

ALESSANDRA MACHADO DA MOTA

**CONVERGÊNCIA DE ALGORITMOS
PARA PROGRAMAÇÃO NÃO LINEAR**

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências, pelo Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Setor de Ciências Exatas, Universidade Federal do Paraná.

Orientadores:

Profa. Dra. Elizabeth Wegner Karas

Prof. MSc. Ademir Alves Ribeiro

CURITIBA

2005

*Complicar aquilo que é simples é lugar-comum;
tornar simples o que é complicado é criatividade.*

Charles Mingus

Aos meus pais e
a todos os meus mestres,
que ao longo de minha vida
me mostraram o valor da
educação.

Agradecimentos

Meus sinceros agradecimentos:

Aos professores Elizabeth Wegner Karas e Ademir Alves Ribeiro, pela orientação, disponibilidade ao esclarecimento de todas as dúvidas, sugestões feitas durante a escrita e revisão final deste trabalho, bem como ao apoio, à paciência, ao incentivo, à confiança, à dedicação e à amizade demonstrados.

Às professoras Sandra Augusta Santos e Marli Cardia, pela leitura cuidadosa da dissertação, participação na banca examinadora e sugestões as quais enriqueceram nossa proposta.

Ao colega Wilfredo Tomaselli, por compartilhar seus conhecimentos durante nossos estudos semanais para a preparação de seminários, pela ajuda na execução de programas computacionais e na elaboração de apresentações para eventos científicos.

À Universidade Federal do Paraná e, em particular, ao programa de Pós-Graduação em Métodos Numéricos em Engenharia, pela oportunidade que me concederam para participar do mestrado, pela disponibilidade do local de estudo e de suas instalações. Ao financiamento concedido para participar do XXVII CNMAC - Congresso Nacional de Matemática Aplicada e Computacional - realizado em Porto Alegre - RS, em setembro de 2004, onde apresentei o trabalho “Convergência local do algoritmo de Cauchy para otimização irrestrita”.

Ao CNPq pelo apoio financeiro e incentivo dado à pesquisa.

Aos meus pais e a toda minha família, pela compreensão e paciência, que permitiram a estabilidade necessária ao longo deste período.

Resumo

O objetivo deste trabalho é reunir os resultados, já provados, sobre convergência dos principais algoritmos para programação não linear. Para tanto, fazemos, inicialmente, uma discussão sobre diferentes definições de taxas de convergência, comparando-as entre si. Estabelecemos algumas implicações entre estas definições e exibimos alguns exemplos, elaborados especificamente para este fim, para mostrar que outras implicações não são válidas.

Em seguida, discutimos métodos para resolver um problema irrestrito, ou seja, métodos para resolver o problema de minimizar uma função diferenciável sem restrições. Uma grande classe desses algoritmos inicia com um ponto dado e determina uma direção ao longo da qual é possível diminuir o valor da função, e nesta direção calcula um comprimento do passo que permite uma redução razoável do valor da função.

Um método clássico de minimização irrestrita é o método de Cauchy, que corresponde a escolher a direção oposta ao gradiente da função no ponto corrente. Para o caso em que a função é quadrática, convexa e o cálculo do comprimento do passo é feito por busca linear exata, mostramos um resultado original: a sequência gerada pelo algoritmo de Cauchy converge com velocidade *q-linear*, na norma euclidiana, para o único minimizador da função e exibimos a taxa de convergência que depende do número de condicionamento da matriz hessiana da quadrática. O que temos na literatura é a convergência *q-linear* na norma induzida pela hessiana, mas apesar das normas em \mathbb{R}^n serem equivalentes, a velocidade de convergência *q-linear* depende da norma.

Outro método para minimização irrestrita é o método de região de confiança. Apresentamos neste trabalho uma demonstração da convergência global deste método.

O problema geral de programação não linear consiste em minimizar uma função sujeita a restrições, lineares ou não, de igualdade ou desigualdade. Aqui apresentamos métodos para resolver o problema com restrições de igualdade. Mostramos que, sob certas condições, o método de programação quadrática sequencial tem convergência local quadrática. Apresentamos também um algoritmo de filtro globalmente convergente.

Abstract

The purpose of this work is to summarize some results about convergence of the main algorithms used for nonlinear programming. We start by discussing different definitions of the convergence rate. We establish some implications and show that others are not valid, using for this aim examples purposely devised.

We then discuss some methods to solve an unconstrained problem, i.e, methods to minimize a differentiable function without constraints. A large class of algorithms developed to solve these problems usually initiates with a given point and determines a direction and a step length in this direction in order to provide a reasonable reduction of the value of the objective function.

A classical method of unconstrained minimization is the Cauchy method, in which the direction opposite to the gradient of the function at the current point is chosen. For the case where the function is quadratic, convex and the calculation of the length of the step is made by accurate linear search, we show an original result: the sequence generated by the Cauchy algorithm converges *q-linearly*, in the Euclidean norm, to the unique minimizer of the function. We show, in that case, that the convergence rate depends on the condition number of the Hessian matrix of the quadratic function. In the literature we have the *q*-linear convergence for the hessian-induced norm. Despite the fact that two norms in \mathbb{R}^n are equivalent, the *q*-linear convergence is a norm-dependent property.

Another method for unconstrained minimization is the trust-region method, for which we present a proof of global convergence.

The general problem of nonlinear programming consists in minimizing an objective function subject to equality and inequality constraints. In this work we present some methods to solve problems with equality constraints. We show that, under certain conditions, the method of sequential quadratic programming has local quadratic convergence. We also present a globally convergent filter algorithm.

Sumário

Introdução	1
1 Revisão	4
1.1 Produto interno e normas	5
1.2 Função de classe \mathcal{C}^n	8
1.3 Sequências em \mathbb{R}^n	9
1.4 Velocidade de Convergência	11
1.4.1 Convergência q	12
1.4.2 Convergência r	24
1.4.3 Convergência B	27
2 Minimização Irrestrita	34
2.1 O problema	35
2.1.1 Caracterização da solução: condições de otimalidade	35
2.1.2 Algoritmo	37
2.1.3 Métodos de busca unidirecional	39
2.2 Método de Cauchy	43
2.2.1 Convergência global	43
2.2.2 Convergência local	46
2.3 Método de Newton	53
2.3.1 Convergência global	56
2.3.2 Convergência local	58
2.4 Método de Região de Confiança	62
2.4.1 Convergência global	65
2.4.2 Convergência local	73

3	Minimização com restrições de igualdade	75
3.1	Método de programação quadrática sequencial	76
3.1.1	Convergência local	80
3.1.2	Convergência global	82
3.2	Método de Filtro	93
	Conclusão	97
	Apêndice	99
	Referências Bibliográficas	101

Introdução

Otimização pode ser definida, segundo Rao [28], como o processo de encontrar o valor máximo ou mínimo de uma função sob certas circunstâncias. A linguagem utilizada pela otimização para expressar os problemas de uma forma declarativa é conhecida universalmente como *programação matemática*.

O problema geral de otimização é expresso em programação matemática como:

$$\begin{aligned} & \text{minimizar} && f_o(x) \\ & \text{sujeito a} && f_{\mathcal{E}}(x) = 0 \\ & && f_{\mathcal{I}}(x) \leq 0, \end{aligned} \tag{1}$$

onde $f_o : \mathbb{R}^n \rightarrow \mathbb{R}$, $f_{\mathcal{E}} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ com $m < n$ e $f_{\mathcal{I}} : \mathbb{R}^n \rightarrow \mathbb{R}^p$.

Caso o problema seja *maximizar* f_o temos que:

$$\text{maximizar } f_o(x) \Leftrightarrow \text{minimizar } -f_o(x).$$

Portanto, sem perda de generalidade, procuramos resolver o problema (1).

A função f_o é chamada função objetivo e o conjunto

$$\Omega = \{x \in \mathbb{R}^n \mid f_{\mathcal{E}}(x) = 0 \text{ e } f_{\mathcal{I}}(x) \leq 0\}$$

é chamado conjunto viável. As soluções $x^* \in \Omega$ do problema (1) são chamadas minimizadores e os valores correspondentes $f_o(x^*)$ são os mínimos do problema. Conforme as características do conjunto Ω e as propriedades das funções objetivo e restrições, teremos os diferentes problemas de otimização. Por exemplo, as funções envolvidas no problema podem ser contínuas ou não, diferenciáveis ou não, lineares ou não.

O caso particular em que a função objetivo e as restrições são funções lineares é conhecido como um problema de *programação linear* e é resolvido por métodos

específicos, os quais não analisaremos neste trabalho.

Estudaremos problemas onde todas as funções usadas para defini-los são continuamente diferenciáveis, isto é, estudaremos problemas de otimização diferenciável. Mais especificamente, estudaremos problemas nos quais as funções envolvidas são, normalmente, não lineares, isto é, estudaremos o problema geral de *programação não linear diferenciável* (PNL).

Um caso particular do problema de PNL é o problema irrestrito (PI) onde o conjunto Ω é todo o \mathbb{R}^n . O problema irrestrito em comparação com o problema geral de PNL pode ser considerado simples, porque consiste em minimizar uma função sem restrições. O estudo dos métodos que resolvem o problema irrestrito é de fundamental importância em otimização, porque muitos métodos para resolver o problema geral de PNL fazem uso dos métodos que resolvem o PI. Em casos muito simples é possível encontrar diretamente as soluções (quando existem) do problema irrestrito. Mas, quando o número de variáveis ou a complexidade das funções aumentam, as manipulações isoladas são insuficientes para achar sequer pontos candidatos à solução. É necessário, então, apelar para métodos numéricos, ou seja, algoritmos.

Um algoritmo é um procedimento iterativo que, a partir de um ponto inicial x^0 , calcula, conforme determinadas regras, um novo ponto x^1 ; a partir de x^1 , pelas mesmas regras, um ponto x^2 , e assim por diante. Os algoritmos existentes diferem entre si justamente nas regras que determinam o ponto seguinte. Para problemas de programação não linear, os algoritmos normalmente nunca atingem exatamente uma solução, mas geram uma sequência de pontos cujos valores de aderência são soluções do problema. Na prática, termina-se o processo da otimização quando um ponto está suficientemente próximo de uma solução.

A teoria de algoritmos iterativos pode ser dividida em dois aspectos. O primeiro encarrega-se da verificação de que o algoritmo, de fato, gerará uma sequência cujos valores de aderência são soluções do problema. Este aspecto pode ser chamado de análise da convergência global, porque se encarrega de examinar a questão importante, se um algoritmo converge a um ponto ótimo (ou solução) de qualquer ponto de partida. Mais formalmente, dizemos que um algoritmo é globalmente convergente quando todo valor de aderência da sequência gerada por ele é uma solução do problema.

O segundo aspecto pode ser chamado de análise da convergência local. Este

se encarrega de determinar a velocidade de convergência à solução do problema. Não é viável considerar um problema resolvido, simplesmente pelo fato de ter um algoritmo convergente. A convergência pode ser tão lenta, que para fins práticos esse algoritmo não pode ser implementado, porque o tempo que o algoritmo gasta para chegar à solução é exorbitante. Portanto, o estudo teórico da convergência local de algoritmos é de fundamental importância para a implementação de algoritmos eficientes.

O estudo da convergência de algoritmos para programação não linear é o principal objetivo deste trabalho.

Organização dos capítulos.

No Capítulo 1 fazemos uma revisão de alguns conceitos, que mais frequentemente aparecem ao longo do trabalho. Ainda neste primeiro capítulo apresentamos as definições, propriedades e equivalências entre três tipos de velocidade de convergência, e através de exemplos, elaborados por nós, mostramos quando as equivalências não são obtidas. No Capítulo 2 apresentamos três métodos que resolvem o problema de programação não linear irrestrito. Para o método clássico de Cauchy apresentamos um resultado, que não aparece na literatura, para uma função quadrática e convexa. Mostramos que a sequência gerada pelo algoritmo de Cauchy com busca exata converge com velocidade q -linear, na norma euclidiana, e exibimos a taxa de convergência. No Capítulo 3 estudamos dois métodos que resolvem o problema de programação não linear com restrições de igualdade. Ao final apresentamos algumas conclusões e deixamos sugestões para trabalhos futuros.

Capítulo 1

Revisão

Neste primeiro capítulo serão revistos apenas os conceitos matemáticos que, mais frequentemente, vamos usar no desenvolvimento do trabalho. Inicialmente, fixamos a notação e apresentamos algumas definições fundamentais. Na Seção 1.3 apresentamos os principais resultados sobre sequências convergentes em \mathbb{R}^n , visto que podemos considerar os iterados de um algoritmo de programação não linear (PNL) como termos de uma sequência em \mathbb{R}^n . Na Seção 1.4 apresentamos o critério usado para analisar o desempenho dos diferentes algoritmos para PNL. Este critério é a *velocidade de convergência* com que a sequência, gerada por um algoritmo, encontra a solução do problema. Definimos três diferentes tipos de velocidade de convergência e mostramos suas equivalências e propriedades.

Notação. Usaremos letras minúsculas e maiúsculas para denotar, respectivamente, vetores colunas e matrizes. Vetores diferentes em \mathbb{R}^n são denotados por superíndices e subíndices denotam as componentes reais de um vetor, assim:

$$x^k \in \mathbb{R}^n, x^k = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ com } x_i \in \mathbb{R}, i = 1, \dots, n.$$

Denotamos a transposta de uma matriz A por A^T e o produto escalar de dois vetores $u, v \in \mathbb{R}^n$ por $u^T v$. Se a matriz A é *semidefinida positiva*, isto é, $x^T A x \geq 0$ para todo $x \in \mathbb{R}^n$, escrevemos $A \geq 0$. Analogamente, se A é *definida positiva*, isto é, $x^T A x > 0$ para todo $x \neq 0$, escrevemos $A > 0$.

1.1 Produto interno e normas

A principal referência desta seção é o livro de Análise em \mathbb{R}^n [18].

Produto interno. Um *produto interno* é uma regra que faz corresponder a cada par de vetores $x, y \in \mathbb{R}^n$ um número real, indicado por $\langle x, y \rangle$, de tal modo que, para quaisquer $x, x', y \in \mathbb{R}^n$ e $\alpha \in \mathbb{R}$, se tenham:

1. $\langle x, y \rangle = \langle y, x \rangle$;
2. $\langle x + x', y \rangle = \langle x, y \rangle + \langle x', y \rangle$;
3. $\langle \alpha x, y \rangle = \alpha \cdot \langle x, y \rangle = \langle x, \alpha y \rangle$;
4. $x \neq 0 \Rightarrow \langle x, x \rangle > 0$.

Um exemplo importante é o *produto interno canônico* o qual é dado por

$$\langle x, y \rangle = x^T y = x_1 y_1 + \dots + x_n y_n.$$

Norma de um vetor. Dado $x \in \mathbb{R}^n$, escrevemos $\|x\| = \sqrt{\langle x, x \rangle}$, ou seja,

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}.$$

Tem-se $\|x\|^2 = \langle x, x \rangle$, de modo que $\|x\| = 0$ se, e somente se, $x = 0$ e $\|x\| > 0$ se, e somente se, $x \neq 0$.

O número $\|x\|$ chama-se *norma euclidiana* ou *comprimento* do vetor $x \in \mathbb{R}^n$. Vamos denotar a norma euclidiana de um vetor $x \in \mathbb{R}^n$ por $\|x\|_2$ ou l_2 .

Dois vetores $x, y \in \mathbb{R}^n$ são *ortogonais* quando o produto interno entre eles é nulo, isto é, $\langle x, y \rangle = 0$. Evidentemente, o vetor nulo 0 é ortogonal a todos os vetores de \mathbb{R}^n . O vetor $e^i \in \mathbb{R}^n$, que tem uma coordenada igual a 1 na posição i e todas as outras nulas, é ortogonal a e^j , se $i \neq j$.

Teorema 1.1 *Desigualdade de Cauchy-Schwarz*

Para quaisquer $x, y \in \mathbb{R}^n$, tem-se $|\langle x, y \rangle| \leq \|x\| \|y\|$. Vale a igualdade se, e somente se, um dos vetores x, y é um múltiplo escalar do outro.

Prova. [18, pág. 4]. □

O teorema acima é válido para qualquer produto interno, com $\|x\| = \sqrt{\langle x, x \rangle}$.

A norma euclidiana $\|x\|_2 = \sqrt{x^T x}$ goza das seguintes propriedades, onde $x, y \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$ e $|\alpha|$ significa o valor absoluto do número real α :

1. $\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$;
2. $\|\alpha \cdot x\|_2 = |\alpha| \cdot \|x\|_2$;
3. $x \neq 0 \Rightarrow \|x\|_2 > 0$;
4. $|\|x\|_2 - \|y\|_2| \leq \|x - y\|_2$.

De um modo geral, uma *norma* é qualquer função real $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$ que cumpra as três primeiras condições acima estipuladas. Se a norma provém de um produto interno, então também vale a quarta propriedade.

Há uma infinidade de normas que se podem considerar em \mathbb{R}^n . A norma euclidiana é motivada pela fórmula do comprimento de um vetor no plano em coordenadas cartesianas e para noções geométricas ela é a mais natural. Por outro lado, há duas normas que são de manipulação formal mais simples, as quais poderemos utilizar em \mathbb{R}^n , quando houver conveniência. Elas são:

$$\begin{aligned} \|x\|_\infty &= \max\{|x_1|, \dots, |x_n|\}, & (\text{norma do máximo ou } l_\infty) \\ \|x\|_1 &= |x_1| + \dots + |x_n| & (\text{norma da soma ou } l_1). \end{aligned}$$

As seguintes relações entre essas normas são importantes:

$$\begin{aligned} \|x\|_\infty &\leq \|x\|_2 \leq \|x\|_1 \quad \text{e} \\ \frac{\|x\|_1}{\sqrt{n}} &\leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty. \end{aligned}$$

É possível mostrar que duas normas quaisquer $|\cdot|$ e $\|\cdot\|$ em \mathbb{R}^n são sempre equivalentes [18, pág. 18], ou seja, existem $\alpha > 0$ e $\beta > 0$ tais que, para todo $x \in \mathbb{R}^n$:

$$\alpha|x| \leq \|x\| \leq \beta|x|. \tag{1.1}$$

Uma norma arbitrária será denotada simplesmente por $\|\cdot\|$. Usando as relações anteriores e a equivalência de normas (1.1) temos que:

$$\|x\|_2 \geq \alpha\|x\|$$

onde $\alpha = 1$ se $\|\cdot\|$ é a norma euclidiana ou a infinito, e $\alpha = 1/\sqrt{n}$ se $\|\cdot\|$ é a norma l_1 .

A Figura 1.1 mostra os pontos de \mathbb{R}^2 com normas l_1 , euclidiana e infinito iguais a 1.

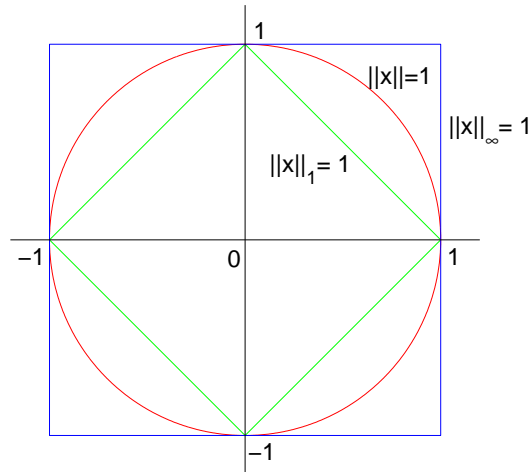


Figura 1.1: Vetores de \mathbb{R}^2 com normas iguais a 1.

Norma de uma matriz. Segundo Gonzaga [12], a norma euclidiana de uma matriz A é dada por

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_{\max}(A^T A)},$$

com $\lambda_{\max}(A)$, e $\lambda_{\min}(A)$ denotando, respectivamente, o maior e menor autovalor de A .

No caso de matrizes simétricas definidas positivas, temos:

$$\begin{aligned} \|A\| &= \lambda_{\max}(A), & \|A^{-1}\| &= \frac{1}{\lambda_{\min}(A)} \\ \|AB\| &\leq \|A\| \|B\|, & A, B &\text{ matrizes} \\ \|Ax\| &\leq \|A\| \|x\|, & A &\text{ matriz, } x \text{ vetor.} \end{aligned}$$

O número de condicionamento κ de uma matriz A é dado por:

$$\kappa = \|A\| \|A^{-1}\| = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

1.2 Função de classe \mathcal{C}^n

Considere $\Omega \subseteq \mathbb{R}^n$ um aberto de \mathbb{R}^n . Segundo Lima [18], temos:

1. Função de classe \mathcal{C}^1 :

Uma função real $f : \Omega \rightarrow \mathbb{R}$ é de classe \mathcal{C}^1 quando existem, em cada ponto $x \in \mathbb{R}^n$, as derivadas parciais $\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x)$ e as n funções $\frac{\partial f}{\partial x_i} : \Omega \rightarrow \mathbb{R}$, assim definidas, são contínuas.

2. Função de classe \mathcal{C}^n

Uma função $f : \Omega \rightarrow \mathbb{R}$ é de classe \mathcal{C}^n quando ela possui derivadas parciais em todos os pontos de Ω e as funções $\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} : \Omega \rightarrow \mathbb{R}$ forem de classe \mathcal{C}^{n-1} , onde n é um inteiro estritamente positivo. Usaremos a notação $f \in \mathcal{C}^n$ para denotar uma função de classe \mathcal{C}^n .

Uma função $f : \Omega \rightarrow \mathbb{R}$ é de classe \mathcal{C}^0 quando ela for contínua, e de classe \mathcal{C}^∞ quando for de classe n para todo $n \geq 0$.

Fórmula de Taylor.

A Fórmula de Taylor para uma função $f : \Omega \rightarrow \mathbb{R}$ é a seguinte:

$$f(a + v) = f(a) + \nabla f(a)^T \cdot v + \frac{1}{2} \nabla^2 f(a) \cdot v^2 + \dots + \frac{1}{p!} \nabla^p f(a) \cdot v^p + r_p(v).$$

Acima, para $v = (\alpha_1 \dots \alpha_n)^T$, escrevemos

$$\begin{aligned} \nabla f(a)^T \cdot v &= \sum_{i=1}^n \frac{\partial f}{\partial x_i}(a) \alpha_i, \\ \nabla^2 f(a) \cdot v^2 &= \sum_{i,j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(a) \alpha_i \alpha_j, \\ \nabla^3 f(a) \cdot v^3 &= \sum_{i,j,k=1}^n \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(a) \alpha_i \alpha_j \alpha_k, \end{aligned}$$

onde

- $\nabla f(a) = \left(\frac{\partial f}{\partial x_1}(a) \quad \dots \quad \frac{\partial f}{\partial x_n}(a) \right)^T$ é o *vetor gradiente* da f ,
- $\nabla^2 f(a) = \left[\frac{\partial^2 f(a)}{\partial x_i \partial x_j} \right]$ $i, j = 1, \dots, n$ é a *matriz hessiana* da f , e assim por diante.

Conforme as hipóteses feitas e os objetivos desejados, temos três situações principais:

1. *Fórmula de Taylor Infinitesimal*

Se f é p vezes diferenciável no ponto a , então $\lim_{v \rightarrow 0} \frac{r_p(v)}{\|v\|^p} = 0$. Neste caso usaremos a notação usual $r_p(v) = o(\|v\|^p)$.

2. *Resto de Lagrange*

Supondo f de classe \mathcal{C}^p , $p+1$ vezes diferenciável no segmento aberto $(a, a+v)$, então existe $\theta \in (0, 1)$, tal que

$$r_p(v) = \frac{1}{(p+1)!} \nabla^{p+1} f(a + \theta v) \cdot v^{p+1}.$$

3. *Resto Integral*

Se f é de classe \mathcal{C}^{p+1} então

$$r_p(v) = \frac{1}{p!} \int_0^1 (1-t)^p \nabla^{p+1} f(a + tv) \cdot v^{p+1} dt.$$

Condição de Lipschitz. Uma função $f : \Omega \rightarrow \mathbb{R}^m$ é *Lipschitziana* se existe uma constante $L > 0$, tal que para todo $x^1, x^2 \in \Omega$, tem-se:

$$\|f(x^1) - f(x^2)\| \leq L \|x^1 - x^2\|.$$

A constante L é chamada constante de *Lipschitz*.

1.3 Sequências em \mathbb{R}^n

A principal referência desta seção é o livro de Análise em \mathbb{R}^n [18].

Uma *sequência* em \mathbb{R}^n é uma aplicação $x : \mathbb{N} \rightarrow \mathbb{R}^n$, definida no conjunto \mathbb{N} dos números naturais. O valor que essa aplicação assume no número k é indicado com x^k e chama-se o *k-ésimo* termo da sequência. Usaremos a notação (x^k) para indicar a sequência cujo *k-ésimo* termo é $x^k \in \mathbb{R}^n$.

Uma *subsequência* de (x^k) é a restrição da sequência a um subconjunto infinito $\mathbb{N}' = \{k_1 < k_2 < \dots < k_i < \dots\} \subset \mathbb{N}$. A subsequência é indicada pela notação $(x^k)_{k \in \mathbb{N}'}$.

Uma sequência (x^k) é *limitada* quando o conjunto dos seus termos é limitado em \mathbb{R}^n , ou seja, quando existe um número real $c > 0$, tal que $\|x^k\| \leq c$ para todo $k \in \mathbb{N}$.

Um ponto $a \in \mathbb{R}^n$ é o limite da sequência de pontos $x^k \in \mathbb{R}^n$ quando, para todo $\epsilon > 0$ dado, é possível obter $k_0 \in \mathbb{N}$, tal que $k > k_0$ implica $\|x^k - a\| < \epsilon$. Neste caso, dizemos que a sequência (x^k) *converge* para a ou *tende* para a , e escrevemos $\lim x^k = a$, $\lim_{k \rightarrow \infty} x^k = a$, ou simplesmente $x^k \rightarrow a$.

Quando existe o limite $a = \lim x^k$, dizemos que a sequência (x^k) é *convergente*. Caso contrário, dizemos que a sequência (x^k) é *divergente*.

Um fato elementar, porém essencial, é que o limite de uma sequência convergente é único, ou seja, se $\lim x^k = a$ e $\lim x^k = b$, então $a = b$. Em particular, se $\lim x^k = a$ e uma subsequência de (x^k) converge para o ponto $b \in \mathbb{R}^n$, então $a = b$. O seguinte teorema é um resultado fundamental relacionado com o problema de otimização:

Teorema 1.2 *Bolzano-Weierstrass*

Toda sequência limitada em \mathbb{R}^n possui uma subsequência convergente.

Prova. [18, pág. 16]. □

Um ponto $a \in \mathbb{R}^n$ é *valor de aderência* de uma sequência de pontos $x^k \in \mathbb{R}^n$ quando alguma subsequência de (x^k) converge para a .

Uma sequência convergente possui um único valor de aderência. A recíproca não vale: a sequência de números reais $(1, 2, 1, 3, 1, 4, 1, 5, \dots)$ possui o único valor de aderência 1, mas não converge. Temos, entretanto, o seguinte teorema.

Teorema 1.3 *Uma sequência limitada em \mathbb{R}^n é convergente se, e somente se, possui um único valor de aderência.*

Prova. [18, pág. 17]. □

Uma sequência (x^k) em \mathbb{R}^n é uma *sequência de Cauchy*, quando para todo $\epsilon > 0$ existe $k_0 \in \mathbb{N}$, tal que $k, r > k_0$ implica $\|x^k - x^r\| < \epsilon$. Isto conduz ao seguinte resultado.

Teorema 1.4 *Uma sequência (x^k) em \mathbb{R}^n é de Cauchy se, e somente se, é convergente.*

Prova. [18, pág. 17]. □

1.4 Velocidade de Convergência

As principais referências desta seção são: [2], usada para definir convergência do tipo B , e [23], usada para definir as convergências do tipo q e r .

Como teremos vários exemplos de seqüências de números reais, denotaremos estas seqüências por (x_k) . Seqüências em \mathbb{R}^n , com $n > 1$, continuaremos denotando por (x^k) .

Um algoritmo precisa, para que seja eficiente, de duas propriedades. Primeiro, que todos os valores de aderência da seqüência gerada pelo algoritmo, independente do ponto inicial, sejam soluções do problema. Neste caso dizemos que o algoritmo é globalmente convergente.

