Michael da Silva Bressiani

Método de Levenberg-Marquardt com correção de segunda ordem

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da "CAPES"

Santo André, 2020

Universidade Federal do ABC

Centro de Matemática, Computação e Cognição

Michael da Silva Bressiani

Método de Levenberg-Marquardt com correção de segunda ordem

Orientadora: Profa. Dra. Erika Alejandra Rada Mora

Coorientador: Prof. Dr. Fedor Pisnitchenko

Dissertação de mestrado apresentada ao Centro de Matemática, Computação e Cognição para obtenção do título de Mestre em Matemática

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Michael da Silva Bressiani, e orientada pela Profa. Dra. Erika Alejandra Rada Mora.

Santo André, 2020

Sistema de Bibliotecas da Universidade Federal do ABC
Elaborada pelo Sistema de Geração de Ficha Catalográfica da UFABC
com os dados fornecidos pelo(a) autor(a).

Silva Bressiani, Michael da
Método de Levenberg-Marquardt com correção de segunda ordem / Michael da Silva Bressiani. — 2020.

92 fls. : il.

Orientadora: Erika Alejandra Rada Mora
Coorientador: Fedor Pisnitchenko

Dissertação (Mestrado) — Universidade Federal do ABC, Programa de Pós Graduação em Matemática, Santo André, 2020.

1. Correção de Segunda Ordem. 2. Método de Levenberg-Marquardt. 3. Quadrados Mínimos Não Lineares. I. Rada Mora, Erika Alejandra. II. Pisnitchenko, Fedor. III. Programa de Pós-Graduação em Matemática, 2020. IV. Título.

Este exemplar foi revisado e alterado em relação à versão original, de acordo com as observações levantadas pela banca no dia da defesa, sob responsabilidade única do(a) autor(a) e com a anuência do(a) orientador(a).

Santo André/SP ☑ 17 de abril de 2020

Assinatura do(a) autor(a): Michael S. Bressiani

Assinatura do(a) orientador(a):

MINISTÉRIO DA EDUCAÇÃO
Fundação Universidade Federal do ABC
Avenida dos Estados, 5001 – Bairro Santa Terezinha – Santo André – SP
CEP 09210-580 · Fone: (11) 4996-0017

FOLHA DE ASSINATURAS

Assinaturas dos membros da Banca Examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato, MICHAEL DA SILVA BRESSIANI realizada em 17 de Abril de 2020:

Prof.(a) LUIS FELIPE CESAR DA ROCHA BUENO
UNIVERSIDADE FEDERAL DE SÃO PAULO

Prof.(a) MAJID FORGHANI ELAHABAD
UNIVERSIDADE FEDERAL DO ABC

Prof.(a) CRISTIAN FAVIO COLETTI
UNIVERSIDADE FEDERAL DO ABC

Prof.(a) JEFERSON CASSIANO
UNIVERSIDADE FEDERAL DO ABC

Prof.(a) FEDOR PISNITCHENKO
UNIVERSIDADE FEDERAL DO ABC

* Por ausência do membro titular, foi substituído pelo membro suplente descrito acima: nome completo, instituição e assinatura

DEDICATÓRIA

Dedico essa tese aos meus familiares, amigos e a Kira.

AGRADECIMENTOS

Agradeço à meus pais, pelo apoio e incentivo, que tornaram possível a realização deste trabalho.

À Erika Alejandra Rada Mora, pela orientação e ajuda nos momentos em que precisei.

Ao Feodor Pisnitchenko, pela coorientação, pelas reuniões, sugestões, questionamentos, ideias e pelas diversas vezes em que me ajudou.

À Momoe Sakamori Pisnitchenko, pelas sugestões e correções deste trabalho.

À Raiane, pelo apoio e companheirismo em todos esses anos, estando sempre presente nos momentos em que mais precisei.

Aos meus amigos, em especial Denis, Diego, Lucas e Tone que estiveram presentes desde o inicio deste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Ohana means family.
Family means nobody gets left behind, or forgotten.
— Lilo & Stitch

RESUMO

Em ciências aplicadas e Computação, existem diversos problemas que consistem em ajustar um modelo teoricamente formulado a um conjunto de dados obtidos por meio de um experimento ou simulação numérica. Esse ajuste consiste em determinar um conjunto finito de parâmetro para os quais o modelo melhor se encaixa no conjunto de dados obtidos previamente.

Uma forma natural de resolver esse tipo de problema é interpreta-lo como um problema de otimização, ou seja, minimizar alguma função objetivo escolhida como a soma dos quadrados das distancias entre modelo e os dados observados.

Neste trabalho propomos um método para encontrar mínimos de tais problemas. De modo geral o método consiste em combinar técnicas de tradicionais do método de Levenberg-Marquardt em conjunto com um vetor de correção que contém informações de segunda ordem dos resíduos. Também propomos três modificações desse método as quais visam melhorar sua eficiência.

Fizemos a teoria do método, obtendo dois teoremas de convergência além de também realizar teste numéricos. Os resultados parecem promissores, diminuindo significativamente o número de iteração em alguns casos.

Palavras-chave: Correção de Segunda Ordem, Método de Levenberg-Marquardt, Quadrados Mínimos Não Lineares

ABSTRACT

In applied and computer science, there are several problems that consist in adjusting a theoretically formulated model to a set of data obtained through an experiment or numerical simulation. This adjustment consists of determining a finite set of parameters for which the model best fits the set of previously obtained data.

A natural way to solve this type of problem is to interpret it as an optimization problem, that is, to minimize any objective function chosen as the sum of the squares of the distances between the model and the observed data.

In this work, we propose a method to find minimums of such problems. In general, the method consists of combining traditional techniques of the Levenberg-Marquardt method together with a correction vector that contains second order information on the residues. We also propose three modifications to this method which aim to improve its efficiency.

We made the theory of the method, obtaining two convergence theorems in addition to also performing numerical tests. The results look promising, significantly reducing the number of iterations in some cases.

Keywords: Levenberg-Marquardt Method, Nonlinear Least Squares, Second Order Correction

CONTEÚDO

1 INTRODUÇÃO

Problemas de quadrados mínimos são tema de grande interesse no campo da otimização irrestrita. De fato, diversas questões em ciências aplicadas e computação consistem em ajustar um modelo, teoricamente formulado, a um conjunto de dados que provêm de alguma simulação numérica ou experimento. Como exemplos, podemos citar o cálculo da matriz fundamental em visão computacional [4], retificação de imagens [28], reconhecimento facial [5], calibração de câmeras [3], assim como diversas aplicações em física e química [11], [38]. Além disso, mais recentemente, houve um aumento no uso de técnicas de otimização não linear para treinar redes neurais, em especial, o uso de técnicas clássicas de mínimos quadrados, como o método de Levenberg-Marquardt, [2].

De modo geral, o objetivo desses problemas consiste em minimizar a discrepância entre um modelo teórico e as observações experimentais. Esses problemas podem ser formulados matematicamente dentro do campo da otimização irrestrita, onde deseja-se minimizar uma função da forma

$$f(x) = \frac{1}{2}\sum_{i=1}^{m} r_i(x)^2,$$

onde resíduo r_i representa a diferença entre y_i associado a i-ésima observação (x_i, y_i), e o valor do modelo em x_i. Essencialmente, queremos associar uma curva a um conjunto de pontos observados, diminuindo a discordância entre o modelo e os dados experimentais, conforme é ilustrado na Figura 1.

Podemos classificar os problemas de quadrados mínimos em dois tipos, os lineares e os não lineares. No caso linear, o problema consiste em resolver um único sistema linear. Já no caso não linear a solução é mais complexa, sendo necessário um estudo mais detalhado.

Em 1944, Levenberg [27] propôs um método para solucionar o caso não linear dos problemas de quadrados mínimos. Porém em sua proposta não obteve bons resultados

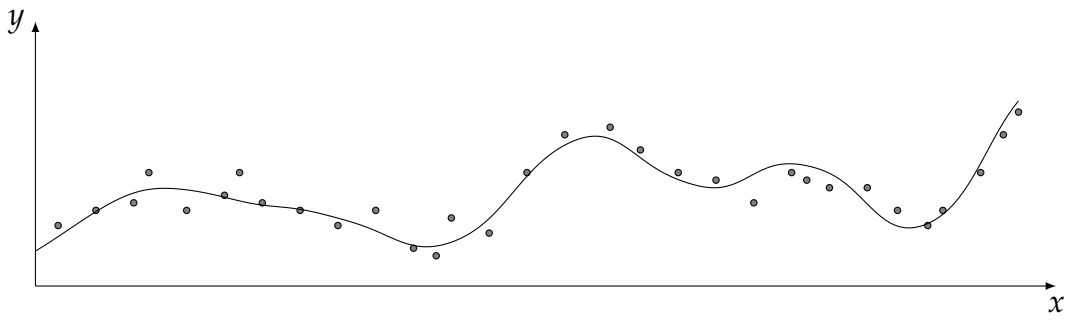

Figura 1: Modelo (curva suave) e as medidas observadas (pontos).

com o problemas mal condicionados, além de necessitar de estudos teóricos mais acurados. Então, em 1963, Marquardt [31] realizou algumas fundamentações teóricas e propôs uma modificação do método a fim de aprimorá-lo, passando a ser chamado de método de Levenberg-Marquardt, ou simplesmente LM.

Nos dias atuais, o algoritmo de LM é amplamente difundido. Este deve ao fato de que há diversas implementações numéricas disponíveis [1], [29], [33], [16], [42], além da simplicidade de sua utilização. Porém, existem alternativas ao método de LM, como por exemplo os Métodos clássicos de Regiões de Confiança aplicados ao problema de mínimos quadrados, tais como Dogleg, Minimização no Subespaço Bidimensional, Método de Newton, Método Quasi-Newton, ou mesmo estratégias de busca linear [36].

Além das boas propriedades numéricas que ele possui, o algoritmo de LM é considerado o precursor dos métodos de regiões de confiança, dando início a toda uma área de estudo, o que enfatiza a sua importância.

De modo geral, o funcionamento do método de LM depende basicamente de dois parâmetros, os quais precisam ser atualizados a cada iteração. Um deles, o parâmetro λ, denominado *damping* ou parâmetro de Levenberg-Marquardt, é razão de intenso estudo na literatura [6], [9], [15], [18], [23], [27], [31], [32], [34], dada a grande influência que este exerce sobre o desempenho do método. Já o outro, é a matriz D, que está associada ao condicionamento da matriz do sistema de LM, sendo fundamental para a robustez do processo.

Em geral, D é escolhida como sendo diagonal positiva definida. Existem diversas formas de construí-la [17], [31], [32], [22], e sua importância reside no fato que existe um grande número de problemas de quadrados mínimos mal condicionados, que é alvo de constante estudo [7], [10], [12], [40], para os quais o algoritmo proposto por Levenberg [27] costuma falhar.

No entanto, mesmo com os numerosos esforços em combinar boas escolhas de λ e D, o método de LM pode encontrar dificuldades de convergência. Por conta disso, técnicas recentes propõem combinar o passo de LM com uma outra direção, obtida sem muito custo adicional, buscando diminuir o número de passos feitos pelo algoritmo [13], [14], [41].

Seguindo essa ideia, propomos uma nova modificação do método de LM que não consta na literatura, que consiste em obter uma correção para método de LM, na qual faremos uso das derivadas de segunda ordem dos resíduos combinado com técnicas tradicionais do método de LM.

Nossa principal contribuição com este trabalho é mostrar, por meio de experimentos numéricos e de teoremas de convergência global, que existe uma maneira viável de utilizar informações referentes à derivada de segunda ordem, a fim de propiciar um melhor desempenho ao método de LM, especialmente nos casos em que a aproximação linear, convencionalmente feita, não apresenta uma boa aproximação local para a função objetivo.

Como uma opção numericamente viável para o cálculo das derivadas de segunda ordem dos resíduos, podemos citar a diferenciação automática, técnica que pode ser explorado a fim de se obter o valor da função, a derivada e a derivada de segunda ordem em um mesmo ponto, simultaneamente, sem muito custo adicional, além de exigir do usuário apenas o fornecimento da função objetivo, dispensando preocupações com as complexidades envolvidas de se fornecer as derivadas necessárias. No escopo dos números duais, também introduzimos uma contribuição, mostrando como obter a hessiana aplicada em uma única direção, simultaneamente a ao valor da função e sua primeira derivada aplicada nessa mesma direção.

No segundo capítulo, apresentamos uma introdução aos problemas de quadrados mínimos abrangendo os métodos de Newton e Gauss-Newton, e como eles inspiram a criação do método de LM. Além disso é exposto o método proposto por Jinyan Fan [13], [14], em que, com base em um Método de Newton Modificado, que possui convergência cúbica [20], [43], é elaborado uma modificação do método de LM, que também possui convergência cúbica. Ainda no segundo capítulo, mostramos como o método de LM pode ser visto como um método de regiões de confiança, tal abordagem pode ser também consultada em [32]. Por fim, como métodos alternativos de regiões de confiança, apresentamos o Ponto de Cauchy e o Método de Dogleg.

No terceiro capítulo apresentamos a proposta do nosso trabalho, estruturada em duas seções. Na primeira explicitamos a construção do método e, na segunda, demonstramos os teoremas de convergência. Como uma aplicação, mostramos que a nossa proposta, quando o parâmetro de Levenberg-Marquardt é zero, converge para o mínimo global da função de Rosenbrock em apenas um passo, independente do ponto inicial escolhido para inicializar o método.

No quarto capítulo, na primeira seção, expomos uma maneira eficiente de realizar a decomposição QR da matriz do sistema de LM, ao explorar a sua estrutura especial. Na segunda seção, exibimos a teoria dos números duais, e explicitamos o fato de como eles podem ser usados para obtermos as derivadas de segunda ordem sem muito custo adicional.

Por fim, na última seção, realizamos os testes numéricos, utilizando o banco de problemas NIST Standard Reference Database 140, [35], e comparamos o número de iterações do método de LM com aqueles obtidos com a nossa proposta e algumas modificações dela. É importante ressaltar que nossos testes numéricos são preliminares, tendo como principal objetivo mostrar que existe uma maneira numericamente viável de usar informações de segunda ordem, juntamente como o método de Levenberg-Marquardt, de maneira eficiente.

2 PROBLEMAS DE QUADRADOS MÍNIMOS NÃO LINEARES

Este capítulo se inicia com algumas definições essenciais para o desenvolvimento do nosso trabalho.

Em seguida, apresentamos a definição formal do problema de quadrados mínimos, isto é, definimos a função a qual desejamos minimizar e calculamos seu gradiente e hessiana, que são fundamentais para o desenvolvimento de alguns métodos, como será visto mais adiante. E mais adiante, apresentamos os principais métodos clássicos para resolvê-los. Os conteúdos presente neste capítulo podem ser consultados nas referências [13], [14], [27], [30], [31], [32], [36].

2.1 DEFINIÇÕES

Problemas de quadrados mínimos são aqueles no qual se deseja encontrar um mínimo para funções $F : \mathbb{R}^n \to \mathbb{R}$ da forma:

$$F(x) = \frac{1}{2}\sum_{i=1}^{m} r_i(x)^2, \tag{1}$$

onde assumimos que $m \geq n$ e $r_i : \mathbb{R}^n \to \mathbb{R}$ é uma função de classe C^2 para todo $i \in \{1, \ldots, m\}$. As funções $r_1, \ldots, r_m$ são chamadas de resíduos.

Usando a norma euclidiana, podemos escrever F da seguinte forma

$$F(x) = \frac{1}{2}\|r(x)\|^2, \quad \forall x \in \mathbb{R}^n, \tag{2}$$

onde $r : \mathbb{R}^n \to \mathbb{R}^m$ é dada por $r(x) = (r_1(x), \ldots, r_m(x))^T$, $\forall x \in \mathbb{R}^n$. A função r é chamada de vetor de resíduos.

Salvo menção do contrário, $\|\cdot\|$ denotará a norma euclideana.

Um fato notório é que na equação (1), pode ter tanto seu gradiente quanto sua hessiana escritos em termos das derivadas parciais dos resíduos, como mostraremos a seguir.

Fixando $x \in \mathbb{R}^n$, o jacobiano $J(x)$ de $r(x)$ em x é dado por

$$J(x) = \begin{bmatrix} \frac{\partial r_1(x)}{\partial x_1} & \dots & \frac{\partial r_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_m(x)}{\partial x_1} & \dots & \frac{\partial r_m(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla r_1(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix}, \tag{3}$$

onde para cada $i \in \{1, \dots, m\}$, o vetor $\nabla r_i(x)$ é o gradiente do resíduo r_i aplicado em x, ou seja, o elemento da linha i e coluna j da matriz $J(x)$ é dado por $\frac{\partial r_i(x)}{\partial x_j}$, para quaisquer $i \in \{1, \dots, m\}$ e $j \in \{1, \dots, n\}$ escolhidos.

Assim, o gradiente de F é dado por

$$\nabla F(x) = \sum_{i=1}^{m} r_i(x) \nabla r_i(x) = J(x)^T r(x), \tag{4}$$

e a hessiana por

$$\begin{aligned} \nabla^2 F(x) &= \sum_{i=1}^{m} \nabla r_i(x)^T \nabla r_i(x) + \sum_{i=1}^{m} r_i(x) \nabla^2 r_i(x) && (5) \\ &= J(x)^T J(x) + \sum_{i=1}^{m} r_i(x) \nabla^2 r_i(x). && (6) \end{aligned}$$

Estes resultados serão amplamente explorados no decorrer deste trabalho. Além disso, vamos convencionar que F será usado para denotar, de modo geral, funções que são da forma (1), enquanto que usaremos f para denotar funções que não são necessariamente dessa forma.

2.2 MÉTODO DE NEWTON

Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função de classe C^2. Fazendo a expansão de Taylor de f em torno de um ponto x e uma direção $p \in \mathbb{R}^n$ temos

$$f(x+p) = f(x) + p^T \nabla f(x) + \frac{1}{2} p^T \nabla^2 f(x) p + \mathcal{O}(\|p\|^3). \tag{7}$$

Tendo em vista a equação (7), podemos intuir a seguinte aproximação quadrática para F em torno de x,

$$\Omega(p) = f(x) + p^T \nabla f(x) + \frac{1}{2} p^T \nabla^2 f(x) p. \tag{8}$$

A ideia do método de Newton é resolver, em cada iteração, um subproblema da forma

$$\min_{p \in \mathbb{R}^n} \Omega(p). \tag{9}$$

Para resolvermos este subproblema, calculamos o seu gradiente e o igualamos a zero o que resulta no seguinte sistema

$$\nabla^2 f(x) p_N = -\nabla f(x). \tag{10}$$

Logo, quando $\nabla^2 f(x)$ é inversível, chamamos p_N de passo de Newton.

Quando p_N está bem definido, ou seja, $\nabla^2 f(x)$ é não singular, o método de Newton consiste em simplesmente efetuar o passo na direção p_N como indica o algoritmo a seguir.

Algoritmo 1: MÉTODO DE NEWTON

Entrada: Dados x_0

Passo 1 : Resolva $\nabla^2 f(x_k) p_N = -\nabla f(x_k)$

Passo 2 : $x_{k+1} = x_k + p_N$

Passo 3 : Faça $k = k + 1$ e retorne ao **Passo 1**

Se supormos que x^* é ponto crítico de f, com $\nabla^2 f(x^*)$ positiva definida e f de classe C^2 com $\nabla^2 f(x)$ Lipschitz em uma vizinhança de x^*, o método de Newton converge quadraticamente para x^* se tomarmos x_0 suficientemente próximo de x^*, [36].

Apesar da convergência quadrática, o método de Newton tem algumas falhas. De modo geral, não é sempre possível garantir que $\nabla^2 f(x_k)$ seja inversível para todo k. Além disso mesmo no caso em que p_N está bem definido, pode ocorrer deste não ser uma direção de descida para a função f. Portanto, diante de tais falhas, se faz necessário buscar novas abordagens.

Uma abordagem é utilizar busca linear na direção de p_N quando esta é de descida, ou na direção de $-p_N$ quando p_N é de subida. No entanto, no caso em que $\nabla f(x)^T p_N \approx 0$, a busca linear pode se tornar ineficiente.

Contudo, a função dada na equação (1) detém uma estrutura especial, a qual exploramos para realizar a seguinte aproximação

$$\nabla^2 F(x) \approx J(x)^T J(x).$$

Dessa forma, estamos contornando a necessidade de se obter as derivadas de segunda ordem dos resíduos, além de aproveitar o fato de que $J(x)^T J(x)$ é sempre positiva semi definida, o que garante que a aproximação quadrática será sempre ao menos uma função convexa.

No entanto, é evidente que no caso em que essas derivadas forem significativas em relação a $J(x)^T J(x)$, essa aproximação pode apresentar baixa confiabilidade. Existem maneiras de se explorar o uso das derivadas de segunda ordem dos resíduos como mostraremos com nossa proposta no Capítulo 3.

Algumas das vantagens dessa aproximação são a facilidade do seu cálculo e o baixo custo, pois não há o cálculo da hessiana de cada um dos resíduos. Uma vez obtido $J(x)$, calculamos tanto o gradiente de F quanto a multiplicação $J(x)^T J(x)$. Além disso, se $J(x)$ possui posto coluna completo, $J(x)^T J(x)$ é positiva definida. Portanto, se considerarmos tal aproximação, a formulação do método de Newton para problemas de mínimos quadrados nos fornece o seguinte passo

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

Observe que, nesse caso, p_g é direção de descida para F. Tal aproximação nos fornece um método Quase-Newton, que será descrito com mais detalhes na próxima subseção.

2.3 MÉTODO DE GAUSS-NEWTON

O método de Gauss-Newton pode ser intuído como feito no final da seção anterior, isto é, via aproximação $\nabla^2 F(x) \approx J(x)^T J(x)$. No entanto, também é possível obtê-lo via teorema de Taylor, como faremos a seguir.

Dados $x, p \in \mathbb{R}^n$, considere a aproximação de primeira ordem em torno de x dada pela expansão de Taylor de r

$$\begin{aligned} F(x+p) &= \frac{1}{2}\|r(x+p)\|^2 \\ &\approx \frac{1}{2}\|r(x)+J(x)p\|^2 \\ &= \frac{1}{2}\|r(x)\|^2 + p^T J(x)^T r(x) + \frac{1}{2}p^T J(x)^T J(x)p \\ &= f(x) + p^T \nabla f(x) + \frac{1}{2}p^T J(x)^T J(x)p, \end{aligned}$$

que nos fornece uma aproximação quadrática em torno de x. Note que essa aproximação é parecida com a expansão de Taylor de segunda ordem de F na direção de p em torno de x, faltando apenas a segunda parte da soma apresentada na equação (6). Assim sendo, temos $J(x)^T J(x)$ como uma aproximação natural para a hessiana de F, o que nos motiva a definir a seguinte função

$$\Pi(p) = \frac{1}{2}\|r(x)\|^2 + p^T J(x)^T r(x) + \frac{1}{2}p^T J(x)^T J(x)p, \tag{11}$$

onde $p \in \mathbb{R}^n$.

De modo geral, o método consiste em construir uma sequência $(x_k)_{k\geq 0}$ em $\mathbb{R}^n$, onde x_0 é dado, e o restante dos elementos são dados pela regra $x_{k+1} = x_k + \alpha_k p_k$, $\forall k \geq 0$, onde α_k é escolhido de maneira conveniente, e p_k é solução do seguinte problema

$$\min_p \Pi_k(p) = \frac{1}{2}\|r(x_k)\|^2 + p^T J(x_k)^T r(x_k) + \frac{1}{2}p^T J(x_k)^T J(x_k)p. \tag{12}$$

Afim de obter p_k, derivamos a equação (11) em relação a p e igualamos a zero. Dessa forma, obtemos

$$J(x_k)^T J(x_k)p_k = -J(x_k)^T r(x_k) = -\nabla f(x_k). \tag{13}$$

Observe que, se $J(x_k)^T J(x_k)$ for inversível, então a solução do sistema acima está sempre bem definida, e vamos denotá-la por p_{gn}. Nesse caso a função quadrática definida na equação (11) é estritamente convexa, pois o fato de $J(x_k)^T J(x_k)$ ser inversível implica que também é positiva definida. Assim, o ponto crítico p_{gn} é o único minimizador global do problema (12). Nesse caso, a solução encontrada é chamada de passo de Gauss-Newton (GN).

Note que, se $\nabla F(x_k) = 0$, então $p_{gn} = 0$. Logo, a partir de k, todos os ponto da sequência serão iguais à x_k, ou seja, o método estagna sempre que encontra um ponto crítico.

Por outro lado, se $\nabla F(x_k) \neq 0$ e supondo que $J(x_k)^T J(x_k)$ é inversível, então temos $p_{gn} \neq 0$. Logo

$$p_g^T \nabla F(x_k) = -p_{gn}^T J(x_k)^T J(x_k) p_{gn} < 0.$$

Usamos o fato de que se $J(x_k)^T J(x_k)$ é positiva definida, p_{gn} está bem definida e é uma direção de descida. Desse modo, podemos implementar o método, chamado de Gauss-Newton, utilizando busca linear na direção de p_{gn}, isto é, escolher α_k utilizando estratégias de busca linear afim de obter uma melhor performance do método, isto é, a fim de obter um decréscimo suficiente no valor de F.

O procedimento a seguir é uma formulação do método de Gauss-Newton fazendo uso de busca linear na direção do passo de GN.

Algoritmo 2: MÉTODO DE GAUSS-NEWTON

Entrada: Dado um ponto inicial x_0
Passo 1 : Resolva $J(x_k)^T J(x_k) p_{gn} = -J(x_k) r(x_k)$
Passo 2 : Faça busca linear na direção de p_{gn} encontrando assim $\alpha_k > 0$
Passo 3 : $x_{k+1} = x_k + \alpha_k p_{gn}$
Passo 4 : Faça $k = k + 1$ e retorne ao **Passo 1**

É possível, mediante imposição de algumas hipóteses, garantir que o algoritmo acima convirja, o teorema a seguir mostra quais suposições são necessárias para garantir tal convergência.

Teorema 2.1. *Considere a sequência $(x_k)_{k\geq 0}$, gerada pelo algoritmo 2, e suponha que cada r_i, $i = 1, \ldots, m$, possua primeira derivada Lipschitz, e que exista $\gamma > 0$ tal que*

$$\|J(x)z\| \geq \gamma \|z\|,$$

para todo $z \in \mathbb{R}^n$ e todo $x \in \mathcal{S}$, onde

$$\mathcal{S} = \{x \in \mathbb{R}^n : F(x_0) \leq F(x)\},$$

é um conjunto limitado.

Suponha também que, para todo α_k, as condições de Wolfe são satisfeitas, isto é,

$$\begin{aligned} F(x_k + \alpha_k p_{gn}) &\leq F(x_k) + c_1 \alpha_k \nabla F(x_k)^T p_{gn} \\ \nabla F(x_k + \alpha_k p_k)^T p_{gn} &\geq c_2 \nabla F(x_k)^T p_{gn}, \end{aligned}$$

com $0 < c_1 < c_2 < 1$. Então,

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

Demonstração. Ver [36]. □

Como pode ser consultado em Nocedal, [36], é possível garantir a existência de α_k sob as mesmas condições do teorema acima, bem como algoritmos especializados em obtê-los de maneira eficiente.

