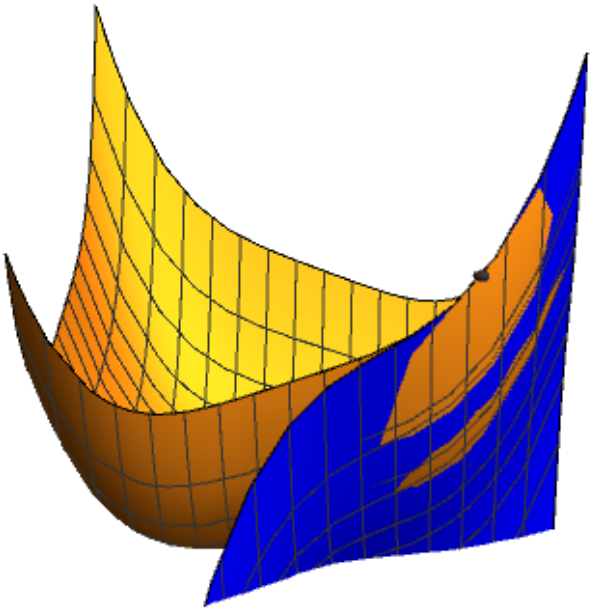


Figura 1.2: *Exemplo em $\mathbb{R}^2$. A superfície em laranja representa o gráfico da função objetivo e a superfície azul representa o gráfico do modelo cúbico em torno do mesmo ponto x^k.*

Porém, o ganho na precisão do modelo pode ser menor que o gasto computacional para se obtê-lo. No caso particular de um modelo cúbico, seria necessário computar não só a hessiana como também o tensor das derivadas terceiras, ambos por iteração, e o custo-benefício pode não valer a pena.

Ainda tendo em vista o gasto computacional, podemos nos perguntar se estamos dispostos a calcular a hessiana a cada iteração do método; pensando nisso uma alternativa mais econômica que pode ser utilizada, são os Métodos Quase-Newton de Região de Confiança, onde a matriz A_k é uma aproximação para a hessiana no ponto x^k. Em geral computar uma aproximação para hessiana usando métodos secantes é mais barato do que computar a hessiana verdadeira, uma vez que as atualizações quase-Newton consistem em correções de posto baixo (1 ou 2). Em [2] encontra-se um estudo mais detalhado sobre os métodos de região de confiança e o artigo [5] trata especificamente de métodos Quase-Newton de região de confiança.

Mais adiante vamos discutir um pouco o subproblema de minimizar o modelo sujeito à região de confiança e veremos o do fato de trabalharmos com um modelo quadrático irá nos auxiliar no cálculo da solução exata do subproblema.

1.3 Verificando a redução

A cada iteração do método podemos atualizar, ou não, o raio Δ_k da região de confiança. Isso irá depender se o passo dado gera, ou não, uma boa redução no valor da função objetivo. Note que temos que levar em consideração dois fatores para avaliar a redução:

1. se a região de confiança de fato é uma região em que podemos confiar no modelo, pois como vimos na secção anterior, quanto mais longe se está do ponto corrente, maior será o erro de aproximação do modelo, então se for o caso, precisamos diminuir o raio da região de confiança para obter uma aproximação melhor;

2. se o minimizador encontrado é de fato o minimizador do modelo, pois ao resolver o subproblema sujeito à região de confiança, pode acontecer que o minimizador irrestrito do modelo não esteja no interior da região de confiança e então o minimizador restrito do modelo estará na fronteira da região de confiança. Veja que pode ocorrer do minimizador restrito encontrado proporcionar uma boa redução na função objetivo. Nesse caso poderíamos ser mais ousados e aumentar o raio da região de confiança e tentar encontrar o minimizador local do modelo, pois provavelmente ele irá proporcionar uma redução ainda melhor que a redução do minimizador restrito. No entanto, não fazemos isso, de fato aceitamos o ponto na fronteira da região como o próximo iterado, e usamos o fato de que o raio poderia crescer para aumentá-lo para a próxima iteração.

Para identificarmos quais desses dois fatores estão interferindo na redução do valor da função objetivo vamos definir *ared* como sendo a redução real da função objetivo (*actual reduction*) e a *pred* como sendo a redução predita pelo modelo (*predicted reduction*) e a razão entre elas, ρ, da seguinte maneira:

$$ared = f(x^k) - f(x^k + d^k), \tag{1.1}$$

$$pred = m_k(0) - m_k(d^k) \quad \text{e} \tag{1.2}$$

$$\rho = \frac{ared}{pred} = \frac{f(x^k) - f(x^k + d^k)}{m_k(0) - m_k(d^k)}. \tag{1.3}$$

O ideal é que $\rho \approx 1$, pois isso implica que $ared \approx pred$, logo tanto o modelo quanto a função objetivo estão tendo a mesma redução.Uma situação peculiar que pode ocorrer é $\rho > 1$, o que implica que $pred < ared$, ou seja, o ponto encontrado gera uma redução na função objetivo que é maior do que a predita pelo modelo, mas em geral, se o método utilizado gera modelos com uma boa aproximação para a função objetivo essa diferença não é muito grande, então não nos preocuparemos com esse tipo de situação em particular.

Por outro lado, uma situação que é muito preocupante é quando $\rho < 0$. Note que, na equação (1.2), $m_k(0)$ é o valor do modelo no ponto corrente, que coincide com o valor da f no ponto corrente, pois

$$m_k(0) = f(x^k) + \underbrace{\nabla f(x^k)^T 0}_{0} + \underbrace{\frac{1}{2} 0^T A_k 0}_{0} = f(x^k),$$

e como d^k é um minimizador de m_k, então o valor de $m_k(0)$ vai ser sempre maior que o valor $m_k(d^k)$, logo a

redução predita pelo modelo sempre será positiva, ou seja, $pred > 0$. Isso significa que $f(x^k) - f(x^k + d^k) < 0$, ou seja, o passo dado aumentou o valor da função objetivo, então o passo deve ser rejeitado (ver figura 1.3).

Vamos estabelecer alguns critérios para que *ared* e *pred* sejam reduções aceitáveis, analisando o valor da razão ρ:

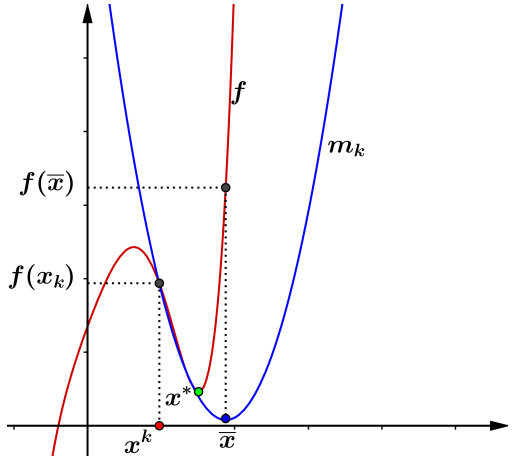
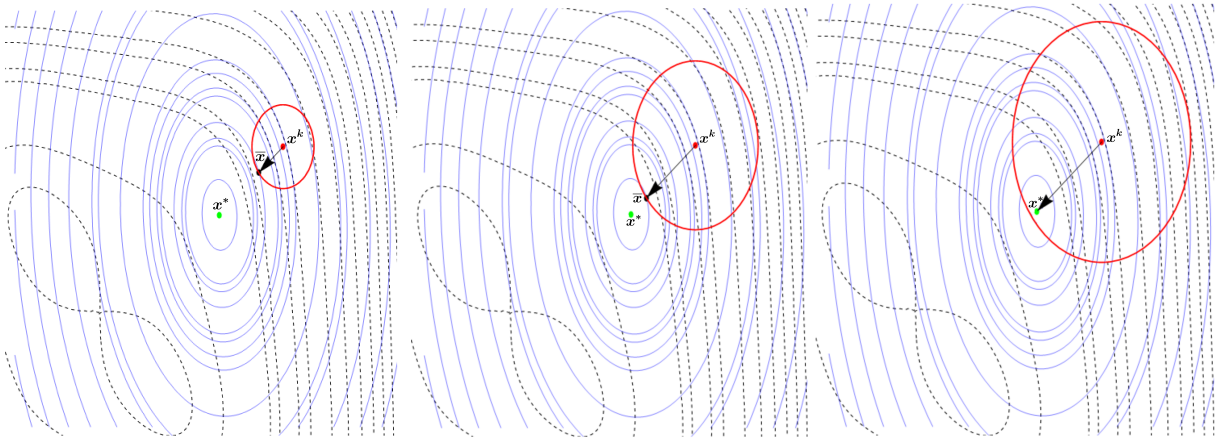


Figura 1.3: *Exemplo gráfico de uma iteração onde a redução predita pelo minimizador $\overline{x}$ é muito boa, porém a redução real é negativa, ou seja, $f(\overline{x}) > f(x^k)$*

- Se $\rho < \frac{1}{4}$, isso implica que a redução predita pelo modelo é maior de a redução real, ou seja, o modelo não representa uma boa aproximação nessa região, então reduzimos o raio da região de confiança e repetimos o processo na esperança de obter uma redução melhor.
- Se $\rho > \frac{3}{4}$ e $\|d^k\| < \Delta_k$ isso implica que redução predita pelo modelo é muito próxima da redução real e o minimizador do modelo está no interior da região de confiança, então aceitamos o passo e mantemos o raio inalterado para a próxima iteração.
- Se $\rho > \frac{3}{4}$ e $\|d^k\| = \Delta_k$, isso implica que redução predita pelo modelo é muito próxima da redução real, mas como $\|d^k\| = \Delta_k$ significa que o passo foi barrado pela fronteira da região de confiança, então aceitamos o minimizador encontrado e aumentamos o raio da região de confiança para a próxima iteração. Na figura 1.4 ilustramos melhor essa situação.



(a) Δ_k (b) Δ'_k (c) Δ''_k

Figura 1.4: *Em **(a)** e **(b)** podemos verificar que o minimizador local do modelo x^* não está no interior da região de confiança, pois em ambos os casos o raio Δ_k era pequeno. Em **(c)** temos o minimizador do modelo no interior da região de confiança.*

Podemos visualizar, na figura 1.5, os elementos presentes em uma iteração do método:

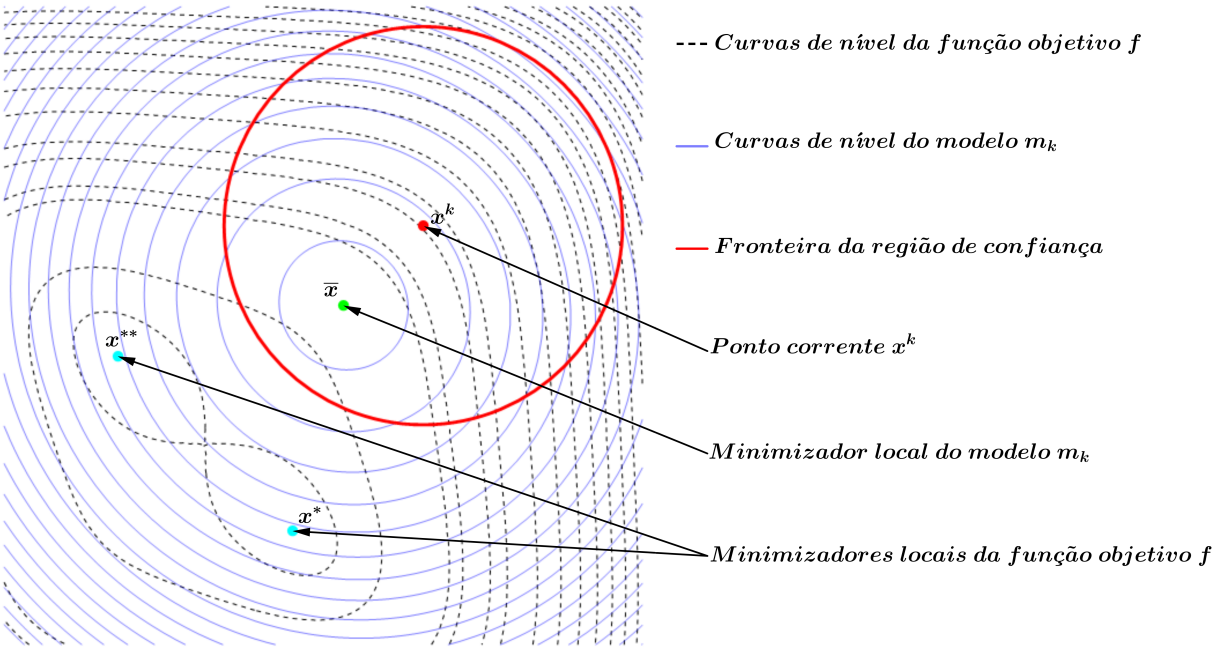


Figura 1.5: *Ilustração gráfico para os elementos de uma iteração do método de região de confiança.*

1.4 Algoritmo

Antes de entendermos as maneiras de se resolver o subproblema, vamos apresentar o algoritmo básico para métodos de região de confiança. Mais adiante, quando formos estudar a convergência global dos métodos de região de confiança veremos que em geral se estabelece uma cota inferior para a redução, denotada por $\eta > 0$, e em geral, um valor dentro do intervalo $[0, 0.25)$. Se $\rho > \eta$ vamos aceitar o passo atual e depois analisamos se vamos manter o raio para a próxima iteração ou se será necessário aumentá-lo de acordo com os critérios acima. Caso $\rho \leq \eta$ então o passo não será aceito e o raio será reduzido.

Abaixo temos o algoritmo proposto em [8]

Algoritmo 1: Região de Confiança

Entrada: $x^0 \in \mathbb{R}^n$, $\Delta_0 > 0$ e $\eta \in [0, \frac{1}{4})$
$k = 0$
repita
  Construir o modelo quadrático $m_k(d)$, em torno de x^k
  Obter d^k, solução de $m_k(d)$
  Calcular ρ_k

  se $\rho_k > \eta$ **então**
    $x^{k+1} = x^k + d^k$
    senão
      $x^{k+1} = x^k$
    fim
  fim
  se $\rho_k < \frac{1}{4}$ **então**
    $\Delta_k = \frac{\Delta_k}{2}$
    senão
      se $\rho_k > \frac{3}{4}$ *e* $\|d^k\| = \Delta_k$ **então**
        $\Delta_{k+1} = 2\Delta_k$
        senão
          $\Delta_{k+1} = \Delta_k$
        fim
      fim
    fim
  fim
  k=k+1
até $\nabla f(x^k) = 0_{\mathbb{R}^n}$;

Na próxima secção, vamos dedicar a nossa atenção à escolha da direção d^k, ou seja, vamos estudar as formas que temos para resolver o subproblema de região de confiança, de maneira aproximada ou exata.

1.5 Encontrando a direção

Após construir o modelo, o segundo passo é encontrar uma direção e o tamanho de passo a ser andado e para isso vamos calcular uma solução aproximada d^k para o modelo quadrático restrito definido a seguir:

$$\min_{d\in\mathbb{R}^n} m_k(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T A_k d \tag{1.4}$$
$$\text{s.a: } \|d\| \leq \Delta_k.$$

Seguindo o desenvolvimento de [9], iremos impor duas hipóteses sobre a direção d^k:

HD1: $\|d^k\| \leq \gamma\Delta_k$, onde $\gamma \geq 1$. Essa hipótese vai permitir que encontremos passos que extrapolem ligeiramente a região de confiança;

HD2: d^k satisfaz a seguinte desigualdade:

$$pred = m_k(0) - m_k(d^k) \geq c_1\|\nabla f(x^k)\|^2 \text{Min}\left\{\Delta_k, \frac{\|\nabla f(x^k)\|}{\|A_k\|}\right\}$$

onde $c_1 \in (0\ 1)$. Essa hipótese é que vai nos permitir provar a convergência global do método, e nos diz que o decréscimo obtido no modelo é, pelo menos, proporcional à norma do gradiente no ponto corrente.

1.5.1 O Passo de Cauchy

O Passo de Cauchy, também conhecido como Método do Gradiente [3] (ou Médodo de Máximo Decrescimento/Máxima Descida), é um dos métodos clássicos de busca linear e consiste em tomar a direção $d^k = -\lambda\nabla f(x^k)$, com $\lambda > 0$ isso porque vetor gradiente aponta na direção de máximo crescimento da função, consequentemente o sentido oposto do gradiente aponta na direção de máximo descrescimento.

Note que para todo $\Delta_k \neq 0$ o passo de Cauchy satisfaz a hipótese **HD1**, então basta mostrarmos que ele também satisfaz **HD2**. Em [9] essa demonstração é apresentada no lema a seguir, que incluímos para completude do texto, com detalhamentos adicionais que nos ajudaram a acompanhar melhor o desenvolvimento da prova.