A outra propriedade refere-se à *velocidade de convergência* com que a seqüência gerada pelo algoritmo encontra a solução. Podemos ter algoritmos que geram seqüências convergentes, mas a velocidade de convergência para a solução pode ser tão lenta que, para fins práticos, esses algoritmos tornam-se inviáveis. Precisamos então analisar a velocidade de convergência das seqüências, para comparar a eficiência dos diferentes algoritmos para programação não linear. Assim, se (x^k) é uma seqüência em \mathbb{R}^n convergente para um ponto $x^* \in \mathbb{R}^n$, isto é, $\lim \|x^k - x^*\| = 0$, onde $\|\cdot\|$ denota uma norma qualquer em \mathbb{R}^n , podemos considerar o número $e(x^k) = e_k = \|x^k - x^*\|$ como uma medida de quanto cada iterado está próximo do limite. A rapidez com que a seqüência (e_k) se aproxima de zero é uma medida de *velocidade de convergência* da seqüência (x^k) para x^* . Este estudo é de *convergência local*, porque somente nos interessa o comportamento da distância $\|x^k - x^*\|$ quando k é grande. Assim, se duas seqüências (x^k) e (\bar{x}^k) são iguais a partir de algum índice $k_0 \in \mathbb{N}$, então as características de convergência e velocidade de convergência são iguais para ambas as seqüências (x^k) e (\bar{x}^k) .

O restante deste primeiro capítulo é dedicado ao estudo de três diferentes tipos de *velocidade de convergência*. Definimos cada um deles e mostramos suas propriedades e equivalências. O primeiro tipo de velocidade de convergência que definimos, convergência q , será usado para estudar a convergência local dos algoritmos para programação não linear, que apresentaremos nos próximos capítulos.

1.4.1 Convergência q

As definições deste tipo de convergência q baseiam-se em Nocedal, Wright [23].

Definição 1.5 *Seja (x^k) uma sequência em \mathbb{R}^n . Dizemos que (x^k) converge q -linearmente para um ponto $x^* \in \mathbb{R}^n$ se existe uma constante $r \in [0, 1)$ e um natural $k_0 \in \mathbb{N}$, tais que*

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq r, \quad (1.2)$$

para todo $k \geq k_0$.

O prefixo “ q ” significa “quociente”, porque este tipo de convergência é definido em termos do quociente de erros sucessivos, onde o número $e(x^k) = e_k = \|x^k - x^*\|$ é o erro cometido na iteração k . É fácil ver que se a sequência (x^k) satisfaz (1.2), então $x^k \rightarrow x^*$. De fato, temos $e_{k_0+1} \leq r e_{k_0}$, $e_{k_0+2} \leq r e_{k_0+1} \leq r^2 e_{k_0}$, etc, e de modo geral $e_{k_0+p} \leq r^p e_{k_0}$, para todo $p \in \mathbb{N}$. Como $r \in [0, 1)$, a sequência (e_k) converge para zero.

O valor da constante r é a *taxa de convergência* da sequência e, quanto mais próximo de zero for, mais rápida será a convergência da sequência (x^k) para x^* .

Exemplo 1 *A sequência (x_k) dada por $x_k = 0,5^k$, converge com velocidade q -linear para 0:*

Temos que:

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \frac{|0,5^{k+1}|}{|0,5^k|} = 0,5^{k+1-k} = 0,5.$$

Assim, para $r = 0,5$, temos $\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq r$. A Figura 1.2 ilustra a sequência (x_k)

em três situações, nos planos $k \times x_k$; $k \times e_{k+1}/e_k$ e $e_k \times e_{k+1}$.

Os dois próximos exemplos mostram que velocidade de convergência q -linear depende da norma. Embora as normas em \mathbb{R}^n sejam equivalentes, podemos ter sequências que convergem com velocidade q -linear em uma norma, mas não com a mesma velocidade em outra norma. Estes exemplos foram elaborados, por nós, especificamente para este fim.

O Exemplo 2 será citado no próximo capítulo, no estudo da convergência local do método de Cauchy para função quadrática, para enfatizar que a velocidade

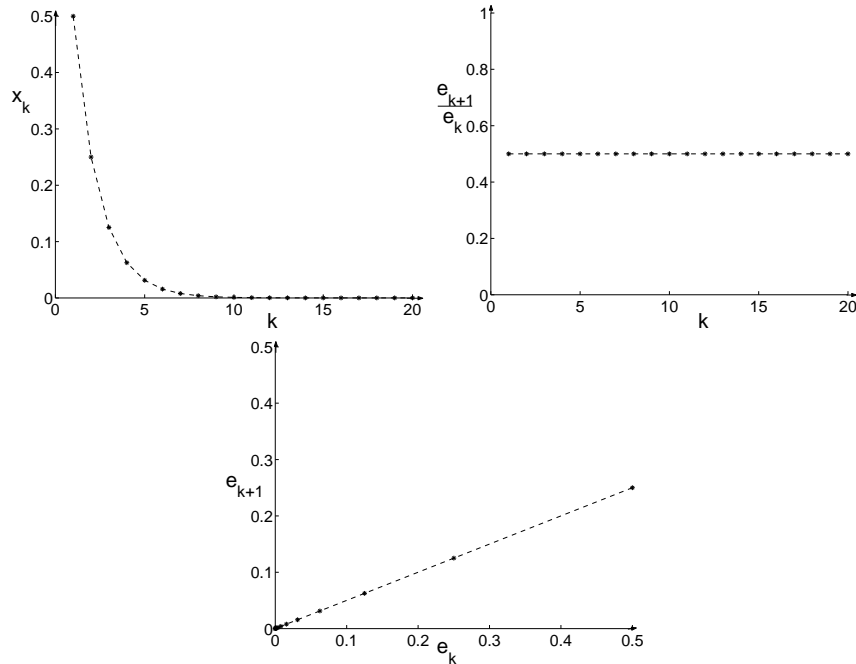


Figura 1.2: Velocidade de convergência: q -linear

de convergência q -linear da sequência (x^k) na norma Q , induzida pela quadrática, não garante a mesma velocidade de convergência na norma euclidiana. A norma Q é definida como $\|x\|_Q^2 = x^T Q x$ com $x \in \mathbb{R}^n$ e $Q \in \mathbb{R}^{n \times n}$.

Exemplo 2 A sequência $(x^k) \in \mathbb{R}^2$ dada por

$$x^{2k} = \left(\frac{1}{(2\sqrt{2})^k} \quad 0 \right)^T \quad e \quad x^{2k+1} = \frac{1}{(2\sqrt{2})^k} \left(\frac{1}{\sqrt{2}} \quad \frac{1}{\sqrt{2}} \right)^T,$$

converge para $(0 \ 0)^T$ com velocidade q -linear na norma Q com $Q = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$, mas não com velocidade q -linear na norma euclidiana.

Na norma euclidiana, temos que

$$\frac{\|x^{2k+1} - x^*\|_2^2}{\|x^{2k} - x^*\|_2^2} = \frac{\frac{1}{(2\sqrt{2})^{2k}} \cdot (1)}{\frac{1}{(2\sqrt{2})^{2k}}} = 1,$$

portanto a sequência (x^k) não tem velocidade q -linear na norma euclidiana. Enquanto que, na norma Q

$$\frac{\|x^{2k+1}\|_Q^2}{\|x^{2k}\|_Q^2} = \frac{\frac{1}{(2\sqrt{2})^{2k} \left(\frac{1}{(\sqrt{2})^2} + \frac{1}{4} \frac{1}{(\sqrt{2})^2} \right)}}{\frac{1}{(2\sqrt{2})^{2k}}} = \frac{5}{8}$$

e

$$\frac{\|x^{2(k+1)}\|_Q^2}{\|x^{2k+1}\|_Q^2} = \frac{\frac{1}{(2\sqrt{2})^{2(k+1)}}}{\frac{1}{(2\sqrt{2})^{2k} \left(\frac{1}{(\sqrt{2})^2} + \frac{1}{4} \frac{1}{(\sqrt{2})^2} \right)}} = \frac{\frac{1}{(2\sqrt{2})^2}}{\frac{1}{(\sqrt{2})^2} + \frac{1}{4} \frac{1}{(\sqrt{2})^2}} = \frac{1}{5},$$

ambas as razões são menores que 1. Logo, a sequência (x^k) converge com velocidade q -linear na norma Q .

A Figura 1.3 mostra a trajetória da sequência (x^k) em relação às duas normas: as elipses representam as curvas de nível da função $x \mapsto \|x\|_Q$, e as circunferências, as curvas de nível da função norma euclidiana, $x \mapsto \|x\|_2$.

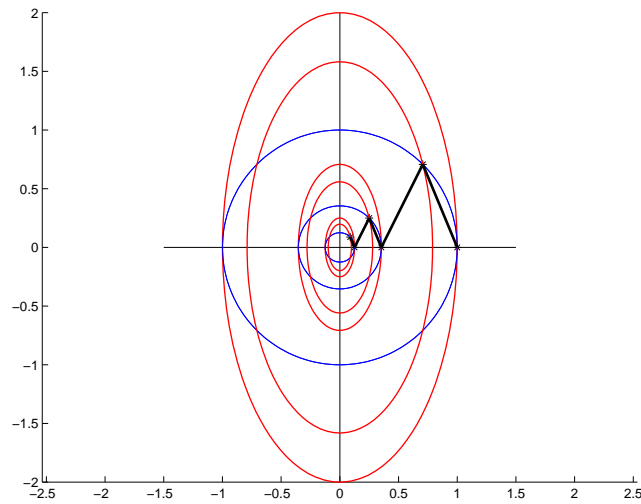


Figura 1.3: Velocidade de convergência q -linear depende da norma.

Note que, essa mesma sequência (x^k) converge com velocidade q -linear na norma do máximo, pois ambas as razões

$$\frac{\|x^{2k+1}\|_\infty}{\|x^{2k}\|_\infty} = \frac{\frac{1}{(2\sqrt{2})^k} \frac{1}{\sqrt{2}}}{\frac{1}{(2\sqrt{2})^k}} = \frac{1}{\sqrt{2}} \quad \text{e} \quad \frac{\|x^{2(k+1)}\|_\infty}{\|x^{2k+1}\|_\infty} = \frac{\frac{1}{(2\sqrt{2})^{k+1}}}{\frac{1}{(2\sqrt{2})^k} \frac{1}{\sqrt{2}}} = \frac{1}{2\sqrt{2}} = \frac{1}{2},$$

são menores que 1.

A Figura 1.4 mostra a trajetória da sequência (x^k) em relação às duas normas: os quadrados representam as curvas de nível da função $x \mapsto \|x\|_\infty$ e as circunferências representam as curvas de nível da função norma euclidiana, $x \mapsto \|x\|_2$.

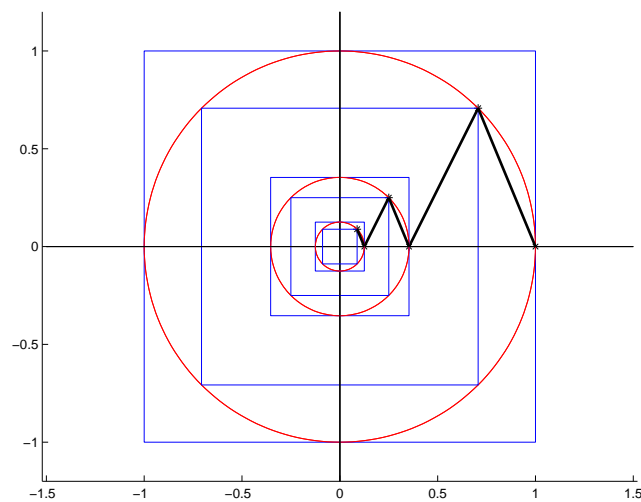


Figura 1.4: Velocidade de convergência q -linear depende da norma.

O seguinte exemplo mostra a situação recíproca: uma sequência que converge com velocidade q -linear na norma euclidiana, mas não com velocidade q -linear na norma do máximo.

Exemplo 3 A sequência $(x^k) \in \mathbb{R}^2$ dada por

$$x^{2k} = \left(\frac{1}{2^k} \quad \frac{1}{2^k} \right)^T \quad \text{e} \quad x^{2k+1} = \left(\frac{1}{2^{k+1}} \quad \frac{1}{2^k} \right)^T$$

converge para $(0 \ 0)^T$ com velocidade q -linear na norma euclidiana, mas não com velocidade q -linear na norma do máximo.

Temos que

$$\frac{\|x^{2k+1}\|_2^2}{\|x^{2k}\|_2^2} = \frac{\frac{1}{(2^{k+1})^2} + \frac{1}{(2^k)^2}}{\frac{1}{(2^k)^2} + \frac{1}{(2^k)^2}} = \frac{\frac{(2^k)^2}{(2^{k+1})^2} + 1}{2} = \frac{\frac{1}{4} + 1}{2} = \frac{5}{8},$$

$$\frac{\|x^{2(k+1)}\|_2^2}{\|x^{2k+1}\|_2^2} = \frac{\frac{1}{(2^{k+1})^2} + \frac{1}{(2^{k+1})^2}}{\frac{1}{(2^{k+1})^2} + \frac{1}{(2^k)^2}} = \frac{2}{1 + \frac{(2^{k+1})^2}{(2^k)^2}} = \frac{2}{1 + 4} = \frac{2}{5}$$

enquanto que

$$\frac{\|x^{2k+1}\|_\infty}{\|x^{2k}\|_\infty} = \frac{\max\left\{\frac{1}{2^{k+1}}, \frac{1}{2^k}\right\}}{\max\left\{\frac{1}{2^k}, \frac{1}{2^k}\right\}} = 1.$$

A Figura 1.5 mostra a trajetória da sequência (x^k) em relação às duas normas: os quadrados representam as curvas de nível da função $x \mapsto \|x\|_\infty$ e as circunferências representam as curvas de nível da função norma euclidiana, $x \mapsto \|x\|_2$.

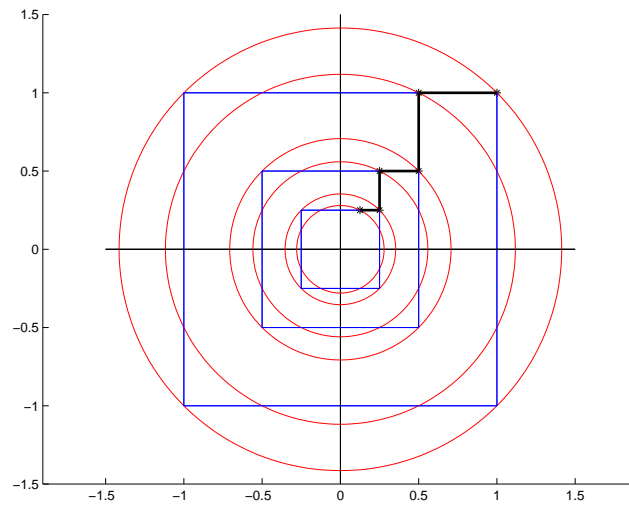


Figura 1.5: Velocidade de convergência q -linear depende da norma.

Definição 1.6 *Seja (x^k) uma sequência em \mathbb{R}^n . Dizemos que (x^k) converge q -superlinearmente para um ponto $x^* \in \mathbb{R}^n$ se*

$$\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0. \quad (1.3)$$

Aqui podemos também provar que a relação (1.3) implica $x^k \rightarrow x^*$.

Exemplo 4 *A sequência (x_k) dada por $x_k = \exp(-k^2)$ converge com velocidade q -superlinear para 0:*

Temos que

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \frac{|\exp(-(k+1)^2)|}{|\exp(-k^2)|} = \exp(-k^2 - 2k - 1 + k^2) = \exp(-2k - 1).$$

Tomando o limite quando $k \rightarrow \infty$:

$$\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \lim_{k \rightarrow \infty} \exp(-2k - 1) = 0.$$

A Figura 1.6 ilustra a sequência (x_k) em três situações, nos planos $k \times x_k$; $k \times e_{k+1}/e_k$ e $e_k \times e_{k+1}$.

A velocidade de convergência da sequência (x^k) pode ser ainda mais rápida. Definimos então:

Definição 1.7 *Seja (x^k) uma sequência em \mathbb{R}^n convergente a x^* . A velocidade de convergência é q -quadrática, se*

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^2} \leq M,$$

para todo k suficientemente grande, onde M é uma constante positiva, não necessariamente menor que 1.

Como vimos, nas definições acima, velocidade de convergência do tipo q depende de r e de M . Entretanto, independentemente destes valores, uma sequência convergente com velocidade q -quadrática realmente converge mais rápido do que uma sequência com velocidade q -linear.

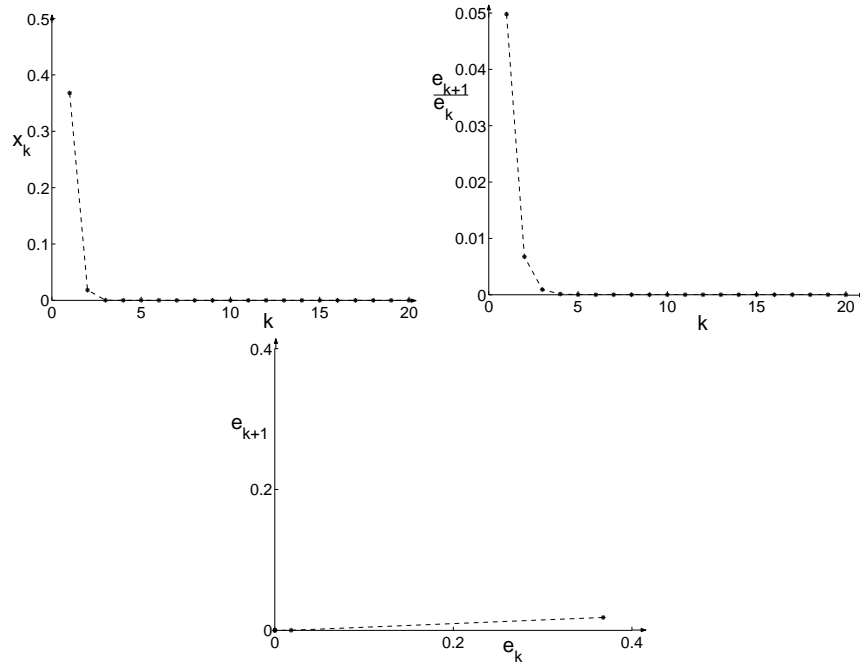


Figura 1.6: Velocidade de convergência: q -superlinear

Exemplo 5 A sequência (x_k) dada por $x_k = 0,5^{2^k}$, converge com velocidade q -quadrática para 0:

Temos que

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} = \frac{|0,5^{2^{k+1}}|}{|0,5^{2^k}|^2} = \frac{|0,5^{2^{k+1}}|}{|0,5^{2^{k+1}}|} = 1.$$

Assim, existe $M = 1$, tal que $\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \leq M$.

A Figura 1.7 ilustra a sequência (x_k) em três situações, nos planos $k \times x_k$; $k \times e_{k+1}/e_k$ e $e_k \times e_{k+1}$.

Podemos definir outras velocidades de convergência (cúbica, quártica, etc), mas não é usual em otimização algoritmos com velocidade superior à quadrática. Em geral, dizemos que a velocidade de convergência q é de ordem p ($p > 1$), se existe uma constante positiva M , tal que:

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^p} \leq M, \tag{1.4}$$

para todo k suficientemente grande e denotaremos por q -ordem p .

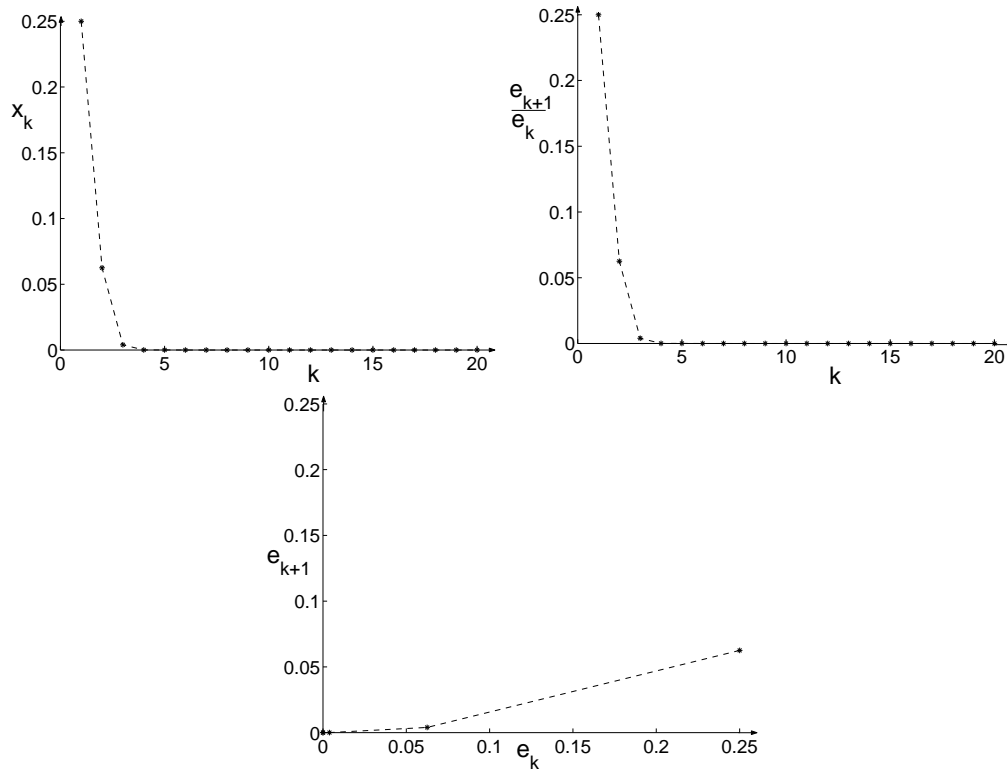


Figura 1.7: Velocidade de convergência: q -quadrática

Obviamente, qualquer sequência que converge com velocidade q -ordem p , também converge com velocidade q -superlinear, e qualquer sequência que converge com velocidade q -superlinear, também converge com velocidade q -linear. De fato, se a sequência (x^k) converge para x^* com velocidade q -ordem p , então existe $M > 0$, tal que $\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^p} \leq M$, para todo $k \geq k_0$ e, podemos escrever

$$0 \leq \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^p} \leq M \|x^k - x^*\|^{p-1}.$$

Tomando o limite quando $k \rightarrow \infty$

$$0 \leq \lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^p} \leq \lim_{k \rightarrow \infty} M \|x^k - x^*\|^{p-1}. \quad (1.5)$$

Denotemos $y_k = \|x^k - x^*\|$ e, como por hipótese $x^k \rightarrow x^*$, temos $y_k \rightarrow 0$. Portanto $y_k^{(p-1)} \rightarrow 0$, pois $p - 1 > 0$. Assim, aplicando o teorema do confronto em (1.5),

obtemos:

$$\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0, \quad (1.6)$$

ou seja, a sequência (x^k) também converge com velocidade *q-superlinear*.

Agora, assumimos que (1.6) se verifica e denotemos $z_k = \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|}$. Assim,

$\lim_{k \rightarrow \infty} z_k = 0$, então dado $\epsilon > 0$ existe $k_0 \in \mathbb{N}$, tal que $|z_k - 0| \leq \epsilon$ para todo $k \geq k_0$.

Escolhendo $\epsilon \in (0, 1)$, temos que

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = |z_k - 0| \leq \epsilon,$$

para todo $k \geq k_0$. Com isto, obtemos que qualquer sequência que converge com velocidade *q-superlinear*, também converge com velocidade *q-linear*.

A afirmação recíproca não é verdadeira. A sequência do Exemplo 1 converge com velocidade *q-linear*, mas não converge com velocidade *q-superlinear*, uma vez que o limite

$$\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = \lim_{k \rightarrow \infty} \frac{0,5^{k+1}}{0,5^k} = 0,5,$$

é diferente de zero. Enquanto que a sequência $x_k = \frac{1}{k^k}$ converge para zero com velocidade *q-superlinear*, mas não com velocidade *q-ordem p* ($p > 1$). De fato, converge superlinearmente para zero, uma vez que o limite

$$\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \lim_{k \rightarrow \infty} \frac{k^k}{(k+1)(k+1)^k}$$

é zero. Porém a razão

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} = \frac{(k^k)^p}{(k+1)(k+1)^k}$$

não é limitada por nenhuma constante $M > 0$, isto é, tende a infinito.

A velocidade de convergência de uma sequência pode ser ainda mais lenta que a *q-linear*. Por exemplo, a sequência $x_k = 1/k$ converge para zero, mas não com velocidade *q-linear*, uma vez que o quociente

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \frac{k}{k+1},$$

tende a 1. Neste caso, dizemos que a velocidade de convergência é *q-sublinear*. Como não existe $r \in (0, 1)$ tal que (1.2) seja satisfeita, a sequência $x_k = 1/k$ converge a zero muito lentamente.

A Figura 1.8 ilustra a sequência (x_k) em três situações, nos planos $k \times x_k$; $k \times e_{k+1}/e_k$ e $e_k \times e_{k+1}$.

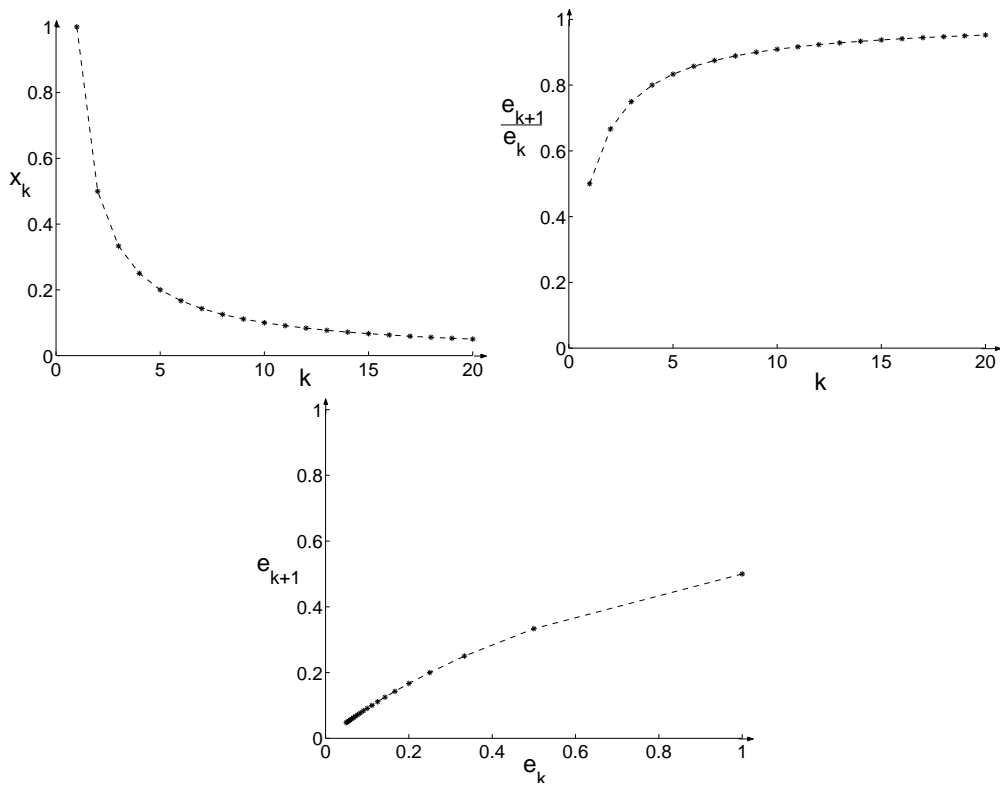


Figura 1.8: Velocidade de convergência: *q-sublinear*

Quando conhecemos a velocidade de convergência de uma sequência (x^k) , podemos determinar a velocidade de convergência de uma outra sequência (y^k) , definida em função de (x^k) , como mostra o seguinte teorema:

Teorema 1.8 *Sejam (x^k) uma sequência em \mathbb{R}^n convergente a x^* e $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ uma aplicação linear injetiva¹. Considere $y^k = Ax^k$ e $y^* = Ax^*$. As afirmações seguintes são verdadeiras:*

1. *Se $x^k \rightarrow x^*$ com velocidade q -quadrática, então $y^k \rightarrow y^*$ com velocidade q -quadrática;*

¹Se a aplicação linear $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ é injetiva, então existe $c > 0$, tal que $\|Ax\| \geq c\|x\|$ para todo $x \in \mathbb{R}^n$.

2. Se $x^k \rightarrow x^*$ com velocidade q -superlinear, então $y^k \rightarrow y^*$ com velocidade q -superlinear.

Prova.

(1) Por hipótese existe $M > 0$, tal que

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^2} \leq M,$$

para todo $k \geq k_0$. Temos que:

$$\frac{\|y^{k+1} - y^*\|}{\|y^k - y^*\|^2} = \frac{\|Ax^{k+1} - Ax^*\|}{\|Ax^k - Ax^*\|^2} = \frac{\|A(x^{k+1} - x^*)\|}{\|A(x^k - x^*)\|^2} \leq \frac{\|A\|\|x^{k+1} - x^*\|}{\|A(x^k - x^*)\|^2}.$$

Como a aplicação linear $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ é injetiva, então existe $c > 0$, tal que $\frac{1}{\|Ax\|^2} \leq \frac{1}{c^2\|x\|^2}$ para todo $x \in \mathbb{R}^n$. Assim:

$$\frac{\|y^{k+1} - y^*\|}{\|y^k - y^*\|^2} \leq \frac{\|A\|\|x^{k+1} - x^*\|}{\|A(x^k - x^*)\|^2} \leq \frac{\|A\|\|x^{k+1} - x^*\|}{c^2\|x^k - x^*\|^2} \leq \frac{\|A\|}{c^2}M.$$

Tomando $\bar{M} = \frac{\|A\|}{c^2}M > 0$, temos que $y^k \rightarrow y^*$ com velocidade q -quadrática.

