MÉTODOS COMPUTACIONAIS DE OTIMIZAÇÃO

José Mario Martínez

Sandra Augusta Santos

Departamento de Matemática Aplicada
IMECC-UNICAMP
1995
Atualizado em dezembro de 1998

ÍNDICE

Chapter 1

Introdução

Otimização é um problema matemático com muitas aplicações no "mundo real". Consiste em encontrar os mínimos ou máximos de uma função de várias variáveis, com valores dentro de uma determinada região do espaço multi-dimensional. Os responsáveis pela tomada de decisões nos mais variados campos da atividade humana defrontam-se, cotidianamente, com esse tipo de necessidade. Às vezes, a índole do problema, a demanda de resultados precisos, ou a própria curiosidade, leva a formalizar variáveis, restrições e objetivos, de maneira que a natureza matemática do problema emerge. Esse é o processo de modelagem, que descobre isomorfismos entre a realidade empírica e o idealismo dos objetos matemáticos. No entanto, a correspondência entre experiência e modelo formal está longe de ser perfeita: a tradução está sujeita a erros, simplificações e falhas de comunicação. Notavelmente, a problemática de adequar um modelo matemático a uma situação real também pode ser formulada como um problema matemático, quase sempre de otimização.

1.1 Uma classificação informal

O problema a ser considerado neste livro é o seguinte:

$$\text{Minimizar } f(x) \text{ sujeita a } x \in \Omega \subset I\!R^n. \tag{1.1.1}$$

A função f é chamada *função objetivo* e o conjunto Ω, freqüentemente definido por um conjunto de igualdades e desigualdades, é o *conjunto factível*. Os pontos de Ω serão os *pontos factíveis* de (1.1.1).

De fato, estamos tão interessados em *minimizar* como em *maximizar* funções, mas falaremos apenas de minimizar dado que, claramente, maximizar $f(x)$ em uma região qualquer do espaço $I\!R^n$ é equivalente a minimizar $-f(x)$ na mesma região. As soluções $x_* \in \Omega$ do problema (1.1.1) serão chamadas *minimizadores* e os valores correspondentes $f(x_*)$ são os *mínimos* do problema. Quase sempre assumiremos a continuidade de f e, com freqüência um pouco menor, a existência de derivadas primeiras contínuas. Às vezes, vamos supor também que f tem derivadas segundas contínuas.

Conforme as características do conjunto Ω, teremos os diferentes problemas de otimização:

Ω	Problema
$I\!R^n$	minimização sem restrições
$\{x \in I\!R^n \mid l \leq x \leq u\}$	minimização em caixas
$\{x \in I\!R^n \mid Ax = b, A \in I\!R^{m \times n}\}$	minimização com restrições lineares de igualdade
$\{x \in I\!R^n \mid Ax = b, Cx \leq d\}$	minimização com restrições lineares
$\{x \in I\!R^n \mid h(x) = 0, h : I\!R^n \to I\!R^m\}$	minimização com restrições de igualdade
$\{x \in I\!R^n \mid h(x) = 0, h : I\!R^n \to R^m$ e $g(x) \leq 0, g : I\!R^n \to I\!R^p\}$	problema geral de programação não linear

Quando v e w são vetores, a notação $v \leq w$ significará sempre $v_i \leq w_i$ para todas suas coordenadas. Assim, quando falamos da "caixa" $l \leq x \leq u$, entendemos o conjunto dos $x \in I\!R^n$ tais que $l_i \leq x_i \leq u_i$ para todo $i = 1, \ldots, n$. O problema geral de programação não linear pode ser reduzido sempre a uma *forma padrão* mediante a introdução de *variáveis de folga*. Com efeito, observamos que o conjunto dos $x \in I\!R^n$ tais que $h(x) = 0$ e $g(x) \leq 0$ coincide com o conjunto

$$\{x \in I\!R^n \quad \mid \quad h(x) = 0 \text{ e } g(x) + z = 0 \text{ para algum } z \geq 0\}.$$

Portanto, o problema

$$\text{Minimizar } f(x) \text{ sujeita a } h(x) = 0, g(x) \leq 0, \tag{1.1.2}$$

onde $h : I\!R^n \to I\!R^m$, $g : I\!R^n \to I\!R^p$, é equivalente a

$$\text{Minimizar } f(x) \text{ sujeita a } h(x) = 0, g(x) + z = 0, z \geq 0. \tag{1.1.3}$$

Agora, mudando os nomes de variáveis e funções, (1.1.3) tem a forma geral

$$\text{Minimizar } f(x) \text{ sujeita a } h(x) = 0, x \geq 0. \tag{1.1.4}$$

A forma (1.1.4) de um problema de programação não linear se denomina *forma padrão*. Quando um problema do tipo (1.1.2) é transformado na sua forma padrão, o número de variáveis é aumentado em p. Às vezes, isso é uma desvantagem. No entanto, a transformação muitas vezes se justifica por considerações algorítmicas, como veremos em capítulos futuros.
Neste livro a ênfase estará colocada em funções objetivo $f(x)$ não lineares. Quando f é linear ($f(x) = c^T x$ para algum $c \in I\!R^n$) o problema de minimização com restrições lineares é chamado de *problema de programação linear*. Na sua forma padrão, este problema é

$$\begin{array}{l} \text{Minimizar } \; c^T x \\ \quad Ax = b \\ \quad x \geq 0 \; . \end{array} \tag{1.1.5}$$

O conteúdo deste livro se aplica a programação linear, embora, pela especificidade deste problema, muito desse conteúdo seja supérfluo. Por outro lado, as particularidades do problema (1.1.5) permitem um tratamento muito mais rico e detalhado, que não será feito aqui. Em menor medida, essa observação vale também no caso em que a função objetivo é quadrática e as restrições lineares, chamado *problema de programação quadrática.*

1.2 Um problema de estimação de parâmetros

Quando o ponto de partida é um problema real, podem existir vários problemas matemáticos de otimização associados, vinculados a diferentes formulações ou a diferentes técnicas de resolução. Nesta seção apresentamos um problema de estimação de parâmetros originado na Ótica, para o qual exibimos algumas formulações sob o ponto de vista da otimização. Ver [189], [33].
Um filme é um material muito fino, cuja espessura, índices de refração e coeficientes de absorção se deseja estimar. Esses parâmetros não são suscetíveis de medição direta, ou seja, devem ser inferidos da medição de outra magnitude física. O experimento que gera a medição indireta consiste, brevemente, no seguinte: coloca-se o material em cima de um substrato transparente e "atravessa-se" filme e substrato com luz de diferentes comprimentos de onda.

Para fixar idéias, esses comprimentos podem ir desde 800 até 2000, com intervalos de 10, nas unidades adequadas. Para cada comprimento de onda λ, mede-se a *transmissão* $T(\lambda) \in [0,1]$, isto é, o quociente, adimensional, entre a luz que atravessa o filme e a luz emitida. Teoricamente, $T(\lambda)$ se relaciona com a espessura (d), o coeficiente de absorção $(\alpha(\lambda))$ e o índice de refração do filme $(n(\lambda))$ através das seguintes fórmulas (por simplicidade, escrevemos $T = T(\lambda)$, $n = n(\lambda)$, $\alpha = \alpha(\lambda)$):

$$T = \frac{A'x}{B' - C'x + D'x^2}, \tag{1.1.6}$$

onde

$$A' = 16s(n^2 + k^2) \tag{1.1.7}$$

$$B' = [(n+1)^2 + k^2][(n+1)(n+s^2) + k^2] \tag{1.1.8}$$

$$\begin{aligned} C' = {} & [(n^2 - 1 + k^2)(n^2 - s^2 + k^2) - 2k^2(s^2+1)]2 \cos \varphi \\ & -k[2(n^2 - s^2 + k^2) + (s^2+1)(n^2 - 1 + k^2)]2 \sin \varphi \end{aligned} \tag{1.1.9}$$

$$D' = [(n-1)^2 + k^2][(n-1)(n-s^2) + k^2] \tag{1.1.10}$$

$$\varphi = 4\pi n d/\lambda, \quad x = exp(-\alpha d), \quad k = \alpha\lambda/(4\pi). \tag{1.1.11}$$

Nas fórmulas (1.1.6)–(1.1.11) s é o índice de refração do substrato, suposto conhecido e constante para todo λ. O experimento físico fornece uma tabela de dados onde a coluna da esquerda são os comprimentos de onda λ_i usados, desde $\lambda_1 = 800$ até $\lambda_m = \lambda_{121} = 2000$, e a coluna da direita está formada pelas medidas correspondentes de transmissão (T_i). As fórmulas (1.1.6)–(1.1.11) definem a função teórica $T(\lambda, d, n, \alpha)$. Portanto, a primeira vista, o objetivo parece ser encontrar d e $n_i, \alpha_i, i = 1, \ldots, m$ tais que, para todo $i = 1, \ldots, m$,

$$T(\lambda_i, d, n_i, \alpha_i) = T_i. \tag{1.1.12}$$

Agora, para cada valor possível da espessura d, a equação (1.1.12) tem duas incógnitas, n_i e α_i. Portanto, o mais provável é que tenha infinitas soluções e que, de fato, não seja difícil encontrar pelo menos uma. Por exemplo, fixando arbitrariamente n_i e resolvendo (1.1.12) para a agora única incógnita α_i. Claro que esse não pode ser o procedimento que resolva o problema físico. Fısicamente, o problema deve ter solução única, enquanto da maneira descrita, infinitas soluções diferentes poderiam ser encontradas. De fato, os graus de liberdade inerentes a (1.1.12) são drasticamente reduzidos incorporando informações fisicamente conhecidas, algumas óbvias, sobre d, α e n. Essas informações são:

(a) Tanto a espessura como os coeficientes n_i e α_i são positivos. Mais ainda, os índices de refração são maiores ou iguais a 1.
(b) $\alpha(\lambda)$ deve ser uma função decrescente e convexa (derivada segunda positiva).
(c) $n(\lambda)$ deve ser uma função decrescente e, também, com derivada segunda positiva.
As condições (a), (b) e (c) devem ser traduzidas como restrições do problema de estimar os parâmetros. Ou seja, devem ser encontradas expressões matemáticas envolvendo d, α_i e n_i que espelhem essas condições. Discretizando as derivadas segundas de $\alpha(\lambda)$ e $n(\lambda)$, essas expressões são:

$$d \geq 0, \quad n_i \geq 1, \quad \alpha_i \geq 0 \text{ para todo } i = 1, \ldots, n; \tag{1.1.13}$$

$$\alpha_{i+1} \leq \alpha_i \text{ e } n_{i+1} \leq n_i \text{ para todo } i = 1, \ldots, m-1; \tag{1.1.14}$$

$$n_i \geq n_{i-1} + \frac{n_{i+1} - n_{i-1}}{\lambda_{i+1} - \lambda_{i-1}}(\lambda_i - \lambda_{i+1}) \text{ e } \alpha_i \geq \alpha_{i-1} + \frac{\alpha_{i+1} - \alpha_{i-1}}{\lambda_{i+1} - \lambda_{i-1}}(\lambda_i - \lambda_{i+1}) \tag{1.1.15}$$

para todo $i = 2, \ldots, m-2$.
Considerando o objetivo (1.1.12) e as restrições (1.1.13), (1.1.14) e (1.1.15), o problema de estimação dos parâmetros pode agora ser modelado assim:

$$\text{Minimizar } \sum_{i=1}^{m} [T(\lambda_i, d, n_i, \alpha_i) - T_i]^2 \text{ sujeita a (1.1.13), (1.1.14) e (1.1.15).} \tag{1.1.16}$$

Observamos que (1.1.16) é um problema de minimização com restrições lineares onde há $2m+1$ variáveis. Se a tabela de dados (λ_i, T_i) obedecesse perfeitamente às fórmulas teóricas deveria existir uma solução de (1.1.16) onde o valor da função objetivo seria 0. Com dados experimentais não é isso o que acontece. De fato, o que se observa nesse caso, usando o método adequado para resolver (1.1.16) é a aparição de "soluções" onde a função objetivo toma um valor sensivelmente maior que 0. Isto se deve, além dos erros de medição que neste caso são, provavelmente, desprezíveis, a que a suposição "substrato transparente com s constante" é essencialmente falsa. Com efeito, para determinadas zonas do espectro (valores de λ) o substrato usado tem um coeficiente de absorção positivo (não é transparente) e, portanto, para essas zonas as equações (1.1.6)-(1.1.11) não se aplicam. Pior ainda, a distinção entre valores de λ para os quais o substrato não é transparente daqueles para os quais é, não é totalmente clara. O grau de aplicabilidade de (1.1.6)-(1.1.11) é de fato, um contínuo, variando entre a aplicabilidade e a não aplicabilidade absoluta. Um experimento adicional, que mede a transmissão produzida apenas pelo substrato (sem o filme), permite quantificar o grau de aplicabilidade das fórmulas. Diremos, então, que algumas equações (1.1.12) devem ser satisfeitas com um peso alto e outras com um peso muito baixo. Atribuindo efetivamente um peso $\theta_i > 0$ a cada equação, de acordo com a transparência do substrato para o comprimento de onda λ_i, o problema (1.1.16) é substituído por

$$\text{Minimizar } \sum_{i=1}^{m} \theta_i [T(\lambda_i, d, n_i, \alpha_i) - T_i]^2 \text{ sujeita a (1.1.13), (1.1.14) e (1.1.15).} \tag{1.1.17}$$

A atribuição de pesos às diferentes linhas da tabela original tem o efeito prático de eliminar a influência dos pontos onde o modelo está claramente errado. Isto aumenta os graus de liberdade do sistema total, e possibilita a existência de muitas soluções de (1.1.17), onde a função objetivo tem praticamente o mesmo valor. O método de otimização encontrou *uma* dessas soluções. Às vezes, pela observação da solução obtida, o físico tem condições de decidir se ela é razoável ou não. Neste problema particular, nosso experimentador encontra uma característica da função α considerada indesejável e sem sentido físico: apesar de ser decrescente e convexa, a função α obtida está formada por 4 segmentos de reta, violando uma suavidade adicional esperável no coeficiente de absorção real. Como os pontos de quebra dos diferentes segmentos de reta podem ser considerados como pontos onde a curvatura da função é muito grande, optamos por limitar o raio de curvatura de α e incluir explicitamente essa limitação no modelo. O cálculo elementar nos ensina que o raio de curvatura $R(\lambda)$ de $\alpha(\lambda)$ é dado por

$$\frac{1}{R(\lambda)} = \frac{\alpha''(\lambda)}{(1 + \alpha'(\lambda)^2)^{\frac{3}{2}}}. \tag{1.1.18}$$

Discretizando α' e α'' da forma usual, para todo $\lambda_i, i = 2, \ldots, m-1$, e estabelecendo uma limitação $\beta > 0$ para a curvatura obtemos as novas restrições

$$\frac{\alpha''(\lambda_i)}{(1 + \alpha'(\lambda_i)^2)^{\frac{3}{2}}} \le \beta, \tag{1.1.19}$$

onde as derivadas devem ser interpretadas como sua discretização usando $\alpha_{i-1}, \alpha_{i+1}$ e α_i.
Acrescentando (1.1.19) no modelo (1.1.17) passamos a ter $m-2$ restrições adicionais, todas elas não lineares. O problema ficou sensivelmente mais

difícil, mas sua solução tem maiores chances de possuir sentido físico. Uma alternativa, motivada pelo fato de que, estritamente falando, a cota β é arbitrária, consiste em incorporar as restrições (1.1.19) na função objetivo. Assim, a função objetivo de (1.1.17) passaria a ser

$$\sum_{i=1}^{m} \theta_i [T(\lambda_i, d, n_i, \alpha_i) - T_i]^2 + \rho \sum_{i=2}^{m-1} \frac{\alpha''(\lambda_i)}{(1 + \alpha'(\lambda_i)^2)^{\frac{3}{2}}}. \tag{1.1.20}$$

Em (1.1.20), ρ é um parâmetro que "castiga" o fato de se ter uma curvatura grande em λ_i. Desta maneira, não é necessário acrescentar as restrições (1.1.19) no problema (1.1.17).
A inclusão de (1.1.19) na sua forma original ou sob a forma (1.1.20) reduz, claramente, os graus de liberdade do problema e, em conseqüência, aumenta a probabilidade de encontrar coeficientes com sentido físico. Se isso é efetivamente conseguido depende de (muita) experimentação numérica, diálogo com os cientistas experimentais e sensibilidade específica. A construção de um bom modelo de otimização raramente se esgota em dois ou três passos de diálogo.

1.3 Definindo minimizadores

Daremos sentidos precisos aos termos *minimizador* e *mínimo* usados nas seções anteriores. Basicamente, veremos que esses termos podem ter dois significados:

(a) Dizemos que x_* é *minimizador global* de (1.1.1) se $f(x_*) \le f(x)$ para todo $x \in \Omega$. Neste caso, $f(x_*)$ é chamado *mínimo* de f em Ω.

(b) Dizemos que x_* é *minimizador local* de (1.1.1) se existe $\varepsilon > 0$ tal que $f(x_*) \le f(x)$ para todo $x \in \Omega$ tal que $\|x - x_*\| \le \varepsilon$.

Também, costuma-se dizer que x_* é *minimizador local estrito* de (1.1.1) se existe $\varepsilon > 0$ tal que $f(x_*) < f(x)$ para todo $x \in \Omega$ tal que $0 < \|x - x_*\| \le \varepsilon$.

Claramente, todos os minimizadores globais também são minimizadores locais. É fácil ver que, por outro lado, apesar de poder admitir muitos minimizadores globais, o valor do mínimo global é sempre o mesmo. Por exemplo, numa função constante, todos os pontos de Ω são minimizadores globais, mas em todos eles o valor de f é igual.

Lembramos que um conjunto Ω compacto é tal que toda seqüência $\{x_k\} \subset \Omega$ admite uma subseqüência convergente. O limite dessa subseqüência deve pertencer a Ω. Por outro lado, em $I\!R^n$, os conjuntos compactos são exatamente os fechados e limitados. Como a imagem inversa de conjuntos fechados por funções contínuas é fechada, o conjunto factível do problema geral de programação linear é fechado no caso usual em que as funções g_i e h_i são contínuas. Portanto, para ser compacto, esse conjunto precisa, apenas, ser limitado. O seguinte teorema, de prova bastante simples, é o mais importante da minimização global.

Teorema 1.3.1 - Bolzano-Weierstrass
Se Ω é compacto, e $f : \Omega \to I\!R$ é contínua, então existe $x_* \in \Omega$ minimizador global do problema (1.1.1).

Prova: Consideremos primeiro a possibilidade de que f não seja limitada inferiormente em Ω. Então, para cada $k \in \mathcal{N}$, existe $x_k \in \Omega$ tal que

$$f(x_k) \le -k,$$

portanto,

$$\lim_{k\to\infty} f(x_k) = -\infty. \tag{1.1.21}$$

Como Ω é compacto, existe K_1 um subconjunto infinito de $\mathcal{N}$ tal que a subseqüência $\{x_k\}_{k\in K_1}$ converge a um ponto de Ω, digamos x_*. Pela continuidade de f, isto implica que

$$\lim_{k\in K_1} f(x_k) = f(x_*),$$

o que entra em contradição com (1.1.21).
Podemos aceitar, portanto, que f é limitada inferiormente em Ω. Seja

$$\gamma = \inf_{x\in\Omega} f(x) > -\infty.$$

Pela definição de ínfimo, para todo $k \in \mathcal{N}$, existe $x_k \in \Omega$ tal que

$$\gamma \le f(x_k) \le \gamma + \frac{1}{k},$$

portanto

$$\lim_{k\to\infty} f(x_k) = \gamma.$$

Seja $\{x_k\}_{k \to K_1}$ uma subseqüência convergente de $\{x_k\}$ e seja x_* seu limite. Então, pela continuidade de f,

$$\gamma = \lim_{k \in K_1} f(x_k) = f(x_*).$$

Ou seja, $f(x_*)$ assume o valor ínfimo de f no conjunto Ω. Isto implica que x_* é minimizador global de (1.1.1). **QED**

Exercício 1.1: As restrições do problema (1.1.17) podem ser expressas como $Ax \geq b, l \leq x \leq u$. Identificar a matriz A e os vetores b, l e u.

Exercício 1.2: Encontrar exemplos onde todos os pontos de Ω são minimizadores locais mas $f(x) \neq f(y)$ se $x \neq y$.

Exercício 1.3: Desenhar conjuntos Ω em $I\!R^2$ e curvas de nível de funções f tais que existam vários minimizadores locais, globais, locais e globais, etc.

Exercício 1.4: Demonstrar o teorema Bolzano-Weierstrass para o caso em que f é semi-contínua inferiormente.

Exercício 1.5: Mostrar, com exemplos, que acontece quando as hipóteses de continuidade e compacidade do teorema Bolzano-Weierstrass são eliminadas.

Exercício 1.6: Provar que se f é contínua em $I\!R^n$ e $\lim_{\|x\| \to \infty} f(x) = \infty$ então f tem minimizador global em $I\!R^n$.

Exercício 1.7: Provar que se f é contínua em $I\!R^n$ e, dado $x_0 \in I\!R^n$, o conjunto de nível $\{x \in I\!R^n \mid f(x) \leq f(x_0)\}$ é limitado, então f tem minimizador global em $I\!R^n$.

Chapter 2

Condições de otimalidade

Neste livro tratamos de métodos para minimizar funções diferenciáveis em conjuntos de $I\!R^n$. As condições de otimalidade são relações entre as derivadas da função objetivo e as derivadas das funções que definem as restrições. As condições necessárias devem ser obrigatoriamente satisfeitas por minimizadores, enquanto as condições suficientes, quando satisfeitas, asseguram que o ponto em consideração é um minimizador local.

As derivadas (sobretudo as primeiras, às vezes também as segundas) da função objetivo e das restrições são o motor da maioria dos algoritmos que estudaremos, da mesma maneira que a potencialidade de movimento de uma partícula se encontra na sua velocidade e aceleração. As condições necessárias de otimalidade vão nos dizer se as derivadas envolvidas contém o germe necessário para imprimir um deslocamento que diminua o valor da função objetivo. Os métodos que estudaremos em capítulos posteriores ficam estáticos em cima de um ponto que satisfaz condições necessárias de otimalidade, mesmo que esse ponto não seja minimizador local nem, muito menos, global. Analogamente, quando estudamos convergência de algoritmos baseados em derivadas, podemos garantir apenas a estacionariedade (isto é, a satisfação de condições necessárias de otimalidade) dos pontos atingíveis no limite.

Freqüentemente, pontos limite de algoritmos são minimizadores, sobretudo quando o método trabalha ativamente diminuindo o valor da função objetivo em cada iteração. No entanto, garantir a condição de minimizador costuma ser difícil. Quando condições suficientes de otimalidade são satisfeitas podemos assegurar que o ponto em questão é minimizador local. A globalidade, no entanto, é muito mais complicada.

Ao longo deste capítulo supomos que f está bem definida e tem derivadas primeiras contínuas em um aberto que contém o conjunto Ω. Denotamos

$$\nabla f(x) = f'(x)^T = (\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x))^T.$$

Indicamos, como é usual, $f \in C^k(\Omega)$ para expressar que f tem derivadas contínuas até a ordem k no aberto que contém Ω. A expressão $f \in C^k$ indica que f tem derivadas contínuas até a ordem k num aberto que contém o domínio não especificado de f.
A notação $A \geq 0$ para $A \in I\!R^{n \times n}$ indica que A é semidefinida positiva. Da mesma forma, $A > 0$ significa que A é definida positiva.

2.1 Restrições em formato geral

Consideremos o problema

$$\begin{array}{c} \text{Minimizar } f(x) \\ x \in \Omega\ . \end{array} \tag{2.1.1}$$

As curvas no conjunto Ω desempenham um papel importante na derivação de condições práticas de otimalidade. A primeira condição de otimalidade que obteremos está baseada apenas no comportamento da função objetivo em cima de curvas factíveis que passam pelo ponto considerado. Apesar de sua generalidade, esta condição de otimalidade é usada no desenvolvimento de algoritmos modernos de minimização (pontos limite desses algoritmos satisfazem a condição). Ver [142], [144].

Definição 2.1.1
Dado $x_* \in \Omega$, chamamos *curva em Ω partindo de x_** a uma função contínua $\gamma : [0, \varepsilon] \to \Omega$ tal que $\varepsilon > 0$ e $\gamma(0) = x_*$.

Definição 2.1.2
Dado $x_* \in \Omega$, chamamos *curva em Ω de classe C^k partindo de x_** a uma função $\gamma : [0, \varepsilon] \to \Omega$ tal que $\varepsilon > 0$, $\gamma(0) = x_*$ e $\gamma \in C^k[0, \varepsilon]$.

Teorema 2.1.3 - Condição necessária de primeira ordem baseada em curvas
Seja x_* minimizador local de (2.1.1), e γ uma curva em Ω de classe C^1 partindo de x_*. Então $\nabla f(x_*)^T \gamma'(0) \geq 0$.

Prova: Definimos $\varphi : [0, \varepsilon] \to I\!R$ por $\varphi(t) = f(\gamma(t))$. Como x_* é minimizador local, existe $\varepsilon_1 \in (0, \varepsilon)$ tal que $\varphi(t) \geq \varphi(0)$ para todo $t \in (0, \varepsilon_1)$. Assim, $(\varphi(t) - \varphi(0))/t \geq 0$ para todo $t \in (0, \varepsilon_1)$ e, então, $\varphi'(0) \geq 0$. Mas, pela regra da cadeia,

$$\varphi'(t) = f'(\gamma(t))\gamma'(t),$$

portanto $\nabla f(\gamma(0))^T \gamma'(0) = \nabla f(x_*)^T \gamma'(0) \geq 0$. **QED**

Corolário 2.1.4

Seja x_* um ponto interior de Ω tal que x_* é minimizador local de (2.1.1). Então $\nabla f(x_*) = 0$.

Exercício 2.1: Demonstrar o Corolário 2.1.4.

Exercício 2.2: Provar que no Corolário 2.1.4 é suficiente que f tenha derivadas para obter a tese.

Corolário 2.1.5

Seja x_* minimizador de f em $I\!R^n$. Então $\nabla f(x_*) = 0$.

Teorema 2.1.6 - Condição necessária de segunda ordem baseada em curvas.

Seja x_* minimizador local de (2.1.1), $f \in C^2(\Omega)$.
(a) Para toda curva γ em Ω de classe C^2 partindo de x_*, $\nabla f(x_*)^T \gamma'(0) = \varphi'(0) \geq 0$, onde $\varphi(t) = f(\gamma(t))$.
(b) Se $\varphi'(0) = 0$, então $\varphi''(0) \geq 0$.

Prova: A prova do item (a) é a dada do Teorema 2.1.3. Em (b), quando $\varphi'(0) = 0$ temos $\varphi(t) = \varphi(0) + \frac{1}{2}\varphi''(0)t^2 + o(t^2)$, onde $\lim_{t\to 0} o(t^2)/t^2 = 0$. Portanto,

$$\lim_{t\to 0} \frac{\varphi(t) - \varphi(0)}{t^2} = \frac{1}{2}\varphi''(0).$$

Por ser x_* minimizador local, temos que $\varphi(t) \geq \varphi(0)$ para t suficientemente pequeno. Portanto, $\varphi''(0) \geq 0$. **QED**

Exercício 2.3: Generalizar o Teorema 2.1.6, definindo o *teorema da condição necessária de otimalidade de ordem k* baseada em curvas.

Definição 2.1.7

Dado $x \in \Omega$, dizemos que γ é uma *curva em* Ω *de classe* C^k *passando por* x se $\gamma : [-\varepsilon, \varepsilon] \to \Omega$, $\varepsilon > 0$, $\gamma(0) = x$ e $\gamma \in C^k$.

Lema 2.1.8
Se $x_* \in \Omega$ é um minimizador local de (2.1.1) e γ é uma curva em Ω de classe C^1 passando por x_*, então $\nabla f(x_*)^T \gamma'(0) = 0$.

Prova: Definimos $\gamma_1 : [0, \varepsilon] \to \Omega$ por $\gamma_1(t) = \gamma(t)$ e $\gamma_2 : [0, \varepsilon] \to \Omega$ por $\gamma_2(t) = \gamma(-t)$. Pelo Teorema 2.1.3,

$$\nabla f(x_*)^T \gamma_1'(0) \geq 0 \quad \text{e} \quad \nabla f(x_*)^T \gamma_2'(0) \geq 0.$$

Mas $\gamma_1'(0) = \gamma'(0)$ e $\gamma_2'(0) = -\gamma'(0)$, logo $\nabla f(x_*)^T \gamma'(0) = 0$. **QED**

Corolário 2.1.9 - Condição necessária de segunda ordem para x_* no interior de Ω (ou $\Omega = I\!R^n$).
Seja x_* minimizador local de (2.1.1), x_* ponto interior de Ω. Se f tem derivadas segundas contínuas numa vizinhança de x_* então $\nabla f(x_*) = 0$ e $\nabla^2 f(x_*) \geq 0$.

Prova: Seja $d \in I\!R^n$, $d \neq 0$, arbitrário. Seja $\gamma : [-\varepsilon, \varepsilon] \to \Omega$ a curva definida por $\gamma(t) = x_* + td$. Pelo Corolário 2.1.4 e o Lema 2.1.8,

$$\nabla f(x_*)^T d \equiv \nabla f(x_*)^T \gamma'(0) = 0.$$

Como d é arbitrário, segue que $\nabla f(x_*) = 0$. Definindo $\varphi : [-\varepsilon, \varepsilon] \to I\!R$ por $\varphi(t) = f[\gamma(t)]$, temos $\varphi'(0) = \nabla f(x_*)^T \gamma'(0) = 0$ e pelo Teorema 2.1.6,

$$0 \leq \varphi''(0) = \gamma'(0)^T \nabla^2 f(x_*) \gamma'(0) = d^T \nabla^2 f(x_*) d.$$

Novamente, a arbitrariedade de d implica em $\nabla^2 f(x_*) \geq 0$. **QED**

Teorema 2.1.10 - Condição suficiente de segunda ordem para x_* no interior de Ω (ou $\Omega = I\!R^n$) *Seja $f \in C^2(\Omega)$ e x_* ponto interior de Ω tal que $\nabla f(x_*) = 0$ e $\nabla^2 f(x_*) > 0$. Então x_* é minimizador local estrito do problema (2.1.1).*

Prova: Escrevendo a expansão de Taylor para f em torno de x_*, como $\nabla f(x_*) = 0$, temos:

$$f(x) = f(x_*) + \frac{1}{2}(x - x_*)^T \nabla^2 f(x_*)(x - x_*) + o(\|x - x_*\|^2) \ ,$$

onde $\lim_{x \to x_*} o(\|x - x_*\|^2)/\|x - x_*\|^2 = 0$ e $\| \cdot \|$ é uma norma qualquer em $I\!R^n$. Como $\nabla^2 f(x_*) > 0$, existe $a > 0$ tal que, para todo $x \neq x_*$,

$$(x - x_*)^T \nabla^2 f(x_*)(x - x_*) \geq a\|x - x_*\|^2 > 0 \; .$$

Logo, $f(x) \geq f(x_*) + \frac{a}{2}\|x - x_*\|^2 + o(\|x - x_*\|^2)$. Portanto, para $x \neq x_*$,

$$\frac{f(x) - f(x_*)}{\|x - x_*\|^2} \geq \frac{a}{2} + o(1),$$

onde $o(1) \equiv \frac{o(\|x - x_*\|^2)}{\|x - x_*\|^2}$ tende a 0 quando $x \to x_*$. Em conseqüência, para x suficientemente próximo e diferente de x_*,

$$\frac{f(x) - f(x_*)}{\|x - x_*\|^2} \geq \frac{a}{4} > 0.$$

Logo, $f(x) > f(x_*)$ para todo x numa vizinhança de $x_*, x \neq x_*$. **QED**

Exercício 2.4: Encontrar exemplos onde:

(a) x_* é minimizador local de f em Ω, mas $\nabla f(x_*) \neq 0$.

(b) x_* é minimizador local de f em Ω, $\nabla f(x_*) = 0$ mas $\nabla^2 f(x_*)$ não é semidefinida positiva.

(c) Ω é aberto, $\nabla f(x_*) = 0$ mas x_* não é minimizador local.

(d) Ω é aberto, $\nabla f(x_*) = 0, \nabla^2 f(x_*) \geq 0$ mas x_* não é minimizador local.

(e) Ω é aberto, x_* é minimizador local estrito mas $\nabla^2 f(x_*)$ não é definida positiva.

2.2 Restrições de igualdade

Consideremos o problema de minimização com restrições gerais de igualdade:

$$\begin{array}{c} \text{Minimizar } f(x) \\ h(x) = 0 \end{array} \tag{2.2.1}$$

onde $h : I\!R^n \to I\!R^m$. Como sempre, chamamos Ω ao conjunto factível do problema. Neste caso $\Omega = \{x \in I\!R^n \mid h(x) = 0\}$.

Definição 2.2.1 Se $x \in \Omega$, chamamos *conjunto tangente a* Ω *por* x (denotado por $M(x)$) ao conjunto dos vetores tangentes a curvas em Ω passando por x, ou seja:

$$M(x) = \{v \in I\!R^n \mid v = \gamma'(0) \text{ para alguma curva } \gamma \text{ passando por } x\} \ .$$

Utilizando a notação

$$h'(x) = \begin{pmatrix} \frac{\partial h_1}{\partial x_1}(x) & \dots & \frac{\partial h_1}{\partial x_n}(x) \\ \vdots & & \\ \frac{\partial h_m}{\partial x_1}(x) & \dots & \frac{\partial h_m}{\partial x_n}(x) \end{pmatrix} = \begin{pmatrix} h_1'(x) \\ \vdots \\ h_m'(x) \end{pmatrix} = \begin{pmatrix} \nabla h_1(x)^T \\ \vdots \\ \nabla h_m(x)^T \end{pmatrix} ,$$

podemos relacionar $M(x)$ com o núcleo do Jacobiano de $h(x)$, denotado por $\mathcal{N}(h'(x))$, pelo seguinte lema:

Lema 2.2.2
Para todo $x \in \Omega$, $M(x) \subset \mathcal{N}(h'(x))$.

Prova: Seja $v \in M(x)$ e $\gamma : [-\varepsilon, \varepsilon] \to \Omega$ tal que $\gamma'(0) = v$, $\gamma(0) = x$. Definimos $\Phi(t) = h(\gamma(t))$, para todo $t \in [-\varepsilon, \varepsilon]$. Portanto, $\Phi(t) = 0$ para todo $t \in [-\varepsilon, \varepsilon]$. Logo, $\Phi'(t) \equiv (\Phi_1(t), \dots, \Phi_m(t))^T = 0$ para todo $t \in (-\varepsilon, \varepsilon)$. Mas, pela regra da cadeia, $\Phi'(t) = h'(\gamma(t))\gamma'(t)$, portanto

$$h'(\gamma(t))\gamma'(t) = 0$$

para todo $t \in (-\varepsilon, \varepsilon)$. Logo, $0 = h'(x)\gamma'(0) = h'(x)v$, ou seja, $v \in \mathcal{N}(h'(x))$.
QED

É natural que nos indaguemos sobre a validade da recíproca do Lema 2.2.2: $\mathcal{N}(h'(x)) \subset M(x)$? Em geral esta relação não é verdadeira, conforme ilustra o seguinte exemplo. Consideremos $h(x_1, x_2) = x_1 x_2$, $x = (\ 0,\ 0\ )^T$. Então $M(x) = \{v \in I\!R^2 \mid v_1 v_2 = 0\}$, mas $h'(x) = (0, 0)$ e, claramente, $\mathcal{N}(h'(x)) = I\!R^2$.

Definição 2.2.3
Dizemos que $x \in \Omega \equiv \{x \in I\!R^n \mid h(x) = 0\}$ é um *ponto regular* se o posto de $h'(x)$ é igual a m ($\{\nabla h_1(x), \dots, \nabla h_m(x)\}$ é um conjunto linearmente independente).

Teorema 2.2.4

Seja $\Omega = \{x \in I\!\!R^n \mid h(x) = 0\}$, $h \in C^k$, $x \in \Omega$ um ponto regular. Então, para todo $v \in \mathcal{N}(h'(x))$, existe uma curva γ de classe C^k passando por x tal que $\gamma'(0) = v$. Portanto, $M(x) = \mathcal{N}(h'(x))$.

Prova: Seja $v \in \mathcal{N}(h'(x))$. Então $h'(x)v = 0$. Queremos encontrar uma curva γ em Ω passando por x tal que $\gamma'(0) = v$. Consideramos o sistema de equações

$$h(x + tv + h'(x)^T u) = 0 \ , \tag{2.2.2}$$

Para x e v fixos, este é um sistema de m equações com $m + 1$ variáveis ($u \in I\!\!R^m$ e $t \in I\!\!R$). Colocando $u = 0, t = 0$ temos uma solução particular deste sistema. O Jacobiano de (2.2.2) em relação a u em $t = 0$ é $h'(x)h'(x)^T \in I\!\!R^{m\times m}$ e é não singular pela regularidade de x. Logo, pelo Teorema da Função Implícita, existe $\bar{\gamma} \in C^k$, definida em $[-\varepsilon, \varepsilon]$, $\varepsilon > 0$, tal que (2.2.2) se verifica se e somente se $u = \bar{\gamma}(t)$. Portanto

$$h(x + tv + h'(x)^T \bar{\gamma}(t)) = 0 \text{ para todo } t \in [-\varepsilon, \varepsilon] \ . \tag{2.2.3}$$

Derivando (2.2.3) em relação a t, para $t = 0$ temos $h'(x)(v + h'(x)^T \bar{\gamma}'(0)) = 0$. Como $h'(x)v = 0$, segue que $h'(x)h'(x)^T \bar{\gamma}'(0) = 0$. Mas $h'(x)h'(x)^T$ é não singular, logo $\bar{\gamma}'(0) = 0$.
Em conseqüência, definindo $\gamma : [-\varepsilon, \varepsilon] \to \Omega$ por

$$\gamma(t) = x + tv + h'(x)^T \bar{\gamma}(t),$$

temos que

$$\gamma'(0) = v + h'(x)^T \bar{\gamma}'(0) = v.$$

Assim, γ é a curva procurada. Como v é arbitrário, temos que $\mathcal{N}(h'(x)) \subset M(x)$. Portanto, $M(x) = \mathcal{N}(h'(x))$. **QED**

Como conseqüência do Teorema 2.2.4 temos o seguinte resultado:

Teorema 2.2.5
Se x_* é minimizador local regular de (2.2.1), então $\nabla f(x_*) \perp \mathcal{N}(h'(x_*))$.

Prova: Seja $v \in \mathcal{N}(h'(x_*))$. Como x_* é regular, existe γ em Ω passando por x_* tal que $\gamma'(0) = v$. Pelo Lema 2.1.8, $\nabla f(x_*)^T v = 0$. **QED**

Teorema 2.2.6 - Multiplicadores de Lagrange

Se x_* é minimizador local regular de (2.2.1), então existem únicos $\lambda_1, \ldots, \lambda_m$ reais tais que $\nabla f(x_*) + \sum_{i=1}^m \lambda_i \nabla h_i(x_*) = 0$. ($\lambda_1, \ldots, \lambda_m$ são chamados multiplicadores de Lagrange do problema.)

Prova: Pelo Teorema 2.2.5, $\nabla f(x_*) \perp \mathcal{N}(h'(x_*))$. Logo, $\nabla f(x_*) \in \mathcal{R}(h'(x_*)^T)$, isto é, existe $\lambda \in I\!R^m$ tal que $\nabla f(x_*) + h'(x_*)^T \lambda = 0$. Como x_* é regular, o Jacobiano $h'(x_*)$ tem posto completo e então esse vetor de multiplicadores $\lambda \in I\!R^m$ é único. **QED**

Considerando os resultados obtidos para o problema (2.2.1), os candidatos a minimizador local para este problema serão os pontos regulares que, ao mesmo tempo, sejam soluções do sistema não linear com $n+m$ equações e $n+m$ incógnitas

$$\begin{array}{rcl} \nabla f(x) + h'(x)^T \lambda & = & 0 \\ h(x) & = & 0 \end{array} \tag{2.2.4}$$

Esses pontos serão chamados *estacionários* ou *críticos*. Naturalmente, os pontos não regulares de Ω também seriam candidatos a minimizador local.

Exercício 2.5: Provar o Teorema 2.2.6 usando o seguinte argumento: como x_* é regular, vale o Teorema da Função Implícita. Logo $h(x) = 0$ é, localmente, $x_B = \varphi(x_N)$. Então o problema (2.2.1) se reduz localmente a um problema sem restrições nas variáveis x_N. A condição necessária de primeira ordem para minimização irrestrita implica a tese do teorema.

Exercício 2.6: Provar que se $h(x) = Ax - b$, a regularidade não é necessária para a existência dos multiplicadores de Lagrange no Teorema 2.2.6.

Exercício 2.7: Provar que se x_* é minimizador local de (2.2.1) então existem $\lambda_0, \lambda_1, \ldots, \lambda_m$ reais tais que $\lambda_0 \nabla f(x_*) + \sum_{i=1}^m \lambda_i \nabla h_i(x_*) = 0$.

Definição 2.2.7
Chamamos Lagrangiano do problema (2.2.1) à função $\ell(x, \lambda) = f(x) + h(x)^T \lambda$.

Exercício 2.8: Relacionar a não singularidade do Jacobiano do sistema (2.2.4) com o comportamento de $\nabla^2_{xx} \ell(x, \lambda)$ no núcleo de $h'(x)$.

Exercício 2.9: Dar um exemplo onde x_* seja minimizador de (2.2.1) mas

x_* seja maximizador de f restrita à variedade tangente afim.

Teorema 2.2.8 - Condições necessárias de segunda ordem para restrições de igualdade.
Suponhamos que $f, h \in C^2, x_*$ é minimizador local regular de (2.2.1) e λ é o vetor de multiplicadores de Lagrange definido no Teorema 2.2.6. Então $v^T \nabla^2_{xx} \ell(x_*, \lambda) v \geq 0$, para todo $v \in \mathcal{N}(h'(x_*))$.

Prova: Pelo Teorema 2.2.6,

$$\nabla f(x_*) + h'(x_*)^T \lambda = 0 \tag{2.2.5}$$

Seja $v \in \mathcal{N}(h'(x_*))$. Pelo Teorema 2.2.4, existe uma curva γ em Ω de classe C^2 passando por x_* $(\gamma(0) = x_*)$ e tal que $v = \gamma'(0)$. Também, $\gamma'(0) \in \mathcal{N}(h'(x_*))$. Definindo $\varphi(t) = f(\gamma(t))$, pelo Lema 2.1.8, $\varphi'(0) = \nabla f(x_*)^T \gamma'(0) = 0$ e então pelo Teorema 2.1.6,

$$\varphi''(0) = \gamma'(0)^T \nabla^2 f(x_*) \gamma'(0) + \nabla f(x_*)^T \gamma''(0) \geq 0 \tag{2.2.6}$$

Agora, definindo $\Phi_i(t) = \lambda_i h_i(\gamma(t)), i = 1, \ldots, m$, temos que $\Phi_i'(t) = 0$ para todo $t \in (-\varepsilon, \varepsilon)$, portanto

$$\Phi_i''(0) = \gamma'(0)^T \lambda_i \nabla^2 h_i(x_*) \gamma'(0) + \lambda_i h_i'(x_*) \gamma''(0) = 0 \ .$$

Logo

$$\sum_{i=1}^m \Phi_i''(0) = \gamma'(0)^T \sum_{i=1}^m \lambda_i \nabla^2 h_i(x_*) \gamma'(0) + \lambda^T h'(x_*) \gamma''(0) = 0 \ . \tag{2.2.7}$$

Somando (2.2.7) e (2.2.6), por (2.2.5) segue que

$$\gamma'(0)^T (\nabla^2 f(x_*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x_*)) \gamma'(0) \geq 0.$$

Por ser v arbitrário a prova está completa. **QED**

Teorema 2.2.9 - Condições suficientes de segunda ordem para restrições de igualdade.
Se f, $h \in C^2$, $x_* \in \Omega$ satisfaz as condições necessárias de primeira ordem para (2.2.1), λ é o vetor de multiplicadores de Lagrange e $y^T \nabla^2_{xx} \ell(x, \lambda) y > 0$ para todo $y \in \mathcal{N}(h'(x_*)), y \neq 0$, então x_* é minimizador local estrito para

(2.2.1).

Exercício 2.10: Usando a redução a problemas irrestritos através do Teorema da Função Implícita, provar os Teoremas 2.2.8 e 2.2.9.

Exercício 2.11: Considerar o problema perturbado MRI(ε)

$$\begin{array}{c} \text{Minimizar } f(x) \\ h(x) = \varepsilon \end{array}$$

e seja x_* solução regular de MRI(0). Chamando $x_* = x(0)$ e usando as condições de otimalidade de MRI(ε) e o Teorema da Função Implícita para definir $x(\varepsilon)$, provar que $\frac{\partial f}{\partial \varepsilon_i}(x(0)) = -\lambda_i, \; i = 1, \ldots, m$.

2.3 Restrições de desigualdade

Consideremos agora o problema de minimização com restrições gerais de desigualdade:

$$\begin{array}{c} \text{Minimizar } f(x) \\ c(x) \leq 0 \end{array} \tag{2.3.1}$$

onde $c : I\!R^n \to I\!R^p$.

Definição 2.3.1
Para cada $x \in \Omega = \{x \in I\!R^n \mid c(x) \leq 0\}$, chamamos de *restrições ativas em* x àquelas para as quais $c_i(x) = 0$. Analogamente, chamamos *restrições inativas em* x àquelas para as quais $c_i(x) < 0$. Como na definição 2.2.4, chamaremos ponto regular a um ponto de Ω onde os gradientes das restrições ativas são linearmente independentes.

A prova do seguinte lema é evidente.

Lema 2.3.2
Se x_* é minimizador local de (2.3.1) e $I = \{i \in \{1, \ldots, p\} \mid c_i(x_*) = 0\}$, então x_* é minimizador local do problema

$$\begin{array}{c} \text{Minimizar } f(x) \\ c_i(x) = 0, \; i \in I \, . \end{array}$$

Com base no Lema 2.3.2, podemos aplicar ao problema (2.3.1) resultados já conhecidos para o problema de minimização com restrições de igualdade.

Lema 2.3.3
Se x_* é minimizador local de (2.3.1), $I = \{i \in \{1, \ldots, p\} \mid c_i(x_*) = 0\}$ e $\{\nabla c_i(x_*), i \in I\}$ é um conjunto linearmente independente, então para todo $i \in I$ existe $\mu_i \in I\!R$ tal que

$$\nabla f(x_*) + \sum_{i \in I} \mu_i \nabla c_i(x_*) = 0 \ .$$

Prova: Análoga à do Teorema 2.2.6. **QED**

O Lemma 2.3.3 nos diz que o gradiente de f é combinação linear dos gradientes das restrições ativas num minimizador local regular do problema. O teorema seguinte mostra que sabemos algo sobre os sinais dos coeficientes dessa combinação linear.

Teorema 2.3.4 - Condições Karush-Kuhn-Tucker (KKT).
Se x_* é minimizador local regular de (2.3.1) ($I = \{i \in \{1, \ldots, p\} \mid c_i(x_*) = 0\}$ e $\{\nabla c_i(x_*), i \in I\}$ é um conjunto linearmente independente) então existem únicos $\mu_i \in I\!R$, $\mu_i \geq 0, i \in I$ tais que

$$\nabla f(x_*) + \sum_{i \in I} \mu_i \nabla c_i(x_*) = 0 \ .$$

Prova: Tendo em vista o Lema 2.3.3, existem $\mu_i \in I\!R, i \in I$ tais que

$$\nabla f(x_*) + \sum_{i \in I} \mu_i \nabla c_i(x_*) = 0 \ . \tag{2.3.2}$$

Falta apenas mostrar que $\mu_i \geq 0, i \in I$. Suponhamos que exista $k \in I$ tal que $\mu_k < 0$. Chamemos

$$\Omega_I = \{x \in I\!R^n \mid c_i(x) = 0, i \in I\},$$

$$\Omega_k = \{x \in I\!R^n \mid c_i(x) = 0, i \in I, i \neq k\},$$

$M_I(x_*)$ o conjunto tangente a Ω_I por x_* e $M_k(x_*)$ o conjunto tangente a Ω_k por x_*. Pela regularidade de x_*, $\nabla c_k(x_*)$ não é combinação linear dos outros gradientes de restrições ativas em x_*. Portanto, existe $y \in M_k(x_*)$ tal que

$$\nabla c_k(x_*)^T y < 0 \; . \tag{2.3.3}$$

Seja $\gamma(t)$ uma curva em Ω_k passando por x_* com $\gamma'(0) = y$. Então, para $t \geq 0$ suficientemente pequeno, $\gamma(t) \in \{x \in I\!R^n \mid c(x) \leq 0\}$. Chamando $\varphi(t) = f(\gamma(t))$, temos que $\varphi'(0) = \nabla f(x_*)^T y$. Logo, por (2.3.2), (2.3.3) e $\mu_k < 0$ segue que $\varphi'(0) < 0$, o que contradiz o fato de x_* ser minimizador local. **QED**

2.4 Restrições de igualdade e desigualdade

Consideremos agora o problema geral de programação não linear:

$$\begin{array}{c} \text{Minimizar } \; f(x) \\ h(x) = 0 \\ c(x) \leq 0 \end{array} \tag{2.4.1}$$

onde $h : I\!R^n \to I\!R^m$ e $c : I\!R^n \to I\!R^p$.

Podemos estabelecer condições análogas às do Teorema (2.3.4) para o problema (2.4.1). De maneira similar aos casos anteriores, definimos ponto regular do conjunto factível como um ponto onde os gradientes das restrições ativas são linearmente independentes.

Teorema 2.4.1 - Condições Karush-Kuhn-Tucker gerais.
Seja x_* um minimizador local regular de (2.4.1). Seja $I = \{i \in \{1, \ldots, p\} \mid c_i(x_*) = 0\}$. Suponhamos que $\{\nabla h_i(x_*), \ldots, \nabla h_m(x_*)\} \cup \{\nabla c_i(x_*), i \in I\}$ é um conjunto linearmente independente. Então existem únicos $\lambda_1 \ldots, \lambda_m \in I\!R$ e $\mu_i \geq 0$ para todo $i \in I$ tais que

$$\nabla f(x_*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x_*) + \sum_{i \in I} \mu_i \nabla c_i(x_*) = 0 \; .$$

Exercício 2.13: Demonstrar o Teorema 2.4.1.

Desta forma, se x é um ponto regular e minimizador local para o problema (2.4.1), definindo $\mu_i = 0$ se $i \notin I$, podemos reescrever as condições KKT da seguinte forma:

$$\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x) + \sum_{i=1}^{p} \mu_i \nabla c_i(x) = 0 \tag{2.4.2}$$

$$h(x) = 0 \tag{2.4.3}$$

$$\mu_i c_i(x) = 0 \,, i = 1, \ldots, p \tag{2.4.4}$$

$$\mu_i \geq 0 \,, i = 1, \ldots, p \tag{2.4.5}$$

$$c_i(x) \leq 0 \,, i = 1, \ldots, p \tag{2.4.6}$$

As $n + m + p$ equações (2.4.2) - (2.4.4) formam um sistema não linear nas incógnitas $x \in I\!R^n, \lambda \in I\!R^m$ e $\mu \in I\!R^p$. As soluções deste sistema que satisfazem (2.4.5) e (2.4.6) são os pontos estacionários de (2.4.1)

Teorema 2.4.2 - Condições necessárias de segunda ordem (restrições de igualdade e desigualdade).
Seja x_* ponto regular e minimizador local de (2.4.1). Seja A a matriz cujas linhas são os gradientes das restrições ativas em x_*, excluindo os gradientes daquelas restrições de desigualdade cujo multiplicador é zero. Então, se λ e μ são os vetores de multiplicadores de Lagrange dados no Teorema 2.4.1,

$$y^T \nabla^2_{xx} \ell(x_*, \lambda, \mu) y \geq 0 \ \ \textit{para todo} \ \ y \in \mathcal{N}(A) \ ,$$

onde

$$\ell(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{i=1}^{p} \mu_i c_i(x) \ .$$

Exercício 2.14: Demonstrar o Teorema 2.4.2.

Exercício 2.16: Refazer os resultados deste capítulo trocando minimizadores por maximizadores.

Exercício 2.17: Interpretar geometricamente todos os resultados deste capítulo, incluindo os relativos ao Exercício 2.16.

Exercício 2.18: Estudar o Lema de Farkas, de um texto adequado sobre convexidade, e deduzir as condições de otimalidade da programação linear. Observar que, desta maneira, a aplicação do Teorema 2.3.4 à programação linear não depende da regularidade do ponto. Usando esse resultado, provar o resultado do Teorema 2.3.4 para minimização com restrições lineares sem a condição de regularidade.

Exercício 2.19: Desenhar um diagrama de conjuntos onde apareçam claramente as relações de inclusão existentes entre pontos regulares, pontos não regulares, minimizadores locais, minimizadores globais, pontos Karush-Kuhn-Tucker e soluções do sistema não linear (2.4.2)-(2.4.4).

Chapter 3

Convexidade e dualidade

Apesar da extensa análise permitida pelos dois temas tratados neste capítulo, procuramos fazer uma abordagem sintética para ambos. Nosso enfoque tem em vista os aspectos teóricos que efetivamente contribuem para o desenvolvimento de algoritmos práticos. Por exemplo, uma das propriedades mais fortes obtidas com hipóteses de convexidade em um problema de minimização é que as condições necessárias de otimalidade passam a ser suficientes. Em outras palavras, um ponto Karush-Kuhn-Tucker torna-se uma solução do problema. A teoria da dualidade, por sua vez, permite uma abordagem do problema original sob um outro ponto de vista. O dual de um problema de otimização tem como variáveis quantidades associadas às restrições do problema original. Em condições adequadas, resolver o problema dual é equivalente a resolver o original (primal) e, às vezes, trabalhar com o dual é mais fácil que com o primal. Mesmo em situações onde o primal e o dual não são equivalentes, problemas duais resolúveis fornecem informações úteis para resolver seus primais correspondentes. Do ponto de vista teórico, convexidade e dualidade fornecem estruturas sob as quais resultados relevantes sobre algoritmos e problemas podem ser obtidos. Por exemplo, as condições de otimalidade podem ser derivadas usando teoremas de separação de conjuntos convexos por hiperplanos (ver [91]). Por outro lado, a teoria de convergência de métodos importantes em programação não linear, como o método do Lagrangeano aumentado (capítulo 10 deste livro) é enriquecida pela consideração do problema dual (ver [175]).

3.1 Convexidade

Um conjunto convexo se caracteriza por conter todos os segmentos cujos extremos são pontos do conjunto. Se x e y são pontos de $\mathbb{R}^n$, o segmento que os une está formado pelos pontos z da forma $y+\lambda(x-y) \equiv \lambda x+(1-\lambda)y$ com $\lambda \in [0,1]$. Isso justifica a seguinte definição.

Definição 3.1.1
O conjunto $K \subset \mathbb{R}^n$ é chamado um *conjunto convexo* se para quaisquer $x, y \in K$ e para todo $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in K$.

Uma caracterização útil para conjuntos convexos é dada pelo seguinte teorema:

Teorema 3.1.2
K é um conjunto convexo se, e somente se, para quaisquer $x_1, \ldots, x_m$ elementos de K e para $\lambda_i \in [0,1], i = 1, \ldots, m$ tais que $\sum_{i=1}^m \lambda_i = 1$, a combinação convexa $\sum_{i=1}^m \lambda_i x_i$ também é um elemento de K.

Exercício 3.1: Demonstrar o Teorema 3.1.2.

Exercício 3.1: Seja $K \subset \mathbb{R}^n$ e $\|\cdot\|$ uma norma arbitrária.
(a) Provar que, se K é fechado e $x \in \mathbb{R}^n$, o problema de minimizar $\|y-x\|$ sujeito a $y \in K$ tem solução.
(b) Provar que, se K é convexo e $\|\cdot\|$ é a norma euclidiana, a solução desse problema é única. Nesse caso, a chamamos $P_K(x)$ (projeção de x em K).
(c) De agora em diante, K é convexo e fechado e a norma é a euclidiana. Provar a "propriedade do ângulo obtuso": Para todo $y \in K$,

$$\langle y - P_K(x), x - P_K(x) \rangle \leq 0.$$

(d) Provar a propriedade de aproximação total: Para todo $y \in K$,

$$\|y - P_K(x)\| \leq \|y - x\|.$$

(e) Provar a propriedade de contração: Para todo $x, z \in \mathbb{R}^n$,

$$\|P_K(x) - P_K(z)\| \leq \|x - z\|.$$

Deduzir a continuidade da função P_K.

Apresentamos a seguir alguns resultados básicos da teoria de convexidade.

Teorema 3.1.3
Se os conjuntos K_i, $i \in I$, são convexos, então $K = \bigcap_{i\in I} K_i$ também é convexo.

Prova: Sejam $x, y \in K = \bigcap_{i\in I} K_i$. Então $x, y \in K_i, i \in I$ e como os conjuntos $K_i, i \in I$ são convexos, para todo $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in K_i, i \in I$. Logo $\lambda x + (1-\lambda)y \in K$ para todo $\lambda \in [0,1]$. **QED**

Exercício 3.2: Se $A \subset I\!R^n$, chamamos de *fecho convexo de A* ao conjunto das combinações convexas dos pontos de A. Provar que o fecho convexo de qualquer conjunto é convexo. Provar que o fecho convexo de $A \subset I\!R^n$ está contido em qualquer convexo K tal que $A \subset K$.

Definição 3.1.4
Se K é um conjunto convexo, $f : K \to I\!R$, é uma *função convexa* se para todo $x,\ y \in K$, $\lambda \in [0,\ 1]$,

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y).$$

Definição 3.1.5
Se K é um conjunto convexo, denominamos *epigrafo* de $f : K \to I\!R$ ao conjunto

$$\{(x,\ y) \in I\!R^n \times I\!R \mid x \in K,\ y \ge f(x)\}.$$

Teorema 3.1.6 *A função $f : K \to I\!R$ é convexa se, e somente se, o epigrafo de f é convexo.*

Prova: Suponhamos que f seja convexa e tomemos $(x,\ \bar{x})$, $(y,\ \bar{y})$ pontos do epigrafo de f. Para $\lambda \in [0,\ 1]$, como K é convexo, $\lambda x + (1-\lambda)y \in K$. Agora, $\lambda\bar{x} + (1-\lambda)\bar{y} \ge \lambda f(x) + (1-\lambda)f(y) \ge f(\lambda x + (1-\lambda)y)$ pois f é convexa. Logo $\lambda(x,\ \bar{x}) + (1-\lambda)(y,\ \bar{y}) = (\lambda x + (1-\lambda)y,\ \lambda\bar{x} + (1-\lambda)\bar{y})$ pertence ao epigrafo de f para todo $\lambda \in [0,\ 1]$. Portanto, o epigrafo é convexo.
Suponhamos agora que f não seja convexa. Então existem $x,\ y \in K$ tais que $f(\lambda x + (1-\lambda)y) > \lambda f(x) + (1-\lambda)f(y)$ para algum $\lambda \in [0,\ 1]$. Assim, $(x,\ f(x))$ e $(y,\ f(y))$ são pontos do epigrafo de f. Então

$$\lambda(x,\ f(x)) + (1-\lambda)(y,\ f(y)) = (\lambda x + (1-\lambda)y,\ \lambda f(x) + (1-\lambda)f(y))\ ,$$

onde $\lambda x + (1-\lambda)y \in K$ mas $\lambda f(x) + (1-\lambda)f(y) < f(\lambda x + (1-\lambda)y)$. Portanto, $\lambda(x,\ f(x)) + (1-\lambda)(y,\ f(y))$ não pertence ao epigrafo de f. Logo o epigrafo de f não é convexo. **QED**

Funções convexas diferenciáveis podem ser caracterizadas pelo teorema a seguir:

Teorema 3.1.7
Sejam $K \subset I\!R^n$ aberto e convexo, $f : K \to I\!R$, $f \in C^1(K)$. Então f é convexa se, e somente se, $f(y) \geq f(x) + \nabla f(x)^T(y-x)$, para todo $x,\ y \in K$.

Prova: Seja f convexa como na hipótese do teorema, $x,\ y \in K$, $\lambda \in [0,\ 1]$. Logo, $f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda)f(x)$. Portanto,

$$f(x + \lambda(y-x)) - f(x) \leq \lambda(f(y) - f(x)) \ .$$

Então

$$\lim_{\lambda \to 0} \frac{f(x + \lambda(y-x)) - f(x)}{\lambda} \leq f(y) - f(x) \ .$$

Logo,

$$\nabla f(x)^T(y-x) \leq f(y) - f(x).$$

Dessa maneira, provamos que

$$f(x) + \nabla f(x)^T(y-x) \leq f(y) \text{ para todo } x, y \in K.$$

Reciprocamente, se $f(y) \geq f(x) + \nabla f(x)^T(y-x)$ para todo $x,\ y \in K$, chamando $z_\lambda = \lambda y + (1-\lambda)x$, temos

$$\begin{array}{rcl} f(x) & \geq & f(z_\lambda) + \nabla f(z_\lambda)^T(x - z_\lambda) \\ f(y) & \geq & f(z_\lambda) + \nabla f(z_\lambda)^T(y - z_\lambda) \ . \end{array}$$

Portanto,

$$\begin{array}{rcl} (1-\lambda)f(x) + \lambda f(y) & \geq & (1-\lambda)(f(z_\lambda) + \nabla f(z_\lambda)^T(x - z_\lambda)) \\ & & + \lambda(f(z_\lambda) + \nabla f(z_\lambda)^T(y - z_\lambda)) \\ & = & f(z_\lambda) + \nabla f(z_\lambda)^T(x - z_\lambda - \lambda x + \lambda z_\lambda + \lambda y - \lambda z_\lambda) \\ & = & f(z_\lambda) + \nabla f(z_\lambda)^T(\lambda y + (1-\lambda)x - z_\lambda) \\ & = & f((1-\lambda)x + \lambda y) \ . \end{array}$$

QED

Outro resultado útil, que estabelece o não decrescimento da derivada direcional para funções convexas, é apresentado a seguir.

Teorema 3.1.8
Seja $K \subset I\!R^n$ aberto e convexo, $f : K \to I\!R$, $f \in C^1(K)$. Então, f convexa se, e somente se, para todo $x,\ y \in K$,

$$\nabla f(x)^T (y - x) \leq \nabla f(y)^T (y - x) \ .$$

Exercício 3.3: Demonstrar o Teorema 3.1.8.

As funções convexas com duas derivadas contínuas são caracterizadas pelo seguinte resultado.

Teorema 3.1.9
Seja $K \subset I\!R^n$ aberto e convexo, $f : K \to I\!R$ e $f \in C^2(K)$. Então f é convexa se, e somente se, $\nabla^2 f(x) \geq 0$ para todo $x \in K$.

Exercício 3.4: Demonstrar o Teorema 3.1.9.

Definição 3.1.10.
Se K é um conjunto convexo, $f : K \to I\!R$ é uma *função estritamente convexa* se, para todo $x,\ y \in K$, $\lambda \in (0,\ 1)$,

$$f(\lambda x + (1 - \lambda) y) < \lambda f(x) + (1 - \lambda) f(y) \ .$$

Exercício 3.5: Provar os teoremas 3.1.7–3.1.9, com as modificações adequadas, substituindo “convexa” por “estritamente convexa”.

Teorema 3.1.11
Seja $f : K \to I\!R$ convexa e $a \in I\!R$. Então o conjunto de nível $\{x \in K \mid f(x) \leq a\}$ é convexo.

Exercício 3.6: Demonstrar o Teorema 3.1.11.

Definição 3.1.12.
Chamamos de *problema de programação convexa* a

$$\begin{array}{c} \text{Minimizar } \ f(x) \\ \text{sujeita a } x \in K \end{array}$$

onde K é um conjunto convexo e f é uma função convexa.

Teorema 3.1.17
Em um problema de programação convexa, todo minimizador local é global. O conjunto dos minimizadores é convexo. Se f é estritamente convexa, não pode haver mais de um minimizador.

Prova: Suponhamos que x_* é uma solução local não global do problema de programação convexa . Então existe $x \in K$ tal que $f(x) < f(x_*)$. Para $\lambda \in [0,\ 1]$, consideremos $x_\lambda = (1-\lambda)x_* + \lambda x$. Pela convexidade de K, $x_\lambda \in K$. Agora, pela convexidade de f,

$$f(x_\lambda) \leq (1-\lambda)f(x_*) + \lambda f(x) = f(x_*) + \lambda(f(x) - f(x_*)) < f(x_*).$$

Assim, para λ suficientemente próximo de 0, x_λ torna-se arbitrariamente próximo de x_*, mas $f(x_\lambda) < f(x_*)$. Portanto, x_* não poderia ser um minimizador local do problema de programação convexa.
Chamemos de S o conjunto dos minimizadores globais do problema. Sejam $x,\ y \in S$. Então $f(x) = f(y) \leq f(\lambda x + (1-\lambda)y),\ \lambda \in [0,\ 1]$. Pela convexidade de f,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) = f(y) + \lambda(f(x) - f(y)) = f(y).$$

Logo, $\lambda x + (1-\lambda)y \in S$ e portanto S é convexo.
Suponhamos agora que existam $x,\ y \in S$, $x \neq y$ e f seja estritamente convexa. Para $\lambda \in [0,\ 1]$, $f(\lambda x + (1-\lambda)y) \geq f(x) = f(y)$ pois $x,\ y$ são minimizadores globais, mas $f(\lambda x + (1-\lambda)y) < f(x) = f(y)$ pelo fato de f ser estritamente convexa. Temos assim a contradição desejada e a prova está completa. **QED**

No próximo teorema consideramos o problema geral de programação não linear (2.4.1). Suponhamos que a função objetivo f e as funções que definem as restrições de desigualdade $g_i, i = 1, \ldots, p$ são convexas e que as $h_i, i = 1, m$ são lineares, isto é, $h_i(x) = a_i^T x + b_i$. Portanto, pelos teoremas 3.1.3 e 3.1.5, o conjunto $\Omega = \{x \in I\!R^n \mid h(x) = 0,\ g(x) \leq 0\}$ é convexo e o problema de programação não linear (2.4.1) é um problema de programação convexa. Com certo abuso de linguagem, ao dizer que (2.4.1) é um problema de programação convexa estaremos sempre supondo que as g_i são convexas e as h_i são lineares. O objetivo do teorema é mostrar que, neste caso, as condições KKT dadas pelo Teorema 2.4.1 são suficientes para caracterizar um minimizador global.

Teorema 3.1.14
Se o problema de minimização com restrições de igualdade e desigualdade (2.4.1) é um problema de programação convexa e em x_* valem as condições KKT gerais (Teorema 2.4.1), então x_* é minimizador global (a regularidade não é necessária).

Prova: Definimos $\Omega = \{x \in I\!R^n \mid h(x) = 0,\ g(x) \leq 0\}$ e tomamos $x \in \Omega$, $x \neq x_*$. Se $\lambda \in I\!R^n$ e $\mu \in I\!R^p$ são os multiplicadores dados pelo Teorema 2.4.1, temos:

$$\nabla f(x_*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x_*) + \sum_{i=1}^{p} \mu_i \nabla g_i(x_*) = 0 \tag{3.1.1}$$

$$h(x_*) = 0 \tag{3.1.2}$$

$$\mu_i g_i(x_*) = 0\ , i = 1, \ldots, p \tag{3.1.3}$$

$$\mu_i \geq 0\ , i = 1, \ldots, p \tag{3.1.4}$$

$$g_i(x_*) \leq 0\ , i = 1, \ldots, p \tag{3.1.5}$$

Agora, $f(x) \geq f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{i=1}^{p} \mu_i g_i(x)$ pois $h_i(x) = 0,\ i = 1, \ldots, m,$ $g_i(x) \leq 0,\ i = 1, \ldots, p$ e vale (3.1.4).

Aplicando a desigualdade do Teorema 3.1.7 às funções f, h_i e g_i segue-se que

$$\begin{aligned} f(x) \quad \geq \quad & f(x_*) + \nabla f(x_*)^T (x - x_*) + \sum_{i=1}^{m} \lambda_i (h_i(x_*) + \nabla h_i(x_*)^T (x - x_*)) \\ & + \sum_{i=1}^{p} \mu_i (g_i(x_*) + \nabla g_i(x_*)^T (x - x_*))\ . \end{aligned}$$

Por (3.1.1) - (3.1.5) temos $f(x) \geq f(x_*)$, ou seja, x_* é minimizador global de (2.4.1). **QED**

3.2 Dualidade

Consideremos o problema geral de programação não linear (problema primal):

$$\begin{array}{cl} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0 \\ & g(x) \leq 0 \end{array} \tag{3.2.1}$$

onde $f : I\!R^n \to I\!R,\ h : I\!R^n \to I\!R^m,\ g : I\!R^n \to I\!R^p$ e $f,\ h,\ g \in C^1(I\!R^n)$.

Definição 3.2.1
Chamamos *Problema Dual (de Wolfe)* (ver [199]) de (3.2.1) ao problema

$$\begin{array}{cl} \text{Maximizar} & \ell(x,\ \lambda,\ \mu) \\ \text{sujeita a} & \nabla_x \ell(x,\ \lambda,\ \mu) = 0 \\ & \mu \geq 0 \end{array} \tag{3.2.2}$$

onde $\ell(x,\ \lambda,\ \mu) = f(x) + \displaystyle\sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{i=1}^{p} \mu_i g_i(x)$.