Embora seja possível garantir a convergência do método de Gauss-Newton mediante algumas hipóteses, o método pode se tornar ineficiente se a busca linear não for bem sucedida nas etapas do método. Assim, outras abordagens são estudadas.

2.4 MÉTODO DE LEVENBERG-MARQUARDT E MODIFICAÇÕES

Como vimos na seção anterior, quando $J(x)$ tem posto coluna completo, então o minimizador do modelo quadrático em x é a solução de $J(x)^T J(x) p_{gn} = -J(x)^T r(x)$, chamado de passo de Gauss-Newton (GN)

No entanto, não existe nenhuma forma inerente ao método de GN de controlar o tamanho do passo e, como vimos, normalmente é feito busca linear na direção encontrada. A busca linear é demasiada cara em casos em que a função objetivo tem um alto grau de não-linearidade. Além disso, a jacobiana pode não ter posto coluna completo ou mesmo ser mal condicionada, e, consequentemente o passo de GN pode não estar bem definido. Sendo assim, outras abordagens podem ser necessárias.

Método de Levenberg-Marquardt

Uma abordagem alternativa ao método de GN foi proposta em 1944 por Levenberg [27] e aprimorada em 1963 por Marquardt [31], conhecida nos dias de hoje como método de

Levenberg-Marquardt. Este método consiste em acrescentar um parâmetro $\lambda > 0$, na aproximação quadrática (11) da seguinte maneira

$$\begin{aligned} m(p) &= \frac{1}{2}\|r(x) + J(x)p\|^2 + \frac{1}{2}\lambda\|p\|^2 \quad (14)\\ &= \frac{1}{2}\|r(x)\|^2 + p^T J(x)^T r(x) + \frac{1}{2}p^T(J(x)^T J(x) + \lambda I)p. \quad (15)\end{aligned}$$

Assim, o minimizador está sempre bem definido, e pode ser obtido resolvendo

$$(J(x)^T J(x) + \lambda I)p_{lm} = -J(x)^T r(x). \quad (16)$$

Observe que $J(x)^T J(x) + \lambda I$ é sempre positiva definida. Portanto, a função quadrática m definida acima é estritamente convexa. Logo, o ponto crítico p_{lm} é único e minimizador global de (15). Tal solução é chamada de passo de Levenberg-Marquardt. Neste trabalho chamaremos o modelo m, definido acima em (14), de modelo quadrático de Levenberg-Marquardt, ou simplesmente de modelo quadrático de LM.

Na literatura é comum o parâmetro λ ser chamado de *damping* ou parâmetro de Levenberg-Marquardt.

Proposição 2.2. *Dado um ponto $x \in \mathbb{R}^n$, o passo p_{lm} é sempre uma direção de descida para a função F nesse ponto.*

Demonstração. Para ver isso basta provar que,

$$\begin{aligned} -p_{lm}^T J(x)^T r(x) &= p_{lm}^T (J(x)^T J(x) + \lambda I)p_{lm}\\ &= \underbrace{p_{lm}^T J(x)^T J(x) p_{lm}}_{\geq 0} + \underbrace{\lambda p_{lm}^T p_{lm}}_{>0} > 0.\end{aligned}$$

Logo, $p_{lm}^T \nabla F(x) = p_{lm}^T J(x)^T r(x) < 0$. Além disso, estamos supondo que $\nabla f(x) \neq 0$, pois caso contrário estaríamos em uma ponto crítico de $F(x)$.

Portanto, p_{lm} é sempre uma direção de descida. □

A proposição a seguir é de suma importância para estabelecer a relação que existe entre a norma do passo de LM e o parâmetro λ.

Proposição 2.3. *Considere a função $\psi : [0, \infty) \longrightarrow \mathbb{R}$, dada por*

$$\psi(\lambda) := \left\|(S + \lambda I)^{-1} g\right\|^2,$$

onde $S \in \mathbb{R}^{n\times n}$ é uma matriz simétrica, positiva definida e $g \in \mathbb{R}^n$ é um vetor qualquer, então a função ψ é contínua e não-crescente em λ. Além disso tem-se que $\lim_{\lambda\to\infty} \psi(\lambda) = 0$.

Demonstração. Como S é uma matriz simétrica, então em particular é diagonalizável. Assim existem matrizes U e Λ matrizes $n \times n$, tais que

$$S = U^T \Lambda U,$$

onde $\Lambda = \text{diag}\,(\mu_1, \ldots, \mu_n)$ e $U^T U = I$, isto é U é uma matriz ortogonal.

Deste modo, temos

$$\begin{aligned}
\psi(\lambda) &= \left\|(S+\lambda I)^{-1} g\right\|^2 \\
&= g^T (S+\lambda I)^{-1})^T (S+\lambda I)^{-1} g \\
&= g^T (S+\lambda I)^{-1} (S+\lambda I)^{-1} g \\
&= g^T (U^T \Lambda U + \lambda U^T U)^{-1} (U^T \Lambda U + \lambda U^T U)^{-1} g \\
&= g^T (U^T (\Lambda + \lambda I) U)^{-1} (U^T (\Lambda + \lambda I) U)^{-1} g \\
&= g^T U^{-1} (\Lambda + \lambda I)^{-1} (U^T)^{-1} U^{-1} (\Lambda + \lambda I)^{-1} (U^T)^{-1} g \\
&= (Ug)^T [(\Lambda + \lambda I)^2]^{-1} (Ug),
\end{aligned}$$

e definindo $\mathbf{v} = Ug = (v_1, \ldots, v_n)^T$, temos

$$\begin{aligned}
\psi(\lambda) &= (Ug)^T [(\Lambda + \lambda I)^2]^{-1} (Ug) \\
&= \mathbf{v}^T [(\Lambda + \lambda I)^2]^{-1} \mathbf{v} \\
&= \sum_{i=1}^{n} \frac{v_i^2}{(\mu_i + \lambda)^2},
\end{aligned}$$

onde fica evidente que $\psi(\lambda)$ é uma função contínua, decrescente em λ (quando este é maior ou igual a zero). Também é claro que $\psi(\lambda) \to 0$, quando $\lambda \to \infty$. □

Fazendo $S = J(x)^T J(x)$ e $g = J(x)^T r(x)$, observamos pela Proposição 2.3 que podemos controlar a norma de p_{lm} por meio do parâmetro λ, aumentando quando queremos diminuir o tamanho do passo de LM, e vice-versa.

Nos casos limites, quando $\lambda \to \infty$, temos

$$p_{lm} \to -\frac{1}{\lambda} J(x)^T r(x) = -\frac{1}{\lambda} \nabla F(x), \tag{17}$$

ou seja, p_{lm} tende a direção de máxima descida. Por outro lado, se $\lambda \to 0$, então

$$p_{lm} \to -(J(x)^T J(x))^{-1} J(x)^T r(x). \tag{18}$$

Nesse caso a tendência é de p_{lm} se tornar cada vez mais próximo ao passo de GN.

Dessa forma, o passo de LM combina a boa convergência global dos métodos de máxima descida com a rápida convergência local do método de Gauss-Newton, ou mesmo Newton, quando os resíduos são nulos na solução encontrada.

De modo geral, a eficiência do método de LM depende de como atualizamos o parâmetro λ. As diversas formas que constam na literatura para fazê-lo levam em consideração o resultado exposto na Proposição 2.3. A seguir vamos discorrer brevemente sobre esse tópico.

Parâmetro de Levenberg-Marquardt

Como visto na Proposição 2.3, é possível controlar o tamanho do passo de LM sem que seja necessário realizar busca linear como era o caso do método de GN.

Na literatura há um intenso estudo desse parâmetro [9], [15], [18], [23], [27], [31], [32], [34], além de haver formas de conciliar atualizações de λ com estratégias de busca linear, como Armijo, Wolfe e Goldstein [6].

Contudo, neste trabalho vamos abordar a estratégia que está contida em Nielsen [34], isto é,

$$\lambda_{k+1} = \begin{cases} 2\lambda_k, & \text{se } \rho_k < 1/4 \\ \lambda_k/3, & \text{se } \rho_k > 3/4 \\ \lambda_k, & \text{se caso contrário} \end{cases}, \tag{19}$$

e para cada $k \geq 0$,

$$\rho_k = \frac{F(x_k) - F(x_k + p_{lm})}{m_k(0) - m_k(p_{lm})}, \tag{20}$$

onde m_k denota o modelo quadrático de LM no ponto x_k. Note que ρ_k é a razão entre a redução de F e a redução prevista pelo modelo quadrático m_k, ambas na direção p_{lm}. Esse escalar pondera o quão bem o modelo aproxima F na direção de p_{lm}.

De forma intuitiva, a atualização λ_{k+1} dada acima se comporta da seguinte maneira. Se $\rho_k > 3/4$, então a função F é bem aproximada pelo modelo m_k na direção de p_{lm} a partir do ponto x_k, logo $\lambda_{k+1} = \lambda_k/3$, permitindo um passo maior em relação ao anterior.

Se $\rho_k < 1/4$, então o modelo quadrático m_k não aproxima bem a função F direção de p_{lm} a partir do ponto x_k, assim $\lambda_{k+1} = 2\lambda_k$, o que implica em um passo menor em relação ao anterior.

Por fim, se $\rho_k \in [1/4, 3/4]$, então m_k aproxima F de forma razoável, logo não faz sentido aumentar nem diminuir o passo em relação ao anterior. Portanto, $\lambda_{k+1} = \lambda_k$.

Contudo, essas atualizações descontínuas de λ quando ρ está próximo de $1/4$ ou $3/4$ podem diminuir a taxa de convergência [30]. Nielsen elimina essas descontinuidades da seguinte forma

$$\lambda_{k+1} = \begin{cases} \lambda_k \max\left\{\frac{1}{3}, 1-(2\rho_k-1)^3\right\}, \ \nu_k = 2, & \text{se } \rho_k > 0 \\ \lambda_k \nu_k, \ \nu_{k+1} = 2\nu_k, & \text{se caso contrário} \end{cases}. \tag{21}$$

As estratégias (19) e (21) são discutidas em mais detalhes em [34] e são ilustradas no gráfico da Figura 2.

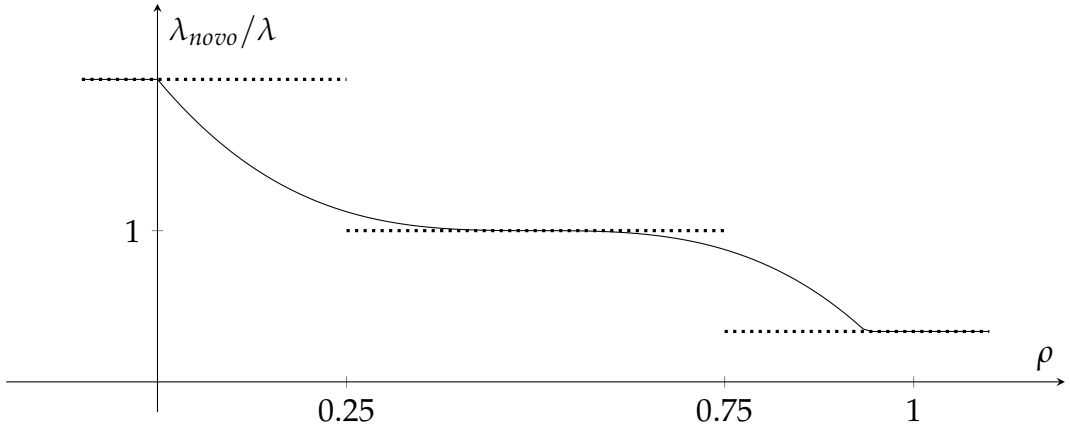

Figura 2: Atualização de λ por (19) (linha tracejada) e por (21) (linha contínua).

Assim, o método de LM pode ser dado por pelo algoritmo a seguir.

Algoritmo 3: MÉTODO DE LM

Entrada: Dados um ponto inicial x_0, $\eta \in \left[0, \frac{1}{4}\right)$, $\lambda_0 > 0$ e $k = 0$

Passo 1 : Resolva o sistema linear (16) para obter p_{lm}

Passo 2 : Obtenha λ_{k+1} usando (21)

Passo 3 : Calcule ρ_k conforme (20) e efetue

$$x_{k+1} = \begin{cases} x_k + p_{lm}, & \text{se } \rho_k > \eta \\ x_k, & \text{se caso contrário} \end{cases}$$

Passo 4 : Faça $k = k + 1$ e retorne ao **Passo 1**

No entanto, existe uma ampla gama de problemas de mínimos quadrados mal-condicionados e seu estudo é frequente na literatura [7], [10], [12], [40]. Em tais casos, se não for feito uma dimensionamento adequado da matriz do sistema de LM, pode ocorrer um grande número de iterações ou mesmo a não convergência.

Uma maneira de melhorar o condicionamento da matriz do do sistema é introduzir mais um parâmetro ao método, uma matriz D inversível. Como o parâmetro λ, este é mais um parâmetro alvo de intenso estudo. A seguir vamos discutir um pouco sobre sua influência e teoria associada.

Dimensionando o sistema de Levenberg-Marquardt

É comum em problemas de quadrados mínimos a função objetivo F ser muito sensível a pequenas alterações em algumas variáveis e pouco em outras, isto pode levar a curvas de nível de F tenderem a elipses altamente excêntricas nas proximidades de um ponto de mínimo, e como nosso objetivo é que o modelo quadrático aproxime da melhor maneira possível as curvas de nível de F, somos levados a considerar um controle da excentricidade das curvas de nível do modelo quadrático de LM, (14), em que os eixos são menores nas direções mais sensíveis e maiores nas direções menos sensíveis. Para isso, fazemos a seguinte modificação

$$m(p) = \frac{1}{2}\|r(x) + J(x)p\|^2 + \frac{1}{2}\lambda\|Dp\|^2 \tag{22}$$

$$= \frac{1}{2}\|r(x)\|^2 + p^T J(x)^T r(x) + \frac{1}{2}p^T(J(x)^T J(x) + \lambda D^T D)p. \tag{23}$$

para alguma matriz inversível D. Assim, o único minimizador desse modelo quadrático é dado pela solução de

$$(J(x)^T J(x) + \lambda D^T D)p = -J(x)^T r(x). \tag{24}$$

Como o modelo quadrático acima é uma generalização do modelo de LM, adotaremos todas nomenclaturas e notações do caso de LM definidos anteriormente. Assim, a solução do sistema acima será denotado por p_{lm}.

Note que se D é diagonal, então quanto maior é D_{ii}, menor é a i-ésima componente da solução do sistema acima, e quanto menor for D_{ii}, maior é essa componente, o que nos permite controlar as componentes de p_{lm}, em vista disso, tipicamente, D é tomada como sendo uma diagonal com os seus elementos relacionados com a escala do problema.

Do ponto de vista teórico, D pode ser tomada como uma matriz não singular, porém no amplo estudo na literatura acerca desse parâmetro, as escolhas mais recorrentes são de matrizes D diagonais, esse fato tem respaldo não apenas na observação do parágrafo anterior, mas também é uma maneira de diminuir o custo da resolução do sistema de LM, como veremos no Capítulo 4.

Recorrendo a literatura, inicialmente Levenberg [27] formulou o método assumindo $D = I$, mas como foi comentado acima, essa escolha pode gerar resultados insatisfatórios. Assim, outros autores sugerem alternativas, que abordaremos a seguir.

Uma característica importante para abordarmos as escolhas de D, é que se F varia pouco na direção da i-ésima variável, então $\|\partial r(x)/\partial x_i\|$ é pequena, por outro lado, se varia muito, essa norma é grande. Dessa forma, como $(J(x)^T J(x))_{ii} = \|\partial r(x)/\partial x_i\|^2$, para $i = 1, \ldots, n$, então uma maneira natural de controlar o tamanho das componentes de p_{lm} de acordo com a sensibilidade de F as suas variáveis é tomar

$$D_k = \operatorname{diag}\left(\sqrt{J(x_k)^T J(x_k)}\right),\ k \geq 0, \tag{25}$$

essa proposta aparece em Marquardt, [31].

Contudo, se $\|\partial r(x_k)/\partial x_i\|$ aumenta com k, então r pode estar variando significativamente em relação a variável i, e portanto os passos obtidos nesse caso podem não ser confiáveis. Este argumento leva à escolha de Moré [32], em que se adota

$$D_k = \text{diag}\left(d_1^{(k)}, \ldots, d_n^{(k)}\right) \quad \text{onde} \tag{26}$$

$$d_i^{(0)} = \sqrt{(J(x_0)^T J(x_0))_{ii}}, \quad i = 1, \ldots, n, \quad \text{para } k = 0. \tag{27}$$

$$d_i^{(k)} = \max\left\{d_i^{(k-1)}, \sqrt{(J(x_k)^T J(x_k))_{ii}}\right\}, \quad i = 1, \ldots, n, \quad \forall k \geq 1. \tag{28}$$

Note que, isso implicará que a variável i de maior influência na iteração k não será privilegiada em relação as outras na obtenção do passo de LM. No caso em que $\|\partial r(x_k)/\partial x_i\|$ diminui com k, temos que r varia lentamente em relação a variável i, portanto, em contrapartida do caso anterior, não há necessidade de diminuir o valor de d_i.

Uma escolha viável quando, para todo i, $\|\partial r(x_k)/\partial x_i\|$ não cresce em relação k, podemos simplesmente tomar D como sendo uma matriz constante, dada por

$$D_k = \text{diag}\left(\sqrt{J(x_0)^T J(x_0)}\right), \; k \geq 0, \tag{29}$$

note que nessa situação essa estratégia nada mais é do que uma aproximação da estratégia de Moré.

Resultados teóricos obtidos em Moré [32] apoiam (27) no lugar de (25), e em alguns casos (29).

Ainda que de modo geral toma-se D como sendo uma matriz diagonal, há autores que estudam o uso de D como sendo não necessariamente diagonal, por exemplo, [22] estuda o caso $D_k = \sqrt{J(x_k)^T J(x_k)} + \epsilon I$ onde $\epsilon > 0$. Embora essa escolha tenha suas vantagens, a opção diagonal é mais recomendada, por facilitar na decomposição matricial do sistema de LM, como veremos em mais detalhes no penúltimo capítulo deste trabalho.

Contudo, uma vez que acrescentamos uma matriz $D \neq I$, a Proposição 2.3 deixa de ser verdadeira, o que remete a questão de como deve ser feita a atualização do parâmetro λ anteriormente discutido. A seguir vamos mostrar que, em relação a convergência do método de LM, podemos reduzir o caso $D \neq I$ ao caso $D = I$ mediante uma simples mudança de variável.

Definindo $\hat{r}(x) := r(D^{-1}x)$ e $\hat{F}(x) := \hat{r}(x)$, temos que a jacobiana de $\hat{r}(x)$ é dada por $\hat{J}(x) := J(D^{-1}x)D^{-1}$, para todo $x \in \mathbb{R}^n$. Dessa forma, o modelo quadrático (14) torna-se

$$\hat{m}(\hat{p}) := \frac{1}{2}\|\hat{r}(x)\|^2 + \hat{p}^T\hat{J}(x)^T\hat{r}(x) + \frac{1}{2}\hat{p}^T(\hat{J}(x)^T\hat{J}(x) + \lambda I)\hat{p}. \tag{30}$$

Note que, definindo $\hat{x} = Dx$, para todo $x \in \mathbb{R}^n$, e aplicando na equação acima, temos que

$$\begin{aligned}
\hat{m}(\hat{p}) &= \frac{1}{2}\|\hat{r}(\hat{x})\|^2 + \hat{p}^T\hat{J}(\hat{x})^T\hat{r}(\hat{x}) + \frac{1}{2}\hat{p}^T(\hat{J}(\hat{x})^T\hat{J}(\hat{x}) + \lambda I)\hat{p} \\
&= \frac{1}{2}\|r(x)\|^2 + \hat{p}^T(D^{-1})^TJ(x)^Tr(x) + \frac{1}{2}\hat{p}^T((D^{-1})^TJ(x)^TJ(x)D^{-1} + \lambda I)\hat{p} \\
&= \frac{1}{2}\|r(x)\|^2 + \hat{p}^T(D^{-1})^TJ(x)^Tr(x) + \frac{1}{2}\hat{p}^T(D^{-1})^T(J(x)^TJ(x) + \lambda D^TD)D^{-1}\hat{p}.
\end{aligned}$$

Assim, o minimizador da função quadrática acima é a solução do sistema linear

$$(J(x)^TJ(x) + \lambda D^TD)D^{-1}\hat{p} = -J(x)^Tr(x)$$

vamos denotar a solução do sistema acima de $\hat{p}_{lm}$. Comparando o sistema acima com o sistema de LM dimensionado (24), temos $\hat{p}_{lm} = Dp_{lm}$. Note ainda que

$$\hat{\rho} := \frac{\hat{F}(\hat{x}) - \hat{F}(\hat{x} + \hat{p}_{lm})}{\hat{m}(0) - \hat{m}(\hat{p}_{lm})} = \frac{F(x) - F(x + p_{lm})}{m(0) - m(p_{lm})} = \rho,$$

dessa maneira, segundo o Algoritmo 3, o passo p_{lm} é aceito se, e só se, $\hat{p}_{lm}$ é aceito.

Dessa forma, se $(\hat{x}_k)_{k\geq 0}$ é uma sequência gerada pelo Algoritmo 3 que satisfaz

$$\liminf_{k\to\infty} \|\hat{J}(\hat{x}_k)^T\hat{r}(\hat{x}_k)\| = 0$$

então, $\liminf_{k\to\infty} \|(D_k^{-1})^TJ(x_k)^Tr(x_k)\| = 0$, onde $x_k = D_k^{-1}\hat{x}_k$, para todo $k \geq 0$. Assim, se $(D_k)_{k\geq 0}$ for limitada e $\|D_k^{-1}\| \not\to 0$, então $\liminf_{k\to\infty} \|J(x_k)^Tr(x_k)\| = 0$.

De maneira análoga e sob as mesmas hipóteses em $(D_k)_{k\geq 0}$ também temos que, se

$$\lim_{k\to\infty} \|(D_k^{-1})^TJ(x_k)^Tr(x_k)\| = 0,$$

então $\lim_{k\to\infty} \|J(x_k)^Tr(x_k)\| = 0$.

A hipótese de que $(D_k)_{k\geq 0}$ é limitada é satisfeita para as escolhas de matrizes D mencionadas nessa subseção, pois, a escolha de Fletcher, (29) é uma matriz constante e para as demais, como J é contínua e como os passos dado pelo Algoritmo 3 estão todos confinados ao subconjunto compacto $\{x \in \mathbb{R}^n : F(x) \leq F(x_0)\}$, então $(J(x_k))_{k\geq 0}$ é limitada. Além disso, para as escolhas de Marquardt, (25) e Moré, (27), se $\|D_k^{-1}\| \to 0$, então $1/\min \operatorname{diag}(D_k) \to 0$, logo $\min \operatorname{diag}(D_k) \to \infty$, o que implica $\max \operatorname{diag}(J(x_k)^T J(x_k)) \to \infty$ e portanto $\left\|J(x_k)^T J(x_k)\right\| \to \infty$, o que não ocorre pois $(J(x_k))_{k\geq 0}$ é limitada. Portanto, temos que ter $\|D_k^{-1}\| \not\to 0$ para as escolhas de D propostas anteriormente.

Mal condicionamento da matriz do sistema de LM

De maneira intuitiva, uma matriz $S \in \mathbb{R}^{n\times n}$, inversível, é dita bem condicionada se pequenas pertubações em seus elementos não afeta significativamente a solução do sistema $Sx = b$, e é dita mal condicionada caso contrário. Assim, o condicionamento da matriz nos dá uma estimativa do quanto o sistema é sensível a erros numéricos. Há diversas formas de definir rigorosamente esse conceito, mas nesse trabalho vamos adotar

$$\kappa(S) = \|S\| \cdot \|S^{-1}\|,$$

onde $\|\cdot\|: \mathbb{R}^{n\times n} \to \mathbb{R}$, é norma definida como

$$\|A\| = \text{maior autovalor de } (A^T A)^{\frac{1}{2}},$$

para toda $A \in \mathbb{R}^{n\times n}$ inversível. Em particular, se S é simétrica e positiva semidefinida, então $\|S\| = \lambda_{\max}$, onde $\lambda_{\max}$ é o maior autovalor de S.

Note que, também que $\kappa(A) = \|A\| \cdot \|A^{-1}\| \geq \|AA^{-1}\| = 1$, para toda $A \in \mathbb{R}^{n\times n}$ inversível, sendo que, quanto mais distante de 1, pior é o condicionamento.

Como é bem conhecido, sistemas mal condicionados podem resultar em muitas iterações ou em soluções de má qualidade, comprometendo o desempenho do método de LM.

Proposição 2.4. *Seja $S \in \mathbb{R}^{n\times n}$ uma matriz simétrica e positiva definida. Então*

$$\kappa(S) \geq \frac{\max \operatorname{diag}(S)}{\min \operatorname{diag}(S)}. \tag{31}$$

Demonstração. Primeiro, relembramos que se S é uma matriz simétrica e positiva definida, então $\|S\| = \lambda_{\max}$, onde $\lambda_{\max}$ é o maior autovalor de S. Vamos provar duas afirmações:

(a) $\|S\| = \max\left\{x^T S x : \|x\| = 1\right\}$.

(b) $\|S^{-1}\| = \max\left\{(x^T S x)^{-1} : \|x\| = 1\right\}$.

Prova de (a) e (b) : Vamos provar apenas **(a)**, pois **(b)** segue de maneira análoga. Seja $\mathcal{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, uma base ortonormal que diagonaliza S e sejam $\lambda_1, \ldots, \lambda_n$ seus respectivos autovalores.

Seja $\mathbf{u} \in \mathcal{B}$, o autovetor unitário associado ao maior autovalor $\lambda_{\max}$ de S. Assim, $\mathbf{u}^T S \mathbf{u} = \lambda_{\max}$, dessa forma temos $\max\left\{x^T S x : \|x\| = 1\right\} \geq \lambda_{\max}$.

Por outro lado, dado $\mathbf{v} \in \mathbb{R}^n$, tal que $\|\mathbf{v}\| = 1$, existem $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, tais que $\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{v}_i$, logo

$$\begin{aligned} \mathbf{v}^T S \mathbf{v} &= \mathbf{v}^T \left(\sum_{i=1}^n \alpha_i S \mathbf{v}_i \right) = \mathbf{v}^T \left(\sum_{i=1}^n \alpha_i \lambda_i \mathbf{v}_i \right) \\ &= \left(\sum_{i=1}^n \alpha_i \mathbf{v}_i \right)^T \left(\sum_{i=1}^n \alpha_i \lambda_i v_i \right) = \sum_{i=1}^n \alpha_i^2 \lambda_i \leq \left(\sum_{i=1}^n \alpha_i^2 \right) \lambda_{\max} = \lambda_{\max} \end{aligned}$$

onde usamos o fato de que $\sum_{i=1}^n \alpha_i^2 = 1$, pois $\|\mathbf{v}\| = 1$ e $\mathcal{B}$ é base ortonormal. Segue que $\lambda_{\max} \geq \max\left\{x^T S x : \|x\| = 1\right\}$ e portanto temos **(a)**.