(2) Por hipótese

$$\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0.$$

Analogamente ao item (1), temos:

$$0 \leq \frac{\|y^{k+1} - y^*\|}{\|y^k - y^*\|} \leq \frac{\|A\|\|x^{k+1} - x^*\|}{c\|x^k - x^*\|}.$$

Tomando o limite quando $k \rightarrow \infty$ e aplicando o Teorema do confronto, obtemos:

$$\lim_{k \rightarrow \infty} \frac{\|y^{k+1} - y^*\|}{\|y^k - y^*\|} = 0.$$

Portanto, $y^k \rightarrow y^*$ com velocidade q -superlinear. \square

O mesmo argumento do Teorema 1.8 não é suficiente para garantir que se $x^k \rightarrow x^*$ com velocidade q -linear, então $y^k \rightarrow y^*$ também com velocidade q -linear.

De fato, se (1.2) é válida para a sequência (x^k) , temos

$$\frac{\|y^{k+1} - y^*\|}{\|y^k - y^*\|} \leq \frac{\|A\| \|x^{k+1} - x^*\|}{c \|x^k - x^*\|} \leq \frac{\|A\|}{c} r,$$

que será menor que 1 se, e somente se, $\frac{\|A\|}{c} \leq \frac{1}{r}$.

O exemplo a seguir mostra uma sequência que converge para zero com velocidade q -linear, mas o limite da razão (1.2) não existe.

Exemplo 6 *Seja a sequência (x_k) definida por*

$$x_{k+1} = \begin{cases} 0, 2x_k & \text{se } k \text{ é par,} \\ 0, 1x_k & \text{se } k \text{ é ímpar.} \end{cases}$$

O menor valor para a razão de convergência é 0, 1. De fato, temos para

$$k \text{ par: } \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \frac{x_{k+1}}{x_k} = \frac{0, 2x_k}{x_k} = 0, 2 \quad \text{e para}$$

$$k \text{ ímpar: } \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \frac{x_{k+1}}{x_k} = \frac{0, 1x_k}{x_k} = 0, 1.$$

Observe que não existe $\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|}$, mas

$$\limsup_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0, 2.$$

Faremos então uma outra definição de velocidade de convergência q -linear, a qual equivale à Definição 1.5:

Definição 1.9 *A sequência (x^k) converge linearmente com taxa de convergência $r < 1$ se, e somente se,*

$$\limsup_{k \rightarrow \infty} \frac{e_{k+1}}{e_k} = r.$$

Observação: $e_k = e(x^k) = \|x^k - x^*\|$ é o erro cometido na iteração k onde $\|\cdot\|$ denota uma norma qualquer em \mathbb{R}^n .

Analogamente, vamos refazer as definições de velocidade *q-superlinear* e *q-quadrática*, as quais também são equivalentes² às Definições 1.6 e 1.7, respectivamente.

Definição 1.10 *A sequência (x^k) converge superlinearmente se, e somente se,*

$$\limsup_{k \rightarrow \infty} \frac{e_{k+1}}{e_k} = 0.$$

Definição 1.11 *A sequência (x^k) converge quadraticamente se, e somente se,*

$$\limsup_{k \rightarrow \infty} \frac{e_{k+1}}{e_k^2} < +\infty.$$

A generalização destas definições é equivalente à (1.4): a ordem de convergência de uma sequência (x^k) , tal que $x^k \rightarrow x^*$, é o supremo dos números $p \in \mathbb{R}$, tais que:

$$\limsup_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^p} < +\infty. \quad (1.7)$$

Difícilmente encontraremos sequências com ordem de convergência superior a 2 (quadrática) em programação não linear. O método de Newton para o problema irrestrito tem convergência quadrática sob certas hipóteses, enquanto que o método de Cauchy tem convergência apenas linear e, frequentemente, com razão próxima de 1 (convergência lenta). A maioria dos algoritmos eficientes na prática exibem convergência superlinear.

1.4.2 Convergência r

As definições deste tipo de convergência r baseiam-se em Nocedal, Wright [23].

Um outro tipo de convergência, caracterizado pelo prefixo “ r ” (de raiz), compara a sequência de erros (e_k) , já definida como

$$e_k = e(x^k) = \|x^k - x^*\|,$$

com uma sequência de escalares não negativos (v_k) , que converge para zero com alguma ordem de convergência do tipo q . Assim, definimos:

²Note que se $\limsup_{k \rightarrow \infty} \frac{e_{k+1}}{e_k} = 0$, então $\lim_{k \rightarrow \infty} \frac{e_{k+1}}{e_k} = 0$.

Definição 1.12 A velocidade de convergência da sequência (x^k) é *r-linear*, se existe uma sequência de escalares não negativos (v_k) , tal que

$$e(x^k) \leq v_k,$$

para todo $k \in \mathbb{N}$ e (v_k) converge para zero com velocidade *q-linear*.

Diz-se que a sequência (e_k) é dominada por (v_k) .

Por exemplo, a sequência

$$x_k = \begin{cases} 1 + (0,5)^k & \text{se } k \text{ é par} \\ 1 & \text{se } k \text{ é ímpar} \end{cases} \quad (1.8)$$

converge para 1 com velocidade *r-linear*, porque a sequência (e_k) é dominada pela sequência $(0,5^k)$, que converge para zero com velocidade *q-linear*. Note que os primeiros termos da sequência são 2; 1; 1,25; 1; 1,0625; etc. O erro realmente aumenta de uma iteração ímpar para uma par. De fato, o erro cometido em cada iteração é dado por $e(x_k) = \|x_k - x^*\|$, assim $e(x_0) = 1$, $e(x_1) = 0$, $e(x_2) = 0,25$, $e(x_3) = 0$, $e(x_4) = 0,0625$, etc. Tal comportamento ocorre quando a taxa de convergência *r-linear* é arbitrariamente alta, mas não ocorre para convergência *q-linear* onde o erro decresce a cada iteração para k suficientemente grande.

O fato de uma sequência convergir com velocidade *r-linear* não garante sua convergência com velocidade *q-linear*. O seguinte exemplo foi elaborado por nós para ilustrar esta situação.

Exemplo 7 A sequência (x_k) , definida como:

$$x_{2k-1} = \frac{1}{4^k} + \frac{1}{4^{2k}}$$

$$x_{2k} = \frac{1}{4^k} - \frac{1}{4^{2k}},$$

converge para zero com velocidade *r-linear*, mas não com velocidade *q-linear*.

Temos:

$$\frac{x_{2k}}{x_{2k-1}} = \frac{\frac{4^k - 1}{4^{2k}}}{\frac{4^k + 1}{4^{2k}}} = \frac{4^k - 1}{4^k + 1} \rightarrow 1.$$

Enquanto que:

$$x_{2k-1} = \frac{1}{4^k} + \frac{1}{4^{2k}} < \frac{1}{4^k} + \frac{1}{4^k} = \frac{2}{2^{2k}} = \frac{1}{2^{2k-1}} \quad e$$

$$x_{2k} = \frac{1}{4^k} - \frac{1}{4^{2k}} < \frac{1}{4^k} = \frac{1}{2^{2k}}.$$

Tomando $v_k = \frac{1}{2^k}$, temos

$$\frac{v_{k+1}}{v_k} = \frac{\frac{1}{2^{k+1}}}{\frac{1}{2^k}} = \frac{1}{2^{k+1}} \cdot \frac{2^k}{1} = \frac{1}{2},$$

para todo $k \in \mathbb{N}$. Assim, existe uma sequência de escalares não negativos (v_k) que converge para zero com velocidade q -linear, tal que

$$e(x_k) = \|x_k - x^*\| \leq v_k \quad \text{para todo } k \in \mathbb{N}.$$

Portanto, a velocidade de convergência da sequência (x_k) é r -linear.

Analogamente ao tipo de convergência q , definiremos velocidade de convergência r -superlinear e r -quadrática.

Definição 1.13 *Seja (x^k) uma sequência em \mathbb{R}^n convergente a x^* . A velocidade de convergência da sequência (x^k) é r -superlinear, se existe uma sequência de escalares não negativos (v_k) , tal que*

$$e(x^k) \leq v_k,$$

para todo $k \in \mathbb{N}$ e (v_k) converge para zero com velocidade q -superlinear.

Definição 1.14 *Seja (x^k) uma sequência em \mathbb{R}^n convergente a x^* . A velocidade de convergência da sequência (x^k) é r -quadrática, se existe uma sequência de escalares não negativos (v_k) , tal que*

$$e(x^k) \leq v_k,$$

para todo $k \in \mathbb{N}$ e (v_k) converge para zero com velocidade q -quadrática.

O próximo teorema relaciona velocidade de convergência do tipo q com velocidade de convergência do tipo r .

Teorema 1.15 *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ uma aplicação Lipschitziana, com $y^k = f(x^k)$ e $y^* = f(x^*)$. Assim,*

1. *se a sequência (x^k) converge para x^* com velocidade q -linear, então a sequência (y^k) converge para y^* com velocidade r -linear;*
2. *se a sequência (x^k) converge para x^* com velocidade q -superlinear, então a sequência (y^k) converge para y^* com velocidade r -superlinear;*
3. *se a sequência (x^k) converge para x^* com velocidade q -quadrática, então a sequência (y^k) converge para y^* com velocidade r -quadrática.*

Prova.

1. Por hipótese, como f é Lipschitziana, existe $L > 0$ tal que

$$\|f(x^k) - f(x^*)\| \leq L\|x^k - x^*\|,$$

para todo $x^k \in \mathbb{R}^n$. Definindo $v_k = L\|x^k - x^*\|$, temos

$$\|y^k - y^*\| \leq v_k.$$

Por hipótese, a sequência (v_k) converge com velocidade q -linear para zero. Portanto, a sequência (y^k) converge para y^* com velocidade r -linear.

De modo análogo, demonstra-se os itens (2) e (3). □

1.4.3 Convergência B

Neste tipo de convergência, devido a Bertsekas [2, pág. 64], a sequência $(e(x^k))$ já definida como

$$e(x^k) = e_k = \|x^k - x^*\|$$

é comparada com a progressão geométrica

$$\beta^k, \quad k = 0, 1, \dots,$$

onde $\beta \in (0, 1)$. Assim, definimos:

Definição 1.16 A sequência $(e(x^k))$ converge linearmente ou geometricamente, se existem $q > 0$ e $\beta \in (0, 1)$, tais que para todo $k \in \mathbb{N}$

$$e(x^k) \leq q\beta^k.$$

A Definição 1.9 implica nesta última definição de convergência linear, segundo Bertsekas. De fato, pela Definição 1.9, temos que:

$$\limsup_{k \rightarrow \infty} \frac{e_{k+1}}{e_k} \leq \beta < 1.$$

Seja $\gamma \in (\beta, 1)$. Então, existe $k_0 \in \mathbb{N}$, tal que $\frac{e_{k+1}}{e_k} \leq \gamma$, para todo $k \geq k_0$. Assim, $e_{k_0+p} \leq \gamma^p e_{k_0}$, ou seja, $e_k \leq \gamma^k (\gamma^{-k_0}) e_{k_0}$, para todo $k \geq k_0$.

Seja $q = \max\{\frac{e_j}{\gamma^j}, j = 1, \dots, k_0\}$. Assim, $e_j \leq q\gamma^j$, para todo $j = 1, \dots, k_0$, e para $k \geq k_0$, temos $e_k \leq \gamma^k \left(\frac{e_{k_0}}{\gamma^{k_0}}\right) \leq \gamma^k q$. Mas a afirmação recíproca não é verdadeira.

Podemos ter sequências que convergem com velocidade linear do tipo Bertsekas, mas não convergem com velocidade q -linear. O exemplo abaixo foi elaborado por nós para ilustrar esta situação.

Exemplo 8 Sejam $\gamma = \frac{1}{2}$ e $q = 1$. Vamos construir $e_k \leq \frac{1}{2^k}$, tal que $\frac{e_{2k}}{e_{2k-1}} \rightarrow 1$.



Figura 1.9: Velocidade linear, segundo Bertsekas, não implica em velocidade q -linear.

Defina:

$$e_{2k-1} = \frac{1}{4^k} + \frac{1}{4^{2k}} \quad \text{e} \quad e_{2k} = \frac{1}{4^k} - \frac{1}{4^{2k}}.$$

Deste modo, temos:

$$\frac{e_{2k}}{e_{2k-1}} = \frac{\frac{4^k - 1}{4^{2k}}}{\frac{4^k + 1}{4^{2k}}} = \frac{4^k - 1}{4^k + 1} \rightarrow 1.$$

Concluimos, então, que a classe de sequências linearmente convergentes, segundo Bertsekas, é maior. E sempre será possível mostrar que convergência linear, segundo

Bertsekas, é obtida, se para algum $\beta \in (0, 1)$, tem-se

$$\limsup_{k \rightarrow \infty} \frac{e(x^{k+1})}{e(x^k)} \leq \beta,$$

isto é, assintoticamente, o erro é dominado por um fator menor que β .

Definição 1.17 *A sequência $(e(x^k))$ converge superlinearmente se para todo $\beta \in (0, 1)$ existe $q > 0$ tal que*

$$e(x^k) \leq q\beta^k,$$

para todo $k \in \mathbb{N}$.

Podemos mostrar que a classe de sequências que convergem superlinearmente, segundo Bertsekas, é maior que a classe de sequências que convergem com velocidade q -superlinear, ou seja, a Definição 1.6 implica na Definição de convergência superlinear, segundo Bertsekas. De fato, da Definição 1.6, temos que:

$$\lim_{k \rightarrow \infty} \frac{e_{k+1}}{e_k} = 0.$$

Assim, dado $\beta \in (0, 1)$, existe $k_0 \in \mathbb{N}$, tal que $\frac{e_{k+1}}{e_k} \leq \beta$, para todo $k \geq k_0$. Analogamente, ao exposto acima de que convergência q -linear implica na convergência linear, segundo Bertsekas, concluímos que $e_k \leq q\beta^k$ para todo $k \geq 1$ e $\beta \in (0, 1)$.

Como na convergência linear, segundo Bertsekas, aqui também a afirmação recíproca não é verdadeira. Para ilustrar este fato, elaboramos o seguinte exemplo.

Exemplo 9 *Considere a sequência (e_k) dada por*

$$\begin{aligned} e_{2k-1} &= \frac{1}{\alpha_k} + \frac{1}{\alpha_k^2} = \frac{\alpha_k + 1}{\alpha_k^2} \\ e_{2k} &= \frac{1}{\alpha_k} - \frac{1}{\alpha_k^2} = \frac{\alpha_k - 1}{\alpha_k^2}, \end{aligned} \tag{1.9}$$

onde $\alpha_k = 2^{2^k}$ e $k \geq 1$.

Dado $\beta \in (0, 1)$, existe $m \in \mathbb{N}$ tal que $\frac{1}{2^m} < \beta$.

Assim $\frac{1}{\beta} < 2^m$, donde

$$\frac{1}{2^{2^k} \beta^k} < \frac{1}{2^{2^k - mk}} \rightarrow 0,$$

porque $\frac{1}{2^{2^k} \beta^k} \leq q$ é limitada, pois converge. Portanto, $e_k \leq \frac{1}{2^{2^k}} \leq q\beta^k$. Além disso,

$$\frac{e_{2k}}{e_{2k-1}} = \frac{\alpha_k - 1}{\alpha_k + 1} = \frac{1 - \frac{1}{\alpha_k}}{1 + \frac{1}{\alpha_k}} \rightarrow 1.$$

No entanto, sempre será possível mostrar que convergência superlinear, segundo Bertsekas, é obtida se

$$\limsup_{k \rightarrow \infty} \frac{e(x^{k+1})}{e(x^k)} = 0.$$

Definição 1.18 *A sequência $(e(x^k))$ converge, pelo menos superlinearmente, com ordem $p > 1$, se existe $q > 0$, $\beta \in (0, 1)$ e $p > 1$, tais que*

$$e(x^k) \leq q\beta^{p^k},$$

para todo $k \in \mathbb{N}$.

O seguinte teorema garante que convergência do tipo q , com ordem $p > 1$, implica na convergência de ordem $p > 1$, segundo Bertsekas.

Teorema 1.19 *A convergência do tipo q , com ordem $p > 1$, implica na convergência de ordem $p > 1$, segundo Bertsekas.*

Prova. Pela equação (1.7), temos que

$$\limsup_{k \rightarrow \infty} \frac{e_{k+1}}{e_k^p} < \infty,$$

existe $k_1 \in \mathbb{N}$, $M > 0$, tal que $e_{k+1} \leq M(e_k)^p$, para todo $k \geq k_1$ e existe $k_0 \geq k_1$, tal que $M^{\frac{1}{p-1}} e_{k_0} < 1$. Assim,

$$\begin{aligned} e_{k_0+1} &\leq M(e_{k_0})^p \\ e_{k_0+2} &\leq MM^p(e_{k_0})^{p^2} \\ e_{k_0+3} &\leq MM^pM^{p^2}(e_{k_0})^{p^3} \\ &\vdots \\ e_{k_0+j} &\leq MM^p \dots M^{p^{j-1}}(e_{k_0})^{p^j}, \end{aligned}$$

onde

$$M^{(1+p+\dots+p^{j-1})}(e_{k_0})^{p^j} = M^{\frac{p^j-1}{p-1}}(e_{k_0})^{p^j} = \frac{1}{M^{\frac{1}{p-1}}}(M^{\frac{1}{p-1}})^{p^j}(e_{k_0})^{p^j} = \frac{1}{M^{\frac{1}{p-1}}}(M^{\frac{1}{p-1}}e_{k_0})^{p^j}.$$

Portanto, $e_k \leq \frac{1}{M^{\frac{1}{p-1}}}(M^{\frac{1}{p-1}}e_{k_0})^{p^{k-k_0}}$ para todo $k \geq k_0$. Sejam $q_1 = \frac{1}{M^{\frac{1}{p-1}}}$,

$\beta_1 = M^{\frac{1}{p-1}}e_{k_0}$. Então, $e_k \leq q_1(\beta_1)^{p^{k-k_0}} = q_1(\beta_1)^{p^k p^{-k_0}} = q_1[(\beta_1)^{p^{-k_0}}]^{p^k}$, para todo $k \geq k_0$.

Sejam $\beta = (\beta_1)^{p^{-k_0}}$ e $q = \max(\{\frac{e_j}{\beta^{p^j}}, j = 1, \dots, k_0\} \cup \{q_1\})$. Assim, $e_k \leq q\beta^{p^k}$, para todo $k \geq k_0$ e $e_k \leq q\beta^{p^k}$, $k = 1, \dots, k_0$. \square

Do mesmo modo que as definições anteriores, de velocidade de convergência, segundo Bertsekas, a afirmação recíproca não é verdadeira. Vamos considerar a mesma sequência (e_k) dada por (1.9) com $\alpha_k = 2^{2^k}$ e $k \geq 1$. Por definição, $e_j \leq \frac{1}{2^{2^j}}$ para j par. Para $j = 2k - 1$, a desigualdade $e_j \leq \frac{1}{2^{2^j}}$, equivale a

$$\frac{\alpha_k + 1}{\alpha_k^2} \leq \frac{1}{\alpha_k^{\frac{1}{2}}}, \quad \text{ou seja,} \quad \alpha_k + 1 \leq \alpha_k^{\frac{3}{2}}.$$

Mas $\alpha_k^{\frac{1}{2}} = 2^{2^{2k-1}} > 1$, logo $\alpha_k^{\frac{3}{2}} > \alpha_k$, e então $\alpha_k^{\frac{3}{2}} - \alpha_k > 0$. Como são inteiros, $\alpha_k^{\frac{3}{2}} - \alpha_k \geq 1$. Vejamos agora que $\frac{e_{2k}}{(e_{2k-1})^2} \rightarrow \infty$. De fato, temos:

$$\frac{e_{2k}}{(e_{2k-1})^2} = \frac{\frac{\alpha_k - 1}{\alpha_k^2}}{\left(\frac{\alpha_k + 1}{\alpha_k^2}\right)^2} = \frac{\alpha_k - 1}{\alpha_k^2} \cdot \frac{\alpha_k^4}{(\alpha_k + 1)^2} = \alpha_k^2 \frac{\alpha_k - 1}{(\alpha_k + 1)^2} = \frac{\alpha_k - 1}{\left(1 + \frac{1}{\alpha_k}\right)^2} \rightarrow \infty.$$

No entanto, é possível mostrar que convergência superlinear com ordem $p > 1$, segundo Bertsekas, é obtida se

$$\limsup_{k \rightarrow \infty} \frac{e(x^{k+1})}{e(x^k)} < \infty.$$

Vejam agora que a relação entre velocidade de convergência tipo *r-linear* com velocidade *linear*, segundo Bertsekas, é mais forte que a relação desta última com velocidade *q-linear*, elas são realmente equivalentes. O seguinte teorema demonstra esta relação.

Teorema 1.20 *Uma sequência $(x^k) \in \mathbb{R}^n$ converge com velocidade *r-linear* se, e somente se, converge com velocidade *linear*, segundo Bertsekas.*

Prova. Se (x^k) converge com velocidade *r-linear*, então existe uma sequência de escalares não negativos (v_k) , tal que $e(x^k) \leq v_k$ e

$$\frac{v_{k+1}}{v_k} \leq \gamma,$$

para todo $k \geq k_0$ e com $\gamma \in (0, 1)$. Assim, $v_{k_0+p} \leq \gamma^p v_{k_0}$, ou seja, $v_k \leq \gamma^k (\gamma^{-k_0}) v_{k_0}$, para todo $k \geq k_0$.

Seja $q = \max\{\frac{v_j}{\gamma^j}, j = 1, \dots, k_0\}$. Assim, $v_j \leq q\gamma^j$, para todo $j = 1, \dots, k_0$ e para $k \geq k_0$, temos $v_k \leq \gamma^k \left(\frac{v_{k_0}}{\gamma^{k_0}}\right) \leq \gamma^k q$. Como $\gamma \in (0, 1)$, $q > 0$ e para todo $k \in \mathbb{N}$, temos $e_k \leq v_k \leq \gamma^k q$, então a velocidade de convergência da sequência (x^k) é *linear*, segundo Bertsekas.

Por outro lado, seja $v_k = q\beta^k$, então

$$\frac{v_{k+1}}{v_k} = \frac{q\beta^{k+1}}{q\beta^k} = \beta.$$

Assim, com $r = \beta$, temos que (v_k) converge com velocidade *q-linear* para zero e $e_k \leq q\beta^k = v_k$. Portanto, a sequência (x^k) converge com velocidade *r-linear*. \square

Enunciaremos o teorema que garante que velocidade de convergência com ordem $p > 1$, segundo Bertsekas, é equivalente à velocidade de convergência do tipo *r* com ordem $p > 1$.

Teorema 1.21 *Uma sequência $(x^k) \in \mathbb{R}^n$ converge com velocidade do tipo *r* com ordem $p > 1$ se, e somente se, converge com velocidade com ordem $p > 1$, segundo Bertsekas.*

Prova. A sequência $e(x^k) = \|x^k - x^*\|$ converge com ordem $p > 1$, segundo Bertsekas, se existe $q > 0$, $\beta \in (0, 1)$ e $p > 1$, tal que para todo $k \in \mathbb{N}$:

$$e(x^k) = \|x^k - x^*\| \leq q\beta^{p^k}.$$

Seja $v_k = q\beta^{p^k}$, então:

$$\frac{\|v_{k+1} - v^*\|}{\|v_k - v^*\|^p} = \frac{\|q\beta^{p^{k+1}} - 0\|}{\|q\beta^{p^k} - 0\|^p} = \frac{|q\beta^{p^{k+1}}|}{|q\beta^{p^k}|^p} = \frac{|q\beta^{p^{k+1}}|}{|q^p \beta^{p^k p}|} = \frac{1}{q^{p-1}}.$$

Assim, com $M = \frac{1}{q^{p-1}} > 0$ temos que (v_k) converge para zero com velocidade do tipo q com ordem $p > 1$ e $e_k \leq q\beta^{p^k} = v_k$. Portanto, a sequência (x^k) converge com velocidade do tipo r com ordem $p > 1$.

A recíproca segue o mesmo raciocínio que usamos para demonstrar o Teorema 1.19 no qual convergência do tipo q com ordem $p > 1$ implica na convergência, segundo Bertsekas, de ordem $p > 1$. \square

Para analisar a velocidade de convergência dos algoritmos, que estudaremos nos próximos capítulos, vamos utilizar o tipo de convergência q .

Capítulo 2

Minimização Irrestrita

Neste capítulo apresentamos um resultado, que não aparece na literatura, para um problema de minimização de uma função quadrática convexa. O resultado é sobre a convergência local da sequência gerada por um algoritmo clássico de minimização irrestrita, o algoritmo de Cauchy. Mostramos que essa sequência converge com velocidade *q-linear*, na norma euclidiana, para a solução do problema, e exibimos a taxa de convergência.

Estrutura do Capítulo. Na primeira seção, apresentamos o problema de minimização irrestrita e demonstramos as condições necessárias e suficientes de otimalidade. Escrevemos um modelo geral de algoritmo para minimização irrestrita e, para resolver o subproblema que aparece no algoritmo, apresentamos dois métodos de busca unidirecional: a busca exata e a busca de Armijo. Na Seção 2.2, apresentamos o método clássico de Cauchy e estudamos sua convergência global e local. Sobre a convergência local do algoritmo de Cauchy para função quadrática, mostramos no Teorema 2.13 um resultado, o qual não aparece na literatura, que garante velocidade de convergência *q-linear* da sequência com busca linear exata na norma euclidiana. Na Seção 2.3, apresentamos o método de Newton e mostramos que, com busca de Armijo, o algoritmo de Newton é globalmente convergente. Provamos, também, que o método de Newton “puro” tem convergência local quadrática. Na Seção 2.4, apresentamos o método de região de confiança e demonstramos que este método é globalmente convergente. Citamos um importante teorema, o qual garante convergência local quadrática para o método região de confiança.

2.1 O problema

Um caso particular do problema geral de programação não linear é o problema de minimização irrestrita, cuja formulação matemática é a seguinte:

$$(P) \quad \begin{array}{l} \text{minimizar } f(x) \\ x \in \mathbb{R}^n \end{array}$$

com $f : \mathbb{R}^n \rightarrow \mathbb{R}$ continuamente diferenciável. A função f é chamada *função objetivo*.

2.1.1 Caracterização da solução: condições de otimalidade

Resolver o problema (P) significa encontrar $x^* \in \mathbb{R}^n$ tal que $f(x^*) \leq f(x)$ para todo $x \in \mathbb{R}^n$. O ponto x^* pode ser caracterizado como:

- Minimizador global: quando $f(x^*) \leq f(x)$ para todo $x \in \mathbb{R}^n$.
- Minimizador global estrito: quando $f(x^*) < f(x)$ para todo $x \in \mathbb{R}^n$ com $x \neq x^*$.
- Minimizador local: quando existe uma vizinhança V de x^* tal que $f(x^*) \leq f(x)$ para todo $x \in V$.
- Minimizador local estrito: quando existe uma vizinhança V de x^* tal que $f(x^*) < f(x)$ para todo $x \in V$ com $x \neq x^*$.

Assim, podemos afirmar que minimizador global implica em minimizador local, mas a recíproca nem sempre será verdadeira.

Em particular se f é duas vezes diferenciável, podemos dizer que x^* é minimizador local (ou minimizador local estrito), examinando o vetor gradiente $\nabla f(x^*)$ e a matriz hessiana $\nabla^2 f(x^*)$ da função f .

As condições necessárias de otimalidade são obtidas, assumindo que x^* é um minimizador local, e com isso provando fatos sobre $\nabla f(x^*)$ e $\nabla^2 f(x^*)$.

Essas demonstrações, das condições necessárias e suficientes de otimalidade, são encontradas em vários livros. Nossa discussão baseia-se em notas de aula, não publicadas, de Gonzaga [12].

Teorema 2.1 *Condição necessária de otimalidade de primeira ordem.*

Se \bar{x} é minimizador local de (P), então $\nabla f(\bar{x}) = 0$.