Lema 1.5.1. *O passo de Cauchy, $d^k = -\lambda\nabla f(x^k)$, satisfaz a hipótese **HD2** com $c_1 = \frac{1}{2}$.*

Demonstração. Para simplificarmos a notação vamos suprimir o índice k e denotaremos $g = \nabla f(x)$. Seja a direção $d = -\lambda\nabla f(x)$, reescrevendo o modelo 1.4, temos:

$$\min\ m(-\lambda g) = f(x^k) - \lambda\|g\|^2 + \frac{1}{2}\lambda^2 g^T A g.$$

Agora queremos encontrar o tamanho de passo ótimo, λ^*, para isso vamos minimizar a seguinte função:

$$\min_{\lambda\in\mathbb{R}^+} \phi(\lambda) = \frac{1}{2}\lambda^2 g^T A g - \lambda\|g\|^2 + f(x)$$
$$\text{s.a: } \|d\| \leq \Delta.$$

Como $\lambda > 0$ e queremos que $\|\lambda g\| \leq \Delta$, temos:

$$\Rightarrow \lambda\|g\| \leq \Delta$$
$$\Rightarrow 0 < \lambda \leq \frac{\Delta}{\|g\|}.$$

Vamos analisar dois casos separadamente:

1. $g^T Ag > 0$, nesse caso ϕ é convexa e seu minimizador satisfaz a condição de otimalidade de primeira ordem:

$$\phi'(\lambda) = \lambda g^T Ag - \|g\|^2 = 0$$
$$\Rightarrow \lambda g^T Ag = \|g\|^2$$
$$\lambda^* = \frac{\|g\|^2}{g^T Ag}.$$

Então, se $\lambda^* \leq \frac{\Delta}{\|g\|}$, ou seja, se o tamanho de passo está dentro da região de confiança, então $\lambda = \lambda^*$. Vamos então calcular a redução do modelo:

$$\begin{aligned} pred &= m(0) - m(d) \\ &= f(x) - \left(f(x) - \frac{\|g\|^2}{g^T Ag}\|g\|^2 + \frac{1}{2}\left(\frac{\|g\|^2}{g^T Ag}\right)^2 g^T Ag \right) \\ &= f(x) - \left(f(x) - \frac{\|g\|^4}{g^T Ag} + \frac{1}{2}\frac{\|g\|^4}{g^T Ag} \right) \\ &= \frac{\|g\|^4}{g^T Ag} - \frac{\|g\|^4}{2g^T Ag} \\ &= \frac{1}{2}\frac{\|g\|^4}{g^T Ag}. \end{aligned}$$

Pela Desigualdade de Cauchy-Schwarz, sabemos que:

$$\Rightarrow g^T Ag \leq \|g\| \cdot \|Ag\|,$$

multiplicando e dividindo por $\|g\|$:

$$\Rightarrow g^T Ag \leq \|g\| \cdot \underbrace{\frac{\|Ag\|}{\|g\|}}_{\leq \|A\|} \cdot \|g\|$$

$$\Rightarrow g^T Ag \leq \|A\| \cdot \|g\|^2.$$

Isso implica que:

$$\frac{1}{2}\frac{\|g\|^4}{g^T Ag} \geq \frac{1}{2}\frac{\|g\|^4}{\|A\|\|g\|^2}$$

$$\Rightarrow pred \geq \frac{1}{2}\frac{\|g\|^2}{\|A\|}. \tag{1.5}$$

Agora se $\lambda^* \geq \frac{\Delta}{\|g\|}$, ou seja, se o tamanho de passo extrapola a região de confiança, então o ponto de Cauchy está definido na fronteira da bola: $\lambda = \frac{\Delta}{\|g\|} < \lambda^* = \frac{\|g\|^2}{g^T Ag}$

$$\frac{\Delta}{\|g\|} < \frac{\|g\|^2}{g^T Ag}$$

$$\Rightarrow \frac{\Delta}{\|g\|} g^T Ag < \|g\|^2$$

$$\Rightarrow \frac{\Delta^2}{\|g\|^2} g^T A g < \Delta \|g\|. \tag{1.6}$$

Calculando o valor do modelo em $d = -\frac{\Delta}{\|g\|} g$:

$$\begin{aligned} m(d) &= f(x) - \frac{\Delta}{\|g\|}\|g\|^2 + \frac{1}{2}\frac{\Delta^2}{\|g\|^2} g^T A g \\ &= f(x) - \Delta\|g\| + \frac{1}{2}\frac{\Delta^2}{\|g\|^2} g^T A g. \end{aligned}$$

Então da relação (1.6) temos:

$$\underbrace{f(x) - \Delta\|g\| + \frac{1}{2}\frac{\Delta^2}{\|g\|^2} g^T A g}_{m(d)} < f(x) - \Delta\|g\| + \frac{1}{2}\Delta\|g\| = \underbrace{f(x)}_{m(0)} - \frac{1}{2}\Delta\|g\|$$

$$\Rightarrow m(d) < m(0) - \frac{1}{2}\Delta\|g\|$$

$$\Rightarrow \underbrace{m(0) - m(d)}_{pred} > \frac{1}{2}\Delta\|g\|$$

$$\Rightarrow pred > \frac{1}{2}\Delta\|g\|. \tag{1.7}$$

2. $g^T A g \leq 0$, nesse caso ϕ é uma função descrescente para $\lambda \geq 0$. Se $g^T A g < 0$, então ϕ é quadrática côncava e possui um maximizador, e se $g^T A g = 0$, ϕ é uma reta decrescente. Em ambos os casos, o ponto de Cauchy está na fronteira da região de confiança, ou seja, $\lambda = \dfrac{\Delta}{\|g\|}$. Logo,

$$\begin{aligned} m(0) - m(d) &= f(x) - \left(f(x) - \Delta\|g\| + \frac{1}{2} g^T A g \right) \\ &= \Delta\|g\| - \frac{1}{2} g^T A g. \end{aligned}$$

Como $g^T A g \leq 0$ segue:

$$\Delta\|g\| - \frac{g^T A g}{2} \geq \Delta\|g\| \geq \frac{1}{2}\Delta\|g\|$$

$$\Rightarrow pred \geq \frac{1}{2}\Delta\|g\|. \tag{1.8}$$

Portanto, de (1.5), (1.7) e (1.8) temos que o Passo de Cauchy satisfaz a hipótese **HD2** com $c_1 = \frac{1}{2}$. □

1.5.2 O Passo de Newton

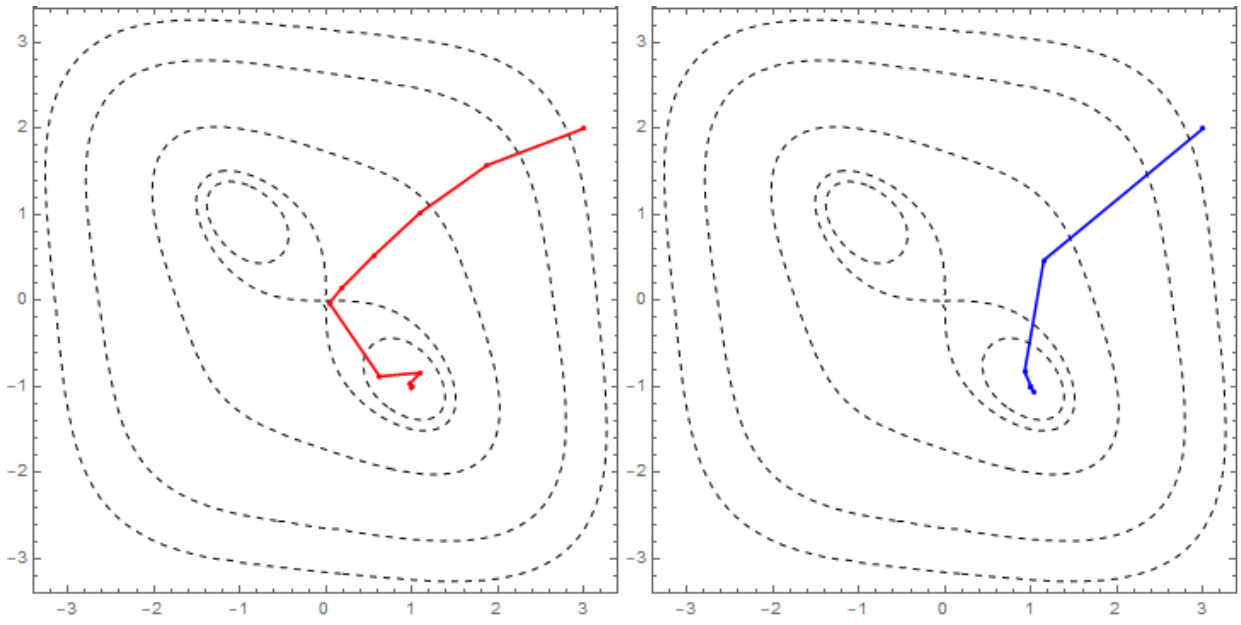
Outra direção também conhecida é o Passo de Newton. Este é, sem dúvidas, o recurso mais importante para otimização, pois quando utilizado nos métodos de busca linear conseguimos obter convergência quadrática (para mais detalhes consultar o capítulo 3 de [8]). A motivação deste método é encontrar o ponto x^k que satisfaz a condição necessária de otimalidade. Para isso se faz necessário resolver um sistema de n equações e n incógnitas, dado por $\nabla f(x^k) = 0^n_{\mathbb{R}}$, em geral esse sistema é não-linear então usamos métodos iterativos para poder resolvê-lo, ou seja, o Método de Newton para resolução de sistemas não-lineares, que consiste em:

$$x^{k+1} = x^k - [J_F(x^k)]^{-1} F(x^k). \tag{1.9}$$

Se considerarmos a função $F(x) = \nabla f(x)$, temos que a Jacobiana $J_F(x) = \nabla^2 f(x)$, então reescrevendo a equação (1.9), temos o Passo de Newton:

$$d = -\nabla^2 f(x)^{-1} \nabla f(x).$$

E aqui entra mais uma vantagem de se utilizar um modelo quadrático, pois ao minimizar uma função quadrática convexa utilizando o passo de Newton, encontramos o minimizador global em apenas 1 iteração, a demonstração pode ser consultada integralmente em [3, Preposição 6.1, p. 43]. Computacionalmente falando, temos um gasto maior ao trabalharmos com esse passo, uma vez que vamos ter um sistema linear para resolver a cada iteração do método, mas a convergência é mais rápida que do Passo de Cauchy, conforme ilustrado na figura 1.6.



(a) Passo de Cauchy (b) Passo de Newton

Figura 1.6: *A partir de um ponto inicial, x^0, geramos 2 sequências de pontos que convergem para o minimizador local da função. A primeira, utilizando o Passo de Cauchy, em* ***(a)****, que convergiu após um total de 20 iterações e a segunda, utilizando o Passo de Newton, em* ***(b)****, que convergiu após um total de 7 iterações.*

Porém, podemos nos deparar com o obstáculo de $\nabla^2 f(x^k)$ não ser inversível. Nesse caso, não é possível

obter o Passo de Newton. Se o gasto computacional não for um problema muito grande, uma forma eficiente de identificar se a hessiana é inversível é resolver o sistema linear utilizando a Fatoração de Cholesky, pois se for possível obter o fator de Cholesky da hessiana, então ela é definida positiva e portanto é não singular. Caso não seja possível obter o fator de Cholesky, então a hessiana não é definida positiva e portanto ela é singular. Conforme veremos na próxima secção, os métodos de região de confiança oferecem uma maneira automática de salvaguardar o Passo de Newton, caso ele não esteja bem definido. Assim como na construção do modelo, podemos trabalhar com aproximações para hessiana utilizando métodos Quase-Newton e obteremos passos Quase-Newton.

1.5.3 A Equação Secular

As direções que estudamos até esse momento se tratam de soluções aproximadas para o subproblema (1.4). Agora, vamos tentar aproveitar a estrutura do subproblema para encontrarmos uma solução exata, ou seja, encontrarmos um passo ótimo. Iremos primeiramente apresentar e discutir as características da solução exata e mais adiante nos preocuparemos em demonstrar formalmente a existência da solução exata. Nesta secção iremos detalhar a discussão apresentada na secção 4.3 de [8]. Para simplificarmos, vamos considerar $g = \nabla f(x)$ e omitir os subíndices k referente à iteração.

A solução global ótima d^* do problema (1.4) satisfaz as seguintes relações:

$$(A + \mu I)d^* = -g, \tag{1.10}$$

$$\mu(\Delta - \|d^*\|) = 0, \tag{1.11}$$

$$(A + \mu I) \quad \text{é semidefinida positiva}, \tag{1.12}$$

para algum $\mu \geq 0$.

Note que a condição (1.11) é referente à complementaridade da restrição do problema, ou seja, se a solução está dentro da região de confiança, então $\mu = 0$ e satisfaz as relações (1.10) e (1.12); agora, se a solução está na fronteira da região de confiança então $\Delta - \|d^*\| = 0$ e então com $\mu \neq 0$ podemos definir um vetor por meio da parametrização:

$$d(\mu) = -(A + \mu I)^{-1} g, \qquad \mu > 0, \tag{1.13}$$

pois para um μ suficientemente grande $A + \mu I$ é definida positiva, então sua inversa está bem definida. E procuramos um μ tal que

$$\|d(\mu)\| = \Delta.$$

Veja que agora nosso problema se resume em encontrar a raiz de uma função de uma variável, ou seja, conseguimos transformar o nosso problema de dimensão n em um problema unidimensional. Então podemos aplicar o Método de Newton para zero de funções para encontrar o valor ótimo μ^* que satisfaz:

$$h(\mu) := \|d(\mu)\|^2 - \Delta^2 = 0 \tag{1.14}$$

Para isso, vamos reescrever o vetor (1.13), utilizando a decomposição espectral de A. Como A é simétrica então podemos fatorar $A = Q\Lambda Q^T$, onde Q é uma matriz ortogonal e Λ é uma matriz diagonal, onde a diagonal principal $(\lambda_1, \lambda_2, ..., \lambda_n)$, com $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ são os autovalores de A. Logo, é fácil ver que

$$A + \mu I = Q(\Lambda + \mu I)Q^T$$

então podemos reescrever o vetor (1.13) como:

$$d(\mu) = -Q(\Lambda + \mu I)^{-1}Q^T g = -\sum_{j=1}^{n} \frac{q_j^T g}{\lambda_j + \mu} q_j, \quad \text{para todo } \mu \neq \lambda_j. \tag{1.15}$$

onde q_j é j-ésima coluna da matriz Q. Agora, como as colunas q_j são ortonormais, podemos escrever a norma de $d(\mu)$, como:

$$\|d(\mu)\|^2 = \sum_{j=1}^{n} \frac{(q_j^T g)^2}{(\lambda_j + \mu)^2} \underbrace{q_j^T q_j}_{\|q_j\|^2=1} \quad \Rightarrow \|d(\mu)\|^2 = \sum_{j=1}^{n} \frac{(q_j^T g)^2}{(\lambda_j + \mu)^2}. \tag{1.16}$$

Substituindo em (1.14):

$$h(\mu) := \sum_{j=1}^{n} \frac{(q_j^T g)^2}{(\lambda_j + \mu)^2} - \Delta^2. \tag{1.17}$$

Agora, analisando a expressão para a norma de $d(\mu)$ obtida em (1.16), podemos extrair algumas informações. Logo em seguida veremos porque essas informações são importantes:

1. Se $\mu > -\lambda_1$, então $\mu + \lambda_j > 0$, para todo $j = 1, 2, ..., n$; então $\|d(\mu)\|$ é uma função contínua e não crescente no intervalo $(-\lambda_1, \infty)$. Além disso,

$$\lim_{\mu \to \infty} \|d(\mu)\| = 0.$$

2. Se $q_j^T g \neq 0$, temos

$$\lim_{\mu \to \lambda_j} \|d(\mu)\| = \infty.$$

E por que essas duas informações são importantes? Veja que queremos encontrar um valor μ suficientemente grande e não negativo para satisfazer a condição (1.12), então a informação **(1)** nos diz que o valor ótimo, μ^*, que estamos procurando se encontra dentro do intervalo $(-\lambda_1, \infty)$. E como queremos que μ também satisfaça $\|d(\mu)\| = \Delta$, podemos observar que se $q_j^T g \neq 0$ sempre vamos conseguir obter um único valor $\mu^* \in (-\lambda_1, \infty)$ tal que $\|d(\mu)\| = \Delta$. Na figura 1.7, podemos verificar essas constatações.

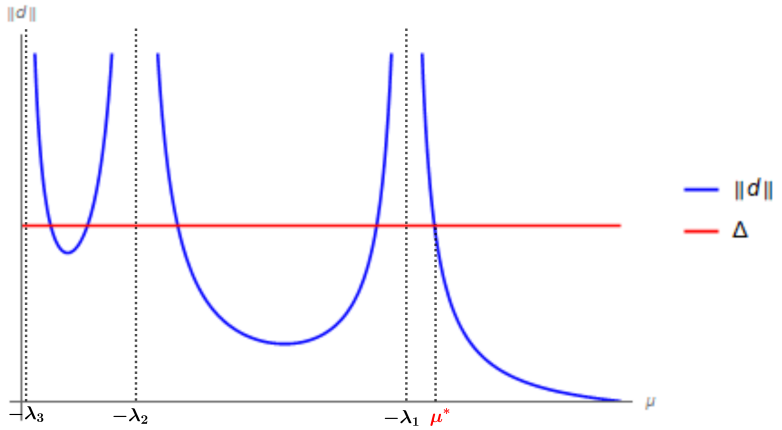


Figura 1.7: *Gráfico de* $\|d(\mu)\|$ *em função de* μ.

A informação **(2)** é importante porque nos mostra que, para valores de μ próximo de $-\lambda_j$ a função cresce indefinidamente, ou seja, essas regiões apresentam grande instabilidade numérica. Então, foi pensando nisso e também no fato de que a função (1.14) é altamente não linear que pensamos numa reformulação para a função h, a equação (1.18) abaixo é chamada de *Equação Secular* e como podemos ver na figura 1.8, a função ϕ apresenta um comportamento quase linear na região do μ^*, o que tornará o método de Newton mais confiável e eficaz.

$$\phi(\mu) = \frac{1}{\Delta} - \frac{1}{\|d(\mu)\|} = 0 \tag{1.18}$$

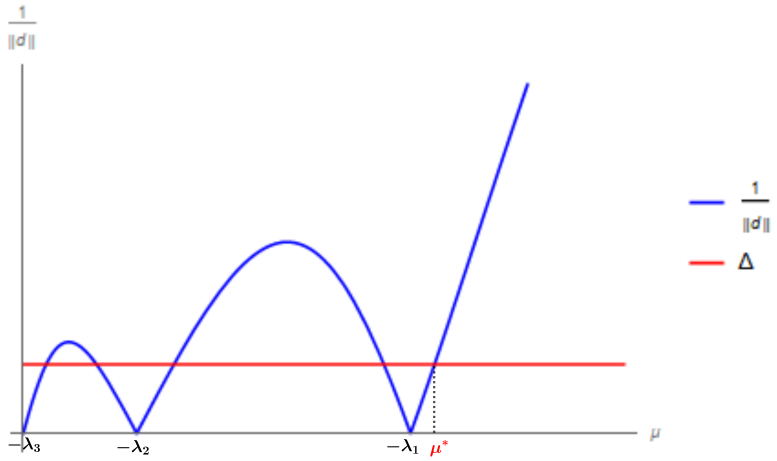


Figura 1.8: *Gráfico de* $\dfrac{1}{\|d(\mu)\|}$ *em função de* μ.