Agora, considere $\mathcal{C} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, base canônica de $\mathbb{R}^n$. De **(a)**, concluímos que

$$\|S\| \geq \mathbf{e}_i^T S \mathbf{e}_i = S_{ii}, \text{ para todo } i = 1, \ldots, n,$$

logo $\|S\| \geq \max \operatorname{diag}(S)$. Por outro lado, de **(b)** temos

$$\|S^{-1}\| \geq (\mathbf{e}_i^T S \mathbf{e}_i)^{-1} = S_{ii}^{-1}, \text{ para todo } i = 1, \ldots, n,$$

logo $\|S^{-1}\| \geq 1/\min \operatorname{diag}(S)$.

Portanto, da definição de condicionamento, temos

$$\kappa(S) = \|S\| \cdot \|S^{-1}\| \geq \frac{\max \operatorname{diag}(S)}{\min \operatorname{diag}(S)},$$

como queríamos demonstrar. □

Mas a proposição acima nos diz que, quanto maior for a razão entre o maior e o menor valor da diagonal de uma matriz simétrica e positiva definida, maior será também o seu número de condicionamento, agora a proposição seguinte fornece uma desigualdade no sentido oposto da anterior, isto é, limitamos superiormente o número de condicionamento em função da razão do maior e menor valor da diagonal. A proposição a seguir pode ser encontrada em [26].

Proposição 2.5. *Sejam $n \geq 1$ e $S \in \mathbb{R}^{n\times n}$ uma matriz simétrica e positiva definida, então*

$$\kappa(S) \leq \frac{1+\sqrt{1-\det(S)n^n/\operatorname{tr}(S)^n}}{1-\sqrt{1-\det(S)n^n/\operatorname{tr}(S)^n}} \leq 4\frac{\operatorname{tr}(S)^n}{n^n \det(S)} - 2. \tag{32}$$

Nas mesmas hipóteses da proposição acima, temos que $\det(S) \geq \min \operatorname{diag}(S)^n$ e $\operatorname{tr}(S) \leq n \max \operatorname{diag}(S)$ e portanto segue que

$$\kappa(S) \leq 4\left(\frac{\max \operatorname{diag}(S)}{\min \operatorname{diag}(S)}\right)^n - 2. \tag{33}$$

Logo, o condicionamento da matriz do sistema de LM está intimamente relacionado com a distribuição dos elementos da diagonal. Vimos na subseção anterior algumas escolhas de matrizes D que tinham por finalidade melhor adequar o modelo quadrático as curvas de nível de F. Observe que podemos obter o passo de LM mediante a resolução do sistema abaixo

$$((D^{-1})^T J(x)^T J(x) D^{-1} + \lambda I)\hat{p}_{lm} = -(D^{-1})^T J(x)^T r(x), \tag{34}$$

onde $\hat{p}_{lm} = Dp_{lm}$. Note que a matriz do sistema acima é positiva definida com os elementos da digonal todos iguais, logo pela desigualdade (33) temos que seu condicionamento é menor ou igual a 2, ou seja, uma melhora substancial no condicionamento do sistema em casos mal condicionados.

Para a proposta de Moré a estratégia feita acima, a estratégia (34) melhoraria no condicionamento da matriz em uma iteração $k \geq 0$, ocorreria apenas se $d_i^{(k-1)}/(J(x_k)^T J(x_k))_{ii} \approx 1$, para todo i. No entanto, se a variação de uma componente diminuir muito nas iterações seguintes em relação ao máximo da diagonal, e este não aumentar muito, é melhor, em termos de condicionamento, resolver diretamente o sistema (24). Mas se o máximo da diagonal cresce muito em k e o mínimo decresce muito, uma estratégia de precondicionamento mais eficiente pode ser necessária para resolver o sistema (24).

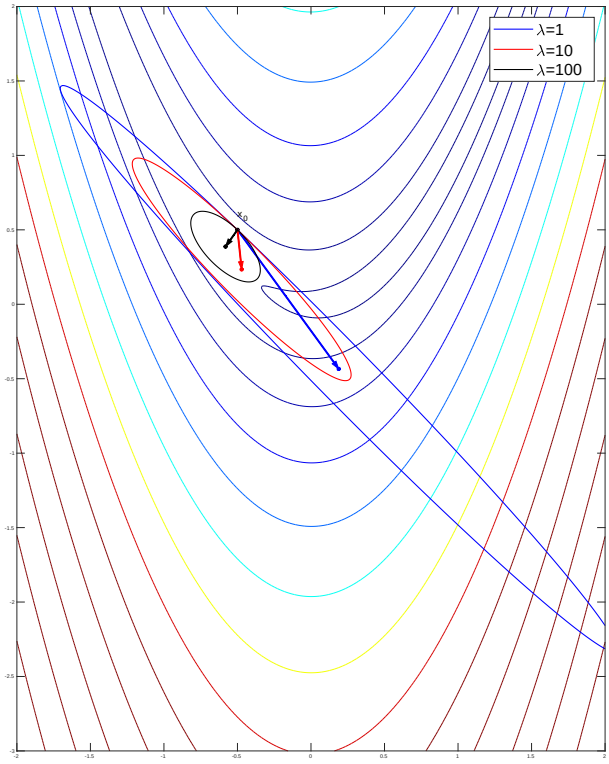

Figura 3: Passos do método de LM para função de Rosenbrock para diferentes valores de λ.

Exemplo com a função de Rosenbrock

A Figura 3 ilustra como são dados os passos de LM a parir do ponto inicial x_0 para o exemplo da função de Rosenbrock, dada por

$$F(x, y) = (1 - x)^2 + 100(y - x^2)^2. \tag{35}$$

Note que o ponto $(1, 1)$ é o único minimizador global dessa função. Além disso, essa função pode ser vista como um problema de mínimos quadrados não lineares. Para isso, basta notar que tomando

$$r_1(x, y) = \sqrt{2}(1 - x) \quad \text{e} \quad r_2(x, y) = 10\sqrt{2}(y - x^2),$$

temos

$$F(x, y) = \frac{1}{2}r_1(x, y)^2 + \frac{1}{2}r_2(x, y)^2.$$

Na Figura 3, é possível observar que a medida que λ aumenta, o tamanho do passo de LM diminui. Desenhamos apenas o nível $f(x_0)$ das curvas de nível dos modelos quadráticos, para $\lambda = 0$ (em azul), $\lambda = 10$ (em vermelho) e $\lambda = 100$ (em preto).

Método de LM com Hessiana completa

Baseado no método de LM, podemos voltar ao método de Newton e nos perguntar se é possível introduzir um parâmetro λ na hessiana de f a fim de torna-la positiva definida, para então aplicar alguma estratégia similar ao método de LM. De fato, isto é possível, ou seja, existe λ tal que $\nabla^2 f(x) + \lambda I$ é positiva definida, como mostraremos a seguir.

Proposição 2.6. *Seja $H \in \mathbb{R}^{n\times n}$ uma matriz simétrica, então existe um $\lambda \in \mathbb{R}$ tal que $H + \lambda I$ é positiva definida.*

Demonstração. Como H é simétrica, então é diagonalizável, isto é, existe U ortogonal, isto é, $U^T U = I$, tal que

$$H = U^T D U,$$

onde

$$C = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix},$$

onde $d_1, \ldots, d_n \in \mathbb{R}$. Seja $\lambda > 0$, temos

$$H + \lambda I = U^T C U + \lambda U^T U = U^T (C + \lambda I) U.$$

Seja $v \in \mathbb{R}^n$ não nulo, defina $u = Uv$ e tome $\lambda > -\min\{d_1, \ldots, d_n\}$, logo

$$\begin{aligned} v^T (H + \lambda I) v &= v^T U^T (C + \lambda I) U v \\ &= (Uv)^T (C + \lambda I)(Uv) \\ &= u^T (C + \lambda I) u = \sum_{i=1}^{n} (d_i + \lambda) u_i^2 > 0, \end{aligned}$$

onde $u = (u_1, \ldots, u_n)^T$, e usamos que $d_i + \lambda > 0,\ \forall i \in \{1, \ldots, n\}$. □

Assim, em cada etapa definimos o seguinte modelo quadrático

$$\Gamma(p) = f(x) + p^T \nabla f(x) + \frac{1}{2} p^T (\nabla^2 f(x) + \lambda I) p, \tag{36}$$

onde λ é escolhido de tal forma a tornar $\nabla^2 f(x) + \lambda I$ positiva definida. Logo, o minimizador global desse modelo é dado por

$$(\nabla^2 f(x) + \lambda I) p_H = -\nabla f(x). \tag{37}$$

Definindo

$$\rho^H = \frac{f(x) - f(x + p_H)}{\Gamma(0) - \Gamma(p_H)}, \tag{38}$$

estabelecemos o seguinte algoritmo.

Algoritmo 4: MÉTODO DE LM COM HESSIANA COMPLETA

Entrada: Dados x_0, η e um valor inicial para λ

Passo 1 : Realize a fatoração Cholesky de $\nabla^2 f(x_k) + \lambda I$

Passo 2 : Se a fatoração falhar faça $\lambda = 2\lambda$ e retorne ao **Passo 1**

Passo 3 : Faça $\lambda_k = \lambda$

Passo 4 : Resolva $(\nabla^2 f(x_k) + \lambda_k I) p_H = -\nabla f(x_k)$ usando a fatoração obtida em **Passo 1**

Passo 5 : Avalie ρ_k^H conforme (38), e faça

$$x_{k+1} = \begin{cases} x_k + p_H, & \text{se } \rho_k^H > \eta \\ x_k, & \text{se caso contrário} \end{cases}$$

Passo 6 : Obtenha λ_{k+1} como em (21)

Passo 7 : $x_{k+1} = x_k + p_H$

Passo 8 : Faça $k = k + 1$ e retorne ao **Passo 1**

Observe que para tornar $\nabla^2 f(x) + \lambda I$ positiva definida, o que fazemos é dobrar λ a cada falha na fatoração Cholesky. Proceder dessa maneira é preferível, mesmo que a Proposição 2.6 nos diz como encontrar λ que estamos buscando, uma vez que é inviável computacionalmente encontrar os autovalores de matrizes com muitas entradas.

Do ponto de vista teórico, o método descrito acima é mais preciso que LM, no entanto, pode ser caro computacionalmente obter λ de tal forma que $\nabla^2 f(x) + \lambda I$ seja positiva definida. De fato, em casos em que a hessiana possui autovalores próximos de zero ou muito negativos, teremos muito consumo de tempo para tornar $\nabla^2 f(x) + \lambda I$ positiva definida, como pode ser constatado no Algoritmo 4.

A seguir apresentaremos dois métodos que trabalham um caso particular do problema de mínimos quadrados não lineares. Estes métodos fazem uso de dois sistemas lineares em cada iteração, com informações apenas de derivadas de primeira ordem dos resíduos, e possuem garantia de convergência cúbica.

LM Modificado

Em [13], [14], Jinyan Fan estuda um caso particular do problema de mínimos quadrados não lineares, onde deseja-se encontrar um zero para funções objetivo cuja a quantidade de resíduos é igual a quantidade de parâmetros. Nesses trabalhos, Fan propõe dois métodos de convergência cúbica que são inspirados no método de Newton modificado [20], [43], que também possui convergência cúbica.

Considere o sistema não linear de equações

$$r(x) = 0, \tag{39}$$

onde $r : \mathbb{R}^n \to \mathbb{R}^m$ é uma função de classe C^2. Devido a não linearidade de r, o sistema (39) pode não ter soluções, no entanto, vamos supor que tais soluções existem e focaremos apenas em construir um método para encontrar ao menos uma delas.

O método de Newton é um algoritmo clássico para resolver esse tipo problema. Dado um ponto inicial x_0, o método consiste em resolver em cada iteração o sistema

$$J(x_k)p_k = -r(x_k),$$

onde $J(x_k)$ é a jacobiana de r aplicada em x_k, e atualizar $x_{k+1} = x_k + p_k$.

Seja x^* tal que $r(x^*) = 0$. A convergência do método será quadrática em uma vizinhança de x^*, se $J(x^*)$ é inversível e lipschitz [25].

No entanto, o cálculo da jacobiana $J(x_k)$ em cada iteração pode se tornar demasiadamente caro dependendo da complexidade de se derivar a função r e do número de variáveis n. Assim, torna-se necessário a criação de novas técnicas que sejam mais eficientes. Por exemplo, há algoritmos que, ao invés de usar J, utiliza uma matriz que a aproxima e que tem um custo menor. Tais métodos são chamados de Quase-Newton.

Porém, há um outro método que ainda usa a matriz J, mas o seu cálculo é aproveitado para resolver um sistema linear adicional. Nesse método, primeiramente se resolve

$$J(x_k)p_k = -r(x_k),$$

e aproveitando da jacobiana já calculada, resolve-se um segundo,

$$J(x_k)q_k = -r(y_k),$$

onde $y_k = x_k + p_k$. Assim, dado um ponto inicial x_0, os passos do método são dados por $x_{k+1} = x_k + s_k$, onde $s_k = p_k + q_k$. Este algoritmo é chamado de Newton Modificado e, em contraste com o método de Newton, tem convergência cúbica se $J(x^*)$ é lipschitz e não singular [20], [43].

Contudo, a matriz jacobiana nem sempre tem posto coluna completo, logo o passo de Newton nem sempre está bem definido. Além disso, a jacobiana pode ser mal condicionada, o que implica na instabilidade do sistema. Para contornar isto, seria interessante darmos passos menores. Uma maneira de fazer isso é acrescentar ao sistema de equações que define p_k um parâmetro $\lambda_k > 0$, da seguinte forma

$$\begin{bmatrix} J(x_k) \\ \sqrt{\lambda_k} I \end{bmatrix} p_k = \begin{bmatrix} r(x_k) \\ 0 \end{bmatrix}. \tag{40}$$

Observe que o sistema linear acima nos fornece o passo de Levenberg-Marquardt, pois implica em

$$(J(x_k)^T J(x_k) + \lambda_k I)p_k = -J^T(x_k)r(x_k), \tag{41}$$

isto é, o próprio sistema de LM. Nesta seção, vamos denotar o passo de LM, na k-ésima iteração, por p_k.

De maneira análoga ao método Newton Modificado, resolvemos mais um sistema linear dado por

$$(J(x_k)^T J(x_k) + \lambda_k I)q_k = -J^T(x_k)r(y_k), \tag{42}$$

onde $y_k = x_k + p_k$.

Note que, de forma semelhante ao método de LM, o passo q_k é o minimizador global da seguinte função quadrática

$$\tilde{m}(p) = \frac{1}{2}\|r(y_k) + J(x_k)p\|^2 + \frac{1}{2}\lambda_k \|p\|^2 . \tag{43}$$

Dessa forma, o passo gerado por esse método é definido como $s_k = p_k + \alpha_k q_k$, onde o parâmetro $\alpha_k > 0$ surge como uma maneira de ajustar o passo q_k.

Neste trabalho, vamos considerar duas maneiras distintas de se definir o parâmetro α_k. A primeira é simplesmente tomar $\alpha_k = 1$ para todo k, e nesse caso temos o chamado método de Levenberg-Marquardt modificado (LMM). O segundo, vamos considerar uma maneira de atualizar α_k de forma a acelerar o método LMM. Vamos descrevê-la em mais detalhes a seguir.

LM Modificado Acelerado

Podemos acelerar o método LMM modificando a maneira como atualizamos α_k. Para tanto, definimos

$$F(x) = \frac{1}{2}\|r(x)\|^2 . \tag{44}$$

Lembrando que $y_k = x_k + p_k$, então observe que

$$F(y_k + q_k) = \frac{1}{2}\|r(y_k + q_k)\|^2 \approx \frac{1}{2}\|r(y_k) + J(x_k)q_k\|^2 .$$

Além disso, devido ao resultado demonstrado por Powell, em [39], temos que

$$\frac{1}{2}\|r(y_k)\|^2 - \frac{1}{2}\|r(y_k) + J(x_k)q_k\|^2 \geq \frac{1}{2}\left\|J(x_k)^T r(y_k)\right\| \min\left\{ \|q_k\| , \frac{\left\|J(x_k)^T r(y_k)\right\|}{\left\|J(x_k)^T J(x_k)\right\|} \right\}, \tag{45}$$

assim,

$$\begin{aligned} F(y_k) - F(y_k + q_k) &\approx \frac{1}{2}\|r(y_k)\|^2 - \frac{1}{2}\|r(y_k) + J(x_k)q_k\|^2 && (46)\\ &\geq \frac{1}{2}\left\|J(x_k)^T r(y_k)\right\| \min\left\{ \|q_k\| , \frac{\left\|J(x_k)^T r(y_k)\right\|}{\left\|J(x_k)^T J(x_k)\right\|} \right\} && (47) \end{aligned}$$

Note que a diferença expressa na desigualdade (45) é positiva, logo, tendo em vista a aproximação (46), é interessante acentuar ainda o valor dessa diferença mediante

um controle no tamanho do passo q_k, em uma tentativa de aumentar a diferença $F(y_k) - F(y_k + q_k)$, para tanto, vamos fazer uma busca linear em y_k ao longo de q_k, resolvendo o seguinte problema

$$\min_{\alpha>0} \frac{1}{2} \|r(y_k) + \alpha J(x_k)q_k\|^2 . \tag{48}$$

Este problema pode ser reescrito de forma equivalente como

$$\max_{\alpha>0} \frac{1}{2} \|r(y_k)\|^2 - \frac{1}{2} \|r(y_k) + \alpha J(x_k)q_k\|^2 .$$

Definindo

$$\phi(\alpha) = \frac{1}{2} \|r(y_k)\|^2 - \frac{1}{2} \|r(y_k) + \alpha J(x_k)q_k\|^2 ,$$

$\phi(\alpha)$ pode ser reescrita como

$$\begin{aligned} \phi(\alpha) &= \frac{1}{2} \|r(y_k)\|^2 - \frac{1}{2} \|r(y_k) + \alpha J(x_k)q_k\|^2 \\ &= \left(- r(y_k)^T J(x_k)q_k \right)\alpha - \left(\frac{1}{2} q_k^T J(x_k)^T J(x_k)q_k \right)\alpha^2 \\ &= \left(q_k^T (J(x_k)^T J(x_k) + \lambda_k I)q_k \right)\alpha - \left(\frac{1}{2} q_k^T J(x_k)^T J(x_k)q_k \right)\alpha^2, \end{aligned}$$

onde usamos o fato de que

$$-r(y_k)^T J(x_k) = q_k^T (J(x_k)^T J(x_k) + \lambda_k I).$$

Observe que ϕ é uma função quadrática em α com máximo em

$$\tilde{\alpha}_k = \frac{q_k^T (J(x_k)^T J(x_k) + \lambda_k I)q_k}{q_k^T J(x_k)^T J(x_k)q_k} = 1 + \frac{\lambda_k q_k^T q_k}{q_k^T J(x_k)^T J(x_k)q_k} > 1,$$

desde que $q_k^T J(x_k)^T J(x_k)q_k \neq 0$.

Note que $\tilde{\alpha}_k$ obtido é sempre maior do que 1, e portanto é razoável supor que q_k é uma boa direção de descida para F a partir do ponto y_k, além disso, pela construção do parâmetro α_k segue que

$$\frac{1}{2} \|r(y_k)\|^2 - \frac{1}{2} \|r(y_k) + \alpha_k J(x_k)q_k\|^2 \geq \frac{1}{2} \|r(y_k)\|^2 - \frac{1}{2} \|r(y_k) + J(x_k)q_k\|^2 ,$$

logo, é razoável supor que a busca linear efetuada acima na direção de q_k a partir de y_k pode de acelerar o método LMM. Mas note, que o escalar $\tilde{\alpha}_k$ é sensível ao fator $q_k^T J(x_k)^T J(x_k) q_k$, pois quando este é próximo de zero, tem-se que $\tilde{\alpha}_k$ é demasiadamente grande, pois

$$\|J(x_k)q_k\|^2 = q_k^T J(x_k)^T J(x_k) q_k \approx 0,$$

Para que haja alguma alteração significativa no valor de $\|r(y_k) + \alpha J(x_k)q_k\|^2$, α tem que ser suficientemente grande, pois caso contrário

$$\|r(y_k) + \alpha J(x_k)q_k\|^2 \approx \|r(y_k)\|^2 .$$

Por outro lado queremos evitar tomar α_k desnecessariamente grande. Para isso, limitamos seu valor máximo por uma constante $\tilde{\alpha} > 1$ e definimos α_k como sendo o passo que soluciona

$$\max_{\alpha \in [1,\tilde{\alpha}]} \frac{1}{2}\|r(y_k)\|^2 - \frac{1}{2}\|r(y_k) + \alpha J(x_k)q_k\|^2 . \tag{49}$$

Assim, combinando essa escolha de α_k ao método LMM temos o chamado de Levenberg-Marquardt modificado acelerado (LMMA).

No caso do método de LM, a direção p_k faz com que $m(0) - m(p_k) > 0$, onde m é dada por (14). Porém o passo s_k não satisfaz, necessariamente, $m(0) - m(s_k) > 0$. Sendo assim, não podemos usar o mesmo preditor do decréscimo de F que foi usado no método de LM para definir o parâmetro ρ_k dado em (20). No entanto, ainda é possível definir um preditor de maneira satisfatória, e consequentemente um parâmetro ρ_k específico para esse caso, para tando, basta definir

$$\begin{aligned} Pred_k &= \Big(m(0) - m(p_k)\Big) + \Big(\tilde{m}(0) - \tilde{m}(q_k)\Big) \\ &= \frac{1}{2}\|r(x_k)\|^2 - \frac{1}{2}\|r(x_k) + J(x_k)p_k\|^2 + \frac{1}{2}\|r(y_k)\|^2 - \frac{1}{2}\|r(y_k) + \alpha_k J(x_k)q_k\|^2 , \end{aligned}$$

é interessante defini-lo dessa maneira, pois, de maneira análoga ao caso da desigualdade (45), temos que

$$\frac{1}{2}\|r(x_k)\|^2 - \frac{1}{2}\|r(x_k) + J(x_k)p_k\|^2 \geq \frac{1}{2}\left\|J(x_k)^T r(x_k)\right\| \min\left\{ \|p_k\| , \frac{\left\|J(x_k)^T r(x_k)\right\|}{\|J(x_k)^T J(x_k)\|} \right\}, \tag{50}$$

e, consequentemente, em vista das desigualdades (45) e (50), temos

$$Pred_k \geq \frac{1}{2}\left\|J(x_k)^T r(x_k)\right\| \min\left\{\|p_k\|, \frac{\left\|J(x_k)^T r(x_k)\right\|}{\left\|J(x_k)^T J(x_k)\right\|}\right\} + \frac{1}{2}\left\|J(x_k)^T r(y_k)\right\| \min\left\{\|q_k\|, \frac{\left\|J(x_k)^T r(y_k)\right\|}{\left\|J(x_k)^T J(x_k)\right\|}\right\},$$

o que nos permite definir,

$$\rho_k = \frac{F(x_k) - F(x_k + s_k)}{Pred_k}. \tag{51}$$

Note que, para todo k, além de termos $Pred_k > 0$, a redução prevista $Pred_k$ é maior do que aquela prevista pelo modelo de LM, assim as direções aceitas apresentam uma redução maior na função objetivo quando comparadas com as direções de LM, o que pode reduzir significativamente o número de passos aceitos ao custo de aumentar os não aceitos, em relação ao algoritmo de LM. Isso apresenta uma vantagem, pois como veremos no penúltimo capítulo desse trabalho, os passos aceitos são mais caros de serem obtidos em relação aos não aceitos.

Fazendo uso de toda teoria exposta nesse capítulo, pode-se construir um algoritmo com algumas propriedades especiais.

Algoritmo 5: LMMA

Entrada: Dados $x_0 \in \mathbb{R}^n$, $\mu_0 > M > 0$, $0 < p_0 \leq p_1 \leq p_2 < 1$, $0 < \delta \leq 2$, $\tilde{\alpha} > 0$ e seja $k = 0$

Passo 1 : Usando $\lambda_k = \mu_k \|r(x_k)\|^\delta$, calcule p_k e q_k via (41) e (42) respectivamente

Passo 2 : Seja $s_k = p_k + \alpha_k q_k$, onde α_k é obtido resolvendo (49)

Passo 3 : Calcule ρ_k usando (51) e efetue

$$x_{k+1} = \begin{cases} x_k + s_k, & \text{se } \rho_k \geq p_0 \\ x_k, & \text{se caso contrário} \end{cases}$$

Passo 4 : Escolha μ_{k+1} da seguinte forma

$$\mu_{k+1} = \begin{cases} 4\mu_k, & \text{se } \rho_k < p_1 \\ \mu_k, & \text{se } \rho_k \in [p_1, p_2] \\ \max\{\mu_k/4, M\}, & \text{se } \rho_k > p_2 \end{cases}$$

Passo 5 : Faça $k = k + 1$ e retorne ao **Passo 1**

Como demonstrado em [13], quando $n = m$, o algoritmo acima tem convergência de ordem $\min\{1 + 2\delta, 3\}$. Em particular, tem convergência cúbica se $1 \leq \delta \leq 2$.

Já para o caso em que se faz a escolha $\alpha_k = 1$ para todo k no algoritmo acima, isto é, considerando o método LMM, tem-se a garantia de convergência cúbica, quando $n = m$, [14].

2.5 MÉTODO DE REGIÕES DE CONFIANÇA

Como vimos, a Proposição 2.3 estabelece a seguinte relação

$$\left\|(J(x)^T J(x) + \lambda I)^{-1} J(x)^T r(x)\right\|^2 = \sum_{i=1}^{n} \frac{v_i^2}{(\mu_i + \lambda)^2} \tag{52}$$

onde $J(x)^T J(x) = U^T \Lambda U$, $\Lambda = \text{diag}\,(\mu_1, \ldots, \mu_n)$, $U^T U = I$ e $v = UJ(x)^T r(x)$. Assim, através do parâmetro λ, é possível controlar o tamanho do passo de LM. Além disso, o passo de LM dado pela solução do sistema (16) é o único minimizador global do modelo quadrático (14) se o restringirmos a uma bola fechada de raio $\|p_{lm}\|$.

Baseado nessas características os métodos de regiões de confiança surgem como generalização do método de LM. Resumidamente, dado um ponto, considera-se uma aproximação quadrática da função objetivo em torno desse ponto e toma-se o passo como sendo aquele que minimiza esse modelo dentro de uma determinada vizinhança.

A seguir vamos definir rigorosamente como são construídos os métodos de regiões de confiança.

Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função de classe C^2 e tome um ponto $x \in \mathbb{R}^n$. Considere o modelo quadrático

$$m(p) = f(x) + p^T \nabla f(x) + \frac{1}{2} p^T A(x) p, \tag{53}$$

onde assumiremos neste trabalho que $A(x)$ é uma matriz simétrica e uniformemente limitada. Em geral, tomamos $A(x)$ como sendo alguma aproximação da hessiana de f no ponto x.

De modo geral, o método de regiões de confiança consiste em resolver, em cada etapa do método, dado $\Delta > 0$, o seguinte subproblema

$$\min_{p \in \mathbb{R}^n} m(p), \quad \text{sujeito à} \quad \|p\| \leq \Delta. \tag{54}$$

O conjunto $R = \{p \in \mathbb{R}^n \ : \ \|p\| \leq \Delta\}$, obtido a cada iteração, é chamado de região de confiança, e o parâmetro Δ é o raio da região de confiança.

Em função da aproximação de Taylor de segunda ordem no ponto x e na direção de p, a diferença entre o $m(p)$ e $f(x+p)$ é da ordem $\mathcal{O}(\|p\|^2)$ para $\|p\|$ suficientemente pequeno.