Prova. Suponha que \bar{x} é minimizador local. Seja $d \in \mathbb{R}^n$ uma direção arbitrária, $d \neq 0$. Temos:

$$f(\bar{x} + \lambda d) = f(\bar{x}) + \lambda \nabla f(\bar{x})^T d + o(\lambda).$$

Vem:

$$\nabla f(\bar{x})^T d = \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} - \frac{o(\lambda)}{\lambda}. \quad (2.1)$$

Por definição de minimizador local, para λ suficientemente pequeno,

$$f(\bar{x} + \lambda d) - f(\bar{x}) \geq 0.$$

Fazendo λ tender a zero, pela direita na expressão (2.1), obtemos:

$$\nabla f(\bar{x})^T d \geq 0.$$

Concluimos que, como d é arbitrária

$$\nabla f(\bar{x})^T d \geq 0,$$

para todo $d \in \mathbb{R}^n$. Isto somente é possível se $\nabla f(\bar{x}) = 0$, completando a demonstração. \square

Observação: Um ponto $x \in \mathbb{R}^n$ tal que $\nabla f(x) = 0$ é um ponto estacionário da função f , podendo ser um ponto de máximo, mínimo ou de sela.

Teorema 2.2 *Condição necessária de otimalidade de segunda ordem.*

Seja f uma função de classe \mathcal{C}^2 e \bar{x} minimizador local de (P) , então a hessiana $\nabla^2 f(\bar{x})$ é semi-definida positiva.

Prova. Considere \bar{x} minimizador local de (P) . Temos

$$f(\bar{x} + \lambda d) = f(\bar{x}) + \lambda \nabla f(\bar{x})^T d + \frac{\lambda^2}{2} d^T \nabla^2 f(\bar{x}) d + o_2(\lambda),$$

com $\lim_{\lambda \rightarrow 0} \frac{o_2(\lambda)}{\lambda^2} = 0$. Pelo Teorema 2.1, $\nabla f(\bar{x}) = 0$. Assim,

$$\frac{d^T \nabla^2 f(\bar{x}) d}{2} = \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda^2} - \frac{o_2(\lambda)}{\lambda^2}. \quad (2.2)$$

Para λ suficientemente pequeno, $f(\bar{x} + \lambda d) - f(\bar{x}) \geq 0$ por hipótese. Fazendo λ tender a zero, na expressão (2.2), temos:

$$\frac{d^T \nabla^2 f(\bar{x}) d}{2} \geq 0.$$

Concluimos que, para todo $d \in \mathbb{R}^n$, $d^T \nabla^2 f(\bar{x}) d \geq 0$, o que é a definição de matriz semi-definida positiva, completando a demonstração. \square

Teorema 2.3 *Condições suficientes de otimalidade.*

Suponha que para $\bar{x} \in \mathbb{R}^n$, $\nabla f(\bar{x}) = 0$ e $\nabla^2 f(\bar{x})$ é definida positiva. Então, \bar{x} é minimizador local estrito de (P).

Prova. Considere $S = \{h \in \mathbb{R}^n \mid \|h\| = 1\}$. Sabemos que uma matriz A é definida positiva se, e somente se, para todo $h \in S$, $h^T A h > 0$.

A função $h \in S \mapsto h^T A h$ é contínua em um conjunto compacto S , e portanto, tem minimizador $\bar{h} \in S$. Isto é, para todo $h \in S$

$$h^T A h \geq \alpha = \bar{h}^T A \bar{h} > 0,$$

onde o valor α é o menor autovalor de A . Com $A = \nabla^2 f(\bar{x})$, temos para todo $h \in S$

$$\begin{aligned} f(\bar{x} + \lambda h) &= f(\bar{x}) + \lambda \nabla f(\bar{x})^T h + \frac{\lambda^2}{2} h^T \nabla^2 f(\bar{x}) h + o_2(\lambda^2) \\ &\geq f(\bar{x}) + \frac{\lambda^2}{2} \alpha + o_2(\lambda^2). \end{aligned}$$

Dividindo por λ^2 , para $\lambda > 0$

$$\frac{f(\bar{x} + \lambda h) - f(\bar{x})}{\lambda^2} \geq \frac{\alpha}{2} + \frac{o_2(\lambda^2)}{\lambda^2}.$$

Para λ suficientemente pequeno, digamos, $0 \leq \lambda \leq \bar{\lambda}$, teremos $f(\bar{x} + \lambda h) - f(\bar{x}) > 0$, portanto, $f(\bar{x}) < f(x)$ para todo x tal que $\|x - \bar{x}\| \leq \bar{\lambda}$ e \bar{x} é minimizador local estrito. \square

2.1.2 Algoritmo

As condições necessárias de otimalidade devem ser obrigatoriamente satisfeitas por minimizadores, enquanto as condições suficientes, quando satisfeitas,

asseguram que o ponto em consideração é um minimizador local. A garantia de que o ponto é um minimizador global, em geral, é muito mais complicada.

Portanto, os métodos desenvolvidos para resolver o problema (P) procuram encontrar pontos candidatos à solução, ou seja, pontos satisfazendo a condição necessária de otimalidade de primeira ordem, isto é, pontos estacionários da função objetivo, os quais, frequentemente, são minimizadores locais.

Em casos muito simples é possível encontrar diretamente as soluções (quando existem) do problema irrestrito (P) . Mas, quando a complexidade da função objetivo ou o número de variáveis aumenta, as manipulações isoladas são insuficientes para achar sequer pontos candidatos à solução. É necessário, então, apelar para métodos numéricos, ou seja, algoritmos.

Os algoritmos que resolvem o problema (P) , quase sempre iterativos, iniciam com um ponto dado x^0 e determinam uma direção d^k , ao longo da qual é possível diminuir o valor da função f , e nesta direção calcula-se um comprimento do passo λ_k , que permita uma redução razoável no valor da função. Os algoritmos geram, portanto, uma sequência (x^k) de valores em \mathbb{R}^n com uma sequência associada $(f(x^k))$ de valores reais monótona decrescente. Vamos escrever este processo como um modelo geral de algoritmo para resolver o problema (P) . Uma discussão mais detalhada pode ser encontrada em [22].

Algoritmo 2.4 *Minimização irrestrita*

Dado $x^0 \in \mathbb{R}^n$

Para $k = 0$

Repita enquanto $\nabla f(x^k) \neq 0$

Escolha uma direção $d^k \in \mathbb{R}^n$ tal que $\nabla f(x^k)^T d^k < 0$;

Calcule o comprimento do passo $\lambda_k > 0$ tal que

$$f(x^k + \lambda_k d^k) < f(x^k);$$

$$x^{k+1} = x^k + \lambda_k d^k$$

$$k = k + 1$$

A Figura 2.1, ilustra uma iteração k do Algoritmo 2.4.

O Algoritmo 2.4 pára quando um ponto estacionário da função objetivo é encontrado, isto é, um ponto que satisfaz a condição necessária de otimalidade de primeira ordem. Como, no algoritmo, escolhemos direções ao longo das quais é

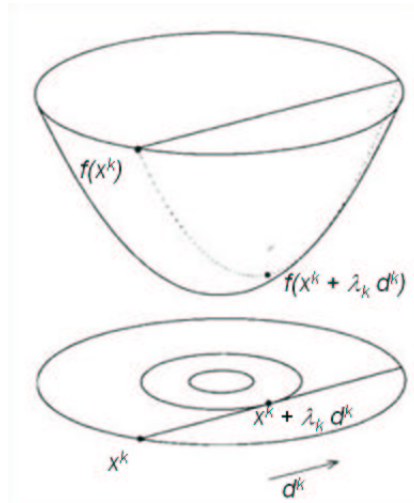


Figura 2.1: Modelo geral de algoritmo para minimização irrestrita

possível fazer decrescer o valor da função objetivo, é de se esperar que o ponto encontrado seja um minimizador local. Assim, as direções $d \in \mathbb{R}^n$ tais que $\nabla f(x)^T d < 0$, são chamadas *direções de descida* a partir de x , ou seja, as direções que formam um ângulo maior que 90 graus com o vetor gradiente da função objetivo.

Para resolver o subproblema de encontrar o comprimento do passo $\lambda_k > 0$ tal que $f(x^k + \lambda_k d^k) < f(x^k)$, existem vários métodos, denominados métodos de busca unidirecional, os quais garantem que o decréscimo no valor da função seja proporcional ao comprimento do passo na direção d^k a partir de x^k . A seguir, apresentamos dois métodos de busca unidirecional: a busca exata e a busca de Armijo.

2.1.3 Métodos de busca unidirecional

Estudaremos agora o seguinte problema: dada uma função $f : \mathbb{R}^n \rightarrow \mathbb{R}$, um ponto \bar{x} e uma direção de descida $d \in \mathbb{R}^n$, encontrar $\bar{\lambda} \geq 0$ (se existir) tal que

$$f(\bar{x} + \bar{\lambda}d) < f(\bar{x}).$$

Queremos na realidade obter $\bar{\lambda}$ que permita uma redução razoável no valor de f . Estudaremos dois métodos: a busca exata e a busca de Armijo. Este estudo baseia-se em notas de aula de Gonzaga [12].

- Busca unidirecional exata:

Dada uma função $f : \mathbb{R}^n \rightarrow \mathbb{R}$, um ponto x e uma direção $d \in \mathbb{R}^n$, $d \neq 0$, o problema de *busca unidirecional exata* ao longo de d a partir de x é

$$\begin{aligned} \text{minimizar } & f(x + \lambda d) \\ & \lambda \geq 0 \end{aligned} \quad (2.3)$$

Usaremos busca linear exata, na próxima seção, para estudar o método de Cauchy.

- Busca de Armijo:

O método de *busca de Armijo* procura uma boa redução da função f ao longo da direção d . Não tenta minimizá-la. Dados d , $x \in \mathbb{R}^n$ e um passo inicial $\rho > 0$, busca-se $\lambda \in [0, \rho]$ tal que

$$f(x + \lambda d) \quad \text{é "bem menor" que} \quad f(x).$$

Para isto usa-se um *modelo linear da função* $m : \mathbb{R} \rightarrow \mathbb{R}$, definido por

$$m(\lambda) = f(x) + \lambda f'(x, d), \quad \text{com} \quad f'(x, d) = \nabla f(x)^T d.$$

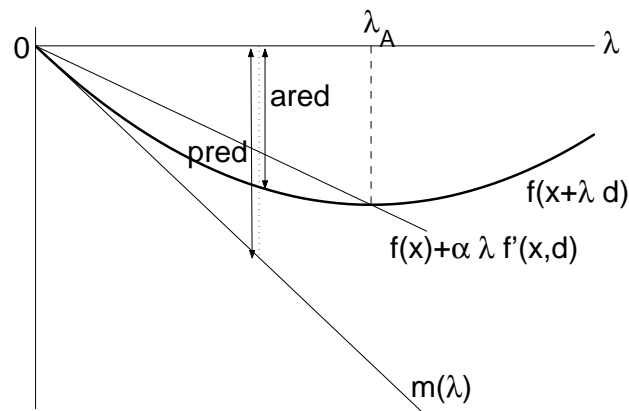


Figura 2.2: Busca de Armijo.

Dado um valor $\lambda \geq 0$, definimos:

$$\begin{aligned} ared &= f(x) - f(x + \lambda d) && \text{redução verdadeira da função,} \\ pred &= m(0) - m(\lambda) && \text{redução predita pelo modelo.} \end{aligned}$$

Consideramos aceitável um ponto em que para $\alpha \in (0, 1)$ fixo

$$\frac{ared}{pred} \geq \alpha, \quad \text{e} \quad \alpha = 0,5 \quad (\text{por exemplo}),$$

isto é, um ponto em que a redução verdadeira está “próxima” da redução do modelo. O algoritmo é simples. Inicia com $\lambda = \rho$ e reduz λ até encontrar um ponto aceitável.

Algoritmo 2.5 *Busca de Armijo*

Dados: x, d, λ e $f'(x, d)$, $\alpha \in (0, 1)$, por exemplo $\alpha = 0,5$.

$$pred = \lambda f'(x, d)$$

$$ared = f(x) - f(x + \lambda d)$$

REPITA enquanto $ared < \alpha pred$

$$\lambda = 0,7\lambda$$

$$pred = \lambda f'(x, d)$$

$$ared = f(x) - f(x + \lambda d)$$

FIM

Observação: No algoritmo, $f'(x, d)$ é um dado. Nunca se calculam derivadas.

O algoritmo pode ser escrito de maneira compacta, sem definir *ared* e *pred*. Isto foi feito para deixar claro o papel do modelo de f . No método região de confiança, utilizaremos um modelo quadrático.

Algoritmo 2.6 *Busca de Armijo, na forma compacta.*

Dados: x, d, λ e $f'(x, d)$, $\alpha \in (0, 1)$, por exemplo $\alpha = 0,5$.

REPITA enquanto $f(x + \lambda d) > f(x) + \alpha \lambda f'(x, d)$

$$\lambda = 0,7\lambda$$

FIM.

Examinemos a função $\lambda > 0 \mapsto f(x + \lambda d)$. Se $f'(x, d) = \nabla f(x)^T d < 0$ então para λ pequeno

$$f(x + \lambda d) < f(x) + \alpha \lambda f'(x, d),$$

como consequência da definição de derivada. É possível que a curva $\lambda \mapsto f(x + \lambda d)$ nunca cruze a reta $\lambda > 0 \mapsto f(x) + \alpha \lambda f'(x, d)$. Neste caso, $\bar{\lambda} = \rho$. Se a curva e a reta se cruzarem, definimos o *ponto de Armijo* λ_A como o primeiro ponto de cruzamento, isto é,

$$f(x + \lambda_A d) = f(x) + \alpha \lambda_A f'(x, d),$$

ou

$$\frac{f(x + \lambda_A d) - f(x)}{\lambda_A} = \alpha f'(x, d).$$

Pelo teorema do valor médio, existe $\tilde{\lambda} \in [0, \lambda_A]$ tal que

$$f'(x + \tilde{\lambda} d, d) = \alpha f'(x, d).$$

Concluimos que em uma busca de Armijo, uma e somente uma das situações abaixo ocorre:

1. $\bar{\lambda} = \rho$ ou
2. $\bar{\lambda} < \rho$, $\bar{\lambda} \leq \lambda_A \leq \frac{\bar{\lambda}}{0,7}$ e existe $\tilde{\lambda} \in [0; \bar{\lambda}/0,7]$ tal que $f'(x + \tilde{\lambda} d, d) = \alpha f'(x, d)$.

O método de Armijo não encontra um ponto próximo a um minimizador unidirecional, mas é muito eficiente. Para algoritmos bem projetados, faz um número muito pequeno de cálculos de função sendo portanto muito rápido.

Para estudar os métodos de minimização irrestrita, que apresentamos nas seções seguintes, vamos utilizar somente estes dois métodos de busca. Existem vários outros métodos de busca que podem ser encontrados, por exemplo, em Bertsekas [2].

O sucesso dos métodos que resolvem o problema (P) , depende da escolha adequada da direção d^k e do método para calcular o comprimento do passo λ_k . A diferença essencial entre os métodos está na escolha da direção d^k , que exerce grande influência no processo iterativo $x^{k+1} = x^k + \lambda_k d^k$. Uma exceção é o método de região de confiança, onde a direção e o comprimento do passo são obtidos simultaneamente.

A seguir discutiremos algumas opções para a escolha da direção d^k do Algoritmo 2.4.

2.2 Método de Cauchy

Um dos métodos, mais conhecido e também um dos mais antigos, que resolve o problema irrestrito é o método clássico do gradiente, também chamado método de Cauchy, já que foi *Augustin Louis Cauchy*, em 1847, quem mostrou que a direção do gradiente é a direção de máximo acréscimo da função. Assim, no algoritmo de Cauchy, a direção é definida em cada iteração como o oposto do gradiente da função objetivo no ponto corrente:

$$d^k = -\nabla f(x^k),$$

e a sequência gerada pelo algoritmo de Cauchy tem a forma

$$x^{k+1} = x^k - \lambda_k \nabla f(x^k),$$

onde λ_k é o comprimento do passo ao longo da direção d^k no ponto x^k . Por exemplo, se usarmos busca linear exata, o algoritmo de Cauchy tem a seguinte forma:

Algoritmo 2.7 *Algoritmo de Cauchy com busca exata*

Dado: $x^0 \in \mathbb{R}^n$

$k=0$

REPITA enquanto $\nabla f(x^k) \neq 0$

$$d^k = -\nabla f(x^k)$$

calcule $\bar{\lambda}$, solução do problema unidirecional (2.3), se existir;

se não existe solução, pare com insucesso.

$$x^{k+1} = x^k + \bar{\lambda}d^k;$$

$$k = k + 1$$

FIM.

2.2.1 Convergência global

Os métodos que utilizam algum tipo de estratégia para a obtenção de convergência independente do ponto inicial são chamados de métodos globais ou métodos globalmente convergentes.

Quando usamos, para determinar o comprimento do passo, busca exata ou busca inexata de Armijo, o algoritmo de Cauchy é globalmente convergente, isto é, todo ponto de aderência da sequência gerada pelo algoritmo é estacionário, ou seja,

todo ponto de aderência da sequência satisfaz a condição necessária de otimalidade de primeira ordem.

A demonstração que apresentamos aqui, sobre a convergência global do algoritmo de Cauchy com busca Armijo, é feita por Gonzaga [12] em notas de aula não publicadas.

Teorema 2.8 *Algoritmo de Cauchy com busca Armijo*

Seja f de classe \mathcal{C}^1 . Todo ponto de aderência da sequência gerada pelo algoritmo de Cauchy com busca Armijo é estacionário.

Prova.

Suponha por absurdo que uma sequência (x^k) , gerada pelo algoritmo de Cauchy com busca de Armijo, possui um valor de aderência \bar{x} não estacionário. Em \bar{x} , definimos $\bar{d} = -\nabla f(\bar{x}) \neq 0$.

Pela condição de Armijo e por continuidade de f , temos que $(f(x^k))$ é decrescente e tem valor de aderência $f(\bar{x})$. Portanto, é convergente e então,

$$\lim_{k \rightarrow \infty} (f(x^k) - f(x^{k+1})) = 0. \quad (2.4)$$

Como \bar{x} é valor de aderência de $(x^k)_{k \in \mathbb{N}}$, existe um conjunto $\mathcal{K} \subset \mathbb{N}$, tal que a sequência $(x^k)_{k \in \mathcal{K}}$ converge para \bar{x} , isto é, $x^k \xrightarrow{\mathcal{K}} \bar{x}$. Primeiramente, vamos mostrar que se $x^{k+1} = x^k + \lambda_k d^k$, para todo $k \in \mathbb{N}$, então $\lambda_k \xrightarrow{\mathcal{K}} 0$. De fato, temos

$$f(x^{k+1}) = f(x^k + \lambda_k d^k) \leq f(x^k) + \alpha \lambda_k \nabla f(x^k)^T d^k,$$

e portanto, $f(x^k) - f(x^{k+1}) \geq -\alpha \lambda_k \nabla f(x^k)^T d^k = \alpha \lambda_k \|\nabla f(x^k)\|^2$, donde segue que

$$\frac{f(x^k) - f(x^{k+1})}{\alpha \|\nabla f(x^k)\|^2} \geq \lambda_k \geq 0, \quad (2.5)$$

pois $\alpha \|\nabla f(x^k)\|^2 > 0$. Tomando limites em (2.5), para $k \in \mathcal{K}$, usando (2.4) e $\alpha \|\nabla f(x^k)\|^2 \xrightarrow{\mathcal{K}} \alpha \|\nabla f(\bar{x})\|^2 \neq 0$, concluímos que $\lambda_k \xrightarrow{\mathcal{K}} 0$. Assim, para k suficientemente grande, $\lambda_k < \rho$. Pelas propriedades da busca unidirecional de Armijo, Seção 2.1.3, para k tal que $\lambda_k < \rho$, temos

$$0 \leq \tilde{\lambda}_k \leq \frac{\lambda_k}{0.7}, \quad \text{tal que,} \quad \nabla f(x^k + \tilde{\lambda}_k d^k)^T d^k = \alpha \nabla f(x^k)^T d^k.$$

Tomando limites para $k \in \mathcal{K}$ nessas duas expressões, obtemos

$$\tilde{\lambda}_k \xrightarrow{\mathcal{K}} 0, \quad d^k = -\nabla f(x^k) \xrightarrow{\mathcal{K}} \bar{d} = -\nabla f(\bar{x}),$$

$$\nabla f(\bar{x})^T \bar{d} = \alpha \nabla f(\bar{x})^T \bar{d},$$

o que contradiz o fato de que $\nabla f(\bar{x})^T \bar{d} = -\|\nabla f(\bar{x})\|^2 \neq 0$, completando a demonstração. \square

Ainda temos que, se a função é convexa, a sequência gerada pelo algoritmo de Cauchy, com busca de Armijo, tem um único valor de aderência, ver [4, 11, 15]. O mesmo não ocorre se usarmos busca linear exata. Em [11], temos um exemplo de uma função convexa continuamente diferenciável em que a sequência, gerada pelo algoritmo de Cauchy com busca exata, tem quatro valores de aderência. A Figura 2.3 ilustra as curvas de nível da função “twisted”, que tem quatro valores de aderência.

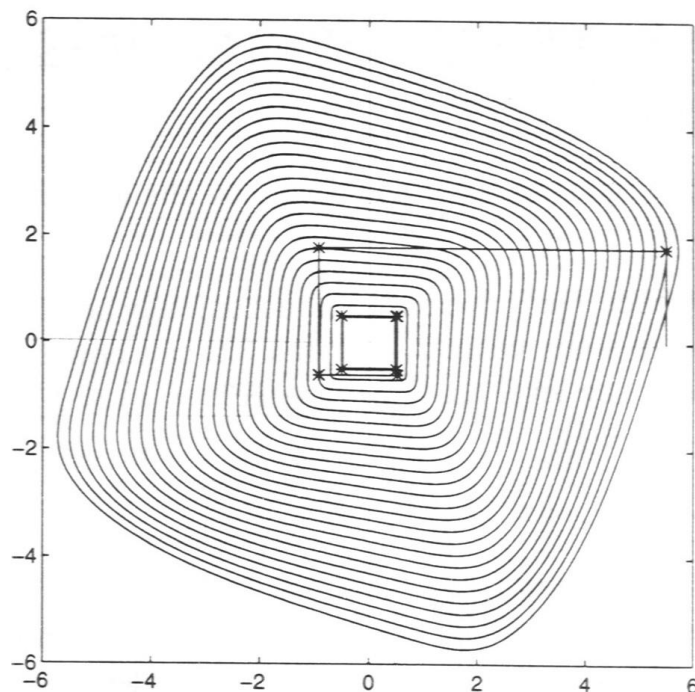


Figura 2.3: Função twisted.

No entanto, como o algoritmo de Cauchy com busca exata é globalmente convergente, esses quatro valores de aderência são estacionários.

2.2.2 Convergência local

Os resultados mais importantes sobre a convergência local do algoritmo de Cauchy são revelados quando a função objetivo é quadrática.

Primeiramente, vamos enunciar um resultado clássico para função não quadrática, o Teorema 2.9, que garante, sob certas hipóteses, que a sequência de valores da função objetivo ($f(x^k)$) converge, para o valor mínimo da função objetivo, com velocidade q -linear e com taxa $\left(\frac{A-a}{A+a}\right)^2$, onde A e a são o maior e o menor autovalor da matriz Hessiana da função objetivo no ponto x^* , minimizador local da f . Quando a função é quadrática e convexa, este resultado implica na convergência q -linear da sequência (x^k), na norma induzida pela Hessiana da quadrática, com taxa $\left(\frac{A-a}{A+a}\right)$. Ver Lema 2.10.

O teorema a seguir fala da ordem de convergência da sequência gerada pelo algoritmo de Cauchy com busca linear exata, para função objetivo não quadrática.

Teorema 2.9 *Método de Cauchy: função não quadrática*

Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f \in \mathcal{C}^2$. Seja $x^* \in \mathbb{R}^n$ um minimizador local da f , tal que a matriz $\nabla^2 f(x^*)$ é definida positiva e a sequência (x^k) gerada pelo algoritmo de Cauchy com busca unidirecional exata converge a x^* , então a sequência ($f(x^k)$) converge q -linearmente a $f(x^*)$ com taxa não superior a $\left(\frac{A-a}{A+a}\right)^2$, onde A e a são o maior e menor autovalor de $\nabla^2 f(x^*)$, respectivamente.

Prova. Veja, por exemplo, [19]. □

Lembramos que a expressão $f \in \mathcal{C}^2$ indica que f tem derivadas contínuas até a ordem 2.

Quando a função objetivo é quadrática

$$f(x) = \frac{1}{2}x^T Qx + b^T x + c,$$

com matriz Hessiana Q definida positiva, a função é convexa e tem um único minimizador x^* , que é global, satisfazendo

$$Qx^* + b = 0. \tag{2.6}$$

Neste caso, em que a função é quadrática, há uma forma explícita para o comprimento do passo com busca linear exata ao longo da direção de Cauchy, que é dado por:

$$\lambda_k = \frac{d^{kT} d^k}{d^{kT} Q d^k}.$$

De fato

$$\begin{aligned} f(x^k + \lambda d^k) &= \frac{1}{2}(x^k + \lambda d^k)^T Q (x^k + \lambda d^k) + b^T (x^k + \lambda d^k) + c \\ &= \frac{1}{2} x^{kT} Q x^k + \lambda x^{kT} Q d^k + \frac{\lambda^2}{2} d^{kT} Q d^k + b^T x^k + \lambda b^T d^k + c \\ &= f(x^k) + \lambda x^{kT} Q d^k + \frac{\lambda^2}{2} d^{kT} Q d^k + \lambda b^T d^k. \end{aligned}$$

Derivando e igualando a zero

$$\frac{d}{d\lambda} f(x^k + \lambda d^k) = x^{kT} Q d^k + b^T d^k + \lambda d^{kT} Q d^k = 0$$

obtemos

$$\lambda_k = \frac{-x^{kT} Q d^k - b^T d^k}{d^{kT} Q d^k} = \frac{-(Q x^k + b)^T d^k}{d^{kT} Q d^k}.$$

Mas $\nabla f(x^k) = Q x^k + b$ e a direção de Cauchy é $d^k = -\nabla f(x^k) = -(Q x^k + b)$. Portanto,

$$\lambda_k = \frac{d^{kT} d^k}{d^{kT} Q d^k}, \quad (2.7)$$

é a forma explícita para o comprimento do passo, que é calculado pelo algoritmo em cada iteração.

O lema, a seguir, mostra que a convergência *q-linear* da sequência $(f(x^k))$ implica na convergência *q-linear* da sequência (x^k) , na norma induzida pela Hessiana da quadrática. A norma induzida por uma matriz Q é definida pela equação:

$$\|x\|_Q^2 = x^T Q x \quad \text{onde } x \in \mathbb{R}^n \text{ e } Q \in \mathbb{R}^{n \times n}.$$

Lema 2.10 *Seja (x^k) uma sequência qualquer em \mathbb{R}^n . Suponha que a sequência $(f(x^k))$ satisfaz $f(x^{k+1}) - f(x^*) \leq \beta(f(x^k) - f(x^*))$, onde $0 \leq \beta < 1$ e x^* é dado*

por (2.6). Então

$$\|x^{k+1} - x^*\|_Q \leq \sqrt{\beta} \|x^k - x^*\|_Q.$$

Prova. Para todo $x \in \mathbb{R}^n$, temos:

$$\begin{aligned} \frac{1}{2} \|x - x^*\|_Q^2 &= \frac{1}{2} (x - x^*)^T Q (x - x^*) \\ &= \frac{1}{2} x^T Q x - x^T Q x^* + \frac{1}{2} x^{*T} Q x^* \\ &= \frac{1}{2} x^T Q x + b^T x + \frac{1}{2} x^{*T} Q x^* \\ &= f(x) - f(x^*). \end{aligned}$$

Por isso,

$$\begin{aligned} \|x^{k+1} - x^*\|_Q &= \sqrt{2(f(x^{k+1}) - f(x^*))} \\ &\leq \sqrt{2\beta(f(x^k) - f(x^*))} \\ &= \sqrt{\beta} \|x^k - x^*\|_Q, \end{aligned}$$

completando a prova. □

Tendo em vista este resultado, gostaríamos de saber se a sequência (x^k) tem convergência *q-linear* na norma euclidiana, que é a norma mais usual.

Embora as normas em \mathbb{R}^n sejam equivalentes, a velocidade de convergência *q-linear* depende da norma. Existem sequências que convergem com velocidade *q-linear* em uma norma, mas não em outra. Os Exemplos 2 e 3, da Seção 1.2.1, ilustraram este fato.

Assim, o fato da sequência (x^k) convergir com velocidade *q-linear* na norma induzida pela hessiana da quadrática, não garante sua convergência *q-linear* na norma euclidiana. Em Bertsekas [2], encontramos um resultado na norma euclidiana, que estabelece a melhor taxa de convergência para uma função quadrática e convexa dada por:

$$\frac{\|x^{k+1} - x^*\|_2}{\|x^k - x^*\|_2} \leq \frac{A - a}{A + a},$$

com passo constante $\alpha = \frac{2}{A + a}$, onde A e a são o maior e o menor autovalor da matriz hessiana da função quadrática. Somente nesta referência encontramos uma taxa de convergência na norma euclidiana. Esta foi uma motivação para estudar a velocidade de convergência, da sequência gerada pelo algoritmo de Cauchy, na norma euclidiana. Outra motivação vem da Figura 2.4, que mostra o decréscimo das distâncias euclidianas entre as iterações da sequência gerada pelo algoritmo de Cauchy.