Reescrevendo (3.2.2), temos:

$$\begin{array}{cl} \text{Maximizar} & f(x) + \displaystyle\sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{i=1}^{p} \mu_i g_i(x) \\ \text{sujeita a} & \nabla f(x) + \displaystyle\sum_{i=1}^{m} \lambda_i \nabla h_i(x) + \sum_{i=1}^{p} \mu_i \nabla g_i(x) = 0 \\ & \mu \geq 0 \end{array} \tag{3.2.3}$$

Antes de estabelecer propriedades do Dual de Wolfe, calculamos os problemas duais de problemas clássicos de otimização.

Exemplo 3.2.2: *Programação Linear.*
Consideremos o problema primal de programação linear no seguinte formato:

$$\begin{array}{cl} \text{Minimizar} & c^T x \\ \text{sujeita a} & Ax \leq b \end{array} \tag{3.2.4}$$

onde $A \in I\!R^{p \times n},\ \ A^T = (a_1, \ldots, a_p)\ ,\ a_i \in I\!R^n,\ i = 1, \ldots, p$.

Neste caso, $\ell(x,\ \lambda,\ \mu) = \ell(x,\ \mu) = c^T x + \displaystyle\sum_{i=1}^{p} \mu_i (a_i^T x - b_i) = c^T x + \mu^T (Ax - b)$.

Logo, $\nabla_x \ell(x,\ \mu) = c + A^T \mu$.

Portanto o problema dual de (3.2.4) é dado por:

$$\begin{array}{rl} \text{Maximizar} & c^T x + \mu^T (Ax - b) \\ \text{sujeita a} & A^T \mu + c = 0 \\ & \mu \geq 0 \; . \end{array} \tag{3.2.5}$$

Utilizando $A^T \mu + c = 0$, podemos eliminar a dependência na variável x na função objetivo. Assim, (3.2.5) fica:

$$\begin{array}{rl} \text{Maximizar} & -b^T \mu \\ \text{sujeita a} & A^T \mu + c = 0 \\ & \mu \geq 0 \; . \end{array} \tag{3.2.6}$$

Substituindo $-\mu$ por $\pi \in I\!R^p$, reescrevemos (3.2.6) da seguinte forma:

$$\begin{array}{rl} \text{Maximizar} & b^T \pi \\ \text{sujeita a} & A^T \pi = c \\ & \pi \leq 0 \; . \end{array} \tag{3.2.7}$$

Podemos observar que, enquanto o problema primal tinha n variáveis e p restrições de desigualdade, o dual tem p variáveis, que devem ser negativas, e n restrições de igualdade. Se o problema primal é levado à forma padrão da programação linear, ele passa a ter $n + p$ variáveis (positivas) e p restrições lineares de igualdade. Esta análise ajuda a decidir em que situações usar o dual pode ser conveniente.

Exercício 3.7: Encontrar o dual de

$$\begin{array}{rl} \text{Maximizar} & c^T x \\ \text{sujeita a} & Ax = b \\ & x \geq 0 \; . \end{array}$$

Exemplo 3.2.3: *Programação quadrática*

Consideremos agora o problema geral de programação quadrática como sendo o problema primal:

$$\begin{array}{rl} \text{Minimizar} & \frac{1}{2} x^T G x + c^T x \\ \text{sujeita a} & Ax = b \\ & Cx \leq d \end{array} \tag{3.2.8}$$

onde $A \in I\!R^{m \times n}$, $C \in I\!R^{p \times n}$ e G simétrica não singular.

Então

$$\ell(x,\ \lambda,\ \mu) = \frac{1}{2}x^T Gx + c^T x + \lambda^T(Ax - b) + \mu^T(Cx - d)$$

$$\text{e} \quad \nabla_x \ell(x,\ \lambda,\ \mu) = Gx + c + A^T\lambda + C^T\mu.$$

Assim, o problema dual de (3.2.8) é

$$\begin{array}{rl} \text{Maximizar} & \frac{1}{2}x^T Gx + c^T x + \lambda^T(Ax - b) + \mu^T(Cx - d) \\ \text{sujeita a} & Gx + c + A^T\lambda + C^T\mu = 0 \\ & \mu \geq 0\ . \end{array} \tag{3.2.9}$$

Substituindo $x = -G^{-1}(c + A^T\lambda + C^T\mu)$, podemos reescrever (3.2.9) da seguinte forma:

$$\begin{array}{rl} \text{Maximizar} & -\frac{1}{2}(c + A^T\lambda + C^T\mu)^T G^{-1}(c + A^T\lambda + C^T\mu) - b^T\lambda - d^T\mu \\ \text{sujeita a} & \mu \geq 0\ . \end{array} \tag{3.2.10}$$

Neste exemplo vemos que o problema dual pode ter uma estrutura diferente do problema primal, neste caso mais simples. A simplicidade do problema dual está associada à possibilidade de calcular $G^{-1}v$. Essa tarefa pode ser muito difícil se G não tem uma estrutura favorável, mas muito fácil em casos bastante comuns nas aplicações. Por exemplo, se o problema primal consiste em encontrar a projeção de um ponto dado no conjunto factível de (3.2.8), a matriz G é a identidade.

Observamos que o dual (3.2.10) está bem definido se G é uma matriz não singular. Isso não significa que sempre seja equivalente ao primal. Para tanto, precisaremos que G seja definida positiva, o que resultará como corolário dos resultados seguintes. Em (3.2.2) e (3.2.3) definimos dualidade sem estabelecer conexões entre o primal e o dual. Com tal generalidade, os problemas primal e dual podem não ser equivalentes. Agora estudaremos relações entre os dois problemas usando hipóteses de convexidade.

Lembramos que chamamos condições Karush-Kuhn-Tucker (KKT) às dadas por (2.4.2)-(2.4.6), isto é:

$$\begin{array}{r} \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla h_i(x) + \sum_{i=1}^p \mu_i \nabla g_i(x) = 0 \\ h(x) = 0 \\ \mu_i g_i(x) = 0\ , i = 1, \ldots, p \\ \mu_i \geq 0\ , i = 1, \ldots, p \\ g_i(x) \leq 0\ , i = 1, \ldots, p \end{array}$$

Um ponto KKT é um ponto onde as condições KKT são satisfeitas.

Teorema 3.2.5
Suponhamos que o problema (3.2.1) é tal que as funções f e $g_i, i = 1, \ldots, p$ são convexas em $I\!R^n$ e que x_* é um ponto KKT com os multiplicadores correspondentes λ_* e μ_*. Então $(x_*,\ \lambda_*,\ \mu_*)$ é solução do dual (3.2.3). Além disso, o valor da função objetivo primal e dual coincidem, isto é $f(x_*) = \ell(x_*,\ \lambda_*,\ \mu_*)$.

Prova: Sabemos que

$$\nabla f(x_*) + \sum_{i=1}^{m} [\lambda_*]_i \nabla h_i(x_*) + \sum_{i=1}^{p} [\mu_*]_i \nabla g_i(x_*) = 0 \ ,$$

com $\mu_* \geq 0$. Das condições KKT se deduz que $f(x_*) = \ell(x_*, \lambda_*, \mu_*)$. Logo, $(x_*,\ \lambda_*,\ \mu_*)$ é um ponto factível para o problema dual (3.2.3). Suponhamos que $(x,\ \lambda,\ \mu)$ seja um outro ponto factível para (3.2.3). Então:

$$\begin{aligned}
\ell(x_*,\ \lambda_*,\ \mu_*) &= f(x_*) + \sum_{i=1}^{m} [\lambda_*]_i h_i(x_*) + \sum_{i=1}^{p} [\mu_*]_i g_i(x_*) \\
&= f(x_*) \\
&\geq f(x_*) + \sum_{i=1}^{m} \lambda_i h_i(x_*) + \sum_{i=1}^{p} \mu_i g_i(x_*) \\
&= \ell(x_*, \lambda, \mu).
\end{aligned}$$

Como (3.2.1) é um problema de programação convexa, é fácil ver que ℓ, como função de x, é convexa para $\mu \geq 0$. Logo, pelo Teorema 3.1.11 e pela factibilidade dual de $(x,\ \lambda,\ \mu)$ segue que

$$\ell(x_*,\ \lambda,\ \mu) \geq \ell(x,\ \lambda,\ \mu) + \nabla_x \ell(x,\ \lambda,\ \mu)^T (x_* - x) = \ell(x,\ \lambda,\ \mu) \ .$$

Isto completa a prova. **QED**

Alguns comentários sobre o Teorema 3.2.5 são pertinentes. Este resultado nos assegura que, se um problema de programação convexa tem um ponto que satisfaz as condições KKT (que portanto, pelo Teorema 3.1.18, será um minimizador global), esse ponto necessariamente vai ser um maximizador global do Dual de Wolfe. Isso não significa que dado um problema de programação convexa, uma solução global do dual corresponda forçosamente a uma solução do primal. No entanto, algumas relações adicionais entre primal e dual podem ser estabelecidas.

Teorema 3.2.6
Suponhamos que (3.2.1) é um problema de programação convexa. Se z é um ponto factível de (3.2.1) e (x, λ, μ) é um ponto factível do problema dual correspondente (3.2.2), então

$$f(z) \geq \ell(x, \lambda, \mu) \ .$$

Prova: Pelo Teorema 3.1.11 aplicado a f e g_i, factibilidade de z em relação a (3.2.1) e de (x, λ, μ) em relação a (3.2.2), temos que

$$\begin{aligned} f(z) - f(x) &\geq \nabla f(x)^T (z - x) \\ &= -\left[\sum_{i=1}^{m} \lambda_i \nabla h_i(x) + \sum_{i=1}^{p} \mu_i \nabla g_i(x)\right]^T (z - x) \\ &\geq -\sum_{i=1}^{m} \lambda_i [h_i(z) - h_i(x)] + \sum_{i=1}^{p} \mu_i [g_i(z) - g_i(x)] \\ &\geq \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{i=1}^{p} \mu_i g_i(x) \ . \end{aligned}$$

Portanto $f(z) \geq f(x) + \sum_{i=1}^{m} \lambda_i h_i(x)] + \sum_{i=1}^{p} \mu_i g_i(x) = \ell(x, \lambda, \mu)$, como queriamos provar. **QED**

O Teorema 3.2.6 implica que, se a região factível do primal (3.2.1) é não vazia mas o problema primal é ilimitado inferiormente, necessariamente a região factível do dual é vazia. Reciprocamente, se o dual é um problema factível mas ilimitado superiormente, então a região factível do primal é vazia. Deste resultado também se deduz que qualquer ponto factível do dual fornece uma cota inferior para o valor da função objetivo numa possível solução do primal. Esse tipo de informação pode ser muito útil na prática.

Exercício 3.8: Supondo que o primal tem apenas restrições lineares, que sua região factível é vazia e que a região factível do dual é não vazia, provar que o supremo da função objetivo do dual é $+\infty$. (Ver [199].)

Exercício 3.9: Considere o problema definido por $n = 1$, $m = 0, p = 1$, $f(x) = 0$ e $g(x) = e^x$. Mostrar que o primal é infactível mas o dual tem solução finita.

Exercício 3.10: Estabelecer as relações entre o dual de Wolfe e o seguinte problema

$$\text{Maximizar } F(\lambda, \mu) \text{ sujeita a } \mu \geq 0,$$

onde $F(\lambda, \mu)$ é o mínimo de $\ell(x, \lambda, \mu)$, em relação a $x \in I\!R^n$.

Chapter 4

Minimização de quadráticas

Uma quadrática é um polinômio em n variáveis com termos até segunda ordem. A minimização dessas funções tem interesse pelo grande número de aplicações que recaem nesse formato. Por exemplo, quando para um conjunto de dados empíricos se postula uma relação *linear* com certos parâmetros desconhecidos, o problema de ajustar esses parâmetros costuma ser resolvido através da minimização da soma dos quadrados dos erros, nesse caso, uma função quadrática. A soma de quadrados não é melhor que outras medidas globais do erro, em termos de qualidade do ajuste. No entanto, é a medida cuja minimização é mais simples do ponto de vista numérico. De fato, a minimização de quadráticas é um dos problemas mais fáceis na arte da otimização, fazendo também com que seja utilizado freqüentemente como *subproblema* auxiliar em algoritmos para resolver problemas mais complicados.

4.1 Quadráticas sem restrições

Dada a matriz simétrica $G \in I\!R^{n \times n}$, o vetor $b \in I\!R^n$ e a constante $c \in I\!R$, o problema tratado nesta seção é:

$$\text{Minimizar } \; q(x) \equiv \frac{1}{2} x^T G x + b^T x + c \, . \qquad (4.1.1)$$

É fácil ver que o gradiente de q é uma função vetorial linear e que a Hessiana é uma matriz constante:

Lema 4.1.1
Se $q(x) = \frac{1}{2}x^T Gx + b^T x + c$, *então* $\nabla q(x) = Gx + b$ *e* $\nabla^2 q(x) = G$ *para todo* $x \in \mathbb{R}^n$.

Exercício 4.1: Identificar G, b e c nos diferentes casos:
(a) $q(x) = 3x_1^2 - 2x_1x_2 + x_1x_3 - x_3^2 + x_3 - x_1 + 5$
(b) $q(x) = x_1^2 - x_2^2 + 4x_1x_3 + 2x_2x_3 + x_1 + x_2 - 8$
(c) $q(x) = 2x_1x_2 + x_1 + x_2$.

Exercício 4.2: Demonstrar o Lema 4.1.1.

Os pontos estacionários de (4.1.1) são aqueles onde se anula o gradiente, portanto, de acordo com o Lema 4.1.1, são as soluções do sistema linear

$$Gx + b = 0. \tag{4.1.2}$$

Sua existência ou unicidade está determinada pelas propriedades desse sistema.

Lema 4.1.2
(a) O problema (4.1.1) admite algum ponto estacionário se, e somente se, $b \in \mathcal{R}(G)$, *onde* $\mathcal{R}(G)$ *é o espaço coluna de* G.
(b) O problema (4.1.1) admite um único ponto estacionário se, e somente se, G *é não singular.*
Exercício 4.3: Demonstrar o Lema 4.1.2.

A equação dos pontos estacionários $Gx + b = 0$ pode ter uma, infinitas ou nenhuma solução. Se (4.1.2) não tem solução, ou seja, b não pertence ao espaço coluna de G, então (4.1.1) não admite nenhum minimizador, local ou global. Esse é o caso, por exemplo, quando q é uma função linear não constante ($G = 0$ e $b \neq 0$). Se (4.1.2) tem solução única, essa solução será o único ponto estacionário de (4.1.1). No entanto, ele pode ser tanto um minimizador, como maximizador ou "ponto sela". Finalmente, se G tem infinitas soluções, o que acontece quando G é singular e b está no seu espaço coluna, todas elas serão pontos estacionários e, como veremos, do mesmo tipo. É interessante observar que um problema com infinitas soluções (G singular e $b \in \mathcal{R}(G)$) pode ser transformado em um problema sem solução por uma perturbação arbitrariamente pequena no vetor b. Por exemplo, o sistema linear $0x + 0 = 0$ tem $\mathbb{R}^n$ como conjunto de soluções, mas o sistema $0x + \varepsilon = 0$ é incompatível para qualquer $\varepsilon \neq 0$. Isso mostra que, muitas

vezes, é difícil distinguir as situações “sem solução” e “infinitas soluções”. Com efeito, devido a erros de arredondamento, pode ser que o vetor b que, “na realidade”, estava no espaço coluna de G, fique fora desse subespaço fazendo que um sistema com infinitas soluções aparente ser incompatível nos cálculos numéricos. Também é possível que uma matriz G singular torne-se inversível , por perturbações de arredondamento, transformando um sistema incompatível, ou indeterminado, em um problema com solução única. Isso mostra que a situação em que G é “claramente não singular”, de maneira que pequenas perturbações não alteram essa condição, é muito mais confortável do ponto de vista da segurança dos cálculos numéricos.
Usando resultados de convexidade do Capítulo 3 e as condições de otimalidade de segunda ordem do Capítulo 2, podemos classificar facilmente os pontos estacionários de (4.1.1). Com efeito, se x_* é um minimizador local, necessariamente teremos $G = \nabla^2 q(x_*) \geq 0$. Por outro lado, se $G \geq 0$, temos que a Hessiana $\nabla^2 q(x)$ é semidefinida positiva para todo $x \in I\!R^n$ e, em conseqüência, q é uma função convexa. Portanto, se $G \geq 0$ e x_* é um ponto estacionário, necessariamente será um minimizador global. Como o mesmo tipo de raciocínio pode ser feito para maximizadores, deduzimos que toda quadrática tem um único tipo de ponto estacionário: minimizadores globais ou maximizadores globais ou ainda pontos sela, que não são maximizadores nem minimizadores locais. A prova do seguinte lema mostra que, devido à simplicidade das funções quadráticas, é fácil obter as conclusões acima sem apelar para os resultados de convexidade.

Lema 4.1.3
Se $G \geq 0$ e x_* é ponto estacionário de (4.1.1), então x_* é minimizador global de (4.1.1).

Prova: Seja x_* ponto estacionário de (4.1.1). Então $b = -Gx_*$. Logo,

$$\begin{aligned} q(x) &= \tfrac{1}{2}x^TGx + b^Tx + c \;=\; \tfrac{1}{2}x^TGx - x_*^TGx + c \\ &= \tfrac{1}{2}(x-x_*)^TG(x-x_*) - \tfrac{1}{2}x_*^TGx_* + c \;\geq\; -\tfrac{1}{2}x_*^TGx_* + c \\ &= \tfrac{1}{2}x_*^TGx_* - x_*^TGx_* + c \;=\; \tfrac{1}{2}x_*^TGx_* + b^Tx_* + c \;=\; q(x_*)\ . \end{aligned}$$

Portanto, $q(x) \geq q(x_*)$ para todo x, ou seja, x_* é minimizador global de (4.1.1). **QED**

Lema 4.1.4

Se (4.1.1) admite um minimizador local, então $G \geq 0$.

Corolário 4.1.5
Todo minimizador local de (4.1.1) é global.

Corolário 4.1.6
Se a matriz G é indefinida, então a quadrática q não tem extremos locais.

Exercício 4.4: Demonstrar o Lema 4.1.4 e os Corolários 4.1.5 e 4.1.6 sem usar as condições de otimalidade do Capítulo 2 nem os resultados de convexidade do Capítulo 3.

Um caso especial muito importante da minimização de quadráticas sem restrições é o problema de *quadrados mínimos* linear. Consiste em, dada uma matriz $A \in I\!R^{m \times n}$ e um vetor $b \in I\!R^m$, encontrar $x \in I\!R^n$ de maneira que Ax se aproxime de b "no sentido dos quadrados mínimos". Isto significa que x deve ser solução de

$$\text{Minimizar } \frac{1}{2}\|Ax - b\|_2^2. \tag{4.1.3}$$

Em (4.1.3), a fração $\frac{1}{2}$ não cumpre nenhum papel, exceto simplificar a expressão do gradiente e da Hessiana. O problema é equivalente a minimizar $q_2(x) \equiv \|Ax-b\|_2$, no entanto, a formulação com a norma ao quadrado é preferível, devido a q_2 não ser diferenciável nos pontos x em que $[Ax - b]_i = 0$. No entanto, (4.1.3) não é equivalente a minimizar outras normas de $Ax - b$. Em muitos ajustes de modelos é necessário estimar parâmetros x de maneira que as observações se aproximem bastante do modelo teórico ($Ax \approx b$). A escolha da norma euclidiana para medir o grau de aproximação se deve, na maioria dos casos, a que essa norma (ao quadrado) fornece o problema de otimização mais simples associado ao ajuste desejado. Algumas propriedades básicas do problema de quadrados mínimos linear são enunciadas no seguinte teorema.

Teorema 4.1.7
Se $q(x) = \frac{1}{2}\|Ax - b\|_2^2$, onde $A \in I\!R^{m \times n}, m \geq n$ e $b \in I\!R^m$, então
(a) $\nabla q(x) = A^T(Ax - b)$;
(b) $\nabla^2 q(x) = A^T A \geq 0$;
(c) As equações normais $A^T Ax = A^T b$ ($\nabla q(x) = 0$) sempre têm solução.
Se posto $(A) = n$, a solução é única e, se posto $(A) < n$, há infinitas soluções.

Exercício 4.5: Demonstrar o Teorema 4.1.7.

4.1.1 Usando fatorações

A forma mais rude de resolver (4.1.1) parte de considerar a decomposição espectral de G. (Ver, por exemplo, [96].) Ao mesmo tempo, ela nos dá toda a informação qualitativa relevante sobre o problema. Com efeito, como G é uma matriz simétrica, existe uma matriz ortogonal Q $(QQ^T = Q^TQ = I)$, e uma matriz diagonal Σ tais que

$$G = Q\Sigma Q^T. \tag{4.1.4}$$

Os autovalores de G, $\sigma_1, \ldots, \sigma_n$, são os elementos da diagonal Σ e os autovetores correspondentes são as colunas de Q. Assim, a matriz G é semidefinida positiva se todas as entradas de Σ são não negativas. Se todos os elementos da diagonal de Σ são maiores que 0, Σ e G são definidas positivas. Portanto, o exame da diagonal Σ fornece a informação sobre o tipo de pontos estacionários que o problema (4.1.1) pode ter. Se estamos interessados em minimizadores, e $\Sigma \geq 0$, analisamos o sistema linear $Gx + b = 0$. Usando (4.1.4), este sistema toma a forma

$$Q\Sigma Q^T x = -b, \tag{4.1.5}$$

que deriva, multiplicando ambos membros por $Q^T = Q^{-1}$, em

$$\Sigma z = -Q^T b \tag{4.1.6}$$

onde $x = Qz$. Agora, (4.1.6) tem solução se, e somente se, um possível zero na diagonal de Σ corresponde a uma coordenada nula do termo independente $-Q^T b$. Se há um zero na diagonal de Σ, digamos σ_i, tal que $[Q^T b]_i \neq 0$ o sistema (4.1.5) não tem solução, e, conseqüentemente, (4.1.1) carece de pontos estacionários. (Lembremos, porém, por um instante, a "advertência numérica" feita acima sobre a falta de estabilidade de conclusões deste tipo.) Se todos os elementos de Σ são estritamente positivos, (4.1.5) tem solução única, e o vetor x calculado através de (4.1.6) e a mudança de variáveis $x = Qz$ é o minimizador global de (4.1.1). Por fim, se o sistema é compatível, mas existe i tal que $\sigma_i = 0$ e $[Q^T b]_i = 0$, teremos infinitas soluções, todas elas minimizadores globais de (4.1.1). Nesse caso, qualquer que seja o valor de z_i escolhido, o vetor x correspondente resolverá (4.1.5) e o conjunto dos x varridos dessa maneira formará uma variedade afim em $\mathbb{R}^n$ de dimensão igual ao número de zeros da diagonal de Σ. O leitor verificará que o vetor de norma mínima dessa variedade afim resulta de escolher $z_i = 0$ toda vez que $\sigma_i = 0$ em (4.1.6).

Quando não existem minimizadores do problema (4.1.1), dado um x arbitrário pertencente a $I\!R^n$, é útil determinar uma direção $d \in I\!R^n$ tal que

$$\lim_{t \to \infty} q(x + td) = -\infty. \tag{4.1.7}$$

Se soubermos achar uma direção que satisfaça (4.1.7) poderemos dizer que *sempre* somos capazes de resolver (4.1.1), até quando o mínimo é $-\infty$ (e o minimizador é "$x + \infty d$"). Analisemos, pois, esse problema. Se algum autovalor de G, digamos σ_i, é menor que 0, tomamos d como o autovetor correspondente (a coluna i da matriz Q). Então,

$$\begin{aligned} q(x + td) &= \frac{1}{2}(x + td)^T G(x + td) + b^T(x + td) + c \\ &= q(x) + t\nabla q(x)^T d + \frac{1}{2}t^2 d^T G d \\ &= q(x) + t\nabla q(x)^T d + \frac{1}{2}\sigma_i t^2. \end{aligned}$$

Portanto, $q(x+td)$ como função de t é uma parábola côncava (coeficiente de segunda ordem negativo) e tende a $-\infty$ tanto para $t \to \infty$ quanto para $t \to -\infty$. Esta escolha de d não é a única que satisfaz (4.1.7). Com efeito, qualquer direção que cumprisse $d^T G d < 0$ teria a mesma propriedade. Direções que satisfazem a desigualdade $d^T G d < 0$ se dizem de *curvatura negativa*.

Consideremos agora o caso em que $\Sigma \geq 0$ mas existe $\sigma_i = 0$ com $[Q^T b]_i \neq 0$. Tomemos, de novo, d a coluna i de Q. Portanto, $b^T d \neq 0$ e $d^T G d = 0$. Se $b^T d > 0$, trocamos d por $-d$, de maneira que sempre podemos supor $b^T d < 0$. Fazendo o mesmo desenvolvimento que no caso anterior, chegamos a

$$\begin{aligned} q(x + td) &= q(x) + t\nabla q(x)^T d + \frac{1}{2}t^2 d^T G d \\ &= q(x) + t(Gx + b)^T d. \end{aligned}$$

Mas d é um elemento do núcleo de G, portanto $x^T G d = 0$ e

$$q(x + td) = q(x) + t b^T d.$$

Logo, $q(x + td)$ é uma reta com coeficiente angular negativo e tende a $-\infty$ quando $t \to \infty$.

A decomposição espectral resolve de maneira totalmente satisfatória o problema (4.1.1). Porém, seu custo computacional é, freqüentemente, intolerável, e a procura de alternativas mais baratas é necessária.
A maneira mais popular de resolver (4.1.1) se baseia na fatoração de Cholesky de G. Tal procedimento funciona e é estável apenas quando G é definida positiva. Nesse caso, a matriz G pode ser decomposta como $G = LDL^T$, onde $L \in I\!R^{n\times n}$ é triangular inferior com diagonal unitária e $D \in I\!R^{n\times n}$ é uma matriz diagonal com elementos positivos. A maneira de encontrar L e D, os fatores de Cholesky, é dada pelo seguinte algoritmo:

Algoritmo 4.1.8 - Fatoração de Cholesky.
Chamemos g_{ij} aos elementos de G, l_{ij} aos de L e d_{ij} aos de D. Defininindo, primeiro, $d_{11} = g_{11}$, as demais entradas de D e L são calculadas pelo seguinte ciclo.
Para $j = 2$ a n faça:

$$d_{jj} = g_{jj} - \sum_{k=1}^{j-1} d_{kk} l_{jk}^2$$

Se $j = n$, termine. Se $j < n$, para $i = j + 1$ a n faça:

$$l_{ij} = \frac{1}{d_{jj}} \left(g_{ij} - \sum_{k=1}^{j-1} d_{kk} l_{jk} l_{ik} \right).$$

O algoritmo de Cholesky termina, produzindo $D > 0$ (e é numericamente estável) se, e somente se, G é definida positiva. De fato, a maneira mais econômica de averiguar se uma matriz simétrica é definida positiva é tentar fazer sua fatoração de Cholesky. Se G é singular ou indefinida, em algum momento aparece um d_{jj} menor ou igual a 0 no cálculo dessas entradas.
Nos casos em que a fatoração de Cholesky de G é completada com sucesso, o único minimizador de (4.1.1) é obtido resolvendo $LDL^T x = -b$, processo que pode ser decomposto em três passos:
(a) resolver $Ly = -b$;
(b) resolver $Dz = y$;
(c) resolver $L^T x = z$.
Os três passos são computacionalmente simples: (a) e (c) consistem em resolver sistemas lineares triangulares, e (b) em dividir cada coordenada de y pela entrada diagonal d_{ii}. Acrescentando a este custo computacional o de fatorar a matriz pelo Algoritmo 4.1.8, a minimização da quadrática consome aproximadamente $n^3/6$ somas e produtos.
Quando, no Algoritmo 4.1.8, detectamos que G não é definida positiva, podemos apelar para o processo muito mais custoso de calcular a decom-

posição espectral. Outras alternativas, baseadas em fatorações mais baratas que a espectral, foram sugeridas na literatura. Ver, por exemplo, a fatoração Bunch-Parlett em [26]. Para efeitos práticos, quando se quer resolver (4.1.7) é, quase sempre, suficiente usar o seguinte problema auxiliar:

$$\text{Minimizar } q(x+d) \text{ sujeita a } \|d\|_2 \leq \Delta, \tag{4.1.8}$$

onde Δ é um número grande. Este problema pode ser resolvido por meio de um número não excessivo de fatorações de Cholesky, como veremos na Seção 4.2.

4.1.2 O caso esparso

A análise teórica feita na sub-seção anterior é válida independentemente da estrutura da matriz G mas, no Algoritmo 4.1.8, usamos, implicitamente, a suposição de que todos as entradas de G e L são armazenadas. Portanto, esse algoritmo usa mais de n^2 posições de memória. Quando G é esparsa, isto é, a grande maioria de suas entradas são nulas, é comum que a matriz L de sua fatoração de Cholesky também o seja. Às vezes, uma permutação conveniente de linhas e colunas de G (que corresponde a re-ordenar as variáveis x_i) faz aumentar consideravelmente o grau de esparsidade (ou "diminuir a densidade") do fator L. Ver, por exemplo, [62]. A fatoração de Cholesky de matrizes esparsas procede da mesma maneira que o Algoritmo 4.1.8, mas toma o cuidado de armazenar apenas os elementos não nulos de G e L, e evita fazer operações com zeros. Dessa maneira, não apenas a memória, mas também o tempo computacional pode diminuir muito e a economia é bastante significativa quando n é grande. Agora, se a fatoração de Cholesky falha, e nos interessa obter uma direção que satisfaça (4.1.7), apelar para a fatoração espectral é quase sempre impossível, porque a matriz Q desta fatoração é geralmente densa, independentemente da esparsidade de G. No entanto, ainda podemos obter uma direção satisfatória, em termos práticos, usando o subprobema (4.1.8).

Exercício 4.6: Obter um exemplo onde G é esparsa mas sua fatoração de Cholesky é densa e um exemplo onde G é esparsa, sua fatoração de Cholesky é esparsa mas sua fatoração espectral é densa.

4.1.3 Métodos iterativos

Os métodos baseados em fatorações, chamados *diretos*, calculam a solução de (4.1.1) em um único passo, através de um processo relativamente trabalhoso. Os métodos *iterativos*, estudados nesta seção, procedem, pelo contrário, computando uma seqüência de aproximações $x_k \in I\!R^n$. A passagem de um iterando para o seguinte se faz através de um conjunto de operações geralmente barato e a solução é obtida depois de um número finito de passos, ou no limite. Existem várias situações nas quais se justifica o uso de métodos iterativos. Às vezes, o problema é suficientemente fácil e pouquíssimas iterações do método podem fornecer uma aproximação muito boa da solução. Nesse caso, minimizaríamos a quadrática com um custo muito baixo, em contraste com os métodos baseados em fatorações, que tem um custo fixo, independentemente da dificuldade do problema. Outras vezes, a precisão requerida para a solução de (4.1.1) é moderada, e pode ser atingida com poucos passos do método iterativo.

No entanto, a principal razão pela qual se utilizam métodos iterativos é outra, e se deve a uma característica da maioria desses métodos que não está, forçosamente, ligada à recursividade. Com efeito, no processo da fatoração de uma matriz, precisamos usar, por um lado, a memória necessária para armazenar seus elementos e, por outro lado, a necessária para armazenar os fatores. Esta última é variável e pode exceder em muito a usada para guardar os dados (embora, naturalmente, certo grau de superposição é possível). Como vimos acima, no caso extremo, os fatores de uma matriz esparsa podem ser densos. Além disso, o tempo usado na fatoração cresce com o número de elementos não nulos dos fatores. Uma estimativa grosseira é que o tempo de fatoração é proporcional a $n \times |L|$, onde $|L|$ é o número de elementos não nulos do fator. Logo, se n é muito grande e as condições para a fatoração não são favoráveis, tanto o tempo quanto a memória necessária podem ser intoleráveis. Por outro lado, a memória usada pelos métodos iterativos é, em geral, muito moderada. Muitas vezes ela é apenas a usada para armazenar os elementos não nulos de G e alguns vetores adicionais, mas, freqüentemente, até menos que isso é preciso. De fato, a operação fundamental realizada por muitos métodos é o produto Gv da matriz por um vetor variável. Quando G tem uma lei de formação, esse produto matriz-vetor pode ser programado sem armazenamento explícito dos elementos de G, isto é, apenas gerando o elemento $[G]_{ij}$ quando é necessário usá-lo. Existem também métodos que podem ser implementados com geração de $[G]_{ij}$ apenas quando é necessário, e onde a operação básica não é o produto Gv.

O método dos gradientes conjugados [119] é o usado mais freqüentemente para resolver (4.1.1). Para motivá-lo, falaremos antes do método de *máxima descida*. Nesta seção, usaremos a notação $g(x) = \nabla q(x) = Gx + b$ e $\| \cdot \|$ será sempre a norma euclidiana. A direção $\bar{d} = -g(x)/\|g(x)\|$ é a de máxima descida a partir do ponto x. De fato, dada uma direção unitária d ($\|d\| = 1$) qualquer, a derivada direcional $D_d q(x)$ é tal que

$$D_d q(x) = g(x)^T d \geq -\|g(x)\| = D_{\overline{d}} q(x) \ .$$

Assim, dentre todas as direções unitárias, a determinada por $-g(x)$ é a que fornece a menor derivada direcional. Portanto, a função objetivo diminuirá se avançarmos nessa direção, e a máxima diminuição será obtida minimizando, ao longo dela, a quadrática q. Isto sugere o seguinte método iterativo:

Algoritmo 4.1.9 - Máxima descida
Seja $x_0 \in I\!R^n$, x_0 arbitrário.
Dado $x_k \in I\!R^n$, defina $d_k = -g(x_k)$ e, se possível, calcule x_{k+1} minimizador de $q(x_k + \alpha d_k)$, para $\alpha \geq 0$.

Exercício 4.7: Demonstrar que, se $d_k^T G d_k > 0$, existe uma fórmula fechada para o passo ótimo no Algoritmo 4.1.9: $\alpha_k = \dfrac{d_k^T d_k}{d_k^T G d_k}$. Provar que as direções de duas iterações consecutivas são ortogonais.

Infelizmente, além do método de máxima descida não produzir a solução do problema em um número finito de iterações, como as direções consecutivas por ele geradas são ortogonais, o método "anda em ziguezague" o que, certamente, nunca é a melhor forma de se acercar de um objetivo. Este comportamento se torna mais desfavorável à medida que as superfícies de nível de q se tornam mais alongadas, o que corresponde a um número de condição grande da matriz G. De fato, a velocidade de convergência deste método depende fortemente da razão entre o maior e o menor autovalor de G. Ver [129]. Nos últimos anos foram introduzidas variações do método de máxima descida onde se conserva o uso das direções dos gradientes mas é mudado o cálculo do passo, com substanciais ganhos de eficiência. Ver [8], [170], [80].

Vamos introduzir o método dos gradientes conjugados como uma espécie de "método de máxima descida com memória". Assim como o método de máxima descida minimiza q na direção $-g(x_0)$, depois na direção de $-g(x_1)$ etc., o método de gradientes conjugados começará minimizando q na direção

$-g(x_0)$, mas depois o fará no plano gerado por $-g(x_0)$ e $-g(x_1)$, depois no *subespaço* gerado por $-g(x_0)$, $-g(x_1)$ e $-g(x_2)$ e assim por diante. Usando a notação $Span\{u_1, \ldots u_\nu\}$ para o subespaço gerado pelos vetores $u_1, \ldots, u_\nu$, apresentamos no Algoritmo 4.1.10 uma primeira descrição geométrica do método dos gradientes conjugados. Nenhuma hipótese adicional sobre a matriz G é assumida além da simetria.

Algoritmo 4.1.10
Começamos o algoritmo com $x_0 \in I\!R^n$ arbitrário. Dado $x_k \in I\!R^n$, definimos

$$\mathcal{S}_k = Span\{-g(x_0), \ldots, -g(x_k)\}$$

e

$$\mathcal{V}_k = x_0 + \mathcal{S}_k = \{v \in I\!R^n \mid v = x_0 + w \text{ com } w \in \mathcal{S}_k\}.$$

Consideramos o problema

$$\text{Minimizar } q(x) \text{ sujeita a } x \in \mathcal{V}_k. \tag{4.1.9}$$

Se (4.1.9) não tem solução, o algoritmo pára "por inexistência de mínimo". Caso contrário, definimos x_{k+1} como uma das soluções de (4.1.9). (Mais tarde, provaremos, que, de fato, (4.1.9) não pode ter mais de uma solução.)

À primeira vista, o Algoritmo 4.1.10 pode parecer pouco prático, pois exige a minimização da quadrática $q(x)$ em variedades de dimensão cada vez maior. Logo, no último caso, estaremos minimizando q em todo $I\!R^n$ (afinal de contas, nosso problema original). No entanto, veremos que os cálculos necessários para computar os sucessivos iterandos são surpreendentemente simples e sem requerimentos de memória. Mais surpreendente é o fato de que, recentemente, foram desenvolvidos métodos iterativos para resolver sistemas lineares *não simétricos* baseados na idéia desse algoritmo, onde os cálculos das iterações não se simplificam, mas que, mesmo assim, parecem ser extremamente eficientes. Ver [179].

Vamos analisar algumas propriedades do Algoritmo 4.1.10. Para simplificar a notação, escreveremos, de agora em diante, $g_k = g(x_k)$ e $s_k = x_{k+1} - x_k$, para todo $k = 0, 1, 2, \ldots$. Da condição de otimalidade para minimização com restrições de igualdade, ou da condição de primeira ordem por curvas, dadas no Capítulo 2, se deduz que, se x_{k+1} está definido, g_{k+1} é ortogonal a $\mathcal{S}_k$. Se, nesse caso, $g_{k+1} \neq 0$, deduzimos que g_{k+1} não pode ser combinação linear de $g_0, g_1, \ldots, g_k$, portanto, com breve raciocínio indutivo, concluímos que o conjunto $\{g_0, g_1, \ldots, g_{k+1}\}$ é linearmente independente.

Por construção, s_k pertence a $\mathcal{S}_k$, o subespaço gerado por $\{g_0, g_1, \ldots, g_k\}$, para todo k. Portanto,

$$Span\{s_0, s_1, \ldots, s_k\} \subset \mathcal{S}_k.$$

Vamos provar, por indução, que a inclusão contrária também é verdadeira. Suponhamos, por hipótese indutiva, que

$$\mathcal{S}_k \subset Span\{s_0, s_1, \ldots, s_k\}.$$

Provaremos que

$$\mathcal{S}_{k+1} \subset Span\{s_0, s_1, \ldots, s_{k+1}\}. \tag{4.1.10}$$

Se $g_{k+1} = 0$ isto é trivial. Se $g_{k+1} \neq 0$, então, como a derivada direcional de q na direção de $-g_{k+1}$ é negativa, se deduz que, tomando $z = x_{k+1} - tg_{k+1} \in \mathcal{V}_{k+1}$ com t positivo e suficientemente pequeno, podemos obter $q(z) < q(x_{k+1})$. Como x_{k+2} é minimizador em $\mathcal{V}_{k+1}$, temos que $q(x_{k+2}) < q(x_{k+1})$. Isto implica que $x_{k+2} \notin \mathcal{V}_k$, já que x_{k+1} era minimizador em $\mathcal{V}_k$. Portanto s_{k+1} não pertence a $\mathcal{S}_{k+1}$. Isso implica que s_{k+1} é linearmente independente de g_0, $g_1 \ldots g_k$. Portanto, o coeficiente correspondente a g_{k+1} de s_{k+1} como combinação de $g_0, \ldots, g_{k+1}$ não pode ser nulo. Portanto, g_{k+1} é combinação de $g_0, \ldots, g_k, s_{k+1}$. Logo, da hipótese indutiva se obtem (4.1.10).

O resultado a seguir estabelece a terminação finita do Algoritmo 4.1.10. Mais precisamente, provaremos que existem duas possibilidades: que, em algum momento, o algoritmo pare "por inexistência" de minimizador de $q(x)$ em $\mathcal{V}_k$ ou que, em um número finito de passos (menor ou igual a n), encontre uma solução do sistema linear $Gx + b = 0$. Quando G é definida positiva ou quando G é semidefinida positiva mas $b \in \mathcal{R}(G)$, os minimizadores dos problemas (4.1.9) sempre existem. Portanto, nesses casos, o algoritmo termina com uma solução de $Gx + b = 0$, que, necessariamente, é minimizador global de (4.1.1). Se $b \notin \mathcal{R}(G)$, não existem soluções de (4.1.2). Logo, nesse caso, o teorema afirma que o algoritmo pára por inexistência de mínimo de (4.1.9) em alguma iteração k. Agora, se $b \in \mathcal{R}(G)$ mas G tem algum autovalor negativo, as duas possibilidades permanecem: que seja encontrada uma iteração que resolva (4.1.2) (ponto crítico de (4.1.1)) ou que o algoritmo pare por inexistência de minimizadores de (4.1.9).

Teorema 4.1.12
Se o Algoritmo 4.1.10 não pára "por inexistência de mínimo", então existe

$k \leq n$ tal que x_k é uma solução do sistema (4.1.2) (ponto estacionário de (4.1.1)).

Prova: Suponhamos que o Algoritmo 4.1.10 não pare por inexistência de mínimo. Então, para cada iteração k em que g_{k+1} é não nulo, temos que

$$\dim(\mathcal{V}_{k+1}) = \dim(\mathcal{V}_k) + 1.$$

Portanto, se chegamos a completar n iterações com gradientes não nulos, teremos $\dim(\mathcal{V}_{n-1}) = n$. Isso implica que $\mathcal{V}_{n-1} = \mathbb{R}^n$ e, portanto, x_n é solução de (4.1.1). **QED**

O resultado a seguir estabelece uma propriedade importante satisfeita pelos incrementos s_k, conhecida como *G-conjugação* ou *G-ortogonalidade.* A denominação gradientes conjugados tem como origem o fato deste método se basear em direções G-conjugadas.

Teorema 4.1.13
Se $\{x_k\}$ é uma seqüência gerada pelo Algoritmo 4.1.10, os incrementos $s_k = x_{k+1} - x_k$, $k = 0, 1, \ldots$ são G-conjugados, *isto é, para todo $k \geq 1$ vale*

$$s_j^T G s_k = 0 \ , \ \ j = 0, 1, \ldots, k-1. \tag{4.1.11}$$

Mais ainda, se g_0, $g_1, \ldots, g_{k-1}$ são não nulos e x_k está bem definido, então

$$s_j^T G s_j > 0 \text{ para todo } j = 0, 1, \ldots, k-1. \tag{4.1.12}$$

Prova: Já sabemos que $g_{k+1} \perp \mathcal{S}_k = Span\{g_0, g_1, \ldots, g_k\} = Span\{s_0, \ldots, s_k\}$. Então,

$$g_{k+1} \perp s_j \ , \ \ j = 0, 1, \ldots, k. \tag{4.1.13}$$

Agora, pela definição de s_k, e por cálculos elementares,

$$g_{k+1} = g_k + G s_k. \tag{4.1.14}$$

Pré-multiplicando (4.1.14) por s_j^T, para $j = 0, \ldots, k-1$, por (4.1.13) segue-se (4.1.11).
Agora provaremos (4.1.12). Se $g_j \neq 0$, temos que x_{j+1} está bem definido, e não pertence a $\mathcal{V}_{j-1}$, portanto $s_j \neq 0$ e $g_j^T s_j < 0$. Mas, pela definição de x_{j+1}, $t = 1$ deve ser minimizador de $q(x_j + t s_j)$. Como esta função

de t é uma parábola, para que exista um minimizador há duas possibilidades, ou é constante ou o coeficiente de segunda ordem é maior que 0. Mas $\frac{d}{dt}q(x_j + ts_j) = g_j^T s_j < 0$ em $t = 0$, portanto a parábola não é constante. Como o coeficiente de segunda ordem é $s_j^T G s_j/2$, segue-se (4.1.12). **QED**

Se x_{k+1} está bem definido, os resultados anteriores garantem que existem λ_0, $\lambda_1 \ldots \lambda_{k-1}, \lambda$ tais que $\lambda \neq 0$,

$$s_k = \lambda_0 s_0 + \ldots + \lambda_{k-1} s_{k-1} - \lambda g_k,$$

e os incrementos s_j são conjugados. Definindo $d_k = s_k/\lambda$, deduzimos que existem escalares $\omega_0, \ldots, \omega_{k-1}$ tais que

$$d_k = -g_k + \omega_0 s_0 + \ldots + \omega_{k-1} s_{k-1}.$$

Pre-multiplicando ambos membros por $s_j^T G$, $j = 0, 1, \ldots, k-1$, e usando a conjugação dos s_j, obtemos

$$0 = s_j^T G d_k = -s_j^T G g_k + \omega_j s_j^T G s_j,$$

ou seja, usando que $s_j^T G s_j > 0$,

$$\omega_j = \frac{g_k^T G s_j}{s_j^T G s_j}, \text{ para } j = 0, 1, \ldots, k-1.$$

Assim, como $Gs_j = g_{j+1} - g_j$, temos que $g_k^T G s_j = 0$ para $j = 0, 1, \ldots, k-2$. Logo, $\omega_j = 0$ para $j = 0, 1, \ldots, k-2$ e, conseqüentemente,

$$d_k = -g_k + \omega_{k-1} s_{k-1} = -g_k + \frac{g_k^T G s_{k-1}}{s_{k-1}^T G s_{k-1}} s_{k-1}. \tag{4.1.15}$$

Por fim, como x_{k+1} deve ser o minimizador de q ao longo da reta que passa por x_k, com direção d_k, obtemos

$$x_{k+1} - x_k = s_k = -\frac{d_k^T g_k}{d_k^T G d_k} d_k. \tag{4.1.16}$$

Antes de organizar um pouco melhor as fórmulas (4.1.15) e (4.1.16), vamos refletir sobre o significado das mesmas em relação ao Algoritmo 4.1.10. O fato mais relevante mostrado por essas expressões é que o cálculo de x_{k+1}, quando esse ponto está bem definido, depende apenas do incremento anterior s_{k-1}, e do gradiente atual g_k. Ou seja, a minimização de q na variedade $\mathcal{V}_k$

pode ser efetuada, contrariamente à intuição inicial, com trabalho e memória mínimos. Além disso, mostramos que a expressão obtida para s_k é única, eliminando a aparente liberdade existente na escolha do minimizador em $\mathcal{V}_k$ no Algoritmo 4.1.10.
Lembrando que $Gs_{k-1} = g_k - g_{k-1}$, e $g_k \perp g_{k-1}$, da fórmula (4.1.15) se deduz que

$$d_k = -g_k - \frac{g_k^T g_k}{s_{k-1}^T g_{k-1}} s_{k-1} = -g_k - \frac{g_k^T g_k}{d_{k-1}^T g_{k-1}} d_{k-1}. \tag{4.1.17}$$

Além disso, como d_{k-1} é a soma de $-g_{k-1}$ mais uma combinação dos gradientes anteriores, e esses gradientes são ortogonais a g_{k-1}, (4.1.17) toma a forma

$$d_k = -g_k + \beta_{k-1} d_{k-1}, \text{ onde } \beta_{k-1} = \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}}. \tag{4.1.18}$$

Finalmente, usando, também, que s_k é combinação de $-g_k$ e dos gradientes anteriores, a fórmula (4.1.16) deriva em

$$x_{k+1} = x_k + \alpha_k d_k \text{ onde } \alpha_k = \frac{g_k^T g_k}{d_k^T G d_k}. \tag{4.1.19}$$

As expressões (4.1.18) e (4.1.19) descrevem o algoritmo de gradientes conjugados de maneira mais operativa. Para fixar idéias, enunciamos de novo o Algoritmo 4.1.10 de maneira computacionalmente adequada.

Algoritmo 4.1.14 - Gradientes conjugados
Começamos com x_0 arbitrário e $d_0 = -g(x_0)$. Dados x_k, g_k e $d_k \in I\!R^n$, a seqüência de pontos x_k (a mesma definida no Algoritmo 4.1.10) é obtida da seguinte maneira:

Se $g_k = 0$, pare declarando convergência. Se $d_k^T G d_k \leq 0$ pare declarando inexistência de mínimo de (4.1.9). Se $g_k \neq 0$ e $d_k^T G d_k > 0$ calcule

$$x_{k+1} = x_k + \alpha_k d_k \ , \tag{4.1.20}$$

$$\text{onde} \quad \alpha_k = \frac{g_k^T g_k}{d_k^T G d_k} \ ; \tag{4.1.21}$$

$$g_{k+1} = g_k + \alpha_k G d_k \ ; \tag{4.1.22}$$

$$d_{k+1} = -g_{k+1} + \beta_k d_k \ , \tag{4.1.23}$$

$$\text{onde} \quad \beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}. \tag{4.1.24}$$

É interessante observar que nos casos em que o algoritmo pára por inexistência de mínimo, o vetor d_k fornece uma direção ao longo da qual q tende a $-\infty$. Com efeito, se $d_k^T G d_k < 0$, a parábola $q(x_k + td_k)$ tem coeficiente de segunda ordem menor que 0 e, em conseqüência, tende a $-\infty$ nos dois sentidos possíveis. Se $d_k^T G d_k = 0$ a expressão (4.1.23) mostra que a derivada direcional ao longo de d_k é negativa e a parábola $q(x_k + td_k)$ é, na realidade, uma reta decrescente. Portanto, a função tende a $-\infty$ quando $t \to \infty$.

Com base nos resultados anteriores sabemos que, no máximo em n passos, o método dos gradientes conjugados encontra uma solução do sistema linear (4.1.2) ou uma direção ao longo da qual a quadrática tende a $-\infty$. Veremos agora que, muitas vezes, o número necessário de passos é bem menor.

Teorema 4.1.15
O "subespaço de Krylov" da matriz G, definido por

$$\mathcal{K}(G, g_0, k) = Span\{g_0, Gg_0, \ldots, G^{k-1}g_0\},$$

coincide com $\mathcal{S}_k$.

Prova: A prova é feita por indução. Para $k = 1$, o resultado claramente vale. Suponhamos que $\mathcal{S}_k = Span\{g_0, Gg_0, \ldots, G^{k-1}g_0\}$ e vamos mostrar que $\mathcal{S}_{k+1} = Span\{g_0, Gg_0, \ldots, G^k g_0\}$. Por (4.1.22), $g_k = g_{k-1} + \alpha_{k-1} G d_{k-1}$. Pela hipótese de indução e pelo fato de que $\mathcal{S}_k = Span\{g_0, \ldots, g_{k-1}\} = Span\{d_0, \ldots, d_{k-1}\}$, tanto g_{k-1} quanto Gd_{k-1} pertencem a $Span\{g_0, \ldots, G^k g_0\}$. Além disso, $g_k \notin \mathcal{S}_k$ pois senão $g_k = 0$, já que $g_k^T d_j = 0\ ,\ \ j = 0, \ldots, k-1$. Portanto, $\mathcal{S}_{k+1} = Span\{g_0, Gg_0, \ldots, G^k g_0\}$, o que completa a prova. **QED**

Lema 4.1.16
A dimensão de $\mathcal{S}_k$ é, no máximo, o número de autovalores distintos da matriz G.

Prova: Seja $Q\Sigma Q^T$ a decomposição espectral da matriz G e chamemos $v = Q^T g_0$. Então, pelo Teorema 4.1.15,

$$\begin{aligned}
\mathcal{S}_k &= Span\{g_0, Gg_0, \ldots, G^{k-1}g_0\} \\
&= Span\{QQ^T g_0, Q\Sigma Q^T g_0, \ldots, Q\Sigma^{k-1}Q^T g_0\} \\
&= Span\{Qv, Q\Sigma v, \ldots, Q\Sigma^{k-1}v\}\ .
\end{aligned}$$

Portanto, a dimensão de $\mathcal{S}_k$ é a mesma que a do subespaço $Span\{v, \Sigma v, \ldots, \Sigma^{k-1} v\}$ e é fácil ver que esta dimensão não pode exceder o número de autovalores distintos de G (elementos da diagonal de Σ). **QED**

Com base no Lema 4.1.16, a terminação finita do Algoritmo 4.1.10 pode ser reescrita da seguinte forma:

Teorema 4.1.17
O método de gradientes conjugados aplicado ao problema (4.1.1) encontra uma solução do sistema $Gx + b = 0$ ou calcula uma direção ao longo da qual a quadrática tende a $-\infty$ em no máximo p passos, onde p é o número de autovalores distintos de G.

Apesar do resultado estabelecido no Teorema anterior, o método dos gradientes conjugados pode ser intoleravelmente lento em problemas de grande porte, se os autovalores diferentes são muitos, ou se o número de condição da matriz é grande. Por exemplo, nas matrizes provenientes de discretizações da equação de Laplace, à medida que o número de pontos cresce, o número de condição de G também aumenta muito e os autovalores são todos diferentes. Nesses casos, estratégias para acelerar o método tornam-se necessárias. Tradicionalmente, o que se faz é construir um problema equivalente ao original mas que seja mais favorável para o método, isto é, no qual a matriz Hessiana tenha um menor número de autovalores distintos e/ou tenha número de condição menor. Tal estratégia é conhecida por *precondicionamento.*
Vamos supor que, de alguma forma, conhecemos uma matriz H "parecida" com G e que H é simétrica definida positiva. Suponhamos que a decomposição espectral de H é $H = Q\Sigma Q^T$. Então, $H^{-\frac{1}{2}} = Q\Sigma^{-\frac{1}{2}} Q^T$ e a matriz $H^{-\frac{1}{2}} G H^{-\frac{1}{2}}$ estaria muito "próxima" da matriz identidade. Desta forma, H seria um precondicionador adequado, já que o problema original (4.1.1) ficaria equivalente ao seguinte problema precondicionado:

$$\text{Minimizar } \frac{1}{2} w^T H^{-\frac{1}{2}} G H^{-\frac{1}{2}} w + d^T w + c$$

onde $w = H^{\frac{1}{2}} x$, $d = H^{-\frac{1}{2}} b$ e o sistema $H^{-\frac{1}{2}} G H^{-\frac{1}{2}} w + d = 0$ teria resolução fácil pois $H^{-\frac{1}{2}} G H^{-\frac{1}{2}} \approx I$.
A arte do precondicionamento consiste em encontrar H parecida com G de maneira que tanto H quanto H^{-1} sejam fáceis de calcular. Um precondicionador clássico é tomar H como a diagonal de G. Também é usual adotar H como uma "fatoração de Cholesky incompleta" de G.

Exercício 4.8: Reescrever as fórmulas do Algoritmo 4.1.14 incorporando precondicionamento e trabalhando com as variáveis originais. Ver [96].

4.2 Quadráticas em bolas

Nesta seção consideramos o seguinte problema:

$$\begin{array}{ll} \text{Minimizar} & q(x) = \frac{1}{2}x^T Gx + b^T x + c \\ & \|x\| \leq \Delta \end{array} \qquad (4.2.1)$$

onde $G = G^T \in I\!R^{n \times n}, b \in I\!R^n, c \in I\!R, \Delta > 0$ e $\| \cdot \| = \| \cdot \|_2$, convenção adotada daqui em diante.
Contrariamente a (4.1.1), este problema sempre tem solução, já que as quadráticas são funções contínuas e a região factível de (4.1.11) é uma bola fechada, portanto, um compacto de $I\!R^n$. Vimos na Seção 4.1 que, quando (4.1.1) não tem solução, existem pontos de $I\!R^n$ ao longo dos quais a função tende a $-\infty$. Portanto, nesse caso, se chamamos $\bar{x}(\Delta)$ a uma solução de (4.2.1), teremos

$$\lim_{\Delta \to \infty} q(\bar{x}(\Delta)) = -\infty.$$

Além disso, é óbvio que $q(\bar{x}(\Delta))$ é não crescente como função de Δ. Logo, uma solução de (4.2.1) para Δ grande fornece uma boa aproximação para uma direção d que verifica (4.1.7).
O estudo do problema (4.2.1) se originou em certos subproblemas que aparecem na minimização irrestrita de funções gerais, como veremos no capítulo 7. Entretanto, recentemente, alguns autores utilizaram (4.2.1) como uma maneira de "regularizar" o problema de minimizar uma quadrática irrestrita. A idéia é que, quando G é muito mal condicionada, a solução exata de (4.1.1) carece de sentido, por ser extremamente sensível aos erros dos dados, ou ao arredondamento. Por outro lado, o problema (4.2.1) é bem condicionado se Δ não é grande. Portanto, substituir (4.1.1) por (4.2.1) representa um certo sacrifício em termos do erro no resíduo do sistema (4.1.2), mas freqüentemente compensado por uma maior estabilidade. Ver [191], [195], [116], [142].

A estrutura muito especial do problema (4.2.1) proporciona caracterizações dos minimizadores muito mais poderosas que no caso geral de minimização restrita. No caso geral, um minimizador deve ser um zero do gradiente do Lagrangiano e a Hessiana desta função deve ser semidefinida positiva num certo subespaço tangente (cf. capítulo 2). No seguinte teorema mostramos que, num minimizador global de (4.2.1), a Hessiana do Lagrangiano deve ser semidefinida positiva globalmente, e não apenas restrita a um subespaço. Ver [88], [187].

Teorema 4.2.1
Se z é solução de (4.2.1), então z é solução da equação

$$(G + \mu I)z = -b \tag{4.2.2}$$

com $\mu \geq 0,\ \mu(z^T z - \Delta^2) = 0$ *e* $(G + \mu I) \geq 0$.

Prova: O problema (4.2.1) é equivalente a

$$\begin{array}{c} \text{Minimizar } \ q(x) \\ x^T x \leq \Delta^2 \ . \end{array} \tag{4.2.3}$$

Como z é solução de (4.2.1), z satisfaz as condições KKT para (4.2.3), isto é, existe $\mu \geq 0$ tal que $Gz + b + \mu z = 0$ e $\mu(z^T z - \Delta^2) = 0$. Portanto, z e μ verificam (4.2.2).
Para vermos que $G + \mu I \geq 0$, suponhamos inicialmente que $z \neq 0$. Como z é solução de (4.2.1), z também é minimizador global de $q(x)$ sujeita a $\|x\| = \|z\|$. Então

$$q(x) \geq q(z) \text{ para todo } x \text{ tal que } \|x\| = \|z\| \ . \tag{4.2.4}$$

Substituindo (4.2.2) em (4.2.4), temos

$$\frac{1}{2}x^T G x - z^T(G + \mu I)x \geq \frac{1}{2}z^T G z - z^T(G + \mu I)z \ . \tag{4.2.5}$$

Rearranjando (4.2.5), segue que

$$\frac{1}{2}(x - z)^T(G + \mu I)(x - z) \geq 0$$

para todo x tal que $\|x\| = \|z\|$. Como $z \neq 0$, as direções $x - z$ tais que $\|x\| = \|z\|$ envolvem todas as direções do espaço exceto as ortogonais a z. Agora, qualquer vetor ortogonal a z é o limite de uma seqüência de vetores v_k para os quais, neste caso $v_k^T(G+\mu I)v_k \geq 0$. Portanto, passando ao limite, a expressão $v^T(G + \mu I)v \geq 0$ vale também para os vetores v ortogonais a z. Portanto, $G + \mu I \geq 0$.

Se $z = 0$, por (4.2.2) temos $b = 0$. Então $z = 0$ é solução de

$$\text{Minimizar } \frac{1}{2}x^T Gx + c \text{ sujeita a } \|x\| \leq \Delta,$$

e, pelo Lema 4.1.4, $G \geq 0$ e $v^T(G + \mu I)v \geq 0$ vale para todo $v \in I\!R^n$ com $\mu = 0$. **QED**

O próximo resultado fornece condições suficientes que garantem que z é solução de (4.2.1).

Teorema 4.2.2
Sejam $\mu \in I\!R$ e $z \in I\!R^n$ tais que

$$(G + \mu I)z = -b \text{ com } (G + \mu I) \geq 0. \tag{4.2.6}$$

(a) Se $\mu = 0$ e $\|z\| \leq \Delta$ então z é solução de (4.2.1).
(b) Se $\|z\| = \Delta$ então z é solução de

$$\textit{Minimizar } q(x) \textit{ sujeita a } \|x\| = \Delta.$$

(c) Se $\mu \geq 0$ e $\|z\| = \Delta$ então z é solução de (4.2.1).
Além disso, se $G + \mu I > 0$, então z é única em (a), (b) e (c).

Prova: Se μ e z satisfazem (4.2.6), z é minimizador da quadrática

$$\widehat{q}(x) = \frac{1}{2}x^T(G + \mu I)x + b^T x + c.$$

Logo,

$$\frac{1}{2}x^T(G + \mu I)x + b^T x + c \geq \frac{1}{2}z^T(G + \mu I)z + b^T z + c \tag{4.2.7}$$

para todo $x \in I\!R^n$.
De (4.2.7) segue que

$$q(x) \geq q(z) + \frac{\mu}{2}(z^T z - x^T x) \tag{4.2.8}$$

para todo $x \in I\!R^n$.
As afirmações (a), (b) e (c) são conseqüências imediatas de (4.2.8). A unicidade segue de (4.2.7) pois se $G + \mu I > 0$, a desigualdade é estrita para $x \neq z$. **QED**

Os teoremas acima mostram que, se existe uma solução z do problema (4.2.1) situada na fronteira da bola, ela deve satisfazer, com seu multiplicador correspondente μ, as seguintes equações:

$$(G + \mu I)z = -b, \quad \|z\| = \Delta. \tag{4.2.9}$$

Além disso, $\mu \geq 0$ e $G + \mu I \geq 0$. Soluções de (4.2.1) no interior da bola só podem existir se G é semidefinida positiva e, nesse caso, z, com norma menor que Δ, deve ser solução de (4.1.2).
Se $\sigma_1 \leq \ldots \leq \sigma_n$ são os autovalores de G, a condição $G+\mu I \geq 0$ é equivalente a $\mu \geq -\sigma_1$. Assim, as duas limitações sobre o multiplicador μ, para detectar soluções na fronteira, se resumem em

$$\mu \geq \text{ máximo } \{0, -\sigma_1\}. \tag{4.2.10}$$

Portanto, para encontrar as soluções de (4.2.1) na superfície da bola de uma maneira ingênua, dividimos o problema em duas questões:
(a) Existem soluções com $\mu > -\sigma_1$?
(b) $-\sigma_1$ é solução de ()?
A segunda questão pode ser eliminada se $\sigma_1 > 0$, ou seja, se G é definida positiva.
Examinemos a questão (a). Na região $\mu > -\sigma_1$ o sistema $(G+\mu I)z = b$ tem como solução única $z = -(G+\mu I)^{-1}b$ já que, neste caso, $G+\mu I$ é inversível. Portanto, encontrar $\mu > -\sigma_1$ satisfazendo () é equivalente a resolver

$$\|(G + \mu I)^{-1}b\| = \Delta. \tag{4.2.11}$$

ou

$$\varphi(\mu) = \Delta^2, \tag{4.2.12}$$

onde $\varphi(\mu) \equiv \|(G + \mu I)^{-1}b\|^2$. Parece bastante relevante, em conseqüência, estudar a forma da função univariada $\varphi(\mu)$. Consideremos a decomposição espectral $G = Q\Sigma Q^T$, onde $Q = (v_1, \ldots, v_n)$, $v_i \in I\!R^n$ e $\Sigma = \text{diag}\,(\sigma_1, \ldots, \sigma_n)$. Pela invariância da norma euclidiana sob transformações ortogonais, a função $\varphi(\mu)$ pode ser escrita como:

$$\varphi(\mu) = d^T(\Sigma + \mu I)^{-2}d = \sum_{i=1}^{n} \frac{d_i^2}{(\sigma_i + \mu)^2}, \tag{4.2.13}$$

onde $d = Q^T b$. A expressão (4.2.13) revela que

$$\lim_{\mu \to \infty} \varphi(\mu) = 0. \tag{4.2.14}$$

Ao mesmo tempo,

$$\lim_{\mu \to -\sigma_1 +} \varphi(\mu) = \infty \tag{4.2.15}$$

se, e somente se, $d_i = [Q^T b]_i \neq 0$ para algum i tal que $\sigma_1 = \sigma_i$. Neste caso, $\varphi(\mu)$ é estritamente decrescente e convexa. Isto significa que, quando b não é perpendicular ao subespaço de autovetores associado ao menor autovalor de G, a equação () tem uma única solução para $\mu > -\sigma_1$, qualquer que seja Δ. Se essa solução μ é maior ou igual a 0, $-(G + \mu I)^{-1} b$ será o único minimizador global de (4.2.1).

Quando b é perpendicular ao subespaço de autovetores associado ao menor autovalor de G a expressão de $\varphi(\mu)$ é

$$\varphi(\mu) = \sum_{i=\nu}^{n} \frac{d_i^2}{(\sigma_i + \mu)^2},$$

onde ν é o índice do menor autovalor diferente de σ_1. Portanto, nesse caso,

$$\varphi(-\sigma_1) = \sum_{i=\nu}^{n} \frac{d_i^2}{(\sigma_i - \sigma_1)^2},$$

e uma única solução de () maior que $-\sigma_1$ existirá se, e somente se, $\varphi(-\sigma_1) > \Delta$. Quando isso acontece, a função φ também é convexa e estritamente decrescente.

A análise acima esgota o exame da existência de soluções de () maiores que $-\sigma_1$. Suponhamos agora que existe z na fronteira da bola tal que $(G - \sigma_1 I)z = -b$. A matriz $G - \sigma_1 I$ é singular, portanto o sistema considerado tem infinitas soluções, e podemos considerar a solução de norma mínima $x^\dagger$. Usando a decomposição espectral, temos

$$(\Sigma - \sigma_1 I) Q^T x^\dagger = -Q^T b = d,$$

ou seja

$$(\sigma_i - \sigma_1)[Q^T x^\dagger]_i = d_i \text{ para } i = \nu, \ldots, n. \tag{4.2.16}$$

Os graus de liberdade da equação (4.2.16) são usados, na solução de norma mínima, escolhendo

$$[Q^T x^\dagger]_i = 0, \text{ para } i = 1, \ldots, \nu - 1. \tag{4.2.17}$$

De (4.2.16) e (4.2.17) é fácil deduzir que

$$\lim_{\mu \to -\sigma_1} (G + \mu I)^{-1} b = x^\dagger$$

e, portanto,

$$\lim_{\mu \to -\sigma_1} \varphi(\mu) = \|x^\dagger\|^2 \le \Delta^2.$$

Portanto, neste caso, não pode haver nenhuma solução de () com μ maior que $-\sigma_1$.
Resumindo, a existência de um minimizador global na fronteira com multiplicador maior que $-\sigma_1$ é incompatível com a existência de outro minimizador global com o multiplicador igual a $-\sigma_1$. Pelo exposto, vemos que, para que $-\sigma_1$ seja o multiplicador ótimo, b deve ser ortogonal ao subespaço de autovetores associado a σ_1. Para encontrar, nesse caso, um minimizador global pode-se proceder encontrando uma solução qualquer de $(G - \sigma_1 I)x = -b$, um autovetor v associado a $-\sigma_1$ e, finalmente, um elemento da fronteira da bola com a forma $x + tv$.

O exposto acima mostra que, possuindo a decomposição espectral de G, resolver o problema (4.2.1) carece de segredos. Como em geral a decomposição espectral é computacionalmente cara, procura-se desenvolver algoritmos que a evitem. Via de regra, esses algoritmos resolvem a equação () calculando φ mediante uma fatoração de Cholesky de $G + \mu I$ para cada tentativa μ. Ver [148]. Mais precisamente, resolve-se a equação

$$\frac{1}{\|(G + \mu I)^{-1} b\|} = \frac{1}{\Delta}$$

que é mais favorável à aplicação do método de Newton para achar zeros de funções que (). Ver [171], [115]. Agora, o caso em que o multiplicador ótimo é $-\sigma_1$, ou está próximo desse valor crítico é complicado numericamente, motivo pelo qual é conhecido como "hard case" na literatura. Atualmente trabalha-se intensamente em métodos para resolver (4.2.1) que usem métodos iterativos lineares, em vez de fatorações de matrizes. Ver [188], [180], [201].

Exercício 4.9: Estabelecer e provar rigorosamente as propriedades de φ e suas derivadas primeira e segunda. Provar que o número total de pontos estacionários de (4.2.1) na fronteira da bola é menor ou igual a $2 \times q$, onde q é o número de autovalores distintos de G.

Exercício 4.10: Estudar as propriedades da função $1/\varphi^{1/2}$ usada para encontrar efetivamente o multiplicador associado a uma solução de (4.2.1).

4.3 Quadráticas em caixas

Em muitos problemas práticos em que se deseja ajustar um modelo linear a um conjunto de dados empíricos, os parâmetros desconhecidos tem sentido físico apenas em uma determinada região do espaço. Nesses casos, em vez de um problema puro de quadrados mínimos teremos um problema de quadrados mínimos com restrições. A situação mais comum é quando cada parâmetro não pode ser inferior a determinada cota, nem superior a outra. Nesse caso, o conjunto de restrições toma a forma

$$l_i \leq x_i \leq u_i \text{ para todo } i = 1, \dots, n,$$

ou, mais brevemente,

$$l \leq x \leq u.$$

O conjunto $\Omega \subset I\!R^n$ formado pelos pontos que satisfazem essas restrições se diz uma *caixa* de $I\!R^n$, denominação mais confortável que a alternativa "hiperparalelepípedo". É conveniente admitir os valores $-\infty$ para l_i e $+\infty$ para u_i, já que, às vezes, apenas algumas variáveis estão naturalmente limitadas e, outras, a limitação é somente inferior, ou superior. Em problemas físicos é muito comum que as incógnitas, representando determinados coeficientes, devam ser positivas, em cujo caso Ω é o ortante $\{x \in I\!R^n \mid x_i \geq 0, i = 1, \dots, n\}$.
Entretanto, como no caso da minimização em bolas, o problema de minimização de quadráticas em caixas não tem interesse apenas por sua aplicação direta. Como veremos mais adiante, este também é um subproblema muito utilizado, de maneira iterativa, quando o objetivo último é resolver um problema mais complicado, por exemplo, a minimização de uma função geral (não quadrática) numa caixa. Nesses casos, a matriz G será a Hessiana da função objetivo num ponto dado e, como nada se sabe a priori sobre os autovalores dessa matriz, é importante considerar não apenas o caso convexo, como também o caso em que a matriz não é semidefinida positiva.
Veremos que, contrariamente à minimização em bolas, em que podíamos reconhecer perfeitamente um minimizador global mesmo no caso não convexo, os algoritmos práticos que apresentaremos deverão se contentar com pontos estacionários. Garantir um minimizador global nestes problemas é possível, mas apenas através de métodos muito caros computacionalmente. Ver [194].