Tomando $A(x) = \nabla^2 f(x)$, teremos o chamado Método de Regiões de Confiança de Newton. Note que, nesse caso, o erro na aproximação de Taylor de segunda ordem, cujo erro é $\mathcal{O}(\|p\|^3)$, ou seja, quando a região de confiança é pequena, o modelo é notoriamente preciso.

Assim, o objetivo desse método é em cada etapa resolver o subproblema (54), isto é, encontrar um minimizador p^* desse modelo quadrático dentro de uma bola de raio Δ. Note que se A_k é positiva definida e $\left\|A(x)^{-1}\nabla f(x)\right\| \leq \Delta$, então $p^* = -A(x)^{-1}\nabla f(x)$. No entanto, se $\left\|A(x)^{-1}\nabla f(x)\right\| > \Delta$ nem sempre é fácil encontrar p^*. Porém, para se ter convergência global e métodos práticos, basta obter soluções aproximadas do subproblema (54), como veremos mais a frente.

Da mesma maneira que o método de LM, em que a atualização do parâmetro λ é de suma importância, a forma como atualizamos o parâmetro Δ também é crucial para o desenvolvimento no método de regiões de confiança. Vamos discutir uma pouco acerca dele a seguir.

Atualização do parâmetro Δ

No inicio desta seção, comentamos o fato de que o método de LM inspirou a formulação dos métodos de regiões de confiança. Para evidenciar ainda mais esse fato, vamos mostrar que o método de LM surge naturalmente como um caso particular de métodos de regiões de confiança.

Tomando as regiões de confiança convenientemente como $\Delta = \|p_{lm}\|$, então o método de LM é equivalente a resolver, em cada iteração, um subproblema da forma

$$\min_{p\in\mathbb{R}^n} m(p) = F(x) + p^T(J(x)^T r(x)) + \frac{1}{2}p^T(J(x)^T J(x) + \lambda I)p, \quad \text{sujeito à} \quad \|p\| \leq \|p_{lm}\|.$$

Onde $F(x) = \frac{1}{2}\|r(x)\|^2$, $\nabla F(x) = J(x)^T r(x)$ e $A(x) = J(x)^T J(x) + \lambda I$. Mais ainda, pela equação (52), temos uma maneira de atualizar Δ a partir do parâmetro λ, ou seja,

$$\Delta^2 = \sum_{i=1}^{n} \frac{v_i^2}{(\mu_i + \lambda)^2}. \tag{55}$$

Observe que a relação acima deixa explícita o fato de que Δ diminui se λ aumenta, e por outro lado, Δ aumenta se λ diminui. E é justamente essa relação que inspira a maneira como atualizamos a região de confiança.

Assim, inspirado na relação inversa dada pela equação (55), e na maneira como a atualização de λ é feita no método de LM, podemos intuir uma maneira similar de definir a como as regiões de confiança são atualizadas, conforme dado em [30], temos

$$\Delta_{k+1} = \begin{cases} \Delta_k/3, & \text{se } \rho_k < 1/4 \\ 2\Delta_k, & \text{se } \rho_k > 3/4 \\ \Delta_k, & \text{se } \rho_k \in [1/4, 3/4] \end{cases}, \tag{56}$$

onde para cada $k \geq 0$,

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}. \tag{57}$$

Outra forma de atualização é descrita em [36], a qual a convergência é garantida, como veremos mais a frente.

Algoritmo 6: ATUALIZAÇÃO DO RAIO DA REGIÃO DE CONFIANÇA Δ

Entrada: Dados $\Delta_0 > 0$, $\hat{\Delta} > 0$, x_0 e $\eta \in \left[0, \frac{1}{4}\right)$

Passo 1 : Seja p_k alguma solução aproximada de (54) e ρ_k usando (57)

Passo 2 : Se $\rho_k \leq \eta$, faça

$$\Delta_{k+1} = \Delta_k/4, \quad x_{k+1} = x_k, \quad k = k+1 \quad \text{e retorne ao} \quad \textbf{Passo 1}$$

Passo 3 : Se $\eta < \rho_k < 1/4$, faça

$$\Delta_{k+1} = \Delta_k/4, \quad x_{k+1} = x_k + p_k, \quad k = k+1 \quad \text{e retorne ao } \textbf{Passo 1}$$

Passo 4 : Se $1/4 \leq \rho_k \leq 3/4$ ou ($\|p_k\| < \Delta_k$ e $\rho_k \geq 1/4$), faça

$$\Delta_{k+1} = \Delta_k, \quad x_{k+1} = x_k + p_k, \quad k = k+1 \quad \text{e retorne ao } \textbf{Passo 1}$$

Passo 5 : Se $\rho_k > 3/4$ e $\|p_k\| = \Delta_k$, faça

$$\Delta_{k+1} = \min\{2\Delta_k, \hat{\Delta}\}, \quad x_{k+1} = x_k + p_k, \quad k = k+1 \quad \text{e retorne ao } \textbf{Passo 1}$$

De maneira análoga ao caso de LM, ρ_k também é a razão entre a redução de f e a redução prevista pelo modelo m_k, ambas na direção p_k. Esse escalar, nos diz o quanto nosso modelo aproxima f na direção de p_k.

De forma intuitiva, a atualização Δ_{k+1} dada acima se comporta da seguinte maneira: Se $\rho_k > 3/4$ e $\|p_k\| = \Delta_k$, então nossa função f, além de ser bem aproximada pelo modelo m_k na direção de p_k a partir do ponto x_k, o passo tomado foi o maior possível. Logo, fazemos $\Delta_{k+1} = \min\{2\Delta_k, \hat{\Delta}\}$, isto é, permitimos um passo maior em relação ao anterior. Se $\rho_k < 1/4$, então a m_k não aproxima bem a função f na direção de p_k a partir do ponto x_k. Assim fazemos $\Delta_{k+1} = \Delta_k/4$, o que implica em um passo menor em relação ao anterior. Por fim, se $\rho_k \in [1/4, 3/4]$ ou $\|p_k\| < \Delta_k$, então m_k aproxima f na direção de p_k a partir do ponto x_k de forma razoável ou o passo tomado é menor que a região de confiança. Portanto, não faz sentido aumentar nem diminuir o passo em relação ao anterior.

Note que o raio da região de confiança só é aumentado quando de fato se atinge a borda da região de confiança. Por outro lado, quando a borda não é atingida, pode-se concluir que o tamanho do raio não está interferindo no cálculo de p_k e portanto não faz sentido altera-lo.

Para que o algoritmo descrito tenha sentido do ponto de vista prático, precisamos entender como são obtidas as soluções do subproblema (54).

Da mesma forma que no método de LM, onde facilmente obtém-se o minimizador global do modelo quadrático através da resolução de único sistema linear, também podemos formular um resultado que generaliza esse fato no contexto dos métodos de regiões de confiança. O teorema abaixo, cuja a demonstração está presente em [36], faz essa generalização, caracterizando as soluções dos subproblemas da forma (53).

Teorema 2.7. *O vetor p^* é uma solução global do problema da região de confiança (54) se e somente se, p^* é solução viável e existe $\lambda_k \geq 0$ tal que*

$$\begin{aligned} (A_k + \lambda_k I)p^* &= -\nabla f(x_k); \\ \lambda_k(\Delta_k - \|p^*\|) &= 0; \\ A_k + \lambda_k I \quad &\text{é positiva definida.} \end{aligned}$$

Note que o passo de LM satisfaz as três equações do teorema se considerarmos $\Delta_k = \|p_{lm}\|$ e $p^* = p_{lm}$.

Exemplo com curvas de nível da função de Rosenbrock

Para ilustrar como se comportam os método de regiões de confiança, vamos utilizar o exemplo da função de Rosenbrock, (35). No exemplo ilustrado na Figura 4, usamos o modelo quadrático dado por

$$m_0(p) = f(x_0) + \nabla f(x_0)^T p + \frac{1}{2} p^T (J(x_0)^T J(x_0) + I)p$$

em torno do ponto $x_0 = (-0.5, 0.5)^T$ e com raios das regiões de confiança variando de acordo com os círculos vermelhos pontilhados: $\Delta = 1$, $\Delta = 0.5$ e $\Delta = 0.25$. A matriz $J(x_0)$ é a jacobiana do vetor de resíduos $r(x, y) = (r_1(x, y), r_2(x, y))^T$ aplicada em $(x, y) = x_0$.

Esboçamos apenas as curvas de nível do modelo quadrático que assumem o valor de mínimo dentro da região de confiança desenhada, isto é, a curva de nível azul é o nível que detém o mínimo do modelo quadrático dentro da região de confiança de raio $\Delta = 1$, e assim por diante. As setas dadas na figura representam os passos tomados a partir de x_0 a fim de minimizar a quadrática dentro da região de confiança. Observe que quanto maior o Parâmetro Δ, é possível obter passos maiores, porém, nem sempre os maiores passos serão os melhores, como pode ser visto na Figura 4, onde o passo que mais diminui o valor da função Rosenbrock é o correspondente a região de confiança $\Delta = 0.5$. Isso mostra a complexidade envolvida em desenvolver métodos eficientes utilizando estratégias de regiões de confiança.

Nas subseções seguintes vamos introduzir dois métodos de regiões de confiança clássicos.

Embora a princípio buscamos a solução ótima do subproblema (54), para propósitos de convergência, é suficiente encontrar uma solução aproximada p_k que esteja dentro da região de confiança e forneça uma redução suficiente no modelo. É nessa ideia que o método de Dogleg se baseia. Antes de apresentar esse método, vamos definir o chamado ponto de Cauchy.

2.5.1 Ponto de Cauchy

Seja $f : \mathbb{R}^n \to R$ um função de classe C^1. Definimos o seguinte subproblema

$$p_k^* = \arg\min_{p \in R_k} f(x_k) + \nabla f(x_k)^T p, \qquad R_k = \{p \in \mathbb{R}^n : \|p\| \leq \Delta_k\}. \tag{58}$$

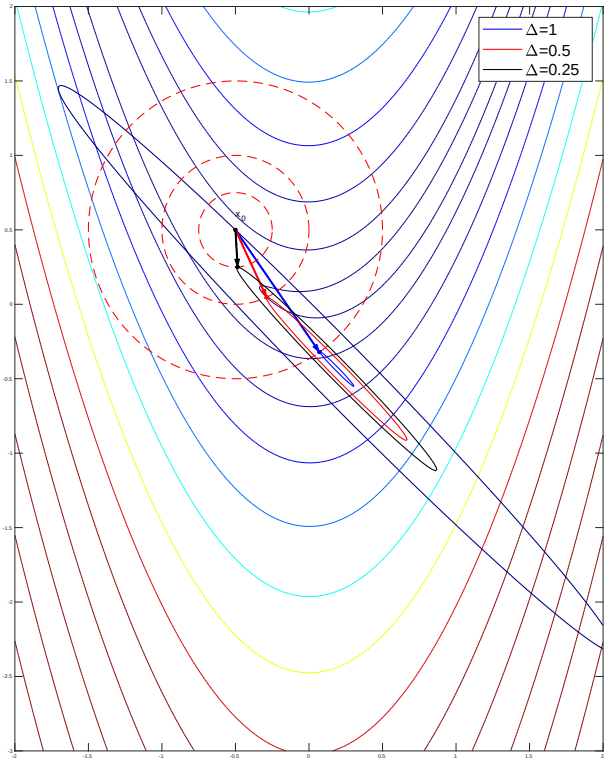

Figura 4: Passos minimizando o modelo quadrático dentro de diferentes regiões de confiança para o caso da função de Rosenbrock.

O método do ponto de Cauchy consiste em minimizar o modelo definido em (53) na direção de p_k^*, isto é, encontrar

$$\alpha_k = \arg\min_{\alpha \geq 0} m_k(\alpha p_k^*), \qquad \|\alpha p_k^*\| \leq \Delta_k, \tag{59}$$

e tomar $p_k^C = \alpha_k p_k^*$. O vetor p_k^C é chamado de passo de Cauchy.

É possível obter p_k^C explicitamente. Para tanto note que, como $-\nabla f(x_k)$ é direção de descida para f a partir do ponto x_k, então os valores do hiperplano $f(x_k) + \nabla f(x_k)^T p$ diminui infinitamente quando seguimos na direção de $-\nabla f(x_k)$. Assim

$$p_k^* = -\frac{\Delta_k}{\|\nabla f(x_k)\|} \nabla f(x_k). \tag{60}$$

Para obter α_k, note que

$$m_k(\alpha p_k^*) = f(x_k) + \left(\nabla f(x_k)^T p_k^*\right)\alpha + \left(\frac{1}{2} {p_k^*}^T A(x_k) p_k^*\right)\alpha^2, \qquad \|\alpha p_k^*\| \leq \Delta_k, \tag{61}$$

ou seja, $m_k(\alpha p_k^*)$ é uma função quadrática em α. Dessa forma, para resolver (49) devemos avaliar o valor de ${p_k^*}^T A(x_k)p_k^*$. Se ${p_k^*}^T A(x_k)p_k^* \leq 0$, então $m_k(\alpha p_k^*)$ decresce infinitamente, e portanto devemos tomar $\alpha_k = 1$. Se ${p_k^*}^T A(x_k)p_k^* > 0$, a parábola acima tem um único ponto de mínimo, que é dado por

$$\alpha_k^* = -\frac{\nabla f(x_k)^T p_k^*}{{p_k^*}^T A(x_k)p_k^*} = \frac{\|\nabla f(x_k)\|^3}{\Delta_k \nabla f(x_k)^T A(x_k)\nabla f(x_k)}.$$

Entretanto, devemos sempre tomar o passo dentro da região de confiança, isto é, devemos satisfazer a condição $\alpha_k \leq 1$. Sendo assim, tomamos

$$\alpha_k = \min\{\alpha_k^*, 1\}.$$

Portanto, abrangendo ambos os casos, temos

$$\alpha_k = \begin{cases} 1, & \text{se } {p_k^*}^T A(x_k)p_k^* \leq 0 \\ \min\{\alpha_k^*, 1\}, & \text{se } {p_k^*}^T A(x_k)p_k^* > 0 \end{cases}, \tag{62}$$

e tomamos $p_k^C = \alpha_k p_k^*$.

Note que o método do ponto de Cauchy nada mais é que o método de máxima descida, onde em cada iteração controlamos o tamanho do passo para que esteja contido dentro da região de confiança. Porém o método de máxima descida apresenta um desempenho ruim, mesmo se o passo ótimo for tomado em cada iteração [36].

Apesar disso, o passo de Cauchy tem grande importância teórica para os métodos de regiões de confiança, pois se um minimizador aproximado do subproblema (54) apresenta redução na função modelo maior ou igual a redução obtida pelo passo de Cauchy, então, sob certas condições, somos capazes de garantir a convergência global do método, como veremos mais adiante.

O próximo método que mostraremos, combina as técnicas de máxima descida com minimizador global do modelo (54).

2.5.2 Método de Dogleg

Para explicitar esse método, vamos denotar por $p_k(\Delta_k)$ o minimizador global de (54) e assumir que $A(x_k)$ é positiva definida para todo $k \geq 0$.

Observamos que, se o raio da região de confiança é pequeno, então

$$p(\Delta_k) \approx -\frac{\Delta_k}{\|\nabla f(x_k)\|}\nabla f(x_k).$$

Por outro lado, se o raio da região de confiança é grande, então $p(\Delta_k) \approx p_k^A$, onde

$$p_k^A = -A(x_k)^{-1}\nabla f(x_k). \tag{63}$$

De maneira análoga ao que foi feito para se obter o passo de Cauchy, podemos minimizar o modelo quadrático (53) na direção de $-\nabla f(x_k)$ (desconsiderando qualquer limitação do tamanho do passo), e temos

$$p_k^B = -\frac{\|\nabla f(x_k)\|^2}{\nabla f(x_k)^T A(x_k)\nabla f(x_k)}\nabla f(x_k). \tag{64}$$

Assim, de maneira intuitiva o minimizador global de (54) é uma combinação entre máxima descida e o minimizador irrestrito do modelo quadrático. A medida que as regiões de confiança vão sendo atualizadas, as soluções $p(\Delta_k)$ desenham uma trajetória, que o método de Dogleg aproxima construindo uma trajetória combinando a máxima descida com minimizador irrestrito do modelo quadrático da seguinte forma:

$$\tilde{p}_k(\tau) = \begin{cases} \tau p_k^A, & \text{se } 0 \leq \tau \leq 1 \\ p_k^A + (\tau - 1)(p_k^B - p_k^A), & \text{se } 1 \leq \tau \leq 2. \end{cases} \tag{65}$$

O método de Dogleg consiste em minimizar $m(\tilde{p}_k(\tau))$ em τ dentro da região de confiança, como é mostrado no próximo lema.

Lema 2.8. *Se A_k é positiva definida, então $\|\tilde{p}_k(\tau)\|$ é crescente em τ e $m(\tilde{p}_k(\tau))$ é uma função decrescente em τ, onde m é o modelo quadrático definido em (53).*

Demonstração. Ver [32]. □

Assim, se $\|p_k^A\| \geq \Delta_k$, então como $\|\tilde{p}_k(\tau)\|$ é crescente em τ, então existe τ_k tal que $\|\tilde{p}_k(\tau_k)\| = \Delta_k$. Além disso, como $m(\tilde{p}_k(\tau))$ é uma função decrescente em τ, então o valor de $m(\tilde{p}_k(\tau_k))$ é o menor possível dentro da região de confiança.

Se $\|p_k^A\| \leq \Delta_k$, então tomamos $\tau_k = 2$, ou seja, fazemos o passo de Dogleg como sendo igual a p_k^A.

Note que o valor de τ_k, no caso em que $\left\|p_k^A\right\| \geq \Delta_k$ pode ser obtido explicitamente, pois a equação

$$\Delta_k^2 - \left\|p_k^A + (\tau - 1)(p_k^B - p_k^A)\right\|^2 = 0$$

é simplesmente uma equação do segundo grau em τ. Logo, sabemos que tem apenas uma raiz real positiva em vista do lema acima.

2.5.3 Teoremas de convergência

Como mencionamos no fim da Subseção 2.5.1, os métodos de regiões de confiança tem sua convergência global para pontos estacionários (pontos de mínimo ou de sela) garantida desde que se tenha uma redução miníma no modelo quadrático (53). Tal redução deve ser ao menos a redução obtida pelo ponto de Cauchy. A desigualdade abaixo descreve formalmente essa redução:

$$m_k(0) - m_k(p_k) \geq c_1 \|\nabla f(x_k)\| \min\left\{\Delta_k, \frac{\|\nabla f(x_k)\|}{\|A(x_k)\|}\right\}, \tag{66}$$

onde $c_1 \in (0, 1]$.

Lema 2.9. *O passo de Cauchy satisfaz a desigualdade (66) com* $c_1 = \frac{1}{2}$.

Demonstração. Ver [36]. □

Definido $\mathcal{S} = \{x \in \mathbb{R}^: f(x) \leq f(x_0)\}$, temos os seguintes teorema de convergência:

Teorema 2.10. *Seja* $\eta = 0$ *no algoritmo 6. Suponha que* $\|A_k\| \leq \beta$ *para alguma constante* $\beta > 0$ *em* $\mathcal{S}$, f *seja* C^2, *limitada inferiormente em* $\mathcal{S}$ *e que todas as soluções aproximadas de (54) satisfaz a desigualdade (66) para alguma positiva* c_1. *Temos então*

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

Demonstração. Ver [36]. □

Teorema 2.11. *Seja* $\eta \in \left(0, \frac{1}{4}\right)$ *no algoritmo 6. Suponha que* $\|A_k\| \leq \beta$ *para alguma constante* $\beta > 0$ *em* $\mathcal{S}$, f *seja* C^2, *limitada inferiormente em* $\mathcal{S}$ *e que todas as soluções aproximadas de (54) satisfaz desigualdade (66) para alguma positiva* c_1. *Temos então*

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

Demonstração. Ver [36]. □

Note que a primeira consequência imediata desses resultados é que o método do ponto de Cauchy tem convergência global para pontos estacionários.

A segunda é que o método de Dogleg também converge globalmente se considerarmos que A_k é positiva definida para todo $k \geq 0$ nos teoremas acima. Para ver isso basta mostrar que o passo de Dogleg satisfaz a desigualdade (66) para algum $c_1 \in (0, 1]$.

De fato, isso ocorre, pois se $\left\|p_k^A\right\| \geq \Delta_k$, então o passo de Dogleg é o passo de minimização global do modelo, e portanto satisfaz trivialmente (66). Se $\left\|p_k^A\right\| < \Delta_k$, então o passo de Dogleg é o passo de Cauchy mais $(\tau - 1)(p_k^A - p_k^B)$. Como vimos, o modelo decresce em τ, assim temos uma redução no modelo maior ou igual a redução prevista pelo passo de Cauchy.

No início deste capítulo mostramos uma forma de interpretar o método de LM como um método de regiões de confiança. Porém em vista do teorema (2.7) há uma outra maneira de fazer isso, para tanto, resolva (54) com modelo dado por (11) usando o teorema (2.7) e atualize o raio da região de confiança com o Algoritmo 6. Dessa forma, também teremos a convergência global garantida, já que a desigualdade (66) é evidentemente satisfeita, pois estamos tomando efetivamente o minimizador global do modelo. Além disso $\left\|J(x_k)^T J(x_k)\right\|$ é limitado em $\mathcal{S}$ para todo $k \geq 0$, pois a jacobiana do vetor de resíduos é contínuo.

Dimensionando o método de regiões de confiança

Como mencionamos na seção anterior, para o caso de problemas de quadrados mínimos, há casos onde a função objetivo f pode ser muito sensível a pequenas alterações em algumas de suas variáveis e pouco em outras, isto pode levar a curvas de nível da função tenderem a elipses altamente excêntricas nas proximidades de um ponto de mínimo.

Algoritmos que não levam em consideração esse aspecto da função objetivo pode ter um mau desempenho. Além disso, o modelo m_k tende a ser uma aproximação mais confiável de f ao longo de direções nas quais f está mudando mais lentamente, como comentamos na seção anterior, para problemas de mínimos quadrados.

Como o formato da nossa região de confiança deve ser de tal forma que a qualidade do modelo é aproximadamente o mesmo em todos os pontos da fronteira dessa região,

então surge, naturalmente, a razão de se considerar regiões de confiança elípticas em que os eixos são menores nas direções mais sensíveis e maiores nas direções menos sensíveis.

As regiões de confiança elípticas podem ser definidas da forma

$$\|Dp\| \leq \Delta,$$

onde D é uma matriz diagonal com elementos diagonais positivos, conduzindo ao seguinte subproblema de região confiável elíptico.

$$\min_{p\in\mathbb{R}^n} m_k(x) = f(x) + p^T\nabla f(x_k) + \frac{1}{2}p^T A(x_k)p, \quad \text{sujeito à} \quad \|D_k p\| \leq \Delta_k. \tag{67}$$

Quando f é sensível a i-ésima componente x_i, definimos que o valor correspondente ao elemento diagonal d_{ii} de D deve ser grande, enquanto que d_{ii} é menor para componentes menos sensíveis.

O problema de regiões de confiança elípticas pode também ser obtido reescalando o caso esférico. Para isso, realizamos a transformação

$$\hat{p} = D_k p,$$

substituindo em (53), temos

$$\min_{p\in\mathbb{R}^n} m_k(x) = f(x) + g_k^T D_k^{-1}\hat{p} + \frac{1}{2}\hat{p}^T(D_k^{-1})^T A(x_k)D_k^{-1}\hat{p}, \quad \text{sujeito à} \quad \|\hat{p}\| \leq \Delta_k. \tag{68}$$

Dessa forma, a teoria dos métodos de regiões de confiança com essa modificação torna-se a mesma da anterior. Para ver isso, basta definir $\hat{g}_k = D^{-1}g_k$, $\hat{B}_k = (D^{-1})^T A(x_k)D^{-1}$. Assim, teremos

$$\min_{p\in\mathbb{R}^n} \hat{m}_k(x) = f(x) + \hat{g}_k\hat{p} + \frac{1}{2}\hat{p}^T\hat{A}(x_k)\hat{p}, \quad \text{sujeito à} \quad \|\hat{p}\| \leq \Delta_k, \tag{69}$$

onde fica evidente que retornamos a teoria padrão dos métodos de regiões de confiança esféricas.

Assim, de maneira análoga ao caso do método de LM, podemos constatar que, aplicando os teoremas de convergência ao modelo (30), temos pelo Teorema 2.10 que

$\liminf_{k\to\infty} \|\hat{g}_k\| = 0$, e pelo Teorema 2.11, que $\lim_{k\to\infty} \|\hat{g}_k\| = 0$, então supondo que $\|D_k^{-1}\| \not\to 0$ e que $(D_k^{-1})_{k\geq 0}$ é limitada, temos $\liminf_{k\to\infty} \|g_k\| = 0$ e $\lim_{k\to\infty} \|g_k\| = 0$.

Portanto, a teoria dos métodos de regiões de confiança com essa modificação torna-se a mesma da anterior.

3 MÉTODO DE LEVENBERG-MARQUARDT COM CORREÇÃO DE SEGUNDA ORDEM

Agora vamos propor um método novo, que consiste em obter uma correção para método de LM. Para essa correção utilizamos as derivadas de segunda ordem dos resíduos combinado com técnicas tradicionais do método de LM.

Nas próximas seções deste capítulo vamos mostrar como o método é construído bem como provar dois teoremas de convergência. Além disso, como uma aplicação teórica do nosso método, vamos demonstrar que é possível convergir para o mínimo global da função Rosenbrock em apenas um passo, independente no ponto inicial escolhido. Também vamos discutir algumas modificações possíveis, e discuti-las com alguns exemplos.

3.1 FUNDAMENTAÇÃO TEÓRICA

Tanto o método de Gauss-Newton como o método de Levenberg-Marquardt tem por base de sua formulação, a aproximação linear do vetor de resíduos, isto é,

$$F(x+p) \approx \frac{1}{2}\left\|r(x)+J(x)p\right\|^2.$$

Seguindo essa ideia, vamos aproximar f em um ponto x usando expansão de Taylor de segunda de ordem do vetor r em torno de x, obtendo

$$F(x+p) \approx \frac{1}{2}\left\|r(x)+J(x)p+\frac{1}{2}K(x)(p,p)\right\|^2, \tag{70}$$

onde a aplicação bilinear $K(x) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ é a segunda derivada de r em x e $K(x)(p,p)$ denota a aplicação de $K(x)$ em (p,p). Observe que, dados $(u,v) \in \mathbb{R}^n \times \mathbb{R}^n$, temos

$$K(x)(u,v) = \begin{bmatrix} u^T \nabla^2 r_1(x) v \\ \vdots \\ u^T \nabla^2 r_m(x) v \end{bmatrix}.$$

Assim, $K(x)$ é simétrica se as hessianas dos resíduos forem simétricos. Como no nosso caso isso ocorre, pois estamos supondo que os resíduos são funções de classe C^2, então $K(x)$ é simétrico.

A fim de propiciar maior clareza nas fórmulas, vamos realizar as seguintes convenções: $r(x) = r$, $J(x) = J$ e $K(x) = K$. Note que não há risco de confusão nessa notação, haja visto que todo processo de construção do método é feita sobre cada iteração.