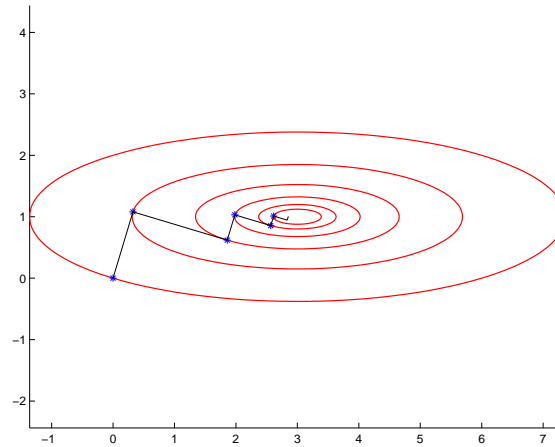


Figura 2.4: O trajeto do algoritmo de Cauchy.

De fato, mostramos que esse decréscimo é *linear*, quando o método de Cauchy com busca linear exata é aplicado a uma função quadrática, convexa. Além disso, exibimos a taxa de convergência que, como no Teorema 2.9, está relacionada com os autovalores da hessiana da quadrática.

Usaremos os resultados dos Lemas 2.11 e 2.12 para mostrar um resultado original: a sequência gerada pelo método de Cauchy com busca linear exata converge, com velocidade *q-linear* na norma euclidiana, para o único minimizador da função quadrática, e que a taxa de convergência da sequência (x^k) é igual a

$$\sqrt{1 - \frac{a}{A}}.$$

Lema 2.11 *Seja $Q \in \mathbb{R}^{n \times n}$, simétrica e definida positiva. Então, para todo vetor $d \in \mathbb{R}^n$, $d \neq 0$, tem-se:*

$$\frac{\|d\|^4}{(d^T Q d)(d^T Q^{-1} d)} \leq 1.$$

Prova. Como Q é definida positiva pode ser escrita na forma

$$Q = P D P^T = (P \sqrt{D} P^T)(P \sqrt{D} P^T) = G G,$$

onde as colunas de P são os autovetores ortonormais de Q e D é a matriz diagonal dos autovalores de Q . A matriz $G = (P \sqrt{D} P^T)$ é simétrica. De fato,

$$G^T = (P \sqrt{D} P^T)^T = P (P \sqrt{D})^T = P \sqrt{D} P^T = G.$$

Assim, podemos escrever:

$$\begin{aligned} Q &= G^T G && \text{logo} \\ d^T Q d &= d^T G^T G d = v^T v && \text{com } v = Gd. \end{aligned}$$

Para obter a expressão análoga com Q^{-1} , escrevemos $Q = GG^T$, assim:

$$\begin{aligned} Q^{-1} &= (GG^T)^{-1} = (G^T)^{-1}G^{-1} = (G^{-1})^T G^{-1}, && \text{e} \\ d^T Q^{-1} d &= d^T (G^{-1})^T G^{-1} d = u^T u && \text{com } u = G^{-1}d. \end{aligned}$$

Com isto, obtemos:

$$\frac{\|d\|^4}{(d^T Q d)(d^T Q^{-1} d)} = \frac{(d^T d)^2}{(v^T v)(u^T u)}.$$

Pela desigualdade de *Cauchy-Schwarz*, temos que

$$|u^T v|^2 \leq (u^T u)(v^T v) \Leftrightarrow \frac{1}{(u^T u)(v^T v)} \leq \frac{1}{|u^T v|^2},$$

onde $|u^T v|^2 = |d^T (G^{-1})^T G d|^2 = |d^T (G^{-1})^T G^T d|^2 = |d^T (GG^{-1})^T d|^2 = |d^T d|^2$. Portanto,

$$\frac{\|d\|^4}{(d^T Q d)(d^T Q^{-1} d)} \leq \frac{(d^T d)^2}{|u^T v|^2} = \frac{(d^T d)^2}{|d^T d|^2} = 1,$$

completando a prova. □

Lema 2.12 Dado $x \in \mathbb{R}^n$, $x \neq 0$, seja $d = -Qx$ com $Q \in \mathbb{R}^{n \times n}$, simétrica e definida positiva e $\lambda = \frac{d^T d}{d^T Q d}$. Então,

$$\lambda \leq \frac{x^T Q x}{x^T Q^2 x}.$$

Prova. Temos que $x^T Q x = x^T (-d) = -d^T x = -d^T (-Q^{-1}d) = d^T Q^{-1}d$, e $x^T Q^2 x = x^T (-Qd) = (-Qx)^T d = d^T d$. Assim,

$$\lambda \frac{x^T Q^2 x}{x^T Q x} = \frac{(d^T d)(d^T d)}{d^T Q d d^T Q^{-1} d} = \frac{\|d\|^4}{(d^T Q d)(d^T Q^{-1} d)}.$$

Pelo Lema 2.11, temos que:

$$\lambda \frac{x^T Q^2 x}{x^T Q x} = \frac{\|d\|^4}{(d^T Q d)(d^T Q^{-1} d)} \leq 1.$$

Portanto,

$$\lambda \leq \frac{x^T Q x}{x^T Q^2 x}.$$

□

Vamos assumir agora, sem perda de generalidade¹, que f é da forma

$$f(x) = \frac{1}{2} x^T Q x. \quad (2.8)$$

Teorema 2.13 *Convergência local do algoritmo de Cauchy com busca exata na norma euclidiana*

Considere a função quadrática, convexa dada em (2.8) e a sequência (x^k) gerada pelo Algoritmo 2.7. Então,

$$\|x^{k+1}\|_2 \leq \beta \|x^k\|_2,$$

para todo $k \in \mathbb{N}$, com

$$\beta = \sqrt{1 - \frac{a}{A}} \quad (2.9)$$

onde a e A são, respectivamente, o menor e maior autovalor da matriz Q .

Prova. Temos que

$$\begin{aligned} \|x^{k+1}\|_2^2 &= (x^{k+1})^T x^{k+1} \\ &= (x^k + \lambda_k d^k)^T (x^k + \lambda_k d^k) \\ &= (x^k)^T x^k + 2\lambda_k (x^k)^T d^k + \lambda_k^2 (d^k)^T d^k, \end{aligned}$$

onde a direção de Cauchy para a função (2.8) é $d^k = -\nabla f(x^k) = -Qx^k$, assim $(d^k)^T d^k = (x^k)^T Q^2 x^k$ e

$$\begin{aligned} \|x^{k+1}\|_2^2 &= \|x^k\|_2^2 - 2\lambda_k (x^k)^T Q x^k + \lambda_k^2 (x^k)^T Q^2 x^k \\ &\leq \|x^k\|_2^2 - 2\lambda_k (x^k)^T Q x^k + \lambda_k (x^k)^T Q x^k \\ &= \|x^k\|_2^2 - \lambda_k (x^k)^T Q x^k, \end{aligned}$$

onde a desigualdade vem do Lema 2.12. Se $x^k = 0$ não há nada a fazer. Assim,

¹Veja Apêndice 1.

assumimos que $x^k \neq 0$. Por isso,

$$\frac{\|x^{k+1}\|_2^2}{\|x^k\|_2^2} \leq 1 - \lambda_k \cdot \frac{(x^k)^T Q x^k}{(x^k)^T x^k} = 1 - \left(\frac{(d^k)^T d^k}{(d^k)^T Q d^k} \cdot \frac{(x^k)^T Q x^k}{(x^k)^T x^k} \right). \quad (2.10)$$

Como Q é definida positiva, então para todo $v \in \mathbb{R}^n$, tem-se

$$av^T v \leq v^T Q v \leq Av^T v,$$

onde a e A representam, respectivamente, o menor e maior autovalor da matriz Q .

$$\text{Para } v = d^k \text{ temos:} \quad (d^k)^T Q d^k \leq A(d^k)^T d^k \quad \text{e}$$

$$\text{para } v = x^k \text{ temos:} \quad (x^k)^T Q x^k \geq a(x^k)^T x^k, \quad \text{assim,}$$

$$\begin{aligned} \frac{(d^k)^T d^k}{(d^k)^T Q d^k} \cdot \frac{(x^k)^T Q x^k}{(x^k)^T x^k} &\geq \frac{1}{A} a \\ - \left(\frac{(d^k)^T d^k}{(d^k)^T Q d^k} \cdot \frac{(x^k)^T Q x^k}{(x^k)^T x^k} \right) &\leq - \left(\frac{a}{A} \right) \end{aligned}$$

Substituindo em (2.10), obtemos:

$$\frac{\|x^{k+1}\|_2^2}{\|x^k\|_2^2} \leq 1 - \frac{a}{A},$$

completando a prova. □

Este teorema é um resultado original desta dissertação. Aqui, novamente a taxa de convergência da sequência está relacionada com os autovalores da matriz Q hessiana da quadrática. De acordo com (2.9), quanto menor for a diferença entre o maior e o menor autovalor da hessiana, mais rápida é a convergência. Se $A = a$, então as curvas de nível da função objetivo são esferas e a convergência ocorre em um passo. Nas situações em que as curvas de nível em torno de x^* são alongadas, isto é, quando a diferença entre o maior e menor autovalor é grande, a velocidade de convergência tende a diminuir.

2.3 Método de Newton

Nesta seção discutiremos outra escolha usual da direção d^k no Algoritmo 2.4: a direção de Newton.

Vimos que uma condição necessária para que um ponto $x \in \mathbb{R}^n$ seja solução do problema irrestrito (P) é

$$\nabla f(x) = 0. \quad (2.11)$$

Ou seja, para encontrar os possíveis pontos candidatos à solução do problema (P) , devemos resolver o sistema (2.11), geralmente não linear.

Sabemos que um método eficiente para resolver sistemas de equações é o método de Newton, veja, por exemplo, [5]. Assim, se a função f é de classe \mathcal{C}^2 , podemos aplicar o método de Newton para a função $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$, onde $g = \nabla f$ e $g \in \mathcal{C}^1$. A sequência gerada pelo método de Newton para resolver o sistema de equações $g(x) = 0$ é da forma

$$x^{k+1} = x^k + d^k,$$

onde d^k é por definição,

$$d^k = -(H(x^k))^{-1}g(x^k).$$

A matriz $H(x^k)$ é a Jacobiana da função g , isto é, $H(x^k) = \nabla^2 f(x^k)$. Com isto, a sequência gerada pelo método de Newton, pode ser escrita como

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

Esta iteração está bem definida se a matriz Jacobiana da função g , ou equivalentemente, a matriz $H(x^k)$ hessiana da função f , for não singular. Além disso, se o ponto inicial x^0 estiver suficientemente próximo da solução, então a sequência assim gerada, herda as mesmas propriedades do método de Newton para sistemas de equações, ou seja, possui convergência local quadrática.

Um modo equivalente de definir o método de Newton para resolver o problema (P) , é como um processo iterativo que consiste em minimizar, em cada iteração, a aproximação quadrática da função objetivo.

Para isto, assumimos que f é de classe \mathcal{C}^2 e escrevemos sua aproximação

quadrática numa vizinhança do ponto corrente $x^k \in \mathbb{R}^n$:

$$f(x^k + d) \approx m(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H(x^k) d \quad (2.12)$$

onde $H(x^k)$ denota a matriz hessiana da função f no ponto x^k . Supondo que m tem um minimizador, d^k , então ele deve ser um ponto estacionário, isto é,

$$\nabla m(d^k) = 0.$$

Derivando (2.12) e igualando a zero, obtemos:

$$\begin{aligned} \nabla m(d^k) = \nabla f(x^k) + H(x^k)d^k &= 0 \\ H(x^k)d^k &= -\nabla f(x^k). \end{aligned}$$

Se $H(x^k) = \nabla^2 f(x^k)$ é invertível, então o passo de Newton é dado por:

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

Assim, como $x^{k+1} = x^k + d^k$, temos que

$$x^{k+1} = x^k - (H(x^k))^{-1} \nabla f(x^k),$$

é a sequência gerada pelo método de Newton “puro”, em que o comprimento do passo λ_k é sempre igual a 1.

O algoritmo de Newton tem a seguinte forma compacta:

Algoritmo 2.14 *Algoritmo de Newton*

Dado $x^0 \in \mathbb{R}^n$

$k = 0$

REPITA enquanto $\nabla f(x^k) \neq 0$

Resolva $H(x^k)d = -\nabla f(x^k)$ (se possível)

$x^{k+1} = x^k + d$

FIM

Se convergir, este processo iterativo converge para um ponto estacionário da f , que pode não ser a solução do problema (P), sendo um maximizador ou ponto de sela.

Além disso, o processo pode divergir se a matriz hessiana for mal condicionada ou singular. Se a matriz $H(x)$ é definida positiva em $x \in \mathbb{R}^n$, então a direção de Newton é uma direção de descida. De fato, $H^{-1}(x)$ é definida positiva e

$$\nabla f(x)^T d = -\nabla f(x)^T H^{-1}(x) \nabla f(x) < 0.$$

É interessante notar que se a função é quadrática com hessiana H não singular, então o método de Newton resolve o problema (P) em uma iteração. De fato, se H é não singular, então a função $f(x) = \frac{1}{2}x^T Hx + b^T x + c$ tem um único minimizador x^* , que é global, e a partir de qualquer ponto inicial x^0 , temos que:

$$\begin{aligned} d &= -(\nabla^2 f(x^0))^{-1} \nabla f(x^0) \\ d &= -H^{-1}(Hx^0 + b) \end{aligned}$$

e a próxima iteração é da forma:

$$\begin{aligned} x^1 &= x^0 + d = x^0 - H^{-1}(Hx^0 + b) \\ x^1 &= x^0 - x^0 - H^{-1}b = -H^{-1}b. \end{aligned}$$

Mas, $\nabla f(x^*) = Hx^* + b = 0 \Leftrightarrow x^* = -H^{-1}b$. Assim, $x^1 = x^*$, ou seja, o método de Newton encontra a solução x^* , em um único passo, a partir de qualquer ponto inicial x^0 . A Figura 2.5 ilustra este fato.

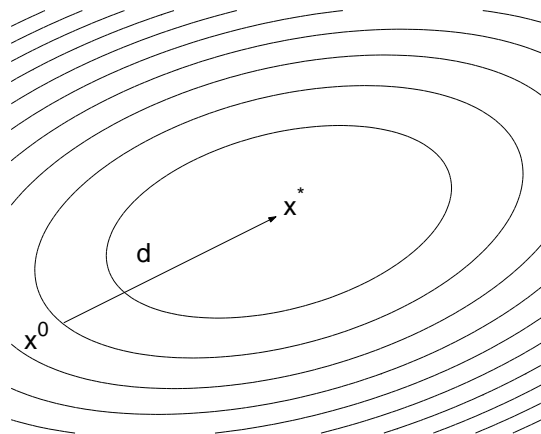


Figura 2.5: Método de Newton para função quadrática.

Assim, para qualquer função f , a direção de Newton é aquela que, em cada iteração k , minimiza o modelo quadrático da f em x^k .

2.3.1 Convergência global

Vimos que, se a matriz $H(x)$, hessiana da função objetivo, é definida positiva em $x \in \mathbb{R}^n$ (e portanto $H^{-1}(\cdot)$ é contínua e definida positiva em \mathbb{R}^n), então a direção de Newton é uma direção de descida. Sendo assim, introduzimos uma busca unidirecional ao longo da direção de Newton $d = -H^{-1}(x^k)\nabla f(x^k)$, em cada iteração, para que o algoritmo seja globalmente convergente.

O teorema a seguir garante que o algoritmo de Newton com busca de Armijo é globalmente convergente, isto é, todo valor de aderência da sequência gerada pelo algoritmo, independente do ponto inicial, é um ponto estacionário da função objetivo.

Teorema 2.15 *Suponha que f é de classe \mathcal{C}^2 e $\nabla^2 f(x)$ é definida positiva para todo $x \in \mathbb{R}^n$. Então o algoritmo de Newton com busca Armijo é globalmente convergente.*

Prova. Suponha por absurdo que uma sequência (x^k) , gerada pelo algoritmo de Newton, possui um valor de aderência \bar{x} não estacionário, isto é, $\nabla f(\bar{x}) \neq 0$. Como \bar{x} é valor de aderência de $(x^k)_{k \in \mathbb{N}}$, existe um conjunto $\mathcal{K} \subset \mathbb{N}$, tal que a sequência $(x^k)_{k \in \mathcal{K}}$ converge para \bar{x} , isto é, $x^k \xrightarrow{\mathcal{K}} \bar{x}$. Em \bar{x} , definimos $\bar{d} = -H^{-1}(\bar{x})\nabla f(\bar{x}) \neq 0$.

Como $x \mapsto H(x)$ é uma função contínua e $H(x)$ é definida positiva para todo $x \in \mathbb{R}^n$ (e portanto o mesmo vale para $H^{-1}(x)$), então a direção de Newton é de descida. Temos que $(f(x^k))_{\mathbb{N}}$ é decrescente e como f é contínua, $f(x^k) \xrightarrow{\mathcal{K}} f(\bar{x})$ e portanto $f(x^k) \xrightarrow{\mathbb{N}} f(\bar{x})$. Assim, para todo $k \in \mathbb{N}$

$$\lim_{k \rightarrow \infty} (f(x^k) - f(x^{k+1})) = 0.$$

Vamos mostrar que, com $x^{k+1} = x^k + \lambda_k d^k$, para todo $k \in \mathbb{N}$, temos $\lambda_k \xrightarrow{\mathcal{K}} 0$. De fato, temos pela condição de Armijo, que para alguma inclinação $\alpha \in (0, 1)$, existe $\lambda_k > 0$ tal que:

$$f(x^{k+1}) = f(x^k + \lambda_k d^k) \leq f(x^k) + \alpha \lambda_k \nabla f(x^k)^T d^k$$

e portanto

$$f(x^k) - f(x^{k+1}) \geq -\alpha \lambda_k \nabla f(x^k)^T d^k.$$

Como $\alpha \nabla f(x^k)^T d^k < 0$, temos:

$$\frac{f(x^k) - f(x^{k+1})}{-\alpha \nabla f(x^k)^T d^k} \geq \lambda_k \geq 0.$$

Tomando limites para $k \in \mathcal{K}$,

$$\lim_{k \rightarrow \infty, k \in \mathcal{K}} \frac{f(x^k) - f(x^{k+1})}{-\alpha \nabla f(x^k)^T d^k} \geq \lim_{k \rightarrow \infty, k \in \mathcal{K}} \lambda_k \geq 0.$$

Como $\lim_{k \rightarrow \infty} (f(x^k) - f(x^{k+1})) = 0$, $\alpha > 0$ e $\nabla f(x^k)^T d^k \xrightarrow{\mathcal{K}} \nabla f(\bar{x})^T \bar{d} < 0$, então

$$0 \geq \lim_{k \rightarrow \infty, k \in \mathcal{K}} \lambda_k \geq 0.$$

Pelo teorema do confronto, conclui-se que λ_k é convergente e $\lambda_k \xrightarrow{\mathcal{K}} 0$.

Concluimos que para k , suficientemente grande, $\lambda_k < \rho$.

Pelas propriedades da busca de Armijo (Seção 2.1.3), para k tal que $\lambda_k < \rho$ temos

$$0 \leq \tilde{\lambda}_k \leq \lambda_A \leq \frac{\lambda_k}{0,7}, \quad \text{tal que,} \quad \nabla f(x^k + \tilde{\lambda}_k d^k)^T d^k = \alpha \nabla f(x^k)^T d^k.$$

Tomando limites para $k \in \mathcal{K}$ nestas duas expressões, obtemos

$$\tilde{\lambda}_k \xrightarrow{\mathcal{K}} 0,$$

e

$$\lim_{k \rightarrow \infty, k \in \mathcal{K}} \nabla f(x^k + \tilde{\lambda}_k d^k)^T d^k = \lim_{k \rightarrow \infty, k \in \mathcal{K}} \alpha \nabla f(x^k)^T d^k.$$

Consequentemente

$$\nabla f(\bar{x})^T \bar{d} = \alpha \nabla f(\bar{x})^T \bar{d},$$

o que contradiz o fato de que $\nabla f(\bar{x})^T \bar{d} = -\nabla f(\bar{x})^T H^{-1}(\bar{x}) \nabla f(\bar{x}) \neq 0$, completando a demonstração. \square

2.3.2 Convergência local

Um dos indicadores da eficiência do método de Newton é a velocidade de convergência que, sob certas hipóteses, é quadrática.

Para demonstrar o importante Teorema 2.17, o qual garante que a sequência gerada pelo algoritmo de Newton “puro” converge quadraticamente, utilizaremos o resultado do seguinte lema:

Lema 2.16 *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f \in \mathcal{C}^2$ com $x \mapsto \nabla^2 f(x)$ Lipschitziana, então existe $\beta > 0$ tal que*

$$\|\nabla f(x^*) - \nabla f(x) - \nabla^2 f(x)(x^* - x)\| \leq \beta \|x - x^*\|^2.$$

Prova. Simplificamos a notação com $(x^* - x) = v$, $\nabla f(x) = g(x)$ e $\nabla^2 f(x) = H(x)$. Como $g(x)$ é de classe \mathcal{C}^1 , então a expansão em série de Taylor com resto integral para $g(x)$ no ponto x em torno de v é dada por:

$$g(x + v) = g(x) + r_0(v) \tag{2.13}$$

com

$$r_0(v) = \frac{1}{0!} \int_0^1 (1 - \theta)^0 d^{0+1} g(x + \theta v) (v)^{0+1} d\theta$$

$$r_0(v) = \int_0^1 dg(x + \theta v) v d\theta = \int_0^1 H(x + \theta v) v d\theta.$$

Substituindo $r_0(v)$ em (2.13), temos:

$$g(x^*) - g(x) = \int_0^1 H(x + \theta v) v d\theta.$$

Subtraindo $H(x)v = \int_0^1 H(x)v d\theta$ e tomando a norma, obtemos:

$$\|g(x^*) - g(x) - H(x)v\| = \left\| \int_0^1 [H(x + \theta v) - H(x)]v d\theta \right\|.$$

Pelas propriedades da integral:

$$\|g(x^*) - g(x) - H(x)v\| \leq \int_0^1 \| [H(x + \theta v) - H(x)]v \| d\theta,$$

e pela desigualdade triangular:

$$\|g(x^*) - g(x) - H(x)v\| \leq \int_0^1 \|H(x + \theta v) - H(x)\| \|v\| d\theta.$$

Como $H(x) = \nabla^2 f(x)$ é Lipschitziana, temos:

$$\int_0^1 \|H(x + \theta v) - H(x)\| \|v\| d\theta \leq \beta \int_0^1 \|\theta v\| \|v\| d\theta.$$

Assim

$$\int_0^1 \|H(x + \theta v) - H(x)\| \|v\| d\theta \leq \beta \int_0^1 \|v\|^2 d\theta = \beta \|v\|^2$$

e portanto:

$$\|g(x^*) - g(x) - H(x)v\| \leq \beta \|v\|^2.$$

Retornando a notação, segue o resultado desejado:

$$\|\nabla f(x^*) - \nabla f(x) - \nabla^2 f(x)(x^* - x)\| \leq \beta \|x^* - x\|^2.$$

□

Com isto, podemos demonstrar o teorema, que estabelece a convergência quadrática do método de Newton para um minimizador local x^* , se a matriz hessiana da função no ponto x^* for não singular e o ponto inicial x^0 estiver suficientemente próximo de x^* .

Teorema 2.17 *Método de Newton*

Seja $f \in \mathcal{C}^2$, com $\nabla^2 f(x^*) > 0$ e $x \mapsto \nabla^2 f(x)$ Lipschitziana. Portanto, existem $\rho > 0$, $M > 0$, tais que se $x^0 \in B(x^*, \rho)$, então a sequência (x^k) gerada pelo método de Newton está bem definida, $x^k \in B(x^*, \rho)$ e satisfaz

$$\|x^{k+1} - x^*\| \leq M \|x^k - x^*\|^2,$$

para todo $k \in \mathbb{N}$.

Prova. Usaremos H para denotar a matriz hessiana da função f , assim:

$$H(x^*) = \nabla^2 f(x^*).$$

Como $H(x^*)$ é definida positiva, existe $\rho_0 > 0$ tal que se $x \in B(x^*, \rho_0)$, então $H(x) > 0$ e, portanto, existe $H^{-1}(x)$. Como $x \mapsto H^{-1}(x)$ é contínua, então para todo $x \in B(x^*, \rho_0)$ existe uma constante β_1 tal que $\|H^{-1}(x)\| \leq \beta_1$.

A sequência (x^k) gerada pelo método de Newton é definida por

$$x^{k+1} = x^k - H^{-1}(x^k)\nabla f(x^k),$$

então:

$$\|x^{k+1} - x^*\| = \|x^k - x^* - H^{-1}(x^k)\nabla f(x^k)\|.$$

Colocando $H^{-1}(x^k)$ em evidência e considerando o fato que $\nabla f(x^*) = 0$, temos que

$$\begin{aligned} \|x^{k+1} - x^*\| &= \|H^{-1}(x^k)[\nabla f(x^*) - \nabla f(x^k) - H(x^k)(x^* - x^k)]\| \\ &\leq \|H^{-1}(x^k)\| \|\nabla f(x^*) - \nabla f(x^k) - H(x^k)(x^* - x^k)\|. \end{aligned}$$

Como $H(x^k)$ é Lipschitziana, então pelo Lema 2.16, temos que existe uma constante, digamos β_2 , tal que:

$$\|x^{k+1} - x^*\| \leq \|H^{-1}(x^k)\| \beta_2 \|x^k - x^*\|^2. \quad (2.14)$$

Como para todo $x \in B(x^*, \rho_0)$, temos $\|H^{-1}(x)\| \leq \beta_1$, então podemos reescrever (2.14) como:

$$\|x^{k+1} - x^*\| \leq \beta_1 \beta_2 \|x^k - x^*\|^2 = \beta_1 \beta_2 \|x^k - x^*\| \|x^k - x^*\|.$$

Denotando $M = \beta_1 \beta_2 > 0$ e escolhendo x^k com $\|x^k - x^*\| < \frac{1}{M}$, temos:

$$\|x^{k+1} - x^*\| \leq \beta_1 \beta_2 \|x^k - x^*\| \|x^k - x^*\| < \|x^k - x^*\|.$$

Tomando $\rho = \min\{\rho_0, \frac{1}{M}\}$, temos que para todo $x^k \in B(x^*, \rho)$, vale:

$$\|x^{k+1} - x^*\| \leq M \|x^k - x^*\|^2 < \|x^k - x^*\| < \rho.$$

A desigualdade $\|x^{k+1} - x^*\| < \|x^k - x^*\|$ mostra que o novo ponto x^{k+1} está mais próximo de x^* que o anterior x^k , e assim aplicam-se as condições outra vez a x^{k+1} . A desigualdade $\|x^{k+1} - x^*\| \leq M \|x^k - x^*\|^2$ estabelece que a ordem de convergência é quadrática, completando a demonstração. \square

Na demonstração, obtivemos a expressão (2.14):

$$\|x^{k+1} - x^*\| \leq \|H^{-1}(x^k)\| \beta_2 \|x^k - x^*\|^2. \quad (2.15)$$

Como vimos no Capítulo 1, a norma de uma matriz H simétrica definida positiva é:

$$\|H\| = \lambda_{\max}(H) \quad \text{e} \quad \|H^{-1}\| = \frac{1}{\lambda_{\min}(H)},$$

onde $\lambda_{\max}(H)$, $\lambda_{\min}(H)$ denotam o maior e menor autovalor de H . Com isto, podemos escrever:

$$\begin{aligned} \|x^{k+1} - x^*\| &\leq \|H^{-1}(x^k)\| \beta_2 \|x^k - x^*\|^2 \\ &= \frac{1}{\lambda_{\min}(H(x^k))} \beta_2 \|x^k - x^*\|^2, \end{aligned}$$

denotando $\beta_2 = L/2$ e como $\lambda_{\min}(H(x^k)) \rightarrow \lambda_{\min}(H(x^*))$, para k suficientemente grande, digamos, $k > k_1$,

$$\frac{\lambda_{\min}(H(x^*))}{2} < \lambda_{\min}(H(x^k)) \Leftrightarrow \frac{1}{\lambda_{\min}(H(x^k))} < \frac{2}{\lambda_{\min}(H(x^*))}.$$

Voltando em (2.15) temos

$$\begin{aligned} \|x^{k+1} - x^*\| &\leq \frac{2}{\lambda_{\min}(H(x^*))} \frac{L}{2} \|x^k - x^*\|^2 \\ \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} &\leq \frac{L}{\lambda_{\min}(H(x^*))} \|x^k - x^*\|. \end{aligned}$$

Como $x^k \rightarrow x^*$, para k grande, $k > k_1$, teremos $\|x_k - x^*\| < \frac{\lambda_{\min}(H(x^*))}{L}$.