O caso difícil (conhecido na literatura como *hard case*) acontece quando $q_j^T g = 0$, pois pode não haver $\mu \in (-\lambda_1, \infty)$ tal que $\|d(\mu)\| = \Delta$. Na figura 1.9 ilustramos essa situação.

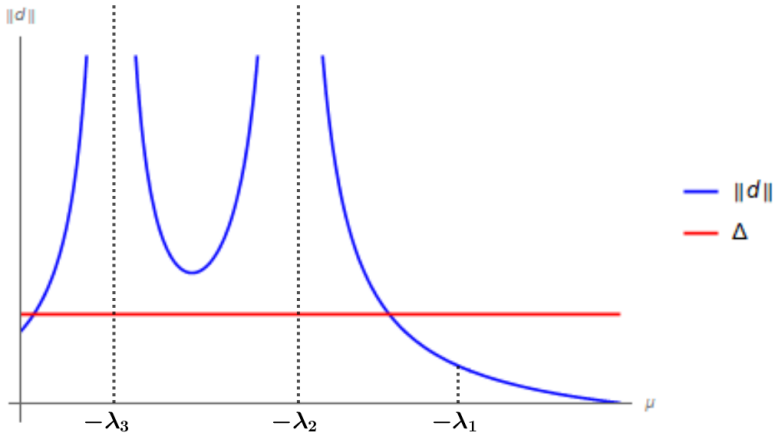


Figura 1.9: *Gráfico do pior caso, quando* $\|d(\mu)\| < \Delta$ *para todo* $\mu \in (-\lambda_1, \infty)$.

Mas o teorema que veremos na secção seguinte nos garante que existe $\mu \geq 0$ que satisfaz as condições (1.10), (1.11 e (1.12), então se considerarmos o intervalo $[-\lambda_1, \infty)$ nossa única possibilidade é $\mu = -\lambda_1$. Porém, vamos reformular a equação (1.16) para determinar o vetor $d(\mu)$ e sua norma. Notemos que a matriz $(A - \lambda_1 I)$ é uma matriz singular, e portanto existe um vetor $z \neq 0$ com $\|z\| = 1$ tal que $(A - \lambda_1 I)z = 0$, note

que z é o autovetor de A associado ao autovalor λ_1. Então podemos reescrever a equação (1.15) da seguinte maneira:

$$d(\mu) = -\sum_{j:\mu\neq\lambda_1} \frac{q_j^T g}{\lambda_j + \mu} q_j + \tau z$$

$$\Rightarrow \|d(\mu)\|^2 = \sum_{j:\mu\neq\lambda_1} \left(\frac{(q_j^T g)^2}{(\lambda_j+\mu)^2} \underbrace{q_j^T q_j}_{\|q_j\|^2=1} + 2\tau \frac{q_j^T g}{\lambda_j + \mu} \underbrace{q_j^T z}_{0} \right) + \tau^2 \underbrace{z^T z}_{\|z\|^2=1}$$

$$\Rightarrow \|d(\mu)\|^2 = \sum_{j:\mu\neq\lambda_1} \frac{(q_j^T g)^2}{(\lambda_j+\mu)^2} + \tau^2,$$

e dessa forma sempre podemos escolher um τ tal que $\|d(\mu)\| = \Delta$. O algoritmo para encontrar o multiplicador μ apresentado baixo foi baseado no algoritmo 4.3 de [8]:

Algoritmo 2: Solução Exata do subproblema de região de confiança

Entrada: $\mu^0 = 0$ e $\Delta > 0$
$l = 0$
repita
  Encontrar o fator de Cholesky tal que $A + \mu^l I = R^T R$;
  Resolva os 2 sistemas lineares: $R^T R p_l = -g$ e $R^T q_l = p_l$;
  Atualize μ:

$$\mu^{l+1} = \mu^l + \left(\frac{\|p_l\|}{\|q_l\|}\right)^2 \left(\frac{\|p_l\| - \Delta}{\Delta}\right);$$

  $l = l + 1$
até $\mu > -\lambda_1$;

1.5.4 O Passo Ótimo

Antes de enunciarmos e provarmos o teorema que irá caracterizar a solução exata do subproblema, vamos enunciar aqui com mais detalhes o lema a seguir, que é apresentado em [8, Lema 4.7, p. 89].

Lema 1.5.2. *Seja $m : \mathbb{R}^n \mapsto \mathbb{R}$ uma função quadrática definida como*

$$m(d) = \frac{1}{2} d^T A d + g^T d, \tag{1.19}$$

onde A é uma matriz simétrica. Então

i *m atinge um valor mínimo se, e somente se, A é semidefinida positiva e g pertence ao espaço imagem de A. Se A é semidefinida positiva então todo $d \in \mathbb{R}^n$ que satisfaça $Ad = -g$, é um minimizador global de m.*

ii *m possui um único minimizador se, e somente se, A é definida positiva.*

Demonstração. Vamos provar separadamente cada um dos dois itens:

(i)

($\Rightarrow$) Suponha que $g \in \mathcal{I}m(A)$, então existe d tal que $Ad = -g$ e assumindo que A seja semidefinida positiva, queremos mostrar que d é minimizador global de m. Temos que, para todo $v \in \mathbb{R}^n$, vale:

$$\begin{aligned} m(d+v) &= \frac{1}{2}(d+v)^T A(d+v) + g^T(d+v) \\ &= \frac{1}{2}(d^T Ad + \underbrace{d^T Av}_{v^T Ad} + v^T Ad + v^T Av) + g^T d + g^T v \\ &= \frac{1}{2}(d^T Ad + 2v^T \underbrace{Ad}_{-g} + v^T Av) + g^T d + g^T v \\ &= \underbrace{\frac{1}{2}d^T Ad + g^T d}_{m(d)} + \frac{1}{2}v^T Av + g^T v - \underbrace{v^T g}_{g^T v}, \\ &= m(d) + \frac{1}{2}v^T Av \end{aligned}$$

e como A é semidefinida positiva, segue

$$m(d+v) = m(d) + \frac{1}{2}v^T Av \geq m(d) \tag{1.20}$$

e portanto d é minimizador global de m.

($\Leftarrow$) Seja d minimizador global de m, queremos mostrar que $Ad = -g$, ou seja, $g \in \mathcal{I}m(A)$ e que A é semidefinida positiva. Sabemos que o gradiente e a hessiana de m são respectivamente:

$$\nabla m(w) = Aw + g \qquad \nabla^2 m = A.$$

Como d é minimizador global de m, então ele satisfaz as condições necessárias de otimalidade para o problema irrestrito de minimizar m(d). Logo:

$$Ad + g = 0 \Rightarrow Ad = -g \qquad e \qquad A \text{ é semidefinida positiva.}$$

(ii)

($\Rightarrow$) Se A é definida positiva, queremos mostrar que o seu minimizador é único. A demonstração desse item é análoga à do item **(i)**, com a única diferença de que agora como A é definida positiva então $v^T Av > 0$ para todo $v \neq 0$, então $m(d) + \frac{1}{2}v^T Av + g^T v > m(d)$. Logo, d é minimizador global estrito de m.

($\Leftarrow$) Agora seja d o minimizador global estrito de m, queremos mostrar que A é definida positiva. Assim como no item **(i)**, as condições necessárias de otimalidade já nos garantem que A é semidefinida positiva, então vamos supor que ela seja singular. Isso implica que existe $v \neq 0$ tal que $Av = 0$. Então, da relação (1.20) temos que $m(d+v) = m(d)$, portanto o minimizador global não é estrito, o que é uma contradição! Logo A é não singular e portanto A é definida positiva.

□

Teorema 1.5.1. *O vetor d^* é solução global do subproblema de região de confiança*

$$\min_{d\in\mathbb{R}^n} m(d) = f(x) + g^T d + \frac{1}{2} d^T A d \tag{1.21}$$

$$s.a:\ \|d\| \leq \Delta,$$

se e somente se d^* é factível e existe um escalar $\mu \geq 0$ tal que as seguintes condições são satisfeitas:

$$(A + \mu I)d^* = -g,$$

$$\mu(\Delta - \|d^*\|) = 0,$$

$$(A + \mu I) \quad \textit{é semidefinida positiva.}$$

Demonstração. ($\Rightarrow$) Primeiramente, assumimos que existe $\mu \geq 0$ tal que as condições (1.10, 1.11 e 1.12) são satisfeitas. Queremos mostrar que d^* é minimizador global do modelo quadrático (1.4). Pelo Lema 1.5.2, temos que d^* é um minimizador global da função quadrática abaixo

$$\begin{aligned}
\hat{m}(d) &= \frac{1}{2} d^T (A + \mu I) d + g^T d \\
&= \frac{1}{2}\left(d^T A d + \underbrace{d^T \mu I d}_{\mu d^T d} \right) + g^T d \\
&= \underbrace{\frac{1}{1} d^T A d + g^T d}_{m(d)} + \frac{\mu}{2} d^T d \\
&= m(d) + \frac{\mu}{2} d^T d \geq m(d*)
\end{aligned}$$

Como $\hat{m}(d) \geq \hat{m}(d^*)$, temos:

$$m(d) + \frac{\mu}{2} d^T d \geq m(d*) + \frac{\mu}{2} (d^*)^T d^*$$

$$\Rightarrow m(d) \geq m(d*) + \frac{\mu}{2} (d^*)^T d^* - \frac{\mu}{2} d^T d$$

$$\Rightarrow m(d) \geq m(d*) + \frac{\mu}{2}\left(\underbrace{(d^*)^T d^*}_{\|d^*\|^2} - \underbrace{d^T d}_{\|d\|^2} \right) \tag{1.22}$$

por hipótese $\mu > 0$ e satistaz a condição de complementaridade, ou seja, $\mu(\Delta - \|d^*\|) = 0$, isso implica que $\Delta - \|d^*\| = 0$ então $\Delta = \|d^*\|$ que também podemos escrever como $\Delta^2 = \|d^*\|^2$, logo

$$\Rightarrow m(d) \geq m(d^*) + \frac{\mu}{2}\left(\Delta^2 - \|d\|^2 \right)$$

para todo $d \leq \Delta$ temos

$$m(d) \geq m(d*) + \frac{\mu}{2}\left(\Delta^2 - \|d\|^2 \right) \geq m(d^*)$$

$$\therefore\ d^* \text{ é minimizador global de (1.4).}$$

($\Leftarrow$) Agora vamos assumir que d^* é minimizador global de (1.4) e vamos mostrar que existe $\mu \geq 0$ que satisfaz as condições (1.10), (1.11 e (1.12). Vamos pensar em 2 situações:

i quando $\|d^*\| < \Delta$ e

ii quando $\|d^*\| = \Delta$.

O caso **i** é o mais simples possível, pois d^* é o minimizador irrestrito do modelo (1.4), logo pelas condições necessárias de otimalidade temos satisfeita as condições (1.10) e (1.12):

$$\nabla m(d^*) = Ad^* + g = 0 \Rightarrow Ad^* = -g \qquad \nabla^2 m(d^*) = A \text{ é semidefinida positiva}$$

e a condição (1.11) é satisfeita com $\mu = 0$.

Agora vamos considerar o caso **ii**. Observe que, nesse caso, a condição (1.11) é automaticamente satisfeita. Notemos também que d^* resolve o problema com restrição de igualdade abaixo:

$$\min m(d) \quad \text{sujeito a } \|d\| = \Delta$$

logo, d^* satisfaz as condições necessárias de otimalidade para problemas com restrições de igualdade, ou seja, existe um μ tal que a função Lagrangiana definida abaixo tem um ponto estacionário em d^*:

$$\mathcal{L}(d, \mu) = m(d) + \frac{\mu}{2}(d^T d - \Delta^2);$$

e ao igualar $\nabla_d \mathcal{L}(d, \mu)$ a zero, obtemos:

$$Ad^* + g + \mu d^* = 0 \quad \Rightarrow (A + \mu I)d^* = -g,$$

e portanto a condição (1.10) também é satisfeita.

Note que podemos escrever $g = -(A + \mu I)d^*$. Como d^* é minimizador de m, temos que $m(d) \geq m(d^*)$ para todo d tal que $d^T d = \Delta^2$, ou seja, para todo d que também esteja na fronteira da região de confiança. Da relação (1.22) temos:

$$m(d) \geq m(d*) + \frac{\mu}{2}\left((d^*)^T d^* - d^T d\right)$$

$$\Rightarrow \frac{1}{2}d^T Ad + g^T d \geq \frac{1}{2}(d^*)^T Ad^* + g^T d^* + \frac{\mu}{2}(d^*)^T d^* - \frac{\mu}{2}d^T d$$

substituindo g na expressão acima, obtemos:

$$\frac{1}{2}d^T Ad - (d^*)^T(A + \mu I)d \geq -\frac{1}{2}(d^*)^T Ad^* - (d^*)^T(A + \mu I)d^* + \frac{\mu}{2}(d^*)^T d^* - \frac{\mu}{2}d^T d,$$

efetuando algumas distribuições, obtemos:

$$\begin{aligned}
&\frac{1}{2}d^T Ad - (d^*)^T(A + \mu I)d \geq -\frac{1}{2}(d^*)^T Ad^* - \frac{\mu}{2}(d^*)^T d^* - \frac{\mu}{2}d^T d \\
\Rightarrow &\frac{1}{2}d^T Ad + \frac{\mu}{2}d^T d - (d^*)^T(A + \mu I)d \geq -\frac{1}{2}(d^*)^T Ad^* - \frac{\mu}{2}(d^*)^T d^* \\
\Rightarrow &\frac{1}{2}d^T(A + \mu I)d - (d^*)^T(A + \mu I)d \geq -\frac{1}{2}(d^*)^T(A + \mu I)d^* \\
\Rightarrow &\frac{1}{2}d^T(A + \mu I)d - (d^*)^T(A + \mu I)d + \frac{1}{2}(d^*)^T(A + \mu I)d^* \geq 0.
\end{aligned}$$

Agora, vamos particionar o termo $d^T(A+\mu I)d^* = \frac{1}{2}(d^*)^T(A+\mu I)d + \frac{1}{2}d^T(A+\mu I)d^*$, substituindo na expressão anterior:

$$\Rightarrow \frac{1}{2}d^T(A+\mu I)d - \frac{1}{2}(d^*)^T(A+\mu I)d - \frac{1}{2}d^T(A+\mu I)d^* + \frac{1}{2}(d^*)^T(A+\mu I)d^* \geq 0$$

$$\Rightarrow \frac{1}{2}\left(\underbrace{d^T(A+\mu I)d - (d^*)^T(A+\mu I)d}_{(d-d^*)^T(A+\mu I)d} - \underbrace{d^T(A+\mu I)d^* + (d^*)^T(A+\mu I)d^*}_{(d-d^*)^T(A+\mu I)d^*}\right) \geq 0$$

$$\Rightarrow \frac{1}{2}(d-d^*)^T(A+\mu I)(d-d^*) \geq 0;$$

e com isso mostramos que a condição (1.12) também é satisfeita.

Por fim, vamos mostrar que de fato $\mu \geq 0$, para isso vamos supor que $\mu < 0$. Como já mostramos que d^* minimizador de m sujeito a $\|d\| \leq \Delta$, então d^* é minimizador irrestrito de m, então pelo Lema 1.5.2, temos:

$$Ad = -g \qquad e \qquad A \text{ é semidefinida positiva.},$$

satisfazendo as condições (1.10), (1.12 e (1.11) com $\mu = 0$, contradizendo a nossa hipótese de que $\mu \geq 0$. □

1.6 Convergência

Agora para garantirmos a convergência global do métodos de região de confiança vamos estabelecer algumas hipóteses sobre a função objetivo, sobre a solução aproximada (ou exata) do subproblema (1.4) e sobre as hessianas da função objetivo:

HF1: A função objetivo f é de classe C^1, com ∇f Lipschitz;

HF2: A função objetivo f é limitada inferiormente no conjunto de nível

$$N = \{x \in \mathbb{R}^n | f(x) \leq f(x^0)\};$$

HD1: $\|d^k\| \leq \gamma\Delta_k$, onde $\gamma \geq 1$;

HD2: d^k satisfaz a seguinte desigualdade:

$$pred = m_k(0) - m_k(d^k) \geq c_1\|\nabla f(x^k)\|\text{Min}\left\{\Delta_k, \frac{\|\nabla f(x^k)\|}{\|A_k\|}\right\},$$

onde $c_1 \in (0\ 1)$.

HM1: As hessianas A_k tem norma uniformemente limitadas, ou seja, exite uma constante $\beta > 0$ tal que $\|A_k\| \leq \beta$ para todo $k \in \mathbb{N}$.

As hipóteses **HF1**, **HF2** e **HM1**, vão nos garantir a diferenciabilidade da função objetivo, a existência de um minimizador local e a convergência para este minimizador; todas elas são bastante usuais na análise de convergência dos métodos de otimização em geral. Enquanto a hipótese **HD1** vai impor que o passo obtido gere uma redução no modelo que seja uma fração da redução gerada pelo Passo de Cauchy e a hipótese **HD2** vai permitir que o passo exceda a região de confiança, contanto ele que seja um múltiplo fixo do raio Δ_k.