Nosso problema é, pois,

$$\begin{array}{ll} \text{Minimizar} & q(x) \\ \text{sujeita a} & x \in \Omega \,, \end{array} \qquad (4.3.1)$$

onde $\Omega = \{x \in I\!R^n \mid l \leq x \leq u \,, l < u\}$, $q(x) = \frac{1}{2}x^T Gx + b^T x + c$. Se G é semidefinida positiva (4.3.1) é um problema convexo e os pontos estacionários coincidem com os minimizadores globais.
Denotaremos $\gamma = \min\{u_i - l_i \,,\ i = 1, ..., n\}$. Veremos que, nas operações em que aparecerá γ, a possibilidade $\gamma = \infty$ terá interpretação unívoca. Outra notação útil será $\bar{g}(x) \equiv -\nabla q(x) \equiv -(Gx + b)$. Em várias situações (nas provas teóricas, não no algoritmo) usaremos uma cota superior $L > 0$ do maior autovalor de G. Teremos assim que, para todo $x, z \in I\!R^n$,

$$q(z) - q(x) - \nabla q(x)^T (z - x) = \frac{1}{2}(z - x)^T G(z - x) \leq \frac{L}{2}\|z - x\|^2. \quad (4.3.2)$$

Definimos uma *face aberta* de Ω como um conjunto $F_I \subset \Omega$, onde I é um subconjunto (talvez vazio) de $\{1, 2, \ldots, 2n\}$ que não contém simultaneamente i e $n + i$, $i \in \{1, 2, \ldots, n\}$, tal que

$$F_I = \{x \in \Omega | x_i = l_i \text{ se } i \in I,\ x_i = u_i \text{ se } n{+}i \in I,\ l_i < x_i < u_i \text{ nos outros casos } \}.$$

Por exemplo, se $\Omega = \{x \in I\!R^3 \mid 1 \leq x_1 \leq 5, 2 \leq x_2\}$ teremos $F_{\{1,2\}} = \{x \in I\!R^3 \mid x_1 = 1, x_2 = 2\}$, $F_{\{4\}} = \{x \in I\!R^3 \mid x_1 = 5, 2 < x_2\}$, $F_{\emptyset} = \{x \in I\!R^3 \mid 1 < x_1 < 5, 2 < x_2\}$ e assim por diante. Claramente, faces abertas correspondentes a sub-índices diferentes são disjuntas ($I \neq J$ implica que a intersecção entre F_I e F_J é vazia) e Ω é a união de todas as suas faces abertas.
Chamamos $\bar{F}_I$ o fecho de cada face aberta, $V(F_I)$ a menor variedade afim que contém F_I, $S(F_I)$ o subespaço paralelo a $V(F_I)$ e dim F_I a dimensão de $S(F_I)$. É fácil ver que dim $F_I = n - |I|$, onde $|I|$ denota o número de elementos de I, ou, em linguagem equivalente, o número de restrições (ou "canalizações") ativas nos pontos de F_I. Lembrando termos usados no Capítulo 2, podemos verificar também que todos os pontos de uma caixa Ω são regulares.
Para cada $x \in \Omega$ definimos o gradiente projetado negativo, ou "vetor de Cauchy" $\bar{g}_P(x) \in I\!R^n$ como

$$\bar{g}_P(x)_i = \begin{cases} 0 & \text{se } x_i = l_i \text{ e } [\nabla q(x)]_i > 0 \\ 0 & \text{se } x_i = u_i \text{ e } [\nabla q(x)]_i < 0 \\ -[\nabla q(x)]_i & \text{nos outros casos.} \end{cases} \quad (4.3.3)$$

Tanto por aplicação da condição necessária de otimalidade de primeira ordem, como por análise direta, podemos verificar que, se x é minimizador local ou global de (4.3.1), teremos

$$\bar{g}_P(x) = 0\ . \quad (4.3.4)$$

Se $G \geq 0$ a quadrática é convexa e (4.3.4) passa a ser uma condição suficiente para minimizador global.
Quando restringimos a função quadrática a uma face aberta F_I, as variáveis livres são apenas as que se encontram estritamente entre os limites definidos pelo conjunto I. O vetor definido a seguir é o inverso aditivo do gradiente em relação a essas variáveis livres. Assim, para cada $x \in F_I$ definimos $\bar{g}_I(x) \in I\!R^n$ como

$$\bar{g}_I(x)_i = \begin{cases} 0 & \text{se } i \in I \text{ ou } n+i \in I \\ -[\nabla q(x)]_i & \text{nos outros casos.} \end{cases} \tag{4.3.5}$$

Observamos que $\bar{g}_I(x)$ é a projeção ortogonal de $-\nabla q(x)$ em $S(F_I)$. Também podemos interpretar $\bar{g}_I(x)$ como "a componente" de $\bar{g}_P(x)$ no subespaço $S(F_I)$. Naturalmente, $\bar{g}_P(x)$ tem uma segunda componente, ortogonal a $S(F_I)$, que chamamos "gradiente chopado" e denotamos por $\bar{g}_I^C(x)$. Dessa maneira, para cada $x \in F_I$,

$$\bar{g}_I^C(x)_i = \begin{cases} 0 & \text{se } i \notin I \text{ e } n+i \notin I \\ 0 & \text{se } i \in I \text{ e } [\nabla q(x)]_i > 0 \\ 0 & \text{se } n+i \in I \text{ e } [\nabla q(x)]_i < 0 \\ -[\nabla q(x)]_i & \text{nos outros casos.} \end{cases} \tag{4.3.6}$$

Como mencionamos acima, é fácil ver que, para todo $x \in F_I$, o gradiente interno $\bar{g}_I(x)$ é ortogonal ao gradiente chopado, e

$$\bar{g}_P(x) = \bar{g}_I(x) + \bar{g}_I^C(x) \ .$$

O algoritmo para minimizar quadráticas em caixas que apresentaremos produz uma seqüência $\{x_k\}$ de aproximações da solução de (4.3.1) baseada na minimização parcial da quadrática nas diferentes faces visitadas. Quando x_k pertence a uma face F_I, um "algoritmo interno" para minimização de quadráticas irrestritas será acionado, trabalhando apenas com as variáveis livres da face. A suposição básica será que esse algoritmo é "convergente" no sentido de que ele produz, em um número finito de passos um ponto externo a Ω (mas pertencente, naturalmente, a $V(F_I)$), ou que todo ponto limite do algoritmo é um ponto estacionário do problema, essencialmente irrestrito, de minimizar $q(x)$ sujeita a $x \in V(F_I)$. Em outras palavras, o algoritmo interno encontra um ponto estacionário restrito a F_I ou viola as restrições inativas dessa face. Em cada passo do algoritmo interno, verificamos se ele já está bastante perto de um ponto estacionário em F_I. Para isso, comparamos o tamanho do gradiente chopado com o tamanho do gradiente projetado.

Se o quociente entre ambos é grande (o valor máximo é 1), significa que o gradiente interno é pequeno em relação ao gradiente chopado e, portanto, continuar explorando a face F_I é pouco econômico, ou seja, abandonar as cotas que estão ativas em F_I parece mais razoável. Isso é feito usando a direção do gradiente chopado. Veremos que a seqüência de pontos assim definida é "convergente" a pontos estacionários de (4.3.1), que são soluções do problema no caso convexo. Este algoritmo é, essencialmente, o definido em [12], com antecedentes nos trabalhos [79], [78], [82], [81], [149].

Provavelmente, agora o leitor percebe mais claramente nosso interesse na propriedade (4.1.7), ou em propriedades análogas. Como o algoritmo irrestrito usado em F_I tem um papel essencial no desempenho do método principal desta seção, vamos estabelecer rigorosamente quais devem ser suas características.

Diremos que um algoritmo para minimizar $q(x)$ em $V(F_I)$ (problema, essencialmente, irrestrito) tem as *propriedades boas para a minimização em caixas* quando produz uma seqüência $\{z_0, z_1, z_2, \ldots\} \subset V(F_I)$, $z_0 \in F_I$ (talvez finita) que cumpre o seguinte:

(a) Se z_k e z_{k+1} estão definidos, então $q(z_{k+1}) < q(z_k)$.

(b) Se z_{k+1} não está definido (a seqüência termina em z_k) isto pode ser devido a dois motivos: z_k é um ponto estacionário da minimização de $q(x)$ em $V(F_I)$ ou foi encontrada uma direção d_k tal que

$$\lim_{t\to\infty} q(z_k + td_k) = -\infty.$$

Neste caso, se $z_k + td_k \in \Omega$ para todo t, a inexistência de solução de (4.3.1) fica caracterizada. Se, pelo contrário, $z_k + td_k \notin \Omega$ para t grande, escolhe-se um "último" $z_{k+1} = z_k + td_k \in \Omega$ tal que $q(z_{k+1}) < q(z_k)$ e dá-se por terminada a seqüência gerada pelo algoritmo interno em z_{k+1}.

(c) Se a seqüência $\{z_k\}$ é infinita, então todo ponto limite da mesma é um ponto estacionário q sujeita a $V(F_I)$. Se não existem pontos limite (logo $\|z_k\| \to \infty$) deve-se satisfazer

$$\lim_{k\to\infty} q(z_k) = -\infty.$$

Vejamos que os algoritmos para minimizar quadráticas sem restrições que estudamos na seção 4.1 satisfazem essas condições. O método direto, baseado na fatoração de Cholesky da matriz G "reduzida" (as variáveis correspondentes às restrições ativas em F_I estão fixas) encontra o minimizador de Q em $V(F_I)$ em um passo, se a quadrática q restrita a $V(F_I)$ é estritamente convexa (a Hessiana reduzida é definida positiva). Portanto, satisfaz claramente (a) e (b) e a hipótese de (c) é vazia porque a seqüência termina em z_1. Quando a Hessiana reduzida não é definida positiva, a fatoração de Cholesky não poderá ser completada. Suponhamos que a fatoração espectral é viável. Nesse caso, já vimos que podemos obter um minimizador irrestrito, quando existe, ou uma direção que satisfaz (4.1.7), portanto, o algoritmo que combina fatoração de Cholesky com decomposição espectral satisfaz as condições acima. Se a fatoração espectral é inviável, podemos usar a fatoração Bunch-Parlett, ou resolver a seqüência de problemas

$$\text{Minimizar } q(z) \text{ sujeita a } z \in V(F_I), \|z - z_k\| \leq \Delta \tag{4.3.7}$$

para Δ grande, usando o método de Moré e Sorensen comentado na seção 4.2, que usa apenas fatorações de Cholesky de matrizes definidas positivas. Se z_k é solução de (4.3.7), então z_k é minimizador de q restrita a $V(F_I)$ e o algoritmo pára. Se (4.3.7) gera uma seqüência infinita, teremos que todo ponto de acumulação da mesma é estacionário de q em $V(F_I)$, ou os valores de $q(x_k)$ tendem a $-\infty$ (exercício para o leitor). Em qualquer caso, as condições (a), (b) e (c) se satisfazem.
As propriedades do método dos gradientes conjugados, para minimizar q em $V(F_I)$ foram estudadas na seção 4.1. Vimos que esse método termina em um ponto estacionário em um número finito de passos ou gera uma direção ao longo da qual a quadrática tende a $-\infty$. Portanto, satisfaz as condições (a), (b) e (c). Em [12] são estudados outros métodos iterativos que satisfazem essas condições em determinadas circunstâncias.

Agora podemos definir o algoritmo para minimizar quadráticas em caixas, com um alto grau de liberdade, devido à flexibilidade na escolha do algoritmo interno a F_I. De fato, observemos que nada obriga a que o mesmo algoritmo interno seja utilizado em todas as caixas. Por exemplo, como observado em [12], diferentes algoritmos podem ser usados em diferentes faces, tendo em conta a dimensão da mesma.

Algorithm 4.3.1 - Minimização de quadráticas em caixas.
Seja $\eta \in (0,1)$ dado indepentemente de k, e $x_0 \in \Omega$ um ponto inicial arbitrário. O algoritmo define uma seqüência $\{x_k\}$ em Ω e pára se $\|\bar{g}_P(x_k)\| = 0$. Suponhamos que $x_k \in \Omega$ é tal que $\|\bar{g}_P(x_k)\| \neq 0$. Seja $I = I(x_k)$ tal que $x_k \in F_I$. Chamemos $\Phi(x) \in \Omega$ ao minimizador de q ao longo do segmento (talvez semi-reta) $\{x \in \Omega \mid x = x_k + t\bar{g}_I^C(x_k), t \geq 0\}$. Os seguintes passos definem o procedimento para encontrar x^{k+1}.

Passo 1: Começando com $z_0 = x_k$, usar um método com as "propriedades boas para minimização de quadráticas em caixas" aplicado ao problema essencialmente irrestrito de minimizar $q(x)$ em $V(F_I)$, obtendo assim $z_0 = x_k, z_1 = x_{k+1}, \ldots$. Interromper esse método quando x_k satisfaz uma das seguintes condições:

(a)

$$x_k \in \Omega \text{ e } \|\bar{g}_P(x_k)\| = 0; \tag{4.3.8}$$

(b) O método interno detectou que (4.3.1) é ilimitado inferiormente.

(c)

$$\|\bar{g}_I^C(x_k)\| > \eta \, \|\bar{g}_P(x_k)\|; \tag{4.3.9}$$

(d)

$$z_{\nu+1} \notin \Omega. \tag{4.3.10}$$

Passo 2: Se o método interno foi interrompido por (4.3.8), parar (x_k é um ponto estacionário de (4.3.1)). Se o método interno detecta que (4.3.1) não tem solução, o algoritmo principal é interrompido com esse mesmo diagnóstico.

Passo 3: Se o teste (4.3.9) foi satisfeito em x_k, e q não é limitada inferiormente no segmento (nesse caso, necessariamente, semi-reta) $\{x+t\bar{g}_I^C(x_k), t \geq 0\}$ o problema (4.3.1) não tem solução. Nesse caso, parar. Em caso contrário, calcular $x_{k+1} = \Phi(x_k)$.

Passo 4: Se $x_k = z_\nu$ e $z_{\nu+1}$ viola os limites de $\bar{F}_I$ (condição (4.3.10)), encontrar x_{k+1} na fronteira de F_I ($\bar{F}_I - F_I$) tal que $q(x_{k+1}) < q(x_k)$ ou detectar que o problema (4.1.3) não tem solução.

Comprovar que o Algoritmo 4.3.1 está bem definido consiste em provar que o Passo 4 é possível. Pelas propriedades do algoritmo interno, temos que $q(z_{\nu+1}) < q(x_k)$. Agora, $\phi(t) \equiv q(x_k + t(z_{\nu+1} - x_k))$ é uma parábola como função de t. Logo, $\phi(t)$ decresce em forma monótona entre $t = 0$ e $t = 1$, ou $\phi(t)$ é estritamente crescente para $t < 0$. No primeiro caso, avançando desde $t = 0$, no sentido positivo, até a fronteira, encontramos um ponto onde a quadrática diminui de valor. Na segunda situação ocorre essencialmente o mesmo, avançando no sentido negativo de t. Nos dois casos, o ponto encontrado está na reta determinada por x_k e $z_{\nu+1}$. Em algoritmos práticos, o ponto da fronteira encontrado será, via de regra, melhor que o definido neste parágrafo.

No seguinte lema vamos considerar a situação em que a condição (4.3.9) é satisfeita e $\Phi(x_k)$ existe, ou seja, pelo menos neste passo não é detectada a eventualidade de que a quadrática seja ilimitada inferiormente, e x_{k+1} é definido como sendo $\Phi(x_k)$. Essencialmente, mostraremos que o decréscimo obtido de x_k até x_{k+1} é proporcional à norma de $\bar{g}_P(x_k)$.

Lema 4.3.2
Se $x_{k+1} = \Phi(x_k)$ é obtido no Passo 3 do Algoritmo 4.3.1, então

$$q(x_k) - q(x_{k+1}) \geq \min\{\frac{\eta\,\gamma}{2}\|\bar{g}_P(x_k)\| \,,\, \frac{\eta^2}{2L}\|\bar{g}_P(x_k)\|^2\} \,.$$

Prova: Como o teste (4.3.9) é satisfeito, então $\bar{g}_I^C(x_k) \neq 0$. Portanto, $x_k + t\bar{g}_I^C(x_k) \in \Omega$ para todo $t \in [0, \tilde{t}]$, onde $\tilde{t} = \gamma/\|\bar{g}_I^C(x_k)\|$. Consideremos a quadrática unidimensional definida por

$$\phi(t) = q(x_k + t\bar{g}_I^C(x_k)) = q(x_k) + t\nabla q(x_k)^T\bar{g}_I^C(x_k) + \frac{1}{2}t^2\bar{g}_I^C(x_k)^T G\bar{g}_I^C(x_k) \,.$$

Se $\bar{g}_I^C(x_k)^T G\bar{g}_I^C(x_k) > 0$ então o único minimizador irrestrito de $\phi(t)$ é dado por

$$t^* = \frac{\|\bar{g}_I^C(x_k)\|^2}{\bar{g}_I^C(x_k)^T G\bar{g}_I^C(x_k)} \,.$$

Se $x_k + t^*\bar{g}_I^C(x_k)$ não está em Ω, então $x_{k+1} = \Phi(x_k)$ é realizado para algum $\bar{t}$ tal que $\tilde{t} \leq \bar{t} < t^*$, e

$$q(x_k + \tilde{t}\bar{g}_I^C(x_k) \geq q(x_k + \bar{t}\bar{g}_I^C(x_k)) \,. \tag{4.3.11}$$

Substituindo $\tilde{t}$ em $\phi(t)$, obtemos

$$\phi(\tilde{t}) = q(x_k) - \gamma\|\bar{g}_I^C(x_k)\| + \frac{\gamma^2\;\bar{g}_I^C(x_k)G\bar{g}_I^C(x_k)}{2\;\|\bar{g}_I^C(x_k)\|^2} \,. \tag{4.3.12}$$

Usando (4.3.12) e o fato de que $t^* > \tilde{t}$, segue-se que

$$q(x_k + \tilde{t}\bar{g}_I^C(x_k)) - q(x_k) < -\frac{\gamma}{2}\|\bar{g}_I^C(x_k)\| \,. \tag{4.3.13}$$

Combinando (4.3.11) e (4.3.13), temos

$$q(x_k) - q(x_{k+1}) > \frac{\gamma}{2}\|\bar{g}_I^C(x_k)\| > \frac{\eta\,\gamma}{2}\|\bar{g}_P(x_k)\| \,. \tag{4.3.14}$$

Agora, se $x_k + t^* \bar{g}_I^C(x_k)$ está em Ω, então esse ponto é x_{k+1} e obtemos

$$q(x_{k+1}) - q(x_k) = -\frac{\|\bar{g}_I^C(x_k)\|^4}{2\,\bar{g}_I^C(x_k)^T G \bar{g}_I^C(x_k)} \ . \tag{4.3.15}$$

Portanto, usando (4.3.2) e (4.3.15), temos:

$$q(x_k) - q(x_{k+1}) > \frac{1}{2L}\|\bar{g}_I^C(x_k)\|^2 > \frac{\eta^2}{2L}\|\bar{g}_P(x_k)\|^2 \ . \tag{4.3.16}$$

Analisemos agora a situação em que $\bar{g}_I^C(x_k)^T G \bar{g}_I^C(x_k) \le 0$. Nesse caso,

$$\phi(t) \le q(x_k) + t\nabla q(x_k)^T \bar{g}_I^C(x_k) \ ,$$

e $q(x_{k+1}) < \phi(\tilde{t}) \le q(x_k) - \gamma\|\bar{g}_I^C(x_k)\|$. Portanto,

$$q(x_k) - q(x_{k+1}) > \gamma\|\bar{g}_I^C(x_k)\| > \eta\gamma\|\bar{g}_P(x_k)\| \ . \tag{4.3.17}$$

Resumindo, existem três casos possíveis: $x_k + t^* \bar{g}_I^C(x_k)$ factível, ou infactível, ou $\bar{g}_I^C(x_k)^T G \bar{g}_I^C(x_k) \le 0$. Em cada caso obtemos, respectivamente, (4.3.14), (4.3.16) e (4.3.17), o que implica a tese. **QED**

Em continuação, provamos a "convergência global" do Algoritmo 4.3.1. Lembramos primeiro as condições nas quais o algoritmo pára, isto é, gera uma seqüência finita: quando encontra um ponto estacionário x_k de (4.3.1) ou quando detecta que o problema é ilimitado inferiormente, e, portanto, sem solução. Basicamente, provaremos que, se o algoritmo gera uma seqüência infinita, haverá, essencialmente, as mesmas duas possibilidades: encontraremos um gradiente projetado arbitrariamente pequeno, ou a seqüência dos valores funcionais em x_k tenderá a $-\infty$.

Teorema 4.3.3
Suponhamos que o Algoritmo 4.3.1 gera uma seqüência infinita $\{x_k\}$. Então, existem duas possibilidades:

$$\liminf_{k\to\infty} \|\bar{g}_P(x_k)\| = 0 \tag{4.3.18}$$

e

$$\lim_{k\to\infty} q(x_k) = -\infty. \tag{4.3.19}$$

Proof. Suponhamos que (4.3.18) não se cumpre. Portanto, existe $\epsilon > 0$ tal que

$$\|\bar{g}_P(x_k)\| > \epsilon \ \text{ para todo } \ k \ . \tag{4.3.20}$$

Consideramos dois casos:
(a) A condição (4.3.9) é satisfeita em um número finito de iterações.
(b) Existe um conjunto infinito de índices $K_1 \subset \mathcal{N}$ tal que (4.3.9) é satisfeita para todo $k \in K_1$.
Se (a) vale, então existe k_0 tal que $x_k \in F_I$ para um I fixo, e para todo $k \geq k_0$. Portanto, a seqüência é gerada pelo algoritmo interno para todo $k \geq k_0$. Pelas propriedades do algoritmo interno, temos que, se $\|x_k\| \to \infty$, vale (4.3.19). Se pelo contrário, $\{x_k\}$ admite uma subseqüência limitada e convergente, $\{x_k\}_{k \in K_2}$, devemos ter

$$\lim_{k \in K_2} \|g_I(x_k)\| = 0.$$

Agora, como (4.3.9) não se satisfaz para nenhum $k \in K_2$, necessariamente $\|g_I^C(x_k)\|$ e $\|g_P(x_k)\|$ também tendem a 0 para $k \in K_2$, o que contradiz (4.3.20). Portanto, a tese do teorema fica provada no caso (a).
Suponhamos agora que vale (b). Seja k_j o j-ésimo índice de K_1, $j \in \mathcal{N}$. Usando (4.3.20), o Lema 4.3.2 e o fato de que $\{q(x_k)\}$ é monotonicamente decrescente, obtemos

$$\begin{aligned} q(x_{k_j}) - q(x_{k_1}) &= \sum_{l=k_1}^{k_j-1} (q(x_{l+1}) - q(x_l)) \\ &\leq \sum_{l \in K_1,\, l=k_1}^{k_j-1} (q(x_{l+1}) - q(x_l)) \\ &\leq \sum_{l \in K_1,\, l=k_1}^{k_j-1} - \min\{\frac{\eta\,\gamma}{2}\|\bar{g}_P(x_l)\|,\ \frac{\eta}{2L}\|\bar{g}_P(x_l)\|^2\} \\ &< -j\ \min\{\frac{\eta\,\gamma}{2}\epsilon,\ \frac{\eta}{2L}\epsilon^2\} \end{aligned} \tag{4.3.21}$$

Usando (4.3.21) concluimos que, neste caso,

$$\lim_{j \to \infty} q(x_{k_j}) = -\infty\ .$$

Portanto, o teorema está provado. **QED**

Examinemos algumas conseqüências do resultado provado no Teorema 4.3.3. Se a seqüência gerada pelo algoritmo é limitada, o que, sem dúvida, acontecerá, por exemplo, quando os limitantes l_i e u_i não assumem valores infinitos, a possibilidade de que a sequência $\{q(x_k)\}$ seja ilimitada inferiormente deve ser excluída. Portanto, nesse caso, temos uma subseqüência $\{x_k\}_{k \in K_1}$ onde os gradientes projetados tendem a 0. Por compacidade, essa subseqüência tem, por sua vez, uma subseqüência convergente. Consideremos agora qualquer subseqüência convergente $\{x_k\}_{k \in K_2}$, com limite, digamos, $x_* \in F_I$. Se $l_i < [x_*]_i < u_i$, segue-se que $l_i < [x_k]_i < u_i$ para todo $k \in K_2$ suficientemente grande. Portanto, a $i-$ésima derivada parcial de q em x_k tende a 0 e, conseqüentemente, $[\nabla q(x_*)]_i = 0$. Se $[x_*]_i = l_i$, teremos que $[x_k]_i \geq l_i$ para todo $k \in K_2$ suficientemente grande, digamos $k \geq k_0$. Definimos $K_3 = \{k \in K_2 \mid [x_k]_i > l_i, k \geq k_0\}$ e $K_4 = \{k \in K_2 \mid [x_k]_i = l_i, k \geq k_0\}$. Claramente, pelo menos um desses conjuntos é infinito. Se K_3 é infinito, teremos que $[-\nabla q(x_k)]_i = [\bar{g}_P(x_k)]_i \to 0$ para $k \in K_3$, portanto $[\nabla q(x_*)]_i = 0$. Se K_4 é infinito, teremos que $\min\,\{0, [\nabla q(x_k)]_i\} \to 0$ para $k \in K_3$, logo $\min\,\{0, [\nabla q(x_*)]_i\} \to 0$. Portanto, em todos os casos chegamos à conclusão que $\bar{g}_P(x_*) = 0$, ou seja, todo ponto limite é um ponto estacionário.

No caso convexo, a situação é mais favorável ainda. Com efeito, como a seqüência $\{q(x_k)\}$ é monótona decrescente os valores de q em todos os pontos limite são iguais. Assim da existência de um ponto limite estacionário (logo, minimizador global) se infere que todos os pontos limite são minimizadores globais. Naturalmente, quando a Hessiana é definida positiva, o minimizador global é único, e a seqüência $\{x_k\}$ completa converge a ele. Outras propriedades deste algoritmo relacionadas com a "degeneração dual" são estudadas em [12].

A eficiência do Algoritmo 4.3.1 em problemas de grande porte está relacionada com a possibilidade de acrescentar ou eliminar em poucas iterações uma grande quantidade de canalizações ativas. A eliminação de canalizações se dá quando a condição (4.3.9) é satisfeita. Quanto menor seja a tolerância η, mais impaciente será o algoritmo com a face na qual está trabalhando, e tratará de sair dela rapidamente. Pelo contrário, se η é próximo de 1, a tendência será sair da face depois de esgotá-la totalmente, inibindo qualquer possibilidade de retorno à mesma. Para problemas grandes, valores pequenos de η são recomendáveis. Por outro lado, as canalizações se acrescentam quando o algoritmo interno fica infactível e se faz necessário achar um ponto na fronteira. No parágrafo anterior onde discutimos este assunto, mostramos que existe um ponto da fronteira com as características desejáveis, mas adiantamos que o indicado não era o melhor possível. De fato, no ponto então definido, via de regra, apenas uma restrição ativa é acrescentada, em relação ao ponto x_k. Uma estratégia mais ávida por restrições ativas se baseia em *buscas projetadas*. A idéia é seguinte: suponhamos que z seja o ponto infactível produzido pelo algoritmo interno. Em vez de considerar a reta

definida por x_k e z, projetamos z na caixa Ω, obtendo, digamos $z^{(1)}$. Este ponto projetado terá como ativas todas as canalizações que eram violadas em z, que são, em geral, mais que as ativas no ponto do segmento que une x_k com z. Assim, testamos se $q(z^{(1)}) < q(x_k)$ e o aceitamos como x_{k+1} em caso positivo. Senão, substituímos z por, digamos, $x_k + (z - x_k)/2$ e repetimos o processo. Se as reduções se repetem tantas vezes que o ponto z fica pertencendo a Ω, nos conformamos com o ponto fronteira da reta $[x_k, z]$, que, como vimos, satisfaz pelo menos a condição requerida para convergência.
Uma última observação é a seguinte. O esquema do Algoritmo 4.3.1 é válido tanto para problemas de grande como de pequeno porte. A diferença entre uns e outros radica apenas na escolha do algoritmo interno. Quando o problema é pequeno, e são usadas fatorações de Cholesky, é fácil ver que o cálculo de x_{k+1} no caso em que (4.3.9) se verifica é quase sempre irrelevante, já que, independentemente de $\Phi(x_k)$, na maioria dos casos x_{k+2} será o mesmo. Mas isto é uma sutileza da qual não precisamos nos ocupar no momento.

Chapter 5

Sistemas de equações não-lineares

As condições de otimalidade de primeira ordem dos problemas de otimização são sistemas não lineares, onde as incógnitas são as variáveis do problema e, às vezes, também os multiplicadores de Lagrange. Alèm disso, quando se trata de minimização com restrições de desigualdade, apenas as soluções que satisfazem determinadas inequações são úteis. Portanto, de certo modo, a arte da otimização está incluída na arte de resolver sistemas não lineares. Por outro lado, quando $F(x) = 0$ ($F : I\!R^n \to I\!R^n$) é resolúvel, encontrar as raízes desse sistema é equivalente a achar o minimizador global de $\|F(x)\|$ onde $\|\cdot\|$ é uma norma qualquer em $I\!R^n$. Desse ponto de vista, a resolução de sistemas não lineares pode ser considerada um caso particular da otimização.

Entretanto, os problemas de otimização tem muita estrutura adicional, o que justifica a introdução de métodos específicos, que transcendem a mera aplicação de algoritmos para resolver sistemas. Com efeito, nas condições necessárias de primeira ordem, apenas as derivadas do problema estão representadas, e não, por exemplo, a função objetivo original. Como conseqüência, os métodos para sistemas não lineares, quando aplicados às condições de otimalidade, tem dificuldades em diferenciar minimizadores de maximizadores já que, freqüentemente, as condições de otimalidade para ambos tipos de extremos são as mesmas. Por outro lado, quando $F(x) = 0$ é transformado em um problema de otimização através da norma da função vetorial, aparecem estruturas próprias do sistema, como o fato da função objetivo ser, geralmente, uma soma de quadrados.

Muitos problemas práticos de física, engenharia, economia e outras ciências são modelados de maneira muito conveniente por sistemas não lineares. É

usual, nesses casos, que alguma versão moderna de um velho algoritmo, o método de Newton, seja usada com sucesso. Esse método, como outros que veremos neste capítulo, é, na sua forma básica, um método iterativo *local*, no sentido de que podemos garantir, apenas, a convergência a uma solução supondo que o ponto inicial usado como aproximação da mesma já é suficientemente bom. A praticidade desses métodos radica em que, geralmente, a visão teórica que exige um ponto inicial muito bom é excessivamente pessimista e, em muitos casos, os métodos locais convergem mesmo se a aproximação inicial não é boa. Um caso extremo é quando o sistema não linear é, de fato, linear, e o método de Newton encontra a solução em uma iteração, independentemente do ponto inicial.
Nos métodos locais para sistemas não lineares encontramos os germes para muitos algoritmos de otimização. Essa é a principal motivação para seu estudo independente neste livro. Algumas afirmações básicas, do tipo "o método de Newton tem convergência quadrática" ou "os métodos quase-Newton são superlineares" formam parte tanto do folclore de otimização quanto de resolução de sistemas. Aqui veremos, com certo rigor, em que condições tais afirmações são válidas.
Neste capítulo, nosso problema será, sempre, resolver

$$F(x) = 0 \ , \quad F : I\!R^n \to I\!R^n \ , \quad F \in C^1(I\!R^n) \ .$$

Utilizaremos a seguinte notação para a função F e para a matriz Jacobiana J:

$$F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} \ \text{ e } \ J(x) = F'(x) = \begin{pmatrix} f_1'(x) \\ \vdots \\ f_n'(x) \end{pmatrix} = \begin{pmatrix} \nabla f_1^T(x) \\ \vdots \\ \nabla f_n^T(x) \end{pmatrix} .$$

5.1 O método de Newton

Em todos os cursos elementares de cálculo numérico, estuda-se o método de Newton (também conhecido como Newton-Raphson) no contexto de achar zeros de funções. Sua generalização para sistemas foi proposta pela primeira vez não por Newton, mas por Simpson, eminente matemático do século XVIII (ver [206]).
O princípio em que se baseia o método é paradigmático na resolução aproximada de problemas matemáticos: o objetivo final é um problema "difícil"

(neste caso $F(x) = 0$), a solução do qual vai sendo aproximada por uma seqüência de pontos $\{x_k\}$. Dada cada aproximação x_k, constrói-se, com a informação disponível nesse ponto, um problema "fácil", que sabemos resolver. A aproximação x_{k+1} é a solução do problema fácil. O problema fácil muda de uma iteração para a seguinte e, via de regra, sua solução está cada vez mais próxima da solução do problema difícil original.

No nosso problema atual, o $k-$ésimo problema fácil vem de considerar a aproximação de Taylor de primeira ordem de $F(x)$, numa vizinhança do ponto atual x_k:

$$F(x) \approx L_k(x) = F(x_k) + J(x_k)(x - x_k)\,. \tag{5.1.1}$$

Seguindo o princípio descrito acima, o ponto seguinte x_{k+1} é uma solução de

$$L_k(x) = 0\ . \tag{5.1.2}$$

Se $J(x_k)$ é não-singular, (5.1.2) tem solução única, e então a iteração Newton consiste em resolver um sistema linear:

$$\begin{aligned} J(x_k)s_k &= -F(x_k) \\ x_{k+1} &= x_k + s_k\ . \end{aligned} \tag{5.1.3}$$

A implementação de (5.1.3) pressupõe o cálculo de $J(x_k)$, isto é, a avaliação das derivadas primeiras das funções $f_i(x),\ i = 1, \ldots, n$. Até poucos anos atrás, o cálculo de derivadas era considerado não só difícil mas também muito suscetível a erros humanos. Atualmente, a possibilidade de falha humana pode ser evitada, através das diferenciações simbólica e automática. É importante ressaltar que, em geral, quando se calculam efetivamente as derivadas, muitos cálculos usados na avaliação da função podem ser reaproveitados. A diferenciação automática é um conjunto de técnicas que produz um programa que avalia $F(x)$ e $J(x)$, com os reaproveitamentos necessários, partindo de um programa que avalia apenas $F(x)$. Ver, por exemplo, [105].

O método de Newton possui uma propriedade única entre os algoritmos para resolver sistemas: a invariância por mudanças de coordenadas, tanto no espaço domínio quanto no contra-domínio. No contra-domínio, isto significa que as iterações de Newton aplicadas a $F(x) = 0$ são as mesmas que as aplicadas ao sistema $AF(x) = 0$, para qualquer matriz A não-singular. A invariância no domínio consiste em que, se $\{x_k\}$ é a seqüência newtoniana para $F(x) = 0$, então os iterandos para o sistema $F(Ax + b) = 0$, com A

não singular e com a aproximação inicial $Ax_0 + b$, são os pontos da forma $Ax_k + b$.
Uma variação de (5.1.3) com praticamente as mesmas propriedades teóricas e práticas que evita o enfadonho cálculo de derivadas é o chamado "método de Newton discreto". O esquema desse método é o descrito em (5.1.3) com a exceção de que as derivadas consideradas não são as analíticas mas suas aproximações por diferenças finitas. Mais precisamente, a coluna j de $J(x_k)$ é substituída por $[F(x_k + he_j) - F(x_k)]/h$, onde h é um passo (de discretização) pequeno e $\{e_1, \ldots, e_n\}$ é a base canônica de $\mathbb{R}^n$. A implementação de uma iteração do método de Newton discreto, embora não exija o cálculo de derivadas, demanda a avaliação da função F em $n + 1$ pontos. Isto pode ser bastante caro computacionalmente, por isso, sempre que possível, as derivadas analíticas devem ser utilizadas.
A resolução do sistema linear (5.1.3) quando a matriz Jacobiana é não-singular pode ser obtida via fatoração LU (variação da clássica eliminação gaussiana), com um custo de $O(\frac{n^3}{3})$ operações. Caso $J(x_k)$ seja singular, deve-se adotar alguma estratégia especial, para não inibir o prosseguimento do método. Ver, por exemplo, [101].
Portanto, o trabalho realizado em uma iteração do método de Newton consiste na avaliação de F em x_k e suas derivadas, mais as $O(\frac{n^3}{3})$ operações necessárias para resolver (5.1.3). O termo em n^3 cresce de maneira dramática com o aumento do porte do problema. Felizmente, em muitos problemas grandes é viável o uso de técnicas de fatoração LU esparsa, utilizando-se estruturas de dados adequadas com previsão de possíveis preenchimentos. Por exemplo, se $J(x_k)$ tem estrutura tridiagonal, sua fatoração e a resolução de sistema correspondente podem ser efetuadas com $O(n)$ operações. Para outros problemas de grande porte, no entanto, o método de Newton pode se tornar inviável.

5.2 Métodos quase-Newton

Se aceitamos a idéia de que o método de Newton é "bom" mas "caro", parece natural a introdução de métodos "quase tão bons" quanto Newton, mas "bem mais baratos". A maioria dos métodos quase-Newton foi estabelecida com esses objetivos. Para ser (quase) tão bons como Newton, esses métodos devem ser parecidos com seu arquétipo sob vários pontos de vista. Por isso, definiremos como métodos quase-Newton aqueles algoritmos para sistemas não lineares cuja iteração tem o seguinte formato:

$$\begin{array}{rcl} B_k s_k & = & -F(x_k) \\ x_{k+1} & = & x_k + s_k. \end{array} \tag{5.2.1}$$

Assim, o próprio método de Newton é um método quase-Newton, o que é esteticamente agradável. Entretanto, métodos quase-Newton práticos serão apenas aqueles em que B_{k+1}^{-1} possa ser obtida facilmente a partir de B_k^{-1}, isto é, com com não mais de $O(n^2)$ operações. Dessa maneira, os cálculos em (5.2.1) poderão ser efetuados com um custo de $O(n^2)$ em termos de tempo por iteração. Algumas implementações de métodos quase-Newton trabalham com fatorações das matrizes B_k, e não com suas inversas. Nesses casos, mostra-se que a fatoração de B_{k+1} pode ser obtida a partir da fatoração de B_k em tempo proporcional a n^2.

Desta forma, vemos que o esforço computacional $O(\frac{n^3}{3})$ empregado por Newton diminui para $O(n^2)$ quando se utilizam métodos quase-Newton adequados. Infelizmente, esta redução nos custos é paga com redução na velocidade de convergência, conforme veremos na seção 5.4.

O método quase-Newton mais simples é o chamado *método de Newton estacionário*, que se obtém fixando $B_k \equiv J(x_0)$. Outra variação bastante tradicional é o *método de Newton estacionário com recomeços a cada m iterações*: Fixado um inteiro m, se k é múltiplo de m, tomamos $B_k = J(x_k)$. Senão, $B_k = B_{k-1}$. Com o objetivo de estabelecer um compromisso entre a eficiência do método de Newton e o baixo custo do método de Newton estacionário, existem estudos teóricos para encontrar o m ótimo no caso de problemas específicos (ver [184]).

Uma outra família de métodos obedecendo a filosofia quase-Newton é a dos *métodos secantes*. Assim como o método de Newton é a generalização para sistemas do algoritmo com o mesmo nome para achar zeros de funções, os métodos secantes são as generalizações dos algoritmos assim denominados para o problema unidimensional. Pensemos, como antes, que na iteração k a função $F(x)$ é aproximada por $L_k(x) = F(x_k) + B_k(x - x_k)$. Escrevendo o mesmo tipo de aproximação para a iteração $k+1$, temos

$$F(x) \approx L_{k+1}(x) = F(x_{k+1}) + B_{k+1}(x - x_{k+1}).$$

A idéia secante consiste em impor que função linear $L_{k+1}(x)$ interpole a função verdadeira nos pontos x_{k+1} e x_k. Em outras palavras,

$$L_{k+1}(x_{k+1}) = F(x_{k+1}) \text{ e } L_{k+1}(x_k) = F(x_k).$$

A condição $L_{k+1}(x_{k+1}) = F(x_{k+1})$ é automaticamente satisfeita pela definição de L_{k+1}. Quanto à condição $L_{k+1}(x_k) = F(x_k)$, podemos ver que é equivalente a

$$F(x_k) = F(x_{k+1}) + B_{k+1}(x_k - x_{k+1}),$$

ou

$$B_{k+1} s_k = y_k \ , \tag{5.2.2}$$

onde $\quad y_k = F(x_{k+1}) - F(x_k)$.
A equação () é chamada *equação secante* por motivos agora óbvios. Podemos pensar () como um sistema linear cuja incógnita é a matriz. Assim interpretado, o sistema tem n^2 variáveis (as entradas de B_{k+1}) e apenas n equações. Portanto, somente no caso $n = 1$ o sistema poderá ter solução única. Se $n > 1$ e $s_k \neq 0$ haverá infinitas matrizes B (uma variedade afim em $\mathbb{R}^{n \times n}$) que satisfazem $Bs_k = y_k$. Diferentes escolhas dessa matriz definem diferentes métodos secantes. Por exemplo, se procuramos B_{k+1} de maneira que a diferença $\Delta B_k \equiv B_{k+1} - B_k$ seja uma matriz de posto unitário, teremos, por (),

$$\Delta B_k s_k = y_k - B_k s_k$$

e poderemos tomar

$$\Delta B_k = \frac{(y_k - B_k s_k) w_k^T}{w_k^T s_k}$$

com $w_k \in \mathbb{R}^n$ arbitrário e não ortogonal a s_k.
A escolha $w_k = s_k$ define o *primeiro método de Broyden.* Se $w_k = y_k - B_k s_k$, o método é conhecido como *correção simétrica de posto um.*
O interessante neste tipo de correção é que B_{k+1}^{-1} também pode ser obtida a partir de B_k^{-1} mediante uma correção de posto um. A expressão para esta correção pode ser calculada usando-se a fórmula de Sherman-Morrison [96], com um custo, facilmente verificável, da ordem de $O(n^2)$ operações.

O fato de que $B_{k+1}^{-1} - B_k^{-1}$ seja uma matriz da forma $u_k v_k^T$ faz com que toda a informação relativa a B_{k+1}^{-1} esteja contida em B_0^{-1}, e nos vetores $u_0, v_0, u_1, v_1, \ldots, u_k, v_k$. (Veremos isso com detalhe no exercício 5.2.) Logo, se B_0 é uma matriz suficientemente simples, de tal forma que a informação relativa a sua inversa ou sua fatoração LU seja armazenável em poucas posições de memória (digamos, $O(n)$), toda a informação necessária para multiplicar B_{k+1}^{-1} por um vetor ocupa $O(kn)$ posições, e o citado produto pode ser efetuado com $O(kn)$ operações. Essa observação fornece os elementos para a utilização de métodos secantes em problemas de grande porte.

De fato, enquanto k é pequeno, o custo da iteração quase-newtoniana é, essencialmente, $O(n)$ e, com sorte, poucas iterações serão suficientes para atingir a convergência, de maneira que k, muitas vezes, não chega a ser grande. Se o índice da iteração k chega a ter valores que fazem a iteração excessivamente cara, sempre cabe o recurso de recomeçar "jogando fora" a informação relativa a iterações velhas. Chamamos "métodos quase-Newton com memória limitada" às implementações dos métodos secantes para problemas de grande porte com armazenamento exclusivo dos vetores u_ℓ, v_ℓ que definem as atualizações das sucessivas aproximações jacobianas B_k.

Exercício 5.1: Provar a fórmula de Sherman-Morrison: se A é não-singular então $A + uv^T$ é não singular se, e somente se, $v^T A^{-1} u \neq -1$. Nesse caso,

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u} \ .$$

Usando essa fórmula, provar que quando se usa uma correção de posto um para gerar B_{k+1},

$$B_{k+1}^{-1} = B_k^{-1} + \frac{(s_k - B_k^{-1} y_k) w_k^T}{w_k^T B_k^{-1} y_k} B_k^{-1} \ .$$

Exercício 5.2: Chamando $u_k = \dfrac{s_k - B_k^{-1} y_k}{w_k^T B_k^{-1} y_k}$, comprovar que

$$B_k^{-1} = (I + u_{k-1} z_{k-1}^T) \ldots (I + u_0 z_0^T) B_0^{-1} \ , \quad k = 1, 2, \ldots$$

isto é, na resolução de (5.2.1) basta armazenar os vetores $u_0,\ z_0,\ \ldots,\ u_{k-1},\ z_{k-1}$.

Exercício 5.3: Caracterizar geometricamente o primeiro método de Broyden, mostrando que $\|B_{k+1} - B_k\|_F \leq \|B - B_k\|_F$, para toda matriz $B \in I\!R^{n \times n}$ tal que $Bs_k = y_k$. $\| \cdot \|_F$ é a norma de Frobenius: para $A \in I\!R^{m \times n}$, $\|A\|_F = (\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2)^{\frac{1}{2}}$. Provar que a mesma propriedade vale usando a norma euclidiana em vez da norma de Frobenius.

5.3 Métodos de Newton truncados

Quando n é muito grande, e a estrutura da matriz $J(x)$ não é favorável para uma fatoração LU esparsa, a resolução do sistema linear newtoniano (5.1.3) por métodos diretos fica impraticável. Os métodos quase-Newton com memória limitada são uma alternativa eficiente em muitos casos, como vimos na seção anterior. No entanto, nesses métodos, necessitamos que B_0^{-1} (ou uma fatoração de B_0) seja simples, o que, freqüentemente, não é o caso para matrizes próximas de $J(x_0)$. Isso significa que, às vezes, para implementar um método quase-Newton com memória limitada, precisamos começar com uma matriz B_0 bem diferente de um Jacobiano verdadeiro, fazendo com que as primeiras iterações do método quase-Newton (sobretudo a primeira) sejam quase aleatórias. Por exemplo, suponhamos que nosso problema original é resolver o problema de contorno tridimensional

$$\Delta u + f(u, x, y, z) = 0, \tag{5.3.1}$$

onde Δ é o operador Laplaciano, u é a função incógnita definida em $[0,1] \times [0,1] \times [0,1]$ e seus valores no contorno do cubo são conhecidos. A discretização por diferenças finitas de (5.3.1) define um sistema não linear de $(N-1)^3$ equações e incógnitas, onde $N = 1/h$ e h é o passo da discretização. Assim, se $h = 0.01$, teremos 970299 variáveis e componentes do sistema. A matriz Jacobiana deste sistema é esparsa. Entretanto, se adotamos a ordem usual lexicográfica para as incógnitas, seus elementos não nulos ocupam as seguintes posições:
(a) As tres diagonais principais;
(b) Duas subdiagonais a distância N da diagonal principal;
(c) Duas subdiagonais a distância N^2 da diagonal principal.
Devido a essa estrutura, a fatoração LU da matriz ocupa $O(N^3)$ posições de memória, o que é intolerável, tanto do ponto de vista de espaço quanto do número de operações que é necessário para sua manipulação. Logo, o método de Newton não pode ser utilizado, e os métodos quase-Newton com memória limitada são forçados a começar com uma matriz B_0 bastante afastada da Jacobiana verdadeira.

Os métodos de Newton truncados representam um ponto de vista radicalmente diferente. Em vez de resolver (5.1.3), como Newton faz, ou substituir esse sistema por outro mais manejável, no estilo quase-Newton, esses métodos abordam a resolução do sistema linear newtoniano através de métodos iterativos lineares que, como sabemos, são geralmente econômicos em termos de memória e custo computacional. Em outras palavras, para resolver

$$J(x_k)s = -F(x_k) \tag{5.3.2}$$

utiliza-se uma seqüência $s^0, s^1, s^2, \ldots$, produzida por um método iterativo linear, onde os sucessivos iterandos s^ℓ são calculados com um custo muito moderado. Vários algoritmos para resolver sistemas lineares podem ser usados. Se $J(x_k)$ é simétrica e definida positiva, resolver (5.3.2) é equivalente a

$$\text{Minimizar } \frac{1}{2}s^T J(x_k)s + F(x_k)^T s. \tag{5.3.3}$$

O método dos gradientes conjugados, que estudamos no Capítulo 4, é, geralmente, o usado para resolver iterativamente (5.3.3).
Se $J(x_k)$ é não-singular mas não é, necessariamente, simétrica a resolução de (5.3.2) é equivalente à de

$$\text{Minimizar } \frac{1}{2}\|J(x_k)s + F(x_k)\|_2^2. \tag{5.3.4}$$

A função objetivo de (5.3.4) também é uma quadrática estritamente convexa, como a de (5.3.3), portanto o método dos gradientes conjugados também pode ser empregado para resolver esse problema. Entretanto, a matriz Hessiana da função objetivo de (5.3.4) é $J(x_k)^T J(x_k)$, e seu número de condição é o quadrado do número de condição de $J(x_k)$. Isso significa que, quando $J(x_k)$ é simétrica e definida positiva, embora tanto (5.3.3) quanto (5.3.4) possam ser empregados, o uso do primeiro é preferível do ponto de vista da estabilidade numérica. Por outro lado, o potencialmente alto número de condição da Hessiana de (5.3.4) faz com que métodos alternativos a gradientes conjugados sejam introduzidos, com a expectativa de um desempenho independente do condicionamento de $J(x_k)^T J(x_k)$. O algoritmo GMRES [179] é, possivelmente, o mais utilizado atualmente para resolver problemas do tipo (5.3.4). A idéia desse método é muito análoga à idéia geométrica dos gradientes conjugados. Trata-se de minimizar a quadrática nos sucessivos subespaços de Krylov gerados por $F(x_k), J(x_k)F(x_k), J(x_k)^2 F(x_k), \ldots$. Contrariamente a gradientes conjugados, em GMRES as iterações não podem ser simplificadas significativamente, de maneira que a implementação do método se baseia diretamente na idéia geométrica e o custo de cada iteração é crescente. Por isso, as implementações correntes procedem descartando informação de passos velhos, e toda uma família de métodos pode ser definida de acordo ao volume de informação descartada.
Outras alternativas promissoras mas pouco testadas para (5.3.3) ou (5.3.4) são os métodos de gradientes com retardos, introduzidos em [80] como generalizações do método Barzilai-Borwein [8], [170], e o próprio método de Broyden aplicado à resolução de sistemas lineares [57], [141]. Os métodos de gradientes com retardos são algoritmos de memória mínima (apenas as direções dos gradientes são usadas), onde o passo de máxima descida é substituído por um coeficiente que aumenta radicalmente sua eficiência. O

método de Broyden como método iterativo linear deve ser implementado com memória limitada, já que, em estado puro, seu custo cresce a cada iteração.

Quando se fala de métodos iterativos lineares, a possibilidade de uma convergência muito lenta está sempre presente. Por isso, freqüentemente sua aplicação é precedida pela manipulação denominada "precondicionamento". Para fixar idéias, o "precondicionamento à esquerda" do sistema (5.3.2) consiste em sua transformação em um sistema equivalente

$$H_k J(x_k) s = -H_k F(x_k) \tag{5.3.5}$$

de maneira que (5.3.5) é mais fácil que (5.3.2) para o método iterativo linear escolhido. A matriz H_k é a precondicionadora de $J(x_k)$ e pretende-se que

$$H_k J(x_k) \approx I. \tag{5.3.6}$$

Naturalmente, a precondicionadora ideal seria $J(x_k)^{-1}$ mas, nos casos em questão, essa matriz não pode ser calculada. Uma boa precondicionadora deve ser, de fato, fácil de computar e manipular, objetivo, em geral, conflitante com (5.3.6). Infelizmente, não é possível fornecer receitas universalmente válidas para o precondicionamento de sistemas lineares. Ver [137], [138].
Qualquer que seja a escolha do método iterativo linear para resolver (5.3.2), deve ser decidido quando um iterando s^ℓ é uma aproximação suficientemente boa do passo newtoniano $-J(x_k)^{-1}F(x_k)$. É oportuno lembrar que, a menos que x_k esteja muito próximo da solução, o "subproblema" $F(x_k)+J(x_k)(x-x_k) = 0$, resolvido por (5.3.2), é bastante diferente do problema original $F(x) = 0$. Portanto, uma precisão muito alta na resolução do subproblema, é, não apenas anti-econômica como, provavelmente, inútil. Dembo, Eisenstat e Steihaug [48], sugeriram um critério de parada para o algoritmo iterativo linear baseado no resíduo $\|J(x_k)s^\ell + F(x_k)\|$. O critério consiste em interromper o algoritmo linear quando este resíduo (em uma norma qualquer) é uma fração η_k da norma do termo independente $F(x_k)$ (que, por outro lado, nada mais é do que o resíduo para $s = 0$). Veremos, na próxima seção, que existem razões teóricas para fazer η_k efetivamente dependente de k, embora, na prática a fração "mágica" $\eta_k \equiv 0.1$ seja geralmente preferida. Resumindo, dada uma seqüência $\eta_k \in (0,1)$, o critério de parada introduzido em [48] produz incrementos que satisfazem

$$\|J(x_k)s_k + F(x_k)\| \le \eta_k \|F(x_k)\|, \tag{5.3.7}$$

onde $\|\cdot\|$ é uma norma qualquer em $\mathbb{R}^n$. Os métodos baseados em (5.3.7) e $x_{k+1} = x_k + s_k$ costumam ser chamados "Newton-inexatos". Quando o incremento s_k é calculado como uma das iterações de um algoritmo iterativo linear falamos de métodos de Newton truncados. Na próxima seção veremos propriedades teóricas dos algoritmos para resolver sistemas não lineares baseados em (5.3.7).

5.4 Convergência local

Nas seções anteriores apresentamos os métodos de Newton, quase-Newton e Newton truncados. Agora veremos resultados de convergência local relacionados com esses algoritmos. Diremos que um método possui convergência local em relação a determinado tipo de soluções do problema considerado se, dada uma solução x_* desse tipo, existe $\varepsilon > 0$ tal que toda seqüência $\{x_k\}$ gerada pelo algoritmo onde $\|x_0 - x_*\| \leq \varepsilon$, converge para x_*. Os resultados de convergência local estão quase sempre associados a resultados de *ordem de convergência.* Diremos que uma seqüência $\{x_k\}$ converge *linearmente* para x_* relativamente à norma $\|\cdot\|$ se existem $k_0 \in \mathcal{N}$ e $r \in (0,1)$ tais que, para todo $k \geq k_0$,

$$\|x_{k+1} - x_*\| \leq r\|x_k - x_*\|. \tag{5.4.1}$$

A convergência de $\{x_k\}$ para x_* será chamada *superlinear* se existe uma seqüência $r_k > 0$ tendendo a 0, tal que

$$\|x_{k+1} - x_*\| \leq r_k\|x_k - x_*\| \tag{5.4.2}$$

para todo $k = 0, 1, 2, \ldots$. Pela equivalência das normas em $\mathbb{R}^n$ podemos ver que a convergência superlinear de uma seqüência é independente da norma. Ao mesmo tempo, se $x_k \to x_*$ superlinearmente, então dado qualquer $r \in (0,1)$ e qualquer norma em $\mathbb{R}^n$, a desigualdade (5.4.1) acabará se verificando para k_0 suficientemente grande, ou seja, teremos convergência linear.
Se $x_k \to x_*$ e existem $k_0 \in \mathcal{N}$, $c > 0$ e $p > 0$ tais que, para todo $k \geq k_0$,

$$\|x_{k+1} - x_*\| \leq c\|x_k - x_*\|^{p+1}, \tag{5.4.3}$$

diremos que $\{x_k\}$ converge para x_* com ordem pelo menos $p+1$. Se $p = 1$, falaremos de *convergência quadrática.* Pela equivalência de normas, (5.4.3) também é independente da norma usada. Além disso, é fácil ver que este tipo de convergência implica a convergência superlinear. Quanto maior seja p mais rapidamente x_k tenderá a x_*. Com efeito, se, para uma iteração

k, o erro $\|x_k - x_*\|$ é da ordem de 0.1, então, na iteração seguinte será da ordem de $c0.1^{p+1}$, e, depois de m iterações será $c0.1^{m(p+1)}$. Portanto, o número de dígitos corretos das componentes da solução crescerá rapidamente se $p \geq 1$. Por isso, costuma-se dizer que, na convergência quadrática, o número de decimais corretos é duplicado em cada iteração. Assim, o tipo de convergência mais desejável é a de ordem $p + 1$ com o maior valor de p possível. Nas seqüências produzidas por métodos numéricos geradas em um computador, a convergência quadrática (ou melhor que quadrática) é observável no rápido crescimento dos dígitos repetidos de uma iteração para outra, ou, equivalentemente, o número de decimais iguais a zero do erro. A convergência superlinear é mais difícil de observar empiricamente. Via de regra, em seqüências teoricamente superlineares (mas não quadráticas), o erro aparece diminuindo de maneira consistente, mas não é usual observar uma queda monótona para zero do quociente entre dois erros consecutivos. Já a apreciação da convergência linear depende integralmente da taxa r. Alguns métodos de tipo ponto fixo para resolver sistemas lineares produzem seqüências com uma taxa linear de convergência tão próxima de 1, que sua utilidade é praticamente nula. Por outro lado, se a taxa for menor que, digamos, 0.5, a convergência pode ser indistinguível, nos experimentos, do comportamento superlinear.

Nesta seção assumiremos as seguintes hipóteses gerais: $F : \Omega \to I\!R^n$, com $\Omega \subset I\!R^n$ aberto e convexo e $F \in C^1(\Omega)$. Portanto, para todo $x \in \Omega$,

$$\lim_{h \to 0} \frac{\|F(x+h) - F(x) - J(x)h\|}{\|h\|} = 0. \tag{5.4.4}$$

Suporemos também que $x_* \in \Omega$ é tal que $F(x_*) = 0$ e $J(x_*)$ é não-singular. Para a prova da convergência quadrática do método de Newton assumimos que existem $L > 0$ e $p > 0$ tais que, em uma vizinhança de x_*,

$$\|J(x) - J(x_*)\| \leq L\|x - x_*\|^p \tag{5.4.5}$$

onde $\| \cdot \|$ é uma norma qualquer em $I\!R^n$ bem como a norma de matrizes consistente associada em $I\!R^{n \times n}$.

Exercício 5.4: Usando (5.4.5), mostrar que para todo $x, z \in \Omega$,

$$\|F(z) - F(x) - J(x_*)(z - x)\| \leq L\|x - z\| \max\{\|x - x_*\|^p, \|z - x_*\|^p\} .$$

Exercício 5.5: Usando (5.4.5), mostrar que para todo $x \in \Omega$,

$$\|F(x) - J(x_*)(x - x_*)\| \leq \frac{L}{1+p}\|x - x_*\|^{p+1} .$$

5.4.1 O teorema das duas vizinhanças

O objetivo desta subseção é mostrar que, se x_0 está próximo de x_* e *todas* as matrizes B_k estão perto de $J(x_*)$, a seqüência gerada por $x_{k+1} = x_k - B_k^{-1}F(x_k)$ converge para x_* com taxa linear. Esse resultado será aplicável aos métodos quase-Newton em geral, e, especificamente, ao próprio método de Newton. Usaremos de maneira essencial que todas as matrizes que se encontram numa certa vizinhança da matriz não-singular $J(x_*)$ são não-singulares. No Lema 5.4.1 vamos precisar o tamanho dessa vizinhança. Um resultado prévio, de álgebra, é o chamado Lema de Banach: dada uma norma arbitrária $\|\cdot\|$ em $I\!R^n$, que denota também a norma matricial subordinada, se $\|A\| < 1$, então $I + A$ é não-singular e

$$\frac{1}{1+\|A\|} \leq \|(I+A)^{-1}\| \leq \frac{1}{1-\|A\|} \ .$$

Exercício 5.6: Demonstrar o Lema de Banach.

Lema 5.4.1
Se $B \in I\!R^{n\times n}$ é tal que $\|B - J(x_*)\| \leq \dfrac{1}{2\|J(x_*)^{-1}\|}$ então B^{-1} existe e satisfaz $\|B^{-1}\| \leq 2\|J(x_*)^{-1}\|$.

Prova: Seja $A = BJ(x_*)^{-1} - I = [B - J(x_*)]J(x_*)^{-1}$. Pela consistência da norma segue que

$$\|A\| = \|[B - J(x_*)]J(x_*)^{-1}\| \leq \|[B - J(x_*)]\| \; \|J(x_*)^{-1}\| \leq \frac{1}{2} < 1 \ ,$$

ou seja, estamos nas condições do Lema de Banach e, então $BJ(x_*)^{-1}$ é não-singular. Logo, existe B^{-1} e vale $[BJ(x_*)^{-1}]^{-1} = J(x_*)B^{-1}$. Além disso,

$$\|J(x_*)B^{-1}\| \leq \frac{1}{1 - \|BJ(x_*)^{-1} - I\|} \leq 2 \ .$$

Como $\|B^{-1}\| = \|J(x_*)^{-1}J(x_*)B^{-1}\| \leq \|J(x_*)^{-1}\| \; \|J(x_*)B^{-1}\|$, segue que $\|B^{-1}\| \leq 2\|J(x_*)^{-1}\|$. **QED**

Lema 5.4.2 - das duas vizinhanças.
Para cada $x \in \Omega$ *e* $B \in I\!R^{n \times n}$*, definimos a função* $\Phi(x, B) = x - B^{-1}F(x)$*. Seja* $r \in (0, 1)$*. Existem* $\varepsilon_1 = \varepsilon_1(r), \delta_1 = \delta_1(r) > 0$ *tais que se* $\|x - x_*\| \leq \varepsilon_1$*,* $\|B - J(x_*)\| \leq \delta_1$*, a função* $\Phi(x, B)$ *está bem definida e satisfaz* $\|\Phi(x, B) - x_*\| \leq r\|x - x_*\|$*.*

Prova: Seja $\delta_1' = \dfrac{1}{2\|J(x_*)^{-1}\|}$. Pelo Lema 5.4.1, se $\|B - J(x_*)\| \leq \delta_1'$ então B^{-1} existe e satisfaz

$$\|B^{-1}\| \leq 2\|J(x_*)^{-1}\| \,. \tag{5.4.6}$$

Assim, $\Phi(x, B)$ está bem definida se $x \in \Omega$ e $\delta_1 \leq \delta_1'$.
Agora

$$\|\Phi(x, B) - x_*\| \leq A_1 + A_2 \tag{5.4.7}$$

onde

$$A_1 = \|x - x_* - B^{-1}J(x_*)(x - x_*)\| \;\text{ e }\; A_2 = \|B^{-1}[F(x) - J(x_*)(x - x_*)]\| \,.$$

Por (5.4.6), temos que

$$\begin{aligned} A_1 &= \|x - x_* - B^{-1}J(x_*)(x - x_*) - B^{-1}B(x - x_*) + B^{-1}B(x - x_*)\| \\ &= \|x - x_* - B^{-1}B(x - x_*) + B^{-1}[B - J(x_*)](x - x_*)\| \\ &= \|B^{-1}[B - J(x_*)](x - x_*)\| \\ &\leq \|B^{-1}\| \, \|B - J(x_*)\| \, \|x - x_*\| \\ &\leq 2\|J(x_*)^{-1}\| \, \delta_1 \, \|x - x_*\| \,. \end{aligned} \tag{5.4.8}$$

Pela diferenciabilidade de F e por (5.4.6), temos:

$$A_2 \leq \|B^{-1}\| \, \|F(x) - J(x_*)(x - x_*)\| \leq 2\|J(x_*)^{-1}\| \, \beta(x) \tag{5.4.9}$$

onde $\displaystyle\lim_{x \to x_*} \frac{\beta(x)}{\|x - x_*\|} = 0$.

Seja ε_1 tal que

$$2\left(\delta_1 + \sup_{\|x - x_*\| \leq \varepsilon_1} \left\{ \frac{\beta(x)}{\|x - x_*\|} \right\}\right) \leq \frac{r}{\|J(x_*)^{-1}\|} \,. \tag{5.4.10}$$

Então, para $\|B - J(x_*)\| \leq \delta_1$ e $\|x - x_*\| \leq \varepsilon_1$, por (5.4.7)–(5.4.10) temos

$$\begin{aligned} \|\Phi(x, B) - x_*\| &\leq 2\|J(x_*)^{-1}\| \, \delta_1 \, \|x - x_*\| + 2\|J(x_*)^{-1}\| \, \beta(x) \\ &= 2\|J(x_*)^{-1}\| \left(\delta_1 + \frac{\beta(x)}{\|x - x_*\|} \right) \|x - x_*\| \\ &\leq r\|x - x_*\| \, . \ \mathbf{QED} \end{aligned}$$

Teorema 5.4.3 - das duas vizinhanças.
Seja $r \in (0,1)$. Existem $\varepsilon = \varepsilon(r)$ e $\delta = \delta(r)$ tais que, se $\|x_0 - x_*\| \leq \varepsilon$ e $\|B_k - J(x_*)\| \leq \delta$ para todo k, então a seqüência gerada por $x_{k+1} = x_k - B_k^{-1} F(x_k)$ está bem definida, converge a x_* e $\|x_{k+1} - x_*\| \leq r\|x_k - x_*\|$ para todo k.

Prova: Considerando a função $\Phi(x, B) = x - B^{-1}F(x)$, temos $x_{k+1} = \Phi(x_k, B_k)$, $k = 0, 1, 2, \ldots$. A prova segue por um argumento de indução e pelo Lema 5.4.2. **QED**

Uma conseqüência imediata do Teorema das duas vizinhanças é a convergência local linear do método de Newton estacionário. Com efeito, dado $r \in (0,1)$, pela continuidade das derivadas de F, existe ε_2 tal que $\|J(x_0) - J(x_*)\| \leq \delta(r)$ sempre que $\|x_0 - x_*\| \leq \varepsilon_2$. Tomemos, então ε como o mínimo entre $\varepsilon(r)$ e ε_2, onde $\delta(r)$ e $\varepsilon(r)$ são os definidos no Teorema das duas vizinhanças. Então, se $\|x_0 - x_*\| \leq \varepsilon$ teremos $\|J(x_0) - J(x_*)\| \leq \delta(r)$ e, portanto, $\|B_k - J(x_*)\| \leq \delta(r)$ para todo k. Logo, estamos dentro das hipóteses do teorema, e, em conseqüência, a seqüência converge com a taxa linear r. É importante observar que esta pequena prova foi iniciada com um $r \in (0,1)$ arbitrário. Portanto, a taxa de convergência linear do método de Newton estacionário poderia ser arbitrariamente pequena, tomando x_0 suficientemente próximo de x_*.

5.4.2 Convergência quadrática de Newton

A aplicação do Teorema das duas vizinhanças ao método de Newton é bastante natural. No entanto, a última observação da subseção anterior, permite vislumbrar que, para este método, resultados mais fortes são possíveis. Aqui vamos usar a condição (5.4.5) para provar que a ordem de convergência de Newton é, pelo menos $p + 1$. É usual que (5.4.5) seja válida com $p = 1$, por isso chamaremos essa propriedade de "convergência quadrática". As situações em que (5.4.5) vale para algum $p \in (0,1)$ mas não para $p = 1$ são um tanto patológicas, e não têm maior importância prática. No entanto, é interessante refletir sobre o caso em que (5.4.5) é satisfeita para algum $p > 1$. Por exemplo, se $p = 2$, essa condição significa que as derivadas segundas de F existem e são nulas em x_*. Nesse caso, a convergência de Newton é de ordem 3. Assim, quanto maior seja a ordem das derivadas que se anulam na solução, acima das segundas, Newton convergirá mais rapidamente. No caso extremo, todas as derivadas de F são nulas em x_* o que, quase sempre, indica que F é uma função linear em uma vizinhança da solução. Nesse caso, a ordem de convergência $p+1$ para todo p significa que x_1 será igual a x_*, ou seja, o método se comportará como um método direto, que é exatamente o que se espera dele quando aplicado a uma função linear.