Nesta região, $\{x \in \mathbb{R}^n \mid \|x - x^*\| < \frac{\lambda_{\min}(H(x^*))}{L}\}$, o erro diminui em todas as iterações $k \geq k_1$, e verifica-se a convergência quadrática:

$$\begin{aligned} \|x^{k+1} - x^*\| &< \|x^k - x^*\| \quad \text{e} \\ \|x^{k+1} - x^*\| &\leq \frac{L}{\lambda_{\min}(H(x^*))} \|x^k - x^*\|^2. \end{aligned}$$

Esta região (normalmente desconhecida) é chamada de “região de convergência quadrática” do método de Newton.

Assim, todo ponto x^0 escolhido na bola $\|x^k - x^*\| < \frac{\lambda_{\min}(H(x^*))}{L}$, converge

quadraticamente para x^* pela sequência:

$$x^{k+1} = x^k - H^{-1}(x^k) \nabla f(x^k).$$

Se, $x^0 \notin B(x^*, \frac{\lambda_{\min}(H(x^*))}{L})$, o erro pode aumentar de uma iteração para outra e, não há garantia de convergência.

2.4 Método de Região de Confiança

Como vimos, a diferença essencial entre os métodos, de Cauchy e de Newton, está na escolha da direção d^k , no Algoritmo 2.4. Nesta seção vamos apresentar um método para minimização irrestrita, o qual difere da estrutura vista no Algoritmo 2.4, que consiste em obter simultaneamente a direção e o comprimento do passo em cada iteração.

O método de região de confiança define uma região em torno do ponto corrente x^k , na qual confia-se num modelo da função objetivo, e então calcula-se o minimizador aproximado do modelo na região de confiança. Assim, a direção e o comprimento do passo são obtidos simultaneamente. Se o ponto não é aceito, reduzimos o tamanho da região e encontramos um novo minimizador. Em geral, a direção do passo muda quando o tamanho da região é alterado.

Aplicaremos o método de região de confiança ao problema irrestrito:

$$(P) \quad \begin{aligned} & \text{minimizar} && f(x) \\ & x \in \mathbb{R}^n, \end{aligned}$$

onde $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é de classe \mathcal{C}^2 .

Um modelo quadrático de f em torno do ponto corrente x^k é:

$$m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p, \quad (2.16)$$

onde $p = x - x^k$, $f_k = f(x^k)$, $\nabla f_k = \nabla f(x^k)$ e $B_k \in \mathbb{R}^{n \times n}$ é uma matriz simétrica, podendo ser uma aproximação da hessiana $\nabla^2 f(x^k)$ ou qualquer outra matriz simétrica que satisfaça

$$\|B_k\| \leq \beta,$$

para alguma constante $\beta > 0$, independente de k .

Como o modelo quadrático (2.16) deixa de ser representativo à medida que x se afasta de x^k , podemos confiar em aproximar $f(\cdot)$ por $m_k(\cdot)$ numa vizinhança de x^k , ou seja, na região:

$$\{p \in \mathbb{R}^n \mid \|p\| \leq \Delta_k\}, \quad (2.17)$$

onde $\Delta_k > 0$ é dito o raio da região de confiança e $\|\cdot\|$ é uma norma qualquer em \mathbb{R}^n .

Dessa forma, uma solução $p^k \in \mathbb{R}^n$ para o problema:

$$\begin{aligned} \text{minimizar } m_k(p) &= f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \\ \text{s.a. } \|p\| &\leq \Delta_k \end{aligned} \quad (2.18)$$

seria uma boa aproximação para o minimizador de f na região (2.17). O passo de região de confiança p^k depende do raio Δ_k da região e pode mudar de direção quando o raio é alterado.

Dado um passo p^k , definimos a redução **predita** produzida pelo passo p^k , como:

$$pred = m_k(0) - m_k(p^k)$$

e a redução **verdadeira** como:

$$ared = f(x^k) - f(x^k + p^k)$$

e então, calculamos a razão:

$$\rho_k = \frac{f(x^k) - f(x^k + p^k)}{m_k(0) - m_k(p^k)}. \quad (2.19)$$

Note que, como o passo p^k é obtido minimizando-se o modelo m_k sobre uma região que inclui o ponto x^k , a redução predita será sempre não negativa.

Assim, se ρ_k é negativo, o novo valor $f(x^k + p^k)$ é maior que o anterior $f(x^k)$ e o passo p^k deve ser rejeitado.

Por outro lado, se ρ_k está próximo de 1, temos que a redução verdadeira está próxima da redução do modelo, assim podemos aumentar a região de confiança na próxima iteração.

Se ρ_k é positivo, mas não está próximo de 1, mantém-se o raio da região, mas se ρ_k está próximo de zero ou é negativo diminuímos a região.

O algoritmo a seguir ligeiramente adaptado de [23], descreve este processo.

Algoritmo 2.18 *Região de confiança*

Dados: $\bar{\Delta} > 0$, $\Delta_0 \in (0, \bar{\Delta})$ e $\eta \in [0, \frac{1}{4})$:

$k = 0$

REPITA enquanto $(\|\nabla f(x^k)\| > \epsilon)$

Obter p^k aproximadamente resolvendo (2.18);

Calcular ρ_k por (2.19)

Se $\rho_k < \frac{1}{4}$

$$\Delta_{k+1} = \frac{1}{4}\Delta_k$$

Senão

Se $\rho_k > \frac{3}{4}$ e $\|p^k\| = \Delta_k$

$$\Delta_{k+1} = \min\{2\Delta_k, \bar{\Delta}\}$$

Senão

$$\Delta_{k+1} = \Delta_k;$$

Se $\rho_k > \eta$

$$x^{k+1} = x^k + p^k$$

Senão

$$x^{k+1} = x^k;$$

$k = k + 1$

FIM

A Figura 2.6 mostra uma iteração do algoritmo de região de confiança. As curvas de nível da função objetivo do problema original estão representadas por linhas tracejadas. As linhas cheias representam as curvas de nível do modelo em x^k . O ponto x^{k+1} é a minimização do modelo na região de confiança centrada em x^k .

No Algoritmo 2.18, $\bar{\Delta}$ é um limitante para o raio da região de confiança, assim, para todo k temos $\Delta_k \leq \bar{\Delta}$.

Em cada iteração, basta resolvermos aproximadamente o problema (2.18), isto é, que a redução no modelo em (2.18) seja tão boa quanto a redução encontrada pelo *ponto de Cauchy*. O *passo de Cauchy*, denotado por p_c^k , é simplesmente o minimizador de $m_k(\cdot)$, ao longo da direção oposta ao gradiente $-\nabla f(x^k)$, sujeito a região de confiança.

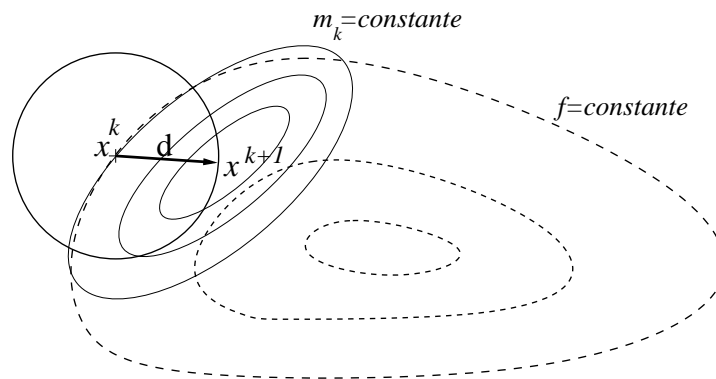


Figura 2.6: Um passo de região de confiança

2.4.1 Convergência global

Para provarmos que o método de região de confiança é globalmente convergente devemos ter

$$pred \geq m_k(0) - m_k(p_c^k). \quad (2.20)$$

A Figura 2.7 mostra o ponto de Cauchy numa iteração k . As elipses representam as curvas de nível do modelo $m_k(\cdot)$. A área hachurada corresponde ao conjunto de pontos que satisfazem (2.20).

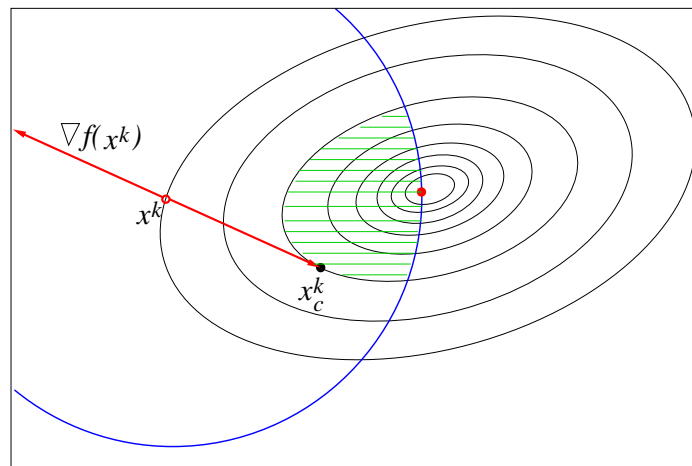


Figura 2.7: Ponto de Cauchy.

Sabemos que o *ponto de Cauchy* (veja, por exemplo, [23]) satisfaz

$$m_k(0) - m_k(p_c^k) \geq c_1 \|\nabla f_k\| \min \left\{ \Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|} \right\} \quad (2.21)$$

para alguma constante $c_1 \in (0, 1]$.

Com isto, o método de região de confiança é convergente se, em cada iteração k , o ponto $x^{k+1} = x^k + p^k$ é pelo menos tão bom quanto o *ponto de Cauchy* $x_c^k = x^k + p_c^k$, no sentido de que

$$pred = m_k(0) - m_k(p^k) \geq m_k(0) - m_k(p_c^k) \geq c_1 \|\nabla f_k\| \min \left\{ \Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|} \right\}. \quad (2.22)$$

Dos algoritmos que resolvem o problema (2.18), podemos usar o método *Dogleg*, que é apropriado quando a matriz B_k é definida positiva, ou o método devido a Steihaug [29], que é mais apropriado quando $B_k = \nabla^2 f(x^k)$ e quando essa matriz é esparsa. Para mais detalhes desses métodos e seus algoritmos veja, por exemplo, [23].

Para generalizar, permitiremos também que o tamanho do passo p^k , solução aproximada de (2.18), exceda o raio da região de confiança, contanto que permaneça dentro de algum múltiplo fixo do raio, isto é,

$$\|p^k\| \leq \gamma \Delta_k, \quad (2.23)$$

para alguma constante $\gamma \geq 1$.

O seguinte teorema estabelecido em [23] garante que o método de região de confiança é globalmente convergente.

Teorema 2.19 *Convergência global do método de região de confiança*

Considere $\eta \in (0, \frac{1}{4})$ no Algoritmo 2.18. Suponha que $\|B_k\| \leq \beta$ para alguma constante $\beta > 0$, f é Lipschitz continuamente diferenciável² e limitada inferiormente no conjunto de nível

$$\{x \mid f(x) \leq f(x^0)\} \quad (2.24)$$

e que toda solução aproximada p^k de (2.18) satisfaz as desigualdades:

$$m_k(0) - m_k(p^k) \geq c_1 \|\nabla f_k\| \min \left\{ \Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|} \right\}, \quad (2.25)$$

$$\|p^k\| \leq \gamma \Delta_k, \quad (2.26)$$

²Uma função f é Lipschitz continuamente diferenciável quando $f \in C^1$ e $x \mapsto \nabla f(x)$ é Lipschitz.

com $c_1 \in (0, 1]$, e $\gamma \geq 1$. Temos, então:

$$\lim_{k \rightarrow \infty} \nabla f_k = 0.$$

Prova. Considere algum índice m , tal que $\nabla f(x^m) = \nabla f_m \neq 0$.

Seja β_1 a constante de *Lipschitz* para ∇f no conjunto de nível (2.33), assim:

$$\|\nabla f(x) - \nabla f_m\| \leq \beta_1 \|x - x^m\|,$$

para todo x no conjunto de nível. Definindo os escalares

$$\epsilon = \frac{1}{2} \|\nabla f_m\|, \quad R = \frac{\|\nabla f_m\|}{2\beta_1} = \frac{\epsilon}{\beta_1}$$

e a bola

$$B(x^m, R) = \{x \mid \|x - x^m\| \leq R\},$$

temos:

$$\begin{aligned} x \in B(x^m, R) &\Rightarrow \|\nabla f(x) - \nabla f_m\| \geq \|\nabla f_m\| - \|\nabla f(x)\| \\ &\Rightarrow \|\nabla f(x)\| \geq \|\nabla f_m\| - \|\nabla f(x) - \nabla f_m\| \\ &\geq \|\nabla f_m\| - \beta_1 \|x - x^m\| \\ &\geq \|\nabla f_m\| - \beta_1 R \\ &= \frac{1}{2} \|\nabla f_m\| = \epsilon. \end{aligned}$$

Se a sequência inteira $\{x^k\}_{k \geq m}$ permanece dentro da bola $B(x^m, R)$, temos da última desigualdade que $\|\nabla f_k\| \geq \epsilon > 0$ para todo $k \geq m$.

Vejamos que isto não ocorre por redução ao absurdo.

Fazendo uma manipulação com o ρ_k da equação (2.19), temos:

$$|\rho_k - 1| = \left| \frac{(f(x^k) - f(x^k + p^k)) - (m_k(0) - m_k(p^k))}{m_k(0) - m_k(p^k)} \right| = \left| \frac{m_k(p^k) - f(x^k + p^k)}{m_k(0) - m_k(p^k)} \right|.$$

Usando a Fórmula de Taylor com resto integral (veja Capítulo 1), temos:

$$f(x^k + p^k) = f(x^k) + \int_0^1 \nabla f(x^k + tp^k)^T p^k dt.$$

Como $\int_0^1 \nabla f(x^k)^T p^k dt = \nabla f(x^k)^T p^k$, podemos escrever:

$$f(x^k + p^k) = f(x^k) + \nabla f(x^k)^T p^k + \int_0^1 [\nabla f(x^k + tp^k) - \nabla f(x^k)]^T p^k dt.$$

Da Definição 2.16: $m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p$, então:

$$\begin{aligned} |m_k(p^k) - f(x^k + p^k)| &= \left| \frac{1}{2} p^{kT} B_k p^k - \int_0^1 [\nabla f(x^k + tp^k) - \nabla f(x^k)]^T p^k dt \right| \\ &\leq \left| \frac{1}{2} p^{kT} B_k p^k \right| + \left| \int_0^1 [\nabla f(x^k + tp^k) - \nabla f(x^k)]^T p^k dt \right| \\ &\leq \frac{\beta}{2} \|p^k\|^2 + \int_0^1 |[\nabla f(x^k + tp^k) - \nabla f(x^k)]^T p^k| dt. \end{aligned}$$

Pela desigualdade de *Cauchy-Schwarz*,

$$\begin{aligned} \int_0^1 |[\nabla f(x^k + tp^k) - \nabla f(x^k)]^T p^k| dt &\leq \int_0^1 \|[\nabla f(x^k + tp^k) - \nabla f(x^k)]\| \|p^k\| dt \\ &= \left(\int_0^1 \|[\nabla f(x^k + tp^k) - \nabla f(x^k)]\|^2 dt \right)^{1/2} \|p^k\| \\ &= (C_4(p^k)) \|p^k\|, \end{aligned}$$

onde

$$C_4(p^k) = \int_0^1 \|[\nabla f(x^k + tp^k) - \nabla f(x^k)]\|^2 dt.$$

Daí

$$|m_k(p^k) - f(x^k + p^k)| \leq \frac{\beta}{2} \|p^k\|^2 + C_4(p^k) \|p^k\|. \quad (2.27)$$

Supondo, então, $\|\nabla f_k\| \geq \epsilon$, temos de (2.25) que para todo $k \geq m$

$$m_k(0) - m_k(p^k) \geq c_1 \|\nabla f_k\| \min \left\{ \Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|} \right\} \geq c_1 \epsilon \min \left\{ \Delta_k, \frac{\epsilon}{\beta} \right\}. \quad (2.28)$$

Usando (2.26), (2.27) e (2.28), temos:

$$|\rho_k - 1| \leq \frac{\gamma \Delta_k \left(\frac{\beta}{2} \gamma \Delta_k + C_4(p^k) \right)}{c_1 \epsilon \min \left\{ \Delta_k, \frac{\epsilon}{\beta} \right\}}. \quad (2.29)$$

Mostraremos, agora, que existe $\tilde{\Delta} > 0$, tal que o lado direito de (2.29) é limitado por $\frac{1}{2}$, para todo $k \geq m$, satisfazendo $\Delta_k \leq \tilde{\Delta}$.

Pela continuidade da função $x \mapsto \nabla f(x)$, para $\bar{\epsilon} = \frac{c_1\epsilon}{4\gamma}$, existe $\delta > 0$, tal que

$$\|p^k\| \leq \delta \Rightarrow \|\nabla f(x^k + tp^k) - \nabla f(x^k)\| \leq \bar{\epsilon},$$

para todo $t \in [0, 1]$.

Seja $\tilde{\Delta} = \min \left\{ \frac{c_1\epsilon}{2\beta\gamma^2}, \frac{\delta}{\gamma}, \frac{\epsilon}{\beta} \right\}$. Deste modo, para $\Delta_k \leq \tilde{\Delta}$, temos:

- $\Delta_k \leq \frac{c_1\epsilon}{2\beta\gamma^2}$, donde segue que $\frac{\beta}{2}\gamma\Delta_k \leq \frac{c_1\epsilon}{4\gamma}$;
- $\Delta_k \leq \frac{\delta}{\gamma}$, portanto $\gamma\Delta_k \leq \delta$ e, como $\|p^k\| \leq \gamma\Delta_k$, temos, para todo $t \in [0, 1]$, $\|\nabla f(x^k + tp^k) - \nabla f(x^k)\| \leq \bar{\epsilon}$. Logo

$$C_4(p^k) = \int_0^1 \|\nabla f(x^k + tp^k) - \nabla f(x^k)\| dt \leq \bar{\epsilon} = \frac{c_1\epsilon}{4\gamma};$$

- $\Delta_k \leq \frac{\epsilon}{\beta}$, o que implica $\min \left\{ \Delta_k, \frac{\epsilon}{\beta} \right\} = \Delta_k$.

Reescrevendo (2.29), temos, para $\Delta_k \leq \tilde{\Delta}$

$$\begin{aligned} |\rho_k - 1| &\leq \frac{\gamma\Delta_k \left(\frac{\beta}{2}\gamma\Delta_k + C_4(p^k) \right)}{c_1\epsilon \min \left\{ \Delta_k, \frac{\epsilon}{\beta} \right\}} = \frac{\gamma}{c_1\epsilon} \left(\frac{\beta}{2}\gamma\Delta_k + C_4(p^k) \right) \\ &\leq \frac{\gamma}{c_1\epsilon} \left(\frac{c_1\epsilon}{4\gamma} + \frac{c_1\epsilon}{4\gamma} \right) = \frac{1}{2}. \end{aligned}$$

Consequentemente, $\rho_k > \frac{1}{4}$ e assim pelo Algoritmo 2.18:

$$\Delta_{k+1} = \min\{2\Delta_k, \tilde{\Delta}\} \quad \text{ou} \quad \Delta_{k+1} = \Delta_k$$

Como $\Delta_k \leq \tilde{\Delta}$, para todo $k \in \mathbb{N}$, temos $\Delta_{k+1} \geq \Delta_k$. Então, se $\Delta_k \leq \tilde{\Delta}$, temos $\Delta_{k+1} \geq \Delta_k$. Caso $\Delta_{k+1} < \Delta_k$, temos $\Delta_k > \tilde{\Delta}$ e $\Delta_{k+1} = \frac{1}{4}\Delta_k \geq \frac{1}{4}\tilde{\Delta}$.

Conclusão:

$$\Delta_k \geq \min \left\{ \Delta_m, \frac{1}{4} \tilde{\Delta} \right\}, \quad (2.30)$$

para todo $k \geq m$. Seja $\mathcal{K} \subset \mathbb{N}$ definido por:

$$\mathcal{K} = \left\{ k \in \mathbb{N} \mid \rho_k \geq \frac{1}{4} \right\}.$$

Caso \mathcal{K} seja infinito, $k \geq m$ e de (2.28), temos:

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

Como, por hipótese, f é limitada inferiormente, segue desta desigualdade que

$$\Delta_k \xrightarrow{\mathcal{K}} 0.$$

Caso \mathcal{K} seja finito, $\rho_k < \frac{1}{4}$ para todo k suficientemente grande, neste caso, Δ_k será reduzido por um fator de $\frac{1}{4}$ em cada iteração, e teremos $\Delta_k \rightarrow 0$. Portanto, pelo menos uma subsequência de $(\Delta_k)_{k \in \mathbb{N}}$ tende a zero, contradizendo (2.30). Assim, a afirmação que $\|\nabla f_k\| \geq \epsilon > 0$ para todo $k \geq m$ é falsa e, conseqüentemente, a sequência $\{x^k\}_{k \geq m}$ sai da bola $B(x^m, R)$. Seja o índice $l \geq m$ tal que x^{l+1} é o primeiro iterando depois de x^m que está fora da bola $B(x^m, R)$.

Assim, $\|\nabla f_k\| \geq \epsilon$ para $k = m, m+1, \dots, l$, e

$$m_k(0) - m_k(p^k) \geq c_1 \|\nabla f_k\| \min \left\{ \Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|} \right\} \geq c_1 \epsilon \min \left\{ \Delta_k, \frac{\epsilon}{\beta} \right\}.$$

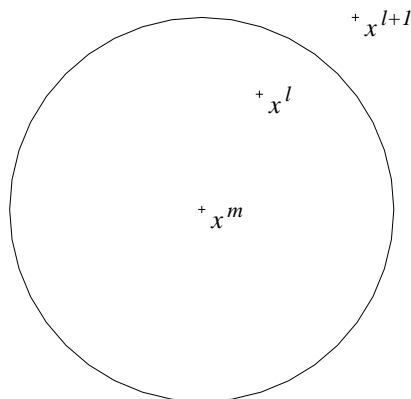


Figura 2.8: Bola de centro x^m e raio R .

Logo,

$$\begin{aligned}
 f(x^m) - f(x^{l+1}) &= \sum_{k=m}^l f(x^k) - f(x^{k+1}) \\
 &= \sum_{k=m}^l \rho^k [m_k(0) - m_k(p^k)] \\
 &\geq \sum_{k=m; x^k \neq x^{k+1}}^l \eta [m_k(0) - m_k(p^k)] \\
 &\geq \sum_{k=m; x^k \neq x^{k+1}}^l \eta c_1 \epsilon \min \left\{ \Delta_k, \frac{\epsilon}{\beta} \right\},
 \end{aligned}$$

onde limitamos o somatório às iterações k com $x^k \neq x^{k+1}$, isto é, às iterações em que o passo p^k foi aceito pelo Algoritmo 2.18.

Há dois casos a considerar:

- Se $\Delta_k \leq \frac{\epsilon}{\beta}$ para todo $k = m, \dots, l$, então:

$$f(x^m) - f(x^{l+1}) \geq \eta c_1 \epsilon \sum_{k=m; x^k \neq x^{k+1}}^l \Delta_k \geq \eta c_1 \epsilon R = \eta c_1 \epsilon^2 \frac{1}{\beta_1}. \quad (2.31)$$

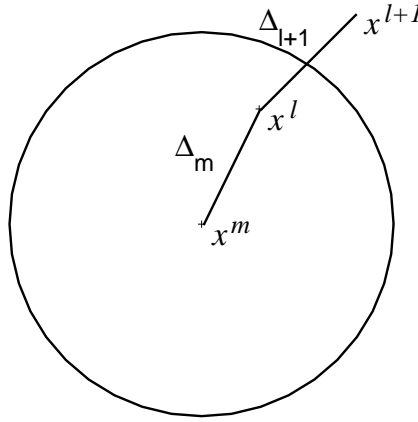


Figura 2.9: A soma dos Δ_k é maior que ou igual a R .

- Se $\Delta_k > \frac{\epsilon}{\beta}$ para algum $k = m, \dots, l$,

$$f(x^m) - f(x^{l+1}) \geq \eta c_1 \epsilon \sum_{k=m; x^k \neq x^{k+1}}^l \min \left\{ \Delta_k, \frac{\epsilon}{\beta} \right\} \geq \eta c_1 \epsilon \frac{\epsilon}{\beta} = \eta c_1 \epsilon^2 \frac{1}{\beta}, \quad (2.32)$$

pois o somatório todo é maior ou igual a uma parcela.

A sequência $\{f(x^m)\}_{m=0}^{\infty}$ é decrescente (pois o algoritmo é de descida) e limitada (por hipótese f é limitada inferiormente no conjunto de nível (2.33)), portanto f é convergente para algum $f^* > -\infty$, isto é,

$$f(x^m) \rightarrow f^*.$$

Consequentemente, usando (2.31) e (2.32), temos:

$$f(x^m) - f^* \geq f(x^m) - f(x^{l+1}) \geq \eta c_1 \epsilon^2 \min \left\{ \frac{1}{\beta}, \frac{1}{\beta_1} \right\}.$$

Como $\epsilon^2 = \frac{1}{4} \|\nabla f_m\|^2$

$$f(x^m) - f^* \geq f(x^m) - f(x^{l+1}) \geq \frac{1}{4} \eta c_1 \min \left\{ \frac{1}{\beta}, \frac{1}{\beta_1} \right\} \|\nabla f_m\|^2.$$

Portanto,

$$\|\nabla f_m\|^2 \leq \left[\frac{1}{4} \eta c_1 \min \left\{ \frac{1}{\beta}, \frac{1}{\beta_1} \right\} \right]^{-1} (f(x^m) - f^*).$$

Como $f(x^m) \rightarrow f^*$, tomando limite na última desigualdade acima, obtemos

$$\lim_{m \rightarrow \infty} \|\nabla f_m\|^2 = 0$$

completando a demonstração. \square

Assim, o teorema mostra que o método de região de confiança é globalmente convergente, isto é, que todo valor de aderência \bar{x} de uma sequência gerada pelo Algoritmo 2.18 é estacionário, ou seja, $\nabla f(\bar{x}) = 0$.

De fato, se \bar{x} é valor de aderência da sequência (x^k) gerada pelo Algoritmo 2.18, existe um subconjunto $\mathcal{K} \subset \mathbb{N}$ tal que $x^k \xrightarrow{\mathcal{K}} \bar{x}$. Como $f(\cdot)$ é continuamente diferenciável, temos

$$\nabla f(x^k) \xrightarrow{\mathcal{K}} \nabla f(\bar{x}),$$

mas do Teorema 2.19, temos que

$$\nabla f(x^k) \rightarrow 0,$$

portanto

$$\nabla f(x^k) \xrightarrow{\mathcal{K}} 0,$$

daí, pela unicidade do limite,

$$\nabla f(\bar{x}) = 0.$$

2.4.2 Convergência local

Apresentamos, sem demonstração, o importante teorema, estabelecido em Conn, Gould, Toint [7], o qual garante convergência local quadrática para o método de região de confiança. Este teorema garante que a região de confiança torna-se inativa, isto é, a partir de um certo ponto o novo iterando sempre está dentro da região de confiança e não na sua fronteira. Assim o algoritmo de região de confiança torna-se equivalente ao método de Newton o qual consiste em minimizar, em cada iteração, a aproximação quadrática da função objetivo sem restrição. Portanto, as propriedades de convergências são as mesmas do método de Newton “puro”, ou seja, velocidade de convergência quadrática para a sequência gerada pelo algoritmo região de confiança.

Teorema 2.20 *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f \in \mathcal{C}^2$ e limitada inferiormente no conjunto de nível*

$$\{x \mid f(x) \leq f(x^0)\} \quad (2.33)$$

com matriz hessiana uniformemente limitada.³ Suponha que m_k é de classe \mathcal{C}^2 na região (2.17) e que os valores $m_k(x^k)$ e $\nabla m_k(x^k)$ são iguais a $f(x^k)$ e $\nabla f(x^k)$, respectivamente, na iteração corrente x^k para todo k . Assuma que a hessiana do modelo m_k permanece limitada na região (2.17), que

$$\lim_{k \rightarrow \infty} \|\nabla^2 f(x^k) - \nabla^2 m_k(x^k)\| = 0,$$

e para todo k o ponto $x^{k+1} = x^k + p^k$ é pelo menos tão bom quanto o passo de Cauchy p_c^k , no sentido de que (2.22) se verifica para alguma constante $c_1 \in (0, 1]$ e que $(x^k)_{k \in \mathbb{N}}$ é uma subsequência gerada pelo Algoritmo 2.18 convergindo para um ponto estacionário x^ . Suponha, além disso, que $p^k \neq 0$ para todo k suficientemente grande. Finalmente, suponha que $\nabla^2 f(x^*)$ é definida positiva. Então a sequência inteira (x^k) converge para x^* , toda iteração está bem definida, e o raio Δ_k da região de confiança é suficientemente positivo.*

Prova. [7, pág. 146] . □

³A matriz hessiana da função objetivo é uniformemente limitada se existe uma constante positiva β tal que, para todo $x \in \mathbb{R}^n$, $\|\nabla^2 f(x)\| \leq \beta$.