Vamos provar o lema a seguir, que é apresentado em [9, Lema 5.37, p. 146], com mais detalhes. Ele irá nos auxiliar a provar a convergência global dos métodos de região de confiança:

Lema 1.6.1. *Supondo que as hipóteses **HF1**, **HM1**, **HD1** e **HD2** são satisfeitas, então existe uma constante $c > 0$ tal que*

$$|\rho - 1| \leq \frac{c\Delta^2}{\|g\| Min\left\{\Delta, \frac{\|g\|}{\beta}\right\}}. \tag{1.23}$$

Demonstração. Seja d uma solução aproximada para o modelo m, vamos calcular as reduções $ared$ e $pred$. Primeiramente, pelo teorema do valor médio, sabemos que existe $\theta \in (0,\ 1)$ tal que

$$f(x+d) = f(x) + \nabla f(x+\theta d)^T d,$$

logo temos:

$$ared = -\nabla f(x+\theta d)^T d \qquad e \qquad pred = -\frac{1}{2} d^T A d - \nabla f(x)^T d.$$

Então, efetuando $ared - pred$, obtemos:

$$\begin{aligned} ared - pred &= \frac{1}{2} d^T A d - \nabla f(x)^T d - \nabla f(x+\theta d)^T d \\ &= \frac{1}{2} d^T A d - (\nabla f(x+\theta) - \nabla f(x))^T\, d, \end{aligned}$$

tirando o valor absoluto, temos:

$$|ared - pred| = \left| \frac{1}{2} d^T A d - (\nabla f(x+\theta) - \nabla f(x))^T\, d \right|.$$

Pela desigualdade triangular, temos:

$$\left| \frac{1}{2} d^T A d - (\nabla f(x+\theta) - \nabla f(x))^T\, d \right| \leq \left| \frac{1}{2} d^T A d \right| + \left| (\nabla f(x+\theta) - \nabla f(x))^T\, d \right|;$$

e pela desigualdade de Cauchy-Schwarz e pelas hipóteses **HF1** e **HM1** temos:

$$\left| \frac{1}{2} d^T A d \right| + \left| (\nabla f(x+\theta) - \nabla f(x))^T\, d \right| \leq \frac{1}{2} \underbrace{\|A\|}_{\leq \beta} \cdot \|d\|^2 + \underbrace{\|\nabla f(x+\theta) - \nabla f(x)\|}_{\leq L\|d\|} \cdot \|d\|$$

reescrevendo, temos:

$$\left| \frac{1}{2} d^T A d \right| + \left| (\nabla f(x+\theta) - \nabla f(x))^T\, d \right| \leq \frac{1}{2}\beta \|d\|^2 + L\|d\|^2,$$

e pela hipótese **HD1**, temos que $\|d\| \leq \gamma\Delta \Rightarrow \|d\|^2 \leq \gamma^2\Delta^2$, então

$$\left(\frac{\beta}{2} + L\right)\|d\|^2 \leq \underbrace{\left(\frac{\beta}{2} + L\right)\gamma^2}_{c_0} \Delta^2,$$

logo,

$$|ared - pred| \leq c_0 \Delta^2. \tag{1.24}$$

Note que $c_0 > 0$, pois $\Delta^2 > 0$. Vamos então finalmente calcular $|\rho - 1|$:

$$|\rho - 1| = \left| \frac{ared}{pred} - 1 \right| = \left| \frac{ared}{pred} - \frac{pred}{pred} \right| = \left| \frac{ared - pred}{pred} \right| = \frac{1}{pred} \cdot |ared - pred|$$

da relação (1.24), temos:

$$\frac{1}{pred} \cdot |ared - pred| \leq \frac{1}{pred} \cdot |ared - pred| \cdot c_0\Delta^2 = \frac{c_0}{pred}\Delta^2,$$

pela hipótese **HD2**, temos:

$$|\rho - 1| \leq \frac{c_0\Delta^2}{c_1\|\nabla f(x)\|\text{Min}\left\{\Delta, \frac{\|\nabla f(x)\|}{\|A_k\|}\right\}},$$

então temos que $c = \dfrac{c_0}{c_1}$. □

O Lema 1.6.1 nos permite afirmar que o Algoritmo 1 está bem definido, ou seja, após uma sequência finita de insucessos, teremos

$$\Delta_k \leq \text{Min}\left\{\frac{\|\nabla f(x^k)\|}{\beta}, \frac{\|\nabla f(x^k)\|}{2c}\right\}. \tag{1.25}$$

E então temos:

$$|\rho - 1| = \frac{c\Delta}{\|\nabla f(x)\|} \leq \frac{1}{2}$$

$$\Rightarrow |\rho - 1| \leq \frac{1}{2} \Rightarrow \rho - 1 \geq -\frac{1}{2}.$$

$$\therefore \rho \geq \frac{1}{2} > \frac{1}{4},$$

logo, pelo Algoritmo 1, o passo será aceito.

Agora vamos para o nosso primeiro resultado de convergência global,

Teorema 1.6.1. *Supondo que todas as hipóteses são satisfeitas, então*

$$\lim_{k\to\infty} \|g_k\| = 0. \tag{1.26}$$

Demonstração. Suponha por absurdo que a afirmação seja falta, então

$$\exists \quad \epsilon > 0 \ / \ \|g_k\| \geq \epsilon, \qquad \forall\ k \in \mathbb{N}.$$

Seja $\tilde{\Delta} = \text{Min}\left\{\dfrac{\epsilon}{\beta}, \dfrac{\epsilon}{2c}\right\}$, onde β é a constante de **HM1** e c é a constante do Lema 1.6.1. Se $\Delta_k \leq \tilde{\Delta}$, então

$$\Delta_k \leq \frac{\epsilon}{\beta} \leq \frac{\|\nabla f(x^k)\|}{\beta} \qquad e \qquad \Delta_k \leq \frac{\epsilon}{2c}.$$

Pelo Lema 1.6.1, temos:

$$|\rho - 1| \leq \frac{c\Delta_k}{\epsilon} \leq \frac{1}{2}.$$

Ou seja, $\rho \geq \frac{1}{2} > \frac{1}{4}$ e pelo Algoritmo 1, temos que $\Delta_{k+1} \geq \Delta_k$. Então, o raio só será reduzido quando $\Delta_k > \tilde{\Delta}$, e nesse caso $\Delta_{k+1} = \frac{\Delta_k}{2} > \frac{\tilde{\Delta}}{2}$.

Logo, temos:

$$\Delta_k \geq \text{Min}\left\{\Delta_0, \frac{\tilde{\Delta}}{2}\right\}, \qquad \forall\ k \in \mathbb{N}. \tag{1.27}$$

Seja o conjunto $\mathcal{K} = \left\{k \in \mathbb{N}/\rho_k \geq \dfrac{1}{4}\right\}$, dado $k \in \mathcal{K}$, então pelo algoritmo 1 e pela hipótese **HD2** temos:

$$\begin{aligned} f(x^k) - f(x^{k+1}) &= f(x^k) - f(x^k + d^k) \\ &\geq \frac{1}{4}\left(m_k(0) - m_k(d^k)\right) \\ &\geq \frac{1}{4} c_1 \epsilon \text{Min}\left\{\Delta_k, \frac{\epsilon}{\beta}\right\}. \end{aligned}$$

Considerando (1.27), sabemos que existe uma constante $\sigma > 0$ tal que

$$f(x^k) - f(x^{k+1}) \geq \sigma, \qquad \forall\ k \in \mathcal{K}. \tag{1.28}$$

Pela hipótese **HF2**, a sequência $(f(x^k))$ é limitada inferiormente, e podemos notar que ela é não crescente, então $f(x^k) - f(x^{k+1}) \to 0$, então podemos concluir que o conjunto $\mathcal{K}$ é finito e então para algum $k \in \mathbb{N}$ suficientemente grande, $\rho_k < \frac{1}{4}$, então o raio Δ_k será reduzido pela metade, a cada iteração. Logo, $\Delta_k \to 0$, o que contradiz 1.27. Portanto,

$$\lim_{k\to\infty} \|g_k\| = 0.$$

□

Agora, quando exigimos que $\eta > 0$ conseguimos obter um resultado ainda mais forte do que o teorema 1.6.1:

Teorema 1.6.2. *Supondo que as hipóteses **HF1**, **HF2**, **HD1**, **HD2** e **HM1** são satisfeitas e que a constante η do Algoritmo 1 seja estritamente positiva, então*

$$\|g_k\| \to 0. \tag{1.29}$$

Demonstração. Vamos supor por absurdo que exista um $\epsilon > 0$ para qual o conjunto

$$\mathcal{K} = \left\{k \in \mathbb{N} / \|\nabla f(x^k)\| \geq \epsilon\right\}$$

seja infinito. Pelo Teorema 1.6.1, dado um $k \in \mathcal{K}$, existe um primeiro índice $l_k > k$ tal que $\|\nabla f(x^{l_k})\| \leq \frac{\epsilon}{2}$. Pela hipótese **HF1** o gradiente da f é Lipschitz, então existe uma constante $L > 0$ tal que

$$\frac{\epsilon}{2} \leq |\ \nabla f(x^k) - \nabla f(x^{l_k})\| \leq L\|x^k - x^{l_k}\|.$$

Então, vamos definir o seguinte conjunto:

$$\mathcal{S}_k = \left\{j \in \mathbb{N} / k \leq j < l_k,\ x^{j+1} \neq x^j\right\}.$$

Pela hipótese **HD2**, obtemos:

$$\frac{\epsilon}{2L} \leq \|x^k - x^{l_k}\| \leq \sum_{j\in\mathcal{S}_k} \|x^j - x^{j+1}\| \leq \sum_{j\in\mathcal{S}_k} \gamma\Delta_j. \tag{1.30}$$

Pelo Algoritmo 1, mais a definição de l_k e as hipóteses **HM1** e **HD2**, temos:

$$\begin{aligned} f(x^k) - f(x^{l_k}) &= \sum_{j\in\mathcal{S}_k} \left(f(x^j) - f(x^{j+1})\right) \\ &> \sum_{j\in\mathcal{S}_k} \eta\left(m_j(0) - m_j(d^j)\right) \\ &> \sum_{j\in\mathcal{S}_k} \eta c_1 \frac{\epsilon}{2} \text{Min}\left\{\Delta_j, \frac{\epsilon}{2\beta}\right\}. \end{aligned}$$

Da relação (1.30), temos

$$f(x^k) - f(x^{l_k}) \geq \eta c_1 \frac{\epsilon}{2} \text{Min} \left\{ \frac{\epsilon}{2\gamma L}, \frac{\epsilon}{2\beta} \right\} > 0, \qquad \forall\ k \in \mathcal{K}. \tag{1.31}$$

Mas novamente, pela hipótese **HF2**, a sequência $(f(x^k))$ é limitada inferiormente, e sabemos que ela é não crescente, então $f(x^k) - f(x^{l_k}) \to 0$, o que contradiz (1.31). Portanto a afirmação do teorema é verdadeira.

□

1.7 Obtendo outras direções

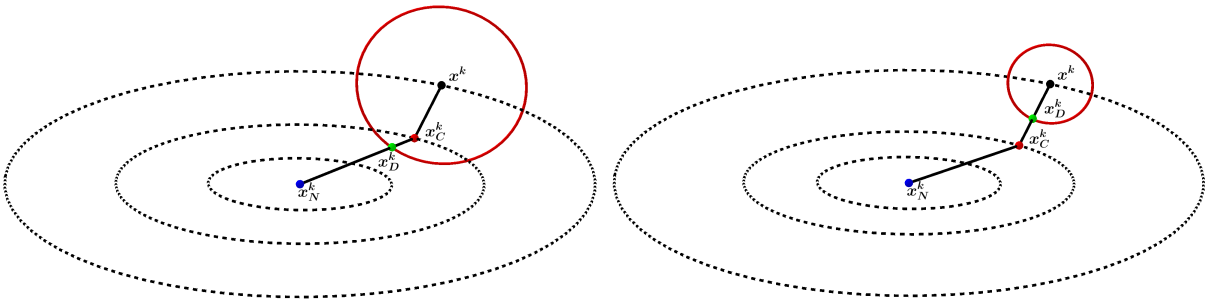
Como vimos anteriormente, o Passo de Cauchy é capaz de fornecer uma redução suficiente para o modelo, mas ainda assim o seu desempenho é muito fraco, pois como podemos verificar, sua redução não é tão eficiente quanto a redução proporcionada pelo Passo de Newton; isso acontece porque a direção do passo de Cauchy é computada sem utilizar nenhuma informação da matriz A_k, ela é usada apenas para determinar o tamanho do passo, enquanto no Passo de Newton, a hessiana A_k tem um papel fundamental no cálculo da direção. Ou seja, se as informações da matriz A_k conseguem contribuir para uma boa representação da curvatura da função e se utilizarmos A_k no cálculo da direção e do tamanho do passo, espera-se que o método convirja mais rapidamente. Nesta secção vamos estudar dois métodos que tentam encontrar um passo melhor que o Cauchy, um deles utiliza o Passo de Newton enquanto o outro é uma generalização do método de gradientes conjugados.

1.7.1 O Método Dogleg

A ideia do método *Dogleg* é encontrar um ponto que proporcione uma redução melhor que a do Passo de Cauchy, ou seja, ele funciona como uma forma de acelerar a convergência do método. Para isso, a cada iteração ele irá calcular o Passo de Cauchy e também o Passo de Newton. Note que para que o método esteja bem definido, a matriz A_k precisa ser definida positiva para todo k, ou seja, em todas as iterações o modelo quadrático precisa ser convexo, pois do contrário não é possível obter o Passo de Newton. O método recebe esse nome, pois a poligonal formada pelo ponto corrente, ponto de Cauchy e ponto de Newton tem uma forma similar a de uma "perna de cachorro".

Assim, em cada iteração, o passo d^k a ser dado será determinado de acordo com as seguintes condições:

- se o ponto de Cauchy x_C^k e o de Newton x_N^k estão no interior da região de confiança, vamos tomar o Passo de Newton como sendo o nosso passo d^k, pois ele já nos fornece uma redução satisfatória;
- se somente a solução de Cauchy está no interior da região de confiança, vamos encontrar o ponto x_D^k, intersecção da poligonal formada pelo ponto corrente, o ponto de Cauchy e o ponto de Newton com a fronteira da região de confiança, pois já que não podemos dar o Passo de Newton, vamos encontrar um ponto que seja melhor que o de Cauchy. A figura 1.10 **(a)** ilustra essa situação.
- agora se ambos pontos estão no exterior da região de confiança vamos adotar d^k como sendo o passo de Cauchy com norma igual ao raio da região de confiança. Essa segunda situação é ilustrada na figura 1.10 **(b)**.



(a) x_C^k no interior da região de confiança (b) x_C^k no exterior da região de confiança

Figura 1.10: *Ilustração para a poligonal da "perna do cachorro" em duas possibilidades distintas para o tamanho do raio da região de confiança (imagem baseada na Figura 5.10 de [9]).*

Abaixo apresentamos o algoritmo para o *Passo Dogleg*, baseado no algoritmo proposto em [9]:

Algoritmo 3: PASSO DOGLEG

Entrada: $x^k \in \mathbb{R}^n$ e $\Delta_k > 0$

Calcule $d_C^k = -\frac{g_k^T g_k}{g_k^T A_k g_k} g_k$

se $\|d_C^k\| > \Delta_k$ **então**
  $d^k = -\frac{\Delta_k}{\|g_k\|} g_k$
senão
  Calcule d_N^k, tal que $A_k d_N^k = -g_k$
  se $\|d_N^k\| \leq \Delta_k$ **então**
    $d^k = d_N^k$
  senão
    Determine $\alpha_k \in [0,\ 1]$ tal que $\|d_C^k + \alpha_k(d_N^k - d_C^k)\| = \Delta_k$
    $d^k = d_C^k + \alpha_k(d_N^k - d_C^k)$
  fim
fim
fim

Podemos verificar na figura 1.10 que, a partir do ponto x^k, o modelo quadrático vai decrescendo ao longo da poligonal e que se o ponto de Newton está no exterior da região de confiança, podemos ver que a poligonal intersecta a fronteira da região uma única vez, o que nos garante que o método está bem definido; a demonstração pode ser vista em [9, Lema 5.40, p. 154]. Há também uma variação desse método chamada de *Double-Dogleg*, este método tenta encontrar um ponto ainda mais próximo do ponto de Newton do que o de Cauchy, e para isso ele cria um quarto vértice para construir a poligonal de tal forma que o ponto x_D^k esteja ainda mais próximo do ponto x_N^k, em [2, Secção 7.5.3, p. 218] os autores detalham as duas variações do método.

1.7.2 O Método GC-Steihaug

Conforme visto anteriormente, o método *Dogleg* só pode ser aplicado quando a matriz A_k é definida positiva, ou seja, a cada iteração o modelo quadrático precisa ser convexo. Mas a priori, não temos como assegurar que, em todas as iterações, o modelo quadrático gerado será convexo, só descobrimos isso aplicando o método *Dogleg* e se ele for interrompido antes de convergir, é porque A_k não é definida positiva. Pensando em relaxar a hipótese da positividade da matriz A_k, em [10] Steihaug propõe um método baseado em Gradientes-Conjugados, que encontra uma solução aproximada para o subproblema (1.4), cuja redução do modelo é pelo menos tão boa quanto o Passo de Cauchy. Abaixo apresentamos o algoritmo *GC-Steihaug*, também baseado no algoritmo proposto em [9]:

Algoritmo 4: Gradientes Conjugados - Steihaug

Entrada: $d_s^0 = 0$, $r^0 = g$ e $p^0 = -r^0$
$j = 0$
repita
  se $(p^j)^T A p^j \leq 0$ **então**
    Calcule $t \in \mathbb{R}$ tal que $d = d_s^j + t p^j$ minimiza m e $\|d\| = \Delta$
    Faça $d^k = d$
    senão
      Calcule $t_j = \frac{(r^j)^T r^j}{(p^j)^T A p^j}$
      Defina $d_s^{j+1} = d_s^j + t_j p^j$
      se $\|d_s^{j+1}\| > \Delta$ **então**
        Calcule $t \in \mathbb{R}$ tal que $d = d_s^j + t p^j$ satisfaz $\|d\| = \Delta$
        Faça $d^k = d$
        senão
          $r^{j+1} = r^j + t_j A p^j$
          se $r^{j+1} = 0$ **então**
            Faça $d^k = d_s^{j+1}$
            senão
              $\beta_j = \frac{(r^{j+1})^T r^{j+1}}{(r^j)^T r^j}$
              $p^{j+1} = -r^{j+1} + \beta_j p^j$
            fim
          fim
        fim
      fim
    fim
  fim
  $j = j + 1$
até *obter o passo* d^k;

Capítulo 2

Um método baseado em autovalores generalizados

2.1 A origem do método

Como vimos no capítulo anterior, os métodos para a resolução do subproblema de região de confiança envolvem procedimentos iterativos. Também vimos, na secção 1.5.4, como o subproblema (1.4) pode ser convertido num problema unidimensional para determinar zeros de função, que também é um problema resolvido de forma iterativa. No entanto, sabemos que existe uma correspondência entre as raízes de um polinômio e os autovalores de uma matriz, então pensando nisso podemos converter um problema de zero de função num problema de autovalores. Baseados nessa intuição, Gander et al. desenvolveram um método para a resolução do subproblema (1.4) resolvendo um único problema de autovalores.