Teorema 5.4.4 - Convergência quadrática de Newton.
Suponhamos que F, L, p satisfazem (5.4.5). Então existem $\varepsilon, \gamma > 0$ tais que para todo x_0 verificando $\|x_0 - x_*\| \leq \varepsilon$, a seqüência gerada por

$$x_{k+1} = x_k - J(x_k)^{-1} F(x_k), \quad k = 0, 1, \ldots$$

está bem definida, converge a x_* e satisfaz

$$\|x_{k+1} - x_*\| \leq \gamma \|x_k - x_*\|^{p+1} .$$

Prova: Escolhemos um r arbitrário entre 0 e 1, digamos, $r = 0.5$. Seja $\varepsilon_1 = \varepsilon_1(r)$, definido pelo Lema das duas vizinhanças. Pela continuidade de $J(x)$, existe $\varepsilon_2 > 0$ tal que, sempre que $\|x - x_*\| \leq \varepsilon_2$, temos $\|J(x) - J(x_*)\| \leq \delta_1(r)$. Tomamos

$$\varepsilon = \text{ mínimo } \{\varepsilon_1, \varepsilon_2\},$$

logo $\|J(x_0) - J(x_*)\| \leq \delta_1(r)$. Então, pelo Lema das duas vizinhanças,

$$\|x_1 - x_*\| \leq r\|x_0 - x_*\| < \varepsilon_1.$$

Portanto, $\|J(x_1) - J(x_*)\| \leq \delta_1(r)$ e o raciocínio pode ser repetido, indutivamente, para provar que $\{x_k\}$ converge para x_* linearmente com taxa r. Agora, por (5.4.6), temos que, para todo k,

$$\begin{aligned} \|x_{k+1} - x_*\| &= \|x_k - x_* - J(x_k)^{-1} F(x_k)\| \\ &= \|J(x_k)^{-1}(-F(x_k) - J(x_k)(x_* - x_k))\| \\ &\leq 2\|J(x_*)^{-1}\| \, \|F(x_k) - J(x_k)(x_k - x_*)\|. \end{aligned}$$

Mas, por (5.4.5) e pelo resultado do exercício 5.5,

$$\begin{aligned} \|F(x_k) - J(x_k)(x_k - x_*)\| &\leq |F(x_k) - J(x_*)(x_k - x_*)| + L\|x_k - x_*\|^{p+1} \\ &\leq 2L\|x_k - x_*\|^{p+1}. \end{aligned}$$

Portanto,

$$\|x_{k+1} - x_*\| \leq 4\|J(x_*)^{-1}\|L\|x_k - x_*\|^{p+1},$$

o que completa a prova. **QED**

Sutilezas maiores que as do Teorema 5.4.4 são possíveis. De fato, o leitor poderá verificar que, mesmo sem supor a condição (5.4.5), mas usando a diferenciabilidade de F, a convergência de Newton é superlinear.

5.4.3 Convergência dos métodos quase-Newton

O Teorema das duas vizinhanças é um elemento essencial na teoria de convergência dos métodos quase-Newton. Com efeito, ele nos diz que em um método desse tipo, se o ponto inicial está suficientemente perto da solução e todas as matrizes B_k estão próximas de $J(x_*)$ a convergência ocorre com taxa linear. A maneira mais fácil de satisfazer as hipóteses desse teorema é escolher uma única vez B_0 próxima de uma Jacobiana e tomar todas as outras B_k iguais a B_0. É o que o método de Newton estacionário faz. A maioria dos métodos quase-Newton tenta uma opção melhor. Por exemplo, os métodos secantes definem $B_{k+1} = B_k + \Delta B_k$ para todo k, onde, quase sempre, ΔB_k tem posto pequeno. Portanto, mesmo que B_0 esteja perto de $J(x_*)$, poderíamos ter o azar de que alguma das B_k's posteriores ficassem fora da vizinhança que garante a convergência linear. Em outras palavras, B_{k+1} pode sofrer uma deterioração em relação a B_k. Para garantir que, apesar dessas possíveis deteriorações, todas as B_k estejam na boa vizinhança de que fala o Teorema 5.4.3, são provados, para os distintos métodos quase-Newton, teoremas de "deterioração limitada". Como seu nome indica, esses teoremas estabelecem que, embora a distância entre B_{k+1} e $J(x_*)$ possa ser maior que $\|B_k - J(x_*)\|$, o grau de degeneração não pode ser tão grande ao ponto de comprometer a convergência. Existem diferentes teoremas de deterioração limitada para os distintos métodos quase-Newton. Enfoques unificados são discutidos em [55], [134] e [135]. Uma propriedade de deterioração limitada típica é:

$$\|B_{k+1} - J(x_*)\| \leq \|B_k - J(x_*)\| + c\|x_k - x_*\| \tag{5.4.11}$$

para algum $c > 0$. A desigualdade (5.4.11) estabelece que a deterioração de B_{k+1} em relação a B_k é de ordem não maior que o erro na iteração k. O método de Broyden, do qual falamos na Seção 5.3, satisfaz uma propriedade desse tipo. Para mostrar como ela contribui para não corromper a convergência de um método quase-Newton, vamos provar o seguinte teorema.

Teorema 5.4.5
Consideramos o método quase-Newton definido por $x_{k+1} = x_k - B_k^{-1} F(x_k)$, onde as matrizes B_k satisfazem (5.4.11). Seja $r \in (0, 1)$. Então, existem $\varepsilon, \delta > 0$ tais que, se $\|x_0 - x_*\| \le \varepsilon$ e $\|B_0 - J(x_*)\| \le \delta$, a seqüência está bem definida, converge a x_* e satisfaz $\|x_{k+1} - x_*\| \le r\|x_k - x_*\|$ para todo k.

Prova: Sejam $\varepsilon_1 = \varepsilon(r)$ e $\delta_1 = \delta(r)$ os definidos no Teorema das duas vizinhanças. Sejam $\varepsilon \le \varepsilon_1$ e $\delta \le \delta_1$ tais que

$$\delta + \frac{c\varepsilon}{1 - r} \le \delta_1. \tag{5.4.12}$$

Vamos provar por indução que

$$\|x_k - x_*\| \le r\|x_{k-1} - x_*\| \le r^k \varepsilon$$

e

$$\|B_k - J(x_*)\| \le \delta + c\varepsilon(1 + r + \ldots + r^{k-1}).$$

O primeiro passo é óbvio. Vejamos o passo indutivo. Por hipótese indutiva temos:

$$\begin{aligned} \|B_k - J(x_*)\| &\le \delta + c\varepsilon(1 + r + \ldots + r^{k-1}) \\ &\le \delta + \frac{c\varepsilon}{1 - r} \le \delta_1. \end{aligned}$$

Como $\|x_k - x_*\| \le r^k \varepsilon \le \varepsilon$, o Teorema das duas vizinhanças implica

$$\|x_{k+1} - x_*\| \le r\|x_k - x_*\| \le r^{k+1}\varepsilon.$$

Agora, por (5.4.11),

$$\begin{aligned} \|B_{k+1} - J(x_*)\| &\le \|B_k - J(x_*)\| + c\|x_k - x_*\| \\ &\le \delta + c\varepsilon(1 + r + \ldots + r^{k-1}) + cr^k\varepsilon, \end{aligned}$$

de onde a tese segue-se trivialmente. **QED**

A maioria dos resultados de deterioração limitada para métodos quase-Newton são obtidos usando propriedades geométricas das fórmulas de atualização das B_k's. O exemplo mais claro é fornecido pelo método de Broyden. Como vimos no Exercício 5.3, nesse algoritmo, B_{k+1} é a projeção segundo a norma de Frobenius de B_k na variedade afim das matrizes que satisfazem a equação secante $Bs_k = y_k$. Se $J(x_*)$ satisfizesse essa equação, a distância entre B_{k+1} e $J(x_*)$ seria menor ou igual à distância entre B_k e $J(x_*)$ e o princípio (5.4.11) seria satisfeito com $c = 0$. Infelizmente, em geral, $J(x_*)$ não é uma das matrizes que satisfazem a equação secante da iteração k. No entanto, se definimos

$$\tilde{B}_k = \int_0^1 J(x_k + t(x_{k+1} - x_k))dt, \tag{5.4.13}$$

podemos verificar, com o teorema fundamental do cálculo, que $\tilde{B}_k s_k = y_k$. Portanto,

$$\|B_{k+1} - \tilde{B}_k\| \leq \|B_k - \tilde{B}_k\|.$$

Assim,

$$\begin{aligned} \|B_{k+1} - J(x_*)\| &\leq \|B_{k+1} - \tilde{B}_k\| + \|\tilde{B}_k - J(x_*)\| \\ &\leq \|B_k - \tilde{B}_k\| + \|\tilde{B}_k - J(x_*)\| \\ &\leq \|B_k - J(x_*)\| + 2\|\tilde{B}_k - J(x_*)\|. \end{aligned} \tag{5.4.14}$$

Por (5.4.13), e usando (5.4.5), podemos verificar que $\|\tilde{B}_k - J(x_*)\| = O(\|x_k - x_*\|)$, portanto a propriedade (5.4.11) segue de (5.4.14).

A interpretação de muitas fórmulas secantes como projeções permite, geralmente, provar outra propriedade importante:

$$\lim_{k\to\infty} \|B_{k+1} - B_k\| = 0. \tag{5.4.15}$$

A idéia é usar, em cada iteração, o Teorema de Pitágoras. Apenas neste parágrafo, $\|\cdot\|$ será a norma de Frobenius,

$$\|B_{k+1} - B_k\|^2 = \|B_k - \tilde{B}_k\|^2 - \|B_{k+1} - \tilde{B}_k\|^2. \tag{5.4.16}$$

Portanto,

$$\|B_{k+1} - B_k\|^2 = \|B_k - J(x_*)\|^2 - \|B_{k+1} - J(x_*)\|^2 + O(\|x_k - x_*\|). \tag{5.4.17}$$

Assim, supondo que o princípio de deterioração limitada já permitiu provar a convergência com taxa linear r da seqüência $\{x_k\}$, e somando todas as igualdades (5.4.17),

$$\sum_{k=0}^{\infty} \|B_{k+1} - B_k\|^2 \leq \|B_0 - J(x_*)\|^2 + \frac{\|x_0 - x_*\|}{1-r}, \qquad (5.4.18)$$

logo, a série da esquerda em (5.4.18) converge e, portanto, (5.4.15) se verifica.

Por enquanto nos limitamos a mostrar que os métodos quase-Newton com deterioração limitada não são piores que o mais simples dos métodos quase-Newton, onde B_k não muda nunca e, portanto, a deterioração é nula. Se os métodos secantes não pudessem oferecer mais do que isso, nunca teriam sido populares. De fato, veremos agora que, via de regra, os métodos secantes não apenas convergem com a taxa linear r de que fala o teorema das duas vizinhanças mas, também, são superlineares. A ferramenta fundamental para essa prova é o seguinte teorema, cujo resultado é conhecido como "condição Dennis-Moré".

Teorema 5.4.6 - Condição Dennis-Moré.
Suponhamos que F satisfaz as hipóteses gerais, incluindo (5.4.5), a seqüência gerada por

$$x_{k+1} = x_k - B_k^{-1} F(x_k)$$

está bem definida, converge a x_*, e satisfaz

$$\lim_{k\to\infty} \frac{\|[B_k - J(x_*)]s_k\|}{\|s_k\|} = 0\,. \qquad (5.4.19)$$

Então a convergência é superlinear.

Antes de provar a condição Dennis-Moré vamos refletir sobre seu significado. Uma primeira observação é que o método de Newton claramente satisfaz (5.4.19) e que, ainda mais, qualquer seqüência de matrizes $\{B_k\}$ tal que $B_k \to J(x_*)$ também satisfaz essa condição. Logo, por este teorema, o método de Newton estacionário com recomeços, do qual falamos na Seção 5.2, é superlinear. No entanto, a condição Dennis-Moré exige menos que a convergência de B_k para $J(x_*)$. Com efeito, o que deve tender para zero não é a diferença $B_k - J(x_*)$ mas a aplicação dessa diferença na direção incremental $s_k/\|s_k\|$. Ou seja, para efeitos de convergência superlinear, é indiferente o que B_k faça com direções diferentes dos incrementos e apenas a ação das matrizes sobre os s_k's tem importância. Assim, um método com essas condições pode ser superlinearmente convergente, mesmo com as matrizes B_k convergindo a algo diferente da Jacobiana na solução. No Teorema 5.4.6 apresentamos a condição Dennis-Moré apenas como uma condição suficiente. Na verdade, o resultado é bem mais elegante (ver [52], [54]): a condição (5.4.19) é também necessária para a convergência superlinear dos métodos quase-Newton e o fato de que x_* é uma raiz pode ser deduzido dela e não apenas assumido como hipótese.

Na prova do Teorema Dennis-Moré, faremos uso de um lema que, brevemente, mostra que $\|F(x)\|$ pode ser utilizado como uma medida da distância entre x e x_* quando $J(x_*)$ é não-singular:

Lema 5.4.7
Existem $\varepsilon, c_1, c_2 > 0$ tais que, sempre que $\|x - x_*\| \leq \varepsilon$,

$$c_1\|x - x_*\| \leq \|F(x)\| \leq c_2\|x - x_*\|.$$

Prova: Pela diferenciabilidade de F,

$$\lim_{x \to x_*} \frac{\|F(x) - J(x_*)(x - x_*)\|}{\|x - x_*\|} = 0.$$

Mas

$$\|x - x_*\| = \|J(x_*)^{-1}J(x_*)(x - x_*)\| \leq \|J(x_*)^{-1}\|\|J(x_*)(x - x_*)\|,$$

portanto

$$\lim_{x \to x_*} \frac{\|F(x) - J(x_*)(x - x_*)\|}{\|J(x_*)^{-1}\|\|J(x_*)(x - x_*)\|} = 0.$$

Logo,

$$\lim_{x \to x_*} \frac{\|F(x) - J(x_*)(x - x_*)\|}{\|J(x_*)(x - x_*)\|} = 0.$$

Mas $| \, \|F(x)\| - \|J(x_*)(x - x_*)\| \, | \leq \|F(x) - J(x_*)(x - x_*)\|$, portanto existe $\varepsilon > 0$ tal que, sempre que $0 < \|x - x_*\| \leq \varepsilon$,

$$-\frac{1}{2} \leq \frac{\|F(x)\| - \|J(x_*)(x - x_*)\|}{\|J(x_*)(x - x_*)\|} \leq \frac{1}{2},$$

ou seja,

$$-\frac{1}{2}\|J(x_*)(x - x_*)\| \leq \|F(x)\| - \|J(x_*)(x - x_*)\| \leq \frac{1}{2}\|J(x_*)(x - x_*)\|,$$

ou ainda,

$$\frac{1}{2}\|J(x_*)(x-x_*)\| \le \|F(x)\| \le \frac{3}{2}\|J(x_*)(x-x_*)\|. \qquad (5.4.20)$$

Mas, $\|J(x_*)(x-x_*)\| \le \|J(x_*)\|\|x-x_*\|$ e

$$\|x-x_*\| = \|J(x_*)^{-1}J(x_*)(x-x_*)\| \le \|J(x_*)^{-1}\|\|J(x_*)(x-x_*)\|,$$

portanto a tese do Lema segue de (5.4.20), com $c_1 = 1/(2\|J(x_*)\|^{-1})$ e $c_2 = \frac{3}{2}\|J(x_*)\|$. **QED**

Prova do Teorema Dennis-Moré: Por (5.4.19), temos:

$$[B_k - J(x_*)](x_{k+1}-x_k) = -F(x_k) - J(x_*)(x_{k+1}-x_k)$$

$$= F(x_{k+1}) - F(x_k) - J(x_*)(x_{k+1}-x_k) - F(x_{k+1}).$$

Agora, pelo resultado do Exercício 5.4,

$$\|F(x_{k+1})-F(x_k)-J(x_*)(x_{k+1}-x_k)\| \le L\|x_{k+1}-x_k\| \max\{\|x_k-x_*\|^p, \|x_{k+1}-x_*\|^p\} .$$

Portanto, pela convergência de $\{x_k\}$ e pela condição (5.4.19),

$$\lim_{k\to\infty} \frac{\|F(x_{k+1})\|}{\|x_{k+1}-x_k\|} = 0 . \qquad (5.4.21)$$

Agora, $\|x_{k+1}-x_k\| \le \|x_{k+1}-x_*\| + \|x_k-x_*\|$ e, pelo Lema 5.4.7, para k suficientemente grande, temos $\|F(x_{k+1})\| \ge c_1\|x_{k+1}-x_*\|$. Portanto, por (5.4.21),

$$\lim_{k\to\infty} \frac{\|x_{k+1}-x_*\|}{\|x_k-x_*\| + \|x_{k+1}-x_*\|} = 0, \qquad (5.4.22)$$

e a convergência superlinear segue de (5.4.22) após breve manipulação algébrica. **QED**

Quando, para um método secante, pode ser provada uma propriedade de deterioração limitada e a forma de definir ΔB_k permite demonstrar também que $\|B_{k+1} - B_k\| \to 0$, a convergência superlinear do método resulta do Teorema Dennis-Moré. Formalizaremos isso no seguinte teorema.

Teorema 5.4.8
Suponhamos as hipóteses gerais desta seção e, também, a condição (5.4.5). Suponhamos que o método quase-Newton definido por $x_{k+1} = x_k - B_k^{-1}F(x_k)$

tem as propriedades (5.4.11) e (5.4.15) e que a equação secante () é satisfeita para todo k. Então, existem $\varepsilon, \delta > 0$ tais que, se $\|x_0 - x_*\| \leq \varepsilon$ e $\|B_0 - J(x_*)\| \leq \varepsilon$, a seqüência $\{x_k\}$ está bem definida, e converge superlinearmente para x_*.

Prova: A boa definição e convergência resultam do Teorema 5.4.3. Para provar a superlinearidade vamos mostrar que a condição Dennis-Moré é satisfeita. Pelo resultado do Exercício 5.4, temos que

$$\|y_k - J(x_*)s_k\| \leq L\|s_k\| \max \{\|x_k - x_*\|^p, \|x_{k+1} - x_*\|^p\}. \quad (5.4.23)$$

Mas, pela condição secante, $B_{k+1}s_k = y_k$. Logo, por (5.4.23) e a convergência de $\{x_k\}$,

$$\lim_{k\to\infty} \frac{\|[B_{k+1} - J(x_*)]s_k\|}{\|s_k\|} = 0. \quad (5.4.24)$$

Claramente, a condição Dennis-Moré (5.4.19) pode ser deduzida de (5.4.24) e (5.4.15). Portanto, a convergência é superlinear. **QED**

5.4.4 Convergência dos Newton inexatos

Como dissemos na Seção 5.3, chamamos métodos de Newton inexatos àqueles baseados na condição (5.3.7). Newton truncados serão aqueles métodos nos quais se utiliza um método iterativo linear para resolver, aproximadamente, o sistema (5.3.2). Freqüentemente, as duas expressões são utilizadas como sinônimos. Entretanto, pode ser que um método de Newton truncado utilize um critério de parada diferente de (5.3.7), e também é possível que o incremento s_k que satisfaz (5.3.7) não seja originado de um processo iterativo linear. Por isso, é conveniente manter as duas denominações com significados diferenciados.
No resultado principal desta subseção, provaremos que os métodos de Newton inexatos são localmente convergentes com taxa linear, em determinada norma, se o valor η_k se mantém fixo ao longo de todo o processo. Se $\eta_k \to 0$, veremos que a convergência é superlinear.

Teorema 5.4.9 - Dembo - Eisenstat - Steihaug.
(a) Se $\eta_k \leq \eta_{\max} < r < 1$, existe $\varepsilon > 0$ tal que se $\|x_0 - x_*\| \leq \varepsilon$, então a seqüência $\{x_k\}$ gerada por um método de Newton inexato converge a x_*. Além disso a convergência é linear com taxa r:

$$\|x_{k+1} - x_*\|_* \leq r\|x_k - x_*\|_* , \quad (5.4.25)$$

onde a norma $\|\cdot\|_*$ *está definida por* $\|y\|_* = \|J(x_*)y\|$.
(b) Se a seqüência $\{x_k\}$ *gerada por um método de Newton inexato converge a* x_* *e se*

$$\lim_{k\to\infty} \eta_k = 0 \ , \tag{5.4.26}$$

então a convergência é superlinear.

Prova: (a) Como $J(x_*)$ é não-singular, para todo $y \in I\!R^n$ vale:

$$\frac{1}{\mu}\|y\| \leq \|y\|_* \leq \mu\|y\| \tag{5.4.27}$$

onde $\mu = \max\{\|J(x_*)\|, \|J(x_*)^{-1}\|\}$.
Como $\eta_{\max} < r$, existe $\gamma > 0$ suficientemente pequeno tal que

$$(1+\mu\gamma)\ [\eta_{\max}(1+\mu\gamma) + 2\mu\gamma] \leq r \ .$$

Agora, escolhemos $\varepsilon > 0$ suficientemente pequeno tal que

$$\|J(y) - J(x_*)\| \leq \gamma \ , \tag{5.4.28}$$

$$\|J(y)^{-1} - J(x_*)^{-1}\| \leq \gamma \ , \tag{5.4.29}$$

$$\|F(y) - F(x_*) - J(x_*)(y - x_*)\| \leq \gamma\|y - x_*\| \tag{5.4.30}$$

se $\|y - x_*\| \leq \mu^2\varepsilon$. A existência de ε é garantida pela diferenciabilidade de F.
Assumindo que $\|x_0 - x_*\| \leq \varepsilon$, vamos provar (5.4.25) por indução. Por (5.4.27), pela hipótese de indução e, novamente por (5.4.27), temos

$$\|x_k - x_*\| \leq \mu\|x_k - x_*\|_* \leq \mu r^k\|x_0 - x_*\|_* \leq \mu^2\|x_0 - x_*\| \leq \mu^2\varepsilon \ ,$$

de tal forma que (5.4.28)–(5.4.30) valem com $y = x_k$. Além disso, a k-ésima etapa de um método de Newton inexato é definida de tal forma que existe s_k satisfazendo

$$J(x_k)s_k = -F(x_k) + R_k, \quad \text{onde} \quad \frac{\|R_k\|}{\|F(x_k)\|} \leq \eta_k \ . \tag{5.4.31}$$

Então,

$$\begin{aligned}
J(x_*)(x_{k+1} - x_*) &= J(x_*)s_k + J(x_*)(x_k - x_*) \\
&= J(x_*)J(x_k)^{-1}[J(x_k)s_k + J(x_k)(x_k - x_*)] \\
&= [I + J(x_*)(J(x_k)^{-1} - J(x_*)^{-1})]\ [J(x_k)s_k + F(x_k) + J(x_k)(x_k - x_*) \\
&\quad - J(x_*)(x_k - x_*) - F(x_k) + F(x_*) + J(x_*)(x_k - x_*)] \\
&= [I + J(x_*)(J(x_k)^{-1} - J(x_*)^{-1})]\ [R_k + [J(x_k) - J(x_*)](x_k - x_*) \\
&\quad - [F(x_k) - F(x_*) - J(x_*)(x_k - x_*)]] \ .
\end{aligned}$$

Usando a definição de μ, (5.4.28), (5.4.29), (5.4.30) e (5.4.31), temos

$$\begin{aligned}\|x_{k+1}-x_*\|_* &\leq [1+\|J(x_*)\|\ \|J(x_k)^{-1}-J(x_*)^{-1}\|]\ [\|R_k\|+\\ &\quad +\|J(x_k)-J(x_*)\|\ \|x_k-x_*\|+\|F(x_k)-F(x_*)-J(x_*)(x_k-x_*)\|]\\ &\leq (1+\mu\gamma)[\eta_k\|F(x_k)\|+\gamma\|x_k-x_*\|+\gamma\|x_k-x_*\|]\ .\end{aligned}$$

Como

$$F(x_k)=[J(x_*)(x_k-x_*)]+[F(x_k)-F(x_*)-J(x_*)(x_k-x_*)]\ ,$$

de (5.4.30) segue que:

$$\begin{aligned}\|F(x_k)\| &\leq \|x_k-x_*\|_*+\|F(x_k)-F(x_*)-J(x_*)(x_k-x_*)\|\\ &\leq \|x_k-x_*\|_*+\gamma\|x_k-x_*\|\ .\end{aligned}$$

Portanto, usando (5.4.27),

$$\begin{aligned}\|x_{k+1}-x_*\|_* &\leq (1+\mu\gamma)[\eta_k[\|x_k-x_*\|_*+\gamma\|x_k-x_*\|]+2\gamma\|x_k-x_*\|]\\ &\leq (1+\mu\gamma)[\eta_{\max}(1+\mu\gamma)+2\mu\gamma]\|x_k-x_*\|_*\ .\end{aligned}$$

Logo, (5.4.25) segue pela escolha de γ.
Para provarmos o item (b), inicialmente, como na k-ésima etapa de um método de Newton inexato vale (5.4.31), (5.4.26) é equivalente a dizer que

$$\|R_k\|=o(\|F(x_k)\|). \tag{5.4.32}$$

Assim, assumindo (5.4.22), analogamente à prova do item (a), segue que

$$\begin{aligned}\|x_{k+1}-x_*\| &\leq [\|J(x_*)^{-1}\|+\|J(x_k)^{-1}-J(x_*)^{-1}\|]\ [\|R_k\|\\ &\quad +\|J(x_k)-J(x_*)\|\ \|x_k-x_*\|+\|F(x_k)-F(x_*)-J(x_*)(x_k-x_*)\|]\\ &= [\|J(x_*)^{-1}\|+o(1)]\ [o(\|F(x_k)\|)+o(1)\|x_k-x_*\|+o(\|x_k-x_*\|)]\ .\end{aligned}$$

Portanto,

$$\|x_{k+1}-x_*\|=o(\|F(x_k)\|)+o(1)\|x_k-x_*\|+o(\|x_k-x_*\|),$$

ou seja $x_k\to x_*$ superlinearmente. **QED**

Outros critérios, além de (5.3.7), têm sido propostos para a parada do método iterativo linear nos algoritmos de Newton truncados. Ypma [205] sugeriu o seguinte critério baseado no erro verdadeiro do sistema linear, e não no resíduo:

$$\|s_k + J(x_k)^{-1}F(x_k)\| \leq \eta_k \|J(x_k)^{-1}F(x_k)\|. \tag{5.4.33}$$

O critério (5.4.33) tem algumas vantagens teóricas sobre (5.3.7) (ver [141]). No entanto, é mais difícil de implementar devido à necessidade de estimar a solução verdadeira do sistema linear.

Uma desvantagem conceitual dos critérios (5.3.7) e (5.4.33) é que, para obter convergência superlinear, a precisão com que se deve resolver o sistema linear deve ser cada vez mais exigente ($\eta_k \to 0$). Através do uso de precondicionadores que satisfazem a equação secante, esta dificuldade é contornada em [137] e [138].

Chapter 6

Minimização irrestrita e busca linear

A minimização de uma função contínua de n variáveis, sem vínculos, é um dos problemas clássicos da otimização não linear. Existem inúmeras situações da realidade que são modeladas dessa maneira. Quando a função é derivável, a condição necessária de primeira ordem para minimizadores estabelece que o gradiente deve se anular. Em casos muito simples, como os tratados nos textos de cálculo multivariado, é possível calcular manualmente todos os pontos críticos o que, geralmente, leva a encontrar soluções globais, quando estas existem. Mas, quando o número de variáveis ou a complexidade da função aumentam, as manipulações isoladas são insuficientes para achar sequer pontos estacionários. É necessário, então, apelar para métodos numéricos, quase sempre iterativos. Os algoritmos estudados neste capítulo funcionam da seguinte maneira: dado o iterando x_k determina-se uma direção d_k ao longo da qual, em princípio, é possível fazer diminuir o valor da função objetivo. A seguir, calcula-se um comprimento de passo que permita uma diminuição razoável. O método de Newton, os quase-Newton, e os chamados métodos de Newton truncados podem ser adaptados para funcionar com este esquema.

6.1 Algoritmos gerais

Vamos considerar o problema de minimização sem restrições

$$\begin{array}{c}\text{Minimizar } f(x) \\ x \in I\!\!R^n\end{array} \tag{6.1.1}$$

com a hipótese inicial de que $f \in C^1(I\!R^n)$.

Neste capítulo consideraremos sempre que $\| \cdot \|$ é a norma euclidiana, embora muitos resultados sejam independentes dessa identificação. Os métodos para resolver (6.1.1) são iterativos. A aproximação x_{k+1} está bem definida e satisfaz $f(x_{k+1}) < f(x_k)$ se $\nabla f(x_k) \neq 0$. Para a definição desses algoritmos, usaremos direções ao longo das quais, pelo menos dando passos muito pequenos, é possível fazer decrescer $f(x)$. Assim, dado $x \in I\!R^n$, $d \in I\!R^n$ é chamada *direção de descida a partir de* x se existe $\varepsilon > 0$ tal que, para todo $t \in (0, \varepsilon]$,

$$f(x + td) < f(x) \ .$$

As direções que formam um ângulo maior que 90 graus com o gradiente são direções de descida, como vemos no seguinte lema.

Lema 6.1.1
Se $\nabla f(x)^T d < 0$ então d é direção de descida.

Prova: Como $\nabla f(x)^T d = \lim_{t \to 0} \dfrac{f(x + td) - f(x)}{t}$ e por hipótese $\nabla f(x)^T d < 0$, então para todo $t > 0$ suficientemente pequeno, temos $f(x + td) < f(x)$.
QED

A direção $d = -\nabla f(x)$ é chamada *direção de máxima descida a partir de* x. Se consideramos todas as direções com norma euclidiana unitária no espaço, é fácil ver que a derivada direcional mais negativa se realiza nessa direção. A solução do problema

$$\text{Minimizar } \bar{f}(x) \text{ sujeita a } \|x - \bar{x}\| \leq \varepsilon,$$

onde $\bar{f}$ é qualquer função tal que $\nabla \bar{f}(\bar{x}) = \nabla f(\bar{x})$, é um ponto $x(\varepsilon)$ tal que $[x(\varepsilon) - \bar{x}]/\|x(\varepsilon) - \bar{x}\|$ tende à direção de máxima descida quando ε tende a 0.

O protótipo de todos os métodos que veremos neste capítulo é o seguinte algoritmo.

Algoritmo 6.1.2 - Algoritmo básico que usa direções de descida.
Dado $x_k \in I\!R^n$ tal que $\nabla f(x_k) \neq 0$, escolher d_k direção de descida e $t_k > 0$ tais que

$$f(x_k + t_k d_k) < f(x_k) \ .$$

Tomar $x_{k+1} = x_k + t_k d_k$.

Exercício 6.1: Mostrar que o Algoritmo 6.1.2 está bem definido, no sentido de que, sempre que $\nabla f(x_k) \neq 0$, é possível encontrar t_k satisfazendo a condição de descida.

Naturalmente, gostaríamos que a aplicação do Algoritmo 6.1.2 nos levasse sempre, depois de um número razoável de iterações, a um minimizador global de f. Isso não vai ser possível. De fato, o algoritmo assim definido é impotente até para nos conduzir a pontos estacionários no limite. Existem exemplos em uma variável que mostram que a seqüência gerada por ele pode convergir a um ponto não estacionário.

Exercício 6.2: Exibir um exemplo do tipo dos mencionados no parágrafo acima.

Uma das razões pelas quais o Algoritmo 6.1.2 fracassa em encontrar minimizadores ou, até, pontos estacionários, é que pedir apenas que $f(x_k + t_k d_k)$ seja menor que $f(x_k)$ é um objetivo excessivamente modesto, pois, na realidade, um descenso mais enérgico pode ser conseguido ao longo de direções de descida. A chamada "condição de Armijo" substitui o descenso simples e serve para invalidar alguns dos contra-exemplos que podem ser construídos para desqualificar a condição de descenso simples. No seguinte teorema mostramos que a obtenção do descenso baseado na condição de Armijo é sempre possível.

Teorema 6.1.3 - Condição de Armijo.
Sejam $x, d \in I\!R^n$ tais que $\nabla f(x) \neq 0$, $\nabla f(x)^T d < 0$ e $\alpha \in (0,1)$. Existe $\varepsilon = \varepsilon(\alpha) > 0$ tal que

$$f(x + td) \leq f(x) + \alpha t \nabla f(x)^T d \tag{6.1.2}$$

para todo $t \in (0, \varepsilon]$.

Prova: Temos

$$0 \neq \nabla f(x)^T d = \lim_{t \to 0} \frac{f(x + td) - f(x)}{t}$$

e portanto

$$\lim_{t \to 0} \frac{f(x + td) - f(x)}{t \nabla f(x)^T d} = 1.$$

Logo, existe $\varepsilon > 0$ tal que para todo $t \in (0, \varepsilon]$,

$$\frac{f(x+td) - f(x)}{t\nabla f(x)^T d} \geq \alpha.$$

Ou seja, para todo $t \in (0, \varepsilon]$, $f(x+td) \leq f(x) + \alpha t \nabla f(x)^T d$. **QED**

Exercício 6.3: Encontrar um exemplo em uma variável onde a seqüência gerada pelo Algoritmo 6.1.2 tenha pontos de acumulação não-estacionários e onde a condição de Armijo não esteja sendo satisfeita em infinitas iterações.

Incorporando a condição de Armijo, o Algoritmo 6.1.2 pode ser reescrito da seguinte maneira.

Algoritmo 6.1.4 - Algoritmo básico de descida com Armijo.
Dado $\alpha \in (0, 1)$ e dados x_k e d_k tais que $\nabla f(x_k)^T d_k < 0$,
escolher $t_k > 0$ como o maior dos números $\{1, 1/2, 1/4, 1/8, \ldots\}$ tal que

$$f(x_k + t_k d_k) \leq f(x_k) + \alpha t_k \nabla f(x_k)^T d_k \ . \tag{6.1.3}$$

Tomar $x_{k+1} = x_k + t_k d_k$.

Novamente, devemos lamentar que a condição (6.1.3), embora mais exigente que a primeira, não garanta as propriedades desejáveis de um método de minimização. Com efeito, até em uma variável é possível encontrar exemplos para os quais o Algoritmo 6.1.4 converge a um ponto não estacionário. A razão é que, na condição de Armijo, nada impede a tomada de passos excessivamente pequenos, produzindo um fenômeno do tipo "Aquiles e a tartaruga".

Exercício 6.4: Encontrar contra-exemplo em $I\!R$ onde o Algoritmo 6.1.4 convirja a um ponto não-estacionário.

Pode ser que passos muito pequenos sejam inevitáveis, simplesmente porque passos grandes não permitem um decréscimo adequado, mas é imperdoável, do ponto de vista do desenho algorítmico, que passos "grandes" não sejam, pelo menos, tentados. Por isso, decidimos tentar sempre, primeiro o passo $t_k = 1$ e diminuir o passo sem exageros apenas quando a condição de Armijo não é satisfeita. Entretanto, esse mecanismo não inibe, por si só, os passos muito curtos, porque poderia ser que o próprio tamanho de d_k fosse muito pequeno. Isso motiva, também, a introdução de uma condição adicional para d_k, que chamaremos "condição β":

$$\|d_k\| \geq \beta \|\nabla f(x_k)\| \tag{6.1.4}$$

com $\beta > 0$.
A condição de Armijo (6.1.2) e a condição (6.1.4) são suficientes para eliminar os inquietantes contra-exemplos unidimensionais, mas ainda não bastam para garantir que todo ponto de acumulação seja estacionário. De fato, se $n \geq 2$, as direções de descida d_k poderiam ser maldosamente escolhidas de maneira que o ângulo entre d_k e $\nabla f(x_k)$ tendesse a 90 graus. Ou seja, o cosseno entre d_k e $\nabla f(x_k)$, embora negativo, tenderia a zero. Essa situação poderia provocar convergência a um ponto não estacionário. Para inibir essa eventualidade, vamos impor que os citados cossenos estejam uniformemente separados de 0. Logo, as direções toleráveis formarão uma espécie de cone agudo com eixo na semi-reta gerada por $-\nabla f(x_k)$. Por razões óbvias, esta será chamada "condição do ângulo":

$$\nabla f(x_k)^T d_k \leq -\theta \|\nabla f(x_k)\| \, \|d_k\|, \tag{6.1.5}$$

com $\theta \in (0, 1)$ e $\| \cdot \| = \| \cdot \|_2$.

Exercício 6.5: Encontrar um contra-exemplo bi-dimensional mostrando que sob (6.1.2) e (6.1.4) ainda podemos ter convergência a um ponto não-estacionário.

Vamos então reformular o Algoritmo 6.1.4, incorporando as condições (6.1.4) e (6.1.5), desculpando-nos por usar o termo "backtracking" sem traduzir.

Algoritmo 6.1.5 - Algoritmo de descida com backtracking.
Sejam $x_0 \in I\!R^n$, $\alpha \in (0, 1)$, $\beta > 0$, $\theta \in (0, 1)$.
Dado x_k, a nova aproximação x_{k+1} é obtida da seguinte maneira:
(1) Se $\nabla f(x_k) = 0$, parar.

(2) Escolher $d_k \in I\!R^n$ tal que

$$\begin{array}{rcl} \|d_k\| & \geq & \beta \|\nabla f(x_k)\| \\ \nabla f(x_k)^T d_k & \leq & -\theta \|\nabla f(x_k)\| \, \|d_k\| \;. \end{array}$$

(3) $t = 1$.

(4) Enquanto $f(x_k + td_k) > f(x_k) + \alpha t \nabla f(x_k)^T d_k$,
escolher novo $t \in [0.1t, 0.9t]$.

(5) $x_{k+1} = x_k + td_k$.

Exercício 6.6: Mostrar que o Algoritmo 6.1.5 está bem definido.

Computacionalmente, quando a condição de Armijo falha no passo (4) do Algoritmo 6.1.5 para $\bar{t}$, a escolha de um novo $t \in [0.1\bar{t}, 0.9\bar{t}]$ pode ser feita minimizando-se a parábola cúbica que interpola $\varphi(0)$, $\varphi(\bar{t})$, $\varphi'(0)$, $\varphi'(\bar{t})$, onde $\varphi(t) = f(x_k + td_k)$ e $\varphi'(t) = \nabla f(x_k + td_k)^T d_k$. Se o minimizador desta cúbica estiver no intervalo de salvaguarda $[0.1\bar{t}, 0.9\bar{t}]$, adotamos t_{novo} como sendo este minimizador. Caso contrário, $t_{\text{novo}} = 0.5\bar{t}$.

Exercício 6.7: A estratégia descrita acima para obter um novo t após um fracasso em Armijo demanda a avaliação extra de $\nabla f(x_k + \bar{t}d_k)$. Propor uma outra estratégia, usando inicialmente uma parábola interpolante em $\varphi(0)$, $\varphi(\bar{t})$ e $\varphi'(0)$ e então, caso ocorra(m) novo(s) fracasso(s) em Armijo, prosseguir com cúbica(s) interpolante(s) em $\varphi(0)$, $\varphi'(0)$, $\varphi(\bar{t})$ e $\varphi(\bar{\bar{t}})$, onde $\bar{t}$ é o último passo fracassado e $\bar{\bar{t}}$ o passo fracassado anterior.

Antes de passar a resultados teóricos, discutiremos a "naturalidade" das condições (6.1.4) e (6.1.5). Vemos que tanto o parâmetro α da condição de Armijo quanto o parâmetro θ em (6.1.5) são adimensionais. Portanto, faz sentido recomendar valores adequados para esses parâmetros. Usualmente $\alpha = 10^{-4}$ ou 0.1 e $\theta = 10^{-6}$. Já o parâmetro β em (6.1.4) tem dimensão física que depende das unidades das variáveis e da função objetivo, o que torna sua escolha dependente do escalamento do problema. Devemos notar, no entanto, que se $B_k d_k = -\nabla f(x_k)$, então $\|B_k\| \, \|d_k\| \geq \|\nabla f(x_k)\|$ ou seja $\|d_k\| \geq \dfrac{1}{\|B_k\|}\|\nabla f(x_k)\|$. Isto sugere um valor natural para β que é o inverso de uma cota superior para a norma da matriz Hessiana, pois assim o algoritmo não inibe a aceitação da direção de Newton.

Exercício 6.8: Supondo $f \in C^2(I\!R^n)$, mostrar que, se o número de condição da matriz $\nabla^2 f(x_k)$ é uniformemente limitado por c, então $1/c$ é um valor natural para θ quando $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$.

Para o Algoritmo 6.1.5 podemos provar um teorema "de convergência global". O sentido da palavra "global" aqui se refere a que a convergência ocorre independentemente do ponto inicial, e, de maneira nenhuma implica convergência a minimizadores globais.

Teorema 6.1.6 - Convergência Global.
Se x_* é ponto limite de uma seqüência gerada pelo Algoritmo 6.1.5, então $\nabla f(x_*) = 0$.

Prova: Denotamos $s_k = x_{k+1} - x_k = td_k$ para todo $k \in \mathcal{N}$. Seja $K_1 \underset{\infty}{\subsetneq} \mathcal{N}$ tal que $\lim\limits_{k\in K_1} x_k = x_*$, onde $\underset{\infty}{\subsetneq}$ denota subconjunto infinito.

Consideramos dois casos:
(a) $\lim\limits_{k\in K_1} \|s_k\| = 0$.
(b) Existem $K_2 \underset{\infty}{\subsetneq} K_1$ e $\varepsilon > 0$ tais que $\|s_k\| \geq \varepsilon$ para todo $k \in K_2$.

Suponhamos inicialmente que valha (a).
(a1) Se existe $K_3 \underset{\infty}{\subsetneq} K_1$, tal que $s_k = d_k$, então

$$\|\nabla f(x_*)\| = \lim_{k\in K_3} \|\nabla f(x_k)\| \leq \lim_{k\in K_3} \frac{\|d_k\|}{\beta} = \lim_{k\in K_3} \frac{\|s_k\|}{\beta} = 0 \ .$$

(a2) Se para todo $k \in K_1, k \geq k_0$ temos $t < 1$, então, para todo $k \in K_1, k \geq k_0$ existe $\bar{s_k}$ um múltiplo de s_k tal que $\|\bar{s_k}\| \leq 10\|s_k\|$ e

$$f(x_k + \bar{s_k}) > f(x_k) + \alpha \nabla f(x_k)^T \bar{s_k} \ .$$

Claramente,

$$\lim_{k\in K_1} \|\bar{s_k}\| = 0$$

e

$$\nabla f(x_k)^T \bar{s_k} \leq -\theta \|\nabla f(x_k)\| \, \|\bar{s_k}\| \tag{6.1.6}$$

para todo $k \in K_1, k \geq k_0$.
Seja v um ponto de acumulação de $\dfrac{\bar{s_k}}{\|\bar{s_k}\|}$. Então $\|v\| = 1$ e existe $K_4 \underset{\infty}{\subsetneq} K_1$ tal que $\lim\limits_{k\in K_4} \dfrac{\bar{s_k}}{\|\bar{s_k}\|} = v$.

Portanto,

$$\nabla f(x_*)^T v = \lim_{k\in K_4} \nabla f(x_k)^T v = \lim_{k\in K_4} \nabla f(x_k)^T \frac{\bar{s_k}}{\|\bar{s_k}\|}$$

e por (6.1.6) segue que

$$\nabla f(x_*)^T v \leq -\theta \lim_{k\in K_4} \|\nabla f(x_k)\| \ . \tag{6.1.7}$$

Agora, para todo $k \in K_4$,

$$f(x_k + \bar{s_k}) - f(x_k) = \nabla f(x_k + \xi_k \bar{s_k})^T \bar{s_k} \ , \ \ \xi_k \in (0, 1).$$

Portanto, pelo fracasso da condição de Armijo para $\bar{s_k}$,

$$\nabla f(x_k + \xi \bar{s_k})^T \bar{s_k} > \alpha \nabla f(x_k)^T \bar{s_k} \ , \ \ \text{para todo} \ \ k \in K_4 \ .$$

Ou seja, para todo $k \in K_4$,

$$\nabla f(x_k + \xi \bar{s_k})^T \frac{\bar{s_k}}{\|\bar{s_k}\|} > \alpha \nabla f(x_k)^T \frac{\bar{s_k}}{\|\bar{s_k}\|} \ .$$

Passando ao limite para $k \in K_4$ temos:

$$\nabla f(x_*)^T v \geq \alpha \nabla f(x_*)^T v$$

ou

$$(1 - \alpha) \nabla f(x_*)^T v \geq 0 \ .$$

Logo

$$\nabla f(x_*)^T v \geq 0$$

e por (6.1.7) segue que $\nabla f(x_*)^T v = 0$. Se $\nabla f(x_*) \neq 0$, novamente por (6.1.7), para $k \in K_4$, k suficientemente grande,

$$0 = \nabla f(x_*)^T v \leq -\theta \|\nabla f(x_k)\| < 0 \ .$$

Portanto, $\nabla f(x_*) = 0$.

Suponhamos agora a validade de (b): $\|s_k\| \geq \varepsilon$ para todo $k \in K_2$. Por Armijo,

$$\begin{aligned} f(x_k + s_k) &\leq f(x_k) + \alpha \nabla f(x_k)^T s_k \\ &\leq f(x_k) - \alpha\theta \|\nabla f(x_k)\| \, \|s_k\| \\ &\leq f(x_k) - \alpha\theta\varepsilon \|\nabla f(x_k)\| \ , \end{aligned}$$

para todo $k \in K_2$.
Portanto,

$$f(x_{k+1}) - f(x_k) \leq -\alpha\theta\varepsilon \|\nabla f(x_k)\|$$

ou seja,

$$\frac{f(x_k) - f(x_{k+1})}{\alpha\theta\varepsilon} \geq \|\nabla f(x_k)\| \ .$$

Passando ao limite para $k \in K_2$, pela continuidade de f temos: $\lim_{k \in K_2} \|\nabla f(x_k)\| = 0$ e portanto $\nabla f(x_*) = 0$. **QED**

Exercício 6.8' Suponha que, no Algoritmo 6.1.5, temos que existe uma constante $c > 0$ tal que

$$\|d_k\| \leq c\|\nabla f(x_k)\|$$

para todo k.
(a) Provar que se x_* é um ponto limite da seqüência e, além disso, numa vizinhança de x_* não existe nenhum outro ponto onde se anule o gradiente, então a seqüência converge a x_*. Sugerência: construa uma "coroa circular" ao redor de x_* onde somente pode existir um número finito de iterandos.
(b) Provar que se, além do suposto em (a), x_* é um minimizador local, então existe $\varepsilon > 0$ tal que a seqüência converge a x_* sempre que $\|x_0 - x_*\| \leq \varepsilon$. (Convergência local.) Sugerência: construa, além da coroa, um conjunto de nível contido dentro da bola menor.
(c) Mostrar que (b) não se cumpre se, em vez de minimizador local, x_* é meramente um ponto sela. (Exemplo unidimensional.) Apesar disso se cumpre (a)! Discutir estes fatos.

6.2 O método de Newton

No Capítulo 5 apresentamos o método de Newton como um método rápido para resolver sistemas não lineares, com convergência local. Como $\nabla f(x) = 0$ é um sistema não linear, esse método pode ser aplicado e, muitas vezes, dará bons resultados. No entanto, o método de Newton para sistemas não dá preferência a minimizadores sobre maximizadores, já que a condição de otimalidade para ambos tipos de extremos é a mesma. Por outro lado, sabemos, pelo Teorema 6.1.6, quais são os elementos que deve possuir um algoritmo globalmente convergente. É natural, em conseqüência, tentar modificar o método local de maneira que manifeste predileção pelos minimizadores e convirja independentemente do ponto inicial.
Observemos primeiro que, quando as direções d_k são geradas como soluções de um sistema linear $B_k d_k = -\nabla f(x_k)$, temos que $d_k^T B_k d_k = -d_k^T \nabla f(x_k)$, portanto, direções de descida são geradas se $B_k > 0$. Logo, é bastante sensato impor que as matrizes que geram direções de busca em métodos de minimização sejam definidas positivas.
Em continuação descrevemos uma modificação do método de Newton local que o converte em caso particular do Algoritmo 6.1.5. Usaremos a notação

$g(x) = \nabla f(x)$.

Algoritmo 6.2.1 - Newton com busca linear.
Dados $\alpha \in (0,1)$, $\beta > 0$, $\theta \in (0,1)$ e $x_k \in I\!R^n$,

(1) Se $g(x_k) = 0$, parar.

(2) Tentar a fatoração de Cholesky: $\nabla^2 f(x_k) = LDL^T$.

(3) Se houve sucesso em (2), obter d_k resolvendo

$$Lz = -g(x_k) \quad \text{e} \quad DL^T d_k = z \ .$$

(4) Se (2) fracassou, definir $B_k = \nabla^2 f(x_k) + \mu I, \mu > 0$, de maneira que $B_k > 0$. Obter a fatoração de Cholesky: $B_k = \bar{L}\bar{D}\bar{L}^T$ e calcular d_k resolvendo

$$\bar{L}z = -g(x_k) \quad \text{e} \quad \bar{D}\bar{L}^T d_k = z \ .$$

(5) Se $g(x_k)^T d_k > -\theta \|g(x_k)\| \ \|d_k\|$, fazer $\mu \leftarrow \max\{2\mu, 10\}$ e repetir o Passo 4, como se tivesse havido fracasso na fatoração de Cholesky.

(6) Se $\|d_k\| < \beta \|g(x_k)\|$, corrigir:

$$d_k \leftarrow \beta \frac{\|g(x_k)\|}{\|d_k\|} d_k \ .$$

(7) Obter t por "backtracking" de modo a satisfazer

$$f(x_k + td_k) \leq f(x_k) + \alpha t g(x_k)^T d_k,$$

definir

$$x_{k+1} = x_k + td_k$$

e voltar para (1).

Quando a Hessiana $\nabla^2 f(x_k)$ é definida positiva, automaticamente teremos que uma condição de tipo (6.1.5) se verifica com θ igual ao recíproco do número de condição de $\nabla^2 f(x_k)$. Ao mesmo tempo, uma condição de tipo (6.1.4) vale com $\beta = 1/\|\nabla^2 f(x_k)\|$. Logo, se θ e β são escolhidos suficientemente pequenos, as condições (6.1.5) e (6.1.4) serão satisfeitas e passaremos diretamente ao Passo 7 com $d_k = -[\nabla^2 f(x_k)]^{-1} g(x_k)$. Portanto, quase sempre, essa será a direção "de busca" no caso definido positivo. Se a Hessiana

não é definida positiva, no Passo 4 a diagonal é aumentada até conseguir que todos os autovalores sejam maiores que 0. Neste caso, é improvável que a condição (6.1.5) não seja satisfeita, mesmo assim, testamos essa desigualdade e continuamos aumentando a diagonal se ela não vale. Para $\mu \to \infty$ a direção $-B_k^{-1}g(x_k)$ tende a ser a direção de $-g(x_k)$, portanto, mais tarde ou mais cedo, conseguiremos um λ para o qual (6.1.5) se satisfaz. Agora, no processo de aumentar λ, o comprimento de d_k diminui, logo, é necessário testar se (6.1.4) continua valendo. Se assim não for, no Passo 6, aumentamos o tamanho de d_k até atingir uma longitude que garanta (6.1.4).
É interessante observar que, devido aos resultados sobre minimização em bolas do Capítulo 4, a direção $d_k = -[\nabla^2 f(x_k) + \lambda I]^{-1} g(x_k)$ é solução do problema quadrático

$$\text{Minimizar } \frac{1}{2} d^T \nabla^2 f(x_k) d + g(x_k)^T d$$
$$\text{sujeita a } \|d\| \leq \Delta,$$

onde $\Delta = \| -[\nabla^2 f(x_k) + \lambda I]^{-1} g(x_k)\|$. Ou seja, entre todas as direções possíveis cujo comprimento é menor ou igual a $\|d_k\|$, em d_k, a aproximação quadrática de segunda ordem de f toma o valor mínimo .

Exercício 6.9: Viabilizar o Passo 4 do Algoritmo 6.2.1, propondo escolhas para μ que explorem o conhecimento de $\nabla^2 f(x_k)$ (por exemplo, usando os discos de Gerschgorin).

Exercício 6.10: Mostrar que as correções propostas nos passos (5) e (6) do Algoritmo 6.2.1 são satisfatórias. Interpretá-las geometricamente. Expor exemplos numéricos.

Exercício 6.11: "Inventar" o método do gradiente, onde $d_k \equiv -g(x_k)$, e outros métodos globais. Discutir possíveis propriedades.

Vimos acima que, quase sempre, se a Hessiana é definida positiva, a direção produzida pelo Algoritmo 6.2.1 coincidirá com o passo que seria calculado pelo método de Newton local aplicado a $g(x) = 0$. No entanto, isso não significa que esse passo será aceito, já que a condição de Armijo poderia não se cumprir, obrigando a uma ou mais reduções de t. Agora, como o método de Newton local, ou puro, tem convergência muito rápida na proximidade de soluções boas, é desejável que, quando x_k está perto de uma dessas soluções, a condição de Armijo se satisfaça, caso contrário estaríamos rejeitando incrementos essencialmente bons. Felizmente, o método de Newton satisfaz esse requisito, como veremos no seguinte teorema. Usaremos, como hipótese, que $f \in C^3(I\!R^n)$ (na realidade, hipóteses mais fracas são suficientes) para podermos utilizar, de maneira bastante forte, uma fórmula de Taylor com resíduo de segunda ordem.

Teorema 6.2.2
Seja $\{x_k\}$ gerada pelo Algoritmo 6.2.1 com $\alpha \in (0,1)$, x_* um ponto limite de $\{x_k\}$ tal que $\nabla f(x_*) = 0$ e $\nabla^2 f(x_*) > 0$. Então a seqüência converge para x_*. Além disso, existe $\varepsilon > 0$ tal que, se $\|x_k - x_*\| \le \varepsilon$, então

$$f(x_k + d_k) \le f(x_k) + \alpha g(x_k)^T d_k, \tag{6.2.1}$$

com $d_k = -\nabla^2 f(x_k)^{-1} g(x_k)$ e $\alpha \in (0, \frac{1}{2})$.

Prova: Sabemos que x_* é minimizador local estrito de f e, pelo Teorema da Função Inversa, existe uma vizinhança de x_* que não contém soluções de $g(x) = 0$ além de x_*. Seja, então, $\varepsilon_0 > 0$ tal que $f(x) > f(x_*)$ e $g(x) \neq 0$ sempre que $0 < \|x - x_*\| \le \varepsilon_0$. Vejamos primeiro que

$$\lim_{k\to\infty} x_k = x_*, \tag{6.2.2}$$

ou seja, x_* é o único ponto limite da seqüência neste caso. Escrevemos, para simplificar, $B_k = \nabla^2 f(x_k)$. Sejam $\varepsilon_1 \in (0, \varepsilon_0), M > 0$ tais que $\|\nabla^2 f(x)^{-1}\| \le M$ sempre que $\|x - x_*\| \le \varepsilon_1$. Portanto, quando $\|x_k - x_*\| \le \varepsilon_1$, temos $\|B_k^{-1}\| \le M$ e

$$\|x_{k+1} - x_k\| \le \|d_k\| \le \|B_k^{-1}\| \|g(x_k)\| \le M \|g(x_k)\|. \tag{6.2.3}$$

Portanto, pela continuidade de $g(x)$, existe $\varepsilon_2 \le \frac{\varepsilon_1}{2}$ tal que

$$\|x_{k+1} - x_k\| \le \frac{\varepsilon_1}{2} \text{ sempre que } \|x_k - x_*\| \le \varepsilon_2. \tag{6.2.4}$$

Agora, f é contínua na coroa $\varepsilon_2 \le \|x - x_*\| \le \varepsilon_1$. Portanto, atinge um valor mínimo m em algum ponto dessa região. Pela suposição feita sobre ε_0, temos que $m > f(x_*)$. Definimos

$$V = \{x \in I\!R^n \mid \|x - x_*\| < \varepsilon_2 \text{ e } f(x) < m\}. \tag{6.2.5}$$

O conjunto V é uma vizinhança aberta de x_*, portanto, como x_* é um ponto limite de $\{x_k\}$, existem infinitos índices k para os quais $x_k \in V$. Se k_0 é um desses índices, então, por (6.2.4),

$$\|x_{k_0+1} - x_*\| \le \|x_{k_0} - x_*\| + \|x_{k_0+1} - x_{k_0}\| \le \varepsilon_2 + \frac{\varepsilon_1}{2} \le \varepsilon_1. \tag{6.2.6}$$

Ao mesmo tempo, exceto no caso trivial em que $x_{k_0} = x_*$, que podemos analisar por separado,

$$f(x_{k_0+1}) < f(x_{k_0}) < m. \tag{6.2.7}$$

Logo, pela definição de m e pelas desigualdades (6.2.6) e (6.2.7), x_{k_0+1} está na bola de raio ε_1 mas não na coroa definida por ε_1 e ε_2. Ou seja, $\|x_{k_0+1} - x_*\| < \varepsilon_2$. Portanto, por (6.2.7) e (6.2.5), $x_{k_0+1} \in V$. Dessa maneira, o raciocínio indutivo usual nos conduz à conclusão de que $x_k \in V$ para todo $k \geq k_0$. Mas, pela suposição inicial feita sobre ε_0, o único possível ponto limite da seqüência na bola $\|x - x_*\| \leq \varepsilon_2$ é o próprio x_*. Portanto, $\{x_k\}$ converge para x_*, como queríamos provar.
Vamos demonstrar a segunda parte do teorema. Tomando o desenvolvimento de Taylor em torno de x_k,

$$f(x_k + d_k) = f(x_k) + g(x_k)^T d_k + \frac{1}{2}(d_k)^T \nabla^2 f(x_k) d_k + r_2(d_k) \tag{6.2.8}$$

onde $\displaystyle\lim_{d_k \to 0} \frac{r_2(d_k)}{\|d_k\|^2} = 0$.

Como $\nabla^2 f(x_k) d_k = -g(x_k)$, substituindo em (6.2.8) temos:

$$f(x_k + d_k) = f(x_k) - \frac{1}{2}(d_k)^T \nabla^2 f(x_k) d_k + r_2(d_k).$$

Suponhamos, por absurdo, que existe um conjunto infinito de índices K_1 tal que, para todo $k \in K_1$,

$$f(x_k + d_k) > f(x_k) + \alpha g(x_k)^T d_k = f(x_k) - \alpha (d_k)^T \nabla^2 f(x_k) d_k.$$

Então

$$f(x_k) - \frac{1}{2}(d_k)^T \nabla^2 f(x_k) d_k + r_2(d_k) > f(x_k) - \alpha (d_k)^T \nabla^2 f(x_k) d_k \ .$$

Ou seja,

$$r_2(d_k) > \left(\frac{1}{2} - \alpha\right)(d_k)^T \nabla^2 f(x_k) d_k \ .$$

Logo,

$$\frac{r_2(d_k)}{\|d_k\|^2} > \left(\frac{1}{2} - \alpha\right) \frac{(d_k)^T \nabla^2 f(x_k) d_k}{(d_k)^T d_k} \geq \left(\frac{1}{2} - \alpha\right) \lambda_1(k) \tag{6.2.9}$$

onde $\lambda_1(k)$ é o menor autovalor de $\nabla^2 f(x_k)$.

Quando $x_k \to x_*$, $d_k \to 0$ e como os autovalores de uma matriz são funções contínuas das componentes desta matriz, temos que $\lambda_1(k)$ converge a λ_1, o menor autovalor de $\nabla^2 f(x_*)$, que, pela hipótese, é maior que 0.
Logo, passando (6.2.9) ao limite para $k \in K_1$, como como $\alpha \in (0, \frac{1}{2})$, chegamos a uma contradição. Ela veio de supor que podiam existir infinitos índices não satisfazendo a condição (6.2.1). Portanto, além da convergência para x_*, temos que (6.2.1) se cumpre para todo k suficientemente grande.
QED

Exercício 6.12: Se $f(x) = \frac{1}{2}x^T Gx + b^T x + c$, com G simétrica e definida positiva, mostre que a partir de qualquer $x_k \in I\!R^n$ a direção de Newton satisfaz Armijo para $\alpha \leq \frac{1}{2}$.

No Teorema 6.2.2 mostramos que, em determinadas condições, o método de Newton globalizado definido nesta seção, acaba coincidindo com o método de Newton local para o sistema $g(x) = 0$, desfrutando, portanto das mesmas propriedades relativas a velocidade de convergência. Vamos resumir tudo isso no seguinte teorema, cuja demonstração limita-se a organizar os resultados anteriores.

Teorema 6.2.3 - Newton Globalizado.
Seja $\{x_k\}$ a seqüência gerada pelo Algoritmo 6.2.1. Então,

(a) Todo ponto de acumulação é estacionário.

(b) Se $f \in C^3(I\!R^n)$, x_* é um ponto limite tal que $\nabla^2 f(x_*) > 0$, $\beta < 1/\|\nabla^2 f(x_*)\|$ e θ é menor que o inverso do número de condição de $\nabla^2 f(x_*)$, então x_k converge para x_* e existe $k_0 \in \mathcal{N}$ tal que para todo $k \geq k_0, t = 1$.

(c) No caso (b), a convergência é quadrática.

Exercício 6.13: Demonstrar o Teorema 6.2.3.

6.3 Métodos quase-Newton

Vimos que a implementação do método de Newton para minimizar funções exige a resolução, em geral via fatoração de Cholesky, do sistema linear

$$\nabla^2 f(x_k) d_k = -g(x_k) \tag{6.3.1}$$

em cada iteração. Às vezes, mais de uma fatoração é necessária para corrigir falta de positividade da matriz Hessiana. Quando não é possível tirar vantagem da estrutura esparsa da matriz, essa fatoração envolve $O(n^3/6)$ operações. Quando n é grande, esse trabalho pode ser intolerável, o que motiva o desenvolvimento de métodos cujo custo por iteração seja $O(n^2)$. Por outro lado, se as derivadas segundas vão ser calculadas manualmente, a probabilidade de erros humanos é considerável, de maneira que o desenvolvimento de algoritmos sem derivadas segundas também se justifica. Mesmo que o cálculo de derivadas segundas não seja um grande problema, por serem fáceis ou pela disponibilidade de programas de diferenciação automática (ver [105]), é possível que o custo de calcular a matriz Hessiana seja muito elevado. Por exemplo, suponhamos que $f(x)$ seja uma soma de (muitos) quadrados:

$$f(x) = \frac{1}{2}\|F(x)\|^2 = \frac{1}{2}\sum_{i=1}^{m} f_i(x)^2, \tag{6.3.2}$$

com $F : I\!R^n \to I\!R^m, J(x) = F'(x) \in I\!R^{m\times n}$. Nesse caso,

$$\nabla f(x) = J(x)^T F(x), \quad \text{e} \quad \nabla^2 f(x) = J(x)^T J(x) + \sum_{i=1}^{m} f_i(x)\nabla^2 f_i(x).$$

Sem considerar possível esparsidade, o cálculo do gradiente envolve pelo menos $O(mn)$ operações. Mas o cálculo da Hessiana precisa $O(mn^2)$ produtos apenas para calcular $J(x)^T J(x)$, ou seja, sem contar a somatória onde aparecem as Hessianas das f_i que, freqüentemente, é mais complicada. Logo, se m é grande, a diferença de custo entre uma iteração $O(n^2)$ e a iteração newtoniana pode ser significativa.

No método de Newton globalizado com buscas lineares, introduzido na Seção 2, a maioria das iterações tem a forma $x_{k+1} = x_k - t_k \nabla^2 f(x_k)^{-1} g(x_k)$. Como esse método tem boas propriedades de convergência local, é natural que os métodos quase-Newton que pretendemos definir tentem se parecer com ele tanto quanto possível, porém, barateando o custo. Assim, "a maioria" das iterações quase-Newton será da forma

$$x_{k+1} = x_k - t_k B_k^{-1} g(x_k). \tag{6.3.3}$$

A idéia é tentar que as matrizes B_k sejam aproximações razoáveis das Hessianas. Os métodos secantes conseguem, geralmente, aproximações satisfatórias exigindo que as B_k's satisfaçam a "equação secante", cujo significado geométrico vimos no Capítulo 5 e que, no caso de minimização sem restrições, toma a forma

$$B_{k+1}s_k = y_k \text{ onde } s_k = x_{k+1} - x_k \text{ e } y_k = g(x_{k+1}) - g(x_k). \qquad (6.3.4)$$

Uma condição para que um método secante tenha baixo custo é que seja possível obter B_{k+1}^{-1} (ou uma fatoração de B_k) facilmente a partir de B_k, s_k e y_k. "Facilmente" significa, via de regra, com $O(n^2)$ operações. Quase sempre é mais cômodo formular os métodos quase-Newton na forma

$$x_{k+1} = x_k - t_k H_k g(x_k), \qquad (6.3.5)$$

com a matriz H_k de (6.3.5) correspondendo a B_k^{-1} de (6.3.3). Dessa maneira, as H_k podem ser interpretadas como aproximações das inversas das Hessianas e a equação secante toma a forma

$$H_{k+1}y_k = s_k. \qquad (6.3.6)$$

Como no caso do método de Newton, a globalização dos métodos quase-Newton será um caso particular do Algoritmo 6.1.6 com as direções d_k calculadas como $-H_k g(x_k)$ (ou $-B_k^{-1}g(x_k)$).

Algoritmo 6.3.1 - Secante globalizado.
Sejam $\alpha \in (0,1),\ \beta > 0,\ \theta \in (0,1)$.
Dados x_k, B_k(ou H_k) e $g_k = \nabla f(x_k) \neq 0$,

(1) Resolver

$$B_k d_k = -g_k \ \ (\text{ou } \ d_k = -H_k g_k)\ .$$

(2) Testar as condições

$$\|d_k\| \geq \beta\|g_k\| \quad \text{e} \quad g_k^T d_k \leq -\theta\|g_k\| \ \ \|d_k\|,$$

corrigindo d_k se necessário.

(3) Fazer "backtracking" até que

$$f(x_k + td_k) \leq f(x_k) + \alpha t g_k^T d_k.$$

(4) Definir $x_{k+1} = x_k + td_k$, $s_k = x_{k+1} - x_k$, $y_k = g_{k+1} - g_k$ e escolher B_{k+1} tal que $B_{k+1}s_k = y_k$ (ou H_{k+1} tal que $H_{k+1}y_k = s_k$).

A correção para d_k mencionada no Passo 2 é inteiramente arbitrária. Por exemplo, qualquer vetor d_k da forma $-\gamma g(x_k)$, com $\gamma \geq \beta$ satisfará, obviamente, as condições (6.1.4) e (6.1.5). Mas, em casos particulares, correções mais inteligentes podem ser tentadas.

Exercício 6.14: Inventar outras correções para d_k no Passo 2 do Algoritmo 6.3.1, de maneira de aproveitar melhor a informação contida na aproximação B_k (ou H_k).

Vamos introduzir fórmulas que satisfazem () ou () e, portanto, geram métodos secantes. Em $I\!R$, existe uma única possibilidade: $B_{k+1} = y_k/s_k$ ou $H_{k+1} = s_k/y_k$. Em geral, qualquer matriz B_{k+1} cumprindo () pertence à variedade afim $Bs_k = y_k$ em $I\!R^{n\times n}$. Pelo mesmo argumento usado em sistemas não lineares, esta variedade é não vazia e, portanto, tem infinitos elementos se $n \geq 2$.

Por razões que veremos mais adiante, é muito freqüente obter B_{k+1} a partir de B_k mediante uma atualização de posto dois. Nesse caso,

$$B_{k+1} = B_k + \Delta B'_k + \Delta B''_k$$

e como $B_{k+1}s_k = y_k$, segue que

$$(B_k + \Delta B'_k + \Delta B''_k)s_k = y_k$$

ou seja,

$$\Delta B'_k s_k + \Delta B''_k s_k = y_k - B_k s_k \tag{6.3.7}$$

Existem muitas maneiras da equação (6.3.7) ser satisfeita. Por exemplo, se $\Delta B'_k s_k = y_k$ e $\Delta B''_k s_k = -B_k s_k$, e impomos que $B_k, \Delta B'_k$ e $\Delta B''_k$ sejam simétricas, temos a seguinte atualização:

$$\Delta B'_k = \frac{y_k y_k^T}{y_k^T s_k} \qquad \text{e} \qquad \Delta B''_k = -\,\frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}\,.$$

Dessa maneira, obtemos a seguinte fórmula secante:

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}\,. \tag{6.3.8}$$

A escolha (6.3.8) é conhecida como *fórmula BFGS*, descoberta independentemente por Broyden, Fletcher, Goldfarb e Shanno em 1970. É a atualização secante mais popular para minimização sem restrições.

Exercício 6.15: Provar que, na fórmula BFGS,

$$B_{k+1}^{-1} = B_k^{-1} + \frac{(s_k - B_k^{-1}y_k)s_k^T + s_k(s_k - B_k^{-1}y_k)^T}{s_k^T y_k} - \frac{(s_k - B_k^{-1}y_k)^T y_k s_k s_k^T}{(s_k^T y_k)^2}.$$

Tendo em vista o Exercício 6.15, a formulação dual da fórmula BFGS efetivamente usada é:

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)s_k^T + s_k(s_k - H_k y_k)^T}{s_k^T y_k} - \frac{(s_k - H_k y_k)^T y_k s_k s_k^T}{(s_k^T y_k)^2}. \tag{6.3.9}$$

Em (6.3.9) observamos que a obtenção de H_{k+1} a partir de H_k (ou B_{k+1}^{-1} a partir de B_k^{-1}) demanda apenas $O(n^2)$ operações, como desejávamos.

Exercício 6.16: Utilizando a mesma heurística usada na obtenção da fórmula BFGS, mas trabalhando inicialmente na formulação dual (matrizes H), "inventar" a fórmula DFP (introduzida por Davidon em 1959 e estudada por Fletcher e Powell em 1963).

A fórmula BFGS e a DFP têm a propriedade de produzir, geralmente, matrizes definidas positivas e, portanto, direções de descida, que, freqüentemente, não precisarão correção. A condição suficiente para tão interessante propriedade é dada no seguinte teorema.

Teorema 6.3.2
Na fórmula BFGS (6.3.8), se B_k é simétrica definida positiva e $s_k^T y_k > 0$, então B_{k+1} também é simétrica e definida positiva.

Prova: Seja $z \neq 0, z \in I\!R^n$. Então

$$z^T B_{k+1} z = z^T B_k z + \frac{(z^T y_k)^2}{y_k^T s_k} - \frac{(z^T B_k s_k)^2}{s_k^T B_k s_k} ,$$

onde $z^T B_k z > 0$ e $\dfrac{(z^T y_k)^2}{y_k^T s_k} \geq 0$. Agora, chamando

$$a = z^T B_k z - \frac{(z^T B_k s_k)^2}{s_k^T B_k s_k} = \frac{s_k^T B_k s_k z^T B_k z - (z^T B_k s_k)^2}{s_k^T B_k s_k},$$

temos que, pela desigualdade de Cauchy-Schwarz, que $a \geq 0$.