Sabemos que, dado $\lambda > 0$ o método de LM consiste em resolver em cada iteração o um subproblema da forma

$$\min_{p \in \mathbb{R}^n} m(p),$$

onde

$$m(p) = \frac{1}{2} \|r + Jp\|^2 + \frac{1}{2} \lambda \|p\|^2.$$

Inspirado no caso acima, introduzindo o termo de segunda ordem, temos

$$\begin{aligned} L(p) &= \frac{1}{2} \left\| r + Jp + \frac{1}{2} K(p,p) \right\|^2 + \frac{1}{2} \lambda \|p\|^2 \\ &= \frac{1}{2} \|r + Jp\|^2 + \frac{1}{2} \lambda \|p\|^2 + \frac{1}{2} (r + Jp)^T K(p,p) + \mathcal{O}(\|p\|^4) \\ &= m(p) + \frac{1}{2} (r + Jp)^T K(p,p) + \mathcal{O}(\|p\|^4). \end{aligned}$$

Com isso, definimos

$$M(p) = m(p) + \frac{1}{2} (r + Jp)^T K(p,p). \tag{71}$$

Seguindo a ideia do método de LM, queremos resolver

$$\min_{p \in \mathbb{R}^n} M(p). \tag{72}$$

Dessa forma, derivando (71) com respeito a p e igualando a zero temos a seguinte equação normal

$$J^T r + (J^T J + \lambda I)p + \frac{1}{2} J^T K(p, p) + K(p, \cdot)^T (r + Jp) = 0 \tag{73}$$

onde

$$K(p, \cdot) = \begin{bmatrix} p^T \nabla^2 r_1(x) \\ \vdots \\ p^T \nabla^2 r_m(x) \end{bmatrix}.$$

Em contraste com o método de LM, a equação para a obtenção do ponto crítico nesse caso é não linear e portanto, não é possível isolar p e encontrá-lo através da resolução de um sistema linear. A fim de contornar esse entrave, vamos realizar uma aproximação linear conveniente de a expressão.

Primeiro vamos começar definindo a função G como sendo

$$G(p) = J^T r + (J^T J + \lambda I)p + \frac{1}{2} J^T K(p, p) + K(p, \cdot)^T (r + Jp)$$

derivando em relação a p temos

$$G'(p) = J^T J + \lambda I + J^T K(p, \cdot) + K(p, \cdot)^T J + K(\cdot, \cdot)^T (r + Jp),$$

onde, dado $u = (u_1, \ldots, u_m)^T \in \mathbb{R}^m$, $K(\cdot, \cdot)^T u = \sum_{i=1}^m u_i \nabla^2 r_i$, é a derivada de $K(p, \cdot)^T u$ em relação a p. Fazendo a expansão de Taylor em G de primeira ordem em p_{lm}, temos

$$G(p_{lm} + p) \approx G(p_{lm}) + G'(p_{lm})p.$$

Convenientemente, é interessante aproximar

$$G'(p_{lm}) \approx J^T J + \lambda I, \tag{74}$$

pois, como $(J^T J + \lambda I)$ é inversível, então existe $p_c \in \mathbb{R}^n$ tal que

$$G(p_{lm}) + (J^T J + \lambda I)p_c = 0,$$

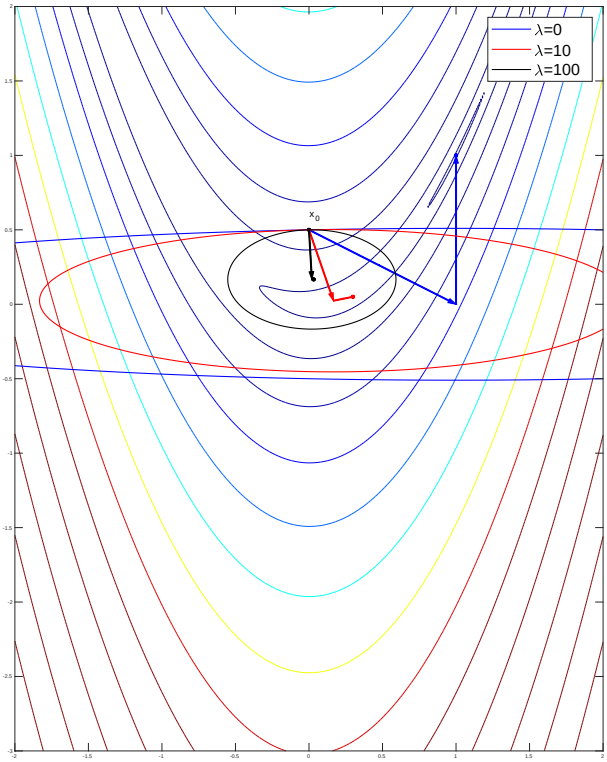

Figura 5: Passos do método LMCS para a função de Rosenbrock para diferentes valores de λ.

assim

$$(J^T J + \lambda I)p_c = -G(p_{lm}). \tag{75}$$

Usando

$$(J^T J + \lambda I)p_{lm} = -J^T r, \tag{76}$$

podemos reescrever a equação (75) como

$$(J^T J + \lambda I)p_c = -\frac{1}{2}J^T K(p_{lm}, p_{lm}) - K(p_{lm}, \cdot)^T (r + J p_{lm}). \tag{77}$$

Em linhas gerais, encontramos uma solução aproximada para (73) via resolução de um sistema linear com mesma matriz do sistema de LM, o que nos permite realizar apenas uma fatoração de matriz a cada iteração do método para encontrar p_{lm} e p_c.

Definimos assim $h_c = p_{lm} + p_c$, como sendo o passo obtido em cada iteração do método. Observar que h_c é uma aproximação de um ponto crítico do modelo M.

Chamaremos o método descrito acima de Levenberg-Marquardt com Correção de Segunda Ordem (LMCS).

Na Figura 5, é possível visualizar o passo p_{lm} partindo do ponto x_0 e p_c do ponto $x_0 + p_{lm}$, compondo assim h_c, para diferentes valores de λ variando de acordo com a legenda. Mostrarmos apenas o nível $F(x_0)$ das curvas de nível dos modelos quadráticos de LM, para $\lambda = 0$ (em azul), $\lambda = 10$ (em vermelho) e $\lambda = 100$ (em preto).

Exemplo: otimizando a função de Rosenbrock em um passo

Considerarmos a função de Rosenbrock, (35). Calculando p_{lm} e p_c através da resolução dos sistemas (76) e (77), respectivamente, com $\lambda = 0$, então, independente do ponto inicial (x, y), teremos $(x, y) + p_{lm} + p_c = (1, 1)$, que é o ponto de mínimo global da função de Rosenbrock.

De fato, observe que se $\lambda = 0$, então o passo de LM se torna o passo de GN, isto é, na notação do Capítulo 2, $p_{lm} = p_{gn}$.

A jacobiana do vetor de resíduos é dada por

$$J(x, y) = \begin{pmatrix} -\sqrt{2} & 0 \\ -20\sqrt{2}x & 10\sqrt{2} \end{pmatrix},$$

observe que $J(x, y)$ é inversível. Assim

$$\begin{aligned} p_{gn} &= -(J(x, y)^T J(x, y))^{-1} J(x, y)^T r(x, y) \\ &= -J(x, y)^{-1} r(x, y), \end{aligned}$$

donde concluímos que $J(x, y)p_{gn} - r(x, y) = 0$, logo

$$p_c = -\frac{1}{2} J(x, y)^{-1} K(x, y)(p_{gn}, p_{gn}),$$

e assim

$$\begin{aligned} h_c &= p_{gn} + p_c \\ &= -J(x, y)^{-1} r(x, y) - \frac{1}{2} J(x, y)^{-1} K(x, y)(p_{gn}, p_{gn}) \\ &= -J(x, y)^{-1} (r(x, y) + \frac{1}{2} K(x, y)(p_{gn}, p_{gn})). \end{aligned}$$

A inversa de $J(x,y)$ é dada por

$$J(x,y)^{-1} = \begin{pmatrix} -1/\sqrt{2} & 0 \\ -\sqrt{2}x & 1/10\sqrt{2} \end{pmatrix},$$

e a derivada de segunda ordem do vetor de resíduos aplicada em (p_{gn}, p_{gn}) é

$$K(x,y)(p_{gn}, p_{gn}) = \begin{pmatrix} 0 \\ 20\sqrt{2}(1-x)^2 \end{pmatrix}.$$

Logo,

$$\begin{aligned} h_c &= -\begin{pmatrix} -1/\sqrt{2} & 0 \\ -\sqrt{2}x & 1/10\sqrt{2} \end{pmatrix} \left(\begin{pmatrix} \sqrt{2}(1-x) \\ 10\sqrt{2}(y-x^2) \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 0 \\ -20\sqrt{2}(1-x)^2 \end{pmatrix} \right) \\ &= -\begin{pmatrix} -1/\sqrt{2} & 0 \\ -\sqrt{2}x & 1/10\sqrt{2} \end{pmatrix} \begin{pmatrix} \sqrt{2}(1-x) \\ 10\sqrt{2}(y-x^2) - 10\sqrt{2}(1-x)^2 \end{pmatrix} \\ &= -\begin{pmatrix} x-1 \\ -2x(1-x) + y - x^2 - (1-x)^2 \end{pmatrix} = -\begin{pmatrix} x-1 \\ y-1 \end{pmatrix} = \begin{pmatrix} 1-x \\ 1-y \end{pmatrix}, \end{aligned}$$

Portanto,

$$\begin{pmatrix} x \\ y \end{pmatrix} + h_c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Dessa forma, mostramos que para o caso da função de Rosenbrock, o método proposto converge para o mínimo global em um passo, independente do ponto inicial escolhido. Isto é inesperado, pois além da função de Rosenbrok ser de quarta ordem, a formulação do passo h_c é construído segundo diversas aproximações, e no entanto obtemos um resultado exato.

Observe que, tanto o passo de LM quanto a correção de segunda ordem são sensíveis ao ajuste do parâmetro λ, sendo que ambas diminuem conforme aumentamos λ, de tal forma que, quando $\lambda \to \infty$, temos

$$p_{lm} \to -\frac{1}{\lambda} J^T r$$

e

$$p_c \to \left(-\frac{1}{2\lambda^3} J^T K(v,v) - \frac{1}{\lambda^2} K(v,\cdot)^T r - \frac{1}{\lambda^3} K(v,\cdot)^T J v \right),$$

onde $v = \lambda p_{lm}$. Note que, quando $\lambda \to \infty$ temos, $v \to -J^T r$, então pela continuidade em v, temos que $K(v, v)$, $K(v, \cdot)^T r$ e $K(v, \cdot)^T Jv$ são limitados em v, logo, para λ suficientemente grande, $\|p_{lm}\| > \|p_c\|$.

Por outro lado, se $\lambda \to 0$, temos

$$p_{lm} \to -(J^T J)^{-1} J^T r = p_{gn}$$

e

$$p_c \to -\frac{1}{2} J^T K(p_{gn}, p_{gn}),$$

onde, como já dito anteriormente, p_{gn} é passo de Gauss-Newton (GN) (passo de p_{lm} com $\lambda = 0$).

Observe também que a relação entre λ e a norma de p_{lm}, obtida via Proposição (2.6), também pode ser aplicada para p_c, já que a matriz do sistema associado a p_c é o mesma do sistema de LM. Assim, temos que $\|p_c\|$ é não crescente em $\lambda > 0$.

Em vista dessas observações, vemos que é razoável tratar o parâmetro λ do mesmo modo no método de LM. Assim, vamos nos apoiar em (21), isto é, a mesma atualização de λ do algoritmo de LM, para atualizar λ para essa nova modificação, porém vamos redefinir ρ da seguinte maneira

$$\rho = \frac{F(x) - F(x + h_c)}{M(0) - M(h_c)}. \tag{78}$$

Observe que, como h_c é uma aproximação do ponto crítico de M, então não é necessariamente verdade que

$$M(0) - M(h_c) > 0.$$

Dessa forma, estamos sujeitos a aceitar passos onde o valor da função objetivo aumenta. Neste trabalho a ideia é aceitar essas direções, o motivo pelo qual o fazemos tem respaldo numérico. Entretanto, para que sejam aceitos essas direções é preciso alguns cuidados.

Fixado no inicio do algoritmo um parâmetro $\epsilon > 0$, avaliamos se $\|\nabla F(x + h_c)\| < \epsilon$, caso afirmativo, então recusamos o passo, caso contrário, aceitamos. Fazemos essa avaliação, pois estamos aumentando o valor da função objetivo, e assim, não desejamos estacionar em um ponto que seja, possivelmente, de máximo local.

Definimos o parâmetro PC que corresponde ao número máximo de aumentos consecutivos permitidos, pois se estamos aceitando muitas direções consecutivas de aumento, é natural supor que o desempenho será prejudicado.

Além do parâmetro PC, também introduzimos PT, que limita o número de passos totais que aumentam o valor da função. Do ponto de vista teórico, esse parâmetro é utilizado a fim de garantir a convergência, no entanto, do ponto de vista do desempenho do método, também é razoável supor que a quantidade total de passos de aumento da função sejam limitados.

Observe que, ao aceitar direções que fazem o valor de F aumentar, estamos sujeitos a convergir para um mínimo local x^*, tal que $F(x^*) \geq F(x_0)$. No entanto, a ideia central para o uso dessas direções de aumento no valor de F, é a possibilidade de poder transpor rapidamente uma região de grande aumento de F que depois cumina em uma descida para regiões mais próximas de um mínimo global.

Um questionamento que poderia surgir nesse momento é se é sempre possível garantir que exista um λ, suficientemente grande, tal que $M(0) - M(h_c) > 0$ e $F(x) - F(x+h_c) > 0$ simultaneamente, ou seja, se é sempre possível obter passos onde o valor da função objetivo é reduzido. De fato, como consequência da Proposição 2.3, temos $\|p_{lm}\| = \mathcal{O}(1/\lambda)$ e $\|p_c\| = \mathcal{O}(1/\lambda^2)$, logo

$$\begin{aligned} M(h_c) &= F(x) + p_{lm}^T \nabla F(x) + \frac{1}{2} p_{lm}^T (J(x)^T J(x) + \lambda I) p_{lm} + \\ &+ p_c^T \nabla F(x) + \frac{1}{2} p_c^T (J(x)^T J(x) + \lambda I) p_c + p_{lm}^T (J(x)^T J(x) + \lambda I) p_c + \\ &+ \frac{1}{2} (r(x) + J(x) p_{lm})^T K(x)(p_{lm}, p_{lm}) + \frac{1}{2} (r(x) + J(x) p_c)^T K(x)(p_c, p_c) \\ &= m(p_{lm}) + \mathcal{O}(1/\lambda^2). \end{aligned}$$

Como $M(0) = m(0)$, então $M(0) - M(h_c) = m(0) - m(p_{lm}) + \mathcal{O}(1/\lambda^2)$, assim existe $\lambda_1 > 0$ tal que $M(0) - M(h_c) > 0$.

Temos também que

$$\begin{aligned} F(x + h_c) &= F(x) + \nabla F(x)^T h_c + \mathcal{O}(\|h_c\|^2) \\ &= F(x) + \nabla F(x)^T p_{lm} + \mathcal{O}(1/\lambda^2), \end{aligned}$$

como p_{lm} é direção de descida para F no ponto x, então existe $\lambda_2 > 0$ suficientemente grande tal que $F(x) - F(x + h_c) > 0$. Logo, tomando $\lambda \geq \max\{\lambda_1, \lambda_2\}$, temos que $F(x) - F(x + h_c) > 0$ e $M(0) - M(h_c) > 0$ simultaneamente.

De maneira análoga ao caso de LM, podemos formular o método em termos de regiões de confiança, o que é mais conveniente para provar a convergência.

Algoritmo 7: MÉTODO DE LMCS (REGIÕES DE CONFIANÇA)

Entrada: Dados $\Delta_0 > 0$, $PT \in \mathbb{N}$, $PC \in \mathbb{N}$, $\eta \in \left[0, \frac{1}{4}\right)$, $\epsilon > 0$, $l_t = 0$, $l_c = 0$ e x_0 um ponto inicial

Passo 1 : Calculo p_{lm} e p_c via (76) e (77), faça $h_c = p_{lm} + p_c$ e avalie ρ_k usando (78)

Passo 2 : Obtenha Δ_{k+1} usando o Algoritmo 6 com ρ_k obtido no passo anterior

Passo 3 : Se $\rho_k > \eta$ e $M_k(0) - M_k(h_c) > 0$, faça

$$x_{k+1} = x_k + h_c, \quad k := k + 1 \quad \text{e retorne ao } \textbf{Passo 1}$$

Passo 4 : Se $\rho_k > \eta$, $M_k(0) - M_k(h_c) < 0$, $l_c \leq PC$ e $l_t \leq PT$, faça

$$x_{k+1} = x_k + h_c, \quad l_c = l_c + 1, \quad l_t = l_t + 1, \quad k := k + 1 \quad \text{e retorne ao } \textbf{Passo 1}$$

Passo 5 : Se $\rho_k > \eta$, $M_k(0) - M_k(h_c) < 0$ e $\|\nabla F(x_k + h_c)\| < \epsilon$, faça

$$x_{k+1} = x_k, \quad k := k + 1 \quad \text{e retorne ao } \textbf{Passo 1}$$

Passo 6 : Se $\rho_k > \eta$, $M_k(0) - M_k(h_c) < 0$ e ($l_c > PC$ ou $l_t > PT$), faça

$$x_{k+1} = x_k, \quad l_c = 0, \quad k := k + 1 \quad \text{e retorne ao } \textbf{Passo 1}$$

Passo 7 : Se $\rho_k \leq \eta$, faça

$$x_{k+1} = x_k, \quad k := k + 1 \quad \text{e retorne ao } \textbf{Passo 1}$$

Mas do ponto de vista prático, implementamos o método acima usando a relação inversa que existe entre λ e Δ, para tanto, basta substituímos a obtenção de Δ_{k+1} no segundo passo, pela atualização de λ feita no método de LM, (21), utilizando ρ_k definido em (78). Convém aqui fazer essa observação, pois é dessa maneira que vamos realizar nossos testes numéricos.

Na próxima seção, propor e provar os teoremas de convergência acerca do método proposto.

3.2 TEOREMAS DE CONVERGÊNCIA

Na Seção 2.5.3 vimos que a convergência global dos métodos de regiões de confiança depende da solução aproximada obtida para minimização da função do modelo m, sendo necessário essas soluções reduzam o maior ou igual a redução do ponto de Cauchy no modelo m, isto é,

$$m(0) - m(p) \geq c_1 \|\nabla f(x)\| \min\left\{\Delta, \frac{\|\nabla f(x)\|}{\|A(x)\|}\right\}, \tag{79}$$

para alguma constante $c_1 \in (0, 1]$. Para os casos dos métodos de Dogleg, LM, LMM e LMMA, essa desigualdade é sempre satisfeita.

No entanto, essa desigualdade não é necessariamente verdadeira no caso em que o passo é h_c e o modelo é M, pois h_c é construído de forma a ser um ponto crítico de M e não necessariamente uma ponto de mínimo. Porém, em nossos resultados numéricos, observamos que essa desigualdade é satisfeita em grande parte das iterações, sendo assim vamos assumi-la em nosso teorema de convergência.

Como em espaços de dimensão finita todas as normas são equivalentes, então, a propriedade $\|J(x)p\| \leq c_1 \|J(x)\| \|p\|$ com $c_1 > 0$ constante, é válida seja qual forem as normas escolhidas. Analogamente, $\|K(x)(p,p)\| \leq c_2 \|K(x)\| \|p\|^2$ com $c_2 > 0$ constante. Assim, para os nossos propósitos nas demonstrações abaixo, podemos tomar a norma canônica de operadores sem perda de generalidade.

A demonstração do teorema abaixo segue uma ideia parecida com a prova do teorema de convergência dos métodos de regiões de confiança para o caso em que $\eta \in [0, 1/4)$ feita em Nocedal, [36].

Teorema 3.1. ***(Convergência do Método LMCS)*** *Seja F definida em (1) e considere a sequência $(x_k)_{k\geq 0}$ dada pelo Algoritmo 7. Se existe $c_1 \in (0, 1]$ tal que*

$$M_k(0) - M_k(h_c) \geq c_1 \|\nabla F(x_k)\| \min\left\{\Delta_k, \frac{\|\nabla F(x_k)\|}{\|J(x_k)^T J(x_k)\|}\right\},$$

para $k \geq k^*$, onde k^* é a iteração a partir da qual aceitamos a direção se $M(0) - M(h_c) > 0$. Suponha que $\mathcal{S} := \{x \in \mathbb{R}^n : F(x) \leq F(x_{k^*})\}$ é limitado, , então $\liminf_{k\to\infty} \|\nabla F(x_k)\| = 0$.

Demonstração. Começamos notando que, como $\mathcal{S}$ é compacto e f é C^2, então existem $\beta > 0$ e $\omega > 0$ tais que $\left\|J(x_k)^T J(x_k)\right\| < \beta$ e $\|K(x_k)\| < \omega$, para todo $k \geq 0$.

Além disso, o fato de F ser de classe C^2 implica que $\nabla^2 F$ é contínua, logo limitada em $\mathcal{S}$ e portanto ∇F é lipschitz em $\mathcal{S}$.

Também temos que, para todo k,

$$|\rho_k - 1| = \left| \frac{M_k(h_c) - F(x_k + h_c)}{M_k(0) - M_k(h_c)} \right| \tag{80}$$

Pelo teorema de Taylor segue que

$$F(x_k + h_c) = F(x_k) + \nabla F(x_k)^T h_c + \int_0^1 (\nabla F(x_k + t h_c) - \nabla F(x_k))^T h_c \, dt. \tag{81}$$

Definimos

$$W_k(h_c) = \frac{1}{2} h_c^T J(x_k)^T J(x_k) h_c + \frac{1}{2} (r(x_k) + J(x_k) h_c)^T K(x_k)(h_c, h_c)$$

e temos

$$|M_k(h_c) - F(x_k + h_c)| \quad = \quad \left| W_k(h_c) - \int_0^1 (\nabla F(x_k + t h_c) - \nabla F(x_k))^T h_c \, dt \right|. \tag{82}$$

Note que

$$\begin{aligned} |W_k(h_c)| &\leq \frac{1}{2}\beta \|h_c\|^2 + \frac{1}{2}\Big(\|r(x_k)\| + \sqrt{\beta}\|h_c\| \Big)\omega \|h_c\|^2 && (83) \\ &\leq \frac{1}{2}\beta \|h_c\|^2 + \frac{1}{2}\Big(\|r(x_{k^*})\| + \sqrt{\beta}\|h_c\| \Big)\omega \|h_c\|^2, && (84) \end{aligned}$$

onde usamos que $\|r(x_k)\| \leq \|r(x_{k^*})\|$, para todo $k \geq k^*$, pois a partir da iteração k^* aceitamos apenas passos h_c que reduzem o valor da função objetivo. Assim

$$|M_k(h_c) - F(x_k + h_c)| \leq \frac{1}{2}\beta \|h_c\|^2 + \frac{1}{2}\Big(\|r(x_{k^*})\| + \sqrt{\beta}\|h_c\| \Big)\omega \|h_c\|^2 + C \|h_c\|^2, \tag{85}$$

onde $C > 0$ é a constante de lipschitz de ∇F em $\mathcal{S}$.

Suponha por absurdo que não ocorre $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$, logo existe $\epsilon > 0$ e $k_0 \geq k^*$ tais que

$$\|\nabla F(x_k)\| \geq \epsilon,$$

para todo $k \geq k_0$. Sendo assim, temos

$$M_k(0) - M_k(h_c) \geq c_1 \epsilon \min\left\{\Delta_k, \frac{\epsilon}{\beta}\right\}, \tag{86}$$

$k \geq k_0$. Assim, usando (85) e (86) em (80), temos

$$|\rho_k - 1| \leq \frac{\left(\frac{1}{2}\left(\beta + \|r(x_{k^*})\| + \sqrt{\beta}\,\|h_c\|\right)\omega + C\right)\|h_c\|^2}{c_1 \epsilon \min\{\Delta_k, \epsilon/\beta\}}.$$

Definindo

$$T = \frac{1}{2}\omega\beta + \frac{1}{2}\|r(x_{k^*})\|\,\omega + C,$$

seja Δ_k tal que $\Delta_k \leq \Delta$, onde

$$\Delta = \frac{\sqrt{T^2 + c_1\omega\epsilon\sqrt{\beta}} - T}{\omega\sqrt{\beta}},$$

o motivo pelo qual definimos Δ dessa forma será justificado mais adiante. Dessa forma,

$$\Delta \leq \frac{T + \sqrt{c_1\omega\epsilon\sqrt{\beta}} - T}{\omega\sqrt{\beta}} \tag{87}$$

$$= \frac{\sqrt{c_1\omega\epsilon\sqrt{\beta}}}{\omega\sqrt{\beta}} \tag{88}$$

$$= \frac{1}{\sqrt{\omega}}\frac{\sqrt{c_1\epsilon}}{\sqrt[4]{\beta}}, \tag{89}$$

redefinindo

$$\omega \mapsto \max\left\{\omega, \frac{c_1\sqrt{\beta^3}}{\epsilon}\right\}$$

(podemos fazer isso já que esse novo ω ainda limitará K), temos

$$\Delta \leq \frac{\epsilon}{\beta},$$

logo, para todo $\Delta_k \in [0, \Delta]$, temos $\min\left\{\Delta_k, \frac{\epsilon}{\beta}\right\} = \Delta_k$. Dessa forma, segue

$$\begin{aligned} |\rho_k - 1| &\leq \frac{\left(\frac{1}{2}\left(\beta + \|r(x_{k^*})\| + \sqrt{\beta}\,\|h_c\|\right)\omega + C\right)\|h_c\|^2}{c_1\epsilon\Delta_k} \\ &\leq \frac{\left(\frac{1}{2}\left(\beta + \|r(x_{k^*})\| + \sqrt{\beta}\Delta_k\right)\omega + C\right)\Delta_k}{c_1\epsilon} \\ &\leq \frac{\left(\frac{1}{2}\left(\beta + \|r(x_{k^*})\| + \sqrt{\beta}\Delta\right)\omega + C\right)\Delta}{c_1\epsilon} \leq \frac{1}{2}, \end{aligned}$$

só é possível concluir a última desigualdade, pois a definição de Δ feita anteriormente é construída para que isso ocorra. Portanto, $\rho_k > 1/4$, então pelo algoritmo (56) temos $\Delta_{k+1} \geq \Delta_k$ se $\Delta_k \leq \Delta$. Assim, $\Delta_{k+1} \leq \Delta_k$ só no caso em que $\Delta_k > \Delta$. Dessa forma, concluímos que

$$\Delta_k \geq \min\{\Delta_{k_0}, \Delta/4\}, \qquad k \geq k_0. \tag{90}$$

Suponha por absurdo que exista um subconjunto infinito $\mathcal{L} \subset \mathbb{N}$, tal que $\forall l \in \mathcal{L}$ tenha-se $\rho_l \geq 1/4$, logo

$$\begin{aligned} F(x_l) - F(x_l + h_c) &\geq \frac{1}{4}(M_l(0) - M_l(h_c(\lambda_l))) \\ &\geq \frac{1}{4}c_1\epsilon \min\left(\Delta_l, \epsilon/\beta\right). \end{aligned}$$

como f é limitada inferiormente, então

$$\lim_{l\in\mathcal{L},\, l\to\infty} \Delta_l = 0.$$

Porém, em vista de (90) temos uma evidente contradição.