Apesar dos resultados teóricos em 1989, quando o método foi apresentado, ele não obteve um bom desempenho computacional; acredita-se que a velocidade lenta e a perda de precisão eram causadas, em grande parte, pelas rotinas utilizadas para calcular autovalores disponíveis na época, o que explica o fato desse método ter sido negligenciado durante todos esses anos.

Em 2017, Adachi et al., baseando-se em [4], elaboraram um método que encontra a solução do subproblema resolvendo um único problema de autovalores generalizados. Vamos focar nossa atenção em compreender e explorar as características desse método e como podemos, a partir da escolha da matriz A, aproveitar a estrutura do modelo. Os testes de desempenho do método foram apresentados pelos autores em [1] e se mostraram bastante promissores.

2.2 Conceitos importantes

Nesta secção iremos dedicar um pouco da nossa atenção para fazer um breve resumo sobre dois conceitos muito importantes que iremos utilizar nas próximas secções para desenvolvermos o método: *autovalores generalizados* e *leques de matrizes*.

2.2.1 Leques de matrizes

Sejam A e B matrizes de dimensão $m \times n$. Podemos definir um conjunto de matrizes, que denominaremos

leque linear ou simplesmente leque[1], no qual cada matriz é da forma $A + \lambda B$, e λ é um escalar complexo. Podemos nos referir ao leque $A + \lambda B$ através do par $(A,\ B)$. Para mais detalhes ver [7].

Definição 2.2.1. *Um leque $(A,\ B)$ é dito regular se A e B são matrizes quadradas de mesma ordem e se $det(A + \lambda B) \neq 0$ para algum $\lambda \in \mathbb{C}$. Todos os outros leques são ditos singulares.*

Dois leques $(A,\ B)$ e $(C,\ D)$ de mesma ordem n são ditos equivalentes se existem matrizes reais P e Q de ordem n, não-singulares tais que

$$C = PAQ \qquad \text{e} \qquad D = PBQ;$$

e chamamos a transformação do leque $(A,\ B)$ no leque $(C,\ D)$ de *transformação de equivalência*. Se o leque $(A,\ B)$ é hermitiano ou simétrico, ou seja, se ambas as matrizes A e B são hermitianas ou simétricas, então podemos escrever

$$C = Q^H AQ \qquad \text{e} \qquad D = Q^H BQ;$$

e nesse caso chamamos a transformação do leque $(A,\ B)$ no leque $(C,\ D)$ de *transformação de congruência* e dizemos que os leques $(A,\ B)$ e $(C,\ D)$ são congruentes.

Definição 2.2.2. *Seja $A \in \mathbb{R}^{m\times n}$. Chamamos de núcleo ou espaço nulo de A ao conjunto de todos os vetores $x \in \mathbb{R}^n$ tais que $Ax = 0_{\mathbb{R}^m}$, ou seja,*

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n \ :\ Ax = 0_{\mathbb{R}^m}\}.$$

Definição 2.2.3. *Seja $A \in \mathbb{R}^{m\times n}$. Chamamos de imagem de A ao conjunto de todos os vetores $y \in \mathbb{R}^m$ tais que $y = Ax$, para algum $x \in \mathbb{R}^n$, ou seja,*

$$\mathcal{I}m(A) = \{y \in \mathbb{R}^m \ :\ y = Ax,\ \text{para algum } x \in \mathbb{R}^n\}.$$

Teorema 2.2.1. *Suponha que A e B são matrizes simétricas de ordem n e seja $C(\mu)$ definida como*

$$C(\mu) = \mu A + (1 - \mu)B, \qquad \mu \in \mathbb{R}. \tag{2.1}$$

Se existe $\mu \in [0,\ 1]$ tal que $C(\mu)$ é semidefinida positiva e

$$\mathcal{N}(C(\mu)) = \mathcal{N}(A) \cap \mathcal{N}(B) \tag{2.2}$$

então existe uma matriz W não singular tal que $W^T AW$ e $W^T BW$ são matrizes diagonais.

Demonstração. Veja [6, Teorema 8.7.1]. □

Na maioria dos casos que analisaremos as condições do Teorema 2.2.1 são satisfeitas, isso porque em geral uma das duas matrizes que definem o leque é definida positiva e como veremos na Secção 2.3 a nossa matriz B sempre será definida positiva. Em [6, Subsecção 8.7.2], os autores apresentam um algoritmo capaz de realizar uma diagonalização por congruência de forma simultânea no leque $(A,\ B)$, de tal forma que

$$W^T(A,\ B)W = (\Lambda,\ I_n),$$

ou seja, $W^T AW = \Lambda$ e $W^T BW = I_n$. Para isso basta que o leque seja simétrico e que a matriz B seja definida positiva. Este procedimento será muito utilizado para as próximas secções, pois ele será a base das demonstrações pertinentes à compreensão do método que estudamos.

[1] Uma livre tradução do termo *pencil*

2.2.2 Autovalores generalizados

Agora que já definimos um leque de matrizes, vamos discutir um pouco sobre uma de suas aplicações, *autovalores generalizados*. Primeiramente vamos recordar o problema de autovalores tradicional. Um escalar $\lambda \in \mathbb{C}$ é um autovalor da matriz $A \in \mathbb{C}^{n\times n}$, se existe um vetor não nulo $v \in \mathbb{C}^n$ tal que

$$(A - \lambda I_n)v = 0_{\mathbb{C}^n}, \tag{2.3}$$

note que podemos generalizar essa ideia substituindo a matriz identidade na equação acima por uma matriz $B \in \mathbb{C}^{n\times n}$, neste caso temos:

$$(A - \lambda B)v = 0_{\mathbb{C}^n}. \tag{2.4}$$

e v é um autovetor associado ao autovalor generalizado λ. Muitos autores costumam chamar o par $(v,\ \lambda)$ de autopar. Como veremos nas próximas secções, o problema de autovalores generalizados que estamos a fim de resolver será sempre um caso particular em que A e B são matrizes reais e simétricas; e B é definida positiva. Vamos então nos concentrar em como podemos resolver o problema de autovalores generalizados apenas para esse caso.

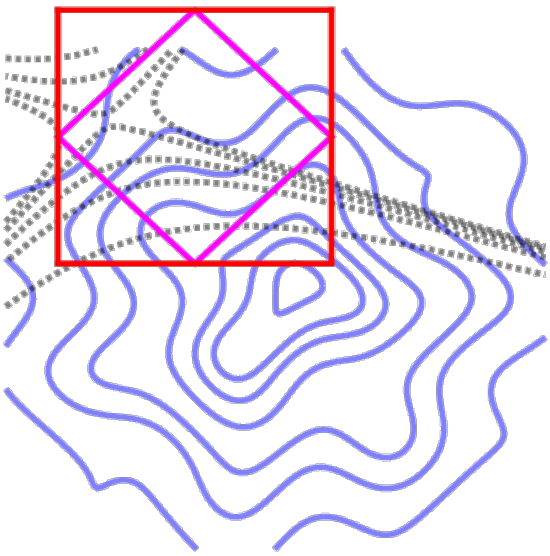
A hipótese da B ser definida positiva, implica que existe uma matriz G triangular inferior tal que $B = GG^T$, então seja $\tilde{A} = G^{-1}AG^{-T}$. Podemos transformar o par (A, B) no par $(\tilde{A}, I_n)$ e dessa forma podemos converter o problema (2.4) num problema de autovalores tradicional (2.3). E mais ainda, como A também é simétrica, podemos escrever $A = VDV^T$, onde V é uma matriz ortogonal e D é uma matriz diagonal e ambas possuem apenas elementos reais, então podemos verificar facilmente que $\tilde{A}$ é similar a D, pois $\tilde{A} = G^{-1}VDV^TG^{-T}$, logo o problema (2.4) possui autovalores reais.

É importante esclarecermos que daqui em diante iremos seguir a convenção adotada por Adachi et al. [1], então toda vez que mencionarmos o par (A, B), estaremos nos referindo ao leque da forma $(A + \lambda B)$ e não $(A - \lambda B)$ como em (2.4). Assim, os autovalores do leque (A, B) formarão o conjunto $\{\mu_1, \cdots, \mu_n\}$, com $-\mu_j$, $j = 1, ..., n$ autovalores generalizados do problema (2.4), e assumimos a ordenação $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$.

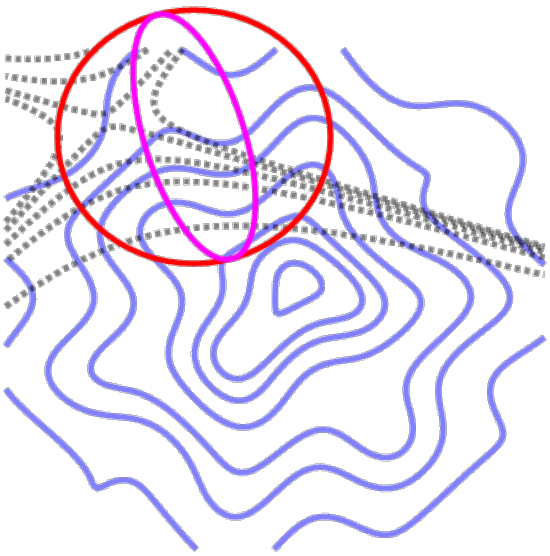
2.3 “Deformando” a região de confiança

No capítulo 1, escolhemos a norma euclidiana para construir a região de confiança, pois em geral ela é a mais utilizada, mas é importante esclarecer que poderíamos ter escolhido qualquer outra norma e a resolução do problema se daria com as peculiaridades decorrentes da norma utilizada.

A norma é responsável por determinar o formato da nossa região de confiança, então no primeiro capítulo, dado o ponto x^k, confiávamos *igualmente* em todas as direções a partir dele, ou seja, todas as direções vão possuir exatamente a mesma norma; na Figura 2.3 abaixo ilustramos algumas regiões de confiança geradas com diferentes normas:



(a) Nesta figura temos 2 regiões de confiança: em rosa gerada com a norma-1 e em vermelho gerada com a norma infinito.



(b) Nesta outra também temos 2 regiões de confiança: em rosa gerada com a *norma da energia* que definiremos a seguir e em vermelho gerada com a norma euclidiana.

A cada iteração iremos resolver o subproblema (1.4) resolvendo um único problema de autovalores generalizados, onde vamos tomar a matriz $A = A_k$ e a matriz B vai ser construída de tal maneira que possamos nos beneficiar da estrutura do modelo a cada iteração para definir a região de confiança. Em [1] os autores não se atem à construção da matriz B, a única hipótese é que ela seja uma matriz simétrica definida positiva. Mas qual seria essa estrutura da qual estamos tentando nos beneficiar?

Na prática é muito comum termos problemas onde as variáveis encontram-se em escalas com ordem de grandeza muito diferentes, esses problemas são chamados de *mal dimensionados* e eles podem trazer algumas dificuldades numéricas, pois a contribuição de uma das variáveis pode ser ínfima em comparação com as outras. Em [2, Secção 6.7.1], os autores trazem uma situação envolvendo circuitos elétricos onde isso acontece.

Como não há uma forma de determinar previamente como vão ser os modelos quadráticos gerados ao longo do processo de minimização, é muito provável que possamos gerar modelos mal dimensionados e, nesses casos, talvez não seja interessante que todas as direções tenham exatamente a mesma norma, agora seria interessante confiar mais em umas direções do que outras, ou seja, estamos interessados em *priorizar* as direções de maior deformação do modelo; e para isso iremos substituir a $\|.\|_2$ por $\|.\|_B$, que é conhecida como norma de energia de deformação, e é definida como $\|x\|_B = \sqrt{x^T Bx}$. Em outras palavras, nossa região de confiança vai deixar de ser uma esfera em $\mathbb{R}^n$ para ser um elipsóide em $\mathbb{R}^n$. Na Figura 2.1 podemos

verificar que o minimizador local do modelo encontra-se no interior da região de confiança gerada com a norma da energia e no exterior da região de confiança gerada com a norma euclidiana.

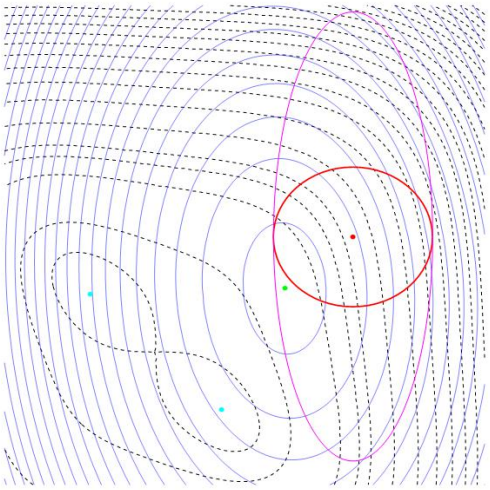


Figura 2.1: *Em magenta temos uma região gerada usando a norma de energia de deformação e, em vermelho, uma região gerada usando a norma euclidiana.*

2.3.1 A escolha da matriz B

Como foi discutido na secção (1.2), a cada iteração estamos definindo a matriz A do modelo $m(d)$ como sendo a hessiana da função objetivo f, pois dessa forma conseguimos que o modelo tenha uma curvatura similar a da função objetivo próximo do ponto corrente. Agora, precisaremos definir a matriz B de tal forma que ela contenha as direções preferenciais do modelo. Mais do que isso, vamos precisar também que B seja definida positiva, pois para que $\sqrt{x^T Bx}$ possa definir uma norma $\sqrt{x^T Bx} \geq 0, \forall x \in \mathbb{R}^n$ e a igualdade ocorre se, e somente se, $x = 0_{\mathbb{R}^n}$.

Como a função objetivo é de classe C^2 sua hessiana A é simétrica, então A admite decomposição espectral, ou seja,

$$A = VDV^T, \tag{2.5}$$

onde V e $D \in \mathbb{R}^{n\times n}$, V é ortogonal e suas colunas são uma base de autovetores de A e D é diagonal e seus elementos não nulos são os autovalores de A. A partir da decomposição obtida em (2.5) conseguimos obter todas as direções preferenciais do modelo $m(d)$ na matriz V, então podemos utilizar essa decomposição para construir a nossa matriz B. Porém, precisamos garantir que a nossa matriz B seja definida positiva, que é equivalente a garantir que todos os autovalores de B sejam estritamente positivos, logo podemos escrever a B da seguinte maneira

$$B = V|D|V^T. \tag{2.6}$$

Veja que a estratégia (2.6) garante apenas que B é uma matriz semidefinida positiva, pois se A possuir algum autovalor nulo somente o valor absoluto não vai ser capaz de torná-lo todos estritamente positivos. Então é necessário criarmos uma salvaguarda para matriz D que pode ser como definimos abaixo

$$\begin{cases} d_{ii} = d_{ii}, & \text{se } d_{ii} \neq 0 \quad \forall i = 1, \cdots, n. \\ d_{ii} = 1, & \text{caso contrário.} \end{cases} \tag{2.7}$$

Como estamos interessados em analisar problemas bidimensionais para termos um apelo geométrico, podemos ser audaciosos e construir a B como descrevemos em (2.6) utilizando a salvaguarda (2.7), pois dessa forma conseguimos extrair o máximo de informações possíveis do modelo. Mas note que para problemas de grande porte essa estratégia pode acabar sendo muito custosa, pois o custo para obter a decomposição espectral de A é da $O(n^3)$ operações.

2.4 Entendendo o método

Antes de mais nada, vamos generalizar o Teorema 1.5.1 para o subproblema

$$\min_{d\in\mathbb{R}^n} m(d) = f(x) + g^T d + \frac{1}{2} d^T A d \tag{2.8}$$
$$\text{s.a: } \|d\|_B \leq \Delta.$$

Teorema 2.4.1. *Um passo d^* é uma solução ótima para o subproblema de região de confiança (2.8) se, e somente se, existe $\lambda^* \geq 0$ tal que:*

$$\|d^*\|_B \leq \Delta, \tag{2.9}$$
$$(A + \lambda^* B)d^* = -g, \tag{2.10}$$
$$\lambda^*(\Delta - \|d^*\|_B) = 0, \tag{2.11}$$
$$A + \lambda^* B \geq 0 \tag{2.12}$$

Como a B é definida positiva, podemos facilmente fazer uma mudança de variáveis e recaímos no problema com $B = I_n$, e sua demonstração foi apresentada em na Secção 1.5 no Teorema 1.5.1.

A partir do leque (A, B) iremos construir 2 outros leques, $M(\lambda)$ de ordem $2n + 1$ e $\tilde{M}(\lambda)$ de ordem $2n$:

$$M(\lambda) = \begin{pmatrix} \Delta^2 & 0 & g^T \\ 0 & -B & A + \lambda B \\ g & A + \lambda B & O_n \end{pmatrix} \tag{2.13}$$

e

$$\tilde{M}(\lambda) = \begin{pmatrix} -B & A + \lambda B \\ A + \lambda B & -\frac{gg^T}{\Delta^2} \end{pmatrix}. \tag{2.14}$$

A construção do método se dá a partir de duas informações fundamentais sobre esses dois leques:

- os autovalores destes dois leques contém os valores de λ que satisfazem as condições de KKT (2.10);
- a solução (λ^*, d^*) pode ser encontrada através do maior autopar real.