Na verdade, $a = 0$ apenas quando z é múltiplo de s_k, mas neste caso, $z^T y_k \neq 0$ e portanto $\frac{(z^T y_k)^2}{s_k^T y_k} > 0$. Logo $z^T B_{k+1} z > 0$. **QED**

Exercício 6.17: Enunciar e provar o resultado análogo ao Teorema 6.3.2 para a fórmula DFP.

O significado de $s_k^T y_k > 0$ precisa ser desvendado. Temos $s_k^T y_k = s_k^T (g_{k+1} - g_k) = s_k^T g(x_k + td_k) - s_k^T g(x_k) = \varphi'(t) - \varphi'(0)$, onde $\varphi(t) = f(x_k + td_k)$. Ou seja, quando $s_k^T y_k > 0$ o passo que acabou satisfazendo (6.1.3) é tal que $\varphi'(t) > \varphi'(0)$. Em outras palavras, a derivada direcional de f na direção de d_k é maior no ponto x_{k+1} que no ponto x_k. É fácil ver que essa condição é satisfeita automaticamente, por exemplo, se a função f é convexa ao longo da direção d_k.

Tanto a fórmula DFP quanto a BFGS satisfazem outra propriedade importante, que foi bastante destacada nos primórdios dos métodos quase-Newton (ver [70]): quando aplicados à minimização de uma quadrática com Hessiana definida positiva e com o passo t calculado como o minimizador da função ao longo da direção d_k, a convergência ao minimizador da quadrática é obtida em no máximo n iterações. Sabe-se, por outro lado, que a fórmula BFGS é preferível à DFP, o que foi verificado experimentalmente ao longo dos anos, e parcialmente explicado do ponto de vista teórico por Powell e outros. Ver [165] e [157]. A teoria de convergência de algoritmos baseados na fórmula BFGS ainda apresenta pontos não elucidados. O Algoritmo 6.3.3 é uma implementação de um esquema BFGS como caso particular do esquema geral da primeira seção deste capítulo, onde, simplesmente, as direções que não satisfazem (6.1.4) e (6.1.5) são descartadas. Com a geração BFGS é possível observar na prática que esse descarte é extremamente raro.

Algoritmo 6.3.3 - BFGS globalizado.
Sejam $\alpha \in (0,1)$, $\beta > 0$, $\theta \in (0,1)$, $x_0 \in I\!R^n$, $H_0 = H_0^T$, $H_0 > 0$ (p. ex., $H_0 = I$).
Dados x_k, H_k e $g_k = \nabla f(x_k) \neq 0$,

(1) $d_k = -H_k g_k$.

(2) Se $(g_k^T d_k > -\theta \|g_k\| \, \|d_k\|)$, substituir d_k por $-g_k$ e H_k por I. Se $(\|d_k\| < \beta \|g_k\|)$ substituir d_k por $\beta \|g_k\| d_k / \|d_k\|$

(3) Fazer "backtracking" até que

$$f(x_k + td_k) \leq f(x_k) + tg_k^T d_k.$$

(4) $x_{k+1} = x_k + td_k, \quad s_k = x_{k+1} - x_k, \quad y_k = g_{k+1} - g_k.$
Se $s_k^T y_k \leq 0$, então $H_{k+1} = H_k$
caso contrário,

$$\begin{aligned} H_{k+1} \;&=\; H_k \;+\; \frac{(s_k - H_k y_k)s_k^T + s_k(s_k - H_k y_k)^T}{s_k^T y_k} \\ &\quad - \frac{(s_k - H_k y_k)^T y_k s_k s_k^T}{(s_k^T y_k)^2}\,. \end{aligned}$$

Exercício 6.18: Uma outra fórmula secante é obtida projetando-se B_k na variedade $Bs_k = y_k$ segundo a norma de Frobenius (ver exercício 5.3). Determinar esta atualização, conhecida como *primeiro método de Broyden*, mostrando que:

(a) $B_{k+1} = B_k + \dfrac{(y_k - B_k s_k)s_k^T}{s_k^T s_k}.$

(b) $B_{k+1}^{-1} = B_k^{-1} + \dfrac{(s_k - B_k^{-1} y_k)s_k^T B_k^{-1}}{s_k^T B_k^{-1} y_k}$, ou seja,

$H_{k+1} = H_k + \dfrac{(s_k - H_k y_k)s_k^T H_k}{s_k^T H_k y_k}.$

(c) $\|B_{k+1} - B_k\|_2 \leq \|B - B_k\|_2$ para toda $B \in I\!R^{n\times n}$ tal que $Bs_k = y_k$.

Exercício 6.19: Para $A \in I\!R^{n\times n}$, mostrar que $\frac{1}{2}(A + A^T)$ é a matriz simétrica mais próxima de A na norma de Frobenius.

Exercício 6.20: Seguindo a mesma idéia do primeiro método de Broyden (Exercício 6.18), mas impondo também simetria, encontrar a *fórmula PSB* ("Powell symmetric Broyden", [162]):

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T + s_k(y_k - B_k s_k)^T}{s_k^T s_k} - \frac{(y_k - B_k s_k)^T s_k s_k s_k^T}{(s_k^T s_k)^2}\,.$$

Exercício 6.21:

(a) Construir a fórmula PSB tipo H.

(b) Infelizmente, a atualização PSB nem sempre gera matrizes definidas positivas. Mostrar que numa vizinhança de x_* tal que $\nabla^2 f(x_*) > 0$, se $B_k > 0$, B_{k+1} dada pela fórmula PSB também é definida positiva.

De maneira análoga ao que fizemos para obter a fórmula BFGS, também podemos determinar uma atualização secante simétrica e de posto unitário. Queremos $B_{k+1}s_k = y_k$, onde $B_{k+1} = B_k + \Delta B_k$. Então, $(B_k + \Delta B_k)s_k = y_k$, ou seja $\Delta B_k s_k = y_k - B_k s_k$. Para que haja simetria, fazemos:

$$\Delta B_k = \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k} \, .$$

Obtemos assim a fórmula chamada *Atualização simétrica de posto um*,

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k} \, . \tag{6.3.10}$$

Exercício 6.22: Mostrar que a formulação dual para a atualização simétrica de posto um é dada por:

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{(s_k - H_k y_k)^T y_k} \, .$$

A atualização simétrica de posto um não gera necessariamente matrizes definidas positivas, e, tampouco há garantia de que o denominador de (6.3.10) seja diferente de zero. Isto sugere que esta atualização é propensa a severa instabilidade numérica. Entretanto, os resultados práticos obtidos são surpreendentemente bons. A descoberta de uma teoria explicativa para o comportamento desta fórmula ainda constitui um desafio. A atualização de posto um foi reinventada várias vezes por diversos autores e já aparecia no artigo pioneiro de Davidon em 1959. Um resultado muito interessante para funções quadráticas é dado no seguinte teorema.

Teorema 6.3.4
Se $f(x) = \frac{1}{2}x^T Gx + b^T x + c$, $G > 0$, se a fórmula (6.3.10) está bem definida em todas as iterações, se os incrementos são linearmente independentes e se o passo $t \equiv 1$ é usado para todo k, então $H_n = G^{-1}$, e portanto, x_{n+1} é

a solução.

Exercício 6.23: Provar o Teorema 6.3.4 (ver, por exemplo, [122]).

Chegamos ao ponto em que é necessário compatibilizar os métodos quase-Newton "locais", estudados no Capítulo 5, que, via de regra, tem convergência superlinear, com a globalização introduzida nos algoritmos 6.3.1 e 6.3.3. Esses algoritmos são casos particulares do Algoritmo 6.1.6, e, portanto, são globalmente convergentes no sentido de que todo ponto limite de uma seqüência gerada por qualquer um deles deve ser estacionário. No entanto, essa propriedade global está baseada nas salvaguardas tomadas para que (6.1.4) e (6.1.5) sejam satisfeitas, e não nas características próprias dos métodos secantes. Como no caso do método de Newton globalizado, seria interessante que, em circunstâncias bem definidas, as iterações puramente locais e as globais fossem as mesmas, para que o método global possa desfrutar da velocidade de convergência do local. No seguinte teorema, resolvemos parcialmente esse problema.

Teorema 6.3.5
Seja $x_* \in I\!R^n$ tal que $\nabla f(x_*) = 0$, $f \in C^3(I\!R^n)$, $\nabla^2 f(x_*) > 0$. Suponhamos que x_* é um ponto limite da seqüência infinita $\{x_k\}$, gerada pelo Algoritmo 6.3.1 com $\alpha \in (0, \frac{1}{2})$, que as condições (6.1.4) e (6.1.5) são sempre satisfeitas por $d_k = -B_k^{-1} g(x_k)$ (ou $d_k = -H_k g(x_k)$ na formulação dual), as matrizes B_k^{-1} (H_k) estão uniformemente limitadas ($\|B_k^{-1}\| \leq M$ ou $\|H_k\| \leq M$ para todo k) e que $\lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x_*)]d_k\|}{\|d_k\|} = 0$ (condição Dennis-Moré). Então,
(a) A seqüência $\{x_k\}$ converge para x_*;
(b) existe $\varepsilon > 0$ tal que, se $\|x_k - x_*\| \leq \varepsilon$,

$$f(x_k + d_k) \leq f(x_k) + \alpha g_k^T d_k,$$

(c) a convergência é superlinear.

Prova: Pela hipótese de limitação uniforme de $\|B_k^{-1}\|$ (ou $\|H_k\|$) a convergência de $\{x_k\}$ para x_* segue exatamente como no Teorema 6.2.2. Suponhamos, por um momento, que (b) se satisfaz. Então, para k suficientemente grande, não é necessário "backtracking" e $t = 1$ é sempre o passo aceito. Assim, para esses valores de k, o algoritmo é um quase-Newton puro que satisfaz a condição Dennis-Moré. Portanto, a convergência superlinear resulta do Teorema Dennis-Moré, provado no Capítulo 5.

Em conseqüência, somente precisamos provar (b).
A expansão de Taylor para f em torno de x_k é dada por:

$$f(x_k + d_k) = f(x_k) + g_k^T d_k + \frac{1}{2} d_k^T \nabla^2 f(x_k) d_k + r_2(d_k) \tag{6.3.11}$$

onde $\lim_{d_k \to 0} \frac{r_2(d_k)}{\|d_k\|^2} = 0$.
Como $B_k d_k = -g_k$, segue que $g_k^T d_k = -d_k^T B_k d_k$ e, substituindo em (6.3.11) temos:

$$f(x_k + d_k) = f(x_k) - d_k^T B_k d_k + \frac{1}{2} d_k^T \nabla^2 f(x_k) d_k + r_2(d_k) \ . \tag{6.3.12}$$

Suponhamos por absurdo, como no Teorema (6.2.9), que existe um conjunto infinito de índices K_1 tal que, para todo $k \in K_1$,

$$f(x_k + d_k) > f(x_k) + \alpha g_k^T d_k = f(x_k) - \alpha d_k^T B_k d_k \ .$$

Então,

$$f(x_k) - d_k^T [B_k - \nabla^2 f(x_k)] d_k - \frac{1}{2} d_k^T \nabla^2 f(x_k) d_k + r_2(d_k)$$

$$> f(x_k) - \alpha d_k^T [B_k - \nabla^2 f(x_k)] d_k - \alpha d_k^T \nabla^2 f(x_k) d_k \ .$$

Ou seja,

$$\frac{r_2(d_k)}{\|d_k\|^2} > (1-\alpha) \frac{d_k^T}{\|d_k\|} (B_k - \nabla^2 f(x_k)) \frac{d_k}{\|d_k\|} + \left(\frac{1}{2} - \alpha\right) \frac{d_k^T \nabla^2 f(x_k) d_k}{d_k^T d_k} \ .$$

Portanto,

$$\frac{r_2(d_k)}{\|d_k\|^2} \geq (1-\alpha) \frac{d_k^T}{\|d_k\|} (B_k - \nabla^2 f(x_k)) \frac{d_k}{\|d_k\|} + \left(\frac{1}{2} - \alpha\right) \lambda_1(k) \ . \tag{6.3.13}$$

Tomando limites para $k \in K_1$ em ambos membros de (6.3.13), usando a condição Dennis-Moré da hipótese do teorema, e a continuidade dos autovalores, obtemos

$$0 = \lim_{k \in K_1} \frac{r_2(d_k)}{\|d_k\|^2} \geq (\frac{1}{2} - \alpha) \lambda_1,$$

onde λ_1 é o menor autovalor de $\nabla^2 f(x_*)$. Isto é uma contradição, porque, por hipótese $\alpha < 1/2$ e a Hessiana em x_* é definida positiva. **QED**

O resultado acima não prova a superlinearidade dos algoritmos 6.3.1 ou 6.3.3. Como vimos no Capítulo 5, a condição Dennis-Moré pode ser deduzida da equação secante e da propriedade $\lim_{k\to\infty} \|B_{k+1} - B_k\| = 0$, mas esta propriedade precisa ser provada para métodos secantes específicos. No entanto, o Teorema 6.3.5 provoca o sentimento de que, em muitos casos, os métodos de minimização caracterizados pela condição secante serão superlinearmente convergentes.

6.4 Métodos de Newton truncados com busca linear

Vimos que, para calcular a direção de busca, o método de Newton precisa resolver um sistema linear, o que demanda $O(n^3/6)$ operações no caso denso, e que o cálculo da direção nos quase-Newton envolve $O(n^2)$ operações. Quando n é grande e a Hessiana é esparsa, o método de Newton pode ser implementado através de fatorações de Cholesky que aproveitem a esparsidade da matriz, armazenando apenas os elementos não-nulos. Também existem implementações de métodos quase-Newton para problemas de grande porte. Nesse caso, em vez de armazenar as matrizes H_k (da formulação dual) são guardados os últimos vetores que contribuem para a definição da atualização, descartando os antigos. Essas implementações se dizem "de memória limitada". Ver [157].

A última alternativa é usar um método iterativo para resolver o sistema linear (6.3.1). Neste caso, o método geralmente recomendado é o de gradientes conjugados, devido à matriz ser simétrica e, muitas vezes, definida positiva. Como no caso de resolução de sistemas, falaremos, neste caso, de métodos de Newton truncados. No entanto, os métodos de Newton truncados com busca linear não desfrutam de grande prestígio no contexto da minimização irrestrita. A razão é, provavelmente, que um tipo diferente de globalização, baseado em *regiões de confiança*, se adapta melhor à resolução iterativa de (6.3.1) que as buscas lineares. Por isso, nos limitaremos aqui a definir um possível método de Newton truncado com buscas lineares e deixaremos suas propriedades para serem analisadas pelo leitor.

Algoritmo 6.4.1 - Newton truncado globalizado.
Sejam $\alpha \in (0,1)$, $\beta > 0$, $\theta \in (0,1)$ e $\eta_k \in (0,1)$ para todo $k = 0, 1, 2, \ldots$.

(1) Dado $x_k \in I\!R^n,\ \nabla f(x_k) \neq 0$, obter d_k satisfazendo:

$$\frac{1}{2} d_k^T \nabla^2 f(x_k) d_k + g(x_k)^T d_k < 0$$

e

$$\|\nabla^2 f(x_k) d_k + \nabla f(x_k)\| \leq \eta_k \|g(x_k)\| \ .$$

(2) Se o cálculo de d_k nas condições acima não é possível num tempo razoável, ou $\|d_k\| < \beta \|\nabla f(x_k)\|$, ou $\nabla f(x_k)^T d_k > -\theta \|\nabla f(x_k)\| \ \|d_k\|$ substituir d_k por $-\nabla f(x_k)$.

(3) Fazer "backtracking" até que

$$f(x_k + t d_k) \leq f(x_k) + t \nabla f(x_k)^T d_k \ .$$

(4) $x_{k+1} = x_k + t d_k$ e voltar para (1).

Exercício 6.26: Analise as propriedades do Algoritmo 6.4.1.

Chapter 7

Regiões de confiança

No Capítulo 5 estudamos, para certo tipo de problemas complexos, o processo iterativo de resolução que consiste em (a) montar um modelo simples do problema original, baseado na informação disponível no ponto atual x_k e (b) definir x_{k+1} como a solução deste modelo.

No Capítulo 6, demos um passo adiante: consideramos a possibilidade de que a solução do modelo simples não fosse suficientemente boa, sendo portanto rejeitada e substituída por uma nova aproximação x_{k+1}, um ponto no segmento cujos extremos são x_k e a solução recusada, produzido pelo processo de "backtracking".

O "backtracking", como outros procedimentos de busca linear, é muito simples e, freqüentemente, efetivo. Entretanto, ele representa uma quebra da filosofia baseada em (a) e (b). De fato, o primeiro ponto tentado nos algoritmos newtonianos do Capítulo 6 é o minimizador de um modelo bastante natural baseado geralmente na fórmula de Taylor, mas os pontos tentados depois da primeira rejeição não podem ser interpretados da mesma maneira. Na realidade, conservando-nos no segmento $[x_k, \text{ponto rejeitado}]$, estamos optando por uma fidelidade parcial ao primeiro subproblema, o que não é fácil de se justificar pois, afinal de contas, sua solução foi descartada.

Os métodos de regiões de confiança, pelo contrário, são radicalizações do esquema (a)–(b). Neles, quando o minimizador do primeiro modelo é recusado, a opção escolhida é modificar o subproblema diminuindo seu domínio de definição e calcular a próxima tentativa como a solução do *novo* subproblema. Assim, o segmento determinado pela primeira rejeição é imediatamente abandonado, com um aumento óbvio no custo, já que esse processo é mais caro.

Contrariamente aos métodos com busca linear, os algoritmos de regiões de confiança se adaptam com bastante naturalidade a diversos problemas com restrições, como veremos no contexto deste capítulo.

7.1 Algoritmo geral

Consideramos o problema genérico de otimização:

$$\begin{array}{c} \text{Minimizar } f(x) \\ x \in \Omega \ , \end{array} \tag{7.1.1}$$

onde Ω é um subconjunto arbitrário de $I\!R^n$. A idéia básica é, a cada iteração, construir uma aproximação quadrática para a função objetivo em torno do ponto atual x_k:

$$f(x) \approx \psi_k(x) \equiv f(x_k) + g(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k) \tag{7.1.2}$$

onde $g(x_k) = \nabla f(x_k)$ e $B_k \in I\!R^{n\times n}$ é simétrica.

Como o modelo quadrático (7.1.2) deixa de ser representativo à medida que x se afasta de x_k, podemos confiar em aproximar $f(x)$ por $\psi_k(x)$ numa vizinhança de x_k, ou seja, no conjunto:

$$\{x \in \Omega \mid \|x - x_k\| \leq \Delta\}, \tag{7.1.3}$$

onde $\Delta > 0$ e $\|\cdot\|$ é uma norma qualquer em $I\!R^n$.

Dessa forma, o minimizador de ψ_k na região (7.1.3) seria uma boa aproximação para o minimizador de f nesta mesma região. No entanto, se o valor de f no minimizador de ψ_k não é suficientemente menor que $f(x_k)$ reduzimos o raio Δ e definimos um novo subproblema com o domínio menor.

O algoritmo conceitual a seguir sistematiza essas idéias.

Algoritmo 7.1.1 - Regiões de Confiança.

Fixar $\Delta_{\min} > 0, \ \alpha \in (0,1), \ x_0 \in \Omega$ dado.

(1) Escolher $\Delta \geq \Delta_{\min}$ e B_k simétrica.
Definir $\psi_k(x) = f(x_k) + g(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k)$.

(2) Encontrar $\overline{x}$ minimizador aproximado de $\psi_k(x)$
sujeito a $x \in \Omega, \|x - x_k\| \leq \Delta$.

(3) Se $f(\overline{x}) \leq f(x_k) + \alpha[\psi_k(\overline{x}) - \psi_k(x_k)]$,
definir $x_{k+1} = \overline{x}$ e terminar a iteração.
Senão, escolher $\Delta_{\text{novo}} \in [0.1\|\overline{x} - x_k\|, 0.9\Delta]$, $\Delta \leftarrow \Delta_{\text{novo}}$ e voltar para (2).

Na forma apresentada, o algoritmo de regiões de confiança se aplica a qualquer problema de otimização, com ou sem restrições. No entanto, os subproblemas de minimizar ψ_k em (7.1.3) podem ser mais difíceis que o problema original, circunstância que é atenuada pela expressão "minimizador aproximado", usada no Passo 2. O raio original da região de confiança na iteração k sempre é maior ou igual a um raio fixo Δ_{min}. Isto representa a necessidade de, pelo menos na primeira tentativa, sermos suficientemente arrojados para não ficarmos com passos muito curtos. Mais ainda, o requisito $\Delta \geq \Delta_{min}$ facilita as provas de convergência, mas não é essencial na metodologia de regiões de confiança. O critério de aceitação da solução do subproblema é dado no Passo 3. Nele se estabelece que a diminuição de f deve ser pelo menos uma fração da diminuição do modelo ψ_k. Usualmente, escolhe-se $\alpha = 0.1$. Existem muitas regras práticas para definir o valor de Δ no começo de cada iteração, em função do êxito ou fracasso na iteração anterior. A idéia é que, se a iteração anterior foi muito bem sucedida, no sentido de que a função objetivo diminuiu quase tanto ou mais que o modelo quadrático, este merece mais confiança e, conseqüentemente, Δ deve ser aumentado. Via de regra, para a definição de Δ_{novo} no Passo 3, são usados procedimentos muitos simples, por exemplo, $\Delta_{novo} = \|\overline{x} - x_k\|/2$.

O algoritmo de regiões de confiança foi analisado com esta generalidade em [142] e [144]. Nas seções seguintes, estudaremos a aplicação desse método para dois tipos de região factível: $I\!R^n$ e caixas $n-$dimensionais.

7.2 Método de Newton

No Capítulo 6 estudamos a globalização por "backtracking" do método de Newton para o problema de minimização sem restrições:

$$\begin{array}{c} \text{Minimizar } f(x) \\ x \in I\!R^n . \end{array} \tag{7.2.1}$$

Vimos que, com as salvaguardas necessárias, o método desfruta das propriedades de convergência global a pontos estacionários de primeira ordem do algoritmo genérico 6.1.5. O esquema de regiões de confiança proporciona uma maneira muito mais natural de globalizar o método de Newton, com a conservação de subproblemas newtonianos para a determinação de tentativas depois de eventuais fracassos. Além disso, o novo procedimento permite um resultado extremamente atraente: os pontos limite são pontos críticos de primeira e segunda ordem.

Algoritmo 7.2.1 - Newton com regiões de confiança.
Fixar $\Delta_{\min} > 0, \alpha \in (0,1)$. Dado $x_0 \in I\!R^n$.

(1) Escolher $\Delta \geq \Delta_{\min}$, calcular $B_k = \nabla^2 f(x_k)$.

(2) Definir $\overline{x}$ como minimizador global de $\psi_k(x)$ sujeito a $\|x - x_k\| \leq \Delta$.

(3) Se $f(\overline{x}) \leq f(x_k) + \alpha(\psi_k(\overline{x}) - \psi_k(x_k))$,
definir $x_{k+1} = \overline{x}$, $\Delta_k = \Delta$ e terminar a iteração.
Senão, escolher $\Delta_{\text{novo}} \in [0.1\|\overline{x} - x_k\|, 0.9\Delta]$, $\Delta \leftarrow \Delta_{\text{novo}}$ e voltar para (2).

O subproblema do Passo 2 consiste em encontrar um minimizador global da quadrática ψ_k na bola $\|x - x_k\| \leq \Delta$. Para uma norma arbitrária, este problema pode ser bastante difícil. No entanto, quando $\| \cdot \|$ é a norma euclidiana, maneiras relativamente simples de resolvê-lo são conhecidas. No capítulo 4 estudamos essa situação com alguma atenção e vimos que $\overline{x}$ pode ser calculada com o custo de algumas fatorações de Cholesky de matrizes da forma $B_k + \mu I$. De fato, apesar de no Passo 2 falarmos de minimizador global "exato" do subproblema, o algoritmo iterativo Moré-Sorensen, geralmente usado, permite certo grau de inexatidão, no sentido de que as sucessivas iterações x^ℓ são soluções exatas de problemas da forma

$$\text{Minimizar } \psi_k(x) \text{ sujeita a } \|x - x_k\| \leq \Delta^\ell,$$

onde $\Delta^\ell \to \Delta$. Como a escolha de Δ no Passo 1 ou no Passo 3 não é rígida, podemos suspender o processo iterativo quando, digamos, $|\Delta^\ell - \Delta| \leq 0.1\Delta$, e redefinir, posteriormente, $\Delta \leftarrow \Delta^\ell$. Dessa maneira, o número de fatorações de Cholesky invocadas pelo método Moré-Sorensen fica bastante moderado. No entanto, é evidente que o custo deste processo é bem maior que o "backtracking".

A seguir vamos mostrar que, a menos que x_k seja um ponto estacionário de segunda ordem, a próxima iteração x_{k+1} está bem definida e satisfaz $f(x_{k+1}) < f(x_k)$. Este será um passo prévio à prova de que todo ponto limite é estacionário de segunda ordem. Ao longo desta seção supomos que

$f \in C^2(I\!R^n)$. Como em outros capítulos, denotamos $g(x) = \nabla f(x)$.

Teorema 7.2.2 - Boa definição.

Se x_k não é um ponto estacionário de segunda ordem de (7.2.1) então x_{k+1} está bem definido e $f(x_{k+1}) < f(x_k)$.

Prova: Se x_k não é estacionário de segunda ordem de (7.2.1), então

$$g(x_k) \neq 0 \tag{7.2.2}$$

ou

$$g(x_k) = 0 \text{ mas } \nabla^2 f(x_k) \not\geq 0 \,. \tag{7.2.3}$$

Suponhamos inicialmente que $g(x_k) \neq 0$. Seja $d \in I\!R^n$ tal que $\|d\| = 1$ e

$$g(x_k)^T d < 0. \tag{7.2.4}$$

Seja $\overline{x}(\Delta)$ minimizador de $\psi_k(x)$ sujeita a $\|x - x_k\| \leq \Delta$. Para simplificar, escreveremos $\overline{x} = \overline{x}(\Delta)$. Como $\|\Delta d\| = \Delta$, temos:

$$\psi_k(\overline{x}) \leq \psi_k(x_k + \Delta d) = f(x_k) + g(x_k)^T \Delta d + \frac{1}{2} \Delta d^T \nabla^2 f(x_k) \Delta d \,.$$

Ou seja,

$$\psi_k(\overline{x}) - f(x_k) \leq g(x_k)^T \Delta d + \frac{1}{2} \Delta d^T \nabla^2 f(x_k) \Delta d \leq g(x_k)^T \Delta d + \frac{\|\nabla^2 f(x_k)\| \ \Delta^2}{2} \,.$$

Logo, como $f(x_k) = \psi_k(x_k)$,

$$\frac{\psi_k(\overline{x}) - \psi_k(x_k)}{\Delta} \leq g(x_k)^T d + \frac{\|\nabla^2 f(x_k)\|}{2} \, \Delta.$$

Portanto, existe $\overline{\Delta} > 0$ tal que para $\Delta \leq \overline{\Delta}$,

$$\frac{\psi_k(\overline{x}) - \psi_k(x_k)}{\Delta} \leq \frac{g_k^T d}{2} = a < 0. \tag{7.2.5}$$

Definimos

$$\rho(\Delta) = \frac{f(\overline{x}) - f(x_k)}{\psi_k(\overline{x}) - \psi_k(x_k)} \tag{7.2.6}$$

e então, de (7.2.5) temos

$$\begin{aligned}
|\rho(\Delta)-1| &= \left|\frac{f(\overline{x})-f(x_k)-|[\psi_k(\overline{x})-\psi_k(x_k)]}{\psi_k(\overline{x})-\psi_k(x_k)}\right| = \left|\frac{f(\overline{x})-\psi_k(\overline{x})}{\psi_k(\overline{x})-\psi_k(x_k)}\right| \\
&= \left|\frac{f(\overline{x})-f(x_k)-g(x_k)^T(\overline{x}-x_k)-\frac{1}{2}(\overline{x}-x_k)^T\nabla^2 f(x_k)(\overline{x}-x_k)}{\psi_k(\overline{x})-\psi_k(x_k)}\right| \\
&\leq o(\Delta^2)/(-a\Delta) \to 0.
\end{aligned}$$

Logo, $\lim\limits_{\Delta\to 0}\rho(\Delta)=1$, ou seja, existe $\overline{\overline{\Delta}}\in(0,\overline{\Delta}]$ tal que para $\Delta\leq\overline{\overline{\Delta}}$,

$$f(\overline{x}(\Delta))\leq f(x_k)+\alpha[\psi_k(\overline{x}(\Delta))-\psi_k(x_k)]. \tag{7.2.7}$$

Portanto, x_{k+1} está bem definido neste caso.

Suponhamos agora que vale (7.2.3). Então existe $d\in I\!R^n$ tal que $\|d\|=1$ e

$$d^T\nabla^2 f(x_k)d<0. \tag{7.2.8}$$

Como antes, seja $\overline{x}=\overline{x}(\Delta)$ minimizador global de $\psi_k(x)$ sujeito a $\|x-x_k\|\leq\Delta$.

Assim, por (7.2.3), segue que para $\Delta\leq\Delta_1$,

$$\psi_k(\overline{x})\leq\psi_k(x_k+\Delta d)=f(x_k)+\frac{1}{2}\Delta d^T\nabla^2 f(x_k)\Delta d\ .$$

Ou seja,

$$\frac{\psi_k(\overline{x})-\psi_k(x_k)}{\Delta^2}\leq\frac{1}{2}d^T\nabla^2 f(x_k)d\ .$$

Portanto, existe $\overline{\Delta}>0$ tal que para $\Delta\leq\overline{\Delta}$,

$$\frac{\psi_k(\overline{x})-\psi_k(x_k)}{\Delta^2}\leq\frac{1}{4}d^T\nabla^2 f(x_k)d=b<0\ . \tag{7.2.9}$$

Portanto,

$$|\rho(\Delta)-1|=\left|\frac{f(\overline{x})-\psi_k(\overline{x})}{\psi_k(\overline{x})-\psi(x_k)}\right|\leq\frac{o(\|\overline{x}-x_k\|^2)}{\Delta^2}\to 0.$$

Logo, $\lim\limits_{\Delta\to 0}\rho(\Delta)=1$. Assim, para Δ suficientemente pequeno, (7.2.7) se verificará, o que completa a prova. **QED**

A convergência global para pontos que satisfazem as condições necessárias de segunda ordem é provada no seguinte teorema.

Teorema 7.2.3 - Convergência global de segunda ordem.

Seja $\{x_k\}$ uma seqüência infinita gerada pelo Algoritmo 7.2.1. Se x_* é um ponto limite de $\{x_k\}$, então $\nabla f(x_*) = 0$ e $\nabla^2 f(x_*) \geq 0$.

Prova: Seja K_1 um conjunto infinito de índices tal que

$$\lim_{k \in K_1} x_k = x_* .$$

Há duas possibilidades a serem consideradas:

$$\inf_{k \in K_1} \Delta_k = 0 \tag{7.2.10}$$

ou

$$\inf_{k \in K_1} \Delta_k > 0 \; . \tag{7.2.11}$$

Assumindo inicialmente (7.2.10), então existe $K_2 \underset{\infty}{\subset} K_1$ tal que

$$\lim_{k \in K_2} \Delta_k = 0 \; . \tag{7.2.12}$$

Desta forma, existe $k_2 \in \mathcal{N}$ tal que $\Delta_k < \Delta_{\min}$ para todo $k \in K_3$, onde $K_3 \equiv \{k \in K_2 \mid k \geq k_2\}$. Mas, em cada iteração k tentamos inicialmente o raio $\Delta \geq \Delta_{\min}$. Então, para todo $k \in K_3$, existem $\overline{\Delta}_k$ e $\overline{x}(\overline{\Delta}_k)$ tais que $\overline{x}(\overline{\Delta}_k)$ é solução global de:

$$\begin{array}{l} \text{Minimizar } \; \psi_k(x) \\ \|x - x_k\| \leq \overline{\Delta}_k \end{array} \tag{7.2.13}$$

mas

$$f(\overline{x}(\overline{\Delta}_k)) > f(x_k) + \alpha[\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)] \; . \tag{7.2.14}$$

Pela atualização do raio de confiança no Passo 3 do Algoritmo 7.2.1, temos

$$\Delta_k > 0.1 \|\overline{x}(\overline{\Delta}_k) - x_k\| \; . \tag{7.2.15}$$

Logo, por (7.2.12) e (7.2.15) segue que

$$\lim_{k \in K_3} \|\overline{x}(\overline{\Delta}_k) - x_k\| = 0 \; . \tag{7.2.16}$$

Suponhamos que x_* não seja um minimizador local de (7.2.1). Então

$$\nabla f(x_*) = g(x_*) \neq 0 \tag{7.2.17}$$

ou

$$g(x_*) = 0 \quad \text{mas} \quad \nabla^2 f(x_*) \not\geq 0 \; . \tag{7.2.18}$$

Se ocorre (7.2.17), então existe $d \in I\!\!R^n$ tal que $\|d\| = 1$ e

$$g(x_*)^T d < 0 \; . \tag{7.2.19}$$

Então, para $k \in K_3$,

$$\psi_k(\overline{x}(\overline{\Delta}_k)) \leq \psi_k(x_k + \overline{\Delta}_k d) = f(x_k) + \overline{\Delta}_k g(x_k)^T d + \frac{\overline{\Delta}_k^2}{2} d^T \nabla^2 f(x_k) d$$

ou seja,

$$\psi_k(\overline{x}(\overline{\Delta}_k)) - f(x_k) \leq \overline{\Delta}_k g(x_k)^T d + \frac{\overline{\Delta}_k^2}{2} \|\nabla^2 f(x_k)\|.$$

Logo, como $f(x_k) = \psi_k(x_k)$,

$$\frac{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}{\overline{\Delta}_k} \leq g(x_k)^T d + \frac{\|\nabla^2 f(x_k)\|}{2} \overline{\Delta}_k \; .$$

Portanto, existe $k_3 \in \mathcal{N}$ tal que para $k \in K_4 \equiv \{k \in K_3 \mid k \geq k_3\}$,

$$\frac{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}{\overline{\Delta}_k} \leq \frac{g(x_*)^T d}{2} \equiv c_1 < 0 \; . \tag{7.2.20}$$

Definimos

$$\overline{\rho}_k = \frac{f(\overline{x}(\overline{\Delta}_k)) - f(x_k)}{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)} \; . \tag{7.2.21}$$

Então

$$\begin{aligned} |\overline{\rho}_k - 1| &= \left| \frac{f(\overline{x}(\overline{\Delta}_k)) - f(x_k) - [\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)]}{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)} \right| \\ &= \left| \frac{f(\overline{x}(\overline{\Delta}_k)) - \psi_k(\overline{x}(\overline{\Delta}_k))}{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)} \right| = \frac{o(\|\overline{x}(\overline{\Delta}_k) - x_k\|^2)}{-c_1 \overline{\Delta}_k} = o(\overline{\Delta}_k) \; . \end{aligned}$$

Portanto,

$$\lim_{k \in K_4} \overline{\rho}_k = 1$$

o que contradiz o fato de que os raios $\overline{\Delta}_k$ eram rejeitados. Logo $\nabla f(x_*) = 0$.

Vamos agora assumir a validade de (7.2.18). Então existe $d \in I\!R^n$ tal que $\|d\| = 1$ e

$$d^T \nabla^2 f(x_*) d < 0 \ . \tag{7.2.22}$$

Para $k \in K_3$, definimos $d_k = \overline{\Delta}_k d$ se $g(x_k)^T d \leq 0$ e $d_k = -\overline{\Delta}_k d$ se $g(x_k)^T d > 0$.

Então,

$$\psi_k(\overline{x}(\overline{\Delta}_k)) \leq \psi_k(x_k + d_k) \leq f(x_k) + \frac{\overline{\Delta}_k^2}{2} d^T \nabla^2 f(x_k) d,$$

logo,

$$\frac{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}{\overline{\Delta}_k^2} \leq \frac{1}{2} d^T \nabla^2 f(x_k) d \ .$$

Portanto, existe $k_4 \in \mathcal{N}$ tal que para $k \in K_5 \equiv \{k \in K_3 \mid k \geq k_4\}$,

$$\frac{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}{\overline{\Delta}_k^2} \leq \frac{1}{4} d^T \nabla^2 f(x_*) d \equiv c_2 < 0 \ .$$

Assim, usando, de novo, a aproximação de Taylor de segunda ordem, temos:

$$|\overline{\rho}_k - 1| = \left| \frac{f(\overline{x}(\overline{\Delta}_k)) - \psi_k(\overline{x}(\overline{\Delta}_k))}{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi(x_k)} \right| \leq \frac{1}{|c_2|} \frac{o(\|\overline{x}(\overline{\Delta}_k) - x_k\|^2)}{\overline{\Delta}_k^2} \ .$$

Portanto $\lim_{k \in K_5} \overline{\rho}_k = 1$, o que contradiz o fato de $\overline{\Delta}_k$ ser um raio rejeitado. Assim, $\nabla^2 f(x_*) \geq 0$, o que conclui a prova quando vale (7.2.10).

Vamos agora considerar a possibilidade (7.2.11). Como $\lim_{k \in K_1} x_k = x_*$ e $\{f(x_k)\}_{k \in \mathcal{N}}$ é monotonicamente decrescente, temos

$$\lim_{k \in K_1} (f(x_{k+1}) - f(x_k)) = 0 \ . \tag{7.2.23}$$

Mas, pelo Passo 3 do Algoritmo 7.2.1,

$$f(x_{k+1}) \leq f(x_k) + \alpha[\psi_k(x_{k+1}) - \psi_k(x_k)] \ . \tag{7.2.24}$$

Então, por (7.2.23) e (7.2.24), segue que

$$\lim_{k \in K_1} (\psi_k(x_{k+1}) - \psi_k(x_k)) = 0 \ . \tag{7.2.25}$$

Definimos $\underline{\Delta} = \inf_{k \in K_1} \Delta_k > 0$ e chamamos $\widehat{x}$ a uma solução global de

$$\begin{array}{ll}\text{Minimizar} & g(x_*)^T(x - x_*) + \frac{1}{2}(x - x_*)^T \nabla^2 f(x_*)(x - x_*) \\ & \|x - x_*\| \leq \underline{\Delta}/2\ .\end{array} \tag{7.2.26}$$

Seja $k_5 \in \mathcal{N}$ tal que

$$\|x_k - x_*\| \leq \underline{\Delta}/2 \tag{7.2.27}$$

para todo $k \in K_6 \equiv \{k \in K_1 \mid k \geq k_5\}$.

Para $k \in K_6$, por (7.2.26) e (7.2.27), temos

$$\|\widehat{x} - x_k\| \leq \underline{\Delta} \leq \Delta_k, \tag{7.2.28}$$

ou seja, $\widehat{x}$ é factível para o subproblema do Passo 2 do Algoritmo 7.2.1. Então, pelo fato de x_{k+1} ser minimizador global de $\psi_k(x)$ em $\|x - x_k\| \leq \Delta_k$, segue que

$$\psi_k(x_{k+1}) \leq \psi_k(\widehat{x}) = f(x_k) + g(x_k)^T(\widehat{x} - x_k) + \frac{1}{2}(\widehat{x} - x_k)^T \nabla^2 f(x_k)(\widehat{x} - x_k) \tag{7.2.29}$$

ou seja,

$$\psi_k(x_{k+1}) - \psi_k(x_k) \leq g(x_k)^T(\widehat{x} - x_k) + \frac{1}{2}(\widehat{x} - x_k)^T \nabla^2 f(x_k)(\widehat{x} - x_k)\ . \tag{7.2.30}$$

Por (7.2.25), passando (7.2.30) ao limite para $k \in K_6$, obtemos:

$$0 \leq g(x_*)^T(\widehat{x} - x_*) + \frac{1}{2}(\widehat{x} - x_*)^T \nabla^2 f(x_*)(\widehat{x} - x_*),$$

portanto x_* é minimizador de (7.2.26) com a restrição $\|x - x_*\| \leq \underline{\Delta}/2$ inativa. Logo $g(x_*) = 0$ e $\nabla^2 f(x_*) \geq 0$ pelas condições necessárias de otimalidade de segunda ordem para minimização sem restrições. Isso completa a prova. **QED**

Como no caso do método de Newton com "backtracking", fica apenas a questão da compatibilização da estratégia global com o algoritmo local. Ou seja, quando $\nabla^2 f(x_*)$ é definida positiva, gostaríamos que a seqüência gerada pelo Algoritmo 7.2.1 convergisse para x_* e coincidisse com a definida pelo algoritmo local aplicado a $g(x) = 0$. Deixamos essa prova, que segue as mesmas linhas do Teorema 6.2.3, como exercício para o leitor.

7.3 Minimização em caixas

Nesta seção vamos considerar o seguinte problema:

$$\begin{array}{c} \text{Minimizar } f(x) \\ l \leq x \leq u \end{array} \tag{7.3.1}$$

com $f : I\!R^n \to I\!R$, $l_i \in I\!R \cup \{-\infty\}$ e $u_i \in I\!R \cup \{\infty\}$ para todo $i = 1, \ldots, n$. A expressão $[x]_i \leq \infty$ (respectivamente $[x]_i \geq -\infty$) deve ser interpretada como $[x]_i < \infty$ (respectivamente $[x]_i > -\infty$). Portanto, o problema de minimização sem restrições, estudado no capítulo 6 e na Seção 7.2, é um caso particular de (7.3.1). Aqui daremos um sentido preciso à expressão "minimizador aproximado", que usamos na definição do Algoritmo 7.1.1. A idéia é definir um algoritmo facilmente adaptável para problemas de grande porte. Os subproblemas que resolveremos serão minimizações de quadráticas em regiões que, via de regra, serão caixas ou bolas, portanto, poderemos usar diferentes métodos estudados no Capítulo 4, dependendendo do tamanho e estrutura do problema. O algoritmo principal pode ser introduzido com qualquer norma para definir a região de confiança. No entanto, quando a região factível é uma caixa limitada, a norma $\|\cdot\|_\infty$ é a mais adequada, porque a intersecção de $l \leq x \leq u$ com $\|x - x_k\|_\infty \leq \Delta$ é, também, uma caixa. Nesse caso, se usássemos, por exemplo, a norma euclidiana o domínio do subproblema seria uma região bem mais complicada.

Algoritmo 7.3.1 - Minimização em caixas.

Sejam $\Delta_{\min} > 0$, $\alpha \in (0, 1)$, $\|\cdot\|$ uma norma arbitrária e x_0 um ponto inicial factível.

Dado x_k tal que $l \leq x_k \leq u$, obter x_{k+1} da seguinte maneira:

(1) Escolher $\Delta \geq \Delta_{\min}$ e $B_k \in I\!R^{n \times n}$ simétrica tal que $\|B_k\|_2 \leq M_k$.

(2) Encontrar x_k^Q solução global de

$$\begin{array}{c} \text{Minimizar } Q_k(x) \equiv f(x_k) + g(x_k)^T(x - x_k) + \frac{M_k}{2}\|x - x_k\|_2^2 \\ l \leq x \leq u \\ \|x - x_k\| \leq \Delta \end{array} \tag{7.3.2}$$

(3) Encontrar $\overline{x}$ tal que

$$\begin{array}{c} \psi_k(\overline{x}) \leq Q_k(x_k^Q) \\ l \leq \overline{x} \leq u \\ \|\overline{x} - x_k\| \leq \Delta \end{array} \tag{7.3.3}$$

(4) Se $f(\overline{x}) \leq f(x_k) + \alpha[\psi_k(\overline{x}) - \psi_k(x_k)]$,
definir $x_{k+1} = \overline{x}$, $\Delta_k = \Delta$ e terminar a iteração.
Senão, escolher $\Delta_{\text{novo}} \in [0.1\|\overline{x} - x_k\|, 0.9\Delta]$, $\Delta \leftarrow \Delta_{\text{novo}}$ e voltar para (2).

O ponto $\overline{x}$ que é computado no Passo 3 é o que chamamos "solução aproximada" de

$$\begin{array}{c} \text{Minimizar } \psi_k(x) \\ \text{sujeita a } l \leq x \leq u, \;\; \|x - x_k\| \leq \Delta. \end{array} \tag{7.3.4}$$

A condição exigida em (7.3.3) para essa solução aproximada é muito fraca. De fato, é fácil ver que, devido a $\|B_k\|_2 \leq M_k$, temos $\psi_k(x) \leq Q_k(x)$ para todo x, portanto o próprio x_k^Q satisfaz as condições de (7.3.3). Por outro lado, M_k e x_k^Q se calculam muito facilmente. M_k pode ser igual a $\|B_k\|_\infty$, que é o máximo da soma dos módulos das linhas de B_k, e x_k^Q é a projeção de $x_k - g(x_k)/M_k$ na caixa $\{x \in I\!R^n \mid l \leq x \leq u\,,\, \|x - x_k\| \leq \Delta\}$. Ou seja, chamando $y_k = x_k - g(x_k)/M_k$, temos que, se $\|\cdot\| = \|\cdot\|_\infty$,

$$[x_k^Q]_i = \;\max\,\{l_i,\; \min\,\{[y_k]_i, u_i\}\}$$

para todo $i = 1, \ldots, n$.

O Algoritmo 7.3.1 foi introduzido em [82]. Outros procedimentos para minimizar em caixas, baseados em problemas fáceis diferentes, podem ser encontrados em [41], [42], [43] e [44]. Qualquer método para minimizar quadráticas em caixas pode ser usado para resolver (aproximadamente) (7.3.4). Esses algoritmos são, geralmente, iterativos. O aconselhável é usar como ponto inicial x_k^Q, de maneira que a satisfação das condições (7.3.3) ficará automaticamente garantida. No entanto, um critério de parada adicional é necessário para interromper o processo combinando uma aproximação razoável na solução de (7.3.4) com um tempo computacional tolerável. As idéias dos métodos de Newton truncados vêm em nossa ajuda. Como em (4.3.3), definimos $\overline{\nabla\psi}_P$ por

$$[\overline{\nabla\psi}_P(x)]_i = \begin{cases} 0 & \text{se } x_i = l'_i \;\text{ e } [\nabla\psi(x)]_i > 0 \\ 0 & \text{se } x_i = u'_i \text{ e } [\nabla\psi(x)]_i < 0 \\ -[\nabla\psi(x)]_i & \text{nos outros casos,} \end{cases} \tag{7.3.5}$$

onde l'_i e u'_i são os limites da caixa $\{x \in \Omega \mid \|x - x_k\|_\infty \leq \Delta\}$. Então, $\overline{x}$ satisfaz as condições de primeira ordem para minimizador de (7.3.4) se

$$\overline{\nabla\psi}_P(\overline{x}) = 0. \tag{7.3.6}$$

Isto sugere que um critério de parada razoável para o processo iterativo aplicado a (7.3.4) seja:

$$\|\overline{\nabla\psi}_P(\overline{x})\| \leq \eta_k \|\overline{\nabla\psi}_P(x_k)\|, \tag{7.3.7}$$

com $\eta_k \in (0,1)$ (em geral, $\eta_k \equiv 0.1$), o que evoca o critério de Dembo, Eisenstat e Steihaug e, de fato, coincide com esse critério no caso em que os limites l'_i e u'_i são infinitos. Por facilidade de exposição, estamos tratando sempre as quadráticas Q e ψ como funções de x. Na prática, elas são manipuladas como funções de $x - x_k$, através de mudanças de variáveis óbvias.

Finalmente, como (7.3.4) é apenas um subproblema, não se justificam esforços enormes para sua resolução. Isto significa que, se por qualquer motivo, o minimizador de quadráticas tem dificuldades para atingir (7.3.7), sua execução deve ser interrompida, lembrando que, de qualquer maneira, as condições (7.3.3) são suficientes para assegurar a continuidade do algoritmo principal. Assim, é freqüente abortar a minimização da quadrática quando o número de iterações excede um número fixo, digamos, 10, para problemas grandes, ou quando o progresso obtido na última iteração é menor que a décima parte do melhor progresso obtido nas iterações anteriores.

Como no caso das quadráticas, definimos a *direção de Cauchy*:

$$[\overline{g}_p(x)]_i = \begin{cases} 0 & \text{se } x_i = l_i \text{ e } [\nabla f(x)]_i > 0 \\ & \text{ou } x_i = u_i \text{ e } [\nabla f(x)]_i < 0 \\ -[\nabla f(x)]_i & \text{caso contrário.} \end{cases}$$

Pelas condições de otimalidade de primeira ordem, obtemos a seguinte caracterização para minimizadores locais de (7.3.1).

Teorema 7.3.2 - Condições de otimalidade para (7.3.1)

Sejam x_* minimizador local de (7.3.1) e $f \in C^1$ em $\Omega = \{x \in I\!R^n \mid l \leq x \leq u\}$. Então $\overline{g}_p(x_*) = 0$.

Exercício 7.1: Demonstrar o Teorema 7.3.2 usando a teoria do Capítulo 2 e fornecer uma prova independente.

Como fizemos com outros métodos, vamos provar agora que, se um iterando não satisfaz as condições de otimalidade de primeira ordem (neste caso $\overline{g}_p(x) = 0$), o ponto seguinte pode ser calculado em tempo finito, e a função objetivo diminui.

Teorema 7.3.3 - Boa definição.
Se $\overline{g}_p(x_k) \neq 0$ *então* x_{k+1} *está bem definido e* $f(x_{k+1}) < f(x_k)$.

Prova: Como $\overline{g}_p(x_k) \neq 0$, existe $d \in I\!R^n, d \neq 0$ tal que d é factível e de descida. Então, existe $\overline{t} > 0$ tal que

$$l \leq x_k + td \leq u$$

paa todo $t \in [0, \overline{t}]$ e

$$g(x_k)^T d < 0 \ .$$

Assim, para Δ suficientemente pequeno, por (7.3.2) temos:

$$Q_k(x_k^Q) \leq Q_k\left(x_k + \frac{\Delta d}{\|d\|}\right) = f(x_k) + \Delta g(x_k)^T \frac{d}{\|d\|} + \frac{M_k \Delta^2}{2} \ .$$

Então

$$\frac{Q_k(x_k^Q) - Q_k(x_k)}{\Delta} = g(x_k)^T \frac{d}{\|d\|} + \frac{M\Delta}{2}.$$

Mas $\psi_k(x_k) = Q_k(x_k)$ e, escrevendo $\overline{x} = \overline{x}(\Delta)$, temos que $\psi_k(\overline{x}) \leq Q_k(x_k^Q)$, portanto existe $\overline{\Delta} > 0$ tal que

$$\frac{\psi_k(\overline{x}) - \psi_k(x_k)}{\Delta} \leq \frac{g(x_k)^T d}{2\|d\|} \equiv c_1 < 0 \tag{7.3.8}$$

para todo $\Delta \in (0, \overline{\Delta}]$.

Definimos, para $\Delta \in (0, \overline{\Delta}]$,

$$\rho(\Delta) = \frac{f(\overline{x}) - f(x_k)}{\psi_k(\overline{x}) - \psi_k(x_k)} \ . \tag{7.3.9}$$

Então, por (7.3.8), temos

$$\begin{aligned} |\rho(\Delta) - 1| &= \left| \frac{f(\overline{x}) - \psi_k(\overline{x})}{\psi_k(\overline{x}) - \psi_k(x_k)} \right| \\ &\leq \left| \frac{f(\overline{x}) - f(x_k) - g(x_k)^T(\overline{x} - x_k)}{c_1 \Delta} \right| + \left| \frac{(\overline{x} - x_k)^T B_k (\overline{x} - x_k)}{2c_1 \Delta} \right| \\ &\leq \frac{o(\|\overline{x} - x_k\|)}{|c_1|\Delta} + \frac{\|B_k\|_2 \|\overline{x} - x_k\|_2^2}{2|c_1|\Delta} \leq \frac{o(\Delta)}{|c_1|\Delta} + \frac{c_2 M_k \Delta}{2|c_1|} \ , \end{aligned}$$

onde $c_2 > 0$ vem da equivalência das normas em $I\!R^n : \|\cdot\|_2 \leq c_2\|\cdot\|$.

Logo, $\lim\limits_{\Delta\to 0}\rho(\Delta) = 1$ e portanto, após um número finito de reduções no raio de confiança Δ, a condição $f(\overline{x}) \leq f(x_k) + \alpha[\psi_k(\overline{x}) - \psi_k(x_k)]$ é satisfeita e o novo ponto x_{k+1} está bem definido. **QED**

No último teorema deste capítulo, mostramos que todo ponto limite de uma seqüência gerada pelo Algoritmo 7.3.1 é estacionário.

Teorema 7.3.4 - Convergência global.

Seja $\{x_k\}$ uma seqüência infinita gerada pelo Algoritmo 7.3.1. Se $\lim\limits_{k\in K_1} x_k = x_*$, onde K_1 é um subconjunto infinito de índices e M_k é limitado para $k \in K_1$, então $\overline{g}_p(x_*) = 0$.

Prova: Devemos considerar duas possibilidades:

$$\inf_{k\in K_1} \Delta_k = 0 \tag{7.3.10}$$

ou

$$\inf_{k\in K_1} \Delta_k > 0\ . \tag{7.3.11}$$

Vamos assumir inicialmente que vale (7.3.10). Então existe $K_2 \underset{\infty}{\subsetneq} K_1$ tal que

$$\lim_{k\in K_2} \Delta_k = 0. \tag{7.3.12}$$

Logo, existe $k_2 \in K_2$ tal que $\Delta_k < \Delta_{\min}$ para todo $k \in K_3 \equiv \{k \in K_2 \mid k \geq k_2\}$. Mas, a cada iteração k, tentamos inicialmente um raio $\Delta \geq \Delta_{\min}$. Logo, para todo $k \in K_3$, existem $\overline{\Delta}_k, x_k^Q(\overline{\Delta}_k)$ e $\overline{x}(\overline{\Delta}_k)$ tais que $x_k^Q(\overline{\Delta}_k)$ é solução global de

$$\begin{array}{c} \text{Minimizar } \ Q_k(x) \\ l \leq x \leq u \\ \|x - x_k\| \leq \overline{\Delta}_k, \end{array}$$

vale a desigualdade

$$\psi_k(\overline{x}(\overline{\Delta}_k)) \leq Q_k(x_k^Q(\overline{\Delta}_k))$$

mas,

$$f(\overline{x}(\overline{\Delta}_k)) > f(x_k) + \alpha[\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)]\ . \tag{7.3.13}$$

Agora, pela atualização do raio de confiança no Passo 4 do Algoritmo 7.3.1,

$$\Delta_k \geq 0.1\|\overline{x}(\overline{\Delta}_k) - x_k\|\ . \tag{7.3.14}$$

Logo, por (7.3.12) e (7.3.14) segue que

$$\lim_{k \in K_3} \|\overline{x}(\overline{\Delta}_k) - x_k\| = 0. \tag{7.3.15}$$

Suponhamos que $\overline{g}_p(x_*) \neq 0$. Então existe $d \in I\!R^n, d \neq 0$ tal que para todo $\lambda \in [0,1]$,

$$l \leq x_* + \lambda d \leq u \tag{7.3.16}$$

e

$$g(x_*)^T d < 0 \; . \tag{7.3.17}$$

Por (7.3.16), existe $k_3 \in K_3, k_3 \geq k_2$ tal que

$$l \leq x_k + \frac{\lambda}{2} d \leq u \tag{7.3.18}$$

para todo $k \in K_4 \equiv \{k \in K_3 \mid k \geq k_3\}, \lambda \in [0,1]$.

Definimos, para $k \in K_4$,

$$d_k = \frac{\|\overline{x}(\overline{\Delta}_k) - x_k\|}{\|d\|} d \; . \tag{7.3.19}$$

Por (7.3.15) e (7.3.18), existe $k_4 \in K_4$ tal que

$$l \leq x_k + d_k \leq u$$

para todo $k \in K_5 \equiv \{k \in K_4 \mid k \geq k_4\}$.

Claramente, $\|d_k\| = \|\overline{x}(\Delta) - x_k\| \leq \overline{\Delta}_k$. Logo, por (7.3.2), (7.3.3) e (7.3.19),

$$\begin{aligned}
\psi_k(\overline{x}(\overline{\Delta}_k)) \;&\leq\; Q_k(x_k^Q(\overline{\Delta}_k)) \;\leq\; Q_k(x_k + d_k) \\
&=\; f(x_k) \,+\, g(x_k)^T d_k \,+\, \frac{M_k}{2}\|d_k\|_2^2 \\
&=\; f(x_k) \,+\, \frac{\|\overline{x}(\overline{\Delta}_k) - x_k\|}{\|d\|} g(x_k)^T d \,+\, \frac{M_k}{2}\|d_k\|_2^2
\end{aligned}$$

para todo $k \in K_5$.

Então,

$$\frac{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}{\|\overline{x}(\overline{\Delta}_k) - x_k\|} \leq g(x_k)^T \frac{d}{\|d\|} + \frac{M_k c_1^2}{2}\|d_k\| \; ,$$

onde $c_1 > 0$ vem da equivalência das normas em $I\!R^n$.

Portanto, por (7.3.15), (7.3.17), a continuidade de g e a limitação de M_k, existem $c_2 < 0$ e $k_5 \in K_5$ tais que

$$\frac{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}{\|\overline{x}(\overline{\Delta}_k) - x_k\|} \leq c_2 < 0 \tag{7.3.20}$$

para todo $k \in K_6 \equiv \{k \in K_5 \mid k \geq k_5\}$.

Definimos, para $k \in K_6$,

$$\overline{\rho}_k = \frac{f(\overline{x}(\overline{\Delta}_k)) - f(x_k)}{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}.$$

Assim, temos

$$\overline{\rho}_k - 1 = a_k + b_k$$

onde

$$a_k = \frac{f(\overline{x}(\overline{\Delta}_k)) - f(x_k) - g(x_k)^T(\overline{x}(\overline{\Delta}_k) - x_k)}{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}$$

e

$$b_k = \frac{1}{2}\frac{(\overline{x}(\overline{\Delta}_k) - x_k)^T B_k(\overline{x}(\overline{\Delta}_k) - x_k)}{\psi_k(\overline{x}(\overline{\Delta}_k)) - \psi_k(x_k)}.$$

Agora, por (7.3.20) e pela equivalência das normas em $I\!R^n$,

$$|a_k| \leq \frac{o(\|\overline{x}(\overline{\Delta}_k) - x_k\|)}{\|\overline{x}(\overline{\Delta}_k) - x_k\|}$$

e

$$|b_k| \leq \frac{M_k c_1^2 \|\overline{x}(\overline{\Delta}_k) - x_k\|}{2|c_2|}.$$

Portanto, $\lim_{k \in K_6} a_k = 0$ e pela limitação de M_k, $\lim_{k \in K_6} b_k = 0$. Ou seja, $\lim_{k \in K_6} \overline{\rho_k} = 1$, o que contradiz (7.3.13). Dessa forma, (7.3.10) não pode se verificar se $\overline{g}_p(x_*) \neq 0$.

Vamos assumir agora a validade de (7.3.11). Como $\lim_{k \in K_1} x_k = x_*$ e $\{f(x_k)\}_{k \in \mathcal{N}}$ é monotonicamente decrescente, temos

$$\lim_{K \in K_1} (f(x_{k+1}) - f(x_k)) = 0.$$

Mas, por (7.3.2), (7.3.3) e pelo Passo 4 do Algoritmo 7.3.1,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \alpha[\psi_k(x_{k+1}) - \psi_k(x_k)] \\ &\leq f(x_k) + \alpha[Q_k(x_k^Q(\Delta_k)) - Q_k(x_k)]. \end{aligned}$$

Logo,

$$\lim_{k\in K_4} Q_k(x_k^Q(\Delta_k)) = 0. \tag{7.3.21}$$

Definimos $\underline{\Delta} = \inf_{k\in K_1} \Delta_k > 0$. Seja $M > 0$ tal que $M_k \leq M$ para todo $k \in K_1$ e seja $\widehat{x}$ solução global de:

$$\begin{array}{rl} \text{Minimizar} & g(x_*)^T(x - x_*) + \frac{M}{2}\|x - x_*\|_2^2 \\ & l \leq x \leq u \\ & \|x - x_*\| \leq \underline{\Delta}/2 \end{array} \tag{7.3.22}$$

Seja $k_6 \in K_1$ tal que

$$\|x_k - x_*\| \leq \underline{\Delta}/2 \tag{7.3.23}$$

para todo $k \in K_7 \equiv \{k \in K_1 \mid k \geq k_6\}$.

Para $k \in K_7$, por (7.3.22) e (7.3.23),

$$\|\widehat{x} - x_k\| \leq \underline{\Delta} \leq \Delta_k\ . \tag{7.3.24}$$

Além disso, por (7.3.22),

$$l \leq \widehat{x} \leq u\ . \tag{7.3.25}$$

Ou seja, por (7.3.24) e (7.3.25) vemos que $\widehat{x}$ é factível para o problema (7.3.2). Então,

$$Q_k(x_k^Q(\Delta_k)) \leq Q_k(\widehat{x}) \tag{7.3.26}$$

para todo $k \in K_7$.

Agora, pela definição de $\widehat{x}$, por (7.3.26) e (7.3.21),

$$\begin{aligned} g(x_*)^T(\widehat{x} - x_*) + \tfrac{M}{2}\|\widehat{x} - x_*\|_2^2 &= \lim_{k\in K_7} g(x_k)^T(\widehat{x} - x_k) + \frac{M}{2}\|\widehat{x} - x_k\|_2^2 \\ &= \lim_{k\in K_7} Q_k(\widehat{x}) \ \geq\ \lim_{k\in K_7} Q_k(x_k^Q(\Delta_k)) \ =\ 0\ . \end{aligned}$$

Mas o valor da função objetivo de (7.3.22) em x_* também é 0, portanto, x_* também é um minimizador global de (7.3.22). Escrevendo a condição de otimalidade para este problema, chegamos a $\overline{g}_p(x_*) = 0$. **QED**

O Algoritmo 7.3.1 pode servir como modelo para a globalização por regiões de confiança de todos os métodos newtonianos. A naturalidade de sua adaptação à filosofia dos Newton truncados já foi comentada. Quando as matrizes B_k são atualizadas por fórmulas secantes, o algoritmo fornece um esquema para globalizar esse tipo de métodos. Tal adaptação merece alguns comentários:

(a) Nos subproblemas (7.3.3) e (7.3.4), o fato de B_k ser definida positiva não tem maior relevância. Por isso, o procedimento de regiões de confiança é mais adequado que o de buscas lineares para globalizar, por exemplo, o algoritmo baseado na atualização de posto 1, e outros métodos onde as aproximações Hessianas não são necessariamente definidas positivas.

(b) O fato de B_{k+1}^{-1} ser facilmente gerado a partir de B_k^{-1} não pode ser explorado em regiões de confiança como nas buscas lineares. Apenas quando os limites do subproblema (7.3.4) são infinitos ou muito grandes, o fato de se ter B_k^{-1} facilmente disponível é uma boa vantagem, pois permite resolver exatamente o subproblema em um passo só, se a matriz é definida positiva.

(c) Apesar da observação (b), se B_k^{-1} é facilmente calculável, o ponto $x_k - B_k^{-1} g(x_k)$ pode representar um excelente ponto inicial alternativo para o algoritmo quadrático, depois de projetado na região factível de (7.3.4). É inevitável, porém, manter simultaneamente na memória B_k e B_k^{-1}.

A compatibilidade do algoritmo global 7.3.1 com os algoritmos locais subjacentes, nos moldes dos Teoremas 6.2.2 e 6.3.5 fica, mais uma vez, para ser discutida pelos leitores.

Chapter 8

Minimização unidimensional

Alguns problemas de otimização consistem em minimizar funções de uma variável. Para esses problemas, podem-se usar os métodos gerais de minimização sem restrições, minimização em caixas, etc. De fato, um bom exercício para o estudante é verificar como se comportam os algoritmos gerais em funções univariadas. No entanto, a unidimensionalidade é uma estrutura extremamente diferenciada, que justifica o desenvolvimento de algoritmos específicos.

Nas versões antigas de algoritmos de minimização de funções de n variáveis com busca linear, esta busca era interpretada quase sempre como minimização unidimensional. Os métodos modernos usam, geralmente, buscas lineares menos exigentes o que, na maioria dos casos é mais eficiente. No entanto, buscas lineares "duras", semelhantes à minimização unidimensional, são ainda usadas em alguns algoritmos atuais com resultados práticos surpreendentemente bons [50].

Neste capítulo, nosso objetivo é apresentar diferentes técnicas para minimização unidimensional, adequadas às propriedades específicas do problema (existência de derivadas, custo de avaliação da função e suavidade). Veremos que, neste caso, a obtenção de minimizadores globais é menos complicada que no caso multivariado.

8.1 Métodos diretos para redução de incerteza

Uma função f de uma variável x no intervalo $[a, b]$ é *unimodal* se existem $\lambda_1, \lambda_2 \in [a, b]$ tais que

(i) f é estritamente descrescente para $x < \lambda_1$,
(ii) f é estritamente crescente para $x > \lambda_2$,
(iii) f é constante para $x \in [\lambda_1, \lambda_2]$.

É fácil ver que os minimizadores locais de uma função unimodal em $[a, b]$ coincidem com os minimizadores globais. Ou seja, este conceito desfruta da mesma propriedade de otimalidade global que a convexidade, com hipóteses menos exigentes sobre a função.

Os métodos diretos para redução de intervalos de incerteza se aplicam bem a funções unimodais. Nada exigem em relação a continuidade ou existência de derivadas. A idéia básica desses métodos é, uma vez conhecido um intervalo $[a, b]$ em que a função f é unimodal, reduzir este intervalo até a precisão desejada. São aplicáveis a problemas com funções cuja avaliação é simples, pois geram um número de iterações (pouco complexas) maior que o produzido pelos métodos polinomiais.

Dada a função $f : I\!\!R \to I\!\!R$, unimodal em $[a, b]$, o algoritmo conceitual a seguir obtém um intervalo reduzido contendo o minimizador de f em $[a, b]$.

Algoritmo 8.1.1 - Redução de incerteza.
Dados $\varepsilon > 0$ e o intervalo $[a, b]$,
definir $k = 0, a_0 = a, b_0 = b$.

(1) Dados a_k e b_k, escolher c_k e d_k tais que

$$a_k < c_k < d_k < b_k \ .$$

(2) Calcular $f(c_k)$ e $f(d_k)$.

(3) Se $f(c_k) < f(d_k)$, fazer $a_{k+1} = a_k, \ b_{k+1} = d_k$
senão $a_{k+1} = c_k, \ b_{k+1} = d_k$.

(4) Se $b_{k+1} - a_{k+1} < \varepsilon$, parar
senão $k = k + 1$ e voltar para (1).

À primeira vista, seriam necessárias duas avaliações da função a cada redução do intervalo. Para que isso não ocorra, podemos escolher c_k e d_k de tal forma que o ponto que permanece no interior do intervalo reduzido seja um dos escolhidos para a próxima avaliação. Apresentaremos duas estratégias para se efetuar estas escolhas: a *busca de Fibonacci* e o *método da seção áurea.*

Para a *busca de Fibonacci* precisamos fixar a priori o número n de avaliações da função a ser feito ou, equivalentemente, a redução desejável no intervalo.

Os números intermediários são então determinados baseados nos números de Fibonacci, definidos de modo recursivo como se segue:

$$F_0 = F_1 = 1; \quad F_k = F_{k-2} + F_{k-1}, \quad k = 2, 3, \ldots \tag{8.1.1}$$

Desta forma, uma vez definido o número n de avaliações, a escolha dos valores c_k e d_k no passo (1) do Algoritmo 8.1.1 é feita da seguinte maneira:

$$\begin{aligned} c_k &= b_k - \tfrac{F_{n-k-1}}{F_{n-k}}(b_k - a_k) \\ d_k &= a_k + \tfrac{F_{n-k-1}}{F_{n-k}}(b_k - a_k) \, . \end{aligned} \tag{8.1.2}$$

Exercício 8.1: Verificar que, com o procedimento (8.1.2), c_{k+1} coincide com d_k e d_{k+1} coincide com c_k.

O *método da seção áurea* é obtido da seqüência de Fibonacci fazendo-se o número n tender para infinito. Assim, no limite, a equação de diferenças de Fibonacci (8.1.1) passa a fornecer a divisão do intervalo $[a, b]$ na razão áurea $\lambda = (\sqrt{5} - 1)/2 \approx 0.618$, que é exatamente a solução do problema da divisão áurea ou do retângulo áureo, proposto pelos gregos por volta de 500 a.C. Na antiguidade, um certo caráter místico foi atribuído a este valor, o que justifica o qualificativo "áureo". Em arquitetura, esta razão, considerada esteticamente agradável, se preserva desde o Parthenon até projetos de Le Corbusier. No método da seção áurea, a escolha dos valores c_k e d_k é feita como se segue:

$$\begin{aligned} c_k &= b_k - \lambda(b_k - a_k) \\ d_k &= a_k + \lambda(b_k - a_k) \, . \end{aligned} \tag{8.1.3}$$

Exercício 8.2: Verificar que c_{k+1} coincide com d_k e d_{k+1} coincide com c_k no procedimento (8.1.3).

É possível provar que, fixado o número de avaliações que será realizado, Fibonacci é o método ótimo para redução de incerteza, pois obtém a máxima redução para o caso mais desfavorável (ver, por exemplo, [197]). No entanto, no método da seção áurea não é necessário fixar-se previamente o número de avaliações de função, o que elimina um pré-requisito pouco natural, do ponto de vista do cálculo numérico, do método de Fibonacci. Na prática de otimização, critérios de parada baseados no valor da função objetivo são mais confortáveis, e esses critérios podem ser implementados sem problemas no método áureo.