Capítulo 3

Minimização com restrições de igualdade

Neste Capítulo, apresentamos dois métodos para resolver o problema de programação não linear com restrições de igualdade.

Ao resolver um problema de programação não linear com restrições, é necessário, simultaneamente, reduzir uma função objetivo e obter viabilidade. Esses dois objetivos são conflitantes, e o balanceamento dos esforços em viabilização e otimização são, normalmente, a maior dificuldade nos algoritmos.

Na primeira seção, apresentamos o método de programação quadrática sequencial (PQS) e, mostramos que a sequência gerada pelo algoritmo “puro” converge quadraticamente para a solução, desde que o ponto inicial esteja suficientemente próximo da solução. Descrevemos uma estratégia para globalizar o método PQS, introduzindo ao subproblema a restrição de região de confiança e usamos a técnica de Byrd e Omojokum [6, 24], que decompõe o passo em duas componentes: normal, que está relacionado com a redução da inviabilidade, e tangencial, com a melhora da otimalidade. Construimos o algoritmo PQS global, escolhendo uma função de mérito, que é uma combinação convexa de uma medida de otimalidade e uma medida de inviabilidade, como critério de aceitação do passo, que segue o mesmo raciocínio do método de região de confiança. Para evidenciar que o algoritmo PQS global com região de confiança pode sofrer do efeito Maratos, o aplicamos ao exemplo de Powell [27].

Na segunda seção, apresentamos o método de filtro (GKV), proposto por Clóvis Gonzaga, Elizabeth Karas e Márcia Vanti [13], em que as fases de viabili-

dade e otimalidade são totalmente independentes e o único acoplamento entre elas é estabelecido pelo filtro. Explicamos seu funcionamento e escrevemos o algoritmo de filtro GKV de uma forma compacta, sem especificar os algoritmos internos. Comentamos sobre sua convergência global para pontos estacionários, que independe dos algoritmos internos usados, em cada iteração, desde que esses algoritmos satisfaçam certas hipóteses sobre sua eficiência. Deixamos em aberto a questão sobre a convergência local do algoritmo de filtro GKV que, assim como o método PQS com função de mérito, pode sofrer do efeito Maratos, se uma correção de segunda ordem não for efetuada.

3.1 Método de programação quadrática sequencial

Um método clássico para resolver o problema geral de programação não linear é o de programação quadrática sequencial (PQS). Neste problema, onde temos uma função objetivo e um conjunto de restrições, geralmente não lineares, a idéia consiste em substituir, em cada passo, a função objetivo por uma aproximação quadrática, e as restrições por aproximações lineares. Dessa maneira, o subproblema a ser resolvido em cada iteração k é um problema de programação quadrática que, em comparação ao original, pode ser considerado simples.

Aplicaremos o método de programação quadrática sequencial ao problema com restrições não lineares de igualdade:

$$(PE) \quad \begin{array}{l} \text{minimizar} \quad f_o(x) \\ \text{sujeito a :} \quad f(x) = 0, \end{array}$$

onde $f_o : \mathbb{R}^n \rightarrow \mathbb{R}$ e $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $m < n$, são funções continuamente diferenciáveis.

Usaremos $A(x)$ para denotar a matriz Jacobiana das restrições, isto é,

$$A(x)^T = [\nabla f_1(x) \quad \nabla f_2(x) \quad \dots \quad \nabla f_m(x)]. \quad (3.1)$$

Para que o conjunto viável de (PE) não seja vazio ou um único ponto vamos supor $\text{posto}(A(x)) = m$, para todo $x \in \mathbb{R}^n$.

A função Lagrangeana para o problema (PE) é definida como:

$$L(x, \lambda) = f_o(x) + \lambda^T f(x). \quad (3.2)$$

Escrevendo as condições necessárias de otimalidade de primeira ordem (condições de KKT) para o problema (PE) , obtemos um sistema de $n+m$ equações com incógnitas (x, λ) em \mathbb{R}^{n+m} :

$$F(x, \lambda) = \begin{bmatrix} \nabla f_o(x) + A^T(x)\lambda \\ f(x) \end{bmatrix} = 0. \quad (3.3)$$

Observe que $\nabla f_o(x) + A^T(x)\lambda$ é igual a $\nabla_x L(x, \lambda)$.

Para encontrar os candidatos à solução do problema (PE) , precisamos resolver o sistema não linear $F(x, \lambda) = 0$. Usaremos o método de Newton para a função $F : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^{n+m}$. Na iteração k , o passo de Newton $[d^k, d^\lambda]$ é a solução do sistema:

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d^k \\ d^\lambda \end{bmatrix} = \begin{bmatrix} -(\nabla f_{o_k} + A_k^T \lambda^k) \\ -f_k \end{bmatrix}, \quad (3.4)$$

onde $W_k = W(x^k, \lambda^k) = \nabla_{xx}^2 L(x^k, \lambda^k) = \nabla^2 f_o(x^k) + \sum_{i=1}^m \lambda_i^k \nabla^2 f_i(x^k)$ é a matriz hessiana da função Lagrangeana, $A_k = A(x^k)$, $\nabla f_{o_k} = \nabla f_o(x^k)$ e $f_k = f(x^k)$.

Note que a matriz Jacobiana de F em (x^k, λ^k) é:

$$J_k = J(x^k, \lambda^k) = \begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix}. \quad (3.5)$$

O passo de Newton está bem definido quando a matriz J_k é não-singular. Para mostrar a não-singularidade de J_k , usaremos as seguintes hipóteses:

(H1) A Jacobiana das restrições tem posto completo, isto é, $\text{posto}(A_k) = m$;

(H2) A matriz W_k é definida positiva no núcleo da Jacobiana das restrições, isto é, $d^T W_k d > 0$ para todo $d \neq 0$ tal que $A_k d = 0$.

Lema 3.1 *Sob as condições acima a matriz dada por (3.5) é não singular.*

Prova. Para simplificar a notação, vamos suprimir o índice k na matriz (3.5). Considere o sistema homogêneo:

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Leftrightarrow \begin{cases} W\alpha + A^T\beta = 0 \\ A\alpha = 0 \end{cases} \quad (3.6)$$

Multiplicando a primeira equação do sistema (3.6) por α^T , temos:

$$\alpha^T W \alpha + (A\alpha)^T \beta = 0.$$

Como $A\alpha = 0$ (segunda equação em (3.6)), $\alpha^T W \alpha = 0$. Mas W é positiva definida em $N(A)$, logo $\alpha = 0$. Substituindo $\alpha = 0$ na primeira equação do sistema (3.6), obtemos $A^T \beta = 0$, logo $AA^T \beta = 0$. Como $\text{posto}(AA^T) = \text{posto}(A) = m$, a matriz (AA^T) é não singular e conseqüentemente $\beta = 0$. Assim, o sistema homogêneo (3.6) só tem a solução trivial, portanto a matriz J_k é não singular. \square

Com as condições acima estabelecidas, o algoritmo de Newton, para resolver um sistema não linear, converge quadraticamente para a solução, constituindo um ótimo algoritmo para resolver problemas não lineares com restrições de igualdade, desde que o ponto inicial esteja suficientemente próximo da solução.

Estrutura PQS. Há uma maneira alternativa para ver o passo estabelecido em (3.4). Dado $(x^k, \lambda^k) \in \mathbb{R}^{n+m}$, considere o subproblema quadrático:

$$\begin{aligned} & \text{minimizar} && \frac{1}{2} p^T W_k p + \nabla f_{o_k}^T p \\ & \text{sujeito a :} && A_k p + f_k = 0. \end{aligned} \quad (3.7)$$

Devemos notar que a função objetivo do problema (3.7) difere apenas por uma constante da aproximação quadrática da função Lagrangeana. De fato, dado $(x^k, \lambda^k) \in \mathbb{R}^{n+m}$ o modelo quadrático para a função Lagrangeana é:

$$M_L(p) = L(x^k, \lambda^k) + \nabla L(x^k, \lambda^k)^T p + \frac{1}{2} p^T W_k p. \quad (3.8)$$

Mas

$$\nabla L(x^k, \lambda^k)^T p = \nabla f_{o_k}^T p + \lambda^{kT} A_k p,$$

logo (3.8) pode ser reescrita como:

$$M_L(p) = \frac{1}{2} p^T W_k p + \nabla f_{o_k}^T p + \nu$$

com a constante $\nu = L(x^k, \lambda^k) + \lambda^{kT} A_k p = f_{o_k} + \lambda^{kT} (f_k + A_k p)$. Considerando a restrição do subproblema (3.7), temos que $\nu = f_{o_k}$. Assim, cada iteração do método de PQS “puro”, consiste em minimizar o modelo quadrático da função Lagrangeana, sujeito à linearização das restrições.

Se as condições usadas para mostrar a não-singularidade de J_k se verificam, o subproblema (3.7) tem uma única solução (p^k, μ^k) , que satisfaz:

$$\nabla L(p^k, \mu^k) = \begin{bmatrix} W_k p^k + \nabla f_{o_k}^T + A_k^T \mu^k \\ A_k p^k + f_k \end{bmatrix} = 0,$$

onde $\nabla L(p^k, \mu^k)$ pode ser reescrito como

$$\nabla L(p^k, \mu^k) = \begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} p^k \\ \mu^k \end{bmatrix} = \begin{bmatrix} -\nabla f_{o_k} \\ -f_k \end{bmatrix}. \quad (3.9)$$

O fato que ocorre aqui é que p^k e μ^k podem ser relacionados com a solução das equações de Newton (3.4). Se adicionarmos $A_k^T \lambda^k$ de ambos os lados da primeira equação em (3.4), obtemos:

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d^k \\ \lambda^k + d^\lambda \end{bmatrix} = \begin{bmatrix} -\nabla f_{o_k} \\ -f_k \end{bmatrix}.$$

Assim, pela não singularidade da matriz J_k , temos que

- $d^k = p^k$ e
- $\lambda^k + d^\lambda = \lambda^{k+1} = \mu^k$.

A interpretação em termos do método de Newton facilita a análise de convergência, enquanto que a estrutura da PQS, permite desenvolver algoritmos práticos para resolver o problema de programação não linear com restrições de igualdade.

O algoritmo a seguir descreve este processo.

Algoritmo 3.2 *PQS local*

Dado (x^0, λ^0)

$k = 0$

REPITA

Calcule f_{o_k} , ∇f_{o_k} , W_k , f_k e A_k ;

Resolva (3.7) e obtenha p^k e μ^k ;

$x^{k+1} = x^k + p^k$; $\lambda^{k+1} = \mu^k$;

se satisfaz teste de convergência PARE com solução aproximada (x^{k+1}, λ^{k+1})

$k = k + 1$

FIM

3.1.1 Convergência local

Os resultados de convergência serão equivalentes aos do método de Newton. Se o ponto inicial (x^0, λ^0) está suficientemente próximo da solução (x^*, λ^*) , as condições de não-singularidade da matriz J_k se verificam e, ainda se as funções f_o e f são de classe \mathcal{C}^2 com derivadas segundas Lipschitz contínuas, então a sequência gerada pelo Algoritmo 3.2 converge quadraticamente para (x^*, λ^*) .

O seguinte teorema, estabelecido em Bonnans, Gilbert, Lemaréchal, Sagastizábal [3], garante que a velocidade de convergência, da sequência gerada pelo algoritmo PQS local é quadrática.

Teorema 3.3 *Convergência local do método de PQS*

Suponha que as funções f_o e f são de classe \mathcal{C}^2 , com derivadas segundas Lipschitz contínuas, em uma vizinhança de um ponto estacionário x^ de (PE) que satisfaz (H1) e (H2), com multiplicador associado λ^* . Então, existe uma vizinhança V de (x^*, λ^*) , tal que, se o primeiro iterado $(x^0, \lambda^0) \in V$, o Algoritmo 3.2 está bem definido, e gera uma sequência (x^k, λ^k) que converge quadraticamente para (x^*, λ^*) .*

Prova. O resultado é obtido aplicando o método de Newton para resolver o sistema não linear:

$$F(x, \lambda) = \begin{bmatrix} \nabla f_o(x) + A^T(x)\lambda \\ f(x) \end{bmatrix}.$$

Seja $J(x, \lambda)$ a matriz Jacobiana de $F(x, \lambda)$. Mostraremos que $J(x, \lambda)$ é não singular numa vizinhança V de (x^*, λ^*) .

Temos, por hipótese, que (x^*, λ^*) satisfaz (H1) e (H2). Portanto, pelo Lema 3.1, $J(x^*, \lambda^*)$ é não singular.

Para mostrar que J é não-singular em uma vizinhança V de (x^*, λ^*) , definimos a função $\varphi : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ por

$$\varphi(x, \lambda) = \text{determinante} \begin{bmatrix} W(x, \lambda) & A^T(x, \lambda) \\ A(x, \lambda) & 0 \end{bmatrix}.$$

Note que $\varphi(x^*, \lambda^*) \neq 0$. Assim, por continuidade, existe uma vizinhança V de (x^*, λ^*) tal que $\varphi(x, \lambda) \neq 0$ para todo $(x, \lambda) \in V$. Com isto, temos que as hipóteses deste teorema satisfazem as do método de Newton, no qual a convergência quadrática da sequência é garantida desde que o ponto inicial $(x^0, \lambda^0) \in V$. \square

O exemplo, a seguir, mostra que a velocidade de convergência quadrática da sequência (x^k, λ^k) não implica na mesma velocidade para a sequência (x^k) .

Exemplo 10 (Bonnans et al [3])

Seja $y_1 \in (0, 1)$ e considere a sequência $(x_k, y_k)_{k \geq 1} \in \mathbb{R}^2$ gerada por $y_{k+1} = y_k^2$ e

$$x_{k+1} = \begin{cases} y_{k+1} & k \text{ ímpar} \\ y_{k+2} & k \text{ par.} \end{cases}$$

A sequência (x_k, y_k) converge quadraticamente para $(0, 0)$, enquanto que a velocidade de convergência da sequência (x_k) para zero não é linear.

Primeiramente, vamos mostrar que $(x_k, y_k) \rightarrow (0, 0)$, quadraticamente.

Para k par, temos:

$$\frac{\|(x_{k+1}, y_{k+1}) - (0, 0)\|_\infty}{\|(x_k, y_k) - (0, 0)\|_\infty^2} = \frac{\|(y_{k+2}, y_{k+1})\|_\infty}{\|(x_k, y_k)\|_\infty^2} = \frac{y_{k+1}}{y_k^2} = \frac{y_{k+1}}{y_{k+1}} = 1,$$

por outro lado, para k ímpar:

$$\frac{\|(x_{k+1}, y_{k+1}) - (0, 0)\|_\infty}{\|(x_k, y_k) - (0, 0)\|_\infty^2} = \frac{\|(y_{k+1}, y_{k+1})\|_\infty}{\|(x_k, y_k)\|_\infty^2} = \frac{y_{k+1}}{y_k^2} = \frac{y_{k+1}}{y_{k+1}} = 1.$$

Como convergência quadrática independe da norma, acabamos de mostrar que a velocidade de convergência da sequência (x_k, y_k) para $(0, 0)$ é quadrática.

Para a sequência (x_k) , quando k é ímpar, temos:

$$\frac{\|x_{k+1} - 0\|}{\|x_k - 0\|} = \frac{\|y_{k+1}\|}{\|x_{(k-1)+1}\|} = \frac{\|y_{k+1}\|}{\|y_{(k-1)+2}\|} = 1.$$

Portanto, a velocidade de convergência da sequência (x_k) para 0 não é sequer linear.

Existe uma versão do método de Newton para resolver (PE), que garante convergência quadrática da sequência (x^k) . Nesta versão, segundo Bonnans et al [3], calcula-se λ_k como função de x^k , isto é,

$$\lambda_k = \lambda(x^k).$$

Podemos escolher a função $\lambda(\cdot)$, de tal modo que a convergência da sequência (x^k)

seja quadrática. Um possível candidato para tal função é :

$$\lambda(x) = -A^-(x)^T \nabla f(x)$$

onde A^- é inversa à direita de $A(x)$.

Assim, refaz-se o Algoritmo 3.2, com λ_k calculado como função de x^k e prova-se [3, pág. 183], sob as mesmas hipóteses do Teorema 3.3 e com a função $\lambda(\cdot)$ contínua, que a sequência (x^k) , gerada por esta nova versão do método de Newton, converge quadraticamente para x^* .

3.1.2 Convergência global

Foi mostrado que o sistema de equações lineares, em (3.9), que surge da resolução do problema de programação quadrática, com $\nabla_{xx}^2 L(x, \lambda)$ em vez de $\nabla^2 f_o(x)$, é equivalente ao sistema (3.4), que resulta da aplicação do método de Newton à resolução do sistema de equações não lineares - condições necessárias de otimalidade de primeira ordem do problema (PE).

Conforme vimos na subseção anterior, o método de programação quadrática sequencial tem convergência local quadrática, tal como se tem quando se aplica o método de Newton à resolução de um sistema de equações não lineares.

Sabe-se que o método de Newton (para a resolução de um sistema não linear), na sua versão básica, não é robusto e não é possível garantir convergência do processo iterativo a partir de qualquer ponto inicial.

Pela equivalência entre os dois métodos, o mesmo acontece com o método de programação quadrática sequencial.

É possível, no entanto, garantir *convergência global*, introduzindo um esquema, que com a ajuda de uma *função de mérito*, possa medir e forçar o progresso do algoritmo em direção ao minimizador. A idéia desse esquema é semelhante à usada para globalizar o método de Newton, para resolver um problema irrestrito, que usa região de confiança.

Assim, o método PQS com região de confiança, adiciona uma restrição ao subproblema quadrático, que é resolvido em cada iteração, definindo uma região em que o modelo quadrático da função Lagrangeana assume valores próximos aos da função objetivo. Podem ocorrer complicações, porque a inclusão da restrição região de confiança pode tornar o subproblema inconsistente.

Em algumas iterações é necessário deslocar as restrições, ocorrendo complicações no algoritmo e acréscimo do custo computacional. O método escolhido, para resolver o subproblema com região de confiança, tem um grande impacto na eficiência e robustez do método PQS, particularmente para problemas de grande porte.

Como vimos, no Capítulo 2, o método região de confiança para o problema irrestrito aceita o passo p^k , somente se a condição de decréscimo suficiente é satisfeita, isto é, se a redução verdadeira (*ared*) é tão boa quanto a redução predita (*pred*). Porém, para o problema geral, essa comparação (*ared/pred*) deve de alguma forma levar em consideração as restrições, as quais não existiam no problema irrestrito. Isto é feito com a ajuda de uma *função de mérito*, por exemplo, $\phi_\mu : \mathbb{R}^n \rightarrow \mathbb{R}$ dada por

$$\phi_\mu(x) = f_o(x) + \frac{1}{\mu} \|f(x)\|_2,$$

onde μ é um peso positivo, que irá forçar o progresso do algoritmo em direção ao minimizador. Em otimização irrestrita, a função de mérito é simplesmente a função objetivo f_o , que está fixada durante o processo de minimização.

O método PQS pode usar qualquer função de mérito, mas uma questão importante é a escolha do parâmetro μ , que deve ser ajustado em algumas iterações, para garantir que a direção obtida no subproblema é necessariamente uma direção de descida para esta função. As regras para adaptar este parâmetro requerem muito cuidado, pois ele tem uma forte influência no desempenho do método PQS.

PQS com região de confiança.

Adicionando a restrição região de confiança ao subproblema quadrático (3.7), obtemos

$$\begin{aligned} \text{minimizar} \quad & \frac{1}{2} p^T W_k p + \nabla f_{o_k}^T p \\ \text{sujeito a:} \quad & A_k p + f_k = 0 \\ & \|p\| \leq \Delta_k \end{aligned} \tag{3.10}$$

A inclusão da restrição de região de confiança torna o subproblema (3.10), consideravelmente, mais difícil de resolver que o problema quadrático (3.7). Além disso, o novo modelo pode nem sempre ter solução, porque as restrições do subproblema (3.10) podem não ter nenhum ponto em comum, como mostra a Figura 3.1.

Para resolver este conflito entre satisfazer a restrição linear e a restrição de bola em (3.10), não basta simplesmente aumentar o raio da região de confiança, pois

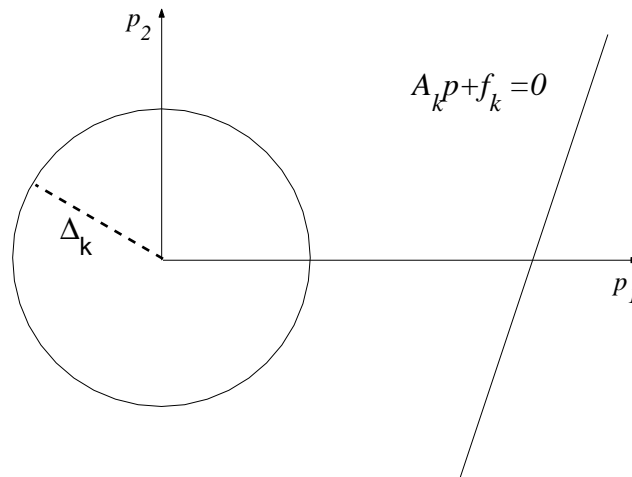


Figura 3.1: Inconsistência entre a linearização das restrições e a região de confiança.

não tem sentido considerar os pontos em que o modelo assume valores distantes ao da função objetivo f_o original, além de prejudicar o progresso do algoritmo.

Dentre os métodos que resolvem o problema (3.10), (veja por exemplo [23, pág. 553]), citamos o método de Byrd e Omojokun [6, 24], que encontra a solução p^k do subproblema (3.10) em duas fases: cada iteração deste método trabalha em uma região centrada na iterada corrente x^k , e é composta de um passo *normal* (passo de viabilidade), seguido de um passo *tangencial* (passo de otimalidade). O passo tangencial deve seguir uma direção no espaço nulo da Jacobiana das restrições em x^k .

Na primeira fase, ignoramos a função objetivo de (3.10), e procuramos satisfazer a linearização das restrições, permanecendo dentro da bola. É preciso então resolver o seguinte subproblema “normal”:

$$\begin{aligned} & \text{minimizar} \quad \|A_k v + f_k\|_2 \\ & \text{sujeito a :} \quad \|v\|_2 \leq \xi \Delta_k \end{aligned} \quad (3.11)$$

onde ξ pode ser qualquer valor no intervalo $(0, 1)$. Uma sugestão é $\xi = 0,8$.

Como para resolver este subproblema (3.11), basta minimizar uma quadrática na bola, podemos usar os algoritmos de *Steihaug* ou *Dogleg*, que podem ser encontrados, por exemplo, em [23].

A solução v^k deste subproblema é o “passo normal”.

Retornamos com a função objetivo do subproblema (3.10) e exigimos que o “passo total” p^k reduza seu valor, sem deixar de satisfazer a linearização das restrições, assim como o passo normal v^k , resultando o novo subproblema:

$$\begin{aligned} \text{minimizar} \quad & \frac{1}{2}p^T W_k p + \nabla f_{o_k}^T p \\ \text{sujeito a :} \quad & A_k p = A_k v^k \\ & \|p\|_2 \leq \Delta_k \end{aligned} \quad (3.12)$$

Agora, é claro que as restrições do subproblema (3.12) são consistentes, pois o ponto $p = v^k$ satisfaz ambas.

Este novo subproblema (3.12) difere do subproblema do método região de confiança (Seção 2.4), por causa da restrição de igualdade $A_k p = A_k v^k$.

Desejamos então eliminar esta restrição de igualdade, para obter um subproblema do tipo (3.11), que consiste em minimizar uma quadrática na bola e, para o qual conhecemos vários métodos de resolução, como os que já foram citados para encontrar o passo normal. Segundo o método de Byrd e Omojokum, conseguimos isto fazendo $p^k = v^k + t$, mas como $A_k v^k + f_k = 0$ então $A_k p^k + f_k = A_k v^k + f_k$ donde,

$$\begin{aligned} A_k p^k &= A_k v^k \\ A_k(v^k + t) &= A_k v^k \\ A_k t &= 0 \Rightarrow t \in N(A_k) \Rightarrow t = Z_k u_k, \end{aligned}$$

onde as colunas da matriz Z_k formam uma base para o espaço nulo de A_k e $u_k \in \mathbb{R}^{n-m}$. Assim, $p^k = v^k + Z_k u_k$.

Substituindo $p^k = v^k + Z_k u$ na função objetivo do subproblema (3.12), obtemos o seguinte subproblema “tangencial” na variável u :

$$\begin{aligned} \text{minimizar} \quad & \frac{1}{2}(v^k + Z_k u)^T W_k (v^k + Z_k u) + \nabla f_{o_k}^T (v^k + Z_k u) \\ \text{sujeito a :} \quad & A_k v^k + A_k Z_k u = A_k v^k \\ & \|v^k + Z_k u\|_2 \leq \Delta_k \end{aligned} \quad (3.13)$$

Reescrevemos a função objetivo do subproblema (3.13), desprezando os termos constantes que são independentes de u , pois eles não alteram a solução. Assim,

$$\frac{1}{2}(v^k + Z_k u)^T W_k (v^k + Z_k u) + \nabla f_{o_k}^T (v^k + Z_k u) =$$

$$\begin{aligned}
&= \frac{1}{2}v^{kT}W_kv^k + \frac{1}{2}(Z_ku)^TW_kv^k + \frac{1}{2}v^{kT}W_k(Z_ku) + \frac{1}{2}(Z_ku)^TW_kZ_ku + \nabla f_{o_k}^T v^k + \nabla f_{o_k}^T Z_ku = \\
&= v^{kT}W_k(Z_ku) + \frac{1}{2}(Z_ku)^TW_kZ_ku + \nabla f_{o_k}^T Z_ku = \\
&= \frac{1}{2}u^T(Z_k^TW_kZ_k)u + (\nabla f_{o_k} + W_kv^k)^T Z_ku,
\end{aligned}$$

que é uma função quadrática.

Como $Z_ku \in N(A_k)$ e $v^k \in I(A_k^T)$, temos que $(Z_ku)^T v^k = 0$, isto é, Z_ku e v^k são ortogonais.

Elevando ao quadrado ambos os membros da restrição de desigualdade em (3.13), obtemos:

$$\begin{aligned}
\|v^k + Z_ku\|_2^2 &\leq \Delta_k^2 \\
(v^k + Z_ku)^T(v^k + Z_ku) &\leq \Delta_k^2 \\
v^{kT}v^k + u^T Z_k^T v^k + v^{kT}Z_ku + (Z_ku)^T Z_ku &\leq \Delta_k^2,
\end{aligned}$$

como Z_ku e v^k são ortogonais,

$$\begin{aligned}
\|v^k\|_2^2 + \|Z_ku\|_2^2 &\leq \Delta_k^2 \\
\|Z_ku\|_2 &\leq \sqrt{\Delta_k^2 - \|v^k\|_2^2}.
\end{aligned}$$

Com isto, o subproblema “tangencial” reduz-se a minimizar uma quadrática com restrição de uma bola, e pode ser escrito como:

$$\begin{aligned}
&\text{minimizar} \quad \frac{1}{2}u^T(Z_k^TW_kZ_k)u + (\nabla f_{o_k} + W_kv^k)^T Z_ku \\
&\text{sujeito a:} \quad \|Z_ku\|_2 \leq \sqrt{\Delta_k^2 - \|v^k\|_2^2}
\end{aligned} \tag{3.14}$$

A solução u^k deste subproblema é o “passo tangencial”.

Dentre os métodos que resolvem o subproblema (3.14), como os citados para resolver o subproblema (3.11), devemos lembrar que só podemos usar o método *Dog-leg* quando temos garantia de que a matriz $(Z_k^TW_kZ_k)$ é positiva definida, enquanto que o método *Steihaug* pode sempre ser aplicado para resolver o subproblema (3.14).

A Figura 3.2 ilustra os passos “normal” e “tangencial” para um problema simples com duas variáveis e uma restrição de igualdade não linear. As curvas tracejadas representam as curvas de nível da função objetivo, cujo minimizador é x^* . A restrição de igualdade está representada pela linha cheia, e a circunferência tracejada centrada em x^k representa a restrição região de confiança. O espaço tangente das

restrições é mostrado pela linha pontilhada passando por x^k (Z_k tem apenas uma coluna neste exemplo), e o passo normal é perpendicular a este espaço. Uma cópia do espaço tangente, passando por $x^k + v^k$, define um conjunto, no qual deve estar o passo tangencial $Z_k u^k$. Assim, o passo total $p^k = v^k + Z_k u^k$ satisfaz a restrição região de confiança.