Primeiro, vamos mostrar que todos os multiplicadores de Lagrange nos pontos que satisfazem a condição (2.10) na fronteira da região de confiança, ou seja, $\|d\|_B = \Delta$, inclusive λ^*, são autovalores de $M(\lambda)$ e $\tilde{M}(\lambda)$. Para isso é suficiente mostrarmos que $det(M(\lambda)) = det(\tilde{M}(\lambda)) = 0$ para todo λ satisfazendo (2.10).

Lema 2.4.1. *Seja (λ, d) um par satisfazendo a condição (2.10) e $\|d\|_B = \Delta$, então temos que $det(M(\lambda)) = 0$ e $det(\tilde{M}(\lambda)) = 0$.*

Demonstração. Vamos mostrar que tanto $M(\lambda)$ quanto $\tilde{M}(\lambda)$ são singulares para o autovalor λ. Primeiro, vamos separar em 2 casos, quando $det(A + \lambda B) = 0$ e quando $det(A + \lambda B) \neq 0$:

1. $det(A + \lambda B) = 0$

   Seja $v \in \mathbb{R}^{2n+1}$, vamos analisar o seguinte sistema linear homogêneo:

$$M(\lambda)\begin{pmatrix} 0_{n+1} \\ x \end{pmatrix} = 0$$

$$\Rightarrow \begin{pmatrix} \Delta^2 & 0 & g^T \\ 0 & -B & A + \lambda B \\ g & A + \lambda B & O_n \end{pmatrix} \cdot \begin{pmatrix} 0_{n+1} \\ x \end{pmatrix} = \begin{pmatrix} g^T x \\ (A + \lambda B)x \\ 0_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0_n \\ 0_n \end{pmatrix}.$$

   Como $det(A + \lambda B) = 0$, então $(A + \lambda B)$ é singular, então existe $x \neq 0_n$ tal que $(A + \lambda B)x = 0_n$ e por hipótese $g \in \mathit{Im}(A + \lambda B)$, logo existe w tal que

$$(A + \lambda B)w = g.$$

   Daí e da simetria de $(A + \lambda B)$ temos:

$$\Rightarrow g^T x = ((A + \lambda B)w)^T x = w^T (A + \lambda B)^T x = w^T \underbrace{(A + \lambda B)x}_{0} = 0.$$

   Então existe um vetor v não nulo que satisfaz o sistema linear homogêneo $M(\lambda)v = 0$, logo $M(\lambda)$ é singular.

   Agora, seja $v \in \mathbb{R}^{2n}$, vamos analisar o seguinte sistema linear homogêneo:

$$\tilde{M}(\lambda)\begin{pmatrix} 0_n \\ x \end{pmatrix} = 0$$

$$\Rightarrow \begin{pmatrix} -B & A + \lambda B \\ A + \lambda B & -\frac{gg^T}{\Delta^2} \end{pmatrix} \cdot \begin{pmatrix} 0_n \\ x \end{pmatrix} = \begin{pmatrix} (A + \lambda B)x \\ -\frac{g}{\Delta^2} g^T x \end{pmatrix} = \begin{pmatrix} 0_n \\ 0_n \end{pmatrix}.$$

   Pelos mesmos argumentos, temos que existe um vetor v não nulo que satisfaz o sistema linear homogêneo $\tilde{M}(\lambda)v = 0$, logo $\tilde{M}(\lambda)$ é singular.

2. $det(A + \lambda B) \neq 0$

   Antes, vamos reescrever o vetor parametrizado em (1.13), substituindo a matriz identidade pela matriz B, logo temos:

$$d(\lambda) = -(A + \lambda B)^{-1} g, \tag{2.15}$$

   assumindo que a inversa esteja bem definida. Agora utilizando o vetor $d(\lambda)$, vamos definir uma matriz auxiliar $X(\lambda)$:

$$X(\lambda) = \begin{pmatrix} 1 & & \\ d(\lambda) & I_n & \\ & & I_n \end{pmatrix}.$$

   Note que $X(\lambda)$ é triangular inferior com diagonal unitária, ou seja, $det(X(\lambda)) = 1$. Então, temos:

$$det(M(\lambda)) = det(X(\lambda)^T) \cdot det(M(\lambda)) \cdot det(X(\lambda)) = det(X(\lambda)^T \cdot M(\lambda) \cdot X(\lambda)),$$

expandindo temos:

$$X(\lambda)^T \cdot M(\lambda) \cdot X(\lambda) = \begin{pmatrix} 1 & d(\lambda) & \\ & I_n & \\ & & I_n \end{pmatrix} \cdot \begin{pmatrix} \Delta^2 & 0 & g^T \\ 0 & -B & A+\lambda B \\ g & A+\lambda B & O_n \end{pmatrix} \cdot \begin{pmatrix} 1 & & \\ d(\lambda) & I_n & \\ & & I_n \end{pmatrix}$$

$$= \begin{pmatrix} \Delta^2 - d(\lambda)^T B d(\lambda) & -d(\lambda)^T B & g^T + d(\lambda)^T (A+\lambda B) \\ -Bd(\lambda) & -B & A+\lambda B \\ g + (A+\lambda B)d(\lambda) & A+\lambda B & O_n \end{pmatrix}.$$

Note que de (2.15) temos $d(\lambda) = -(A+\lambda B)^{-1} g$, substituindo no último bloco da primeira coluna e primeira linha:

$$g + (A+\lambda B)(-(A+\lambda B)^{-1} g)$$
$$\Rightarrow g - \underbrace{(A+\lambda B)(A+\lambda B)^{-1}}_{I_n} g = g - g = 0.$$

Obtemos então a seguinte matriz:

$$\begin{pmatrix} \Delta^2 - d(\lambda)^T B d(\lambda) & -d(\lambda)^T B & 0 \\ -Bd(\lambda) & -B & A+\lambda B \\ 0 & A+\lambda B & O_n \end{pmatrix},$$

calculando o seu determinante chegamos na seguinte expressão:

$$det(M(\lambda)) = (-1)^n \cdot det(A+\lambda B)^2 \cdot (\Delta^2 - d(\lambda)^T B d(\lambda)),$$

por hipótese $det(A+\lambda B) \neq 0$ e $d(\lambda)$ está na fronteira da região de confiança, isso implica que $d(\lambda)^T B d(\lambda) = \Delta^2$ e portanto $det(M(\lambda)) = 0$.

Vamos agora pensar na $\tilde{M}$. Por ora, vamos utilizar apenas a matriz $M(\lambda)$ e uma segunda matriz auxiliar, triangular superior T, definida da seguinte maneira:

$$T = \begin{pmatrix} 1 & & -\frac{1}{\Delta^2} g^T \\ & I_n & \\ & & I_n \end{pmatrix}. \tag{2.16}$$

T também é unimodular, então temos:

$$det(M(\lambda)) = det(T^T) \cdot det(M(\lambda)) \cdot det(T) = det(T^T \cdot M(\lambda) \cdot T),$$

expandindo, obtemos:

$$T^T \cdot M(\lambda) \cdot T = \begin{pmatrix} \Delta^2 & & \\ & -B & A+\lambda B \\ & A+\lambda B & -\frac{gg^T}{\Delta^2} \end{pmatrix} = \begin{pmatrix} \Delta^2 & \\ & \tilde{M}(\lambda) \end{pmatrix}.$$

Como já sabemos que $det(M(\lambda)) = 0$, conseguimos facilmente calcular o determinante de $\tilde{M}(\lambda)$:

$$det(M(\lambda)) = \Delta^2 det(\tilde{M}(\lambda)) \quad \Rightarrow \quad 0 = \Delta^2 det(\tilde{M}(\lambda)) \quad \Rightarrow \quad det(\tilde{M}(\lambda)) = 0$$

□

Relembrando da Definição 2.2.1, os leques $M(\lambda)$ e $\tilde{M}(\lambda)$ são regulares, isso implica que a quantidade de autovalores é igual à sua dimensão, logo $M(\lambda)$ vai possuir $2n+1$ autovalores enquanto $\tilde{M}(\lambda)$ vai possuir $2n$. Agora note que a equação racional (1.17) é um somatório de n funções racionais e cada uma delas pode ser transformada em uma função polinomial de grau 2, então podemos afirmar que (1.17) possui $2n$ raízes complexas contando as multiplicidades que é exatamente a dimensão de $\tilde{M}(\lambda)$.

Agora, note que uma das matrizes que compõem o leque $M(\lambda)$ possui uma linha e uma coluna nulas, ou seja, é uma matriz singular:

$$M(\lambda) = \begin{pmatrix} \Delta^2 & 0 & g^T \\ 0 & -B & A \\ g & A & O_n \end{pmatrix} + \lambda \begin{pmatrix} 0 & 0 & 0 \\ 0 & O_n & B \\ 0 & B & O_n \end{pmatrix}. \tag{2.17}$$

Isso implica que $M(\lambda)$ vai ter um autovalor no infinito e podemos ver facilmente que $\tilde{M}(\lambda)$ é o complemento de Schur de $M(\lambda)$, então temos que os autovalores de $M(\lambda)$ vão ser os $2n$ autovalores de $\tilde{M}(\lambda)$ mais o autovalor no infinito, totalizando $2n+1$.

Tendo isso em vista, vamos mostrar que λ^* está entre o maior autovalor real desses leques e o infinito, ou seja, $\lambda^* \in [\mu_n,\ \infty)$.

Teorema 2.4.2. *Para um par (λ^*, d^*) solução do subproblema de região de confiança (2.8) satisfazendo (2.10)- (2.12), tal que d^* está na fronteira. Então o multiplicador λ^* é igual ao maior autovalor real finito de $M(\lambda)$ e $\tilde{M}(\lambda)$; e $\lambda^* \in [\mu_n,\ \infty)$, onde μ_n é o maior autovalor do leque $A + \lambda B$.*

Demonstração. Da relação (2.12) temos que $A + \lambda^* B \geq 0$, ou seja, o conjunto de autovalores generalizados $\Lambda(A + \lambda^* B) = \{\mu_1, \mu_2, \cdots, \mu_n\}$ possui apenas elementos não negativos e com pelo menos 1 deles não nulo. Como μ_n é o maior deles então $\mu_n > 0$. E como discutimos na secção 1.5.3, $\lambda^* \geq \mu_n$.

Sabemos que se d^* está na fronteira, $\|d^*\|_B = \Delta$, é equivalente a buscar o zero de uma equação secular apropriada, salvo o caso $\lambda = \mu_n$. De fato, pelo Teorema 2.2.1, se $A + \lambda B$ é semidefinida positiva então existe W não singular (não necessariamente ortogonal) tal que

$$W^T(A + \lambda B)W = D_A + \lambda D_B,$$

com D_A e $D_B \in \mathbb{R}^{n\times n}$ diagonais. Com a hipótese que B é definida positiva, e usando o fato que matrizes diagonais comutam entre si, temos:

$$\underbrace{D_B^{-\frac{1}{2}} W^T}_{\overline{W}^T}(A + \lambda B)\underbrace{W D_B^{-\frac{1}{2}}}_{\overline{W}} = \underbrace{D_B^{-1} D_A}_{D} + \lambda I,$$

de onde segue que $\overline{W}^T(A + \lambda B)\overline{W} = D + \lambda I$, e portanto $\overline{W}^{-1}(A + \lambda B)^{-1}\overline{W}^{-T} = (D + \lambda I)^{-1}$, ou ainda, $(A + \lambda B)^{-1} = \overline{W}(D + \lambda I)^{-1}\overline{W}^T$.

Agora, como $d(\lambda) = -(A + \lambda B)^{-1} g$, obtemos

$$\begin{aligned} \|d(\lambda)\|_B^2 &= g^T(A + \lambda B)^{-T} B (A + \lambda B)^{-1} g \\ &= g^T \overline{W}(D + \lambda I)^{-1} \underbrace{\overline{W}^T B \overline{W}}_{I} (D + \lambda I)^{-1} \overline{W}^T g \\ &= (\overline{W}^T g)(D + \lambda I)^{-2} \overline{W}^T g \\ &= \sum_{j=1}^{n} \frac{(\overline{w}_j^T g)^2}{(-\mu_j + \lambda)^2}. \end{aligned}$$

Dessa forma, definimos

$$\hat{h}(\lambda) := \sum_{j=1}^{n} \frac{(\overline{w}_j^T g)^2}{(-\mu_j + \lambda)^2} - \Delta^2 \tag{2.18}$$

Para mostrar que λ^* é o maior autovalor real de $M(\lambda)$ vamos separar em dois casos:

1. $\lambda^* > \mu_n$

   Note que a função $\hat{h}(\lambda)$ definida em (2.18) é estritamente decrescente no intervalo $(\mu_n,\ \infty)$, ou seja, ela vai possuir apenas um único zero nesse intervalo. Então, $\tilde{M}(\lambda)$ tem exatamente um autovalor real maior que μ_n, que necessariamente será λ^*.

2. $\lambda^* = \mu_n$

   Da condição (2.10), se $\lambda^* = \mu_n$ então $(A + \lambda^* B)d^* = 0_{\mathbb{R}^n}$, isso implica que $g \in \mathcal{N}(A + \lambda^* B)$ e $\hat{h}(\lambda)$ não vai ter um pólo em $\lambda = \mu_n$, porém sabemos que ela é estritamente decrescente no intervalo $(\mu_{n-l},\ \infty)$, onde $l \geq 1$ é um inteiro tal que $\mu_{n-l} < \mu_{n-l+1} = \mu_n$. Veja que neste caso, temos que reescrever $\hat{h}(\lambda)$:

   $$\hat{h}(\lambda) + \Delta^2 = \sum_{j=1}^{n-l} \frac{(\overline{w}_j^T g)^2}{(-\mu_j + \lambda)^2}. \tag{2.19}$$

   Mais adiante iremos provar que $\hat{h}(\lambda) + \Delta^2 = \sum_{j=1}^{n-l} \frac{(\overline{w}_j^T g)^2}{(-\mu_j+\lambda)^2}$ é igual ao tamanho do vetor d com a menor B-norma tal que (2.10) é satisfeita.

   Agora se $\hat{h}(\lambda^*) > 0$ então de (2.19), temos que não existe uma solução que satisfaça (2.10) e (2.11), então $\hat{h}(\lambda^*) \leq 0$, e como ela é estritamente decrescente no intervalo $(\mu_{n-l},\ \infty)$, não temos uma solução para $\hat{h}(\lambda) = 0$ com $\lambda > \mu_n$. Logo λ^* também é o maior autovalor real de $M(\lambda)$ e $\tilde{M}(\lambda)$ nesse caso.

□

E por fim, vamos mostrar que podemos obter a solução do subproblema (2.8) através do autovetor de $M(\lambda)$ e $\tilde{M}(\lambda)$ associado a λ^*.

Teorema 2.4.3. *Os autovetores de $\tilde{M}(\lambda)$ e $M(\lambda)$, respectivamente, associados ao maior autovalor real finito $\lambda = \lambda^*$ correspondem a*

$$\tilde{M}(\lambda)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 0_n \qquad \Leftrightarrow \qquad M(\lambda)\begin{pmatrix} -\frac{1}{\Delta^2} g^T y_2 \\ y_1 \\ y_2 \end{pmatrix} = 0_{n+1}. \tag{2.20}$$

Além disso, se $g^T y_2 \neq 0$, então a solução do subproblema de região de confiança (2.8) pode ser obtida por:

$$d^* = -\frac{\Delta^2}{g^T y_2} y_1 \tag{2.21}$$

Demonstração. ($\Rightarrow$)
Seja $\tilde{v} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ autovetor do leque $\tilde{M}(\lambda)$ associado a λ, então

$$\tilde{M}(\lambda)\tilde{v} = 0_n \tag{2.22}$$

Agora seja $v = \begin{pmatrix} -\frac{1}{\Delta^2} g^T y_2 \\ y_1 \\ y_2 \end{pmatrix}$. Vamos expandir o produto matricial $M(\lambda)v$:

$$\begin{pmatrix} \Delta^2 & 0 & g^T \\ 0 & -B & A + \lambda B \\ g & A + \lambda B & O_n \end{pmatrix} \begin{pmatrix} -\frac{1}{\Delta^2} g^T y_2 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} -g^T y_2 + g^T y_2 \\ \tilde{M}(\lambda)\tilde{v} \end{pmatrix} = \begin{pmatrix} 0 \\ 0_n \end{pmatrix} \tag{2.23}$$

($\Leftarrow$)

Agora vamos mostrar que se $v = \begin{pmatrix} \alpha \\ y_1 \\ y_2 \end{pmatrix}$ é autovetor de $M(\lambda)$ então $\alpha = -\frac{1}{\Delta^2} g^T y_2$. Vamos utilizar novamente a matriz auxiliar T definida em (2.16), note que T é uma matriz inversível, então podemos facilmente verificar que:

$$M(\lambda) = T^{-T} \begin{pmatrix} \Delta^2 & \\ & \tilde{M}(\lambda) \end{pmatrix} T^{-1},$$

então

$$\begin{aligned} M(\lambda)v = T^{-T} \begin{pmatrix} \Delta^2 & \\ & \tilde{M}(\lambda) \end{pmatrix} T^{-1} v &= 0_{n+1} \\ = T^{-T} \begin{pmatrix} \Delta^2(\alpha + \frac{g^T y_2}{\Delta^2}) \\ \tilde{M}(\lambda) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \end{pmatrix} &= 0_{n+1}, \end{aligned}$$

como $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ é autovetor de $\tilde{M}(\lambda)$ então temos:

$$T^{-T} \begin{pmatrix} \Delta^2(\alpha + \frac{g^T y_2}{\Delta^2}) \\ 0_n \end{pmatrix} = 0_{n+1}$$

logo

$$\begin{aligned} \underbrace{\Delta^2}_{\geq 0}(\alpha + \frac{g^T y_2}{\Delta^2}) &= 0 \\ \Rightarrow \alpha + \frac{g^T y_2}{\Delta^2} &= 0 \\ \Rightarrow \alpha &= -\frac{g^T y_2}{\Delta^2}. \end{aligned}$$

Por último vamos mostrar a relação (2.21). Primeiro, vamos escrever o leque $\tilde{M}(\lambda)$:

$$\begin{aligned} \begin{pmatrix} -B & A \\ A & -\frac{gg^T}{\Delta^2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= -\lambda \begin{pmatrix} O_n & B \\ B & O_n \end{pmatrix} \\ \Rightarrow \begin{pmatrix} -By_1 + Ay_2 \\ Ay_1 - \frac{gg^T}{\Delta^2} y_2 \end{pmatrix} &= -\lambda \begin{pmatrix} By_2 \\ By_1 \end{pmatrix} \\ \Rightarrow \begin{pmatrix} -By_1 + (A + \lambda B)y_2 \\ (A + \lambda B)y_1 - \frac{gg^T}{\Delta^2} y_2 \end{pmatrix} &= \begin{pmatrix} 0_n \\ 0_n \end{pmatrix}. \end{aligned}$$

Agora analisando o último bloco, temos:

$$(A + \lambda B)y_1 = \frac{gg^T}{\Delta^2}y_2,$$

Podemos observar que o lado direito é um múltiplo escalar de g. Então multiplicando os dois membros da equação por Δ^2, depois dividindo por $g^T y_2$, que não é nulo por hipótese, e multiplicando tudo por (-1) obtemos:

$$(A + \lambda B)\underbrace{(-\frac{\Delta^2}{g^T y_2}y_1)}_{d^*} = -g$$

relembrando da expressão em (2.10), concluímos que

$$d^* = -\frac{\Delta^2}{g^T y_2}y_1$$

□

Com isso provamos que a solução do subproblema (2.8) pode ser obtida através do maior autopar real $M(\lambda)$, sempre que $g^T y_2 \neq 0_{\mathbb{R}^n}$. Porém, ainda precisamos mostrar que o mesmo pode ser feito no caso $g^T y_2 = 0_{\mathbb{R}^n}$, ou seja, o *caso difícil*. Neste caso a obtenção da solução vai ser um pouco mais trabalhosa, porém ainda vamos conseguir obtê-la através através do maior do maior autopar real $\tilde{M}(\lambda)$.