Uma outra estratégia para redução de incerteza, bastante simples e intuitiva, é o *método da bisseção.* Este método é usado quando a função $f : [a, b] \to I\!R$ é diferenciável, unimodal e tem derivada com avaliação computacionalmente viável.

Algoritmo 8.1.2 - Método da Bisseção.
Dado ε (tolerância para redução do intervalo $[a, b]$),

(1) $a_0 = a, \ b_0 = b.$

(2) Dados $a_i, \ b_i$, calcular $c_i = \frac{1}{2}(a_i + b_i)$.

(3) Calcular $f(c_i)$.
Se $f'(c_i) = 0$, parar.
Se $f'(c_i) < 0, \ a_{i+1} = c_i, \ b_{i+1} = b_i$,
senão $a_{i+1} = a_i, \ b_{i+1} = c_i$.

(4) Se $b_{i+1} - a_{i+1} < \varepsilon$, parar,
senão $i = i + 1$ e voltar para (2).

Exercício 8.3: Provar que todas as funções convexas são unimodais.

Exercício 8.4: Obter uma função cúbica real que seja unimodal mas não convexa para $0 \leq x \leq 1$.

8.2 Aproximações polinomiais

Muitas vezes podemos assegurar um "bom comportamento" da função a ser minimizada, ainda que apenas nas vizinhanças do minimizador. Desta maneira, temos garantia de uma boa aderência entre a função e uma aproximação por polinômios. A idéia dos métodos que utilizam aproximações polinomiais é, a partir de $k + 1$ informações sobre a função (valores da função, das derivadas, etc), determinar um polinômio de ordem k, estimando-se o minimizador da função a partir do minimizador do polinômio. Em geral, trabalha-se iterativamente e a estratégia de redução de incerteza utilizada nos métodos diretos também é empregada como salvaguarda. As aproximações polinomiais geram um número de iterações inferior ao dos métodos diretos, sendo porém de maior complexidade. No que se segue, vamos apresentar quatro maneiras de efetuar aproximações polinomiais: *o método de Newton*, *o método secante*, *o método DSC-Powell* e *o método da aproximação cúbica.*

O *método de Newton* consiste em aproximar f em torno do ponto x_k pela parábola construída com as informações $f(x_k), f'(x_k)$ e $f''(x_k)$, ou seja,

$$f(x) \approx q(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{f''(x_k)}{2}(x - x_k)^2 . \tag{8.2.1}$$

Para se empregar o método de Newton é preciso que a função seja duas vezes diferenciável. Trata-se de um esquema iterativo localmente convergente, portanto o ponto inicial x_0 deve estar suficientemente próximo da solução x_* para a convergência ser garantida.
Se $f''(x_k) > 0$, a parábola $q(x)$ é estritamente convexa e x_{k+1} será um minimizador global de $q(x)$ se, e somente se,

$$q'(x_{k+1}) = f'(x_k) + f''(x_k)(x_{k+1} - x_k) = 0.$$

Desta forma, o novo ponto x_{k+1} é dado por:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}. \tag{8.2.2}$$

Observamos que (8.2.2) não depende de $f(x_k)$. Na verdade, este método é equivalente ao método da tangente para resolver a equação $f'(x) = 0$. Por isso, quando $f''(x_k) < 0$, o algoritmo pode convergir para um maximizador.

No *método secante* também aproxima-se f em torno de x_k por uma parábola, agora construída a partir de $f(x_k)$, $f'(x_k)$ e $f'(x_{k-1})$. Neste caso, o novo ponto do esquema iterativo é dado por:

$$x_{k+1} = x_k - \frac{f'(x_k)(x_k - x_{k-1})}{f'(x_k) - f'(x_{k-1})}. \tag{8.2.3}$$

Comparando (8.2.2) com (8.2.3), vemos que a informação de segunda ordem do método de Newton é calculada em (8.2.3) usando-se diferenças finitas. Assim, para funções cuja avaliação é trabalhosa, o esquema iterativo (8.2.3) torna-se mais eficiente. Analogamente ao método de Newton, o método secante terá convergência assegurada quando o ponto inicial estiver suficientemente próximo da solução x_*, e pode convergir para um maximizador em vez de um minimizador se não se usam salvaguardas adequadas.

O *método DSC-Powell* é uma combinação, sugerida por Box, Davies e Swann [21], de um algoritmo de Davies, Swann e Campey (DSC) com um algoritmo de Powell.

Em ambos ajusta-se f por uma quadrática conhecidos os valores da função f em três pontos.
Inicialmente o algoritmo cerca a solução x_*, fazendo então uma interpolação quadrática com pontos igualmente espaçados. Esta etapa corresponde ao método DSC. As iterações seguintes, devidas ao método de Powell, consistem em prosseguir interpolando quadraticamente, mas com pontos desigualmente espaçados.

Algoritmo 8.2.1 - DSC-Powell.
Dados o ponto inicial x_0, o tamanho do passo Δx e a precisão ε;

(1) Avaliar $f(x_0)$ e $f(x_0 + \Delta x)$
Se $f(x_0 + \Delta x) > f(x_0),\ \Delta x \leftarrow -\Delta x$.

(2) $x_{k+1} = x_k + \Delta x$.

(3) Calcular $f(x_{k+1})$.

(4) Se $f(x_{k+1}) \leq f(x_k),\ \Delta x = 2\Delta x,\ k \leftarrow k+1$, voltar para (2)
senão $x_m = x_{k+1},\ x_{m-1} = x_k,\ x_{m-2} = x_{k-1},\ \Delta x \leftarrow \frac{\Delta x}{2}$
e repetir (2) e (3) pela última vez, determinando $x_{m+1} = x_{k+2}$.

(5) Dentre os quatro pontos igualmente espaçados $\{x_{m+1},\ x_m,\ x_{m-1},\ x_{m-2}\}$, descartar o mais distante do ponto com menor valor da função.
Renomear os valores restantes por $x_a,\ x_b,\ x_c$,
onde x_b é o ponto central, $x_a = x_b - \Delta x$ e $x_c = x_b + \Delta x$.

(6) Fazer uma interpolação quadrática para estimar x_*:

$$\widehat{x}_* = x_b + \frac{\Delta x(f(x_a) - f(x_c))}{2(f(x_a) - 2f(x_b) + f(x_c))}.$$

(7) Repetir:
redefinir $\{x_a, x_b, x_c\}$ como $\{x_a, \widehat{x}_*, x_b\}$ ou $\{x_b, \widehat{x}_*, x_c\}$,
calcular $f(x_b)$ e estimar x_* por uma interpolação quadrática para pontos desigualmente espaçados:

$$\widehat{x}_* = -\frac{1}{2}\frac{(x_b^2 - x_c^2)f(x_a) + (x_c^2 - x_a^2)f(x_b) + (x_a^2 - x_b^2)f(x_c)}{(x_b - x_c)f(x_a) + (x_c - x_a)f(x_b) + (x_a - x_b)f(x_c)},$$

até que $|x_c - \widehat{x}_*| < \varepsilon$.

Mostra-se que a seqüência gerada pelo Algoritmo 8.2.1 converge para o minimizador quando a função f é convexa. Para mais detalhes sobre o método DSC-Powell, ver Himmelblau [120].

Na *aproximação cúbica* são necessárias quatro informações para construir um polinômio de grau três para aproximar a função f. A escolha mais clássica envolve o conhecimento de $f(x_k)$, $f'(x_k)$, $f(x_{k-1})$ e $f'(x_{k-1})$ e resulta no seguinte minimizador para a cúbica (Luenberger (1984), p.206):

$$x_{k+1} = x_k - \frac{(x_k - x_{k-1})[f(x_k) + \sigma_2 - \sigma_1]}{f(x_k) - f'(x_{k-1}) + 2\sigma_2}, \tag{8.2.4}$$

$$\begin{array}{ll} \text{onde} & \sigma_1 = f'(x_{k-1}) + f'(x_k) - 3\dfrac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \\ \text{e} & \sigma_2 = \sqrt{\sigma_1^2 - f'(x_{k-1})f'(x_k)}\ . \end{array}$$

Se a função é unimodal no intervalo $[a, b]$, $f'(a) < 0$ e $f'(b) > 0$, a aproximação cúbica pode ser combinada com técnicas de redução de incerteza para obter um algoritmo globalmente convergente.

Esse tipo de combinação é computacionalmente necessária em qualquer algoritmo baseado em aproximações polinomiais. De fato, com salvaguardas adequadas, é possível garantir uma efetiva redução do intervalo de incerteza, evitando-se passos muito pequenos quando se está longe da solução. Assim, a interpolação polinomial pode se combinar com o método da bisseção, quando as derivadas são disponíveis, ou com o método da seção áurea, quando se conhecem apenas os valores da função.

Exercício 8.5: Mostrar que no método secante a convergência local é superlinear, mostrando que existe $a > 0$ tal que

$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|^r} \le a\ , \quad r = \frac{1+\sqrt{5}}{2} \approx 1.618\ .$$

Exercício 8.6: Escrever um algoritmo de interpolação cúbica com salvaguardas que garantam uma redução efetiva do intervalo de incerteza em cada iteração.

8.3 Técnicas de minimização global

Quase sempre, o objetivo do otimizador diante de um determinado problema, é obter um minimizador *global.* No entanto, a maioria dos algoritmos práticos e eficientes não possuem convergência garantida para esse tipo de "verdadeiros" minimizadores. Na maioria dos casos, é possível provar convergência, em algum sentido, para pontos estacionários que, muito provavelmente, são minimizadores locais. Freqüentemente, pelas próprias características do problema, os pontos estacionários assim encontrados são minimizadores globais, o que possibilita a solução efetiva de muitos problemas práticos de otimização.

No entanto, existem problemas com infinidade de minimizadores locais, cuja resolução por algoritmos como os mencionados acima é extremamente difícil. Isso motiva o desenvolvimento de *métodos globais*, isto é, algoritmos com convergência garantida para um minimizador global do problema. Infelizmente, os métodos globais assim desenvolvidos perdem muito de sua eficácia quando aplicados a problemas de grande porte. Freqüentemente, o tempo e a memória requeridos por uma iteração são proibitivos até para computadores avançados.

A situação é diferente quando o número de variáveis é pequeno, especialmente, quando a função é de uma variável só, como as que estudamos neste capítulo. Assim, é possível que técnicas globais unidimensionais, combinadas com técnicas "locais" baseadas em buscas lineares ou até regiões de confiança consigam aumentar muito a potencialidade global destas últimas. Neste capítulo, vamos destacar as técnicas de minimização global utilizando *envelopes convexos* e *análise intervalar* [146], [145], [114].

A obtenção de um minimizador global de $f : [a, b] \to I\!R$ através de *envelopes convexos* baseia-se na partição do intervalo $[a, b]$ e, conseqüentemente, do problema original, em subproblemas. A seguir, utilizando-se uma subestimativa convexa para a função objetivo no subintervalo, determina-se facilmente um limitante inferior para o minimizador do subproblema através do minimizador do envelope convexo. Acrescentando-se uma estratégia para eliminar subintervalos, com base nos valores "mínimos" encontrados para a função, mostra-se que o ponto correspondente ao menor dos limitantes inferiores determinados para a função converge para a solução global do problema original.

Com relação à determinação dos envelopes convexos, o fundamental é encontrar os pontos em que a representação da subestimativa convexa muda de forma. Quando a função tem trechos convexos, muitas vezes o envelope convexo coincide com a função original num subintervalo. Pode ainda ser uma reta unindo um ponto ao trecho adjacente, convertendo-se novamente na função num trecho seguinte, e assim por diante. A determinação de quantas representações diferentes são necessárias depende tanto dos tamanhos dos subintervalos quanto do comportamento da própria função. Para se conhecer os pontos exatos em que o envelope convexo muda de representação (de uma reta para a curva da função ou vice-versa), basta fazer um ajuste entre as declividades da curva e da reta. Em outras palavras, se a é o ponto inferior do intervalo, queremos encontrar $x \in [a, b]$ tal que $\frac{f(x)-f(a)}{x-a} = f'(x)$, que é equivalente a

$$f(x) - f(a) - (x - a)f'(x) = 0. \qquad (8.3.1)$$

Dentre as diversas estratégias para se resolver (8.3.1), o método de Newton implementado com salvaguardas geralmente funciona bem e tem o seguinte esquema iterativo:

$$x_{k+1} = x_k + \left(\frac{f(x_k) - f(a)}{x_k - a} - f'(x_k)\right) \left[f''(x_k)\right]^{-1}. \qquad (8.3.2)$$

A idéia básica da *análise intervalar* aplicada à minimização global é o refinamento dos intervalos contendo o valor extremo, descartando-se as regiões em que o minimizador global não pode estar. Assim, na determinação do minimizador global de $f : [a, b] \to I\!R$, suponhamos que $[a, b]$ foi subdividido em $[a, c]$ e $[c, b]$. Suponhamos também que conhecemos $[u, v]$ contendo a imagem do intervalo $[c, b]$ pela f, isto é $f([c, b]) \subset [u, v]$ e conhecemos $[w, z]$ contendo $f(x_1)$, com $x_1 \in [a, c]$. Se $z < u$, então todo o intervalo $[c, b]$ pode ser descartado, já que não existe $x \in [c, b]$ tal que o valor $f(x)$ seja menor que $f(x_1) \leq z$. Assim, o minimizador de f em $[a, b]$ está em $[a, c]$ e não em $[c, b]$. Portanto, com este tipo de teste pode-se excluir regiões que seguramente não contém o minimizador global procurado.

Exercício 8.7: Aplicar as técnicas de envelopes convexos e análise intervalar para obter o minimizador global de
(a) $f(x) = e^{-x} + \operatorname{sen}(\pi x) + x^2\,,\ x \in [-1, 2]$.
(b) $f(x) = -x(1 + x)\cos(x)\,,\ x \in [-2, 2]$.

Chapter 9

Restrições lineares

Vamos considerar o problema de otimização em que a região factível é um politopo em $I\!R^n$, ou seja, um conjunto definido por equações e inequações lineares. A minimização em caixas é um caso particular desse problema. No capítulo 7, aplicamos o algoritmo geral de regiões de confiança ao caso $l \leq x \leq u$, dando um sentido (o do "subproblema fácil") à minimização aproximada do modelo quadrático. Aqui, em princípio, podemos proceder da mesma maneira, com a dificuldade de que o problema fácil não é tão fácil como no caso das caixas. Com efeito, quando o conjunto factível é um politopo, o ponto x_k^Q do Algoritmo 7.3.1 é a projeção de $x_k - g(x_k)/M_k$ na intersecção desse conjunto com a caixa de confiança. Embora haja razões para supor que essa projeção não é difícil de se calcular, certamente é bem mais complicada que quando a região é uma caixa $n-$dimensional. Também, neste caso, é mais conflitante a decisão sobre o algoritmo a ser usado para determinar o ponto-tentativa $\bar{x}$. Portanto, embora as questões teóricas relativas à aplicação de regiões de confiança a minimização com restrições lineares estejam essencialmente resolvidas em [142], não existem ainda implementações práticas amplamente reconhecidas. Ver, também [89] e [40].

Os métodos mais tradicionais para otimização em politopos estão baseados na *estratégia de restrições ativas*. A idéia é similar à usada no capítulo 4 para minimizar quadráticas em caixas. A região é dividida em faces, de maneira que, dentro de cada uma delas, o problema é, essencialmente, irrestrito. Uma face pode ser abandonada apenas quando o trabalho sobre ela se revela improdutivo. Ver [75], [92], [94], [99], [154], [155], [172], [173], [174] e o artigo pioneiro de Rosen [178].

Os problemas de programação linear e programação quadrática são casos particulares do tratado neste capítulo. No primeiro, a função objetivo é linear ($f(x) = c^T x$) e, no segundo, é uma quadrática. O método mais usado para programação linear é o Simplex [46] que é, de fato, um algoritmo de restrições ativas. O programa MINOS para minimização com restrições ([154], [155]) é, quando aplicado a problemas lineares, uma das implementações mais eficientes do método Simplex para grande porte. O conteúdo deste capítulo se aplica, em conseqüência a programação linear e quadrática, mas a estrutura especial destes problemas, e o tratamento da "degeneração primal" justifica o desenvolvimento de textos específicos. Ver [46], [9], etc.

A programação linear e outras áreas da otimização foram sacudidas, a partir de 1984, com o desenvolvimento dos "métodos de pontos interiores". Ver [103]. Algumas indicações sobre a aplicação desses métodos à minimização de funções gerais com restrições lineares serão dadas neste capítulo.

9.1 Igualdades

O problema geral deste capítulo é:

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & x \in \Omega \end{array} \tag{9.1.1}$$

onde $f \in C^1(\Omega)$ e $\Omega = \{x \in I\!\!R^n \mid A_1 x = b_1 \, , \; A_2 x \geq b_2\}$, com $A_1 \in I\!\!R^{m_1 \times n}$ e $A_2 \in I\!\!R^{m_2 \times n}$. O conjunto Ω definido pelas restrições lineares de igualdade e desigualdade é denominado *politopo*.

Um politopo geral Ω sempre pode ser levado à forma $\{x \in I\!\!R^n \mid Ax = b \, , \; x \geq 0\}$ ou à forma $\{x \in I\!\!R^n \mid Ax = b \, , \; l \leq x \leq u\}$, mediante a introdução de "variáveis de folga". Alguns algoritmos trabalham exclusivamente com essa formulação, chamada "padrão".

Exercício 9.1: Converter $\Omega = \{x \in I\!\!R^n \mid A_1 x = b_1 \, , \; A_2 x \geq b_2\}$ para o formato $\{y \in I\!\!R^N \mid Ay = b \, , \; y \geq 0\}$.

Na definição de Ω, estamos incluindo as possibilidades $m_1 = 0$ e $m_2 = 0$. Se ambas dimensões são nulas, o problema é irrestrito. Se apenas $m_2 = 0$ temos o problema de minimização com restrições de igualdade:

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & Ax = b \, . \end{array} \tag{9.1.2}$$

Suponhamos que a região factível de (9.1.2) é não vazia e seja $\overline{x} \in I\!R^n$ tal que $A\overline{x} = b$. Então, todos os pontos da que verificam $Ax = b$ satisfazem $x = \overline{x} + Zz$, onde $Z \in I\!R^{n\times(n-m_p)}$ é uma matriz cujas colunas formam uma base para o núcleo da matriz A e m_p é o posto de A. Assim, (9.1.2) pode ser reescrito como um problema irrestrito num espaço de dimensão menor:

$$\begin{array}{c}\text{Minimizar } \varphi(z) \equiv f(\overline{x} + Zz) \\ z \in I\!R^{n-m_p}\ .\end{array} \tag{9.1.3}$$

Exercício 9.2: Mostrar que

$$\nabla \varphi(z) = Z^T \nabla f(\overline{x} + Zz)$$

e

$$\nabla^2 \varphi(z) = Z^T \nabla^2 f(\overline{x} + Zz) Z.$$

O vetor $\nabla\varphi$ é denominado *gradiente reduzido* e a matriz $\nabla^2\varphi$, *Hessiana reduzida.*

Uma vez encontrado $\overline{x}$ tal que $A\overline{x} = b$ e Z tal que $\mathcal{R}(Z) = \mathcal{N}(A)$, a resolução de (9.1.2) pode ser tentada usando um método direcional (Newton, quase-Newton) ou um método de regiões de confiança para minimização sem restrições. Ver [74].

Para a viabilidade de métodos baseados em (9.1.3) para problemas de grande porte é fundamental que a matriz Z seja esparsa. Ainda mais, se a intenção é implementar o método de Newton, também é necessário que $Z^T\nabla^2 f(x)Z$ o seja. Se Z é grande e densa, (9.1.3) não pode ser utilizado. Nesse caso, observamos que, se B é uma matriz definida positiva (melhor, esparsa e talvez diagonal), a solução de

$$\text{Minimizar } \frac{1}{2} d^T B d + g(\overline{x})^T d \text{ sujeita a } Ad = 0 \tag{9.1.4}$$

corresponde a uma solução (d, π) do sistema linear

$$Bd + g(\overline{x}) + A^T \pi = 0, \quad Ad = 0. \tag{9.1.5}$$

Portanto, a direção d computada por (9.1.5) é uma direção de descida para f, pertencente ao núcleo de A. Se $B = \mu I$, $d = d(\mu)$ se aproxima de uma direção de máxima descida no núcleo, quando μ tende a infinito. Agora, (9.1.5) pode ser resolvido usando apenas a esparsidade de A ou, talvez, um método iterativo linear. Idéias análogas às invocadas no capítulo 6 podem ser adaptadas para provar que um algoritmo baseado em direções d_k calculadas por (9.1.5), com "backtracking", é globalmente convergente a um ponto estacionário de (9.1.2). Uma vantagem adicional de usar iterativamente (9.1.5) é que os sucessivos π_k são estimativas dos multiplicadores de Lagrange na solução. A importância desse fato emergirá no tratamento de restrições de desigualdade.

9.2 Estratégia de restrições ativas

Para facilitar a exposição, consideraremos o problema geral de minimização em politopos apenas na forma

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & Ax \geq b \, , \end{array} \tag{9.2.1}$$

onde $A \in I\!R^{m \times n}$, $A^T = (a_1 \ldots a_m)$, $a_i \in I\!R^n$, $i = 1, \ldots, m$. A transposição das idéias desta seção para o formato geral (9.1.1) é rotineira, e será deixada como exercício para o leitor. Como antes, escrevemos $\Omega = \{x \in I\!R^n \mid Ax \geq b\}$. As definições a seguir são paralelas às dadas quando introduzimos algoritmos para minimizar quadráticas em caixas.

Definição 9.2.1

Dado $I \subset \{1, 2, \ldots, m\}$, chamamos de *face relativa ao conjunto I* ao conjunto

$$F_I = \{x \in \Omega \mid a_i^T x = b_i \text{ se } i \in I \text{ e } a_i^T x > b_i \text{ se } i \notin I\} \, .$$

Como sempre, chamamos $\overline{F}_I$ ao fecho de F_I.

As restrições que são satisfeitas por x na igualdade, isto é, tais que $a_i^T x = b_i$, $i \in I$, são chamadas *ativas em x*. As outras são denominadas *inativas* .

Exercício 9.3: Provar que

(a) $\Omega = \bigcup_{I \in \mathcal{P}} F_I$, onde $\mathcal{P}$ é o conjunto das partes de $\{1, 2, \ldots, m\}$.

(b) Se $I_1 \neq I_2$, $F_{I_1} \cap F_{I_2} = \emptyset$.

Vamos definir agora um algoritmo conceitual que implementa a estratégia de restrições ativas. Nesse algoritmo, trabalhamos com "super-iterações", que permitem passar diretamente de um ponto qualquer a um minimizador global irrestrito. Naturalmente, a existência dessas super-iterações na prática está restrita a problemas simples, como os lineares ou quadráticos. Chamamos $\mathcal{S}$ ao conjunto de minimizadores globais de (9.2.1) e partimos de um ponto inicial arbitrário e factível.

Algoritmo 9.2.2 - Estratégia de restrições ativas.

Dado $x_k \in \Omega$, $x_k \in F_I$, $x_k \notin \mathcal{S}$,
se x_k é minimizador de f em F_I,
então
(1) $x_{k+1} \notin \overline{F}_I$ e $f(x_{k+1}) < f(x_k)$.
Senão
(2) $x_{k+1} \in F_I$ e x_{k+1} é minimizador de f em F_I, ou
(3) $x_{k+1} \in [\overline{F}_I - F_I]$ (a fronteira de F_I) e $f(x_{k+1}) < f(x_k)$, ou
(4) f é ilimitada inferiormente em F_I e o algoritmo pára.

O leitor familiarizado com o Simplex poderá reconhecer que esse método está no escopo do Algoritmo 9.2.2. As faces visitadas nesse caso são vértices, formadas por um único ponto. Portanto x_k sempre é "minimizador de f em F_I", o fecho de F_I é a própria F_I e o ponto seguinte é um ponto diferente onde a função objetivo diminui. Para interpretar corretamente o caso em que o Simplex detecta que o problema é ilimitado, a partir do vértice x_k, pensemos na introdução de uma "iteração" fictícia x_{k+1} factível e situada na semi-reta ao longo da qual f tende a $-\infty$. Essa "última" iteração está numa "aresta" F_I na qual a função é ilimitada inferiormente. A situação, portanto, corresponde ao Passo 4 do Algoritmo 9.2.2.

No seguinte teorema, provamos que a estratégia de restrições ativas é sempre bem sucedida. A dificuldade estará, em conseqüência, em sua implementação.

Teorema 9.2.3

Em um número finito de iterações, o método das restrições ativas encontra a solução de (9.2.1) ou detecta que o problema não tem solução.

Prova: Suponhamos que o Passo 4 do Algoritmo 9.2.2 não acontece em nenhuma iteração da seqüência $\{x_k\}$. Quando uma face F_I é abandonada no Passo 1, então, como x_k é minimizador global para $x \in F_I$ e $f(x_j)$ é monótona decrescente, temos que $x_j \notin F_I$ para todo $j > k$. Como o número de faces é finito, a partir de certo k_0 o Passo 1 não é mais executado. Pela finitude do número de restrições, o Passo 3 também pode ser executado apenas um número finito de vezes se $k \geq k_0$. Portanto, a partir de certo $k_1 \geq k_0$, apenas o Passo 2 é possível. Isso implica que x_{k_1+1} é minimizador global na sua face. Como o Passo 1 não é mais possível, resulta que x_{k_1+1} deve ser minimizador global do problema. **QED**

Apesar do Algoritmo 9.2.2 ter convergência finita, o Passo 2 é, quase sempre, impossível de ser executado em um número finito de etapas. Assim, uma iteração do Algoritmo 9.2.2 é, na verdade, uma super-iteração, pois pode embutir um procedimento infinito.

Suponhamos que $x_k \in F_I$ não é minimizador global de f em F_I. Para obter x_{k+1} pelo Passo 2 ou pelo Passo 3, definimos $\mathcal{V}(F_I) = \{x \in I\!R^n \mid a_i^T x = b_i, i \in I\}$ e consideramos o problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & x \in \mathcal{V}(F_I) \end{array}$$

ou, equivalentemente,

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & a_i^T x = b_i \, , \; i \in I \, . \end{array} \tag{9.2.2}$$

Este problema é do tipo (9.1.2). Para "resolvê-lo" aplicamos um método iterativo, começando com $x_k^0 = x_k$, e gerando uma seqüência $x_k^1, x_k^2, \ldots$ de maneira que, antes de parar, $x_k^j \in \mathcal{V}(F_I)$ e $f(x_k^{j+1}) < f(x_k^j)$ para todo j . Suponhamos que, antes da parada, aparece j tal que $x_k^{j+1} \notin \Omega$. Neste caso, chamamos $d_k^j = x_k^{j+1} - x_k^j$ e t_j o máximo $t > 0$ tal que $[x_k^j, x_k^j + t d_k^j] \subset \Omega$. Uma suposição sobre o processo para (9.2.2) que garante que o Passo 3 do Algoritmo 9.2.2 pode ser completado é que

$$f(x_k^j + t_j d_k^j) < f(x_k^j).$$

Nessa situação, chamamos $x_{k+1} = x_k^j + t_j d_k^j$. O método iterativo aplicado a (9.2.2) será interrompido, no melhor caso, quando x_k^j seja minimizador global de f em F_I, mas é difícil que consigamos essa propriedade em tempo finito. (Uma exceção é quando f é uma quadrática estritamente convexa.) Portanto, o Algoritmo 9.2.2 não poderá ser rodado em estado puro, e a condição "se x_k é minimizador de f em F_I" deverá ser substituída, na prática, por "se x_k é minimizador aproximado de f em F_I". A decisão sobre o que se considera "minimizador aproximado" define diferentes métodos implementáveis de restrições ativas.

9.3 Saindo da face

Nesta seção, descrevemos uma das possíveis maneiras de viabilizar o Passo 2 do Algoritmo de restrições ativas. Mais precisamente, vamos supor que x_k^j é uma das iterações do algoritmo interno usado dentro de F_I, que devemos decidir se x_k^j já é minimizador aproximado nessa face, e, em caso afirmativo, que precisamos mostrar como conseguir $x_{k+1} \notin \overline{F}_i$ e $f(x_{k+1}) < f(x_k^j)$. Para simplificar a notação, escreveremos x_k em vez de x_k^j.

Vamos supor, a princípio, que os gradientes das restrições que definem a face F_I são linearmente independentes. Sem perda de generalidade, suponhamos que $I = \{1, \ldots, \nu\}$, $\bar{A}^T = (a_1, \ldots, a_\nu)$. Portanto, $\bar{A}$ tem posto ν e admite uma submatriz $B \in I\!R^{\nu \times \nu}$ não singular. Por simplicidade, vamos supor que $\bar{A} = (\, B \; N \,)$. Consideramos a mudança de variáveis

$$\begin{array}{rcl} y_1 & = & a_1^T x \\ & \vdots & \\ y_\nu & = & a_\nu^T x \\ y_{\nu+1} & = & x_{\nu+1} \\ & \vdots & \\ y_n & = & x_n \end{array}$$

ou seja,

$$y = \begin{pmatrix} B & N \\ 0 & I \end{pmatrix} x = \bar{B} x \, .$$

É fácil ver que $\bar{B}$ é não-singular. Então, temos $x = \bar{B}^{-1} y$ e podemos reformular o problema (9.2.1) da seguinte maneira

$$\begin{array}{ll} \text{Minimizar} & \bar{f}(y) \equiv f(\bar{B}^{-1} y) \\ \text{sujeita a} & \begin{array}{rcl} y_i & \geq & b_i \, , \; i = 1, \ldots, \nu \\ a_i^T \bar{B}^{-1} y & \geq & b_i \, , \; i = \nu + 1, \ldots, n \, . \end{array} \end{array} \tag{9.3.1}$$

Seja $\overline{y} = \bar{B} x_k$. Como $x_k \in F_I$, temos que $\overline{y}_i = b_i$ se $i \in I$ e $a_i^T \bar{B}^{-1} \overline{y} > b_i$ se $i \notin I$. Portanto, as direções factíveis de descida, a partir de $\bar{y}$, para (9.3.1) são as mesmas que as do problema onde as restrições inativas são eliminadas:

$$\begin{array}{ll} \text{Minimizar} & \bar{f}(y) \\ \text{sujeita a} & y_i \geq b_i \, , \; i = 1, \ldots, \nu \, . \end{array} \tag{9.3.2}$$

Agora, como fizemos no capítulo 4 com as quadráticas em caixas, podemos definir aqui a direção de Cauchy $\bar{\nabla}\bar{f}(\bar{y})$ por

$$[\bar{\nabla}\bar{f}(\bar{y})]_i = 0 \text{ se } \bar{y}_i = b_i \text{ e } [\nabla\bar{f}(\bar{y})]_i \geq 0 \ ;$$

$$[\bar{\nabla}\bar{f}(\bar{y})]_i = -[\nabla\bar{f}(\bar{y})]_i \text{ nos outros casos.}$$

O ponto $\bar{y}$ será estacionário de primeira ordem de (9.2.1), (9.2.2) e (9.3.1) se, e somente se,

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

Se $\bar{\nabla}\bar{f}(\bar{y}) \neq 0$ esse vetor é uma direção factível e de descida a partir de $\bar{y}$. Escrevendo

$$\bar{\nabla}\bar{f}(\bar{y}) = (\bar{\nabla}_C\bar{f}(\bar{y})^T, \bar{\nabla}_I\bar{f}(\bar{y})^T)^T,$$

com $\bar{\nabla}_C\bar{f}(\bar{y}) \in I\!R^\nu$, $\bar{\nabla}_I\bar{f}(\bar{y}) \in I\!R^{n-\nu}$, teremos também que x_k é ponto estacionário de (9.1.2) se, e somente se, $\bar{\nabla}_I\bar{f}(\bar{y}) \in I\!R^{n-\nu} = 0$. Portanto, é natural que a decisão sobre abandonar a face ou não dependa de uma avaliação do quociente

$$quoc = \frac{\|\bar{\nabla}_I\bar{f}(\bar{y})\|}{\|\bar{\nabla}\bar{f}(\bar{y})\|}.$$

Claramente, $quoc \in [0, 1]$ e a decisão de abandono será obrigatória quando $quoc = 0$, já que nesse caso nada mais podemos esperar de um algoritmo que use apenas derivadas primeiras para minimizar (9.1.2). Por outro lado, se $quoc = 1$ deveremos ficar dentro da face, pois todo o potencial de descida se encontra dentro dela. Assim, nada mais sensato que decidir pela saida (Passo 2) quando $quoc \leq TOL$ onde TOL é uma tolerância entre 0 e 1. Toda analogia com o algoritmo dado no capítulo 4 para minimizar quadráticas em caixas é proposital. Uma vez decidido o abandono da face, temos bastante liberdade para escolher a direção de saída, já que, em princípio, qualquer direção no espaço y que seja factível, de descida, e tenha alguma das ν primeiras coordenadas maiores que 0, servirá para esse fim. Uma candidata natural é $\bar{d} = \bar{\nabla}\bar{f}(\bar{y})$. Assim, tomando $t > 0$ suficientemente pequeno, teremos que $x_k + t\bar{B}^{-1}\bar{d} \in (\Omega - F_I)$ e $f(x_k + t\bar{B}^{-1}\bar{d}) < f(x_k)$.

A pressa em sair da face, provocada, por exemplo, por um valor de TOL muito próximo de 1, pode ocasionar um fenômeno chamado de "ziguezague". Uma face pode ser abandonada e retomada um número infinito de vezes, impedindo a convergência do método. Existem muitos procedimentos "anti-ziguezague", introduzidos para driblar tão desagradável comportamento. Ver [69]. Na minimização de quadráticas em caixas, por exemplo, vimos que a saída pelo gradiente chopado elimina toda possibilidade de não-convergência.

Quando os gradientes das restrições que definem I são linearmente dependentes, dizemos que estamos em um ponto degenerado. Grande parte da teoria do método Simplex em programação linear (ver, por exemplo [34]) está destinada a analisar esse caso. Felizmente, se a função objetivo é não-linear, podemos usar um artifício que nos permite resolver a situação evocando o caso linear. Com efeito, suponhamos que, em $x_k \in F_I$, temos $I = 1, \ldots, \nu$ e $\{a_1, \ldots, a_\nu\}$ dependentes. Consideramos o problema auxiliar

$$\text{Minimizar } \nabla f(x_k)^T d, \text{ sujeita a } a_i^T d \geq 0, i \in I. \tag{9.3.3}$$

Se aplicamos o método Simplex para resolver (9.3.3) com o ponto inicial 0, sabemos que esse método detectará, em tempo finito, que 0 é solução de (9.3.3), ou encontrará d factível tal que $\nabla f(x_k)^T d < 0$, usando procedimentos contra a ciclagem, se for necessário. Tal direção é uma direção factível e de descida para (9.2.1), que nos permitirá continuar o processo.

Exercício 9.5: Justificar cuidadosamente as afirmações no texto relativas à mudança de variáveis, em particular, provar a não singularidade de $\bar{B}$.

Exercício 9.6: Analisar a estratégia de escape definida pelos métodos do tipo gradiente projetado para restrições lineares (ver, por exemplo, [129], p.330).

Exercício 9.7: Justificar a estratégia de escape adotada pelo método Simplex.

Exercício 9.8: Analisar o comportamento do método Simplex para pontos não regulares.

Exercício 9.9: Refazer a análise das seções 9.2 e 9.3 com outras formas de descrever o politopo Ω.

9.4 Redução a caixas

O leitor incomodado com as fatorações, a convergência duvidosa e as perigosas degenerações da estratégia das restrições ativas, se sentirá confortado pelos resultados desta seção. Provaremos que, quando f é convexa e o politopo é limitado, o problema (9.1.1) pode ser reduzido a um problema de minimização em caixas, cuja teoria, como vimos, é bastante sólida e adaptável a situações de grande porte. Aqui, mediante a introdução de variáveis de folga, se necessário, (9.1.1) terá sempre a forma padrão:

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & Ax = b\,,\ x \geq 0\,, \end{array} \tag{9.4.1}$$

onde $f \in C^2(I\!R^n)$ é convexa e $\Omega = \{x \in I\!R^n \mid Ax = b\,,\ x \geq 0\}$.

As condições de otimalidade de primeira ordem de (9.4.1) são

$$\begin{array}{rcl} \nabla f(x) + A^T y - z & = & 0 \\ Ax - b & = & 0 \\ x^T z & = & 0 \\ x \geq 0\,,\ z \geq 0\,. & & \end{array} \tag{9.4.2}$$

Definimos, para $\|\cdot\| = \|\cdot\|_2$,

$$\Phi(x,y,z) = \frac{1}{2}\left(\|\nabla f(x) + A^T y - z\|^2 + \|Ax - b\|^2 + (x^T z)^2\right)\,,$$

e consideramos o problema

$$\begin{array}{ll} \text{Minimizar} & \Phi(x,y,z) \\ \text{sujeita a} & x \geq 0\,,\ z \geq 0\,. \end{array} \tag{9.4.3}$$

À primeira vista, ao resolvermos (9.4.3), esperamos apenas encontrar pontos estacionários, não necessariamente minimizadores globais, já que $\Phi(x,y,z)$ não é uma função convexa. No entanto, o resultado a seguir assegura que todo ponto estacionário de (9.4.3) é um minimizador global para este problema satisfazendo (9.4.2) e, portanto, resolver (9.4.3) é equivalente a resolver (9.4.1). Ver [83], [84] e [85] para extensões e variações deste teorema.

Teorema 9.4.1

Se $f \in C^2(I\!R^n)$ é convexa e o politopo Ω é não vazio e limitado, então (9.4.3) admite pelo menos um ponto estacionário (KKT) e todo ponto estacionário (x_*, y_*, z_*) de (9.4.3) é um minimizador global com $\Phi(x_*, y_*, z_*) = 0$.

Prova: A primeira parte é imediata. Como Ω é limitado e f é contínua, existe um minimizador global para o problema (9.4.1). Este minimizador tem que satisfazer (9.4.2) e, portanto, é um minimizador global de (9.4.3).

Vamos supor que (x, y, z) seja um ponto estacionário do problema (9.4.3). Então existem γ, $\mu \in I\!R^n$ tais que

$$A^T(Ax-b) + \nabla^2 f(x)(\nabla f(x) + A^T y - z) + (x^T z)z - \gamma = 0 \ , \tag{9.4.4}$$

$$A(\nabla f(x) + A^T y - z) = 0 \ , \tag{9.4.5}$$

$$-(\nabla f(x) + A^T y - z) + (x^T z)x - \mu = 0 \ , \tag{9.4.6}$$

$$\gamma^T x = 0 \ , \tag{9.4.7}$$

$$\mu^T z = 0 \ , \tag{9.4.8}$$

$$x \geq 0 \, , \ z \geq 0 \, , \ \gamma \geq 0 \, , \ \mu \geq 0 \ . \tag{9.4.9}$$

Por (9.4.5) e (9.4.6) temos que

$$(x^T z)x - \mu \in \mathcal{N}(A) \ , \tag{9.4.10}$$

onde $\mathcal{N}(A)$ é o núcleo da matriz A.

Portanto, pré-multiplicando (9.4.4) por $(x^T z)x - \mu$ e usando (9.4.6), obtemos

$$((x^T z)x-\mu)^T \nabla^2 f(x)((x^T z)x-\mu)+((x^T z)x-\mu)^T((x^T z)z-\gamma) = 0 \ . \tag{9.4.11}$$

Como $\nabla^2 f$ é semi-definida positiva, (9.4.11) implica em

$$((x^T z)x - \mu)^T((x^T z)z - \gamma) \leq 0 \ .$$

Logo, por (9.4.7) e (9.4.8) segue que

$$(x^T z)^3 + \mu^T \gamma \leq 0 \ . \tag{9.4.12}$$

Assim, por (9.4.9) temos

$$x^T z = 0 \tag{9.4.13}$$

e

$$\mu^T \gamma = 0 \ . \tag{9.4.14}$$

Por (9.4.6) e (9.4.13),

$$-(\nabla f(x) + A^T y - z) = \mu \geq 0 \ . \tag{9.4.15}$$

Mas, por (9.4.5), $-(\nabla f(x) + A^T y - z) \in \mathcal{N}(A)$. Portanto, como Ω é limitado, a equação (9.4.15) implica necessariamente em

$$-(\nabla f(x) + A^T y - z) = 0 \ . \tag{9.4.16}$$

Então, por (9.4.4), (9.4.13) e (9.4.16) temos

$$A^T(Ax - b) = \gamma \geq 0 \; . \tag{9.4.17}$$

Agora, (9.4.17) e (9.4.7) são as condições de otimalidade (necessárias e suficientes) do problema quadrático convexo

$$\begin{array}{ll} \text{Minimizar} & \frac{1}{2}\|Ax - b\|^2 \\ \text{sujeita a} & x \geq 0 \; . \end{array} \tag{9.4.18}$$

Como Ω é não vazio, temos que $Ax = b$. Esta igualdade, juntamente com (9.4.13) e (9.4.16) completam a prova. **QED**

O problema

$$\begin{array}{ll} \text{Minimizar} & \frac{1}{2}\left(\|\nabla f(x) + A^T y - z\|^2 + \|Ax - b\|^2 + x^T z\right) \\ \text{sujeita a} & x \geq 0 \, , \; z \geq 0 \end{array} \tag{9.4.19}$$

é obviamente equivalente a (9.4.3). No entanto, (9.4.19) pode admitir pontos estacionários que não são minimizadores globais. De fato, basta considerarmos o problema de minimizar x sujeito a $0 \leq x \leq 2$ ou, no formato (9.4.1), minimizar x_1 sujeito a $x_1 + x_2 = 2$, $x_1 \geq 0$, $x_2 \geq 0$. O problema da forma (9.4.19) associado a este problema trivial admite o ponto estacionário $x = (2, 0)^T$ e $z = (0, 0)^T$, que naturalmente não é um minimizador global.

9.5 Pontos interiores

A revolução dos métodos de pontos interiores começou em 1984 com o lançamento do "método de Karmarkar" [124]. Por primeira vez na história era anunciado um algoritmo eficiente na prática e, ao mesmo tempo, polinomial, para o problema de programação linear. Desde então, foram escritos centenas de artigos introduzindo e analisando algoritmos desse tipo. O "survey" [103] é, provavelmente, a melhor referência disponível para o estado da arte até 1992. Nesta seção nos limitaremos a introduzir a idéia "affine-scaling" ([58], [3], [193], [5]), uma das mais fecundas geradoras de algoritmos de pontos interiores, no contexto da minimização de funções gerais com restrições lineares.

A idéia dos métodos de pontos interiores é provocativamente contraditória com o método Simplex, e com as estratégias de restrições ativas em geral. Mesmo sabendo que, com alta probabilidade, a solução está na fronteira (com certeza em um vértice no caso da programação linear), esses algoritmos geram iterandos que permanecem sempre no interior do conjunto. Em vez de apostar na face em que provavelmente se encontra o minimizador, de acordo com a informação disponível, os métodos de pontos interiores evitam o fracasso de repetidos abandonos seguindo caminhos curvos na região onde nenhuma restrição é ativa.

A tática "affine-scaling" se baseia em subproblemas onde a região Ω é substituida por um elipsóide interno, que nos permitiremos identificar com uma região de confiança. Primeiro, acrescentemos variáveis de folga em (9.1.1), de maneira que nosso problema é

$$\text{Minimizar } f(x) \text{ sujeita a } Ax - z = b, \quad z \geq 0. \tag{9.5.1}$$

O ponto inicial x_0, assim como todos os iterandos x_k, será interior a Ω, ou seja, $Ax_k > b$ ($z_k > 0$) para todo k.

O maior elipsóide no espaço z, centrado em z_k, contido no ortante positivo e com eixos paralelos aos eixos coordenados é dado por

$$\sum_{i=1}^{m} \frac{(z_i - [z_k]_i)^2}{[z_k]_i^2} \leq 1, \tag{9.5.2}$$

ou seja

$$(z - z_k)^T Z_k^{-2} (z - z_k) \leq 1, \tag{9.5.3}$$

onde Z_k é a matriz diagonal cujas entradas são $[z_k]_i, i = 1, \ldots, m$. Portanto, é bastante natural considerar o subproblema

$$\text{Minimizar } \tilde{f}(x) \text{ sujeita a } Ax - z = b, \quad (z - z_k)^T Z_k^{-2} (z - z_k) \leq 1. \tag{9.5.4}$$

onde $\tilde{f}(x)$ é uma aproximação de $f(x)$, construída com a informação disponível em x_k. Por exemplo,

$$\tilde{f}(x) = f(x), \tag{9.5.5}$$

$$\tilde{f}(x) = f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k) \tag{9.5.6}$$

ou

$$\tilde{f}(x) = f(x_k) + \nabla f(x_k)(x - x_k). \tag{9.5.7}$$

Em todos os casos, $\nabla \tilde{f}(x_k) = \nabla f(x_k)$. Desenvolvendo (9.5.4), o subproblema toma a forma

$$\text{Minimizar } \tilde{f}(x) \text{ sujeita a } (Ax - b - z_k)^T Z_k^{-2} (Ax - b - z_k) \leq 1, \tag{9.5.8}$$

ou, usando que $z_k = Ax_k - b$,

$$\text{Minimizar } \tilde{f}(x) \text{ sujeita a } (x - x_k)^T A^T Z_k^{-2} A(x - x_k) \leq 1. \tag{9.5.9}$$

Suponhamos que o posto de A é n. O subproblema (9.5.9) pode ser resolvido com apenas uma fatoração de Cholesky no caso (9.5.7). Se $\tilde{f}(x)$ é quadrática, pode ser reduzido, pela mudança de variáveis $y = [A^T Z_k^{-2} A]^{\frac{1}{2}}(x - x_k)$ a minimizar quadráticas em bolas, problema que estudamos no capítulo 4 e relembramos no capítulo 7 . Soluções aproximadas de (9.5.9) no caso (9.5.5) podem ser obtidas usando os algoritmos de minimização em bolas descritos em [142].

Chamamos $\tilde{d}_k = \bar{x} - x_k$ a uma solução aproximada de (9.5.9). A aproximação deve ser, pelo menos no sentido de que

$$\tilde{f}(x_k + \tilde{d}_k) < \tilde{f}(x_k) \text{ sempre que } \nabla f(x_k) \neq 0.$$

Se $\tilde{f}$ é convexa ou quadrática, isto implica que $\nabla f(x_k)^T \tilde{d}_k < 0$. Nesse caso, definimos

$$\widehat{d}_k = \alpha_k \tilde{d}_k$$

onde α_k é o maior α tal que $[x_k, x_k + \alpha \tilde{d}_k] \subset \Omega$ e

$$d_k = \beta \widehat{d}_k$$

onde $\beta \in (0, 1)$ é muito próximo de 1, digamos 0.95, de maneira que $x_k + d_k$ é interior mas está próximo da fronteira. Finalmente, $x_{k+1} = x_k + td_k$, com $t \in [0, 1]$, é obtido por um processo de backtracking, até satisfazer uma condição de tipo Armijo.

Quando $\tilde{f}$ não é convexa nem quadrática, é mais coerente, no caso de $f(x_k + d_k)$ não ser suficientemente menor que $f(x_k)$, definir sucessivos subproblemas mediante diminuição do tamanho da "região de confiança".

Exercício 9.10: Estabelecer a relação entre o posto de A e a limitação do politopo Ω. Justificar a suposição de que o posto de A é n.

Exercício 9.10: Formular o Algoritmo da Seção 9.5 para o caso linear $f(x) = c^T x$. Mostrar que a solução do subproblema é a de um sistema linear com matriz definida positiva. Relacionar "quase-singularidade" dessa matriz com pontos degenerados (não-regulares) da fronteira.

Exercício 9.11: Modificar o algoritmo (colocando salvaguardas) de maneira que sua convergência possa ser provada usando técnicas de regiões de confiança.

Exercício 9.12: Detalhar a mudança de variáveis que faz com que o subproblema tenha como domínio uma bola.

Exercício 9.13: Justificar a afirmação "se $\tilde{f}$ é convexa ou quadrática, d_k é uma direção de descida". Mostrar que não é verdade no caso não-convexo.

Chapter 10

Penalidade

Não apenas em otimização, mas também em outras áreas da atividade humana, procura-se converter problemas complexos em outros cuja resolução é conhecida. Os leitores satisfeitos com as estratégias introduzidas até aqui para minimização sem restrições, minimização em caixas e em politopos se sentiriam agradecidos se qualquer outro problema de otimização *com restrições não lineares* pudessse ser reduzido àqueles. A penalidade é o procedimento mais radical para isso. Mediante ele, a não-satisfação (ou o "risco de não-satisfação") de uma restrição é sancionada com um acréscimo da função objetivo, de maneira que a função que define a restrição é eliminada como tal e substituída por um termo introduzido no objetivo.

Na chamada "penalidade interna" a função objetivo é modificada agregando um termo funcional que tende a infinito quando o ponto se aproxima da fronteira do conjunto factível. Forma-se assim uma espécie de barreira intransponível: métodos irrestritos começando no interior da região são desencorajados de se aproximar do contorno devido a valores muito altos do objetivo. Por esse motivo, os métodos de penalidade interna são também conhecidos por *métodos de barreira.* Esses são, por outro lado, os mais antigos métodos de pontos interiores, com prestígio radicalmente incrementado após a revolução que seguiu-se ao trabalho de Karmarkar [124].

Na penalidade externa, muitas vezes denominada simplesmente de *penalidade*, acrescenta-se na função objetivo um termo cujo custo aumenta com a violação das restrições. A solução de um problema penalizado externamente está, geralmente, fora do conjunto factível, mas se aproxima dele quando o termo de penalidade é muito grande.

A razão pela qual a penalidade não é o procedimento universal para lidar com restrições é que o parâmetro que deve multiplicar à função-restrição para castigar violação (na externa) ou o risco de violação (na interna) provoca, ao tomar valores extremos, pesado mal-condicionamento do problema. Também peca a filosofia penalizadora por outro defeito essencial: a própria estrutura do problema é transtornada quando uma restrição é acrescida à função objetivo, em geral, complicando a fisionomia desta. No entanto, todas as estratégias de penalidade estão vivas na otimização contemporânea por sua simplicidade, adaptabilidade para problemas de grande porte, e capacidade de se enriquecer automaticamente com os progressos realizados na resolução de problemas mais simples.

10.1 Métodos de barreiras

Os métodos de penalidade interna ou barreiras foram originalmente propostos para lidar com restrições não lineares de desigualdade, quando, via de regra, o conjunto factível tem interior não vazio.

Consideraremos, para a introdução dos métodos de penalidade interna, problemas de otimização da seguinte forma:

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & c(x) \geq 0\,,\ x \in \mathcal{D}\,, \end{array} \tag{10.1.1}$$

onde $\mathcal{D}$ é um subconjunto de $I\!R^n$, $c : I\!R^n \to I\!R^m$, $f\,,\ c \in C^0(\mathcal{D})$ e

$$\Omega = \{x \in \mathcal{D} \mid c(x) \geq 0\}$$

tem *interior relativo* não vazio, denotado por $\Omega^\circ = \{x \in \mathcal{D} \mid c(x) > 0\}$. Vamos supor que (10.1.1) tem minimizador global.

Podemos transformar (10.1.1) em um problema irrestrito com função objetivo $f(x) + tB(x)$, $t > 0$, onde a *função barreira* B satisfaz os seguintes *axiomas*:

(i) $B(x)$ está definida e é contínua para todo $x \in \Omega^\circ$.

(ii) $B(x) \geq 0$ para todo $x \in \Omega^\circ$.

(iii) Se $\{x_k\} \subset \Omega$, $c(x_k) > 0$ para todo k e $\lim_{k\to\infty} c_i(x_k) = 0$ para algum $i \in \{1, \ldots, m\}$, então $\lim_{k\to\infty} B(x_k) = \infty$.

A diferenciabilidade da função barreira não é essencial para o método em si. Entretanto, se a função objetivo original é diferenciável, torna-se interessante que B também o seja, pois assim podem ser aplicadas técnicas para minimização sem restrições que explorem ao máximo a estrutura do problema.

Tendo por princípio os três axiomas acima, podemos estabelecer o método básico de penalidade interna:

Algoritmo 10.1.1 - Barreiras.

Dados $t_1 > 0$, $x_0 \in \Omega^\circ$, $k = 1$.

(1) Calcular $x_k \equiv x(t_k)$ solução global de

$$\begin{array}{ll} \text{Minimizar} & f(x) + t_k B(x) \\ \text{sujeita a} & x \in \Omega^\circ. \end{array} \tag{10.1.2}$$

(2) Escolher t_{k+1} tal que $0 < t_{k+1} < t_k$, $k \leftarrow k + 1$ e voltar para (1).

Para obter x_k, no Passo 1 do algoritmo, usamos um método qualquer para minimizar funções com a restrição $x \in \mathcal{D}$. Quase sempre, se tratará de um algoritmo iterativo, e o ponto inicial recomendável nesse caso será x_{k-1}, embora diversas estratégias de aceleração possam ser implementadas. Estritamente falando, no problema penalizado (10.1.2) aparecem as restrições $c_i(x) > 0$ além de $x \in \mathcal{D}$. No entanto, como a função objetivo de (10.1.2) tende a infinito quando x tende à fronteira, estamos autorizados a supor que um algoritmo irrestrito (ou melhor, restrito apenas a $\mathcal{D}$), não sentirá a menor atração por pontos muito próximos ao contorno, e que, portanto, permanecerá também afastado de pontos externos. Às vezes, pode ser necessária alguma modificação leve do algoritmo "irrestrito" para garantir a permanência no interior de Ω. Sabemos, por outro lado, que encontrar minimizadores globais costuma ser muito difícil e que, usando métodos iterativos, não poderemos, de fato, atingir exatamente a solução de (10.1.2). Por isso, na prática, x_k será apenas uma solução "aproximada" de (10.1.2). As propriedades do método, no entanto, emergem de maneira poderosa e surpreendentemente simples quando consideramos sua versão exata.

A seqüência de parâmetros de penalidade t_k deve tender a 0. Uma regra mágica é fazer $t_1 = 1$ e $t_{k+1} = t_k/10$ para todo k. Para problemas não muito complicados, pode-se tentar resolver um único subproblema com um parâmetro muito pequeno, na expectativa que a solução computada esteja próxima da solução do problema original. Esta estratégia é chamada "shortcut" (atalho) em [69] e, às vezes, pode ser fragorosamente ineficiente.

Existem dois exemplos clássicos de funções barreira: a *função barreira* *inversa*

$$B(x) = \sum_{i=1}^{m} \frac{1}{c_i(x)} \tag{10.1.3}$$

e a *função barreira logarítmica*

$$B(x) = -\sum_{i=1}^{m} \log\left(c_i(x)\right) . \tag{10.1.4}$$

A função (10.1.4) pode assumir valores negativos, e portanto, não cumpre o axioma (ii). Porém, no caso em que Ω é limitado, veremos que trabalhar com ela é equivalente a fazê-lo com uma outra função que sim satisfaz os axiomas. Observemos, primeiro, que quando o interior de Ω é limitado, então a função (10.1.4) é limitada inferiormente.

Exercício 10.1: Provar a afirmação anterior.

Seja $M \in I\!R$ tal que $B(x) > M$ para todo $x \in \Omega^\circ$ e consideremos

$$\tilde{B}(x) = -\sum_{i=1}^{m} \log\left(c_i(x)\right) - M . \tag{10.1.5}$$

É fácil ver que $\tilde{B}$ satisfaz os três axiomas da função barreira. Agora, o problema com barreira associado a $\tilde{B}$:

$$\begin{array}{ll} \text{Minimizar} & f(x) + t\tilde{B}(x) \\ \text{sujeita a} & x \in \Omega^\circ \ , \end{array}$$

coincide com

$$\begin{array}{ll} \text{Minimizar} & f(x) + tB(x) - tM \\ \text{sujeita a} & x \in \Omega^\circ \ , \end{array}$$

que é equivalente a

$$\begin{array}{ll} \text{Minimizar} & f(x) + tB(x) \\ \text{sujeita a} & x \in \Omega^\circ \ . \end{array}$$

Assim, a função logarítmica (10.1.4) pode ser usada como barreira sem nenhum prejuízo.

De agora em diante, definimos

$$\mathcal{Q}(x,t) = f(x) + tB(x) \ , \tag{10.1.6}$$

e passamos a provar as propriedades fundamentais do Algoritmo 10.1.1.

Lema 10.1.2

Seja $\{x_k\}$ *a seqüência gerada pelo Algoritmo 10.1.1. Então*

$$\begin{aligned} \mathcal{Q}(x_{k+1}, t_{k+1}) &\leq \mathcal{Q}(x_k, t_k) & (10.1.7)\\ B(x_k) &\leq B(x_{k+1}) & (10.1.8)\\ f(x_{k+1}) &\leq f(x_k)\ . & (10.1.9) \end{aligned}$$

Prova: Como a seqüência de parâmetros penalizadores é monótona decrescente, pelo axioma (ii) da função barreira B e pelo fato de $\{x_k\}$ ser uma seqüência de minimizadores globais de (10.1.2) temos:

$$\begin{aligned} \mathcal{Q}(x_{k+1}, t_{k+1}) &= f(x_{k+1}) + t_{k+1} B(x_{k+1})\\ &\leq f(x_k) + t_{k+1} B(x_k)\\ &\leq f(x_k) + t_k B(x_k)\\ &= \mathcal{Q}(x_k, t_k)\ . \end{aligned}$$

Para mostrarmos a validade de (10.1.8), temos:

$$\mathcal{Q}(x_{k+1}, t_{k+1}) = f(x_{k+1}) + t_{k+1} B(x_{k+1}) \leq f(x_k) + t_{k+1} B(x_k)\ . \quad (10.1.10)$$

Por outro lado,

$$\mathcal{Q}(x_k, t_k) = f(x_k) + t_k B(x_k) \leq f(x_{k+1}) + t_k B(x_{k+1})\ . \quad (10.1.11)$$

Subtraindo (10.1.11) de (10.1.10) obtemos

$$(t_{k+1} - t_k) B(x_{k+1}) \leq (t_{k+1} - t_k) B(x_k)$$

e como $t_{k+1} - t_k \leq 0$ segue que $B(x_k) \leq B(x_{k+1})$.

Por fim, usando (10.1.8) temos

$$\begin{aligned} f(x_{k+1}) + t_{k+1} B(x_{k+1}) &\leq f(x_k) + t_{k+1} B(x_k)\\ &\leq f(x_k) + t_{k+1} B(x_{k+1})\ . \end{aligned}$$

Logo, $f(x_{k+1}) \leq f(x_k)$, o que completa a prova. **QED**

No Teorema 10.1.3 (ver [29]), provaremos que, se usarmos o Algoritmo 10.1.1, conseguiremos uma aproximação arbitrariamente próxima de um minimizador global do problema original, para k suficientemente grande.

Teorema 10.1.3

Seja $\{x_k\}$ *a seqüência de minimizadores (10.1.2) gerada pelo Algoritmo 10.1.1, com* $\lim_{k\to\infty} t_k = 0$*. Então, todo ponto limite de* $\{x_k\}$ *é minimizador global de (10.1.1).*

Prova: Chamemos, para $k = 0,\, 1,\, 2, \ldots,$

$$b_k = \min\{\mathcal{Q}(x, t_k) \mid x \in \Omega^\circ\}\ . \tag{10.1.12}$$

Então, $b_k \ \geq\ b_{k+1}$ para todo k.

Agora, seja

$$b = \min\{f(x) \mid x \in \Omega\}\ .$$

Claramente,

$$b_0 \ \geq\ b_1 \ \geq\ \cdots \ \geq\ b_k \ \geq\ b_{k+1} \ \cdots \ \geq\ b\ .$$

Como $\{b_k\}$ é uma seqüência decrescente e inferiormente limitada, é convergente:

$$\lim_{k\to\infty} b_k = \bar{b}\ . \tag{10.1.13}$$

Se $\bar{b} \neq b$, então $\bar{b} > b$.

Seja x_* um minimizador global do problema (10.1.1). Como f é contínua, existe uma bola $\mathcal{B}$ com centro em x_* tal que para todo $x \in \mathcal{Q} \cap \Omega^\circ$,

$$f(x) \ <\ \bar{b} - \frac{1}{2}(\bar{b} - b)\ . \tag{10.1.14}$$

Agora, como $0 < t_{k+1} < t_k$ e $B(x) \geq 0$ para $x \in \Omega^\circ$, temos

$$0 < t_{k+1}\, B(x) < t_k\, B(x)$$

para todo $x \in \Omega^\circ$. Portanto, $\lim\limits_{k\to\infty} t_k\, B(x) = 0$ para $x \in \Omega^\circ$. Assim, tomemos $x' \in \mathcal{Q} \cap \Omega^\circ$. Para k suficientemente grande,

$$t_k\, B(x') \ <\ \frac{1}{4}(\bar{b} - b)\ . \tag{10.1.15}$$

Então, por (10.1.14) e (10.1.15), para k suficientemente grande,

$$\mathcal{Q}(x', t_k) \ <\ \bar{b} - \frac{1}{4}(\bar{b} - b) \ <\ \bar{b}\ ,$$

o que contradiz (10.1.12)-(10.1.13). Portanto, $\bar{b} = b$.

Agora, seja K um subconjunto infinito de $\mathcal{N}$ tal que

$$\lim_{k\in K} x_k = \bar{x}$$

onde $\bar{x} \in \Omega$. Suponhamos que $\bar{x} \neq x_*$, solução global de (10.1.1), com $f(\bar{x}) > f(x_*)$.

Então, a seqüência $\{(f(x_k) - f(x_*)) + t_k\, B(x_k)\}_{k \in K}$ não pode convergir a zero, o que contradiz o fato de que $b_k - b \to 0$. Logo, $\bar{x} = x_*$ ou $\bar{x} \neq x_*$ mas $f(\bar{x}) = f(x_*)$. Ou seja, todo ponto limite da seqüência gerada pelo Algoritmo 10.1.1 é uma solução global do problema (10.1.1). **QED**

Um defeito estrutural dos métodos de penalidade interna é que restrições de igualdade não podem participar da definição da função $B(x)$. Assim, se no problema original aparecem restrições desse tipo, elas devem ser conservadas no conjunto $\mathcal{D}$, mas não podem contribuir na penalidade. Portanto, se não soubermos minimizar funções com a restrição $\mathcal{D}$, a barreira é inaplicável.

Não menos importante é a questão da estabilidade numérica, já que os subproblemas tornam-se computacionalmente mais difíceis de se resolver à medida que o parâmetro t_k diminui. Vejamos porque isso ocorre no seguinte exemplo:

$$\begin{array}{ll} \text{Minimizar} & f(x_1, x_2) = (x_1+1)^2 + (x_2-1)^2 \\ & x_1 \geq 0 \ , \end{array} \tag{10.1.16}$$

cuja solução é $x_* = (0 \ \ 1)^T$. Vamos considerar a função barreira

$$B(x) = -\log(x_1).$$

Então

$$\mathcal{Q}(x,t) = (x_1+1)^2 + (x_2-1)^2 - t\, \log(x_1).$$

Portanto,

$$\nabla_x \mathcal{Q}(x,t) = \begin{pmatrix} 2(x_1+1) - \frac{t}{x_1} \\ 2(x_2-1) \end{pmatrix}$$

e

$$\nabla^2_{xx} \mathcal{Q}(x,t) = \begin{pmatrix} 2 + \frac{t}{x_1^2} & 0 \\ 0 & 2 \end{pmatrix}.$$

Os pontos estacionários com $x_1 > 0$ são da forma $\bar{x} = \left(\frac{-1+\sqrt{1+2t}}{2} \ \ 1 \right)^T$, $t > 0$ e então

$$\nabla^2 \mathcal{Q}(\bar{x},t) = \begin{pmatrix} 2 + \frac{2t}{t+1-\sqrt{1+2t}} & 0 \\ 0 & 2 \end{pmatrix}.$$

Assim, como $\lim\limits_{t \to 0} \frac{2t}{t+1-\sqrt{1+2t}} = \infty$, segue que o número de condição da matriz Hessiana $\nabla^2 \mathcal{Q}(x,t)$ tende a infinito quando $t \to 0$, o que retrata algebricamente a dificuldade crescente dos subproblemas. Geometricamente, as curvas de nível das funções $\mathcal{Q}$ ficam cada vez mais alongadas, o que torna mais e mais imprecisa a determinação do minimizador.

O ponto de vista tradicional (até meados da década de 80) era que as restrições incorporadas na função objetivo deviam ser as mais complicadas, pela dificuldade intrínseca a sua manipulação direta. Penalizar em relação a restrições simples teria sido considerado um sacrilégio. A aparição dos métodos de pontos interiores em programação linear mostrou que a situação é bem mais confusa, pois muitos desses métodos podem ser interpretados como penalidade logarítmica em relação às restrições *extremamente simples* $x_i \geq 0$. Consideremos o problema de minimização com restrições de igualdade na sua forma padrão :

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & Ax = b\,,\ x \geq 0\,, \end{array} \tag{10.1.17}$$

onde $A \in I\!R^{m \times n}$, $m \leq n$ e posto$(A) = m$.

Utilizando a função barreira logarítmica, temos o seguinte subproblema, apenas com restrições lineares de igualdade:

$$\begin{array}{ll} \text{Minimizar} & f(x) - t \displaystyle\sum_{i=1}^{n} \log(x_i) \\ \text{sujeita a} & Ax = b\,. \end{array} \tag{10.1.18}$$

As condições de otimalidade de (10.1.18) correspondem a um sistema não-linear com $n + m$ equações e $n + m$ variáveis:

$$\begin{array}{rcl} \nabla f(x) - t \begin{pmatrix} \frac{1}{x_1} \\ \vdots \\ \frac{1}{x_n} \end{pmatrix} + A^T y & = & 0 \\ Ax & = & b\,. \end{array} \tag{10.1.19}$$

A matriz Jacobiana do sitema (10.1.19) é dada por

$$\begin{pmatrix} \nabla^2 f(x) + t\,X^{-2} & A^T \\ A & 0 \end{pmatrix} \tag{10.1.20}$$

onde $X = \text{diag}(x_1, \ldots, x_n)$. O número de condição desta matriz cresce quando $t \to 0$ e alguma componente $x_i\,,\ i = 1, \ldots, n$ se aproxima de zero.

O mal-condicionamento inerente ao método de barreira pode ser contornado com a seguinte mudança de variáveis:

$$z_i = \frac{t}{x_i}\,, \quad i = 1,\ \ldots\,,\ n\,.$$

Então (10.1.19) pode ser reescrito como

$$\begin{array}{rcl} \nabla f(x) \; - \; z \; + \; A^T y & = & 0 \\ Ax & = & b \\ x_i \, z_i \; - \; t & = & 0 \,, \quad i = 1, \; \ldots \,, \; n \,. \end{array} \tag{10.1.21}$$

O sistema aumentado (10.1.21), com $2n+m$ equações e $2n+m$ incógnitas, tem o seguinte Jacobiano:

$$\begin{pmatrix} \nabla^2 f(x) & A^T & -I \\ A & 0 & 0 \\ Z & 0 & X \end{pmatrix} \tag{10.1.22}$$

onde $Z = \text{diag}(z_1, \ldots, z_n)$. Além de (10.1.22) independer de t, se tivermos complementariedade estrita, isto é, se $x_i \, z_i = 0$ com $x_i \neq 0$ ou $z_i \neq 0$, então (10.1.22) tem posto completo (um bom exercício para o leitor). O sistema (10.1.21) só será mal condicionado se o problema original (10.1.17) o for.

Assim, se ao invés de trabalharmos com (10.1.18), resolvermos (10.1.21), quando $t = 0$ teremos as condições Karush-Kuhn-Tucker do problema original (10.1.17). No caso em que (10.1.17) é mal-condicionado, (10.1.21) pode ser resolvido monitorando-se a homotopia obtida quando $t \to 0$, através de alguma variação do método de Newton inexato para sistemas não lineares. Em programação linear, a homotopia (10.1.21) é o fundamento dos métodos primais-duais, que, na década dos 90 são os algoritmos de pontos interiores com maior prestígio para esse problema.

10.2 Penalidade externa

Os métodos de penalidade externa ou, simplesmente, penalidade, podem ser aplicados ao problema de otimização em seu formato mais geral:

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & x \in \Omega_1 \,, \; x \in \Omega_2 \,, \end{array} \tag{10.2.1}$$

onde Ω_1 e Ω_2 são subconjuntos arbitrários de $I\!R^n$. Suponhamos, como antes, que (10.2.1) admite minimizador global.

O princípio é a utilização de uma função contínua que se anula no conjunto a ser penalizado e é positiva fora dele. Assim, se no problema (10.2.1) quisermos penalizar em relação ao conjunto Ω_1, basta escolhermos

$P : I\!R^n \to I\!R$, $P \in C^0(I\!R^n)$ tal que

$$P(x) \quad \begin{cases} = & 0 \quad \text{se} \quad x \in \Omega_1 \\ > & 0 \quad \text{se} \quad x \notin \Omega_1 \; . \end{cases} \tag{10.2.2}$$

Com a introdução de um parâmetro $\rho \geq 0$, temos o seguinte problema penalizado associado a (10.2.1):

$$\begin{array}{ll} \text{Minimizar} & f(x) + \rho \, P(x) \\ \text{sujeita a} & x \in \Omega_2 \; . \end{array} \tag{10.2.3}$$

Quando ρ torna-se muito grande, a violação das restrições fica cada vez mais cara, de tal forma que as soluções dos problemas (10.2.3), para uma seqüência controlada de aumentos em ρ, produz uma seqüência cujos pontos de acumulação resolvem o problema original, conforme provaremos adiante.