Logo $\mathcal{L}$ tem que ser finito. Dessa forma existe $k_1 \geq k_0$ tal que $\rho_k < 1/4$ para todo $k \geq k_0$. Nesse caso temos que $\Delta_{k+1} = \Delta_k/4$, para todo $k \geq k_1$, logo

$$\lim_{k\to\infty} \Delta_k = 0,$$

o que caracteriza outra contradição em vista de (90).

Portanto $\liminf_{k\to\infty} \|\nabla F(x_k)\| = 0$. □

No teorema acima, o fato do parâmetro η poder assumir o valor nulo faz com que não tenhamos a garantia que a sequência dos gradientes tende a zero.

Assumindo que $\eta \in (0, 1/4)$, podemos provar que a sequência das normas dos gradientes gerados pelo algoritmo tende a zero, como será verificado no Teorema a seguir.

Novamente, a prova do teorema abaixo é análoga à demonstração do teorema de convergência do método de regiões de confiança para o caso $\eta \in (0, 1/4)$ contido em Nocedal, [36].

Teorema 3.2. *Seja F a função definida em (1) e considere a sequência $(x_k)_{k\geq 0}$ dada pelo Algoritmo 7 com $\eta \in (0, 1/4)$. Se existe $c_1 \in (0, 1]$ tal que*

$$M_k(0) - M_k(h_c) \geq c_1 \|\nabla F(x_k)\| \min\left\{\Delta_k, \frac{\|\nabla F(x_k)\|}{\|J(x_k)^T J(x_k)\|}\right\},$$

para $k \geq k^*$, onde k^* é a iteração a partir da qual aceitamos a direção se $M(0) - M(h_c) > 0$. Suponha que $\mathcal{S} := \{x \in \mathbb{R}^n : F(x) \leq F(x_{k^*})\}$ é limitado, , então $\lim_{k\to\infty} \|\nabla F(x_k)\| = 0$.

Demonstração. Suponha por absurdo que $\lim_{k\to\infty} \|\nabla F(x_k)\| \neq 0$. Logo, existe um índice t tal que $\|\nabla F(x_t)\| \neq 0$. Seja $\gamma > 0$ a constante de lipschitz de ∇F, logo temos que

$$\|\nabla F(x) - \nabla F(x_t)\| \leq \gamma \|x - x_t\|.$$

Defina

$$\epsilon = \frac{\|\nabla F(x_t)\|}{2}, \qquad R = \frac{\epsilon}{\gamma}.$$

Assim, se x pertence à bola fechada de raio R centrada em x_t, denotada aqui por $\mathcal{B}[x_t, R]$, então

$$\|\nabla F(x)\| \geq \|\nabla F(x_t)\| - \|\nabla F(x) - \nabla F(x_t)\| \geq \frac{1}{2}\|\nabla F(x_t)\| = \epsilon > 0.$$

Caso toda a sequencia $\{x_k\}_{k\geq t}$ esteja contida em $\mathcal{B}[x_t, R]$, então teremos que

$$\|\nabla F(x_k)\| \geq \epsilon > 0,$$

para todo $k \geq t$. Mas pelo do teorema acima vimos que isso não ocorre, logo infinitos elementos sequência $\{x_k\}_{k\geq t}$ não pertençam à $\mathcal{B}[x_t, R]$.

Seja $l \geq t$ tal que x_{l+1} é o primeiro elemento que não pertence à $\mathcal{B}[x_t, R]$. Como $\|\nabla F(x_k)\| \geq \epsilon > 0$, para todo $k = t, t+1, \ldots, l$, usando (86) temos

$$\begin{aligned} F(x_t) - F(x_{l+1}) &= \sum_{k=t}^{l} \Big(F(x_k) - F(x_{k+1}) \Big) \\ &\geq \sum_{k=l,\ x_k \neq x_{k+1}}^{l} \eta (M_k(0) - M_k(h_c)) \\ &\geq \sum_{k=l,\ x_k \neq x_{k+1}}^{l} \eta c_1 \epsilon \min \left\{ \Delta_k, \frac{\epsilon}{\beta} \right\}, \end{aligned}$$

onde $\beta \geq \|J(x_k)^T J(x_k)\|$, para todo $x_k \in \mathcal{S}$. Se $\Delta_k \leq \epsilon/\beta$, para todo $k = t, t+1, \ldots, l$, temos

$$F(x_t) - F(x_{l+1}) \geq \eta c_1 \epsilon \sum_{k=t,\ x_k \neq x_{k+1}}^{l} \Delta_k.$$

Pela desigualdade triangular, o somatório das normas de todos passos aceitos de t até l é maior do que R, pois $x_{l+1} \notin \mathcal{B}[x, R]$, e como esse mesmo somatório é menor que o somatório das regiões de confiança de t até l, temos

$$F(x_t) - F(x_{l+1}) \geq \eta c_1 \epsilon \sum_{k=t,\ x_k \neq x_{k+1}}^{l} \Delta_k \geq \eta c_1 \epsilon R = \eta c_1 \frac{\epsilon^2}{\gamma}. \tag{91}$$

Caso contrário, isto é, se $\Delta_k > \epsilon/\beta$, para algum $k = t, t+1, \ldots, l$, assim

$$F(x_t) - F(x_{l+1}) \geq \eta c_1 \epsilon^2/\beta. \tag{92}$$

Como, a partir da iteração k^*, aceitamos apenas passos que reduzem o valor da função objetivo, então sequência $(F(x_k))_{k \geq k^*}$ é não crescente. Como é limitada inferiormente, então esta sequência converge para algum $L \in \mathbb{R}$, pois $\mathcal{S}$ é compacto. Usando as expressões (91) e (92), temos

$$\begin{aligned} F(x_t) - L &\geq F(x_t) - F(x_{l+1}) \\ &\geq \eta c_1 \epsilon \min \left\{ \frac{\epsilon}{\beta}, \frac{\epsilon}{\gamma} \right\} \\ &= \frac{1}{2} \eta c_1 \|\nabla F(x_t)\| \min \left\{ \frac{\|\nabla F(x_t)\|}{2\beta}, \frac{\|\nabla F(x_t)\|}{2\gamma} \right\} > 0. \end{aligned}$$

Note que na nossa hipótese de absurdo, isto é, $\lim_{k\to\infty} \|\nabla F(x_k)\| \neq 0$, implica que existem infinitas escolhas possíveis de índices como t, então tomando $t \geq 0$ cada vez maiores, concluímos, mediante a desigualdade acima juntamente com o fato de que $(F(x_k) - L)_{k\geq 0}$ converge para zero, que $(\|\nabla F(x_k)\|)_{k\geq 0}$ converge para zero, que contradiz a hipótese de absurdo. Portanto $\lim_{k\to\infty} \|\nabla F(x_k)\| = 0$. □

Do mesmo que no caso do método de LM, podemos dimensionar o método de LMCS introduzindo uma matriz diagonal D, inversível. Os teoremas de convergência acima continuam válidos com essa modificação, e a prova disso é a mesma do caso do método de LM, já feita no Capítulo 2.

3.3 MODIFICAÇÕES

A seguir apresentamos algumas modificações do método LMCS. Tais modificações teve sua motivação puramente numérica, isto é, por meio de tentativa e erro, com objetivo de diminuir o número de passos em alguns problemas por nós testados.

Modificação M1

Vamos começar descrevendo a primeira modificação M1, em que resolvemos em cada iteração do método os seguintes sistemas

$$(J^TJ + \lambda D^TD)p_{lm} = -J^Tr \tag{93}$$

$$(J^TJ + \lambda D^TD + 2(\lambda+1)H)p_c = -\frac{1}{2}J^TK(p_{lm}, p_{lm}) - K(p_{lm}, \cdot)^T(r + Jp_{lm}). \tag{94}$$

onde $H = \text{diag}\,(\sqrt{Z})$ e Z é uma matriz dada por

$$Z = (J^TK(p_{lm}, \cdot))^T(K(p_{lm}, \cdot)^TJ).$$

A razão de acrescentarmos a matriz H dessa maneira é puramente numérica. O objetivo era encontrar uma forma viável de acrescentar informações de segunda ordem do lado direito do sistema (94).

Modificação M2

Do ponto de vista da diferenciação automática, que veremos no capítulo 4, dado uma função f de classe C^2, um ponto $x \in \mathbb{R}^n$ e uma direção $p \in \mathbb{R}^n$, podemos obter, simultaneamente, $f(x)$, $\nabla f(x)^T p$ e $\nabla^2 f(x)p$. Em vista dessa estratégia, na modificação M2 buscamos minimizar o uso da diferenciação automática em cada iteração.

Dessa forma, seja x_0 o ponto inicial. Na primeira iteração resolvemos normalmente os sistemas

$$(J^T J + \lambda D^T D)p^0_{lm} = -J^T r \tag{95}$$

$$(J^T J + \lambda D^T D)p_c = -\frac{1}{2}J^T K(p^0_{lm}, p^0_{lm}) - K(p^0_{lm}, \cdot)^T (r + Jp^0_{lm}). \tag{96}$$

Nas iterações seguintes prosseguimos da seguinte maneira

$$(J^T J + \lambda D^T D)p_{lm} = -J^T r \tag{97}$$

$$(J^T J + \lambda D^T D)p_c = -\frac{1}{2}J^T K(-p^a_{lm}, p_{lm}) - K(-p^a_{lm}, \cdot)^T (r + Jp_{lm}), \tag{98}$$

onde p^a_{lm} denota o passo de LM da iteração anterior à iteração presente e substituímos p_{lm} por $-p^a_{lm}$ no cálculo da derivada segunda direcional.

A diferença entre o primeiro passo e os seguinte está no fato de que, quando formos avaliar $r(x)$ para a resolução do sistema (98), do ponto de vista da diferenciação automática, aproveitamos esse momento para obter $K(-p^a_{lm}, \cdot)$, enquanto que no primeiro passo a derivada segunda é calculada no passo de LM da iteração presente.

A motivação do sinal de menos nem $-p^a_{lm}$, é a seguinte: se o passo de p^a_{lm} tem aproximadamente a direção oposta de p_{lm}, então use a direção oposta p^a_{lm}.

Existem métodos que fazem uso de estratégias onde busca-se aproveitar informações das direções anteriores, como é o caso do momento de Polyak, [37], onde a direção anterior é usada para melhorar o passo.

Modificação M3

A terceira modificação M3, trata de combinar as duas técnicas anteriores, isto é, substituímos a matriz dos sistemas (96) e (98) pela matriz do sistema (94) e prosseguimos da mesma maneira.

Modificação M4

Por fim, a modificação M4, é semelhante com a M2, mas dessa vez substituímos $K(p_{lm}, \cdot)$ por $K(p^a_{lm}, \cdot)$. Assim, prosseguimos nos passos seguintes, sempre usando a segunda derivada K aplicada na direção do passo de LM da iteração anterior. Além disso, avaliamos o modelo M usando $K(h^a_c, h^a_c)$ onde, h^a_c é o passo do método da iteração anterior a iteração presente, porém na primeira iteração utilizamos $K(h^0_c, h^0_c)$ onde, h^0_c é o passo correspondente ao ponto x_0. Novamente, aqui estamos fazendo uso de informação de uma passo anterior

Também é possível realizar o controle do tamanho da correção, a motivação dessa estratégia é numérica, mas pertinente, pois em alguns casos, em decorrência desse controle, temos um melhor desempenho, como nos exemplos de 2 a 4 que veremos na próxima seção.

Controlando o tamanho da correção

Por vezes, o tamanho do passo da correção, assim como o ângulo entre p_c e p_{lm}, se não forem adequadamente controlados, podem acarretar, em algumas situações, em uma ineficiência do método. Nesses casos, levamos em consideração a robustez do passo de LM para construir uma estratégia, de fácil implementação, e que ainda aproveita o passo p_c. Tal estratégia é dada a seguir.

Dado $\theta \in [-1, 1]$, se

$$\frac{p_c^T p_{lm}}{\|p_c\| \, \|p_{lm}\|} \geq \theta,$$

então fazemos $h_c = p_{lm}$.

Se não, se $\|p_c\| \geq \|p_{lm}\|$, então fazemos $\tilde{p}_c = b_2 p_c$, onde

$$b_2 = a_2 \frac{\|p_{lm}\|}{\|p_c\|}.$$

com $a_2 \in [0, 1]$ e fazemos $h_c = p_{lm} + \tilde{p}_c$.

A seguir, expomos alguns exemplos, bem como imagens do comportamento dos LM, LMCS e M4 nesses exemplos.

3.4 EXEMPLOS

Nesta Seção, vamos estudar o comportamento de três métodos em três exemplos de funções, por meio de imagens, vamos ilustrar o comportamento dos métodos LM, LMCS e M4, bem como a contagem de passos aceitos e rejeitados. Porém, nesses exemplos fizemos uso do controle do tamanho do passo de correção utilizando a estratégia dada acima.

Nestes exemplos, utilizamos $D = I$, $\lambda_0 = 10^{-4}$. Já para o parâmetro η, usamos $\eta = 0$ dos exemplos 1 ao 2, sendo que no exemplo 3, fizemos $\eta = 0$ para LM e LMCS, mas para M4 usamos $\eta = 1/4$. Como critério de parada, sendo $(x_k)_{k\geq 0}$ a sequência gerada pelo método, o algoritmo se encerra quando $\|h_c\| \leq \epsilon_1(x_k + \epsilon_1)$ ou $\|\nabla f(x_k)\| < 10^{-8}$, onde tomamos $\epsilon_1 = 10^{-8}$. Já para os parâmetros $PC = \infty$ e $PT = \infty$, ou seja, aceitamos todos os passos que aumentam o valor da função objetivo.

Exemplo 1

Considere a função,

$$F(x, y) = (1 - x^4)^2 + 100(y - x^2)^2 + 100(\sin(y^2) - x)^2.$$

Logo, seu gradiente e dado por

$$g(x, y) = \begin{bmatrix} -8(1 - x^4)x^3 - 400(y - x^2)x - 200(\sin(y^2) - x) \\ 200(y - x^2) + 400(\sin(y^2) - x)\cos(y^2)y \end{bmatrix}.$$

Essa função pode ser vista como um problema de mínimos quadrados, pois

$$F(x, y) = \frac{1}{2}(r_1(x, y)^2 + r_2(x, y)^2 + r_3(x, y)^2),$$

onde

$$r(x, y) = \begin{bmatrix} r_1(x, y) \\ r_2(x, y) \\ r_3(x, y) \end{bmatrix} = \begin{bmatrix} \sqrt{2}(1 - x^4) \\ 10\sqrt{2}(y - x^2) \\ 10\sqrt{2}(\sin(y^2) - x) \end{bmatrix}.$$

Assim, jacobiana do vetor de resíduos é

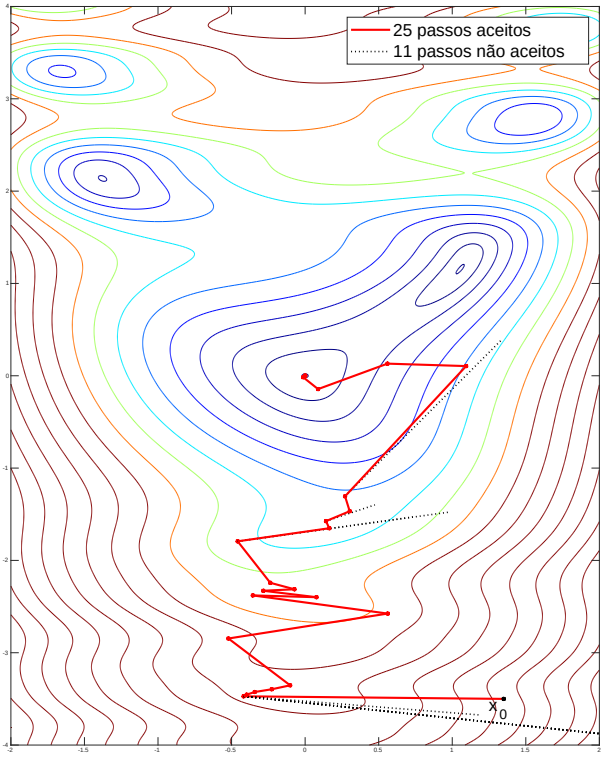

Figura 6: LM para o Exemplo 1.

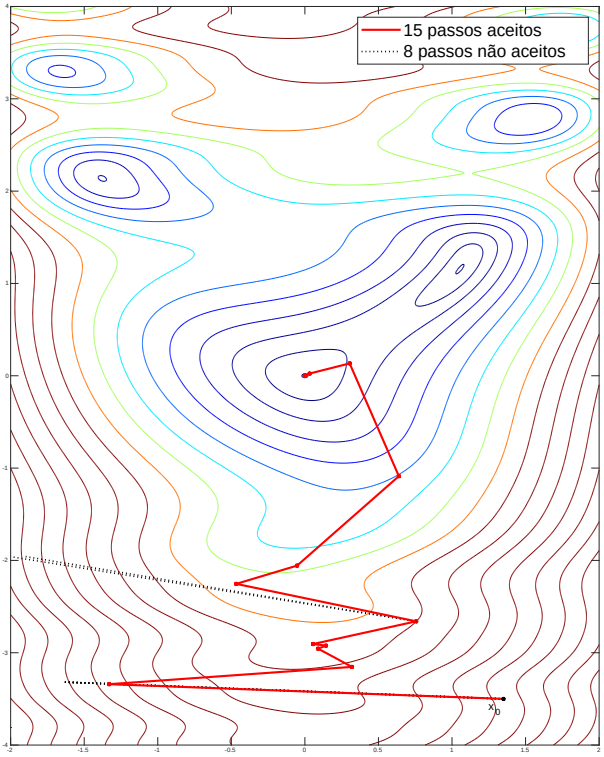

Figura 7: LMCS para o Exemplo 1.

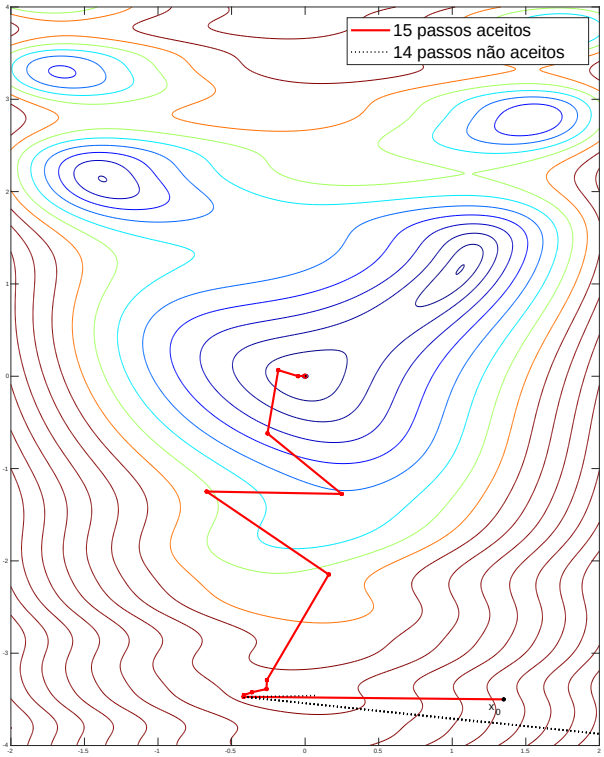

Figura 8: M4 para o Exemplo 1.

$$J(x,y) = \begin{bmatrix} -\sqrt{2}4x^3 & 0 \\ -20\sqrt{2}x & 10\sqrt{2} \\ -10\sqrt{2} & 20\sqrt{2}y\cos(y^2) \end{bmatrix},$$

e as hessianas dos resíduos são dadas por

$$\nabla^2 r_1(x,y) = \begin{bmatrix} -12\sqrt{2}x^2 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\nabla^2 r_2(x,y) = \begin{bmatrix} -20\sqrt{2} & 0 \\ 0 & 0 \end{bmatrix}$$

e

$$\nabla^2 r_3(x,y) = \begin{bmatrix} 0 & 0 \\ 0 & 20\sqrt{2}(\cos(y^2 - 2(y^2)\sin(y^2))) \end{bmatrix}.$$

Nas Figuras 6, 7 e 8 estão esboçados os passos dos algoritmos de LM, LMCS e M4 com controle do tamanho da correção, aplicados para resolver o problema do Exemplo 1. Nesse exemplo, utilizamos $\theta = 1$ e $a_2 = 0.7$ tanto para o LMCS quanto para M4. Todos os passos obtidos são esboçados até a convergência para um ponto estacionário e são iniciados com os mesmo ponto inicial x_0. Observe que os três algoritmos convergem para o mesmo ponto estacionário.

Note que neste caso, o método LMCS realiza um número bem menor de iterações, reduzindo tanto a quantidade de passos aceitos quanto de rejeitados quando comparado ao método de LM. Já M4 apresenta uma redução de passos aceitos as custas de um aumento no número de passos rejeitados, em relação ao método de LM. Tal fenômeno é, em geral, preferível, pois, a estrutura especial da matriz do sistema de LM nos permite, como veremos no próximo capítulo, realizar a decomposição QR de maneira pouco custosa, já que, para os passos rejeitados, o ponto da sequência gerado pelo método permanece o mesmo, sendo atualizado apenas o parâmetro λ. Dessa maneira, a obtenção da solução do sistema associado tem um custo menor do que nos casos dos passos aceitos.

Exemplo 2

Considere a função,

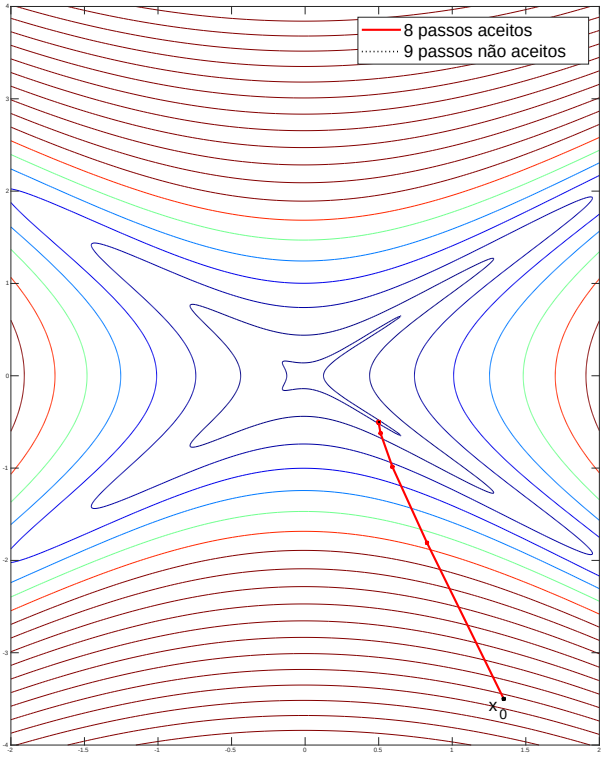

Figura 9: LM para o Exemplo 2.

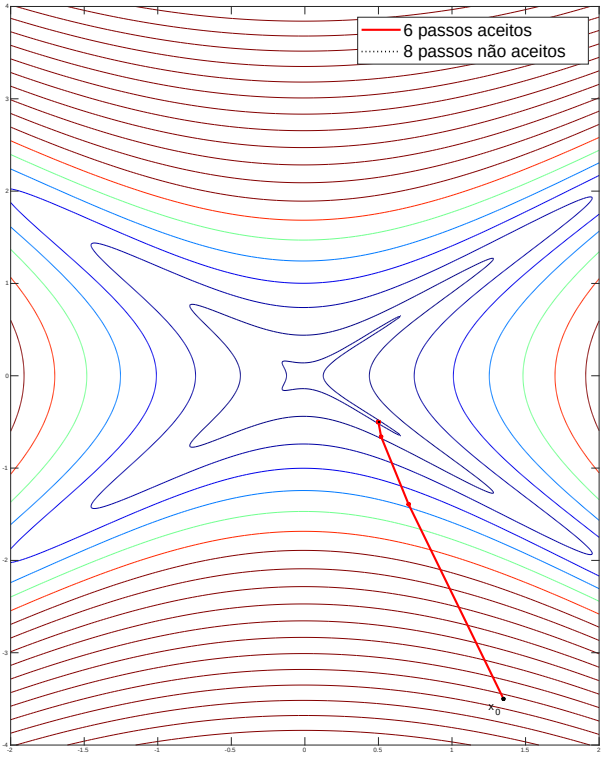

Figura 10: LMCS para o Exemplo 2.

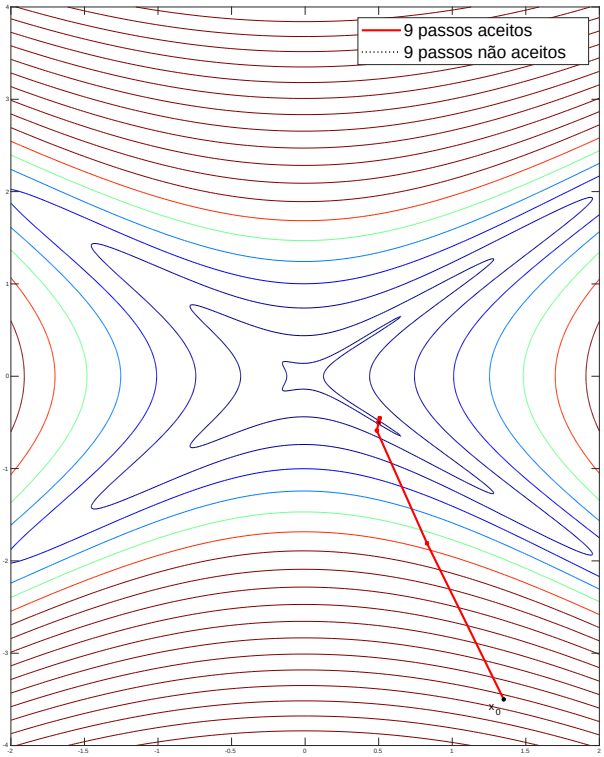

Figura 11: M4 para o Exemplo 2.

$$F(x,y) = (2y^2 - x)^2 + 100(y^2 - x^2)^2.$$

Logo, seu gradiente e dado por

$$g(x,y) = \begin{bmatrix} -2(2y^2 - x) - 400(y^2 - x^2)x \\ 8(y^2 - x)y + 400(y - x^2)y \end{bmatrix}.$$

Essa função pode ser vista como um problema de mínimos quadrados, pois

$$F(x,y) = \frac{1}{2}(r_1(x,y)^2 + r_2(x,y)^2),$$

onde

$$r(x,y) = \begin{bmatrix} r_1(x,y) \\ r_2(x,y) \end{bmatrix} = \begin{bmatrix} \sqrt{2}(2y^2 - x) \\ 10\sqrt{2}(y^2 - x^2) \end{bmatrix}.$$

Assim, jacobiana do vetor de resíduos é

$$J(x,y) = \begin{bmatrix} -\sqrt{2} & 4\sqrt{2}y \\ -20\sqrt{2}x & 20\sqrt{2}y \end{bmatrix},$$

e as hessianas dos resíduos são dadas por

$$\nabla^2 r_1(x,y) = \begin{bmatrix} 0 & 0 \\ 0 & 4\sqrt{2} \end{bmatrix}$$

e

$$\nabla^2 r_2(x,y) = \begin{bmatrix} -20\sqrt{2} & 0 \\ 0 & 20\sqrt{2} \end{bmatrix}.$$

Nas Figuras 9, 10 e 11 estão esboçados os passos dos algoritmos de LM, LMCS e M4 com controle do tamanho da correção, aplicados para resolver o problema do Exemplo 2. Nesse exemplo, usamos $\theta = 1$, tanto para LMCS quanto para M4 e $a_2 = 0.7$ para LMCS e $a_2 = 0.9$ em M4. Todos os passos obtidos são esboçados até a convergência para um ponto estacionário e são iniciados com os mesmo ponto inicial x_0. Observe que os três algoritmos convergem para o mesmo ponto estacionário.