Teorema 2.4.4. *Suponha que o subproblema da região de confiança corresponda a um* caso difícil *e o par* $(\lambda^*,\ d^*)$ *satisfaça (2.10)-(2.12) e* $\|d^*\|_B = \Delta$ *com* $\lambda^* = \mu_n$. *Seja* $r = dim(\mathcal{N}(A + \lambda^* B)$ *e* $V = [v_1, \cdots, v_r]$ *uma base para* $\mathcal{N}(A + \lambda^* B)$ $B-$*ortogonal, ou seja,* $V^T BV = I_r$. *Para um* $\alpha > 0$ *arbitrário, definimos:*

$$H := \left(A + \lambda^* B + \alpha \sum_{i=1}^{r} Bv_i v_i^T B\right). \tag{2.24}$$

Então H é definida positiva, $q = -H^{-1}g$ *é a solução de norma B mínima para o sistema linear* $(A + \lambda^* B)d = -g$, *ou seja,*

$$q = argmin_d\{\|s\|_B \ :\ (A + \lambda^* B)d = -g\}. \tag{2.25}$$

Além disso, para qualquer $v \in \mathcal{N}(A + \lambda^* B)$ *não-nulo existe um escalar* $\eta \in \mathbb{R}$ *tal que* $d^* = q + \eta v$ *é uma solução para o subproblema de região de confiança (2.8).*

Demonstração. Primeiro iremos mostrar que $H \succ 0$. Seja W tal que $W^T(A, B)W = (\Lambda, I_n)$, então temos

$$W^T(A + \lambda^* B)W = diag(\lambda^* - \mu_1, \cdots, \lambda^* - \mu_{n-d}, 0, \cdots, 0) \tag{2.26}$$

Vamos particionar a matriz W da seguinte maneira $W = [W_1 \ \ W_2]$, com $(W_1)_{n\times(n-r)}$ e $(W_2)_{n\times r}$ e de (2.26) temos

$$(A + \lambda^* B)W_2 = (A + \lambda^* B)V = O_{n-r}; \tag{2.27}$$

$$W_2^T BW_2 = V^T BV = I_r. \tag{2.28}$$

De (2.27) e (2.28) podemos concluir que V e W_2 são $B-ortogonais$ e ambas geram o mesmo subespaço, o gerado pelos autovetores de $(A + \lambda^* B)$ associados ao autovalor λ^*. Logo, existe uma matriz ortogonal $Q \in \mathbb{R}^{r\times r}$ tal que $V = W_2 Q$, e temos

$$W^T(BVV^T B)W = [W_1 \ \ W_2]^T BW_2 Q(W_2 Q)^T B[W_1 \ \ W_2]$$

$$= [W_1 \;\; W_2]^T BW_2W_2^TB[W_1 \;\; W_2]$$
$$= \begin{pmatrix} O_{(n-r)\times r} \\ I_r \end{pmatrix}\begin{pmatrix} O_{r\times(n-r)} & I_r \end{pmatrix} = \begin{pmatrix} O_{n-r} & \\ & I_r \end{pmatrix}.$$

Em que a última igualdade decorre de W ser tal que $W^TBW = [W_1 \;\; W_2]^TB[W_1 \;\; W_2] = I_n$. Daí temos

$$W^THW = \left(W^TAW + \lambda^*W^TBW + \alpha\sum_{i=1}^{r} W^TBv_iv_i^TBW\right)$$
$$= \left(W^TAW + \lambda^*I_n + \alpha\sum_{i=1}^{r} W^TBv_iv_i^TBW\right)$$
$$= diag(\lambda^* - \mu_1, \cdots, \lambda^* - \mu_{n-r},\ \alpha, \cdots, \alpha); \tag{2.29}$$

ou seja, W^THW é uma matriz diagonal e seus autovalores são os elementos da diagonal. Como a diagonal possui elementos estritamente positivos, então W^THW é definida positiva, pela Lei da Inércia de Sylvester, os autovalores de H também são todos positivos, e portanto H também é definida positiva.

Agora vamos mostrar que $q = -H^{-1}g$ é uma solução para o sistema linear $(A + \lambda^*B)d = -g$. Primeiro definimos $\tilde{\Lambda} = -diag(\mu_1, \cdots, \mu_{n-r})$ e de (2.26) podemos escrever:

$$(A + \lambda^*B) = W^{-T}\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & O_r \end{pmatrix}W^{-1}$$

daí e de (2.29) temos:

$$W^THW = \begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & \alpha I_r \end{pmatrix}$$

então

$$H^{-1} = W\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & \alpha I_r \end{pmatrix}^{-1}W^T.$$

Tendo isso vamos expandir o seguinte produto

$$(A + \lambda^*B)q = -W^{-T}\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & O_r \end{pmatrix}W^{-1}W\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & \alpha I_r \end{pmatrix}^{-1}W^Tg$$
$$= -W^{-T}\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & \alpha I_r \end{pmatrix}\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & \alpha I_r \end{pmatrix}^{-1}W^Tg$$
$$= -W^{-T}\begin{pmatrix} I_{n-r} & \\ & O_r \end{pmatrix}W^Tg,$$

mas sabemos que

$$g = -W^{-T}\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & O_r \end{pmatrix}W^{-1}d$$

substituindo obtemos

$$(A + \lambda^*B)q = W^{-T}\begin{pmatrix} I_{n-r} & \\ & O_r \end{pmatrix}W^TW^{-T}\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & O_r \end{pmatrix}W^{-1}d$$
$$= W^{-T}\begin{pmatrix} I_{n-r} & \\ & O_r \end{pmatrix}\begin{pmatrix} \tilde{\Lambda} + \lambda^*I_{n-r} & \\ & O_r \end{pmatrix}W^{-1}d = -g.$$

Logo $q = -H^{-1}g$ é uma solução do sistema linear $(A + \lambda^* B)d = -g$. Agora vamos provar que q é solução com norma B mínima. Podemos escrever uma solução genérica d como sendo $d = q + v$ onde $v \in \mathcal{N}(A + \lambda^* B)$, então vamos mostrar que

$$\|q\|_B \leq \|q + v\|_B \qquad \forall\ v \in \mathcal{N}(A + \lambda^* B).$$

Observe que

$$\|q + v\|_B^2 = (q + v)^T B(q + v) = q^T Bq + 2v^T Bq + v^T Bv,$$

como B é definida positiva, sabemos que o último termo da equação acima é positivo, então basta mostrarmos que $2v^T Bq \geq 0$ para todo $v \in \mathcal{N}(A + \lambda^* B)$. Para isso, vamos observar que q pode ser escrito como $q = -H^{-1}g = H^{-1}(A + \lambda^* B)d$ e vamos substituí-lo em Bq, obtendo:

$$Bq = BH^{-1}(A + \lambda^* B)d,$$

e aí com a decomposição em W podemos verificar que $BH^{-1}(A + \lambda^* B) = (A + \lambda^* B)H^{-1}B$. De fato,

$$\begin{aligned}
BH^{-1}(A + \lambda^* B) &= W^{-T}W^{-1}\left[W\begin{pmatrix} \tilde{\Lambda} + \lambda^* I_{n-r} & \\ & \alpha I_r \end{pmatrix}^{-1} W^T\right]\left[W^{-T}\begin{pmatrix} \tilde{\Lambda} + \lambda^* I_{n-r} & \\ & O_r \end{pmatrix} W^{-1}\right] \\
&= W^{-T}\begin{pmatrix} \tilde{\Lambda} + \lambda^* I_{n-r} & \\ & O_r \end{pmatrix} W^{-1} \\
&= \left[W^{-T}\begin{pmatrix} \tilde{\Lambda} + \lambda^* I_{n-r} & \\ & O_r \end{pmatrix} W^{-1}\right]\left[W\begin{pmatrix} \tilde{\Lambda} + \lambda^* I_{n-r} & \\ & \alpha I_r \end{pmatrix}^{-1} W^T\right] W^{-T}W^{-1} \\
&= (A + \lambda^* B)H^{-1}B.
\end{aligned}$$

Então temos que para todo $v \in \mathcal{N}(A + \lambda^* B)$

$$v^T Bq = v^T(BH^{-1}(A + \lambda^* B)q) = v^T(A + \lambda^* B)H^{-1}Bq = \underbrace{(v^T(A + \lambda^* B))}_{0_{1\times n}} H^{-1}Bq = 0,$$

então

$$\|q\|_B \leq \|q + v\|_B \qquad \forall\ v \in \mathcal{N}(A + \lambda^* B).$$

E por último, temos por hipótese que $\|d^*\|_B = \Delta$. Seja $d^* = q + \eta v$, como acabamos de provar $\|q\|_B \leq \|q + v\|_B$, obtemos a seguinte equação quadrática em η

$$\eta^2 v^T Bv + 2\eta v^T Bq + q^T Bq = \Delta^2,$$

como $v^T Bq = 0$, podemos reescrever da seguinte forma:

$$\eta^2 v^T Bv + q^T Bq - \Delta^2 = 0,$$

$$\Rightarrow \eta = \pm\sqrt{\frac{\Delta^2 - \|q\|_B}{v^T Bv}},$$

e como $\|q\|_B^2 - \Delta^2 \leq 0$ então $\eta \in \mathbb{R}$. □

2.5 Visualização dos Experimentos

Nesta secção apresentaremos algumas visualizações geradas com *software Mathematica 11.1*, porém todos os dados foram obtidos no *MatLab R2018a*. Nas figuras vamos utilizar pontos próximos da solução apenas para conseguirmos visualizar melhor a convergência do método, mas iremos apresentar algumas tabelas/gráficos com pontos iniciais mais distantes da solução.

O código original elaborado pelos autores Adachi et al. em 2017, encontra-se disponível online na página: `https://www.opt.mist.i.u-tokyo.ac.jp/~nakatsukasa/codes/TRSgep.m`. É importante destacar que o código A.2, que consta no Apêndice é uma adaptação do original, que foi implementado para problemas de grande porte. Em nosso caso, foi preciso adaptar o cálculo dos autovalores generalizados, pois trabalhamos com problemas de baixa dimensão.

2.5.1 Função de Rosenbrock

Nosso primeiro teste foi com a função de Rosenbrock, $f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$. A Figura 2.2 ilustra um dos um dos testes que foi feito a partir do ponto inicial $x_0 = (0,\ 2)$ e com $\Delta_0 = \dfrac{\|\nabla f(x_0)\|_2}{100}$. Essa função é bastante conhecida na literatura por seu caráter mal dimensionado e por isso escolhemos um ponto mais próximo do minimizador pra obtermos a convergência em poucas iterações. O registro dessa rodada pode ser acompanhado na Tabela 2.1, em que vemos o decréscimo dos valores de função a cada iteração (também ilustrado na Figura 2.3), bem como a dinâmica de atualização dos raios das regiões de confiança.

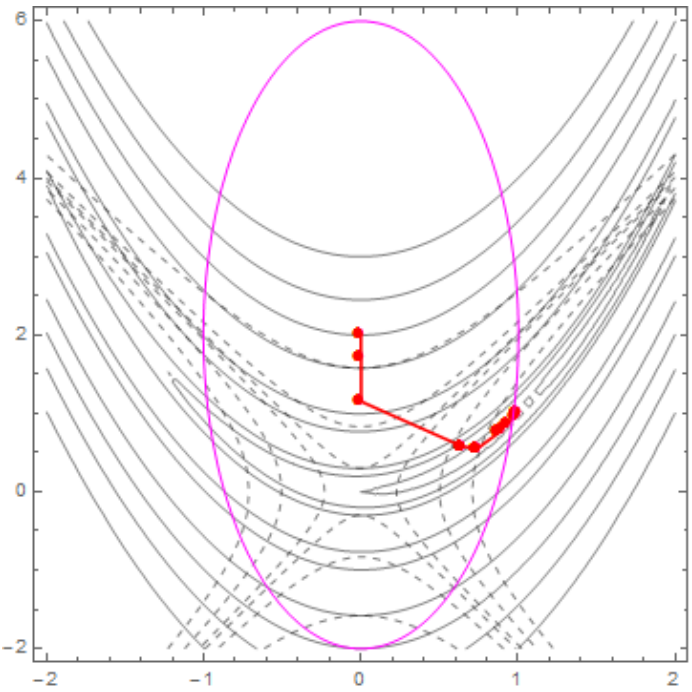


Figura 2.2: *Curvas de nível da função (traço contínuo), do modelo (tracejado), a região de confiança da primeira iteração e a trajetória dos pontos até o minimizador local da função.*

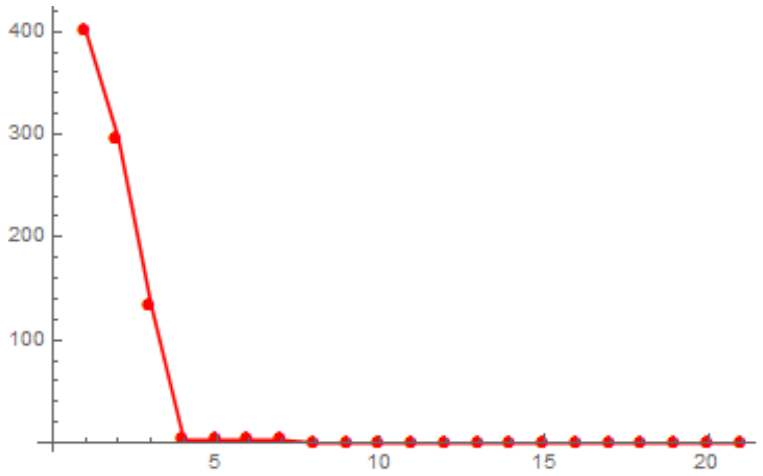


Figura 2.3: *gráfico do valor da função objetivo a cada iteração, num total de 21 iterações.*

Tabela 2.1: *Registro das iterações do método na função de Rosenbrock*

Iteração k	***Ponto*** x_k	***Valor funcional*** $f(x_k)$	***Status do Raio***
0	(0.000000, 2.000000)	401	—
1	(0.000276, 1.717154)	295.861307	Aumentou
2	(0.000969, 1.151463)	133.584566	Aumentou
3	(0.645368, 0.578350)	2.745306	Inalterado
4	(0.645368, 0.578350)	2.745306	Reduziu
5	(0.645368, 0.578350)	2.745306	Reduziu
6	(0.645368, 0.578350)	2.745306	Reduziu
7	(0.741436, 0.552648)	0.067709	Aumentou
8	(0.741436, 0.552648)	0.067709	Reduziu
9	(0.741436, 0.552648)	0.067709	Reduziu
10	(0.741436, 0.552648)	0.067709	Reduziu
11	(0.741436, 0.552648)	0.067709	Reduziu
12	(0.741436, 0.552648)	0.067709	Reduziu
13	(0.878170, 0.754765)	0.041794	Inalterado
14	(0.892585, 0.788403)	0.018434	Aumentou
15	(0.932953, 0.868772)	0.004761	Inalterado
16	(0.983519, 0.964753)	0.000925	Inalterado
17	(0.994424, 0.988760)	0.000033	Inalterado
18	(0.999870, 0.999711)	0.000000	Inalterado
19	(0.999999, 0.999998)	0.000000	Inalterado
20	(1.000000, 1.000000)	0.000000	Inalterado

2.5.2 Função Quártica

Trabalhamos agora com uma função mais bem comportada, a função quártica $f(x) = x^4 + y^4 + 4xy + 1$, a partir do ponto inicial $x_0 = (5,\ 4)$ e com $\Delta_0 = \frac{\|\nabla f(x_0)\|_2}{500}$. Note que agora iniciamos de um ponto mais distante do minimizador e com um raio inicial bem pequeno, ainda assim obtivemos convergência. E em comparação com o primeiro teste, mesmo estando mais distante e iniciando com um raio inicial muito menor, a convergência se deu em menos iterações.