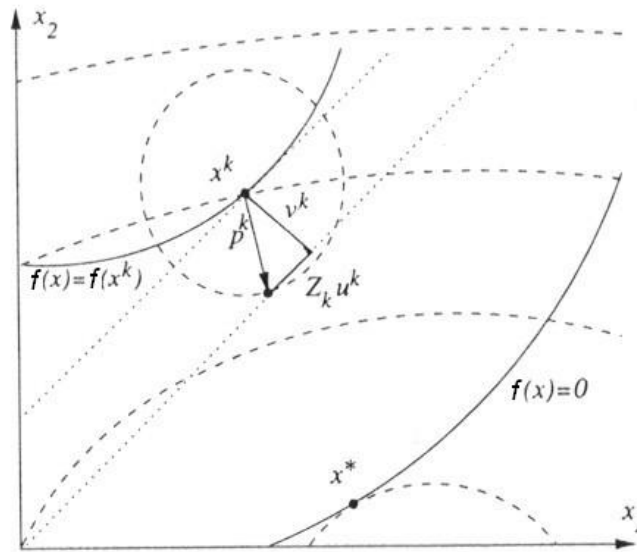


Figura 3.2: O passo $p^k = v^k + Z_k u^k$ do método região de confiança.

Função de mérito.

Uma vez calculado o passo normal e o passo tangencial, temos o passo total

$$p^k = v^k + Z_k u^k,$$

e o próximo ponto gerado pelo algoritmo, a partir de x^k , será da forma $x^{k+1} = x^k + p^k$.

Precisamos, então, de alguma estratégia para aceitar ou não esse novo ponto, a fim de garantir convergência do algoritmo para a solução do problema (PE).

Introduzimos um esquema que, com a ajuda de uma função de mérito, possa medir e forçar o progresso do algoritmo em direção ao minimizador.

A função de mérito escolhida é:

$$\phi_\mu(x) = f_o(x) + \frac{1}{\mu} \|f(x)\|_2. \quad (3.15)$$

Outras funções de mérito podem ser usadas, veja por exemplo [23, Cap. 15].

Se o passo p^k produz um decréscimo suficiente na função de mérito, então, o passo é aceito. Caso contrário, o raio da região de confiança é reduzido e o passo recalculado, repetindo-se todo o processo. Em geral, a direção do passo muda quando o tamanho da região é alterado.

Dados $\mu > 0$, o ponto corrente x^k e um passo p^k , definimos a redução **verdadeira** como:

$$ared = \phi_\mu(x^k) - \phi_\mu(x^k + p^k).$$

Observe que, para o problema geral, a redução verdadeira avalia a variação da função de mérito, que para o problema irrestrito é a própria função objetivo.

Para simplificar a notação denotaremos:

- $\sigma = \mu^{-1}$,
- $g = \nabla f_o(x^k)$,
- $f = f(x^k)$,
- $A = A(x^k)$,
- $W = \nabla^2 L(x^k, \lambda^k)$,
- $q(s) = \frac{1}{2}s^T W s + g^T s$.

Dado x^k , o modelo para a função de mérito (3.15), segundo Conn, Gould, Toint [7, pág. 660], é:

$$m_\phi(p) = f_o(x^k) + g^T p + \frac{1}{2}p^T W p + \sigma \|f + Ap\|.$$

Assim a redução **predita** produzida pelo passo p^k é definida como:

$$pred = m_\phi(0) - m_\phi(p^k) = \Delta m^t + \sigma \Delta m^n + \Delta q^n \quad (3.16)$$

onde

$$\Delta m^n = \|f\| - \|f + Av^k\|$$

é a redução no modelo (3.11), obtida pelo passo normal v^k ,

$$\Delta m^t = -\left[\frac{1}{2}t^{kT} W t^k + (g + W v^k)^T t^k\right]$$

é a redução no modelo (3.14) obtida pelo passo tangencial $t^k = Z_k u^k$, e

$$\Delta q^n = \frac{1}{2} v^{kT} W v^k + g^T v^k$$

é a variação no modelo quadrático do Lagrangeano, obtido pelo passo normal v^k .

Os dois primeiros componentes do lado direito de (3.16) são positivos, pois são obtidos minimizando-se a norma da aproximação linear das restrições na região de confiança e a aproximação quadrática do Lagrangeano no espaço nulo da Jacobiana das restrições e, sujeito ainda à região de confiança, respectivamente. Mas o terceiro termo em (3.16) pode ser positivo ou negativo. O passo normal v^k pode produzir um aumento no valor da função objetivo do subproblema (3.10), o qual nem sempre será compensado pela redução dada pelo passo tangencial.

Para garantir redução no modelo (3.16,) escolhemos σ suficientemente grande para que $pred$ seja positivo:

$$pred = \Delta m^t + \sigma \Delta m^n + \Delta q^n \geq 0.$$

Para produzir um decréscimo significativo no modelo (3.16), exigimos:

$$pred = \Delta m^t + \sigma \Delta m^n + \Delta q^n \geq \nu \sigma \Delta m^n$$

para algum $\nu \in (0, 1)$.

Assim, escolhemos σ , tal que:

$$\sigma \geq \frac{-(\Delta m^t + \Delta q^n)}{(1 - \nu) \Delta m^n}$$

Agora, estamos aptos para escrever o algoritmo de PQS com região de confiança, para resolver o problema (PE).

Algoritmo 3.4 PQS global

Escolha as constantes $\epsilon > 0$ e $\eta, \xi, \gamma \in (0, 1)$;

Escolha um ponto inicial x^0 e um raio inicial da região de confiança $\Delta_0 > 0$;

$k = 0$

REPITA

Calcule $f_{o_k}, f_k, \nabla f_{o_k}, A_k$;

Calcule estimativas do multiplicador $\hat{\lambda}^k$ por $\hat{\lambda}^k = (A_k A_k^T)^{-1} A_k \nabla f_{o_k}$;

se $\|\nabla f_{o_k} - A_k^T \hat{\lambda}^k\|_\infty < \epsilon$ e $\|f_k\|_\infty < \epsilon$

Pare com solução aproximada x^k .

Resolva o subproblema normal (3.11) para v^k ;

Calcule a matriz Z_k , cujas colunas formam uma base para o espaço nulo de A_k ;

Calcule ou atualize W_k ;

Resolva o subproblema tangencial (3.14) para u^k ;

Faça $p^k = v^k + Z_k u^k$;

Calcule $\rho_k = \frac{\text{ared}}{\text{pred}}$

se $\rho_k > \eta$

Faça $x^{k+1} = x^k + p^k$;

Escolha Δ_{k+1} tal que $\Delta_{k+1} \geq \Delta_k$

senão

Faça $x^{k+1} = x^k$

Escolha Δ_{k+1} tal que $\Delta_{k+1} \leq \gamma \|p^k\|$;

$k = k + 1$

FIM

A constante ξ , que define um raio menor para a região de confiança no subproblema (3.11), é comumente escolhida igual a 0,8.

A prova da convergência global do Algoritmo 3.4 pode ser encontrada, por exemplo, em Conn, Gould, Toint [7].

Efeito Maratos.

Infelizmente, o método PQS com função de mérito pode rejeitar passos de Newton a partir de pontos arbitrariamente próximos à solução, tornando lenta a convergência do algoritmo.

Este indesejável fenômeno é chamado *efeito Maratos*, pois foi primeiramente observado por Maratos [20].

Usaremos o exemplo ([23, pág. 567]), de Powell [27], para ilustrar este fato:

$$\begin{aligned} \text{minimizar} \quad & f_o(x_1, x_2) = 2(x_1^2 + x_2^2 - 1) - x_1 \\ \text{sujeito a:} \quad & x_1^2 + x_2^2 - 1 = 0 \end{aligned} \quad (3.17)$$

A solução deste problema é $x^* = (1, 0)^T$, com multiplicador de Lagrange correspondente $\lambda^* = \frac{3}{2}$ e $\nabla_{xx}^2 l(x^*, \lambda^*) = I$. A função Lagrangeana (3.2) para este problema é

$$L(x, \lambda^*) = (2 - \lambda^*)(x_1^2 + x_2^2 - 1) - x_1.$$

Como $\lambda^* = \frac{3}{2}$, temos:

$$L(x, \frac{3}{2}) = \frac{1}{2}(x_1^2 + x_2^2 - 1) - x_1.$$

A Figura 3.3 mostra a solução x^* , as circunferências tracejadas representam as curvas de nível da função objetivo, a circunferência em linha cheia representa a restrição de igualdade e, uma iteração x^1 a partir de x^0 .

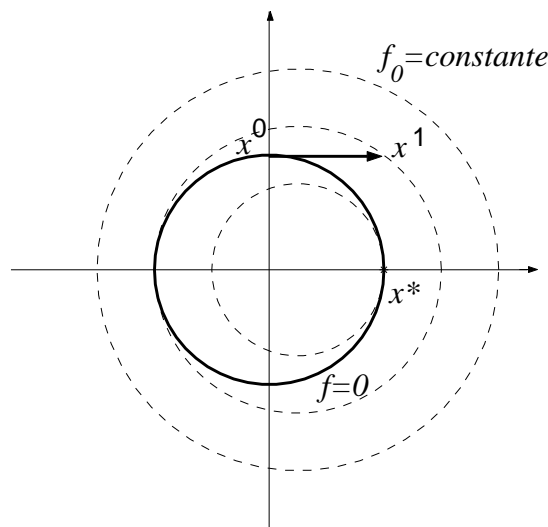


Figura 3.3: Efeito *Maratos*: Problema (3.17).

Consideremos x^k da forma $x^k = (\cos \theta, \sin \theta)^T$, que é possível para qualquer valor de θ . Para encontrar o passo p^k , resolvemos o subproblema (3.7) com $W_k =$

$\nabla_{xx}^2 l(x^*, \lambda^*) = I$. Assim,

$$f_o(x^k) = -\cos \theta, \quad \nabla f_o(x^k) = \begin{pmatrix} 4 \cos \theta - 1 \\ 4 \sin \theta \end{pmatrix}, \quad A^T(x^k) = \begin{pmatrix} 2 \cos \theta \\ 2 \sin \theta \end{pmatrix},$$

o subproblema quadrático (3.7) tem a forma:

$$\begin{aligned} \text{minimizar} \quad & -\cos \theta + (4 \cos \theta - 1)p_1 + 4 \sin \theta p_2 + \frac{1}{2}p_1^2 + \frac{1}{2}p_2^2 \\ \text{sujeito a :} \quad & p_2 = -\cot \theta p_1. \end{aligned}$$

Resolvendo este subproblema, obtemos:

$$p^k = \begin{pmatrix} \sin^2 \theta \\ -\sin \theta \cos \theta \end{pmatrix}, \quad (3.18)$$

que produz o novo ponto

$$x^k + p^k = \begin{pmatrix} \cos \theta + \sin^2 \theta \\ \sin \theta(1 - \cos \theta) \end{pmatrix}.$$

Se $\sin \theta \neq 0$, temos

$$\|x^k + p^k - x^*\|_2 = 2 \sin^2(\theta/2), \quad \|x^k - x^*\|_2 = 2 |\sin(\theta/2)|,$$

e a razão

$$\frac{\|x^k + p^k - x^*\|_2}{\|x^k - x^*\|_2^2} = \frac{1}{2}.$$

Assim, este passo aproxima-se da solução com uma taxa de convergência quadrática.

Porém,

$$\begin{aligned} f_o(x^k + p^k) &= \sin^2 \theta - \cos \theta > -\cos \theta = f_o(x^k) \quad \text{e} \\ f(x^k + p^k) &= \sin^2 \theta > c(x^k) = 0, \end{aligned}$$

ambos os valores, da função objetivo e da restrição, aumentaram no ponto $x^k + p^k$, que será rejeitado pelo algoritmo.

Este exemplo mostra que qualquer função de mérito da forma

$$\phi_\mu(x) = f_o(x) + \frac{1}{\mu} h(f(x))$$

(em (3.15) tomamos $h(\cdot) = \|\cdot\|_2$), rejeitaria o passo (3.18), e qualquer algoritmo, baseado em função de mérito deste tipo, sofre do efeito Maratos, que pode levar a passos curtos em pontos arbitrariamente próximos à solução ótima.

Técnicas para evitar o efeito Maratos podem ser encontradas em [23, pág. 570], como por exemplo:

- Correção de segunda ordem ;
- Estratégia Watchdog.

3.2 Método de Filtro

Vimos, na seção anterior, que para garantir a convergência global do método de programação quadrática sequencial (PQS), é necessário usar algum critério de aceitação do passo.

O critério usado foi uma função de mérito. As funções de mérito são, de modo geral, uma combinação linear de uma medida de otimalidade e uma de inviabilidade. Outro critério possível é um filtro.

Algoritmos de filtro definem uma *região proibida*, memorizando a cada iteração, pares ordenados, onde as componentes representam a função objetivo e a inviabilidade. Os pontos cujos valores da função objetivo e da inviabilidade forem maiores que os dos pares ordenados memorizados, ou seja, que são dominados por esses pares ordenados, fazem parte da região proibida. Cada iteração gera um ponto não pertencente à região proibida e atualiza a lista de pares ordenados, que formam o filtro. Os algoritmos de filtro foram propostos por Fletcher e Leyffer [9] e são baseados em programação quadrática seqüencial (PQS), usando o conceito de dominância como critério para aceitação do ponto obtido a cada iteração. Uma prova de convergência global de métodos de filtro foi apresentada por Fletcher, Gould, Leyffer, Toint e Wächter [8].

Por outro lado, Martínez e Pilotta [21] propõem um algoritmo baseado em PQS com restauração inexata. É usada uma função de mérito para avaliar, a cada iteração, se o ponto obtido é aceito ou não. Cada iteração é decomposta em duas fases: a primeira, dita fase de restauração, procura reduzir a inviabilidade, e a segunda procura melhorar a otimalidade.

Gonzaga, Karas e Vanti [13] propõem um algoritmo de filtro (GKV) para programação não linear de restauração inexata, no sentido de Martínez e Pilotta,

que usa filtro em vez de uma função de mérito. Cada iteração é composta de uma fase de restauração e uma fase de otimalidade. Essas duas fases são totalmente independentes, e o único acoplamento entre elas é estabelecido pelo filtro.

Assim, ao contrário do método de filtro de Fletcher e Leyffer, no método de filtro GKV não há nenhuma dependência entre as fases de viabilidade e otimalidade, que podem ser baseadas em qualquer algoritmo que satisfaça algumas hipóteses razoáveis. A única conexão entre ambas as fases é que elas não permitem gerar pontos proibidos. Devemos ainda notar que o método de filtro GKV é um método com memória, e sempre que um novo par é introduzido no filtro, podemos eliminar todos os outros pares dominados por este novo par, cujas componentes representam o valor da função objetivo e o valor da medida de inviabilidade no ponto corrente.

Consideramos uma medida de inviabilidade das restrições $x \in \mathbb{R}^n \mapsto h(x)$, que é uma penalidade exata aplicada às restrições. Usaremos a medida de inviabilidade dada por

$$h(x) = \|f(x)\|, \quad (3.19)$$

onde $\|\cdot\|$ denota uma norma arbitrária.

Algoritmo de Filtro GKV.

No algoritmo de filtro GKV, cada iteração é constituída de uma fase de viabilidade e uma de otimalidade.

Numa dada iteração, o filtro armazena um par ordenado temporário formado pelo valor da função objetivo e pela medida de inviabilidade do ponto obtido na iteração anterior.

Na fase de viabilidade, é obtido um ponto, não proibido pelo filtro, cuja medida de inviabilidade é menor que a do ponto da iteração anterior. Caso o ponto corrente seja viável, essa fase é desnecessária.

A seguir é feita a fase de otimalidade, em que é obtido outro ponto, não proibido pelo filtro, no qual o valor da função objetivo é menor que o valor do ponto obtido na fase de viabilidade. Esse novo ponto deve ser obtido de forma a não perder demais a viabilidade adquirida na fase anterior. O ponto obtido na fase de otimalidade é o ponto a ser usado na próxima iteração.

Se o valor da função objetivo no novo ponto for maior ou igual que no ponto corrente, o par ordenado temporário presente no filtro torna-se permanente; caso contrário, é eliminado do filtro.

O algoritmo pára com sucesso quando é obtido um ponto viável e estacionário. No caso em que, na fase de viabilidade, não se consegue obter um ponto mais viável que o da iteração anterior, o algoritmo pára com insucesso.

Apresentamos o algoritmo de filtro GKV de forma compacta, sem qualquer especificação dos algoritmos internos usados nos passos de viabilidade e otimalidade (para detalhes, ver [13]).

Algoritmo 3.5 *Algoritmo de filtro - GKV*

Dados: $x^0 \in \mathbb{R}^n$, $F_0 = \emptyset$, $\mathcal{F}_0 = \emptyset$, $\alpha \in (0, 1)$.

$k = 0$

REPITA

$$(\tilde{f}_o, \tilde{h}) = (f_o(x^k) - \alpha h(x^k), (1 - \alpha)h(x^k)).$$

Construa o conjunto $\bar{F}_k = F_k \cup \{(\tilde{f}_o, \tilde{h})\}$.

Defina o conjunto $\bar{\mathcal{F}}_k = \mathcal{F}_k \cup \{x \in \mathbb{R}^n \mid f_o(x) \geq \tilde{f}_o, h(x) \geq \tilde{h}\}$.

Fase de viabilidade:

se $h(x^k) = 0$, então faça $z^k = x^k$

senão, calcule $z^k \notin \bar{\mathcal{F}}_k$ tal que $h(z^k) < (1 - \alpha)h(x^k)$.

se impossível, então pare com insucesso.

Fase de otimalidade:

se z^k é estacionário, então pare com sucesso

senão, calcule $x^{k+1} \notin \bar{\mathcal{F}}_k$ tal que $x^{k+1} \in \mathcal{N}(A(z^k))$ e $f_o(x^{k+1}) \leq f_o(z^k)$.

Atualização do filtro:

se $f_o(x^{k+1}) < f_o(x^k)$, então

$$F_{k+1} = F_k, \mathcal{F}_{k+1} = \mathcal{F}_k \quad (\text{iteração - } f_o)$$

senão,

$$F_{k+1} = \bar{F}_k, \mathcal{F}_{k+1} = \bar{\mathcal{F}}_k \quad (\text{iteração - } h)$$

$k = k + 1$.

Um dos resultados inéditos apresentado por Karas [14], em sua tese de doutorado, é a prova de convergência global do algoritmo de filtro GKV, que depende dos algoritmos internos usados em cada iteração, desde que esses algoritmos satisfaçam certas hipóteses sobre sua eficiência.

Assim como o método PQS com função de mérito, o método de filtro GKV pode sofrer do efeito Maratos, quando aplicado ao exemplo de Powell.

Atualmente, estuda-se a incorporação de passos de segunda ordem ao algoritmo de filtro GKV para evitar o efeito Maratos. Com isto, acredita-se ser possível analisar as propriedades de convergência local do algoritmo de filtro GKV com passos de segunda ordem.

Conclusão

Neste trabalho reunimos os resultados, já provados, sobre convergência dos métodos clássicos de programação não linear para os problemas irrestrito e com restrições de igualdade.

Para analisar a convergência local dos algoritmos foi necessário, primeiramente, estudar velocidade de convergência. Estudamos três diferentes definições de velocidade de convergência estabelecendo, através das nossas demonstrações, suas implicações e equivalências. Vários exemplos foram elaborados, por nós, para evidenciar quando as equivalências entre as definições não são válidas.

No estudo do método de Cauchy, mostramos um resultado que não aparece na literatura, apesar deste ser um dos métodos mais antigos para otimização irrestrita. O resultado é sobre a convergência local da sequência gerada pelo algoritmo de Cauchy com busca linear exata, quando aplicado a uma função quadrática e convexa. Na literatura temos a convergência *q-linear* na norma induzida pela hessiana. Entretanto, apesar das normas em \mathbb{R}^n serem equivalentes, a velocidade de convergência *q-linear* depende da norma. Mostramos neste trabalho que essa sequência converge com velocidade *q-linear*, na norma euclidiana (ao invés da norma induzida pela hessiana), para o único minimizador da função, e exibimos a taxa de convergência, que está relacionada com os autovalores da matriz da quadrática. Quanto melhor condicionada for a matriz da função quadrática, mais rápida será a convergência. Ainda temos que, se a função é quadrática, então o comprimento do passo calculado por Armijo com $\alpha = \frac{1}{2}$, coincide com o passo da busca exata. Isto implica que a sequência gerada pelo algoritmo de Cauchy com busca Armijo também convergirá com velocidade *q-linear* e com a mesma taxa da sequência com busca exata.

No estudo do método de Newton, para o problema irrestrito, vimos que o algoritmo “puro” é muito eficiente, possui convergência quadrática, mas somente quando o ponto inicial está suficientemente próximo da solução. A convergência

global do método de Newton é garantida quando introduzimos no algoritmo a busca unidirecional de Armijo.

Demonstramos que o método de região de confiança é globalmente convergente, adaptando duas provas dadas por Nocedal e Wright, onde incorporamos certas correções e detalhamos algumas passagens. Além disso, indicamos um resultado que garante convergência local quadrática, quando a região de confiança torna-se inativa.

Para o problema de programação não linear com restrições de igualdade, mostramos que o método de programação quadrática sequencial (PQS) “puro” converge quadraticamente para a solução, desde que o ponto inicial esteja suficientemente próximo da solução. No entanto, para globalizar o método PQS, introduzimos ao subproblema a restrição de região de confiança e usamos a técnica de Byrd e Omojokum, que decompõe o passo em duas componentes: normal e tangencial. Porém, se uma correção de segunda ordem não for aplicada, o algoritmo PQS com região de confiança pode sofrer do efeito *Maratos*, fazendo com que o algoritmo tenha propriedades de convergência local ruins.

Com o estudo do método de filtro, globalmente convergente, proposto por Gonzaga, Karas e Vanti, encerramos a discussão dos métodos para programação não linear. Assim como o método PQS com região de confiança, também o método de filtro pode sofrer do efeito *Maratos*, se uma correção de segunda ordem não for implementada. Sugere-se, então, a incorporação de correções de segunda ordem ao algoritmo de filtro, para eliminar o efeito *Maratos*, o qual ocorre quando o algoritmo rejeita passos muito bons, ocasionando numa convergência local lenta. Assim, acredita-se ser possível estudar as propriedades de convergência local do algoritmo de filtro com correção de segunda ordem.

Apêndice

O teorema, a seguir, mostra que usando a mudança de variável $y = x - x^*$ e subtraindo a constante $f(x^*)$ de $f(x)$, podemos assumir sem perda de generalidade que a função f tem um minimizador $x^* = 0$ e que $f(x^*) = 0$.

Teorema 3.6 *Sejam $g(x) = f(x) - c$, $h(x) = g(x + x^*)$ onde $f : \mathbb{R}^n \rightarrow \mathbb{R}$, então:*

- 1. x^* é minimizador local de f se, e somente se, x^* é minimizador local de g ;*
- 2. x^* é minimizador local de g se, e somente se, 0 é minimizador local de h .*

Prova.

1. Se x^* é minimizador local de f , então existe uma vizinhança V de x^* , tal que $f(x^*) \leq f(x)$ para todo $x \in V$. Assim, temos:

$$g(x^*) + c = f(x^*) \leq f(x) = g(x) + c \Leftrightarrow g(x^*) + c \leq g(x) + c \Leftrightarrow g(x^*) \leq g(x),$$

para todo x pertencente a uma vizinhança V de x^* . Portanto x^* é minimizador local de g .

Por outro lado, se x^* é minimizador local de g , então existe uma vizinhança V de x^* , tal que $g(x^*) \leq g(x)$, para todo $x \in V$. Assim

$$g(x^*) = f(x^*) - c \leq f(x) - c = g(x) \Leftrightarrow f(x^*) - c \leq f(x) - c \Leftrightarrow f(x^*) \leq f(x)$$

para todo x pertencente a uma vizinhança V de x^* . Portanto x^* é minimizador local de f .

2. Se x^* é minimizador local de g , então existe uma vizinhança $V_{x^*} = \{x \mid \|x - x^*\| < \epsilon\}$, tal que $g(x^*) \leq g(x)$, para todo $x \in V_{x^*}$. Seja $V_0 = \{y \mid \|y\| < \epsilon\}$,

ou seja, $y = x - x^*$, daí:

$$h(0) = g(x^*) \leq g(x) = g(y + x^*) = h(y) \Leftrightarrow h(0) \leq h(y),$$

para todo $y \in V_0$. Portanto, 0 é minimizador local de h .

Por outro lado, se 0 é minimizador local de h , então existe uma vizinhança $V_0 = \{y \mid \|y\| < \epsilon\}$, tal que $h(0) \leq h(y)$, para todo $y \in V_0$.

Seja $V_{x^*} = \{x \mid \|x - x^*\| < \epsilon\}$, assim,

$$g(x^*) = h(0) \leq h(y) = h(x - x^*) = g(x - x^* + x^*) = g(x) \Leftrightarrow g(x^*) \leq g(x),$$

para todo $x \in V_{x^*}$. Portanto, x^* é minimizador local de g .

□

Referências Bibliográficas

- [1] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: theory and algorithms*. John Wiley and Sons, New York, 1993.
- [2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Massachusetts, 1995.
- [3] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical Optimization*. Springer-Verlag, Berlin, Heidelberg, 2003.
- [4] R. Burachik, L. Drummond, A. Iusem, and B. Svaiter. Full convergence of the steepest descent method with inexact line searches. *Optimization*, 32:137–146, 1995.
- [5] R. L. Burden and J. D. Faires. *Numerical analysis*. Pacific Grove, USA, 1997.
- [6] R. H. Byrd. Robust trust region methods for constrained optimization. Third SIAM Conference on Optimization, 1987.
- [7] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. MPS/SIAM Series on Optimization, Philadelphia, 2000.
- [8] R. Fletcher, N. Gould, S. Leyffer, P. Toint, and A. Wächter. Global convergence of trust-region and SQP-filter algorithms for general nonlinear programming. *SIAM J. Optimization*, 13(3):635–659, 2002.
- [9] R. Fletcher and S. Leyffer. Nonlinear programming without a penalty function. *Mathematical Programming - Ser. A*, 91(2):239–269, 2002.
- [10] A. Friedlander. *Elementos de Programação Não-Linear*. Editora da Unicamp, Campinas, SP, Brasil, 1994.

- [11] C. C. Gonzaga. Two facts on the convergence of the Cauchy algorithm. *Journal on Optimization Theory and Applications*, 107(3):591–600, 2000.
- [12] C. C. Gonzaga. Um curso de programação não linear. Notas de aula, Florianópolis, SC, Brasil, 2004.
- [13] C. C. Gonzaga, E. W. Karas, and M. Vanti. A globally convergent filter method for nonlinear programming. *SIAM J. Optimization*, 14(3):646–669, 2003.
- [14] E. W. Karas. *Exemplos de trajetória central mal comportada em otimização convexa e um algoritmo de filtros para programação não linear*. PhD thesis, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil, 2002.
- [15] K. Kiwiel and K. G. Murty. Convergence of the steepest descent method for minimizing quasi convex functions. *Journal of Optimization Theory and Applications*, 89:221–226, 1996.
- [16] S. J. Leon. *Álgebra Linear: com aplicações*. LTC, Rio de Janeiro, 1999.
- [17] E. L. Lima. *Curso de Análise vol. 1*. IMPA, Rio de Janeiro, 1981.
- [18] E. L. Lima. *Curso de Análise vol. 2*. IMPA, Rio de Janeiro, 1981.
- [19] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison - Wesley Publishing Company, New York, 1986.
- [20] N. Maratos. *Exact Penalty Function Algorithms for Finite Dimensional and Control Optimization Problems*. PhD thesis, Imperial College Science Technology, University of London, 1978.
- [21] J. M. Martínez and E. A. Pilotta. Inexact restoration algorithms for constrained optimization. *Journal of Optimization Theory and Applications*, 104:135–163, 2000.
- [22] J. M. Martínez and S. A. Santos. Métodos computacionais de otimização. 20.^o Colóquio Brasileiro de Matemática - IMPA, 1995.
- [23] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, 1999.

- [24] E. Omojokun. *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*. PhD thesis, Dept. of Computer Science, University of Colorado, 1991.
- [25] J. M. Ortega and W. G. Rheinboldt. *Iterative solution of nonlinear equation in several variables*. San Diego. Academic Press, 1970.
- [26] A. L. Peressini, F. E. Sullivan, and Jr. J. J. Uhl. *The mathematics of nonlinear programming*. New York. Springer-Verlag, 1988.
- [27] M. J. D. Powell. Convergence properties of algorithms for nonlinear optimization. *SIAM Review*, 28:487–500, 1986.
- [28] S. S. Rao. *Engineering optimization: theory and practice, 3rd Edition*. New York. Wiley-Interscience, 1996.
- [29] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM J. Numer. Anal.*, 20:626–637, 1983.