Os resultados neste caso são equilibrados, sendo que LMCS realiza uma quantidade de passos ligeiramente menor que LM.

Exemplo 3

Considere a função,

$$F(x,y) = (1-\sin(x))^2 + 100(\cos(y)-x^2)^2.$$

Logo, seu gradiente e dado por

$$g(x,y) = \begin{bmatrix} -2(1-\sin(x))\cos(x) - 400(\cos(y)-x^2)x \\ -200(\cos(y)-x^2)\sin(y) \end{bmatrix}.$$

Essa função pode ser vista como um problema de mínimos quadrados, pois

$$F(x,y) = \frac{1}{2}(r_1(x,y)^2 + r_2(x,y)^2),$$

onde

$$r(x,y) = \begin{bmatrix} r_1(x,y) \\ r_2(x,y) \end{bmatrix} = \begin{bmatrix} \sqrt{2}(1-\sin(x)) \\ 10\sqrt{2}(\cos(y)-x^2) \end{bmatrix}.$$

Assim, jacobiana do vetor de resíduos é

$$J(x,y) = \begin{bmatrix} -\sqrt{2}\cos(x) & 0 \\ -20\sqrt{2}x & -10\sqrt{2}\sin(y) \end{bmatrix},$$

e as hessianas dos resíduos são dadas por

$$\nabla^2 r_1(x,y) = \begin{bmatrix} \sqrt{2}\sin(x) & 0 \\ 0 & 0 \end{bmatrix}$$

e

$$\nabla^2 r_2(x,y) = \begin{bmatrix} -20\sqrt{2} & 0 \\ 0 & -10\sqrt{2}\cos(y) \end{bmatrix}.$$

Nas Figuras 12, 13 e 14 estão esboçados os passos dos algoritmos de LM, LMCS e M4 com controle do tamanho da correção, aplicados para resolver o problema do Exemplo 3. Nesse exemplo, usamos $\theta = 0.95$ e $a_2 = 0.9$ tanto em LMCS quanto em M4. Todos os passos obtidos são esboçados até a convergência para um ponto estacionário e são iniciados com os mesmo ponto inicial x_0. Observe que os três algoritmos convergem para o mesmo ponto estacionário.

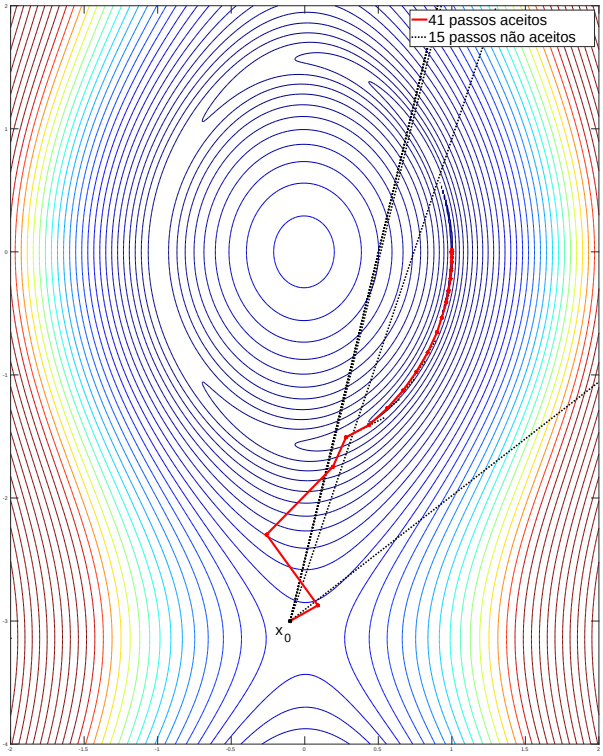

Figura 12: LM para o Exemplo 3.

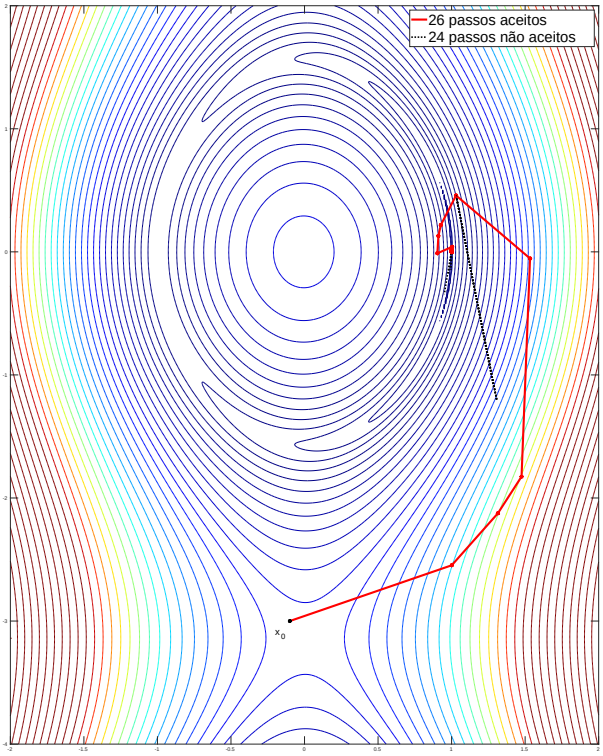

Figura 13: LMCS para o Exemplo 3.

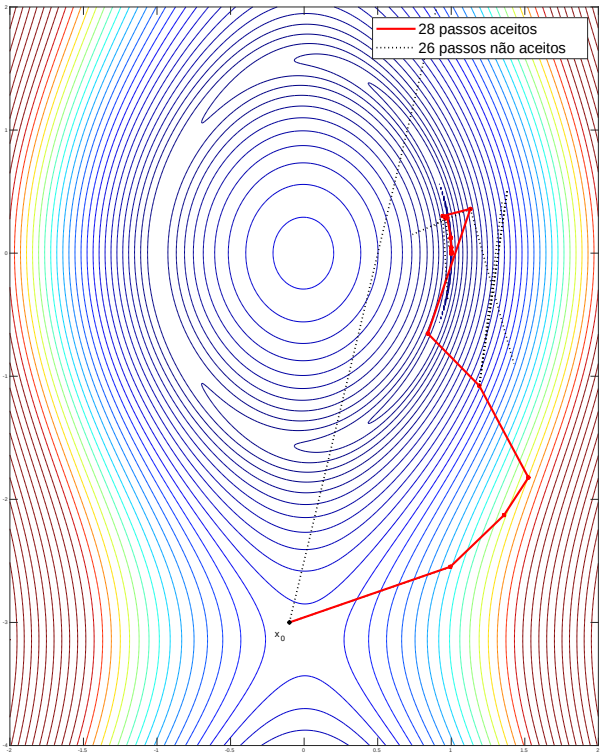

Figura 14: M4 para o Exemplo 3.

Note que neste caso, tanto LMCS quanto M4 realizam um número total de iterações menor que o método de LM. Além disso, tanto LMCS quanto M4 apresentam um redução significativa de passos aceitos as custas de um aumento no número de passos rejeitados, quando comparado com o método de LM, entretanto, como discutimos no Exemplo 1, isso é mais vantajoso do ponto de vista dos custos computacionais envolvidos na solução do sistema de LM.

4 IMPLEMENTAÇÃO E RESULTADOS NUMÉRICOS

Observe que, no método de Levenberg-Marquardt com correção de segunda ordem, é necessário resolver, em cada iteração, dois sistemas lineares. Porém, como a matriz de ambos os sistemas são iguais, podemos aplicar técnicas de fatoração matricial a fim de reduzir o custo do processo, haja visto que, se decompormos uma das matrizes uma vez para resolver um dos sistemas, a outra será obtida imediatamente de maneira muito econômica. Para essa parte, podemos usar técnicas especificas para este caso, em que a decomposição matricial pode tornar a obtenção da solução dos sistemas mais eficiente do ponto de vista numérico, [32], [36].

Além da resolução dos sistemas, em nossa proposta de modificação do método de LM, exigimos o cálculo da derivada de segunda ordem aplicada em uma direção, para obtê-la, propomos, em detrimento da tradicional técnica de diferenças finitas, a diferenciação automática [8], [21] e [24]. Nesse sentido, também apresentamos uma contribuição, fornecendo a fórmula necessária para obter, dados um ponto x e uma direção p, a derivada de segunda ordem em x e aplicado na direção p, simultaneamente com os valores da função em x, e a derivada de primeira em x na direção p.

4.1 FATORANDO A MATRIZ DO SISTEMA DE LM

A fim de propiciar maior clareza nas fórmulas, vamos realizar as seguintes convenções: $r(x) = r$, $J(x) = J$ e $K(x) = K$. Observe que não há risco de confusão nessa notação, haja visto que todo processo de construção do método é feita sobre cada iteração.

Note que o problema de resolver

$$(J^T J + \lambda D^T D)p = -J^T r, \tag{99}$$

pode ser escrito de forma equivalente como

$$\min_{p\in\mathbb{R}^n} \left\| \begin{bmatrix} J \\ \sqrt{\lambda}D \end{bmatrix} p + \begin{bmatrix} r \\ 0 \end{bmatrix} \right\|^2 . \tag{100}$$

A princípio, podemos usar a fatoração Cholesky para resolver (99). Entretanto, devido à estrutura especial da matriz $J^TJ + \lambda D^TD$, vamos apresentar uma técnica eficiente para encontrar uma decomposição QR, com permutação das colunas, da matriz do sistema (100), isto é, queremos obter

$$\begin{bmatrix} J \\ \sqrt{\lambda}D \end{bmatrix} P_\lambda = Q_\lambda \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix}, \tag{101}$$

onde $Q_\lambda \in \mathbb{R}^{(n+m)\times(n+m)}$ é uma matriz ortogonal, $R_\lambda \in \mathbb{R}^{n\times n}$ triangular superior e $P_\lambda \in \mathbb{R}^{n\times n}$ é a matriz de permutação nas colunas.

Para tanto, primeiramente calculamos a decomposição QR de J com permutação das colunas, que pode ser obtida por meio das transformações de Householder [19],

$$JP = Q \begin{bmatrix} R \\ 0 \end{bmatrix},$$

onde $Q \in \mathbb{R}^{m\times m}$ é uma matriz ortogonal, $R \in \mathbb{R}^{n\times n}$ é uma matriz triangular superior e $P \in \mathbb{R}^{n\times n}$ é a matriz de permutação das colunas de J.

Vamos obter Q_λ, R_λ e P_λ em função de Q, R, P e λ. Para isso, note que podemos escrever

$$\begin{bmatrix} Q^T & 0 \\ 0 & P^T \end{bmatrix} \begin{bmatrix} J \\ \sqrt{\lambda}D \end{bmatrix} P = \begin{bmatrix} Q^T & 0 \\ 0 & P^T \end{bmatrix} \begin{bmatrix} JP \\ \sqrt{\lambda}DP \end{bmatrix} = \begin{bmatrix} Q^TJP \\ D_\lambda \end{bmatrix} = \begin{bmatrix} R \\ 0 \\ D_\lambda \end{bmatrix},$$

onde definimos $D_\lambda = \sqrt{\lambda}P^TDP$, que é uma matriz diagonal com os mesmos elementos da matriz D, porém permutados. Logo,

$$\begin{bmatrix} Q^T & 0 \\ 0 & P^T \end{bmatrix} \begin{bmatrix} J \\ \sqrt{\lambda} D \end{bmatrix} P = \begin{bmatrix} R \\ 0 \\ D_\lambda \end{bmatrix}. \tag{102}$$

O fato de D_λ ser diagonal permite que, com $n(n+1)/2$ rotações de Givens, todos os n elementos presentes na diagonal são zerados [19], [36]. Assim, definimos a matriz ortogonal $\bar{Q}_\lambda$ como sendo aquela obtida pelo produto das $n(n+1)/2$ rotações de Givens. Dessa forma, temos

$$\bar{Q}_\lambda^T \begin{bmatrix} R \\ 0 \\ D_\lambda \end{bmatrix} = \begin{bmatrix} R_\lambda \\ 0 \\ 0 \end{bmatrix},$$

e da equação (102), segue que

$$\bar{Q}_\lambda^T \begin{bmatrix} Q^T & 0 \\ 0 & P^T \end{bmatrix} \begin{bmatrix} J \\ \sqrt{\lambda} D \end{bmatrix} P = \bar{Q}_\lambda^T \begin{bmatrix} R \\ 0 \\ D_\lambda \end{bmatrix} = \begin{bmatrix} R_\lambda \\ 0 \\ 0 \end{bmatrix}.$$

De onde concluímos que

$$\begin{bmatrix} J \\ \sqrt{\lambda} D \end{bmatrix} P = \begin{bmatrix} Q & 0 \\ 0 & P \end{bmatrix} \bar{Q}_\lambda \begin{bmatrix} R_\lambda \\ 0 \\ 0 \end{bmatrix},$$

e portanto comparando com a equação (101), Q_λ pode ser dada por

$$Q_\lambda = \begin{bmatrix} Q & 0 \\ 0 & P \end{bmatrix} \bar{Q}_\lambda, \tag{103}$$

e $P_\lambda = P$. Sabemos que a matriz de permutação P é ortogonal, isto é, $P^T P = I$. Das equações (101) e (103), e usando o fato de que transformações ortogonais preservam a norma, temos que

$$\begin{aligned}
\left\| \begin{bmatrix} J \\ \sqrt{\lambda}D \end{bmatrix} P(P^T p) + \begin{bmatrix} r \\ 0 \end{bmatrix} \right\|^2 &= \left\| Q_\lambda \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} P^T p + \begin{bmatrix} r \\ 0 \end{bmatrix} \right\|^2 \\
&= \left\| Q_\lambda^T \left(Q_\lambda \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} P^T p + \begin{bmatrix} r \\ 0 \end{bmatrix} \right) \right\|^2 \\
&= \left\| \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} P^T p + Q_\lambda^T \begin{bmatrix} r \\ 0 \end{bmatrix} \right\|^2 \\
&= \left\| \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} P^T p + \bar{Q}_\lambda^T \begin{bmatrix} Q^T r \\ 0 \end{bmatrix} \right\|^2 .
\end{aligned}$$

Como $\begin{bmatrix} J \\ \sqrt{\lambda}D \end{bmatrix}$ tem posto coluna completo, pois D é inversível, então existe $p \in \mathbb{R}^n$ tal que

$$\begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} P^T p = -\bar{Q}_\lambda^T \begin{bmatrix} Q^T r \\ 0 \end{bmatrix},$$

e esse p é o mínimo global que soluciona (100), e portanto é o passo de LM e podemos denotá-lo por p_{lm}. Assim, temos

$$R_\lambda (P^T p_{lm}) = -u, \tag{104}$$

onde $u \in \mathbb{R}^n$ é tal que,

$$-\bar{Q}_\lambda^T \begin{bmatrix} Q^T r \\ 0 \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix},$$

para algum $v \in \mathbb{R}^m$.

Contudo, quando m é grande o custo de guardar a matriz Q é grande. A estratégia seguinte, nos fornece uma maneira de encontrar p_{lm}, sem a necessidade de armazenar Q.

Usando a decomposição QR dada em (101), temos

$$Q_\lambda \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} (P^T p_{lm}) = - \begin{bmatrix} r \\ 0 \end{bmatrix}. \tag{105}$$

Multiplicando à esquerda ambos os lados da equação (105) por $P^T \begin{bmatrix} J^T & \sqrt{\lambda}D^T \end{bmatrix}$, temos

$$P^T \begin{bmatrix} J^T & \sqrt{\lambda}D^T \end{bmatrix} Q_\lambda \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} (P^T p_{lm}) = -P^T \begin{bmatrix} J^T & \sqrt{\lambda}D^T \end{bmatrix} \begin{bmatrix} r \\ 0 \end{bmatrix}. \tag{106}$$

Usando a decomposição a QR dada pela equação (101) em (106), obtemos

$$\begin{bmatrix} R_\lambda^T & 0 \end{bmatrix} Q_\lambda^T Q_\lambda \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} (P^T p_{lm}) = -P^T J^T r.$$

Como Q_λ é ortogonal, então $Q_\lambda^T Q_\lambda = I$, logo

$$\begin{bmatrix} R_\lambda^T & 0 \end{bmatrix} \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix} (P^T p_{lm}) = -P^T J^T r.$$

Finalmente, o sistema (99) pode ser escrito como

$$R_\lambda^T R_\lambda (P^T p_{lm}) = -P^T J^T r, \tag{107}$$

o qual pode ser resolvido via solução de dois sistemas lineares triangulares.

Entretanto, apesar da vantagem de não precisar armazenar a matriz Q, se a matriz de do sistema de LM for mal condicionada, R_λ também será, e o condicionamento de $R_\lambda^T R_\lambda$ será o quadrado do condicionamento de R_λ.

Para o caso do sistema para encontrar o passo de correção p_c, podemos simplesmente usar a decomposição QR, (101), já efetuada e resolvemos

$$R_\lambda^T R_\lambda (P^T p_{lm}) = -P^T (J^T K(p_{lm}, p_{lm}) + K(p_{lm}, \cdot)^T (r + J p_{lm})). \tag{108}$$

Observe que, no caso em que o passo falha, o ponto da sequência permanece o mesmo, logo apenas apenas λ é alterado. Nesse caso, não é necessário recalcular a fatoração QR de $\begin{bmatrix} J \\ \sqrt{\lambda}D \end{bmatrix} P_\lambda$, basta apenas atualizar as rotações de Givens.

Essa estratégia apresenta uma grande vantagem quando m é muito maior que n, pois para realizar decomposição QR da matriz completa, teríamos $\mathcal{O}((m+n)n^2)$ operações, no entanto, se realizarmos a estratégia acima, teríamos que realizar $\mathcal{O}(mn^2)$ operações, para obter a QR de J, mais $\mathcal{O}(n^3)$ da atualização das rotações de Givens, dessa forma,

se rejeitamos o passo, realizamos apenas $\mathcal{O}(n^3)$ operações para atualizar decomposição QR, no lugar de realizar $\mathcal{O}((m+n)n^2)$.

4.2 NÚMEROS DUAIS

Definição

De forma semelhante a definição dos números complexos, os números duais [8], [21], [24], podem ser vistos como uma extensão dos números reais, isto é, dados $x, y \in \mathbb{R}$ definimos um número dual como

$$z = x + y\epsilon, \tag{109}$$

com a propriedade de que o elemento ϵ é nilpotente, isto é, $\epsilon^2 = 0$.

Analogamente ao caso dos números complexos, chamamos $x = Re(z)$ de parte real e $y = Im(z)$ de parte imaginária do número dual, e escrevemos z usando a notação vetorial

$$z = (x, y). \tag{110}$$

Tomando dois números duais, z_1 e z_2, tais que

$$z_1 = x_1 + y_1\epsilon = (x_1, y_1) \text{ e } z_2 = x_2 + y_2\epsilon = (x_2, y_2),$$

definimos a aritmética de números duais da seguinte forma

$$z_1 + z_2 = (x_1 + x_2) + (y_1 + y_2)\epsilon = (x_1 + x_2, y_1 + y_2), \tag{111}$$

$$z_1 - z_2 = (x_1 - x_2) + (y_1 - y_2)\epsilon = (x_1 - x_2, y_1 - y_2), \tag{112}$$

$$z_1 z_2 = (x_1 x_2) + (x_1 y_2 + x_2 y_1)\epsilon = (x_1 x_2, x_1 y_2 + x_2 y_1), \tag{113}$$

$$\frac{z_1}{z_2} = \left(\frac{x_1}{x_2}\right) + \left(\frac{y_1 x_2 - x_1 y_2}{x_2^2}\right)\epsilon = \left(\frac{x_1}{x_2}, \frac{y_1 x_2 - x_1 y_2}{x_2^2}\right). \tag{114}$$

A equação (113), isto é, a multiplicação de números duais, pode ser intuída pela propriedade distributiva, ou seja,

$$\begin{aligned} z_1 z_2 &= (x_1 + y_1\epsilon)(x_2 + y_2\epsilon) \\ &= x_1x_2 + x_1y_2\epsilon + y_1x_2\epsilon + y_1y_2\epsilon^2 \\ &= x_1x_2 + x_1y_2\epsilon + y_1x_2\epsilon \\ &= x_1x_2 + (x_1y_2 + y_1x_2)\epsilon. \end{aligned}$$

Já a última relação dada pela equação (114) é facilmente obtida multiplicando-se o numerador e o denominador pelo conjugado de z_2, onde o conjugado de z_2, $\bar{z}_2$, é definido de maneira análoga ao caso dos números complexos, ou seja, $\bar{z}_2 = x_2 - y_2\epsilon$.

Assim, a divisão de números duais pode ser dada por

$$\begin{aligned} \frac{z_1}{z_2} &= \frac{x_1 + y_1\epsilon}{x_2 + y_2\epsilon} \\ &= \frac{(x_1 + y_1\epsilon)(x_2 - y_2\epsilon)}{(x_2 + y_2\epsilon)(x_2 - y_2\epsilon)} \\ &= \frac{x_1x_2 + (y_1x_2 - x_2y_1)\epsilon}{x_2^2 - y_2^2\epsilon^2} \\ &= \frac{x_1x_2 + (y_1x_2 - x_2y_1)\epsilon}{x_2^2} \\ &= \frac{x_1}{x_2} + \frac{y_1x_2 - x_2y_1}{x_2^2}\epsilon. \end{aligned}$$

A aritmética de números duais nos permite definir de forma natural um anel que estende $\mathbb{R}$, o qual chamaremos de espaço dual e denotaremos por $\mathbb{E}$.

Note que os números duais cuja parte real é igual a zero não possuem inverso multiplicativo, o que implica que o conjunto $\mathbb{E}$ não é um corpo.

Extensão de funções reais diferenciáveis

A princípio queremos estender as funções reais diferenciáveis para o anel dos números duais, mas antes, a fim de motivar a ideia geral de como isso será feito, vamos começar estendendo um polinômio sobre $\mathbb{R}$ qualquer, para um polinômio sobre o anel $\mathbb{E}$.

Considere $P : \mathbb{R} \to \mathbb{R}$, um polinômio de grau n dado por

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0, \quad x \in \mathbb{R}.$$

Estendendo o domínio de P para $\mathbb{E}$, obtemos o polinômio $\tilde{P} : \mathbb{E} \to \mathbb{E}$ dado por

$$\tilde{P}(z) = \alpha_n z^n + \alpha_{n-1} z^{n-1} + \ldots + \alpha_1 z + \alpha_0, \quad z = (x, y) \in \mathbb{E}.$$

Usando o binômio de Newton e a relação (109), verificamos que

$$\begin{aligned} z^n &= (x + y\epsilon)^n \\ &= \sum_{k=0}^{n} \frac{n!}{k!\,(n-k)!} x^{n-k} (y\epsilon)^k \\ &= x^n + n x^{n-1} y\epsilon \\ &= (x^n, n x^{n-1} y). \end{aligned} \tag{115}$$

Dessa forma,

$$\begin{aligned} \tilde{P}(z) =& \alpha_n z^n + \alpha_{n-1} z^{n-1} + \ldots + \alpha_1 z + \alpha_0 \\ =& \alpha_n (x^n, n x^{n-1} y) + \alpha_{n-1} (x^{n-1}, (n-1) x^{n-2} y) + \ldots + \alpha_1 (x, y) + \alpha_0 \\ =& \Big(\underbrace{\alpha_n x^n + \alpha_{n-1} x^{n-1} + \ldots + \alpha_1 x + \alpha_0}_{\text{parte real}}, \\ & \underbrace{(n\alpha_{n-1} x^{n-1} + (n-1)\alpha_{n-2} x^{n-2} \ldots + 2\alpha_2 x^2 + \alpha_1 x) y}_{\text{parte imaginária}} \Big) \\ =& (P(x), P'(x) y), \end{aligned} \tag{116}$$

onde P' é a derivada de P. Portanto,

$$\tilde{P}|_{\mathbb{R}} = P,$$

logo $\tilde{P}$ é uma extensão de P e, além disso,

$$Im(\tilde{P}(z)) = P'(x) y, \quad \forall z = (x, y) \in \mathbb{E}.$$

A relação (116) pode ser generalizada para qualquer função analítica

$$f(z) = f(x + y\epsilon) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!} (y\epsilon)^n = f(x) + f'(x) y\epsilon = (f(x), f'(x) y). \tag{117}$$

Esta relação é a fundamental para teoria de diferenciação automática, pois a extensão do domínio de uma função real para números duais permite calcular o valor da função e da derivada simultaneamente.

Note que, se tomamos a parte imaginária da variável dual como 1, obtemos

$$f((x,1)) = (f(x), f'(x)). \tag{118}$$

Podemos verificar todas as regras de derivação usando a análise dual apresentada. Por exemplo, de (118)

$$f(g((x,1))) = f((g(x), g'(x))) = (f(g(x)), f'(g(x))g'(x)), \tag{119}$$

de (118) e (114)

$$\frac{f((x,1))}{g((x,1))} = \left(\frac{f(x)}{g(x)}, \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}\right), \tag{120}$$

e assim por diante.

Da mesma forma, como estendemos o domínio de funções reais de uma variável para o espaço dual, podemos estender uma função definida em $\mathbb{R}^n$ para $\mathbb{E}^n$. Sejam $f : \mathbb{E}^n \to \mathbb{E}$ e $\mathbf{z} \in \mathbb{E}^n$, com $\mathbf{z} = \mathbf{x} + \mathbf{y}\epsilon = (\mathbf{x}, \mathbf{y})$. Usando o mesmo resultado de (117) com a notação matricial, obtemos

$$f(\mathbf{z}) = f(\mathbf{x} + \mathbf{y}\epsilon) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{y}\epsilon = (f(\mathbf{x}), \nabla f(\mathbf{x})^T \mathbf{y}), \tag{121}$$

com a parte imaginária dada como uma derivada direcional na direção de $Im(\mathbf{z}) = \mathbf{y}$.

Duais Generalizados

Assim como os números complexos [24], os números duais podem ser generalizados da seguinte forma

$$z = x + y_1\epsilon_1 + y_2\epsilon_2 + \ldots + y_n\epsilon_n = (x, \mathbf{y}), \tag{122}$$

onde $\mathbf{y} = y_1\epsilon_1 + y_2\epsilon_2 + \cdots + y_n\epsilon_n$ pode ser interpretado como um vetor com os coeficientes $y_1, y_2, \ldots, y_n$, ou seja, a parte imaginária do dual generalizado é um vetor.