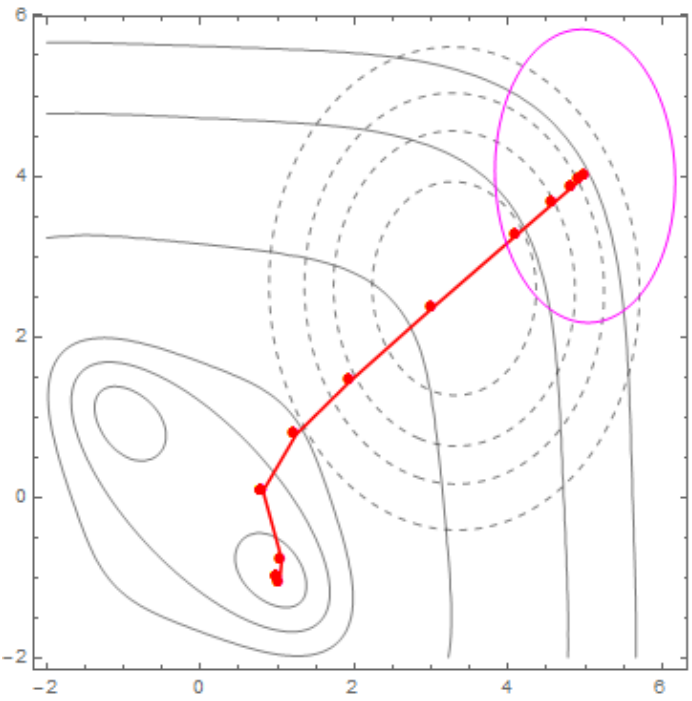


Figura 2.4: *Curvas de nível da função (traço contínuo), do modelo (tracejado), a região de confiança da primeira iteração e a trajetória dos pontos até o minimizador local da função.*

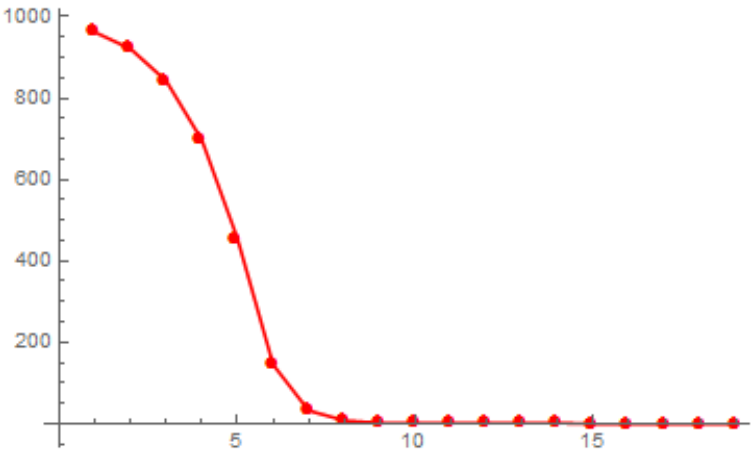


Figura 2.5: *Gráfico do valor da função objetivo a cada iteração, num total de 19 iterações.*

Tabela 2.2: *Registro das iterações do método na função Quártica*

Iteração	***Ponto***	***Valor funcional***	***Raio***
0	(5.000000, 4.000000)	962	—
1	(4.944014, 3.953861)	921.056623	Aumentou
2	(4.830794, 3.860522)	842.311185	Aumentou
3	(4.599139, 3.669390)	697.205201	Aumentou
4	(4.112938, 3.267494)	454.903776	Aumentou
5	(3.028376, 2.366292)	145.125018	Aumentou
6	(1.965834, 1.460500)	31.968724	Inalterado
7	(1.243331, 0.779372)	7.634745	Inalterado
8	(0.813060, 0.073399)	1.675749	Inalterado
9	(0.813060, 0.073399)	1.675749	Reduziu
10	(0.813060, 0.073399)	1.675749	Reduziu
11	(0.813060, 0.073399)	1.675749	Reduziu
12	(0.813060, 0.073399)	1.675749	Reduziu
13	(0.813060, 0.073399)	1.675749	Reduziu
14	(1.057105, -0.783565)	-0.687537	Aumentou
15	(1.026853, -1.079866)	-0.963824	Inalterado
16	(1.002767, -1.006552)	-0.999768	Inalterado
17	(1.000024, -1.000051)	-1.000000	Inalterado
18	(1.000000, -1.000000)	-1.000000	Inalterado

2.5.3 Função Trigonométrica

Também testamos uma função altamente não linear, $f(x) = (x_1 - \cos x_2)^2 + (-x_2 + \sin x_1)^2$, gerada a partir do ponto inicial $x_0 = (-3,\ 6.5)$ e com $\Delta_0 = \dfrac{\|\nabla f(x_0)\|_2}{5}$.

E chamamos atenção para a região de confiança apresentada na Figura 2.6, pois esse foi um caso em que as direções de maior deformação do modelo m_0 não foram as direções canônicas.

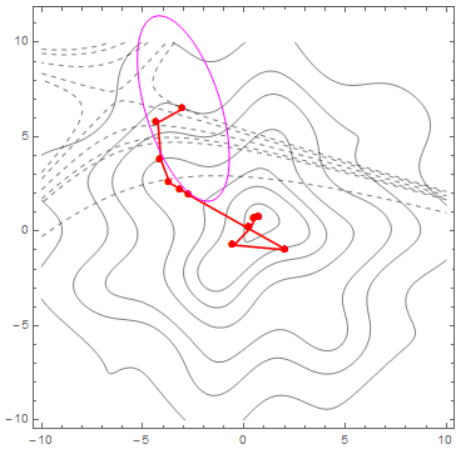


Figura 2.6: *Gráfico do valor da função objetivo a cada iteração, num total de 20 iterações.*

Como podemos observar na Figura 2.7, a alta não linearidade fez com que o método tivesse algumas iterações com mesmo valor de função; e se verificarmos a Tabela 2.3 veremos que nesses casos tivemos sucessivas reduções do raio da região de confiança.

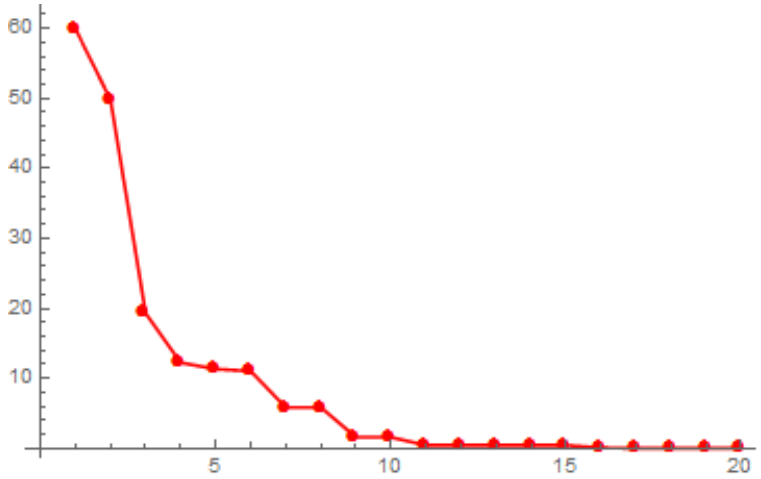


Figura 2.7: *Curvas de nível da função (traço contínuo), do modelo (tracejado), a região de confiança da primeira iteração e a trajetória dos pontos até o minimizador local da função.*

Tabela 2.3: *Registro das iterações do método na função Trigonométrica*

Iteração k	***Ponto*** x_k	***Valor funcional*** $f(x_k)$	***Status do Raio***
0	(-3.000000, 6.500000)	59.917700	——
1	(-4.257673, 5.752799)	49.782638	Aumentou
2	(-4.100128, 3.746798)	19.319405	Inalterado
3	(-3.670372, 2.553604)	12.254937	Inalterado
4	(-3.108995, 2.207262)	11.340350	Inalterado
5	(-2.658703, 1.879453)	11.039034	Inalterado
6	(2.104525, -0.993246)	5.867001	Inalterado
7	(2.104525, -0.993246)	5.867001	Reduziu
8	(-0.496477, -0.736361)	1.598760	Inalterado
9	(-0.496477, -0.736361)	1.598760	Reduziu
10	(0.322733, 0.176351)	0.457750	Aumentou
11	(0.322733, 0.176351)	0.457750	Reduziu
12	(0.322733, 0.176351)	0.457750	Reduziu
13	(0.322733, 0.176351)	0.457750	Reduziu
14	(0.322733, 0.176351)	0.457750	Reduziu
15	(0.530245, 0.636600)	0.092132	Aumentou
16	(0.699045, 0.697643)	0.007464	Aumentou
17	(0.766004, 0.695678)	0.000008	Inalterado
18	(0.768166, 0.694820)	0.000000	Inalterado
19	(0.768169, 0.694820)	0.000000	Inalterado

2.6 Discussão e Conclusão

É importante destacar que fizemos uma abordagem geométrica do método e nesse caso trabalhos com problemas de baixa dimensão, então foi necessário fazermos alguns ajustes no código fornecido pelos autores, uma vez que este foi originalmente implementado para problemas de grande porte. Ou seja, além da implementação da rotina para resolver o problema do tipo (PI) houve também um trabalho de aprimoramento do código original para que fosse possível resolver o subproblema (2.8).

O primeiro ajuste realizado eliminou um teste obsoleto, pois no nosso caso a matriz B sempre era uma matriz densa então as primeiras linhas do código onde testava-se a esparsidade da B pra construir o leque $\tilde{M}(\lambda)$ era desnecessário.

O segundo ajuste, trata-se do mais importante, pois nele trocamos a função `eigs` por `eig`. Essa alteração foi feita em dois trechos do código: o primeiro quando estamos no caso fácil e o segundo quando estamos no caso difícil. E esse ajuste se faz necessário, porque para que a função `eigs` consiga calcular os autovalores generalizados ela exige que as matrizes do leque sejam grandes e esparsas, o que não é o nosso caso. Os autores de [1] comentam no artigo que para os casos grandes e esparsos a melhor rotina a ser utilizada é o `eigs`.

Um ponto importante a ser comentado é que apesar de estarmos extraindo o máximo de informações possíveis do modelo para construirmos a matriz B, essa estratégia só foi adotada porque estamos trabalhando com duas variáveis, pois para os problemas de grande porte a construção da matriz B pode ficar muito custosa. Os autores não se ativeram à construção da matriz B no artigo deles, mas eles também comentam que essa é uma estratégia possível.

Apêndice A

Códigos

A.1 Rotina para o problema de minimização

```
function[k, x, raio]=TRS(x0, eta)

%escolhe uma funcao do portfolio pra ser minimizada
funcao=escolhe();


[f0, g0, h0]=feval(funcao, x0);
aux=norm(g0);
Delta=aux/100;
contador=0;

k=0;
while aux>10^(-6)
% declaramos a matriz definida positiva B:
[U, T]=eig(h0);
D=abs(T);

for i=1:2
    if D(i,i)==0
       D(i,i)=1;
    end
end

B=U*D*U';

%aqui usamos o TRSgepM para encontrar o passo p
[p, ~,contador]=TRSgepM(h0,g0,B,Delta,contador);

xt=x0+p;

%construimos o modelo em x0:
[~, mk]=modelo(x0,xt,f0,g0,h0);

[fk, gk, hk]=feval(funcao, xt);

ared=f0-fk;
```

```
pred=f0-mk;

rho=ared/pred;

if rho>eta % sucesso: o ponto vai ser atualizado
    x0=xt;
    f0=fk;
    g0=gk;
    h0=hk;
end


if rho<0.25 % Reduzir o raio
    Delta=Delta*0.5;

else %Ampliar o raio
    if (rho>0.75 && (abs((p'*B*p)-Delta^2)<10^(-4)))
        Delta=2*Delta;
    end
end

aux=norm(g0);
k=k+1;
end
x=x0;
raio=Delta;
```

A.2 Rotina para o subproblema

```
function [x,lam1,contadordehardcase] = TRS(A,a,B,Del,contadordehardcase)
% Solves the trust-region subproblem by a generalized eigenproblem without
% iterations
%
% minimize (x^TAx)/2+ ax
% subject to x^TBx <= Del^2
%
% A: nxn symmetric, a: nx1 vector
% B: nxn symmetric positive definite
%
% Yuji Nakatsukasa, 2015

n = size(A,1);

MM1 = [zeros(n) B;B zeros(n)];
tolhardcase = 1e-4; % tolerancia para o caso dificil

p1 = pcg(A,-a,1e-12); % possivel solucao interior
if norm(A*p1+a)/norm(a)<1e-5
    if p1'*B*p1>=Del^2
        p1 = nan;
    end
else
    p1 = nan;
end
```

```
26
27 % Esta eh a alma do codigo
28
29 MM0=[-B A; A -a*a'/Del^2]; %construimos MM0
30 [Q,D] = eig(MM0,-MM1); %encontramos seus autovalores generalizados
31 d=diag(D); %vetorizamos a diagonal com os autovalores
32 [o, seq]=sort(d,'descend'); %ordenamos em ordem decrescente
33 lam1=max(o); %selecionamos o maior autovalor
34 V=Q(:,seq(1)); %selecionamos o autovetor associado ao maior autovalor
35
36 if norm(real(V)) < 1e-3 % as vezes complexo
37     V = imag(V);
38 else
39     V = real(V);
40 end
41
42 lam1 = real(lam1);
43 x = V(1:length(A)); % esta eh paralela a solucao
44 normx = sqrt(x'*(B*x));
45 x = x/normx*Del; % no caso facil, esta simples normalizacao melhora precisao
46 if x'*a>0 % toma o sinal correto
47     x = -x;
48 end
49
50 if normx < tolhardcase % caso dificil
51     contadordehardcase=contadordehardcase+1;
52     disp(['hard case!',num2str(normx)])
53     x1 = V(length(A)+1:end);
54     alpha1 = lam1;
55     Pvect = x1;  %primeiro tente k=1, quase sempre o suficiente
56     x2 = pcg(@(x)pcgforAtilde(A,B,lam1,Pvect,alpha1,x),-a,1e-12,500);
57     if norm((A+lam1*B)*x2+a)/norm(a)>tolhardcase % residuo muito grande,
           repetir
58         [Pvect,~] = eig(A,B);
59         x2 = pcg(@(x)pcgforAtilde(A,B,lam1,Pvect,alpha1,x),-a,1e-8,500);
60         if norm((A+lam1*B)*x2+a)/norm(a) < tolhardcase
61             return
62         end
63     end
64
65     Bx = B*x1; Bx2 = B*x2; aa = x1'*(Bx); bb = 2*x2'*Bx; cc = (x2'*Bx2-Del^2);
66     alp = (-bb+sqrt(bb^2-4*aa*cc))/(2*aa); %norm(x2+alp*x)-Delta
67     x = x2+alp*x1;
68     disp(contadordehardcase)
69 end
70
71 % escolhe entre uma solucao interior ou na fronteira
72
73 if sum(isnan(p1))==0
74     if (p1'*A*p1)/2+a'*p1 < (x'*A*x)/2+a'*x
75         x = p1; lam1 = 0;
76     end
77 end
78 end
```

A.3 Rotinas auxiliares

```
function[funcao]=escolhe()

disp('Escolha a funcao desejada:')
disp('1 = Funcao de Rosenbrock: 100(y-x^2)^2+(1-x)^2')
disp('2 = Funcao Quartica:      x^4-y^4-2x^2+2y^2+1/16')
disp('3 = Funcao Quartica:      x^4-2x^2+y^2-17/16')
disp('4 = Funcao Quartica:      x^4+y^4+4xy+1')

n=input('');

switch n
    case 1
        funcao='rosenbrock';
    case 2
        funcao='quartica1';
    case 3
        funcao='quartica2';
    case 4
        funcao='quartica3';
end

function [y] = MM0timesx(A,B,g,Delta,x)
% MM0 = [-B A;
%          A -g*g'/Delta^2];
n = size(A,1);
x1 = x(1:n); x2 = x(n+1:end);
y1 = -B*x1 + A*x2;
y2 = A*x1-g*(g'*x2)/Delta^2;
y = [y1;y2];
end


function [y] = pcgforAtilde(A,B,lamA,Pvect,alpha1,x)

[n,m] = size(Pvect);
y = A*x+lamA*(B*x);

for i=1:m
    y = y+(alpha1*(x'*(B*Pvect(:,i))))*(B*Pvect(:,i));
end
end

function[m0, mk]=modelo(x0, x, f0, g0, h0)

mk=f0+g0'*[x(1)-x0(1); x(2)-x0(2)]+1/2*[x(1)-x0(1),x(2)-x0(2)]*h0*[x(1)-x0(1);
   x(2)-x0(2)];

m0=f0;
```

Referências Bibliográficas

[1] S. Adachi, S. Iwata, Y. Nakatsukasa, and A. Takeda. Solving the Trust-Region Subproblem By a Generalized Eigenvalue Problem. *SIAM Journal on Optimization*, 27(1):269–291, 2017. doi: https://doi.org/10.1137/16M1058200.

[2] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust region methods*. SIAM, Philadelphia, 2000.

[3] A. Friedlander. *Elementos de Programação Não Linear*. Editora da Unicamp, Campinas, SP, 1994.

[4] W. Gander, G. H. Golub, and U. von Matt. A Constrained Eigenvalue Problem. *Linear Algebra and its Applications*, 114:815–839, 1989. doi: https://doi.org/10.1016/0024-3795(89)90494-1.

[5] E. M. Gertz. A quasi-Newton trust-region method. *Mathematical Programming*, 100(3):447–470, 2004. doi: https://doi.org/10.1007/s10107-004-0511-1.

[6] G. H. Golub and C. F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, Maryland, 4th edition, 2013.

[7] K. D. Ikramov. Matrix pencils: Theory, applications, and numerical methods. *Journal of Soviet Mathematics*, 64(2):783–853, 1993. doi: https://doi.org/10.1007/BF01098963.

[8] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, New York, second edition, 2006.

[9] A. A. Ribeiro and E. W. Karas. *Otimização Contínua: Aspectos Teóricos e Computacionais*. Cengage Learning, São Paulo, 2014.

[10] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983. doi: https://doi.org/10.1137/0720042.