Sistematizando as idéias acima em forma algorítmica, com a função de penalidade P obedecendo (10.2.2), temos:

Algoritmo 10.2.1 - Penalidade externa.

Dados $\rho_1 \geq 0$, $x_0 \in I\!R^n$, $k = 1$.

(1) Calcular $x_k \equiv x(\rho_k) \in I\!R^n$ como a solução de

$$\begin{array}{ll} \text{Minimizar} & f(x) + \rho_k \, P(x) \\ \text{sujeita a} & x \in \Omega_2 \; . \end{array} \tag{10.2.4}$$

(2) Escolher $\rho_{k+1} > \rho_k$, $k \leftarrow k+1$ e voltar para (1).

De maneira análoga ao que ocorre com o Algoritmo 10.1.1, na seqüência $\{x_k\}$ gerada pelo Algoritmo 10.2.1 os pontos são desvinculados, e é apenas aconselhável que x_{k-1} seja o ponto inicial para o algoritmo que resolve (10.2.4). O monitoramento dos parâmetros penalizadores é, em geral, feito da seguinte forma: $\rho_1 = 1$ e $\rho_k = 10\rho_{k-1}$. Da mesma forma que em penalidade interna, a estratégia "shortcut" pode ser usada, tomando ρ_1 muito grande (por exemplo, 10^{24}) e resolvendo um único problema do tipo (10.2.4). Infelizmente, isso nem sempre funciona.

Vamos apresentar alguns exemplos de funções de penalidade. Se o conjunto factível a ser penalizado é dado por:

$$\Omega_1 = \{x \in I\!R^n \mid h(x) = 0\} \; ,$$

onde $h : I\!R^n \to I\!R^m$, podemos tomar

$$P(x) = \sum_{i=1}^{m} h_i(x)^2 = \|h(x)\|_2^2 \; .$$

Se abrirmos mão da diferenciabilidade, podemos definir

$$P(x) = \sqrt{\sum_{i=1}^{m} h_i(x)^2} = \|h(x)\|_2 \ ,$$

ou ainda

$$P(x) = \sum_{i=1}^{m} |h_i(x)| = \|h(x)\|_1 \ .$$

Para

$$\Omega_1 = \{x \in I\!R^n \mid c(x) \geq 0\} \ ,$$

onde $c : I\!R^n \to I\!R^p$, temos

$$P(x) = \sum_{i=1}^{p} (\min\{0\, , \ c_i(x)\})^2 \ .$$

Agora, se

$$\Omega_1 = \{x \in I\!R^n \mid h(x) = 0\, , \ c(x) \geq 0\} \ ,$$

onde $h : I\!R^n \to I\!R^m$ e $c : I\!R^n \to I\!R^p$, a função P pode ser dada por:

$$P(x) = \sum_{i=1}^{m} h_i(x)^2 + \sum_{i=1}^{p} (\min\{0, c_i(x)\})^2 \ .$$

Quando

$$\Omega_1 = \{x \in I\!R^n \mid g(x) \leq 0\} \ ,$$

com $g : I\!R^n \to I\!R^p$, é usual a notação

$$g_i(x)_+ = \max\{0\, , \ g_i(x)\}\, , \ i = 1\, , \ \ldots\, , \ p$$

e então $g(x)_+$ é o vetor p-dimensional cuja i-ésima componente é $g_i(x)_+$. Assim, podemos considerar uma classe geral de funções de penalidade

$$P(x) = \gamma(g(x)_+) \tag{10.2.5}$$

onde $\gamma : I\!R^p \to I\!R$ é uma função contínua definida de forma a satisfazer (10.2.2). Por exemplo, $\gamma(y) = \frac{1}{2}\|y\|_2^2$ ou $\gamma(y) = y^T A y$, onde $A \in I\!R^{p \times p}$ é simétrica definida positiva.

Denotando a função objetivo do problema penalizado por

$$\mathcal{P}(x, \rho) = f(x) + \rho\, P(x) \ , \tag{10.2.6}$$

temos as seguintes propriedades:

Lema 10.2.2

Seja $\{x_k\}$ a seqüência gerada pelo Algoritmo 10.2.1.

Se x_k é a solução global de (10.2.4), então

$$\begin{aligned} \mathcal{P}(x_k, \rho_k) &\leq \mathcal{P}(x_{k+1}, \rho_{k+1}) && (10.2.7)\\ P(x_{k+1}) &\leq P(x_k) && (10.2.8)\\ f(x_k) &\leq f(x_{k+1}) \; . && (10.2.9) \end{aligned}$$

Prova: Como para todo k temos $0 \leq \rho_k \leq \rho_{k+1}$ e x_k é minimizador global de (10.2.4) temos:

$$\begin{aligned} \mathcal{P}(x_k, \rho_k) &= f(x_k) \; + \; \rho_k \, P(x_k) \\ &\leq f(x_{k+1}) \; + \; \rho_k \, P(x_{k+1}) \\ &\leq f(x_{k+1}) \; + \; \rho_{k+1} \, P(x_{k+1}) \\ &= \mathcal{P}(x_{k+1}, \rho_{k+1}) \; . \end{aligned}$$

Agora,

$$\mathcal{P}(x_k, \rho_k) \; = \; f(x_k) + \rho_k \, P(x_k) \; \leq \; f(x_{k+1}) + \rho_k \, P(x_{k+1}) \qquad (10.2.10)$$

e

$$\mathcal{P}(x_{k+1}, \rho_{k+1}) \; = \; f(x_{k+1}) + \rho_{k+1} \, P(x_{k+1}) \; \leq \; f(x_k) + \rho_{k+1} \, P(x_k) \; . \qquad (10.2.11)$$

Subtraindo (10.2.11) de (10.2.10) temos

$$(\rho_k - \rho_{k+1}) \; P(x_k) \leq (\rho_k - \rho_{k+1}) \; P(x_{k+1})$$

e como $\rho_k < \rho_{k+1}$, segue que $P(x_{k+1}) \leq P(x_k)$.

Finalmente, usando (10.2.8) temos

$$f(x_k) + \rho_k \, P(x_k) \leq f(x_{k+1}) + \rho_k \, P(x_{k+1}) \leq f(x_{k+1}) + \rho_k \, P(x_k)$$

ou seja, $f(x_k) \leq f(x_{k+1})$ e a prova está completa. **QED**

Temos ainda uma outra relação para as seqüências de valores das funções objetivo original e penalizada, de onde se deduz que, se $\{x_k\}$ não é solução de (10.2.1), necessariamente deve ser um ponto externo a Ω.

Lema 10.2.3

Se x_* é um minimizador global do problema (10.2.1), então, para $k = 0, 1, \ldots$ temos

$$f(x_k) \;\leq\; \mathcal{P}(x_k, \rho_k) \;\leq\; f(x_*)\;. \tag{10.2.12}$$

Como conseqüência, $x_k \in \Omega$ se, e somente se, é uma solução global de (10.2.1).

Prova: Como $\rho_k \geq 0$, $P(x) \geq 0$ para todo $x \in I\!R^n$ e x_k é minimizador global de (10.2.4) temos:

$$f(x_k) \;\leq\; f(x_k) + \rho_k\, P(x_k) \;\leq\; f(x_*) + \rho_k\, P(x_*) = f(x_*)\;.$$

QED

No que se segue, apresentamos o resultado clássico de convergência dos métodos de penalidade externa.

Teorema 10.2.4

Seja $\{x_k\}$ a seqüência de minimizadores globais de (10.2.4), gerada pelo Algoritmo 10.2.1 com $\rho_k \to \infty$. Então, todo ponto limite de $\{x_k\}$ é minimizador global do problema (10.2.1).

Prova: Seja K um subconjunto infinito de $\mathcal{N}$ tal que $\lim_{k \in K} x_k = \bar{x}$. Pela continuidade de f temos

$$\lim_{k \in K} f(x_k) \;=\; f(\bar{x})\;. \tag{10.2.13}$$

Seja f_* o valor ótimo associado ao problema (10.2.1), isto é,

$$f_* = \min\{f(x) \mid x \in \Omega_1\,,\; x \in \Omega_2\}.$$

Pelos Lemas 10.2.2 e 10.2.3, a seqüência $\{\mathcal{P}(x_k, \rho_k)\}$ é não-decrescente e limitada superiormente por f_*. Então,

$$\lim_{k \in K} \mathcal{P}(x_k, \rho_k) \;=\; p_* \;=\; f_*\;. \tag{10.2.14}$$

Subtraindo (10.2.13) de (10.2.14) temos:

$$\lim_{k \in K} \rho_k\, P(x_k) \;=\; p_* - f(\bar{x})\;. \tag{10.2.15}$$

Como $P(x_k) \geq 0$ e $\rho_k \to \infty$, por (10.2.15) segue que

$$\lim_{k \in K} P(x_k) \;=\; 0\;.$$

Pela continuidade de P, $P(\bar{x}) = 0$, ou seja, $\bar{x} \in \Omega_1$. Para provarmos que $\bar{x}$ é ótimo, basta notarmos que pelo Lema 10.2.3, $f(x_k) \leq f_*$ e então

$$f(\bar{x}) \ = \ \lim_{k \in K} f(x_k) \ \leq \ f_* \ ,$$

o que completa a prova. **QED**

Vamos nos concentrar agora na função de penalidade externa mais popular, que consiste em elevar ao quadrado cada restrição violada. Para fixar idéias, pensaremos apenas na minimização com restrições de igualdade:

$$\text{Minimizar } f(x) \text{ sujeita a } h(x) = 0,$$

onde $h : I\!R^n \to I\!R^m$ e todas as funções tem derivadas contínuas. A função de penalidade será

$$P(x) = \frac{1}{2}\|h(x)\|_2^2.$$

Portanto, a condição necessária de otimalidade em x_k é

$$\nabla f(x_k) + h'(x_k)^T \rho_k h(x_k) = 0.$$

Logo, o vetor $\rho_k h(x_k)$ desempenha, em relação a x_k, o mesmo papel que os multiplicadores de Lagrange na solução têm em relação a x_*. Essa propriedade, que provaremos rigorosamente a seguir, autoriza o uso de $\rho_k h(x_k)$ como estimador dos multiplicadores, o que, como se verá na próxima seção, tem sua utilidade.

Teorema 10.2.5

Suponhamos que o Algoritmo 10.2.1 seja aplicado ao problema (10.2.1) com $\Omega_1 = \{x \in I\!R^n \mid h(x) = 0\}$, $h : I\!R^n \to I\!R^m$, $h \in C^1$, $\Omega_2 \equiv I\!R^n$ com a função de penalidade $P(x) = \frac{1}{2}\|h(x)\|_2^2$. Correspondendo à seqüência $\{x_k\}$ gerada por este algoritmo, definimos $\lambda_k = \rho_k h(x_k)$. Se $x_k \to x_*$, onde x_* é solução global de (10.2.1) e ponto regular, então $\lambda_k \to \lambda_*$, onde λ_* é o vetor dos multiplicadores de Lagrange associado a x_*.

Prova: O subproblema (10.2.4), sob as hipóteses acima, converte-se no seguinte problema irrestrito:

$$\text{Minimizar } \ f(x) + \rho_k \frac{1}{2}\|h(x)\|_2^2 \ . \qquad (10.2.16)$$

Portanto, anulando o gradiente, temos:

$$\nabla f(x_k) + h'(x_k)^T \lambda_k \ = \ 0 \ . \qquad (10.2.17)$$

Como x_* é solução regular de (10.2.1), existe um único $\lambda_* \in I\!\!R^m$ tal que

$$\nabla f(x_*) + h'(x_*)^T \lambda_* \;=\; 0 \,. \tag{10.2.18}$$

Ou seja,

$$\lambda_* \;=\; -(h'(x_*)^T)^\dagger \nabla f(x_*) \,, \tag{10.2.19}$$

onde $(h'(x_*)^T)^\dagger \;=\; (h'(x_*)h'(x_*)^T)^{-1} h'(x_*)$. Logo, como $h \in C^1$, para k suficientemente grande, $h'(x_k)$ tem posto m e, por (10.2.17), segue que

$$\rho_k h(x_k) \;=\; -(h'(x_k)^T)^\dagger \nabla f(x_k) \,. \tag{10.2.20}$$

Portanto, passando (10.2.20) ao limite quando $k \to \infty$, pela continuidade de $[h'(x)]^\dagger$ numa vizinhança de x_*, por (10.2.19) temos

$$\lim_{k\to\infty} \lambda_k \;=\; \lim_{k\to\infty} \rho_k h(x_k) \;=\; \lambda_* \,.$$

QED

Exercício 10.2: Generalizar o Teorema 10.2.5 para desigualdades.

Infelizmente, de maneira análoga aos método de barreiras, a dificuldade em se resolver os subproblemas cresce com o aumento do parâmetro penalizador ρ. Vejamos como isso acontece no exemplo (10.1.16), para o qual o problema penalizado pode ser dado por:

$$\text{Minimizar } \mathcal{P}(x,\rho) = (x_1+1)^2 + (x_2-1)^2 + \rho\, P(x_1,x_2) \,, \tag{10.2.21}$$

onde $P(x_1,x_2) = \begin{cases} 0 & \text{se} \quad x_1 \geq 0 \\ x_1^2 & \text{se} \quad x_1 < 0 \,. \end{cases}$

Como a função objetivo de (10.2.21) é convexa, basta determinar os pontos em que $\nabla_x \mathcal{P}(x,\rho) = \begin{pmatrix} 2(x_1+1) + 2\rho x_1 \\ 2(x_2-1) \end{pmatrix}$ se anula, obtendo $x_1 = \frac{-1}{1+\rho}$, $x_2 = 1$ e então $\lim_{\rho\to\infty} x_1 = 0$. Agora, $\nabla^2_{xx}\mathcal{P}(x,\rho) = \begin{pmatrix} 2+2\rho & 0 \\ 0 & 2 \end{pmatrix}$, ou seja, $\text{cond}(\nabla^2_{xx}\mathcal{P}(x,\rho)) \to \infty$ quando $\rho \to \infty$. Numericamente, o termo penalizador absorve o termo relativo à função objetivo original.

Vamos agora analisar a Hessiana do problema penalizado associado ao problema geral de minimização com restrições de igualdade:

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0 \,, \end{array} \tag{10.2.22}$$

onde $h : I\!R^n \to I\!R^m$ e $f, h \in C^2(I\!R^n)$. Se $P(x) = \frac{1}{2}h(x)^T h(x)$, temos $\mathcal{P}(x,\rho) = f(x) + \frac{\rho}{2}h(x)^T h(x) \equiv \Phi(x(\rho))$. Então, se $x \equiv x(\rho)$, temos

$$\nabla\Phi(x) = \nabla f(x) + \rho\, h'(x)^T h(x)$$

e

$$\nabla^2\Phi(x) = \nabla^2 f(x) + \rho\,[h'(x)^T h'(x) + \sum_{i=1}^{m} h_i(x)\nabla^2 h_i(x)]\ . \tag{10.2.23}$$

Se $x_* \in I\!R^n$ é uma solução regular de (10.2.22) e $\lambda_* \in I\!R^m$ é o multiplicador de Lagrange associado, pelo Teorema 10.2.5 sabemos que

$$\lim_{\rho\to\infty} \rho\, h(x(\rho)) = \lambda_*.$$

Então, para ρ suficientemente grande,

$$\nabla^2\Phi(x) \approx \nabla^2 f(x) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x) + \rho\, h'(x)^T h'(x)\ .$$

Embora $\nabla^2 f(x) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x)$ independa de ρ, o termo dominante $\rho\, h'(x)^T h'(x)$ tem posto deficiente, fazendo com que o número de condição de $\nabla^2\Phi(x)$ cresça ilimitadamente quando $\rho \to \infty$.

Vamos tentar contornar esta dificuldade, analisando o sistema não linear que representa as condições de otimalidade de problema penalizado com mais cuidado (ver [143]). Escrevendo esse problema como

$$\text{Minimizar}\ \ \Phi(x(\rho)) = f(x) + \frac{\rho}{2}\|h(x)\|_2^2\ , \tag{10.2.24}$$

temos que seus pontos estacionários são os que verificam

$$\nabla f(x) + \rho\, h'(x)^T h(x)\ =\ 0\ . \tag{10.2.25}$$

Fazendo a mudança de variáveis $y = \rho\, h(x)$, o sistema (10.2.25) se converte em

$$\begin{array}{rcl} \nabla f(x) + h'(x)^T y & = & 0 \\ h(x) - \frac{y}{\rho} & = & 0 \end{array} \tag{10.2.26}$$

cuja Jacobiana, membro da esquerda da seguinte expressão, verifica

$$\begin{pmatrix} \nabla^2 f(x) & h'(x)^T \\ h'(x) & -\frac{1}{\rho}I \end{pmatrix} \xrightarrow[\rho\to\infty]{} \begin{pmatrix} \nabla^2 f(x) & h'(x)^T \\ h'(x) & 0 \end{pmatrix}\ . \tag{10.2.27}$$

Assim, no limite, o Jacobiano (10.2.27) não é, necessariamente, mal-condicionado. A instabilidade proveniente do parâmetro penalizador ρ deixa de existir, e (10.2.27) só será mal-condicionado se $h'(x)$ tiver posto deficiente, o que é uma característica do problema, e não um defeito do processo de penalidade. Uma discussão do uso do sistema (10.2.26) do ponto de vista do raio de convergência do método de Newton pode ser encontrada em [143]. O próprio método de Newton aplicado a (10.2.24) pode ser estabilizado com um artifício similar ao usado aqui (ver [104]), mas a velocidade de convergência é maior quando usamos (10.2.26) como estratégia estabilizadora.

Infelizmente, com esta abordagem via sistemas não lineares perdemos a estrutura de minimização inerente ao problema (10.2.24). Com efeito, a matriz Jacobiana (10.2.27) é simétrica, mas não é semidefinida positiva. Assim, resolver o sistema (10.2.26) não é equivalente a um problema de minimização em (x, y). Embora exista uma função potencial

$$\mathcal{F}(x, y) = f(x) + h(x)^T y - \frac{1}{\rho} y^T y,$$

o problema primitivo não seria minimizá-la pois $\nabla^2_{yy}\mathcal{F}(x, y) = -\frac{1}{\rho} I < 0$. Temos, portanto, uma motivação para pensarmos numa abordagem um pouco diferente da penalidade externa, que será tratada na próxima seção.

Para finalizarmos a análise dos métodos de penalidade externa, vamos considerar as chamadas *funções de penalidade exatas*, em que a solução do problema penalizado é exatamente a solução do problema original para um valor finito do parâmetro penalizador. Assim, com estas funções não seria preciso resolver uma seqüência infinita de subproblemas. Infelizmente, a maioria da funções de penalidade exatas são não-diferenciáveis na solução. Um exemplo diferenciável, mas de interesse sobretudo teórico devido a sua complexidade, é a função de introduzida por Fletcher ([67], [68]) que, para o problema (10.2.22), é

$$\mathcal{P}(x, \rho) = f(x) - h(x)^T \lambda(x) + \frac{\rho}{2} h(x)^T h(x) \ ,$$

onde $\lambda(x) = (h'(x)^T)^\dagger \nabla f(x)$.

A função de penalidade exata não diferenciável mais conhecida é baseada na norma $\|\cdot\|_1$ e, para o problema (10.2.22), toma a forma

$$P(x) = \sum_{i=1}^{m} |h_i(x)| = \|h(x)\|_1,$$

portanto

$$\mathcal{P}(x, \rho) = f(x) + \rho \, \|h(x)\|_1 \ . \qquad (10.2.28)$$

A função (10.2.28) tem derivadas descontínuas em todos os pontos factíveis, e portanto, uma solução x_* para (10.2.22) é um ponto de descontinuidade do seu gradiente. Desta forma, os métodos de minimização irrestrita convencionais não se aplicam a (10.2.26) e são necessários algoritmos específicos que utilizam informações do problema original (10.2.22) (ver, por exemplo, [36] e [37]).

O resultado a seguir estabelece a convergência dos subproblemas penalizados associados a (10.2.22) para um parâmetro ρ finito quando se usa a função de penalidade exata baseada na norma $\| \cdot \|_1$.

Teorema 10.2.6

Se x_* é um ponto que satisfaz as condições suficientes de segunda ordem para minimizador local de (10.2.22) (capítulo 2) e $\lambda_* \in I\!R^m$ é o vetor dos multiplicadores de Lagrange correspondente, então, para $\rho > \max\{|(\lambda_*)_i| \, , \, i = 1 \, , \, \ldots \, , \, m\}$, x_* também é um minimizador local da função (10.2.28).

Prova: Ver Luenberger [129], p.389.

No resultado acima, vemos que o valor crucial para ρ a partir do qual o subproblema passa a admitir como minimizador a solução do problema original depende dos multiplicadores ótimos, sendo portanto desconhecido. Podem surgir dificuldades por uma escolha inadequada de ρ. Se ρ for muito pequeno, a função penalizada pode ser inferiormente ilimitada. Por outro lado, se ρ for muito grande, surgem os problemas de mal-condicionamento. Outras tentativas de amortecer o mal-condicionamento provocado por grandes parâmetros podem ser encontradas na literatura. Ver, por exemplo, [45] e [203].

10.3 Lagrangiano aumentado

Na seção anterior, vimos que o grande defeito dos métodos de penalidade externa é a necessidade de que o parâmetro penalizador ρ cresça ilimitadamente provocando instabilidade numérica. Ainda que se trabalhe com funções de penalidade exatas, estas são, freqüentemente, pouco práticas (não-diferenciáveis ou muito complicadas). Por outro lado, considerando-se o problema original de minimização com restrições de igualdade (10.2.22), se ao invés de resolvermos o problema penalizado (10.2.24), trabalharmos com o sistema não-linear aumentado (10.2.26), perdemos a estrutura inerente do problema pois a matriz Jacobiana não é semidefinida positiva. Os métodos de Lagrangiano aumentado têm por objetivo conciliar estes dois aspectos: contornar o mal-condicionamento proveniente de $\rho \to \infty$ e evitar a perda da estrutura de minimização. Foram sugeridos independentemente por Hestenes [118] e Powell [161].

Para fixar idéias, vamos considerar o problema de minimização com restrições de igualdade

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0 \ , \end{array} \tag{10.3.1}$$

onde $f : I\!\!R^n \to I\!\!R$, $h : I\!\!R^n \to I\!\!R^m$, $f, h \in C^1(I\!\!R^n)$. No entanto, as idéias de Lagrangiano aumentado se aplicam ao problema que também contém restrições de desigualdade. De fato, o caso mais importante é o definido pela forma padrão

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0 \, , \ \ l \le x \le u \ , \end{array}$$

usado por Conn, Gould e Toint ([43], [44]) no desenvolvimento do pacote LANCELOT para programação não-linear de grande porte.

As condições de Lagrange para (10.3.1) são dadas pelo bem-conhecido sistema não-linear com $n + m$ equações e $n + m$ variáveis:

$$\begin{array}{rcl} \nabla f(x) + h'(x)^T y & = & 0 \\ h(x) & = & 0 \ . \end{array} \tag{10.3.2}$$

Se $(x_*^T \, , \ y_*^T)^T$ satisfaz (10.3.2), então, definindo a função Lagrangiana da maneira usual,

$$\ell(x, y) = f(x) + h(x)^T y,$$

temos

$$\nabla \ell(x_* \, , \ y_*) = 0 \ .$$

Infelizmente, x_* pode não ser minimizador de $\ell(x \, , \ y_*)$, conforme ilustra o seguinte exemplo:

$$\begin{array}{ll} \text{Minimizar} & x^3 \\ \text{sujeita a} & x + 1 = 0 \ , \end{array}$$

onde $x_* = -1$, $y_* = -3$, $\ell(x \, , \ y_*) = x^3 - 3(x + 1)$, $\ell'(x \, , \ y_*) = 3x^2 - 3$, $\ell''(x \, , \ y_*) = 6x$ e portanto $\ell''(x_* \, , \ y_*) = -6 < 0$.

Agora, as condições necessárias de otimalidade de segunda ordem estabelecem que a Hessiana, em relação a x, da função Lagrangiana é semidefinida positiva no núcleo de $h'(x_*)$ (ver capítulo 2). Portanto, as direções de curvatura negativa de ℓ como função de x podem ser encontradas, preferencialmente, no subespaço ortogonal a esse núcleo, o espaço coluna $\mathcal{R}(h'(x_*)^T)$. Isto nos sugere que um subproblema irrestrito conveniente pode ser obtido se as características de estacionariedade de x_* forem mantidas, mas alterando-se a Hessiana $\nabla^2 \ell$ no espaço imagem de $h'(x_*)^T$. Mostraremos abaixo que esse é precisamente o efeito produzido acrescentando-se à função Lagrangiana o termo $\frac{\rho}{2}\|h(x)\|_2^2$, $\rho > 0$. Veremos que, nesse caso, existe ρ finito para o qual a função Lagrangiana aumentada é localmente convexa em torno de $(x_*^T, y_*^T)^T$. Antes vamos precisar do seguinte lema:

Lema 10.3.1

Seja $G = G^T \in \mathbb{R}^{n \times n}$ *tal que* $z^T G z > 0$ *para todo* $z \in \mathcal{N}(A)$, $z \neq 0$, $A \in \mathbb{R}^{m \times n}$.

Existe $\bar{\lambda} \geq 0$ *tal que* $G + \lambda A^T A > 0$ *para todo* $\lambda \geq \bar{\lambda}$.

Prova: Suponhamos que, para todo $k \in \mathcal{N}$, exista $x_k \in \mathbb{R}^n$, $\|x_k\| = 1$, tal que

$$x_k^T (G + k A^T A) x_k \leq 0 . \tag{10.3.3}$$

Pela compacidade dos x_k's, existe K subconjunto infinito de $\mathcal{N}$ tal que $\lim_{k \in K} x_k = \bar{x}$. Como $x_k A^T A x_k \geq 0$ para todo k, por (10.3.3) segue que $\bar{x}^T A^T A \bar{x} = 0$, ou seja, $\bar{x} \in \mathcal{N}(A)$. Então, por (10.3.3), $\bar{x}^T G \bar{x} \leq 0$, com $\bar{x} \in \mathcal{N}(A)$, o que é uma contradição. **QED**

Agora mostraremos que é suficiente um valor finito de ρ para transformar x_* num minimizador local estrito do Lagrangiano, em relação à variável x.

Teorema 10.3.2

Se x_* *satisfaz as condições suficientes de segunda ordem para o problema (10.3.1) (ver capítulo 2) e* $y_* \in \mathbb{R}^m$ *é o vetor dos multiplicadores correspondente, então existe* $\bar{\rho} \geq 0$ *tal que a função*

$$\bar{\ell}(x) = f(x) + y_*^T h(x) + \frac{\rho}{2}\|h(x)\|_2^2 \tag{10.3.4}$$

tem um minimizador local estrito em x_* *para todo* $\rho \geq \bar{\rho}$.

Prova: Temos que $\nabla \bar{\ell}(x) = \nabla f(x) + h'(x)^T y_* + \rho\, h'(x)^T h(x)$. Portanto,

$\nabla \bar{\ell}(x_*) = 0$, ou seja, x_* também é ponto estacionário de (10.3.4). Agora,

$$\nabla^2 \bar{\ell}(x) = \nabla^2 f(x) + \sum_{i=1}^{m} y_i^* \nabla^2 h_i(x) + \rho \left(h'(x)^T h'(x) + \sum_{i=1}^{m} h_i(x) \nabla^2 h_i(x) \right) .$$

Logo, $\nabla^2 \bar{\ell}(x_*) = \nabla^2 \ell(x_*) + \rho\, h'(x_*)^T h'(x_*)$, e o resultado desejado segue pelo Lema 10.3.1. **QED**

O Teorema 10.3.2 é animador no seguinte sentido. Se os multiplicadores de Lagrange na solução nos fossem dados de presente, bastaria um valor finito de ρ para transformar nosso problema original em um problema irrestrito. Infelizmente, não sabemos, a priori, qual seria esse valor finito (pelo qual corremos o risco, de instabilidade por superestimá-lo ou de funções não-limitadas por subestimá-lo) e, muito menos, qual é o vetor de multiplicadores de Lagrange. No entanto, o resultado sugere que, se em vez do vetor verdadeiro de multiplicadores, tivermos uma estimativa, os valores de ρ necessários para uma boa aproximação da solução não precisariam ser astronômicos. Para elaborar melhor este ponto de vista, observemos que o problema (10.3.1) é equivalente a

$$\begin{array}{ll} \text{Minimizar} & f(x) + y^T h(x) \\ \text{sujeita a} & h(x) = 0 \, , \end{array} \tag{10.3.5}$$

para qualquer $y \in I\!R^m$. (Podemos ler, se quisermos, "para qualquer estimador dos multiplicadores de Lagrange y".)

Aplicando penalidade quadrática a (10.3.5), temos

$$\text{Minimizar } \; f(x) + y^T h(x) + \frac{\rho}{2} h(x)^T h(x) \; , \tag{10.3.6}$$

que, para cada $y \in I\!R^m$ é um problema diferente.

Quando resolvemos (10.3.6), obtemos

$$\nabla f(x) + h'(x)^T y + \rho\, h'(x)^T h(x) = 0$$

ou

$$\nabla f(x) + h'(x)^T (y + \rho\, h(x)) = 0 \, .$$

Por comparação direta com (10.3.2) e, também, amparados pelo Teorema 10.2.5, deduzimos que $y + \rho\, h(x)$ pode ser uma estimativa razoável para y_*. Isto sugere o seguinte algoritmo:

Algoritmo 10.3.3 - Lagrangiano aumentado.

Dados $x_0 \in I\!R^n$, $\rho_1 > 0$, $y_1 \in I\!R^m$, $k = 1$.

(1) Minimizar $f(x) + y_k^T h(x) + \frac{\rho_k}{2}\|h(x)\|_2^2$,
tomando x_{k-1} como ponto inicial e obtendo x_k.

(2) Se $\|h(x_k)\| > 0.1\|h(x_{k-1})\|$ então $\rho_k \leftarrow 10\,\rho_k$.

(3) Reestimar $y_{k+1} = y_k + \rho_k\, h(x_k)$, $\rho_{k+1} = \rho_k$, $k \leftarrow k+1$ e voltar para (1).

Em cada passo do método é garantido, pelo processo de minimização, que $\nabla f(x_k) + h'(x_k)^T(y_k + \rho_k\, h(x_k)) = 0$. No entanto, a condição $h(x_k) = 0$ pode estar sendo muito "mal-satisfeita". Por isso, no Passo 2, é incrementado o parâmetro de penalidade, depois de um monitoramento de $h(x)$. Como rascunhamos numa seção anterior, o método de penalidade pode ser interpretado como uma maneira de acompanhar a homotopia

$$f(x(\rho)) + \frac{\rho}{2}\|h(x)\|_2^2 = \text{ mínimo,}$$

que desenha uma curva $\{x(\rho)\}$ em $I\!R^n$, culminando na solução do problema original quando $\rho = \infty$. Pela equivalência (10.3.6), para cada $y \in I\!R^m$, temos uma curva homotópica diferente, dada por

$$f(x(\rho)) + h'(x)^T y + \frac{\rho}{2}\|h(x)\|_2^2 = \text{ mínimo,}$$

que, também, "termina" em x_* quando $\rho = \infty$. Portanto, o método de Lagrangiano aumentado pode ser interpretado como uma maneira de saltar entre diferentes homotopias. A diferença entre uma e outra está em que, quanto mais próximo estiver y do vetor de multiplicadores de Lagrange correto, menor será o valor de ρ necessário para aproximar x_* com uma precisão dada.

Na prática, os subproblemas que conduzem às iterações x_k raramente podem ser resolvidos exatamente. Portanto, x_k deve ser interpretado, na maioria dos casos de aplicação prática do Algoritmo 10.3.3, como um minimizador aproximado. Assim, algoritmos computacionais baseados no Lagrangiano aumentado incorporam critérios de parada explícitos para os subproblemas (10.3.6). Quando x_k é apenas uma aproximação do minimizador do subproblema, a estimativa $y_k + \rho_k h(x_k)$ para os multiplicadores é mais difícil de justificar. De fato, outras estimativas mais robustas podem ser implementadas (ver Exercício 10.4) e a eficiência dos métodos está bastante ligada à qualidade de tais estimadores.

Exercício 10.4: Interpretar geometricamente o método de Lagrangiano aumentado do Algoritmo 10.3.3. Através desta interpretação, sugerir estimativas mais sofisticadas para os multiplicadores.

Exercício 10.5: Usando apenas argumentos de penalidade, provar a convergência do Algoritmo 10.3.3.

Exercício 10.6: Mostrar que a atualização $y_{k+1} = y_k + \rho_k \, h(x_k)$ corresponde ao método de máxima subida (gradiente) aplicado ao problema dual:

$$\text{Maximizar } \; \Phi(y) = f(x) + h(x)^T y + \frac{1}{2}\|h(x)\|_2^2 \, .$$

Exercício 10.7: Sugerir e interpretar a estimativa de "quadrados mínimos" para os multiplicadores quando o subproblema do passo (1) do Algoritmo 10.3.3 é resolvido aproximadamente.

Exercício 10.8: Desenvolver um método de Lagrangiano aumentado para o problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0 \, , \; c(x) \leq 0 \, , \end{array}$$

onde $f : I\!\!R^n \to I\!\!R$, $h : I\!\!R^n \to I\!\!R^m$, $c : I\!\!R^n \to I\!\!R^m$.

Exercício 10.9: Desenvolver um método de Lagrangiano aumentado para

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0 \, , \; l \leq x \leq u \, , \end{array}$$

onde os subproblemas são

$$\begin{array}{ll} \text{Minimizar} & f(x) + h(x)^T y + \frac{\rho}{2}\|h(x)\|_2^2 \\ \text{sujeita a} & l \leq x \leq u \, . \end{array}$$

Esta é a abordagem do pacote LANCELOT ([43, 44]).

Exercício 10.10: Desenvolver e discutir um método de Lagrangiano aumentado para

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0 \, , \; Ax = b \, , \; l \leq x \leq u \, , \end{array}$$

onde os subproblemas tenham a forma

$$\begin{array}{ll} \text{Minimizar} & f(x) + h(x)^T y + \frac{\rho}{2}\|h(x)\|_2^2 \\ \text{sujeita a} & Ax = b \, , \; l \leq x \leq u \, . \end{array}$$

Exercício 10.11: Discutir diferentes formas de aplicar Lagrangiano aumentado a programação linear e a programação quadrática.

Chapter 11

Gradiente reduzido generalizado

Contrariamente aos métodos de penalidade, cujo princípio básico é evitar a manipulação das restrições, mediante sua inclusão na função objetivo, os métodos analisados neste capítulo optam por conservar a factibilidade, lidando diretamente com as restrições "como elas são". A idéia fundamental é enxergar o problema original, pelo menos localmente, como um problema irrestrito num espaço de dimensão menor.

Wolfe [200] propôs o método de *gradiente reduzido*, para problemas de minimização com restrições lineares. Este método foi estendido por Abadie e Carpentier [1] para o problema geral de programação não-linear, originando os métodos de *gradiente reduzido generalizado* (GRG). Abadie e Carpentier são também responsáveis pela primeira implementação computacional do método básico. Com a mesma filosofia dos métodos de restrições ativas para problemas com restrições lineares, os métodos do tipo GRG buscam diminuir o valor da função objetivo mantendo factibilidade dos iterandos. A idéia básica é que um conjunto de restrições de igualdade não lineares é um sistema de equações onde, de maneira implícita, é possível colocar algumas variáveis em função de outras. Assim, minimizar com esse conjunto de restrições passa a ser um problema irrestrito cujas variáveis são, justamente, as variáveis selecionadas como independentes. Quando há restrições de desigualdade procedimentos adequados para mudar de face devem ser introduzidos.

Os métodos de tipo GRG têm analogia computacional com o método Simplex para programação linear. Usando técnicas de fatoração de matrizes e de manipulação de esparsidade similares às usadas no Simplex, foram desenvolvidos programas GRG extremamente eficientes do ponto de vista prático e, inclusive, com valor comercial. Ver, por exemplo, [128]. Este é um caso onde o alto investimento realizado nos aspectos de implementação compensa a relativa falta de desafios teóricos do método.

11.1 Restrições de igualdade

Analisaremos os métodos do tipo GRG aplicados ao seguinte problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0\,, \end{array} \tag{11.1.1}$$

onde $f : I\!R^n \to I\!R$, $h : I\!R^n \to I\!R^m$, f, $h \in C^1(I\!R^n)$, $m \leq n$.
Seja $\bar{x}$ um ponto factível e regular para o problema (11.1.1). Logo $h'(\bar{x})$ tem posto completo m. Assim, podemos considerar uma partição em m componentes *dependentes* ou *básicas* e $n-m$ componentes *independentes* ou *não-básicas*. As componentes básicas correspondem a uma sub-matriz não singular de $h'(\bar{x})$. Sem perda de generalidade, vamos supor que as primeiras m colunas de $h'(\bar{x})$ são linearmentes independentes. Então, podemos escrever $h'(\bar{x}) = (B\ N)$, com $B \in I\!R^{m\times m}$, B não-singular, $N \in I\!R^{m\times(n-m)}$ e $\bar{x} = (\bar{x}_B^T\ \bar{x}_N^T)^T$. Portanto, $h(\bar{x}) = h(\bar{x}_B, \bar{x}_N) = 0$ e, localmente, vale o Teorema da Função Implícita: existem vizinhanças $V_1 \subset I\!R^{n-m}$ e $V_2 \subset I\!R^m$ de $\bar{x}_N$ e $\bar{x}_B$ respectivamente, e uma função $\varphi : V_1 \to V_2$ tais que $\varphi \in C^1(V_1)$, $\varphi(\bar{x}_N) = \bar{x}_B$, $h(\varphi(x_N), x_N) = 0$ para todo $x_N \in V_1$, e

$$\varphi'(x_N) = -\left[\frac{\partial h}{\partial x_B}(x_B, x_N)\right]^{-1} \frac{\partial h}{\partial x_N}(x_B, x_N)$$

para todo $x_N \in V_1$.

Desta forma, se nos restringíssemos aos pares (x_B, x_N) para os quais o sistema $h(x_B, x_N) = 0$ é equivalente a $x_B = \varphi(x_N)$ (o que inclui os pares (x_B, x_N) tais que $x_N \in V_1$ e $x_B = \varphi(x_N)$) o problema (11.1.1) seria equivalente a

$$\begin{array}{ll} \text{Minimizar} & \Phi(x_N) \equiv f(\varphi(x_N), x_N) \\ \text{sujeita a} & x_N \in I\!R^{n-m}\,. \end{array} \tag{11.1.2}$$

Com isto estamos simplesmente formalizando o procedimento mais óbvio para minimizar funções com restrições de igualdade: colocar algumas variáveis em função das outras e substituir na função objetivo. O problema é que, na maioria dos casos, não conhecemos a forma explícita da função φ.

Usando a regra da cadeia, podemos calcular $\nabla\Phi$. Assim:

$$\Phi'(x_N) = \frac{\partial f}{\partial x_B}(x_B, x_N)\varphi'(x_N) + \frac{\partial f}{\partial x_N}(x_B, x_N)$$

para todo $x_N \in V_1$. Em particular, para $x_N = \bar{x}_N$,

$$\Phi'(\bar{x}_N) = \frac{\partial f}{\partial x_B}(\bar{x}_B, \bar{x}_N)(-B^{-1}N) + \frac{\partial f}{\partial x_N}(\bar{x}_B, \bar{x}_N).$$

Logo, transpondo a expressão acima:

$$\begin{aligned} \nabla\Phi(\bar{x}_N) &= -N^T B^{-T} \nabla_{x_B} f(\bar{x}) + \nabla_{x_N} f(\bar{x}) \\ &= (\, -N^T B^{-T} \quad I \,) \begin{pmatrix} \nabla_{x_B} f(\bar{x}) \\ \nabla_{x_N} f(\bar{x}) \end{pmatrix} \\ &= (\, -(B^{-1}N)^T \quad I \,)\, \nabla f(\bar{x}) \;. \end{aligned}$$

A expressão $\nabla\Phi(\bar{x})$ calculada acima é chamada o gradiente reduzido generalizado do problema (11.1.1), no ponto factível $\bar{x}$, relativo à partição $(B \quad N)$. As direções $d \in I\!R^{n-m}$ que formam um ângulo obtuso com $\nabla\Phi(\bar{x})$ são direções de descida para essa função. Se a vizinhança V_1 fosse igual a $I\!R^{n-m}$, a aplicação de um método de minimização sem restrições a (11.1.2) estaria plenamente justificada. Como freqüentemente $V_1 \neq I\!R^{n-m}$, algumas providências devem ser tomadas. Com base nos nossos conhecimentos de minimização irrestrita, estabelecemos o seguinte algoritmo conceitual para o método do tipo GRG aplicado ao problema (11.1.1):

Algoritmo 11.1.1 - GRG para igualdades com busca linear.
Sejam $\alpha \in (0,1), (\alpha \approx 10^{-4}\ \beta > 0,\ \theta \in (0,1)$ e $x_0 \in I\!R^n$ tal que $h(x_0) = 0$.
Dado $x_k \in I\!R^n$ tal que $h(x_k) = 0$, x_{k+1} é obtido da seguinte maneira:
Passo 1. Escolher uma partição $h'(\bar{x_k}) = (B_k \quad N_k)$, com $B_k \in I\!R^{m \times m}$ não singular. Então $x_k = \begin{pmatrix} x_k^B \\ x_k^N \end{pmatrix}$.
Calcular $\nabla\Phi(x_k^N) = (\, -(B_k^{-1}N_k)^T \quad I \,)\nabla f(x_k)$. Se $\nabla\Phi(x_k^N) = 0$, parar.
Passo 2. Escolher $d_k \in I\!R^{n-m}$ tal que

$$\|d_k\|_2 \geq \beta \|\nabla\Phi(x_k^N)\|_2 \tag{11.1.3}$$

e

$$\nabla\Phi(x_k^N)^T d_k \leq -\theta \|\nabla\Phi(x_k^N)\|_2 \, \|d_k\|_2 \;. \tag{11.1.4}$$

Passo 3. Começar o "backtracking" com $t = 1$.
Passo 4. Calcular $z = \varphi(\bar{x}_k^N + td_k) \in I\!R^m$, resolvendo o sistema (geralmente não linear), de $m \times m$,

$$h(z, \bar{x}_k^N + td_k) = 0. \tag{11.1.5}$$

Se não é possível resolver (11.1.5) (o que certamente acontecerá se esse sistema não tem solução), reduzir d_k (por exemplo, $d_k \leftarrow d_k/2$), e voltar ao Passo 3.
Passo 5. Se

$$f(z, x_k^N + td_k) \leq f(x_k^B, x_k^N) + \alpha t \nabla \Phi(x_k^N)^T d_k, \tag{11.1.6}$$

definir $x_{k+1}^N = x_k^N + td_k$, $x_{k+1}^B = z = \varphi(\bar{x}_k^N + td_k)$ e dar por terminada a iteração k.
Se (11.1.6) não se verifica, escolher um novo $t \in [0.1t, 0.9t]$ e retornar ao Passo 4.

No Passo 2 do Algoritmo 11.1.1, diferentes escolhas para d_k produzem os diferentes métodos do tipo GRG. Embora a direção de máxima descida $d_k = -\nabla \Phi(x_k^N)$ seja uma escolha possível, alternativas quase-Newton ou o próprio método de Newton nas coordenadas reduzidas poderiam ser consideradas. O cálculo de $\varphi(\bar{x}_k^N + td_k)$, no Passo 3, cuja existência numa vizinhança de x_k^N é assegurada pelo Teorema da Função Implícita, é o ponto crucial dos métodos. De fato, calcular $\varphi(x_k^N + td_k)$ corresponde a resolver o sistema (11.1.5). Para resolver esse sistema, usa-se qualquer método local para sistemas não lineares. (Para fixar idéias suponhamos que usamos o método de Newton.) Agora, (11.1.5) pode não ter solução, ou pode ser que, depois de um número razoável de iterações de Newton, não tenha sido possível chegar a uma solução com uma precisão adequada. Em ambos casos, o algoritmo reduz a direção d_k e recomeça o "backtracking". Teoricamente, este processo necessariamente termina, porque, mais tarde ou mais cedo, $x_k^N + td_k$ entra na vizinhança V_1. Porém, devido à impaciência em esperar um número suficientemente grande de iterações de Newton, ou a problemas de convergência desse método, é possível que o tamanho de td_k chegue a ser tão pequeno, que a condição (11.1.3) deixe de ser satisfeita. Nesse caso, o diagnóstico é que nossa escolha da partição $(B_k \;\; N_k)$ foi infeliz, no sentido da vizinhança V_1, onde a função φ existe, ser muito pequena. Provavelmente, neste caso, B_k é quase-singular. O recomendável, é tentar uma partição diferente, mas o sucesso também não é garantido.
Um problema de ordem prática que aparece na resolução do sistema (11.1.5) é a determinação de um bom ponto inicial z_0 para usar Newton, ou o algoritmo escolhido para resolver sistemas não lineares neste caso. Muitas vezes,

tomar $z_0 = x_k^B$ é suficientemente bom, mas não é difícil arquitetar uma estratégia melhor. A idéia é seguir a mesma filosofia do passo corretor no método preditor-corretor para equações diferenciais. Um ponto inicial sensato na resolução de (11.1.5) é o ponto "preditor" definido pela aproximação linear para $h(x) = 0$ em torno de x_k:

$$h'(x_k)(x - x_k) + h(x_k) = 0$$

ou seja,

$$\left(B_k \;\; N_k \right) \begin{pmatrix} z_0 - x_k^B \\ td_k \end{pmatrix} + h(x_k) = 0,$$

e então

$$z_0 = x_k^B - B_k^{-1}(N_k d_k + h(x_k)) \; .$$

O Algoritmo 11.1.1, aplicado ao caso $m = 0$ (sem restrições) é globalmente convergente, como vimos em um capítulo anterior. A garantia dessa convergência global é fornecida pelas condições (11.1.3) e (11.1.4). Se a mesma função φ estivesse bem definida para todo $x_N \in I\!R^{n-m}$ a mesma teoria de convergência se aplicaria no problema (11.1.1), já que, globalmente, o problema consistiria em minimizar, em $I\!R^{n-m}$, a (única) função Φ. Por isso, se justifica exigir, também neste caso, as condições (11.1.3) e (11.1.4). No entanto, a necessidade de "mudar de base" B_k em determinadas situações impede que a análise de convergência sem restrições possa ser estendida de maneira trivial ao caso geral. Uma complicação adicional é que, estritamente falando, como a solução de (11.1.5) é obtida por um método iterativo, devemos considerar que a avaliação de Φ está sujeita a um erro, cuja influência deveríamos contemplar. Uma discussão sobre convergência do método GRG pode ser encontrada em Sargent [181].

Cabe reforçar que, essencialmente, cada avaliação da função objetivo Φ do problema irrestrito (11.1.1) tem o custo da resolução do sistema não-linear (11.1.5). Vemos portanto que os métodos do tipo GRG são vantajosos quando o grau de não linearidade das restrições é pequeno. À medida que a não linearidade de h cresce, sua eficácia diminui. No entanto, GRG produz uma seqüência de pontos factíveis para o problema original o que é muito interessante para problemas onde é essencial conservar a factibilidade. Teorias abrangentes das quais podem ser deduzidas implementações promissoras de métodos do tipo GRG podem ser encontradas em [142] e [144].

Exercício 11.1: Simplificar o Algoritmo 11.1.1 para que resolva o problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & Ax = b \, , \end{array}$$

onde $A \in I\!R^{m \times n}$, $m < n$, posto$(A) = m$, $f \in C^2(R^n)$, sugerindo escolhas para d_k e completando todos os detalhes.

Exercício 11.2: Calcular, no Algoritmo 11.1.1, d_k usando Newton. Definir, cuidadosamente, o método "Newton-GRG com busca linear" para o problema (11.1.1).

11.2 GRG com desigualdades

O tratamento de restrições de desigualdade pelas estratégias do tipo GRG procede através da transformação do problema original à "forma padrão"

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0\,,\ l \leq x \leq u\,, \end{array} \tag{11.2.1}$$

onde $f : I\!R^n \to I\!R$, $h : I\!R^n \to I\!R^m$, $f,\ h \in C^1(I\!R^n)$. De fato, qualquer problema de minimização com restrições de igualdade e desigualdade pode ser levado à forma (11.2.1) pela introdução de variáveis de folga nas restrições do tipo $c(x) \geq 0$.
Neste capítulo introduziremos um método do tipo GRG para o problema (11.2.1). Nossa estratégia será similar à usada no caso de (11.1.1). Com efeito, um caso particular de (11.2.1) é quando $m = 0$. Nesse caso, o problema consiste em minimizar uma função com restrições de caixa. É natural, então, que o algoritmo do tipo GRG aplicado a (11.2.1) tenha como caso particular um bom algoritmo para minimizar em caixas, quando as restrições de igualdade não estão presentes. Como no caso (11.1.1), o método funcionará gerando uma seqüência de iterações factíveis ($h(x_k) = 0$, $l \leq x_k \leq u$). Em particular, um ponto inicial x_0 factível será necessário. O problema de encontrar esse ponto pode ser resolvido mediante a resolução de

$$\text{Minimizar } \|h(x)\|_2^2, \text{ sujeita a } l \leq x \leq u. \tag{11.2.2}$$

Este é um problema de minimizar em caixas, que, em princípio, poderia ser resolvido pelo mesmo método usado para (11.2.1).
Uma das dificuldades adicionais que aparecem devido às canalizações em (11.2.1) é que as variáveis declaradas dependentes (básicas) na iteração k não podem estar nos limites l_i ou u_i. A razão para essa restrição é que precisamos garantir que os pontos da forma $(\varphi(x_N), x_N)$ estejam dentro das canalizações para pequenas variações de x_N numa vizinhança (V_1) de x_k^N. Se uma variável básica estivesse num limite, qualquer movimento das variáveis

não básicas, por menor que fosse, poderia levar o ponto fora da caixa. É importante observar que essa é exatamente a condição de regularidade do conjunto $\Omega = \{x \in I\!R^n \ | h(x) = 0, \ \ l \leq x \leq u\}$. Com efeito, se as colunas de $h'(x)$ podem ser particionadas de maneira que (sem perda de generalidade) $h'(x) = (B \quad N)$, com B não singular e $l_i < [x]_i < u_i$ para todo $i = 1, \ldots, m$, então os gradientes das restrições ativas de Ω são linearmente independentes em x. Fica a cargo do leitor provar que, se x é um ponto regular de Ω, então pode ser encontrada uma partição com as condições desejadas.

Algoritmo 11.2.1 - GRG para o problema padrão.
Sejam $\alpha \in (0, 1)$ $(\alpha \approx 0.1)$, $M > 0$ (grande), $\Delta_{min} > 0$, e $x_0 \in I\!R^n$ tal que $h(x_0) = 0$, $l \leq x_0 \leq u$ e x_0 regular.
Dado $x_k \in I\!R^n$ tal que $h(x_k) = 0$, $l \leq x \leq u$, e x_k regular, vamos supor, sem perda de generalidade que $h'(x_k) = (B_k \quad N_k)$, com B_k não singular e $l_i < [x_k]_i < u_i$ para todo $i = 1, \ldots, m$. Nesse caso, x_{k+1} é obtido da seguinte maneira:
Passo 1. Escrevemos, como sempre, $x_k = \begin{pmatrix} x_k^B \\ x_k^N \end{pmatrix}$.
Calcular $\nabla\Phi(x_k^N) = (\, -(B_k^{-1}N_k)^T \quad I\,)\nabla f(x_k)$. Calcular H_k, uma aproximação de $\nabla^2\Phi(x_k^N)$ tal que $\|H_k\| \leq M$.
Passo 2. Iniciar o processo de encontrar uma região de confiança adequada escolhendo $\Delta \geq \Delta_{min}$.
Passo 3. Resolver, aproximadamente, o problema quadrático

$$\begin{array}{ll} \text{Minimizar} & \frac{1}{2}(w - x_k^N)^T H_k(w - x_k^N) + \nabla\Phi(x_k^N)^T(w - x_k^N) \\ \text{sujeita a} & l \leq x \leq u, \ \ \|w - x_k^N\|_\infty \leq \Delta. \end{array} \tag{11.2.3}$$

Se x_k^N é um ponto estacionário do problema (11.2.3), parar.
Passo 4. Calcular $z = \varphi(w) \in I\!R^m$, resolvendo o sistema (geralmente não linear), de $m \times m$

$$h(z, w) = 0. \tag{11.2.4}$$

Se não é possível resolver (11.2.4) (o que certamente acontecerá se esse sistema não tem solução), ou se a solução z encontrada está fora dos limites l e u, reduzir Δ (por exemplo, $\Delta \leftarrow \Delta/2$), e voltar ao Passo 2.
Passo 5. Se

$$f(z, w) \leq f(x_k^B, x_k^N) + \alpha[\frac{1}{2}(w - x_k^N)^T H_k(w - x_k^N) + \nabla\Phi(x_k^N)^T(w - x_k^N)] \tag{11.2.5}$$

definir $x_{k+1}^N = w$, $x_{k+1}^B = z$ e dar por terminada a iteração k.

Se (11.2.5) não se verifica, escolher um novo $\Delta \in [0.1\Delta, 0.9\|w - x_k^N\|_\infty]$ e retornar ao Passo 3.

Todas as observações feitas sobre o Algoritmo 11.1.1 são válidas, também, para este algoritmo. No Algoritmo 11.1.1 escolhemos, como método sem restrições subjacente, um algoritmo de buscas lineares. No Algoritmo 11.2.1 escolhemos um método de regiões de confiança com norma ∞ porque esse tipo de método se ajusta melhor ao formato de uma região em forma de caixa. A convergência global desse método, quando não aparecem as restrições $h(x) = 0$, dando um sentido preciso à resolução "aproximada" de (11.2.3), foi estudada num capítulo anterior deste livro. Naturalmente, também podíamos ter usado como algoritmo subjacente no caso do problema (11.1.1) um método de regiões de confiança. No entanto, as buscas lineares são mais tradicionais quando se fala de GRG aplicado a minimização com restrições de igualdade.

Exercício 11.2: Escrever um algoritmo de gradiente reduzido para o problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & Ax = b\,,\ x \geq 0\,, \end{array}$$

onde $A \in I\!R^{m \times n}$, $m < n$, posto$(A) = m$, $f \in C^2(R^n)$.

Exercício 11.3: Escrever um algoritmo de gradiente reduzido para o problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & Ax = b\,,\ l \leq x \leq u\,, \end{array}$$

onde $A \in I\!R^{m \times n}$, $m < n$, posto$(A) = m$, $f \in C^2(R^n)$. Estudar o caso em que $f(x) = c^T x$.

Exercício 11.4: Provar que, se x_k^N é um ponto estacionário de (11.2.3), então x_k é um ponto estacionário de (11.2.1).

11.3 Implementação computacional

Como comentamos na Seção 11.1, o funcionamento dos métodos do tipo GRG depende fortemente de sua implementação e a fama dos métodos se deve, provavelmente, ao aproveitamento da "experiência Simplex" para produzir bom software.

Embora sejam difíceis de ser implementados, os métodos GRG mereceram a atenção de equipes muito competentes. Atualmente, existem programas desenvolvidos com eficiência comprovada. Por exemplo, o pacote GRG2 [128], desenvolvido em FORTRAN, usa uma implementação robusta de BFGS para obter a direção d_k. Este programa também possui uma opção para trabalhar com métodos de gradientes conjugados com memória limitada, o que permite lidar com milhares de variáveis, mas a matriz Jacobiana das restrições é armazenada de forma densa, o que limita a resolução a problemas com, no máximo, duzentas restrições ativas.
Com o objetivo de complementar a atuação do pacote GRG2 para problemas de grande porte, foi desenvolvido recentemente o pacote LSGRG2 [186], utilizando estruturas esparsas para armazenamento e fatorações esparsas para as bases B_k. Lasdon [127] apresenta um resumo dos avanços relativamente recentes no uso de métodos do tipo GRG, bem como uma comparação dos desempenhos de GRG, programação linear sequencial e programação quadrática sequencial.
Finalmente, deve ser mencionado que a estratégia GRG tem, historicamente, despertado o interesse de pesquisadores devotados a resolver problemas de controle discreto (ou de controle contínuo por meio de discretização). Nesses casos, as variáveis do problema (11.1.1) são as variáveis de controle junto com as variáveis de estado do sistema, as restrições $h(x) = 0$ são as equações de estado e, talvez, restrições adicionais, e a caixa $l \leq x \leq u$ representa cotas nas variáveis, tanto de estado como de controle. O atrativo do GRG para esse tipo de problemas radica em que, por um lado, é essencial neles a manutenção da factibilidade, pois uma solução parcial que não satisfaça uma equação de estado carece totalmente de sentido. Por outro lado, as variáveis de controle são variáveis independentes naturais do problema o que, provavelmente, garante em muitos casos a necessidade de um número pequeno de mudanças de bases ao longo de todo o processo. Existem implementações especiais de métodos de tipo GRG para a estrutura particular de determinados problemas de controle. Um exemplo de método desse tipo, e bibliografia mais ampla, podem ser encontrados em [77].

Chapter 12

Programação quadrática seqüencial

Um dos procedimentos fundamentais do cálculo numérico consiste na resolução de problemas relativamente complicados através de uma seqüência de problemas mais simples. Dada uma aproximação x_k da solução do problema difícil, define-se um problema "fácil" que é parecido com o problema original, pelo menos numa região próxima de x_k. Freqüentemente, a solução do problema fácil é uma melhor aproximação da solução do problema colocado originariamente. A versão mais simples dessa idéia é o método de Newton para achar zeros de funções. Os métodos de programação quadrática seqüencial são as generalizações do método de Newton para o problema geral de otimização. 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 equações ou inequações lineares. Dessa maneira, o *subproblema* a ser resolvido em cada iteração k é um problema de programação quadrática que, em comparação ao problema original, pode ser considerado *simples*. Assim como acontece com o método de Newton para zeros de funções, a versão mais ingênua da idéia não tem boas propriedades de convergência global, e modificações são necessárias para melhorar essas propriedades. Neste capítulo procuramos combinar uma visão didática dos princípios da programação quadrática seqüencial com uma introdução a um método moderno, onde as principais dificuldades da idéia fundamental são contornadas.

12.1 Programação quadrática seqüencial "pura"

Ao longo deste capítulo vamos considerar o problema geral de otimização na forma padrão:

$$\text{Minimizar } f(x) \text{ sujeita a } h(x) = 0, \;\; l \leq x \leq u, \tag{12.1.1}$$

onde $f : I\!R^n \to I\!R$, $h : I\!R^n \to I\!R^m$. Os vetores l e u podem ter componentes $-\infty$ ou $+\infty$ respectivamente. Nesses casos, o símbolo $\leq$ deve ser interpretado como $<$. Sabemos que, de fato, qualquer problema de otimização com igualdades e desigualdades pode ser levado à forma (12.1.1) através da introdução de variáveis de folga. Por exemplo, toda restrição do tipo

$$c(x) \geq 0, \tag{12.1.2}$$

pode ser transformada em

$$c(x) - z = 0, \;\; z \geq 0.$$

Dessa maneira, uma variável (z) é acrescentada ao problema para cada restrição do tipo (12.1.2), o que pode ser uma desvantagem. Por outro lado, o tratamento de restrições na forma padrão é geralmente mais simples e muitos algoritmos eficientes, com software bem desenvolvido, se baseiam na forma padrão.

Suponhamos que x_k é uma aproximação da solução de (12.1.1). Provavelmente conseguiremos uma aproximação melhor se, usando a informação disponível em x_k, transformarmos o problema (12.1.1) em um problema mais simples, e resolvermos este último.

Se, lembrando o paradigma newtoniano, substituirmos a função objetivo f por sua melhor aproximação linear numa vizinhança de x_k, e fizermos a mesma coisa com as restrições, o "problema simples" associado a (12.1.1) será

$$\begin{array}{ll} \text{Minimizar} & f(x_k) + \nabla f(x_k)^T (x - x_k) \\ \text{sujeita a} & h'(x_k)(x - x_k) + h(x_k) = 0, \;\; l \leq x \leq u. \end{array} \tag{12.1.3}$$

As substituições efetuadas para chegar a (12.1.3) se baseiam no fato de que, para funções f e h diferenciáveis, temos $f(x) \approx f(x_k) + \nabla f(x_k)(x - x_k)$ e $h(x) \approx h(x_k) + h'(x_k)(x - x_k)$. Agora, (12.1.3) é um problema de programação linear, portanto, métodos baseados nessa aproximação podem ser chamados de "programação linear seqüencial". Um pouco mais de generalidade é obtida se, em vez de aproximar f por uma função linear, o fazemos por uma aproximação quadrática:

$$f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k).$$

Neste caso, em vez do problema simples (12.1.3), teremos que resolver, em cada iteração k, o seguinte subproblema:

$$\begin{array}{ll} \text{Minimizar} & f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k) \\ \text{sujeita a} & h'(x_k)(x - x_k) + h(x_k) = 0, \;\; l \leq x \leq u. \end{array} \tag{12.1.4}$$

O subproblema (12.1.4) é um problema de programação quadrática. Ele é simples em termos relativos, ou seja, em comparação com o problema original (12.1.1). (Via de regra, sua resolução eficiente pode demandar técnicas bastante sofisticadas.) Quando usamos a aproximação quadrática de f neste contexto, a primeira tentação é definir $B_k = \nabla^2 f(x_k)$. Veremos mais adiante que, contrariamente a intuição, esta não é a escolha mais adequada de B_k. Pelo momento, no entanto, não faremos nenhuma suposição sobre esta matriz.

Uma das dificuldades mais sérias para a implementação de algoritmos práticos baseados no subproblema (12.1.4) é que este problema pode não ter solução. Isto acontece em duas situações:

(a) Quando a região factível de (12.1.4) é vazia. Com efeito, a variedade afim $h'(x_k)(x - x_k) + h(x_k) = 0$ pode não ter intersecção com a caixa $l \leq x \leq u$. Também, quando o posto de $h'(x_k)$ é menor que m e $h(x_k)$ não está no espaço coluna de $h'(x_k)$, a própria variedade afim é vazia.

(b) Quando a função objetivo de (12.1.4) não é limitada inferiormente na região factível. Neste caso, pela continuidade da função quadrática, a região factível não pode ser compacta, em particular, alguma componente de l_i ou u_i deve ser infinita.

Um problema menor é que, mesmo quando o subproblema (12.1.4) tem solução, ela pode não ser única.

Exercício 12.1: Provar que quando a região factível é não vazia, o subproblema (12.1.4) tem solução. Provar que a solução é única quando a matriz B_k é definida positiva. Exibir exemplos onde a solução é única mesmo sem essa hipótese. Considerar o caso $l_i = -\infty$, $u_i = \infty$ para todo i. Analisar, nesse caso, em que situações o problema tem solução e em que situações a solução é única. Exibir exemplos.

Exercício 12.2: Analisar o método iterativo baseado no subproblema (12.1.4) nos seguintes casos particulares: (a) quando $m = n$ e $f(x)$ é constante; (b) quando $l_i = -\infty$, $u_i = \infty$ para todo i; (c) quando $m = 0$ (não há restrições $h(x) = 0$; (d) quando (c) e (d) acontecem juntos. Em cada caso, observar que o método resultante é conhecido. Identificar o método e estabelecer propriedades em cada caso.

12.2 Forçando solubilidade do subproblema

Na seção anterior vimos que a região factível de (12.1.4) pode ser vazia, ou seja, é possível que não exista nenhuma solução do sistema linear

$$h'(x_k)(x - x_k) + h(x_k) = 0$$

que pertença à caixa $l \leq x \leq u$. Existem várias maneiras de contornar esta dificuldade. Em todas elas, o problema deve ser modificado de maneira tal que, por um lado, o novo subproblema tenha solução e, por outro lado, que a nova solução coincida com a solução do subproblema (12.1.4) nos casos em que aquela existia. Ambos pré-requisitos são preenchidos da seguinte maneira. Primeiro, definimos o seguinte "subproblema prévio":

$$\begin{array}{ll} \text{Minimizar} & \|h'(x_k)(x - x_k) + h(x_k)\|_2^2 \\ \text{sujeita a} & l \leq x \leq u\ . \end{array} \tag{12.2.1}$$

O problema (12.2.1), que consiste em minimizar uma quadrática convexa numa caixa, sempre tem solução. (A prova disto será deixada como exercício para o leitor.) Chamemos x_k^{nor} a uma das soluções de (12.2.1). Portanto, o politopo definido pela intersecção da caixa $l \leq x \leq u$ com a variedade afim

$$h'(x_k)(x - x_k) = h'(x_k)(x_k^{nor} - x_k)$$

é não vazio. Claramente, no caso em que a região factível de (12.1.4) é não vazia, temos que $h'(x_k)(x_k^{nor} - x_k) + h(x_k) = 0$. É natural, em conseqüência, substituir o subproblema (12.1.4) pelo seguinte problema de programação quadrática:

$$\begin{array}{ll} \text{Minimizar} & f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k) \\ \text{sujeita a} & h'(x_k)(x - x_k) = h'(x_k)(x_k^{nor} - x_k), \\ & l \leq x \leq u. \end{array} \tag{12.2.2}$$

Pelo exposto, a região factível de (12.2.2) é não vazia. Persiste, porém, a possibilidade de que a função objetivo de (12.2.2) seja ilimitada inferiormente no seu conjunto de factibilidade. Portanto, para que exista solução do subproblema de programação quadrática, este precisa de uma modificação adicional.

A pista para a nova modificação vem da seguinte consideração: nosso objetivo final é resolver (12.1.1), e para isso nos baseamos em que *perto de* x_k, os subproblemas (12.1.4) ou (12.2.2) são *parecidos* com o problema de otimização original. Em conseqüência, mesmo que (12.2.2) tenha uma solução x, é provável que, se $\|x - x_k\|$ for muito grande, essa solução tenha pouca relação com boas aproximações para a solução de (12.1.1). Logo, é justificável, do ponto de vista dos nossos objetivos últimos, exigir uma limitação na distância entre a solução de (12.2.2) e a aproximação atual x_k. Expressaremos essa necessidade, acrescentando, em (12.2.2), a restrição adicional $\|x - x_k\|_\infty \leq \Delta$, onde $\Delta > 0$ (o "raio da região de confiança") será ajustado em cada iteração k. Assim, nosso subproblema de programação quadrática seria:

$$\begin{array}{ll} \text{Minimizar} & f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k) \\ \text{sujeita a} & h'(x_k)(x - x_k) = h'(x_k)(x_k^{nor} - x_k), \\ & l \leq x \leq u, \quad \|x - x_k\|_\infty \leq \Delta. \end{array} \tag{12.2.3}$$

Infelizmente, a imposição da restrição limitante $\|x - x_k\|_\infty \leq \Delta$ em (12.2.3) pode ser incompatível com a definição de x_k^{nor} em (12.2.1). De fato, com essa definição, poderia ser que o problema (12.2.3) fosse infactível. Portanto, se queremos a limitação de $\|x - x_k\|_\infty$ em (12.2.3), precisamos modificar a definição de x_k^{nor}. Para tanto, vamos redefinir x_k^{nor} como uma solução de

$$\begin{array}{ll} \text{Minimizar} & \|h'(x_k)(x - x_k) + h(x_k)\|_2^2 \\ \text{sujeita a} & l \leq x \leq u, \quad \|x - x_k\|_\infty \leq 0.8\Delta. \end{array} \tag{12.2.4}$$

A restrição $\|x - x_k\|_\infty \leq 0.8\Delta$ em (12.2.4) obriga a região factível do problema (12.2.3) a ser não vazia. Isto também seria conseguido se, em vez dessa restrição tivéssemos colocado $\|x - x_k\|_\infty \leq r\Delta$ para qualquer $r \in [0, 1]$. A escolha $r = 0.8$ parece satisfazer simultaneamente os requisitos de que $\|h'(x_k)(x - x_k) + h(x_k)\|_2^2$ seja suficientemente pequeno, e que a região factível de (12.2.3) seja suficientemente ampla para permitir um decréscimo de sua função objetivo.

Do ponto de vista da existência e limitação da solução do subproblema a escolha da norma $\| \cdot \|_\infty$ não tem nenhum papel. Essa escolha se justifica porque, com ela, os subproblemas (12.2.4) e (12.2.3) continuam sendo de programação quadrática, o que não aconteceria, por exemplo, se escolhêssemos a norma euclidiana para limitar a distância entre x e x_k.

Exercício 12.3: Provar que (12.2.1) e (12.2.4) sempre têm solução. Provar que, mesmo quando a solução não é única, o vetor $h(x_k^{nor})$ independe da solução escolhida x_k^{nor}.

Exercício 12.4: Analisar o par de subproblemas (12.2.4)–(12.2.3) nos seguintes casos: (a) todos os l_i são $-\infty$ e todos os u_i são $+\infty$; (b) não há restrições $h(x) = 0$; (c) a função $f(x)$ é constante; (d) as restrições $h(x) = 0$ são lineares.

Exercício 12.5: Analisar os subproblemas (12.2.4)–(12.2.3) substituindo $\| \cdot \|_\infty$ por $\| \cdot \|_2$. Considerar x_k^{nor} como uma função de Δ e desenhar uma trajetória típica $x_k^{nor}(\Delta)$ para $\Delta \in [0, \infty)$. Interpretar geometricamente.