A propriedade

$$\epsilon_i\epsilon_j = 0, \qquad \forall i, j \in \{1, \ldots, n\}, \tag{123}$$

fornece a aritmética de números duais generalizados. Dados

$$z_1 = (x_1, \mathbf{y}_1) \text{ e } z_2 = (x_2, \mathbf{y}_2),$$

temos

$$z_1 + z_2 = (x_1 + x_2, \mathbf{y}_1 + \mathbf{y}_2) \tag{124}$$

$$z_1 - z_2 = (x_1 + x_2, \mathbf{y}_1 - \mathbf{y}_2) \tag{125}$$

$$z_1 z_2 = (x_1 x_2, x_1 \mathbf{y}_2 + x_2 \mathbf{y}_1) \tag{126}$$

$$\frac{z_1}{z_2} = \left(\frac{x_1}{x_2}, \frac{x_2 \mathbf{y}_1 - x_1 \mathbf{y}_2}{x_2^2} \right), \tag{127}$$

onde as expressões (126) e (127) podem ser intuídas da mesma forma que no caso não generalizado. Denotaremos o espaço dual generalizado com n elementos imaginários por $\mathbb{E}_n$.

A extensão de uma função com domínio real para o dual generalizado é dada por

$$f(z) = f((x, \mathbf{y})) = (f(x), f'(x)\mathbf{y}), \tag{128}$$

onde $z = (x, \mathbf{y}) \in \mathbb{E}_n$. Usando o mesmo raciocínio aplicado em (121), definimos $f : \mathbb{E}_n^n \to \mathbb{E}_n$, extensão de uma função real de n variáveis, e $\mathbf{z} = (z_1, z_2, \ldots, z_n) \in \mathbb{E}_n^n$, com $z_i = (x_i, \mathbf{y}_i)$, $i \in \{1, \ldots, n\}$, por

$$f(\mathbf{z}) = f(z_1, z_2, \ldots, z_n) = f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{n} \nabla f(x_1, x_2, \ldots, x_n)^T \mathbf{y}_i \epsilon_i = (f(\mathbf{x}), \mathbf{u}), \tag{129}$$

com

$$\mathbf{x} = (x_1, x_2, \ldots, x_n) \text{ e } \mathbf{u} = (\nabla f(\mathbf{x})^T \mathbf{y}_1, \nabla f(\mathbf{x})^T \mathbf{y}_2, \ldots, \nabla f(\mathbf{x})^T \mathbf{y}_n). \tag{130}$$

Tomando $\mathbf{y}_i = \mathbf{e}_i$, como o i-ésimo vetor da base canônica, obtemos

$$\nabla f(\mathbf{x})^T \mathbf{y}_i = \frac{\partial f(\mathbf{x})}{\partial x_i}$$

e, portanto,

$$f(\mathbf{z}) = (f(\mathbf{x}), \nabla f(\mathbf{x})), \text{ com } z_i = (x_i, \mathbf{e}_i). \tag{131}$$

Note que dessa maneira obtemos o valor da função na parte real e o vetor gradiente na parte imaginária.

Fazendo um passo a mais, podemos obter a hessiana de f aplicada em uma direção dada. Supondo que $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{y}\epsilon$ é um vetor no espaço dual $\mathbb{E}^n$ (não generalizado), com $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$ e para cada $k \in \{1, \ldots, n\}$ estendendo $\partial f(x)/\partial x_k$ para o espaço dual $\mathbb{E}_n$, então por (121) segue que

$$\begin{aligned}
\frac{\partial f(\tilde{\mathbf{x}})}{\partial x_k} &= \frac{\partial f(\mathbf{x})}{\partial x_k} + \nabla\left(\frac{\partial f(\mathbf{x})}{\partial x_k}\right)^T \mathbf{y}\epsilon \\
&= \frac{\partial f(\mathbf{x})}{\partial x_k} + \left(\sum_{i=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i \partial x_k} y_i\right)\epsilon \\
&= \left(\frac{\partial f(\mathbf{x})}{\partial x_k}, \sum_{i=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i \partial x_k} y_i\right).
\end{aligned}$$

Note que

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

e, além disso, usando o fato de que

$$\frac{\partial f(\mathbf{x})}{\partial x_k \partial x_i} = \frac{\partial f(\mathbf{x})}{\partial x_i \partial x_k} \text{ para todo } i, k \in \{1, \ldots, n\},$$

temos

$$(\nabla^2 f(\mathbf{x})\mathbf{y})_k = \sum_{i=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_k \partial x_i} y_i = \sum_{i=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i \partial x_k} y_i,$$

para todo $k \in \{1, \ldots, n\}$. Dessa forma, definindo

$$\nabla f(\tilde{\mathbf{x}}) = \left(\frac{\partial f(\tilde{\mathbf{x}})}{\partial x_1}, \ldots, \frac{\partial f(\tilde{\mathbf{x}})}{\partial x_n}\right)^T,$$

concluímos que

$$\nabla f(\tilde{\mathbf{x}}) = (\nabla f(\mathbf{x}), \nabla^2 f(\mathbf{x})\mathbf{y}). \tag{132}$$

Portanto, podemos inserir a parte imaginária de duais generalizados no espaço dual e obter a extensão de uma função real de n variáveis. Esta extensão fornece, além do valor da função, o vetor gradiente e o produto da hessiana por um vetor.

Produto de hessiana por vetor através do uso da aritmética de números duais

Agora vamos apresentar outra contribuição de nosso trabalho. Assim, como em (117), também é possível obter a derivada de segunda ordem em uma direção específica simultaneamente com o cálculo de f e sua derivada de primeira ordem em duas direções dadas. Para tanto, precisamos definir um outro espaço, que denotaremos por $\widehat{\mathbb{E}}_n$ e é constituído por elementos que são da forma

$$z = x + y\epsilon + \sum_{i=1}^{n} u_i\epsilon_i + \sum_{i=1}^{n} v_i\epsilon_i\epsilon = (x, y, \mathbf{u}, \mathbf{v}), \tag{133}$$

onde $x, y, v_i, u_i \in \mathbb{R}$, para todo $i \in \{1, \ldots, n\}$ e com as propriedades

$$\epsilon^2 = 0 \quad \text{e} \quad \epsilon_i\epsilon_j = 0 \qquad \forall i, j \in \{1, \ldots, n\}.$$

Dados

$$z_1 = (x_1, y_1, \mathbf{u}_1, \mathbf{v}_1) \text{ e } z_2 = (x_2, y_2, \mathbf{u}_2, \mathbf{v}_2),$$

a aritmética dual é definida como

$$z_1 + z_2 = (x_1 + x_2, y_1 + y_2, \mathbf{u}_1 + \mathbf{u}_1, \mathbf{v}_1 + \mathbf{v}_1) \tag{134}$$

$$z_1 - z_2 = (x_1 - x_2, y_1 - y_2, \mathbf{u}_1 - \mathbf{u}_2, \mathbf{v}_2 - \mathbf{v}_2) \tag{135}$$

$$z_1 z_2 = (x_1x_2, x_1y_2 + x_2y_1, x_1\mathbf{u_2} + x_2\mathbf{u_1}, y_1\mathbf{u}_1 + y_2\mathbf{u}_2 + x_2\mathbf{v}_1 + x_1\mathbf{v}_2) \tag{136}$$

$$\frac{z_1}{z_2} = \left(\frac{x_1}{x_2}, \frac{x_2y_1 - x_1y_2}{x_2^2}, \frac{x_2\mathbf{u}_1 - x_1\mathbf{u}_2}{x_2^2}, \frac{x_2\mathbf{v}_1 - x_1\mathbf{v}_2 - (y_1\mathbf{u}_2 + y_2\mathbf{u}_1)}{x_2^2}\right), \tag{137}$$

onde a expressão (136) pode ser intuída da mesma forma que no caso não generalizado, enquanto que (137) é obtido multiplicando o denominador e o numerador por pelo dual generalizado dado pelo produto

$$(x_2, -y_2\epsilon, -\mathbf{u}_2, -\mathbf{v}_2) \cdot (x_2^2, 0, 0, -2y_2\mathbf{u}_2).$$

Usando o mesmo raciocínio aplicado em (129), é possível de definir $f : \widehat{\mathbb{E}}_n^n \to \widehat{\mathbb{E}}_n$, isto é, a extensão de uma função real de n variáveis a qual vamos supor ser ao menos C^2, onde $\mathbf{z} = (z_1, z_2, \ldots, z_n) \in \widehat{\mathbb{E}}_n^n$, com $z_i = (x_i, y_i, \mathbf{u}_i, \mathbf{v}_i)$, $i \in \{1, \ldots, n\}$. Para tanto, defina os elementos

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{y}\epsilon \quad \text{e} \quad \mathbf{z} = \tilde{\mathbf{x}} + \sum_{i=1}^{n} \mathbf{u}_i \epsilon_i.$$

Note que, usando a mesma ideia para se obter (129), podemos escrever

$$\begin{aligned} f(\mathbf{z}) &= f(\tilde{\mathbf{x}}) + \sum_{i=1}^{n} \nabla f(\tilde{\mathbf{x}})^T \mathbf{u}_i \epsilon_i \\ &= (f(\tilde{\mathbf{x}}), \nabla f(\tilde{\mathbf{x}})^T \mathbf{U}), \end{aligned}$$

onde $\mathbf{U} := [\mathbf{u}_1, \ldots, \mathbf{u}_n]$. Por outro lado, usando a mesma ideia para se calcular (121), obtemos

$$f(\tilde{\mathbf{x}}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{y}\epsilon = (f(\mathbf{x}), \nabla f(\mathbf{x})^T \mathbf{y})$$

e

$$\nabla f(\tilde{\mathbf{x}}) = \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})^T \mathbf{y} = (\nabla f(\mathbf{x}), \nabla^2 f(\mathbf{x})\mathbf{y}).$$

Dessa forma, temos

$$\begin{aligned} f(\mathbf{z}) &= f(\tilde{\mathbf{x}}) + \sum_{i=1}^{n} \nabla f(\tilde{\mathbf{x}})^T \mathbf{u}_i \epsilon_i \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{y}\epsilon + \sum_{i=1}^{n} (\nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\mathbf{y}\epsilon)^T \mathbf{u}_i \epsilon_i \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{y}\epsilon + \sum_{i=1}^{n} \nabla f(\mathbf{x})^T \mathbf{u}_i \epsilon_i + \sum_{i=1}^{n} \mathbf{y}^T \nabla^2 f(\mathbf{x}) \mathbf{u}_i \epsilon_i \epsilon \\ &= (f(\mathbf{x}), \nabla f(\mathbf{x})^T \mathbf{y}, \nabla f(\mathbf{x})^T \mathbf{U}, \mathbf{y}^T \nabla^2 f(\mathbf{x}) \mathbf{U}). \end{aligned}$$

Tomando $\mathbf{u}_i = \mathbf{e}_i$, obtemos

$$f(\mathbf{z}) = (f(\mathbf{x}), \nabla f(\mathbf{x})^T \mathbf{y}, \nabla f(\mathbf{x}), \mathbf{y}^T \nabla^2 f(\mathbf{x})). \tag{138}$$

Dessa maneira, no método de LM com correção de segunda ordem apresentado no capitulo 3, tomando $\mathbf{y} = p_{lm}$ e aplicando a relação obtida acima em cada resíduo $r_i(\mathbf{x})$, $i = 1, \ldots, n$, obtemos $r(\mathbf{x})$, $J(\mathbf{x})p_{lm}$, $J(\mathbf{x})$ e $K(\mathbf{x})(p_{lm}, \cdot)$ simultaneamente.

4.3 TESTES NUMÉRICOS

Os nossos testes numéricos utilizamos o banco de problemas NIST Standard Reference Database 140, [35]. Estes problemas fazem parte de um projeto que visa fornecer um conjuntos de dados de referência, com resultados computacionais certificados, para que os usuários possam testar a viabilidade de seus métodos. Nele, existem vários problemas de mínimos quadrados não lineares, em todos são fornecidos a solução, o valor da função objetivo nessa solução e dois pontos iniciais de referência, denotados por x_1 e x_2.

Para os nossos testes numéricos comparamos os resultados obtidos de cinco métodos diferentes, o método de LM, o método de LM com correção de segunda ordem e as modificações M1, M2 e M3. Diferente dos exemplos do capítulo anterior, em nenhum dos testes deste capítulo fizemos uso de controle do tamanho da correção, pois o intuito dos nossos testes é mostrar, de maneira preliminar, a performance da nossa proposta com o mínimo de alterações possíveis.

Resultados numéricos

Organizamos nossos testes em duas tabelas. Na tabela 1 estão aqueles realizados iniciando os algoritmos com ponto x_1 da referência, enquanto que na tabela 2 os algoritmos são iniciados no ponto x_2.

Em ambas as tabelas a coluna que segue abaixo do nome de cada método se refere ao número de iterações que aquele método realizou, no caso de convergência. Quando não há convergência, usamos nc (não convergiu).

Em todos os problemas das tabelas a seguir, utilizamos $D = I$, $\lambda_0 = 10^{-4}$ e $\eta = 0$. Como critério de parada, sendo $(x_k)_{k\geq 0}$ a sequência gerada pelo método, o algoritmo se encerra quando $\|h_c\| \leq \epsilon_1(x_k + \epsilon_1)$ ou $\|\nabla f(x_k)\| < 10^{-8}$, onde tomamos $\epsilon_1 = 10^{-8}$. Já para os parâmetros $PC = \infty$ e $PT = \infty$, isto é são aceitas todas as direções que aumentam o valor da função objetivo.

Tabela 1: Número de iterações quando o ponto inicial é igual a x_1

Problemas	Número de parâmetros	Número de resíduos	LM	LMCS	M1	M2	M3
BoxBod	2	6	30	nc	34	nc	nc
Chwirut1	3	214	36	9	35	23	31
Chwirut2	3	54	36	22	11	3	10
DanWood	2	6	5	5	5	6	6
Gauss1	8	250	5	4	5	5	4
Gauss2	8	250	5	5	5	5	5
Gauss3	8	250	6	6	6	7	6
Kirby2	5	151	9	8	10	10	9
Lanczos1	6	24	262	67*	93*	20	16
Lanczos2	6	24	249	67*	92*	20	17
Lanczos3	6	24	267	69*	97*	23	31
Misra1a	2	14	22	21	14	58	11
Misra1b	2	14	15	18	22	32	15

nc Não convergiu.
* Convergiu para um ponto estacionário diferente do contido em [35].

Como pode ser observado nas tabelas 1 e 2, o método de LM convergiu em todos os casos, porém realizou uma quantidade expressiva de passos nos problemas Lanczos1, Lanczos2, Lanczos3. Em contrapartida, os métodos LMCS e M1, realizaram cerca de 50% de iterações a menos que LM, e o ponto crítico encontrado fornece uma redução bem próxima da contida na referência. Já M2 e M3, além de apresentarem o mesmo valor na função objetivo do banco de questões e fizeram pelo menos 10 vezes menos iterações que LM.

O caso de BoxBod foi o problema onde nossas propostas apresentaram mais falhas. No entanto, onde houve convergência dos métodos propostos, os resultados foram compatíveis com LM. A razão pela qual nossos métodos não convergiram nesse problema, está no fato de que esse problema é muito mal condicionado, sendo necessário alguma escolha pertinente de uma matriz D de dimensionamento, como discutimos no Capítulo 2.

Para os problemas Chwirut1 e Chwirut2, nossas propostas apresentaram menos iterações em todos os casos, sendo que na maioria o número de iterações foi 50% inferior em relação a LM.

Tabela 2: Número de iterações quando o ponto inicial é igual a x_2

Problemas	Número de parâmetros	Número de resíduos	LM	LMCS	M1	M2	M3
BoxBod	2	6	12	12	11	13	10
Chwirut1	3	214	17	17	17	21	16
Chwirut2	3	54	22	14	21	9	15
DanWood	2	6	4	4	4	4	4
Gauss1	8	250	5	4	5	5	4
Gauss2	8	250	5	4	5	5	5
Gauss3	8	250	9	10	9	nc	10
Kirby2	5	151	8	7	8	8	7
Lanczos1	6	24	151	50*	64*	14	15
Lanczos2	6	24	147	50*	63*	14	15
Lanczos3	6	24	168	52*	69*	21	16
Misra1a	2	14	13	10	13	18	6
Misra1b	2	14	18	9	15	16	16

nc Não convergiu.
* Convergiu para um ponto estacionário diferente do contido em [35].

Para DanWood, Gauss1, Gauss2, Gauss3 e Kirby2 nossas propostas são compatíveis com LM, exceto em Gauss3 quando é utilizado M2 iniciado com x_2, nesse caso não houve convergência.

No Misra1a com exceção dos casos de M2 e M3, as propostas são compatíveis com LM. No caso de M2, houve um número maior de iterações em relação ao LM, mais de 50% quando iniciado com x_1, por outro lado, M3 apresentou uma redução de 50% no número de iterações. Já em Misra1b, temos uma redução de 50% no número de iterações quando usa-se LMCS iniciando com x_2, mas temos um aumento de 50% quando usa-se M2 iniciando com x_1. Os restantes das propostas são é compatível com LM para o caso Misra1b.

Portanto, em vista dos testes numéricos realizados juntamente com os teoremas de convergência provados, mostramos que há casos em que o uso da correção de segunda ordem pode ser viável, como é o caso dos problemas Lanczos. Mas também pode apresentar situações compatíveis com LM, ou mesmo desvantajosas, como pode ser visto nas tabelas 1 e 2.

5 CONCLUSÃO

Nesta dissertação consideramos o problema de encontrar um minimizador para problemas de mínimos quadrados não lineares. Para tanto, estudamos alguns métodos clássicos da literatura, como Método de Newton, Gauss-Newton, Levenberg-Marquardt, Modificações do Método de Levenberg-Marquardt, bem como métodos de regiões de confiança, como Ponto de Cauchy e Dogleg. Além do estudo dos métodos anteriormente citados, propomos um novo voltado a resolução de problemas de mínimos quadrados não lineares, explicitado em conjunto com algumas modificações do mesmo, inspiradas em testes numéricos.

Para nossa proposta e suas modificações, realizamos testes numéricos para alguns casos do banco de problemas NIST e comparamos com o método clássico de LM. Mostramos também, por meio de imagens e contagem de número de passos, como os métodos propostos evoluem em comparação com o método clássico de LM, além de provar que, para o caso da função de Rosenbrock, nossa proposta converge para o mínimo global em um único passo, independente do ponto inicial escolhido, o que coloca em perspectiva a validade do uso de tal função para teste de eficiência de métodos de otimização, prática amplamente difundida na área da otimização irrestrita. Além dos resultados numéricos, mostramos a convergências dos métodos por nós propostos.

Também expomos que existe uma maneira eficiente de implementar tanto o método de LM como a nossa proposta, para tanto observamos que a fatoração QR da matriz do sistema de LM é uma alternativa viável, pois a estrutura especial desta matriz permite que as rotações de Givens sejam utilizadas para obter a decomposição no caso em que o passo obtido na iteração é rejeitado, [32], [36]. Além disso, para o cálculo das derivadas de segunda ordem direcionais que aparecem no método novo, utilizando aritmética dual, é possível obter as derivadas de segunda ordem em uma direção dada através da diferenciação automática.

Os testes numéricos mostram que nossas propostas são promissoras, mas que ainda existem aprimoramentos a serem feitos, pois, o método novo exibe sensibilidade em

relação a atualização do parâmetro λ, a escolha adequada da matriz D e nas formas de controlar o tamanho da correção. Além disso, também podem ser aprimoradas as técnicas para diminuir o custo no que diz respeito o uso das derivadas de segunda ordem, neste âmbito, propomos as modificações M2, M3 e M4 as quais demonstraram bons resultados numéricos.

Por fim, as modificações propostas apresentaram bom desempenho, como pode ser visto nas tabelas 1 e 2, porém carecem de uma justificativa teórica, representando um tópico interessante para estudos posteriores.

BIBLIOGRAFIA

[1] AGARWAL, S.; MIERLE K.; Others., *Ceres solver,* <http://ceres-solver.org>.

[2] ARIF, J. N.; CHAUDHURI, B.; CHAUDHURI, R.; RAY, S., *Online levenberg-marquardtalgorithm for neural network based estimation and control of power systems,* In 2009 International Joint Conference on Neural Networks, pp. 199–206, 2009.

[3] ARMANGUÉ, X.; BATLLE, J.; SALVI, J., *A comparative review of camera calibrating methods with accuracy evaluation,* Pattern Recognition, Ezmsford, v. 35, n. 7, p. 1617-1635, 2002.

[4] ARMANGUÉ, X.; SALVI, J., *Overall view regarding fundamental matrix estimation,* Image and Vision Computing, Guildford, v. 21, n. 2, p. 205–220, 2003.

[5] BAEK, J.; KIM, M., *Face recognition using partial least squares components,* Pattern Recognition, 37:1303–1306, 2004.

[6] BAO, J.; YU, C. K. W.; WANG, J.; et al., *Modified inexact Levenberg–Marquardt methods for solving nonlinear least squares problems,* Comput Optim Appl 74, 547–582, 2019.

[7] BRANHAM, R. L. Jr., *Least-squares solution of ill-conditioned systems. II,* 85:1520–1527, 1980.

[8] CLIFFORD, M. A., *Preliminary sketch of biquaternions,* Proceedings of the London Mathematical Society, s1- 4(1):381–395, 1871.

[9] DAVIES, M.; WHITTING I. J., *A modified form of Levenberg's correction,* Em Numerical Methods for Nonlinear Optimization, F. A. Lootsma (ed.) London, Academic Press, pp. 191–201, 1972.

[10] DENG, X.; DING, M.; PENG, S.; YIN, L., *An iterative algorithm for solving ill-conditioned linear least squares problems,* Geodesy and Geodynamics, 6(6):453 –459, 2015.

[11] DORAO, C. A.; JAKOBSEN, H. A., *Application of the least-squares method for solving population balance problems in* $\mathbb{R}^{d+1}$, Chemical Engineering Science, 61(15):5070 – 5081, 2006.

[12] Eldén, L., *Algorithms for the regularization of ill-conditioned least squares problems,* BIT, 17:134–145, 1977.

[13] FAN, J., *Acceleration the Modified Levenberg-Marquardt Method for Nonlienar Equations,* Mathematics of Computation, vol. 83, no. 287, pp. 1173–1187, 2014.

[14] FAN, J., *The Modified Levenberg-Marquardt Method for Nonlinear Equations With Cubic Convergence,* Mathematics of Computation, vol. 81, no. 277, pp. 447–466, 2012.

[15] FAN, J.; YUAN, Y., *On the Quadratic Convergence of the Levenberg-Marquardt Method without Nonsingularity Assumption,* Computing 74, 23–39, 2005.

[16] FLANNERY, B. P.; PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING,W. T., *Numerical Recipes in C: the Art of Scientific Computing,* [S.l.]: Cambridge University Press, 1992.

[17] FLETCHER, R., *Modified Marquardt Subroutine for Non-Linear Least Squares,* United Kingdom: N. p., 1971.

[18] FUKUSHIMA, M.; YAMASHITA, N., *On the rate of convergence of the Levenberg-Marquardt method,* Technical Report 2000-008, Kyoto University, Kyoto, Japão, 2000.

[19] GOLUB, G. H.; VAN LOAN, C. F., *Matrix Computations,* The Johns Hopkins University Press, 1996.

[20] GRAU-SÁNCHEZ, M.; GRAU, A.; DIAZ-BARRERO, J. L., *On Computational Order of Convergence of some Multi-Precision Solvers of Nonlinear Systems of Equations,* arXiv preprint arXiv:1106.0994, 2011.

[21] HARKIN, A. A.; HARKIN, J. B., *Geometry of generalized complex numbers,* Mathematics Magazine, 77(2):118–129, 2004.

[22] HAYAMI, Y.; KUWAHARA, A.; MATSUMOTO, K.; SAKAMOTO, H., *Acceleration and stabilization techniques for the Levenberg-Marquardt method,* IEICE Trans. Fundamentals E88-A, pp. 1971–1978, 2005.

[23] HWANG, J.; KWAK, Y.; YOO, C., *A new damping strategy of Levenberg-Marquardt algorithm for Multilayer Perceptrons,* Neural Network World, v. 21, 2011.

[24] KANTOR, I. L.; SOLODOVNIKOV, A. S., *Hypercomplex Numbers: An Elementary Introduction to Algebras,* Springer, 1989.

[25] KELLEY, C. T., *Iterative Methods for Optimization, Frontiers in Applied Mathematics,* 18, SIAM, Philadelphia, 1999.

[26] LEMEIRE, F., *Bounds for condition numbers of triangular and trapezoid matrices,* BIT, 15(1), pp. 58–64, 1975.

[27] LEVENBERG, K., *A method for the solution of certain problems in least squares,* Quart. Ap. Math. 2, pp. 164–168, 1944.

[28] LOOP, C.; ZHANG, Z., *Computing rectifying homographies for stereo vision,* Conference on Computer Vision and Pattern Recognition, v. 1, p. 125–131, 1999.

[29] LOURAKIS, M. I. A., *Levmar: Levenberg-Marquardt nonlinear least squares algorithms in c/c++,* 2004. Disponível em: <http://users.ics.forth.gr/ lourakis/levmar/>.

[30] MADSEN, K.; NIELSEN; H. B., *Introduction to Optimization and Data Fitting,* Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2010.

[31] MARQUARDT, D., *An algorithm for least squares estimation of nonlinear parameters,* J. Soc. Indust. Ap. Math. 11, pp. 431–441, 1963.

[32] MORÉ, J.J., *The Levenberg-Marquardt Algorithm: Implementation and Theory,* In: Numerical analysis, pp. 105–116. Lecture Notes in Math., v. 630, 1978.

[33] MORÉ, J. J.; SORENSEN, D. C.; HILLSTROM, K. E.; GARBOW, B. S., *The MINPACK Project: in Sources and Development of Mathematical Software,* [S.l.]: Prentice Hall, 1984.

[34] NIELSEN, H. B., *Damping Parameter in Marquardt's Method,* IMM Department of Mathematical Modeling, DTU, Tech. Report IMM-REP-1999-05, 1999.

[35] *NIST Standard Reference Database 140 Last update to Data Content: 2003,* DOI: <http://dx.doi.org/10.18434/T43G6C>.

[36] NOCEDAL, J.; WRIGHT, S. J., *Numerical Optimization,* Springer, 2000.

[37] NOWOZIN, S.; SRA, S.; WRIGHT, S. J., *Optimization for Machine Learning,* The MIT Press, 2011.

[38] OPRISAN, S. A., *An application of the least-squares method to system parameters extraction from experimental data,* Chaos: An Interdisciplinary Journal of Nonlinear Science, 12(1): 27–32, 2002.

[39] POWELL, M. J. D., *Convergence Properties of a Class of Minimization Algorithms,* In: O. L. Mangasarian, R. R. Meyer and S. M. Robinson Eds., Nonlinear Programming 2, Academic Press, New York, pp. 1-27, 1975.

[40] SEBER, G. A. F.; WILD, C. J., *Nonlinear Regression,* John Wiley and Sons, New York, 1989.

[41] SETHNA, J. P.; TRANSTRUM, M. K., *Improvements to the Levenberg-Marquardt algorithm for nonlinear least-squares minimization,* arXiv preprint arXiv:1201.5885v1, 2012.

[42] SHEARER, J. M.; WOLFE, M. A., *ALGLIB: a simple symbolmanipulation package,* Communications of the ACM, New York, v. 28, n. 8, p. 820-825, 1985.

[43] VATTI, V. B. K., *Cubic convergent modified newton's method,* International Journal of Advanced Research, 4:48–52, 2016.