Exercício 12.6: Estabelecer rigorosamente em que sentido a solução de (12.2.4)–(12.2.3) coincide com a solução de (12.1.4) quando este problema é solúvel.

Exercício 12.7: Refazer os argumentos das Seções 12.1 e 12.2 para o problema de otimização definido na forma

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) \leq 0, \end{array}$$

onde $h : I\!R^n \to I\!R^m$. Refazer, mais uma vez, os argumentos para considerar misturas de restrições de igualdade e desigualdade.

12.3 A função de mérito

A argumentação das seções 12.1 e 12.2 parece consolidar a seguinte forma para um algoritmo de programação quadrática seqüencial destinado a resolver o problema (12.1.1):

Algoritmo 12.3.1
Suponhamos que $x_0 \in I\!R^n$ $(l \leq x \leq u)$ é uma aproximação inicial da solução de (12.1.1). Se x_k $(k = 0, 1, 2, \ldots)$ é a aproximação obtida na $k-$ésima

iteração ($l \leq x_k \leq u$), $B_k \in I\!R^{n \times n}$ é uma matriz simétrica e $\Delta > 0$, então x_{k+1} é obtida da seguinte maneira:
Passo 1. Resolver (12.2.4) e (12.2.3).
Passo 2. Se $\bar{x}$, a solução obtida no Passo 1, é "suficientemente boa" em relação a x_k, então definir $x_{k+1} = \bar{x}$ e terminar a iteração. Caso contrário, diminuir Δ e retornar ao Passo 1.

A principal questão que o "Algoritmo" 12.3.1 deixa em aberto é: que significa "suficientemente boa"? Se não houvesse restrições do tipo $h(x) = 0$, o único critério para julgar se $\bar{x}$ é melhor que x_k seria o valor de $f(\bar{x})$ em relação ao valor de $f(x_k)$. Por outro lado, se a função objetivo de (12.1.1) fosse constante, o critério deveria estar baseado em alguma norma de $h(x)$. De um modo geral, nas iterações destinadas a resolver (12.1.1) existem dois objetivos a serem melhorados simultaneamente: a factibilidade (medida por $\|h(x)\|$) e a otimalidade (medida por $f(x)$). Claramente, se $f(\bar{x}) \ll f(x_k)$ e $\|h(\bar{x})\| \ll \|h(x_k)\|$ devemos decidir que $\bar{x}$ "é melhor" que x_k em relação ao objetivo de resolver (12.1.1). A situação não é clara quando

$$f(\bar{x}) < f(x_k) \text{ e } \|h(\bar{x})\| > \|h(x_k)\|$$

ou

$$f(\bar{x}) > f(x_k) \text{ e } \|h(\bar{x})\| < \|h(x_k)\|.$$

No primeiro caso nos perguntamos: será que o ganho em otimalidade compensa a perda de factibilidade? No segundo: o ganho em factibilidade compensa o aumento de f?
Uma função de mérito combina $f(x)$ e $h(x)$ de maneira a permitir possíveis respostas às perguntas acima. Elementos adicionais para a construção de uma função de mérito vêm de considerar as condições de otimalidade do problema (12.1.1). Definimos, como é habitual, o Lagrangiano, $\ell(x, \lambda)$ por

$$\ell(x, \lambda) = f(x) + h(x)^T \lambda \tag{12.3.1}$$

para todo $x \in I\!R^n, \lambda \in I\!R^m$. As condições necessárias de primeira ordem (Karush-Kuhn-Tucker) estabelecem que um minimizador local x junto com seu vetor de multiplicadores λ deve satisfazer:

$$[\nabla_x \ell(x, \lambda)]_i = 0, \;\; [\nabla_x \ell(x, \lambda)]_i \geq 0, \;\; [\nabla_x \ell(x, \lambda)]_i \leq 0 \tag{12.3.2}$$

se $l_i < [x]_i < u_i$, $[x]_i = l_i$ ou $[x]_i = u_i$ respectivamente. Além disso, a factibilidade da solução implica que

$$\nabla_\lambda \ell(x, \lambda) = h(x) = 0. \tag{12.3.3}$$

As condições (12.3.2) e (12.3.3) são satisfeitas se o par (x, λ) é um minimizador de $\ell(x, \lambda)$ para $l \leq x \leq u$.

Exercício 12.8: Estabelecer rigorosamente as condições nas quais valem (12.3.2) e (12.3.3).

As considerações acima parecem sugerir que $\ell(x, \lambda)$ definida em (12.3.1) seria uma função de mérito adequada, porém, envolvendo as duas variáveis, x e λ. No entanto, podemos observar que, se $h(x) \neq 0$, valores de $\ell(x, \lambda)$ muito grandes e negativos podem ser obtidos apenas variando λ, por exemplo, fazendo $\lambda = -\rho h(x)$ para ρ muito grande (embora, talvez, limitado). Isso significa que, se usássemos o Algoritmo 12.3.1 com um critério de aceitação baseado na função de mérito ℓ, a solução $\bar{x}$ de (12.2.4)-(12.2.3) sempre seria aceita se apenas tomássemos a providência de escolher de maneira oportuna, as novas estimativas dos multiplicadores.
Examinemos, pois, uma segunda possibilidade, que contempla a função ℓ, combinando-a com uma segunda função que se preocupa, fundamentalmente, com a factibilidade da iteração. Esta segunda função é, simplesmente,

$$\varphi(x) = \frac{1}{2}\|h(x)\|^2. \tag{12.3.4}$$

A "combinação" aludida acima é uma combinação convexa de ℓ e φ. Dado $\theta \in [0, 1]$, definimos

$$\Phi(x, \lambda, \theta) = \theta\ell(x, \lambda) + (1 - \theta)\varphi(x). \tag{12.3.5}$$

A confiança que depositamos em Φ como função de mérito se baseia no seguinte: "se for necessário" (o que será estabelecido precisamente mais adiante) θ será escolhido perto de 0, de maneira que φ será dominante na combinação (12.3.5). Assim as componentes de $h(x)$ serão obrigatoriamente empurradas para valores pequenos. Agora, para valores pequenos de $\|h(x)\|$, se a aproximação dos multiplicadores é mantida limitada, o efeito redutor devido a variação destes, de que falamos antes, será desprezível. Portanto, a diminuição do primeiro termo da combinação convexa Φ será devido à diminuição de f.
Essas considerações nos levam a especificar um pouco mais o Algoritmo 12.3.1, agora baseado na função de mérito Φ.

Algoritmo 12.3.2
Suponhamos que $L > 0$ (grande), $x_0 \in I\!\!R^n$ $(l \leq x \leq u)$ é uma aproximação inicial da solução de (12.1.1) e $\lambda_0 \in I\!\!R^m (\|\lambda_0\| \leq L)$ é uma aproximação

inicial dos multiplicadores de Lagrange. Se x_k, λ_k ($k = 0, 1, 2, \ldots$) são as aproximações obtidas na $k-$ésima iteração ($l \leq x_k \leq u$, $\|\lambda_k\| \leq L$), $B_k \in \mathbb{R}^{n \times n}$ é uma matriz simétrica e $\Delta > 0$, então x_{k+1} é obtida da seguinte maneira:

Passo 1. Resolver (12.2.4) e (12.2.3).

Passo 2. Escolher um valor adequado para $\theta \in [0, 1]$ e estimar novos multiplicadores $\bar{\lambda}$ ($\|\bar{\lambda}\| \leq L$).

Passo 3. Se $\bar{x}$, a solução obtida no Passo 1 é tal que

$$\Phi(\bar{x}, \bar{\lambda}, \theta) \ll \Phi(x_k, \lambda_k, \theta), \tag{12.3.6}$$

definir $x_{k+1} = \bar{x}$, $\lambda_{k+1} = \bar{\lambda}$ e terminar a iteração. Caso contrário, diminuir Δ e retornar ao Passo 1.

12.4 Decréscimo suficiente

No Algoritmo 12.3.2 ainda existem vários aspectos não definidos:

(a) A escolha dos "novos multiplicadores" $\bar{\lambda}$ no Passo 2.

(b) A determinação do parâmetro θ, no mesmo passo.

(c) O significado preciso da expressão "$\ll$" no Passo 3.

(d) A escolha do valor inicial Δ em cada iteração e a forma de diminuir Δ, quando isso é necessário.

(e) A escolha da matriz simétrica B_k.

A decisão sobre a escolha de B_k será adiada para uma seção posterior. O monitoramento do "raio de confiança" Δ não oferece grandes dificuldades conceituais. Nosso procedimento, neste caso, é o seguinte: estabelece-se a priori (independentemente do número da iteração k) um "raio de confiança mínimo inicial" Δ_{min}. O primeiro Δ testado ao começar a iteração k deve ser maior ou igual a Δ_{min}. Isso possibilita que, ao menos no começo, iterações suficientemente arrojadas sejam efetuadas, evitando passos excessivamente curtos. Agora, quando precisamos diminuir Δ no Passo 3 (devido a função de mérito não ter decrescido suficientemente), determinamos o "novo" Δ no intervalo $[0.1\Delta, 0.9\Delta]$. Naturalmente, fazer "Novo" $\Delta = \Delta/2$ é uma escolha admissível.

O vetor de multiplicadores $\bar{\lambda}$ pode ser escolhido de maneira totalmente arbitrária, sujeito à restrição $\|\bar{\lambda}\| \leq L$. Existem, no entanto, escolhas mais eficientes que outras, como veremos numa seção posterior. Por exemplo, uma boa idéia é escolher esse vetor de multiplicadores como o próprio vetor de multiplicadores associado à condição de otimalidade do subproblema (12.2.3). Mas o leitor pode aproveitar a liberdade que é admitida na escolha de $\bar{\lambda}$ para, numa primeira leitura deste capítulo, supor que $\bar{\lambda} = 0$. De fato, esta é uma escolha admissível e a maior parte da teoria funciona com ela.

Exercício 12.9: O leitor verificará que o procedimento indicado para diminuir Δ pode levar, se implementado de maneira ingênua, a repetir de maneira desnecessária a resolução de problemas de programação quadrática. Efetuar as modificações necessárias no Algoritmo 12.3.2 para que essa repetição seja claramente evitada.

Exercício 12.10: A definição da função Φ foi motivada na Seção 12.3. Refazer, na medida do possível, essa motivação esquecendo que os multiplicadores de Lagrange existem (ou seja, supondo que $\bar{\lambda} \equiv 0$). Analisar quais argumentos podem ser reutilizados e quais não.

Na Seção 12.5 veremos como calcular um parâmetro "de penalidade" θ adequado para cada iteração. Nesta seção, nos limitaremos a definir significado do símbolo $\ll$ em (12.3.6). Em nosso jargão, $a \ll b$ significa *a é "suficientemente menor" que b*, ou *a é menor que algo "claramente menor" que b*. Para especificar o significado de $\ll$ no caso de (12.3.6) precisamos de algumas considerações gerais sobre expansões de Taylor e, em particular, sobre expansões de f, h e φ. Vamos supor que tanto f como h têm derivadas segundas contínuas para todo $x \in I\!R^n$. (Esta é uma suposição desnecessariamente forte para nossos objetivos, mas suficientemente simples para fazer claro o raciocínio.) Ao mesmo tempo, as deduções serão mais legíveis se usamos, livremente, a notação $O(.)$. Lembramos que "$f = O(g)$" significa que existe uma constante c, independente da variável independente, tal que $f \leq cg$. Nosso objetivo agora é mostrar que a função de mérito Φ se aproxima bem por uma quadrática nas variáveis x e λ. O leitor interessado em fixar idéias, pode identificar x com x_k e s com $\bar{x} - x_k$ na seguinte seqüência de limitantes.
Pelo desenvolvimento de Taylor de h, temos que

$$\ell(x+s, \bar{\lambda}) - \ell(x+s, \lambda) = h(x+s)^T(\bar{\lambda} - \bar{\lambda}) = [h(x) + h'(x)s]^T(\bar{\lambda} - \lambda) + O(\|s\|^2). \tag{12.4.1}$$

Pelo desenvolvimento de Taylor de f e h e supondo que as matrizes B_k estão uniformemente limitadas, temos:

$$\ell(x+s,\lambda) - \ell(x,\lambda) = f(x+s) + h(x+s)^T\lambda - [f(x) + h(x)^T\lambda]$$

$$= f(x+s)-f(x)+[h(x+s)-h(x)]^T\lambda = \nabla f(x)^T s+\frac{1}{2}s^T B_k s+[h'(x)s]^T\lambda+O(\|s\|^2)$$

$$= [\nabla f(x)+h'(x)^T\lambda]^T s+\frac{1}{2}s^T B_k s+O(\|s\|^2) = \nabla_x \ell(x,\lambda)^T s+\frac{1}{2}s^T B_k s+O(\|s\|^2) \tag{12.4.2}$$

Somando membro a membro (12.4.1) e (12.4.2), obtemos:

$$\ell(x+s,\bar\lambda)-\ell(x,\lambda) = \nabla_x \ell(x,\lambda)^T s+\frac{1}{2}s^T B_k s+[h(x)+h'(x)s]^T(\bar\lambda-\lambda)+O(\|s\|^2). \tag{12.4.3}$$

Por outro lado, pelo desenvolvimento de Taylor de h,

$$h(x+s) = h(x) + h'(x)s + O(\|s\|^2),$$

portanto,

$$\|h(x+s)\|_2^2 = \|h(x) + h'(x)s\|_2^2 + O(\|s\|^2),$$

e, pela definição de φ,

$$\varphi(x+s) - \varphi(x) = \frac{1}{2}\|h(x) + h'(x)s\|_2^2 - \frac{1}{2}\|h(x)\|_2^2 + O(\|s\|^2). \tag{12.4.4}$$

Multiplicando (12.4.3) por θ, (12.4.4) por $1-\theta$, e somando membro a membro as duas expressões resultantes, obtemos:

$$\Phi(x,\lambda,\theta) - \Phi(x+s,\bar\lambda,\theta) = Pred(x,s,\lambda,\bar\lambda,B_k,\theta) + O(\|s\|^2), \tag{12.4.5}$$

onde

$$Pred(x,s,\lambda,\bar\lambda,B_k,\theta)$$

$$= -\{\theta[\nabla_x \ell(x,\lambda)^T s + \frac{1}{2}s^T B_k s + [h(x) + h'(x)s]^T(\bar\lambda - \lambda)]$$

$$+(1-\theta)[\frac{1}{2}\|h(x) + h'(x)s\|_2^2 - \frac{1}{2} + \|h(x)\|_2^2]\}. \tag{12.4.6}$$

Portanto, podemos considerar que a expressão $Pred$ é uma boa aproximação do decréscimo $\Phi(x,\lambda,\theta)-\Phi(x+s,\bar\lambda,\theta)$ na função de mérito Φ. Daí a denominação $Pred$, abreviatura de "predicted reduction". Brevemente, (12.4.5) significa que $\Phi(x,\lambda,\theta) - \Phi(x+s,\bar\lambda,\theta)$ coincide com $Pred$ para $s = 0$ junto com suas primeiras derivadas. Portanto, pelo menos quando $\|s\|$ é pequena, um decréscimo da ordem de $Pred$ na função de mérito Φ é de se esperar. Adiando, por um momento, a prova de que $Pred$ é, efetivamente, positivo, e adotando uma postura conservadora, diremos que $\Phi(\bar{x}, \bar{\lambda}, \theta) \ll \Phi(x_k, \lambda_k, \theta)$ quando

$$\Phi(x_k, \lambda_k, \theta) - \Phi(\bar{x}, \bar{\lambda}, \theta) \geq 0.1 Pred(x_k, \bar{x} - x_k, \lambda_k, \bar{\lambda}, B_k, \theta). \qquad (12.4.7)$$

Incorporando o critério de aceitação (12.4.7), definimos agora uma modificação do Algoritmo 12.3.2, com o qual finalizamos esta seção. O Algoritmo 12.4.1 é idêntico ao Algoritmo 12.3.2, com o critério impreciso (12.3.6) substituído por (12.4.7).

Algoritmo 12.4.1
Suponhamos que $x_0 \in I\!R^n$ $(l \leq x \leq u)$ é uma aproximação inicial da solução de (12.1.1) e $\lambda_0 \in I\!R^m, \|\lambda_0\| \leq L$ é uma aproximação inicial dos multiplicadores de Lagrange. Se x_k, λ_k $(k = 0, 1, 2, \ldots)$ são as aproximações obtidas na $k-$ésima iteração $(l \leq x_k \leq u,\ \|\lambda_k\| \leq L)$, $B_k \in I\!R^{n \times n}$ é uma matriz simétrica e $\Delta > 0$, então x_{k+1} é obtida da seguinte maneira:
Passo 1. Resolver (12.2.4) e (12.2.3).
Passo 2. Escolher um valor adequado para $\theta \in [0, 1]$ e estimar novos multiplicadores $\bar{\lambda}$ $(\|\bar{\lambda}\| \leq L)$.
Passo 3. Se $\bar{x}$, a solução obtida no Passo 1, satisfaz (12.4.7), definir $x_{k+1} = \bar{x}$, $\lambda_{k+1} = \bar{\lambda}$ e terminar a iteração. Caso contrário, diminuir Δ e retornar ao Passo 1.

12.5 O parâmetro de penalidade

Nesta seção discutiremos a escolha do parâmetro de penalidade θ, no Passo 2 do nosso algoritmo básico. A denominação "parâmetro de penalidade" se justifica, depois de observar que

$$\Phi(x, \lambda, \theta) = \theta[\ell(x, \lambda) + \frac{1 - \theta}{\theta} \varphi(x)].$$

Portanto, exigir decréscimo de $\Phi(x, \lambda, \theta)$ equivale a exigir decréscimo da função

$$\bar{\Phi}(x, \lambda, \rho) = \ell(x, \lambda) + \rho \varphi(x),$$

com $\rho = (1 - \theta)/\theta$. A função $\bar{\Phi}$ é um Lagrangiano aumentado, onde ρ é o parâmetro de penalidade clássico. Assim, $\rho \to \infty$ corresponde a $\theta \to 0$ e

$\rho \to 0$ corresponde a $\theta \to 1$. Pelos mesmos motivos, nos sentiremos livres para chamar Lagrangiano aumentado também à função de mérito Φ.
Na seção anterior observamos que, para que a condição (12.4.7) possa ser chamada com justiça de "decréscimo suficiente" era necessário que $Pred$ fosse maior que zero. No entanto, a resolução dos subproblemas (12.2.4) e (12.2.3) implica necessariamente que

$$\|h(x_k)\|_2^2 - \|h(x_k) + h'(x_k)(\bar{x} - x_k)\|_2^2 \geq 0.$$

Portanto, da definição de $Pred$ surge que, para $\theta = 0$,

$$Pred(x_k, \bar{x} - x_k, \lambda_k, \bar{\lambda}, B_k, \theta) \geq 0.$$

Ou seja, $Pred$ é uma combinação convexa do tipo $\theta a + (1-\theta)b$ onde, necessariamente, $b \geq 0$. No entanto, o elemento a dessa combinação convexa, não é necessariamente positivo. Para que $Pred$ seja, garantidamente, maior ou igual a 0, e maior que zero quando $b > 0$, vamos exigir que

$$Pred(x_k, \bar{x} - x_k, \lambda_k, \bar{\lambda}, B_k, \theta) \geq \frac{1}{2}[\|h(x_k)\|_2^2 - \|h(x_k) + h'(x_k)(\bar{x} - x_k)\|_2^2]. \tag{12.5.1}$$

Como (12.5.1) vale para $\theta = 0$, resulta que podemos definir $\theta^{sup} \geq 0$ por

$$\theta^{sup} = \sup \{\theta \in [0,1] \text{ tais que (12.5.1) se verifica}\}. \tag{12.5.2}$$

Se, no Passo 2 do algoritmo, escolhermos sempre $\theta \leq \theta^{sup}$, então, por (12.5.1), a condição (12.4.7) implicará descida simples da função de mérito. $(\Phi(\bar{x}, \bar{\lambda}, \theta) \leq \Phi(x_k, \lambda_k, \theta))$. Como valores maiores que θ^{sup} não satisfazem (12.5.1) parece bastante sensato, impor a condição

$$\theta \leq \theta^{sup} \tag{12.5.3}$$

para a escolha de θ no Passo 2. No entanto, o requisito (12.5.3) deixa ainda bastante liberdade, quando $\theta^{sup} > 0$. Outras considerações serão necessárias para fazer uma eleição adequada, dentro das possíveis.
O algoritmo baseado na função de mérito Φ poderia ser interpretado como um método destinado a minimizar a Φ sujeita apenas as restrições de canalização $l \leq x \leq u$. Esta interpretação parece ser compatível com o conceito genérico do significado de uma função de mérito. No entanto, neste caso, tal interpretação não parece totalmente adequada, devido à função Φ mudar de uma iteração para outra, de acordo com a escolha de θ. Com efeito, θ estabelece pesos relativos para a factibilidade e a otimalidade no algoritmo (com

θ perto de 0 o método privilegia factibilidade e com θ perto de 1 privilegia otimalidade). Grandes variações de θ de uma iteração para outra pareceriam indicar que o método não consegue decidir qual é o peso adequado para cada um dos objetivos que são visados. Essa é uma motivação para limitar, pelo menos assintoticamente, as variações de θ. A maneira mais óbvia de forçar uma variação limitada de θ, consiste em impor, além da condição (12.5.3), a seguinte:

$$\theta \leq \theta_{k-1}, \tag{12.5.4}$$

onde, para todo $k = 0, 1, 2, \ldots$, θ_k é o valor de θ escolhido na última passada pelo Passo 2. Juntando as condições (12.5.3) e (12.5.4), teremos que a seqüência $\{\theta_k\}$ é monótona não crescente e positiva, portanto convergente. Isso implicaria que, a longo prazo, a função de mérito seria, essencialmente, a mesma, e a interpretação criticada acima passaria a ser válida.

No entanto, a escolha monótona de θ também não é plenamente satisfatória. Lembrando que θ estabelece uma ponderação entre factibilidade e otimalidade, seria possível que, sobretudo nas primeiras iterações, valores muito pequenos de θ fossem impostos por (12.5.3) devido à necessidade de reforçar factibilidade, e que esses valores muito pequenos fossem herdados por todas as iterações posteriores, onde valores maiores seriam toleráveis. Em outras palavras, a condição (12.5.4) carrega demasiadamente a história de dificuldades passadas do algoritmo, que podem não existir mais na iteração atual. Essas considerações nos levam à definição da seguinte estratégia "não monótona" para θ: escolhe-se, independentemente de k um número $N > 0$ que representará o "grau de não-monotonicidade" de $\{\theta_k\}$. $N = 0$ corresponderá à escolha monótona, baseada em (12.5.4), e valores grandes de N aproximarão θ de θ^{sup}. Definimos

$$\theta_k^{min} = \min\,\{1, \theta_0, \ldots, \theta_{k-1}\}, \tag{12.5.5}$$

$$\theta_k^{grande} = (1 + (N/k)^{1.1})\theta_k^{min}, \tag{12.5.6}$$

e, finalmente,

$$\theta = \min\,\{\theta_k^{grande}, \theta^{sup}\}. \tag{12.5.7}$$

Apesar de (12.5.5)–(12.5.7) não implicar monotonia de $\{\theta_k\}$, essa escolha implica convergência da seqüência $\{\theta_k\}$ (ver Exercício 12.13), o que, do ponto de vista da interpretação da função de mérito, é igualmente satisfatório.

Como fizemos nas seções anteriores, a discussão realizada aqui nos permite especificar um pouco mais o algoritmo principal.

Algoritmo 12.5.1

Suponhamos que $x_0 \in I\!R^n$ $(l \leq x \leq u)$ é uma aproximação inicial da solução de (12.1.1) e $\lambda_0 \in I\!R^m, \|\lambda_0\| \leq L$ é uma aproximação inicial dos multiplicadores de Lagrange, $N, \Delta_{min} > 0$. Se x_k, λ_k $(k = 0, 1, 2, \ldots)$ são as aproximações obtidas na $k-$ésima iteração $(l \leq x_k \leq u$, $\|\lambda_k\| \leq L)$, $B_k \in I\!R^{n\times n}$ é uma matriz simétrica e $\Delta \geq \Delta_{min}$, então x_{k+1} é obtida da seguinte maneira:
Passo 1. Resolver (12.2.4) e (12.2.3).
Passo 2. Escoher $\theta \in [0, 1]$ usando (12.5.5)–(12.5.7) e estimar novos multiplicadores $\bar{\lambda}$ $(\|\bar{\lambda}\| \leq L)$.
Passo 3. Se $\bar{x}$, a solução obtida no Passo 1, satisfaz (12.4.7), definir $x_{k+1} = \bar{x}$, $\lambda_{k+1} = \bar{\lambda}$, $\theta_k = \theta$ e terminar a iteração. Caso contrário, diminuir Δ, (por exemplo, dividir Δ por 2) e retornar ao Passo 1.

Exercício 12.11: Em que caso o único parâmetro de penalidade que verfica $Pred \geq 0$ é $\theta = 0$?

Exercício 12.12: Obter uma fórmula explícita para θ^{sup}.

Exercício 12.13: Provar que a seqüência $\{\theta_k\}$ definida por (12.5.5)–(12.5.7), é convergente.

12.6 O algoritmo está bem definido

O método apresentado até aqui é muito análogo ao introduzido em [98]. A diferença fundamental é que em [98], visando aplicação a problemas de grande porte, os subproblemas (12.2.4) e (12.2.3) são resolvidos apenas "aproximadamente", com critérios adequados para a precisão da sua resolução. Para simplificar a exposição, apresentamos neste capítulo o algoritmo supondo solução exata de (12.2.4) e (12.2.3). A análise de convergência do algoritmo é complicada, e daremos apenas indicações sobre a mesma na Seção 12.7. Nesta seção, provaremos que o algoritmo está bem definido, isto é, que sob hipóteses adequadas, que incluem o fato de x_k ainda não ser uma solução, pode-se encontrar x_{k+1} em tempo finito. Em outras palavras, mostraremos que o ciclo através dos passos 1, 2 e 3 do algoritmo é finito.
Provaremos que o algoritmo está bem definido em duas situações:
(a) x_k *não é* um ponto estacionário do problema

$$\text{Minimizar } \varphi(x) \text{ sujeita a } \ell \leq x \leq u; \tag{12.6.1}$$

(b) x_k é um ponto factível, regular e não estacionário de (12.1.1).

Assim, ficam as seguintes situações em que o algoritmo *não está bem definido* e que, portanto, devem ser identificadas antes de começar o ciclo principal de cada iteração para evitar "loops" infinitos:
(c) x_k é um ponto estacionário de (12.6.1) mas $h(x_k) \neq 0$. (Lembremos que, por construção, $l \leq x_k \leq u$ para todo k.
(d) x_k é um ponto factível de (12.1.1) mas não é regular (os gradientes das restrições ativas em x_k, incluindo as canalizações, são linearmente dependentes).
(e) x_k é um ponto regular e estacionário de (12.1.1).
Nessas situações, o algoritmo deveria "parar". Delas, apenas (e) pode ser considerada um sucesso. A situação (c) representa, claramente, um "fracasso". Uma situação duvidosa é (d), já que um ponto não regular de (12.1.1) poderia ser minimizador global de (12.1.1). Não entraremos nesse tipo de sutileza.
Começaremos provando que o algoritmo está bem definido quando x_k não é um ponto estacionário de (12.6.1).

Teorema 12.6.1 - Boa definição em pontos não factíveis
Se x_k não é um ponto estacionário de (12.6.1), então o Algoritmo 12.5.1 calcula um novo ponto x_{k+1} através de uma quantidade finita de passagens pelos passos 1–3.

Prova: Definimos

$$M(x) = \frac{1}{2}\|h'(x_k)(x - x_k) + h(x_k)\|_2^2.$$

Claramente, $\nabla\varphi(x_k) = \nabla M(x_k) = h'(x_k)^T h(x_k)$, portanto x_k não é ponto estacionário de $M(x)$ sujeita a $l \leq x \leq u$. Portanto, existe uma direção factível e de descida para M na caixa $l \leq x \leq u$. Seja, pois, $d \in I\!R^n$ tal que $\|d\|_\infty = 1$ e $\nabla M(x_k)^T d < 0$.
A função $\beta(t) = M(x_k + td)$ é uma parábola convexa tal que $\beta'(0) = d^T \nabla M(x_k) < 0$. Se a parábola é estritamente convexa (coeficiente de segunda ordem estritamente positivo), admite um minimizador irrestrito $\hat{t} > 0$. Propriedades elementares das parábolas garantem, nesse caso, que

$$\beta(t) \leq \beta(0) + \frac{1}{2}\beta'(0)t \tag{12.6.2}$$

para todo $t \in [0, \hat{t}]$. Se $\beta(t)$ não é estritamente convexa, então é uma reta, e (12.6.2) se satisfaz trivialmente para todo $t \geq 0$.

Seja $\bar{t}$ o máximo dos t positvos tais que $l \leq x_k + td \leq u$ e $\tilde{t} = \min\{\hat{t}, \bar{t}\}$. Naturalmente, (12.6.2) vale para todo $t \in [0, \tilde{t}]$. Mais ainda, como $\|d\|_\infty = 1$, temos que $t = \|td\|_\infty$ e, em conseqüência, (12.6.2) implica a seguinte proposição: Para todo $\Delta \leq \tilde{t}/0.8 = \bar{\Delta}$, existe x tal que $l \leq x \leq u$ e $\|x - x_k\|_\infty \leq 0.8\Delta$ verificando

$$M(x) \leq M(0) - c\Delta,$$

onde $c = -0.4\beta'(0) > 0$.
Portanto, para $\Delta \leq \bar{\Delta}$, escrevendo $x^{nor} = x^{nor}(\Delta)$, temos que

$$\frac{1}{2}[\|h(x_k)\|_2^2 - \|h(x_k) + h'(x_k)(x^{nor}(\Delta) - x_k)\|_2^2] \geq c\Delta.$$

Logo, escrevendo $\bar{x} = \bar{x}(\Delta)$, deduzimos, pela forma do subproblema (12.2.3), que

$$\frac{1}{2}[\|h(x_k)\|_2^2 - \|h(x_k) + h'(x_k)(\bar{x}(\Delta) - x_k)\|_2^2] \geq c\Delta.$$

Portanto, de (12.5.1) inferimos que, para todo $\Delta \in (0, \bar{\Delta}]$,

$$Pred(x_k, \bar{x}(\Delta) - x_k, \lambda_k, \bar{\lambda}, B_k, \theta) \geq \frac{c}{2}\Delta > 0. \tag{12.6.3}$$

De (12.4.5) e (12.6.3) deduzimos que

$$\lim_{\Delta \to 0} \left| \frac{\Phi(x_k) - \Phi(\bar{x}(\Delta)}{Pred(x_k, \bar{x}(\Delta) - x_k, \lambda_k, \bar{\lambda}, B_k, \theta)} - 1 \right| = 0.$$

Este limite implica que, para Δ suficientemente pequeno o teste (12.4.7) é satisfeito. Portanto, a iteração termina depois de um número finito de reduções de Δ. **QED**

Nosso próximo passo consiste em provar que, se x_k é um ponto factível, regular e não estacionário de (12.1.1), então a iteração definida pelo algoritmo 12.5.1 também termina em tempo finito.

Teorema 12.6.2 - Boa definição em pontos factíveis
Suponhamos que x_k é um ponto factível, regular e não estacionário de (12.1.1). Então o Algoritmo 12.5.1 calcula um novo ponto x_{k+1} através de uma quantidade finita de passagens pelos passos 1–3.

Prova: Definimos, analogamente ao Teorema 12.6.1,

$$Q(x) = \frac{1}{2}(x - x_k)^T B_k (x - x_k) + \nabla f(x_k)(x - x_k) + f(x_k).$$

Consideramos o problema de programação quadrática

$$\text{Minimizar } Q(x), \text{ sujeita a } h'(x_k)(x - x_k) = 0, l \leq x \leq u. \tag{12.6.4}$$

Claramente, x_k é um ponto factível e regular do problema (12.6.4). Mais ainda, as condições de otimalidade de (12.1.1) e de (12.6.4) em x_k são idênticas. Como, por hipótese, elas não são cumpridas para (12.1.1), segue-se que x_k não é um ponto estacionário de (12.6.4). Portanto, existe uma direção factível, unitária ($\|d\|_\infty = 1$) e de descida para o problema (12.6.4). Logo, $\nabla Q(x_k)^T d < 0$. Definimos

$$\beta(t) = Q(x_k + td).$$

Pelo mesmo raciocínio do Teorema 12.6.1, podemos garantir que existem $\tilde{t} > 0$ e $c > 0$ tais que para todo $t \in [0, \tilde{t}]$, $x_k + td$ é factível para o problema (12.6.4) e

$$Q(x_k) - Q(x_k + td) \geq ct.$$

Portanto, como $\|td\|_\infty = t$, podemos afirmar que, para todo Δ suficientemente pequeno, digamos $\Delta \leq \bar{\Delta}$, existe um ponto $\tilde{x}$ factível para (12.6.4) tal que

$$Q(x_k) - Q(\tilde{x}) \geq c\Delta.$$

De acordo com a definição de $\bar{x} = \bar{x}(\Delta)$ no subproblema (12.2.3), isto implica que

$$Q(x_k) - Q(\bar{x}) \geq c\Delta. \tag{12.6.5}$$

Agora, como $\bar{x} - \bar{x_k}$ está, neste caso, no núcleo de $h'(x_k)$ e $h(x_k) = 0$, a desigualdade (12.6.5) implica que

$$-[\nabla_x \ell(x_k, \lambda_k)^T(\bar{x} - x_k) + \frac{1}{2}(\bar{x} - x_k)^T B_k(\bar{x} - x_k)$$

$$+[h(x_k) + h'(x_k)(\bar{x} - x_k)]^T(\bar{\lambda} - \lambda_k)] \geq c + \Delta > 0.$$

Logo, pela definição de $Pred$ temos que

$$Pred(x_k, \bar{x} - x_k, \lambda_k, \bar{\lambda}, B_k, \theta) \geq \theta c\Delta > 0.$$

Agora, como $h(x_k) = h'(x_k)(\bar{x} - x_k) = 0$, temos que todos os $\theta \in (0, 1]$ satisfazem o teste (12.5.1) para $\Delta \leq \bar{\Delta}$. Isto implica que, para esses valores de Δ, o parâmetro θ não precisa ser reduzido. Portanto, existe $\theta' > 0$ tal que

$$Pred(x_k, \bar{x} - x_k, \lambda_k, \bar{\lambda}, B_k, \theta) \geq \theta' c\Delta > 0 \tag{12.6.6}$$

para todo $\Delta \in (0, \bar{\Delta}]$. Como no caso do Teorema 12.6.1, segue que

$$\lim_{\Delta \to 0} \left| \frac{\Phi(x_k) - \Phi(\bar{x}(\Delta)}{Pred(x_k, \bar{x}(\Delta) - x_k, \lambda_k, \bar{\lambda}, B_k, \theta)} - 1 \right| = 0.$$

Logo, para Δ suficientemente pequeno o teste (12.4.7) é satisfeito e, assim, a iteração termina depois de um número finito de reduções de Δ. **QED**

12.7 A prova de convergência global

É comum que a prova da convergência global de um algoritmo esteja muito relacionada com a prova de boa definição. Isto é bastante natural já que, na boa definição, provamos que os pontos onde o algoritmo deve *parar* têm determinadas características, e nos teoremas de convergência, geralmente, provamos que os pontos limite da seqüência gerada têm essas mesmas características. Logo, os teoremas de convergência dizem sobre o limite a mesma coisa que os resultados de boa definição dizem sobre os iterandos. Muitas vezes, as provas de convergência global reproduzem, com variadas complicações analíticas, as idéias usadas para provar boa definição.

Nesta seção procuraremos dar as idéias essenciais da prova de convergência do Algoritmo 12.5.1. Os argumentos completos podem ser encontrados em [98].

A prova tem duas partes, que correspondem aos teoremas 12.6.1 e 12.6.2. Nos dois casos usa-se como hipótese a seqüência gerada estar totalmente contida em um compacto de $I\!R^n$. Evidentemente, quando as cotas l e u são finitas, esta é uma hipótese perfeitamente razoável. Na primeira parte se prova que *todos* os pontos limites de uma seqüência gerada pelo algoritmo são pontos estacionários de (12.6.1). Para demonstrar esse fato, passa-se por um processo comparável ao usado para provar o Teorema 12.6.1:

(a) Prova-se que, se x_* não é um ponto estacionário de (12.6.1), então, nos iterandos x_k próximos a x_*, a quantidade $Pred$, pensada como função de Δ é proporcional a Δ. Isto é análogo a (12.6.3), mas a constante da proporcionalidade é, neste caso, independente de k.

(b) Usa-se a fórmula de Taylor para mostrar que $Pred$ é uma aproximação de segunda ordem da redução da função de mérito. Junto com o resultado (a), isso implica, como no Teorema 12.5.1, que

$$\left| \frac{\Phi(x_k) - \Phi(\bar{x})}{Pred(\Delta)} - 1 \right| = O(\Delta).$$

(c) Supondo que x_* é um ponto limite não estacionário para (12.6.1), o resultado (b) implica que, em todos os iterandos numa vizinhança de x_*, o raio de confiança finalmente aceito Δ_k é uniformemente maior que um número positivo fixo $\tilde{\Delta}$. Junto com (b), isto implica que a redução da função de mérito em uma quantidade infinita de iterações vizinhas de x_* é superior a uma quantidade positiva fixa.
(d) Se a função de mérito fosse sempre a mesma para todo k suficientemente grande, o resultado (c) seria suficiente para chegar a um absurdo (função de mérito tendendo a $-\infty$ em condições de compacidade). Como a função de mérito muda de uma iteração para outra, esse absurdo se consegue apenas pela propriedade de convergência da seqüência θ_k que, como vemos aqui, é crucial do ponto de vista teórico.
Na segunda parte da prova de convergência se demonstra a existência de *pelo menos um* ponto limite que é estacionário para o problema (12.1.1). Não existe ainda uma prova de que *todos* os pontos limites são estacionários e, ao longo de toda a demonstração desta segunda parte, é usada, por absurdo, a hipótese de que nenhum ponto limite da seqüência é estacionário. Outras suposições sobre o problema também são necessárias nesta parte:
(i) Todos os pontos estacionários de (12.6.1) são factíveis.
(ii) Todos os pontos factíveis de (12.1.1) são regulares.
Devido à hipótese (i), pode-se supor, ao longo da prova, que

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

Na primeira parte da prova por absurdo, demonstra-se que a função (quadrática) objetivo de (12.2.3) tem um bom decréscimo (proporcional a Δ) desde $x^{nor}(\Delta)$ até $\bar{x}(\Delta)$. Chamamos a esta variação de "decréscimo tangencial". O argumento se baseia em x^{nor} ser um ponto factível de (12.2.3) e, devido a $\|x^{nor} - x_k\| \le 0.8\Delta$, existir uma folga (brevemente, de 0.2Δ) para um bom decréscimo da quadrática.
Na segunda parte da prova, examinamos a composição da quantidade crucial que chamamos *Pred*. Como na prova da estacionariedade em relação a φ dos pontos limite, necessitamos que *Pred* seja positivo e proporcional a Δ. O decréscimo proporcional a Δ da função objetivo de (12.2.3), entre x^{nor} e $\bar{x}$ é um bom passo. Agora, observando a definição (12.4.6) de *Pred*, vemos que o termo que multiplica θ está composto, além do decréscimo da quadrática entre x^{nor} e $\bar{x}$, pela variação dessa quadrática entre x_k e x^{nor} e pelo termo que envolve a variação dos multiplicadores de Lagrange. Esses dois termos "estorvam" o objetivo de ter um *Pred* suficientemente positivo. Por outro lado, o termo que multiplica a $1-\theta$ é, claramente, proporcional a $\|h(x_k)\|$,

que tende a zero. Portanto, para ter um $Pred$ positivo e proporcional a Δ, precisaremos que θ não evolua para valores próximos de zero, e, por outro lado, que o "estorvo" seja dominado pelo decréscimo tangencial da quadrática.
Não é difícil provar que o "estorvo" está limitado, em módulo, por um múltiplo de $\|h(x_k)\|$. Escrevendo

$$|\text{Estorvo}| \leq c_1\|h(x_k)\|$$

e

$$\text{Decréscimo tangencial } \geq c_2\Delta,$$

e, desde que

$$Pred(\Delta) \geq \text{ Decréscimo tangencial } - |\text{Estorvo}|,$$

se deduz que

$$Pred(\Delta) \geq c_2\Delta - c_1\|h(x_k)\|.$$

Portanto, se $\|h(x_k)\| \leq \alpha\Delta$, com $\alpha = c_2/(2c_1)$, obtemos que $Pred(\Delta)$ é positivo e proporcional a Δ.
Pensamos agora no "plano" $(\Delta, h(x))$. O argumento acima nos leva a considerar uma "zona boa" do plano, formado pelos pares (Δ, x_k) tais que $\|h(x_k)\| \leq \Delta$ e uma "zona ruim", onde o contrário acontece. Na zona boa, o fator de θ em (12.4.6) é tão grande, e o fator de $1-\theta$ tão pequeno, assintoticamente, que a condição (12.5.1) se satisfaz com valores grandes de θ. Portanto, sempre que o par se encontre na zona boa θ não precisará ser diminuído.
Por outro lado, o mesmo raciocínio usado na prova de estacionariedade em relação a φ leva a que $\theta_k \to 0$. Com efeito, se assim não fosse, os valores de $Pred$ para esses k seriam superiores a um múltiplo de Δ, já que o fato do primeiro Δ testado ser superior ao valor fixo Δ_{min}, obriga a que a seqüência de possíveis Δ's fracassados dentro de uma mesma iteração não possa tender a zero. Teríamos assim, infinitos θ superiores a um valor fixo e infinitos Δ_k superiores a um valor fixo. As duas coisas juntas levam a uma função de mérito tendendo a $-\infty$, o que é absurdo.
O argumento central continua com uma propriedade surpreendente da zona ruim: uma análise cuidadosa da aproximação de Taylor da função de mérito Φ, junto com a propriedade $\theta_k \to 0$, provam que, nessa zona, para k suficientemente grande, o raio de confiança Δ é necessariamente aceito. Em outras palavras, para cada iteração k pode haver apenas uma tentativa Δ dentro da zona ruim. Por outro lado, como vimos antes, é apenas nesta situação que

pode ser necessário diminuir θ. Uma propriedade adicional da zona ruim é que, nessa zona, θ^{sup} é sempre superior a um múltiplo de Δ. Juntando as duas propriedades acima, diríamos que é possível entrar na indesejável zona ruim, mas pouco, e que é possível ter que diminuir θ na zona ruim, mas de maneira controlada.

Não é de se estranhar, em conseqüência, que os efeitos perniciosos da zona ruim estejam também limitados. De fato, usando as propriedades acima e, de novo, a expansão de Taylor da função de mérito, chega-se a conclusão de que o quociente entre a variação desta e $Pred$ converge a 1 considerando apenas raios na zona boa. Isso é uma flagrante contradição, porque implicaria em jamais ser necessário entrar na zona ruim. Tais contradições se originam na suposição errônea original que, como lembramos, consistia em assumir que nenhum ponto limite era estacionário para o problema (12.1.1).

12.8 A Hessiana da quadrática

Os algoritmos estudados neste capítulo permitem uma grande liberdade na escolha na matriz B_k, Hessiana da quadrática função objetivo de (12.2.3). O Algoritmo 12.5.1 exige apenas que a seqüência de matrizes B_k esteja uniformemente limitada. Por exemplo, a teoria é válida se todas as B_k são nulas, caso no qual poderíamos falar, mais apropriadamente, de "programação linear seqüencial".

No entanto, como acontece na resolução de sistemas não lineares e na minimização de funções sem restrições, existem escolhas ótimas para as matrizes que definem os algoritmos, e outras escolhas francamente desaconselháveis. Nos algoritmos de regiões de confiança sem restrições a melhor escolha é a Hessiana da função objetivo. Apesar disso, a teoria de convergência global para condições de primeira ordem funcionaria mesmo que escolhêssemos sua inversa aditiva!

De um modo geral, estamos acostumados a pensar que a escolha ótima de uma matriz é a que se relaciona mais diretamente com o método de Newton. Vejamos aonde nos leva este tipo de argumento no caso da programação quadrática seqüencial.

Para fixar idéias, vamos considerar nesta seção problemas do tipo (12.1.1) apenas com as restrições de igualdade, ou seja:

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & h(x) = 0 \end{array} \qquad (12.8.1)$$

À primeira vista, a escolha mais "newtoniana" para B_k é a própria Hessiana da função objetivo: $B_k = \nabla^2 f(x_k)$. No entanto, o seguinte problema simples ajuda a levantar alguma suspeita sobre essa eleição:

$$\begin{array}{rl} \text{Minimizar} & 4(x_1-1)^2 + x_2^2 \\ \text{sujeita a} & x_1 - x_2^2 = 0 \ , \end{array} \tag{12.8.2}$$

Neste problema, o ponto $(0,0)$ seria um minimizador para

$$\begin{array}{rl} \text{Minimizar} & 4(x_1-1)^2 + x_2^2 \\ \text{sujeita a} & x_1 = 0 \ , \end{array}$$

mas um maximizador para (12.8.2). Em outras palavras, quando tomamos $B_k = \nabla^2 f(x_k)$ em (12.2.3), perdemos informações sobre a curvatura das restrições. Isto nos sugere que devemos incorporar em B_k as derivadas segundas de h.

Vejamos a situação sob outro ponto de vista, mais claramente newtoniano. Consideremos as condições de otimalidade do problema (12.8.1). Se x_* é um ponto regular minimizador local de (12.8.1), então existe $\lambda_* \in \mathbb{R}^m$ tal que

$$\begin{array}{rcl} \nabla f(x_*) + h'(x_*)^T \lambda_* & = & 0 \\ h(x_*) & = & 0 \ . \end{array} \tag{12.8.3}$$

Pensando (12.8.3) como um sistema não linear nas variáveis (x,λ) $(F(x,\lambda)=0)$, seu Jacobiano é

$$F'(x,\lambda) = \begin{pmatrix} \nabla^2 f(x) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x) & h'(x)^T \\ h'(x) & 0 \end{pmatrix}$$

Portanto, o método de Newton aplicado a $F(x,\lambda)=0$ vem dado por

$$[\nabla^2 f(x_k) + \sum_{i=1}^m [\lambda_k]_i \nabla^2 h_i(x_k)](x-x_k) + h'(x_k)^T(\lambda-\lambda_k) = -(\nabla f(x_k) + h'(x_k)^T \lambda_k)$$

e

$$h'(x_k)(x-x_k) = -h(x_k),$$

ou seja,

$$\begin{array}{rcl} [\nabla^2 f(x_k) + \sum_{i=1}^m [\lambda_k]_i \nabla^2 h_i(x_k)](x-x_k) + h'(x_k)^T\lambda + \nabla f(x_k) & = & 0 \\ h'(x_k)(x-x_k) + h(x_k) & = & 0 \ . \end{array} \tag{12.8.4}$$

Agora, as condições de otimalidade de (12.1.4), sem as restrições de canalização $l \le x \le u$, são

$$\begin{aligned} B_k(x - x_k) + \nabla f(x_k) + h'(x_k)^T y &= 0 \\ h'(x_k)(x - x_k) + h(x_k) &= 0 \end{aligned} \tag{12.8.5}$$

onde $y \in I\!R^m$. Logo, comparando (12.8.4) com (12.8.5), o método de Newton nos sugere que

$$B_k = \nabla^2 f(x_k) + \sum_{i=1}^{m} [\lambda_k]_i \nabla^2 h_i(x_k) \tag{12.8.6}$$

onde $\lambda_k \in I\!R^m$ é uma estimativa para os multiplicadores de Lagrange. Com a escolha (12.8.6) para B_k, a curvatura das restrições está sendo contemplada. A matriz B_k ideal seria portanto a Hessiana do Lagrangiano, para a qual as propriedades de convergência local do método definido pelo subproblema (12.1.4) seriam as mesmas que as do método de Newton aplicado ao sistema definido por (12.8.3). Para outras aproximações para B_k, a convergência local seria a mesma que a de um método quase-Newton. Boggs, Tolle e Wang [19] deram uma condição análoga à condição Dennis-Moré para a convergência superlinear de métodos quase-Newton aplicados a (12.8.3). Uma conseqüência dessa condição é que, supondo não singularidade da Jacobiana do sistema (12.8.3), se as matrizes B_k convergem à Hessiana do Lagrangiano na solução, então a convergência *do par* (x_k, λ_k) para (x_*, λ_*) é superlinear.

Exercício 12.14: Discutir duas alternativas para o coeficiente linear de (12.1.4)–(12.2.3): $\nabla f(x_k)$ e $\nabla \ell(x_k)$. Justificar a afirmação de que, em um caso, (12.1.4)–(12.2.3) fornece diretamente a nova estimativa dos multiplicadores de Lagrange, e no outro, fornece seu incremento.

Exercício 12.15: Relacionar a não singularidade da Jacobiana do sistema (12.8.3) na solução com as propriedades da Hessiana do Lagrangiano no núcleo de $h'(x_*)$. Relacionar com as condições suficientes de otimalidade de segunda ordem para minimização com restrições de igualdade.

Uma abordagem quase-newtoniana bastante empregada é atualizar B_k com algo análogo à popular fórmula BFGS de minimização sem restrições:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k}$$

onde $s_k = x_{k+1} - x_k$ e $y_k = \nabla_x \ell(x_{k+1}, \lambda_{k+1}) - \nabla_x \ell(x_k, \lambda_k)$. Se B_k é definida positiva, como no caso de minimização sem restrições, a condição $s_k^T y_k > 0$

garante que B_{k+1} é definida positiva. No entanto, pode ser que s_k e y_k não satisfaçam essa desigualdade. Powell [164] propõe que y_k seja substituído por

$$\bar{y}_k = \theta y_k + (1-\theta) B_k s_k \ ,$$

onde

$$\theta = \begin{cases} 1 & , \quad s_k^T y_k \geq 0.2 s_k^T B_k s_k \\ \dfrac{0.8 s_k^T B_k s_k}{s_k^T B_k s_k - s_k^T y_k} & , \quad s_k^T y_k < 0.2 s_k^T B_k s_k \ . \end{cases}$$

No entanto, o mesmo autor [167] observa que a substituição de y_k por $\bar{y}_k$ pode ser instável. Boggs e Tolle [17], por sua vez, propõem que $B_{k+1} = B_k$ quando $s_k^T y_k < 0$.

Exercício 12.16: Supor que o problema (12.1.4) sem canalizações é factível. Mostrar, usando uma base do núcleo de $h'(x_k)$, como esse problema pode ser reduzido à minimização de uma quadrática sem restrições. Em que condições esse problema tem solução única? Supondo que B_k é definida positiva, e escrevendo $A_k = h'(x_k)$, $h_k = h(x_k)$, $g_k \nabla f(x_k)$, provar que a solução desse problema é

$$\bar{x} = x_k - B_k^{-1}(g_k + A_k^T z)$$

onde

$$z = (A_k B_k^{-1} A_k^T)^{-1}(h_k - A_k B_k^{-1} g_k).$$

Discutir a praticidade dessas fórmulas. Por exemplo, analisar o que acontece em relação à conservação da possível esparsidade de A_k e B_k.

12.9 Outras funções de mérito

No Algoritmo 12.5.1 usamos a função de mérito

$$\Phi(x, \lambda, \theta) = \theta \ell(x, \lambda) + (1-\theta)\varphi(x),$$

com $\varphi(x) = \|h(x)\|_2^2/2$. Usar esta função, com o parâmetro θ entre 0 e 1, é essencialmente equivalente a usar

$$\Phi_\rho(x, \lambda, \rho) = \ell(x, \lambda) + \rho\varphi(x), \qquad (12.9.1)$$

que é a forma tradicional do Lagrangiano aumentado. Agora, vimos que a teoria de convergência global permite um enorme liberdade para as aproximações dos multiplicadores λ_k. Em particular, é admissível usar sempre

$\lambda_k = 0$, o que, por outro lado, permite uma leitura mais simples da teoria. Agora, usar $\lambda_k = 0$ corresponde a trabalhar com a função de mérito

$$\Phi_{quad}(x, \rho) = f(x) + \rho\varphi(x). \tag{12.9.2}$$

Claramente, Φ_{quad} é a clássica função de penalidade quadrática. Com a introdução dos multiplicadores na função (12.9.2) esperamos que o parâmetro de penalidade ρ não precise crescer muito, eliminando possíveis fontes de instabilidade numérica, o que não é refletido numa teoria de convergência global.

No entanto, podemos analisar o comportamento da função Φ_{quad} sob outro aspecto. Como sabemos, a aplicação do método de Newton ao sistema (12.8.3), tem propriedades de convergência local quadrática, no par (x, λ), quando a Jacobiana na solução é não singular. Nessas condições, o método de Newton pode ser interpretado como a resolução recursiva do subproblema de programação quadrática (12.1.4) com as matrizes B_k sendo as Hessianas dos Lagrangianos. Como este método é localmente rápido, é desejável que, dado x_k, a solução $\bar{x}$ aportada pela resolução de (12.1.4) seja aceita como nova iteração x_{k+1} e que não seja necessário apelar, neste caso, para diminuições do raio de confiança Δ. Agora, para que isso aconteça, é necessário, pelo menos, que a função de mérito calculada em $(\bar{x}, \bar{\lambda})$ (solução de (12.1.4) e multiplicador correspondente) seja menor que a mesma função em (x_k, λ_k). Caso contrário, a função de mérito estaria recomendando rejeitar um ponto essencialmente bom.

Infelizmente, muitas funções de mérito têm essa desagradável propriedade, que é denominada *efeito Maratos*. Ver [131]. O efeito Maratos reflete, assim, um conflito entre o ponto de vista *Cauchy*, que exige diminuição de uma função objetivo, e o ponto de vista *Newton* que produz convergência local rápida. Em particular, a função de mérito Φ_{quad} sofre dessa propriedade e inibe convergência rápida do método de Newton em circunstâncias onde ela seria perfeitamente possível.

Exercício 12.17: Considerar o problema

$$\begin{array}{rl} \text{Minimizar} & x_2 \\ \text{sujeita a} & x_1^2 + x_2^2 = 1 \end{array}$$

e a função de mérito $\Phi_\rho(x) = x_2 + \rho|x_1^2 + x_2^2 - 1|$ para ρ suficientemente grande de maneira que o minimizador de Φ_ρ seja $(0, -1)^T$. Verificar o efeito Maratos.

Diferenciabilidade, parâmetros de penalidade moderados, simplicidade e ausência de efeito Maratos são qualidades desejáveis das funções de mérito aplicadas a programação quadrática seqüencial. Vejamos como aparecem (ou não) essas propriedades em outras funções sugeridas na literatura.
(a) A função de penalidade com $\| \cdot \|_1$, dada por

$$\Phi_\rho(x) = f(x) + \rho \|h(x)\|_1$$

é interessante por ser *exata*, isto é, para um valor finito do parâmetro, seu minimizador é a solução do problema de otimização original, como vimos no Exercício 12.17. No entanto, ela não é diferenciável e sofre do efeito Maratos.

(b) A função de penalidade exata de Fletcher

$$\Phi_\rho(x) = f(x) - h(x)^T \lambda(x) + \frac{\rho}{2} \|h(x)\|_2^2 \ ,$$

onde $\lambda(x) = (h'(x)^T)^\dagger \nabla f(x)$, não tem efeito Maratos, mas é computacionalmente cara, o que a faz pouco interessante para problemas grandes.

(c) A soma de quadrados associada ao sistema não linear:

$$\Phi(x, \lambda) = \|\nabla f(x) + h'(x)^T \lambda\|_2^2 + \|h(x)\|_2^2$$

não tem efeito Maratos, é diferenciável e simples. Porém, praticamente não é usada porque seus minimizadores resultam tanto em minimizadores quanto em maximizadores do problema original.

O Lagrangiano aumentado usado neste capítulo é simples e diferenciável. No entanto, a moderação nos parâmetros de penalidade e o efeito Maratos dependem da escolha dos multiplicadores λ. No momento em que escrevemos este capítulo, a teoria de convergência local do Algoritmo 12.5.1 não está completa, mas é previsível que ela incluirá os seguintes resultados:
(a) Em condições adequadas de regularidade local do problema (12.1.1) (i) o subproblema (12.1.4) coincide com (12.2.3); (ii) (12.1.4) tem solução única; (iii) com uma boa escolha dos multiplicadores λ_k e das matrizes B_k os parâmetros de penalidade θ_k são todos maiores que um número positivo fixo e a solução de (12.1.4) é aceita como próxima iteração x_{k+1}.
(b) Nas condições acima, se as B_k são Hessianas dos Lagrangianos, a convergência de (x_k, λ_k) para (x_*, λ_*) é quadrática. Para escolhas quase-newtonianas adequadas de B_k, a convergência é superlinear. Para determinadas estimativas de λ_k a convergência é quadrática no caso Newton e superlinear no

caso quase-Newton considerando apenas a variável x.

Exercício 12.18: Discutir convergência quadrática ou superlinear na variável x e no par (x, λ). Qual é mais forte? Qual é mais desejável? Dar exemplos mostrando quando uma não implica a outra.

Exercício 12.19: Schittkowski (1981) e Gill, Murray, Saunders e Wright (1992), entre outros, estudaram o problema (12.1.1) na forma

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeita a} & c(x) \leq 0 \end{array} \tag{12.9.3}$$

Para construir uma função de mérito, esses autores introduzem variáveis de folga nas restrições, apenas para efetuar a busca na função de mérito

$$c_i(x) = 0 \quad \Leftrightarrow \quad c_i(x) + s_i = 0 \ , \ s_i \geq 0 \ , \ i = 1, \ldots, p$$

e então

$$\Phi_\rho(x, \lambda, s) = f(x) + \mu^T(c(x) + s) + \frac{\rho}{2}\|c(x) + s\|_2^2$$

onde $\mu \in I\!\!R^p$ é uma estimativa para os multiplicadores. Discutir as propriedades dessa função.

Exercício 12.20: Existem duas estratégias para a formulação dos subproblemas quadráticos num método PQS aplicado a (12.9.3). Na primeira, baseada em desigualdades, trabalha-se com problemas quadráticos com restrições lineares de desigualdade, e a decisão acerca do conjunto de restrições ativas é tomada internamente durante a resolução do problema quadrático. A segunda estratégia, baseada em igualdades, consiste em fixar-se a priori quais serão as restrições ativas e então trabalhar com subproblemas quadráticos com restrições de igualdade. O conjunto de restrições ativas $I_k \subset \{1, \ldots, p\}$ é atualizado a cada iteração pela análise dos multiplicadores de Lagrange do subproblema e pelo exame dos valores $c_i(x_{k+1})$ para $i \notin I_k$. É possível ainda adotar-se uma estratégia híbrida, isto é, baseada em desigualdades, mas com um "warm start" para o conjunto das restrições ativas, com o objetivo de melhorar a eficiência do algoritmo. Fazer uma análise a priori das possíveis vantagens e desvantagens das duas estratégias.

12.10 Notas históricas

A primeira proposta de um método de programação quadrática seqüencial foi feita por Wilson (1963) em sua tese de doutorado, para problemas convexos. Ele trabalhou com subproblemas quadráticos com restrições de desigualdade e utilizou a própria matriz Hessiana do Lagrangiano no modelo quadrático. Como estimativa para os multiplicadores, Wilson utilizou os multiplicadores do subproblema na iteração anterior.

A abordagem de Wilson foi retomada e interpretada por Beale (1967), originando o algoritmo SOLVER. Bard e Greenstadt (1969) reinterpretaram SOLVER, mostrando que o algoritmo de Wilson-Beale pode ser dividido em dois passos: primeiro fixar os multiplicadores λ_k e obter $x(\lambda_k)$ minimizando o Lagrangiano do subproblema e a seguir obter λ_{k+1} e a correção $x(\lambda_{k+1}) - x(\lambda_k)$ pela maximização deste mesmo Lagrangiano. Murray ([150], [151]) estendeu os trabalhos anteriores, incluindo aproximações quase-Newton para a Hessiana do modelo quadrático e estimativas diferentes para os multiplicadores de Lagrange. Além disso, Murray também considerou a resolução parcial do subproblema e sugeriu uma busca linear a cada iteração utilizando a função de penalidade quadrática (ver também Gill e Murray(1974), cap.8, parte III).

Biggs (1972, 1974, 1975) propôs uma variação do método de Murray, com subproblemas quadráticos apenas com restrições de igualdade e sugeriu estimativas especiais para os multiplicadores. García-Palomares e Mangasarian (1976) sugeriram um método baseado em programação quadrática derivado da aplicação de técnicas quase-Newton ao sistema não linear proveniente das condições de otimalidade do problema original. Han (1976 e 1977) retomou a idéia original de Wilson, trabalhando com restrições de desigualdade nos subproblemas quadráticos, mas sugerindo atualizações quase-Newton definidas positivas para a matriz Hessiana do Lagrangiano. As estimativas para os multiplicadores são tomadas como os multiplicadores da iteração anterior. No algoritmo de Han, superlinearmente convergente sob certas hipóteses, a função de penalidade exata ℓ_1 é usada pela primeira vez como função de mérito.

Powell (1977 e 1978) propôs um algoritmo de programação quadrática seqüencial semelhante ao de Han, com aproximações quase-Newton definidas positivas para a Hessiana do Lagrangiano e também superlinearmente convergente sob algumas hipóteses. Nesta linha de trabalho baseada em aproximações quase-Newton para a matriz Hessiana do Lagrangiano destacam-se as estratégias de Powell (1977), Murray e Wright (1978), Schittkowski (1980) e Boggs, Tolle e Wang (1982). Como afirmamos em outra seção, Boggs, Tolle e Wang obtiveram uma condição necessária e suficiente para convergência superlinear do tipo da condição de Dennis-Moré para sistemas não lineares. Maratos (1978) e Chamberlain (1979) descrevem algumas dificuldades decorrentes do uso da função de penalidade exata baseada em $\|\cdot\|_1$ como função de mérito. Chamberlain, Lemaréchal, Pederson e Powell (1980) também analisam alguns aspectos dessa penalidade exata como função de mérito. Uma proposta para evitar o efeito Maratos, baseada em buscas lineares não monótonas, é feita por Panier e Tits (1991) e complementada por Bonnans, Panier, Tits e Zhou (1992).

Murray e Wright (1980) fazem uma discussão de diferentes formulações para o subproblema. Métodos de programação quadrática seqüencial cujo subproblema lida apenas com restrições de igualdade (fixando-se a priori as restrições ativas) são tratados por Wright (1976), que introduziu o uso da função Lagrangiano aumentado como função de mérito, e por Murray e Wright (1978).

Cabe observar que muitos métodos para minimização de funções de penalidade não diferenciáveis têm a mesma forma que métodos de programação quadrática seqüencial nos quais a direção de busca é obtida pela decomposição em dois passos ortogonais: um no núcleo e outro no espaço linha do Jacobiano das restrições. Nesta classe se enquadram os métodos de Coleman (1979), Coleman e Conn (1980 e 1984), Fontecilla (1983) e Nocedal e Overton (1985). Em termos de atualizações para a matriz Hessiana do modelo quadrático, Fontecilla (1983), Colemman e Conn (1984) e Nocedal e Overton (1985) conservaram as matrizes B_k definidas positivas apenas no subespaço tangente às restrições. Já Celis, Dennis e Tapia (1985) trabalharam com métodos de região de confiança, nos quais não se precisa de matrizes positivas definidas como garantia para existência de solução nos subproblemas.

Powell e Yuan (1986) trabalharam com Lagrangiano aumentado como função de mérito, em problemas com restrições de igualdade. Os multiplicadores são estimados por quadrados mínimos, sendo portanto tratados como funções do ponto atual. Neste trabalho, Powell e Yuan provam propriedades de convergência global e local.

Outras funções de mérito suaves foram consideradas por Dixon (1979), Di Pillo e Grippo (1979), Schittkowski (1981), Boggs e Tolle (1984,1985), Bartholomew-Biggs (1987) e Gill, Murray, Saunders e Wright (1992).

Os multiplicadores como variáveis adicionais, com busca linear com relação ao vetor aumentado que contém as variáveis originais e os multiplicadores foram usados por Tapia (1977) no contexto de Lagrangiano aumentado e subproblemas irrestritos. Também foi aplicada por Schittkowski (1981) e Gill,

Murray, Saunders e Wright (1992) em algoritmos de programação quadrática seqüencial.
Em programação quadrática seqüencial, é possível truncar o procedimento iterativo para resolução do subproblema quadrático sem alterar a taxa de convergência assintótica. Neste sentido, critérios práticos de parada são apresentados por Dembo e Tulowitzki (1985), Fontecilla (1985, 1990) e Yabe, Yamaki e Takahashi (1991).
Para problemas de grande porte, Nickel e Tolle (1989) propõem um algoritmo baseado no problema dual associado ao subproblema quadrático.
Com o objetivo de contornar a possibilidade de se ter subproblemas infactíveis, Burke (1989) propõe um método robusto e estável com propriedades de convergência global. Outros trabalhos combinam idéias de programação quadrática seqüencial, no sentido de usar resoluções aproximadas do sistema linear newtoniano associado às condições de otimalidade com idéias de pontos interiores para restrições de desigualdade. Ver [117], e suas referências.
Apesar da extensa teoria desenvolvida em torno dos métodos principalmente em aspectos relativos a convergência, pouco tem sido feito em termos de experimentos numéricos comparativos. Isto se deve, provavelmente, à diversidade de detalhes próprios da implementação dos diferentes algoritmos existentes, o que os torna pouco comparáveis. O trabalho de Shanno e Phua (1989) é pioneiro neste sentido. Eles comparam o desempenho de um algoritmo geral de de programação quadrática seqüencial combinando diferentes escolhas da atualização secante para a matriz Hessiana do modelo quadrático, diferentes maneiras de estimar os multiplicadores e diferentes funções de mérito. Como conclusões, Shanno e Phua recomendam uma variante do algoritmo básico de Boggs e Tolle (1984) e observam que as experiências numéricas indicam a necessidade de se investir na obtenção de melhores estimativas para os multiplicadores de Lagrange.
O algoritmo no qual nos concentramos neste capítulo, essencialmente introduzido em [98], reúne várias das características desejáveis em bons métodos de programação quadrática seqüencial:

(a) O uso de regiões de confiança, que aumentam a estabilidade dos subproblemas quando é necessário reduzir o raio.

(b) Aplicabilidade a igualdades e desigualdades, através da formulação (12.1.1).

(c) O uso do Lagrangiano aumentado, diferenciável, simples, estável e, provavelmente, livre do efeito Maratos.

(d) Estratégia não monótona para o parâmetro de penalidade.

(e) Os subproblemas não precisam ser resolvidos exatamente, o que viabiliza a aplicabilidade a problemas de grande porte.

(f) Embora não tenha sido destacado neste capítulo, a existência de segundas derivadas de f e h não é necessária. (Em [169] encontramos exemplos de problemas importantes de otimização onde essa característica é relevante.)

Bibliography

[1] J. Abadie e J. Carpentier (1969): Generalization of the Wolfe Reduced Gradient Method to the Case of Nonlinear Constraints, em *Optimization* (R.Fletcher ed.), Academic Press, London e New York, 37-47.

[2] J. Abadie e J. Carpentier (1967): Some numerical experiments with the GRG method for nonlinear programming, Paper HR7422, Electricité de France.

[3] I. Adler, M. Resende, G. Veiga e N. Karmarkar (1989): An implementation of Karmarkar's algorithm for linear programming, *Mathematical Programming* 44, 297-335.

[4] Y. Bard e J. L. Greenstadt (1969): A Modified Newton Method for Optimization with Equality Constraints, em *Optimization* (R. Fletcher ed.), Academic Press, London e New York, 299-306.

[5] E. R. Barnes (1986): A variation of Karmarkar's algorithm for solving linear programming problems, *Mathematical Programming* 36, 174-182.

[6] R. Barrett, M. Berry, T.F. Chan, J. Demmel, J.Donato, J. Dongarra, V. Eijkhout, R. Pozo, Ch. Romine e H. van derVorst (1994): *Templates for the solution of linear systems: buildin gblocks for iterative methods*, SIAM Publications, Philadelphia.

[7] M. C. Bartholomew-Biggs (1987): Recursive quadratic programming methods based on the augmented Lagrangian, *Mathematical Programming Studies* 31, 21-41.

[8] J. Barzilai e J.M. Borwein (1988): Two point step size gradient methods, *IMA Journal of Numerical Analysis* 8, 141-148.

[9] M. S. Bazaraa, J. J. Jarvis e H. D. Sherali (1977): *Linear programming and network flows*, John Wiley and sons, New York, Chichester, Brisbane, Toronto e Singapore.

[10] E. M. L. Beale (1967): Numerical Methods, em *Nonlinear Programming* (J. Abadie ed.), North-Holland, Amsterdam, 132-205.

[11] D.P. Bertsekas (1982): Projected Newton methods for optimization problems with simple constraints, *SIAM Journal on Control and Optimization* 20, 141-148.

[12] R. H. Bielschowsky, A. Friedlander, F. M. Gomes, J. M. Martínez e M. Raydan (1997): An adaptive algorithm for bound constrained quadratic minimization, *Investigación Operativa*, Vol 7, Número 1–2, pp. 67-102.

[13] M. C. Biggs (1972): Constrained minimization using recursive equality quadratic programming, em *Numerical Methods for Nonlinear Optimization* (F. A. Lootsma ed.), Academic Press, London e New York, 411-428.

[14] M. C. Biggs (1974): *The Development of a Class of Constrained Optimization Algorithms and their Application to the Problem of Electric Power Scheduling*, Ph. D. Thesis, University of London.

[15] M. C. Biggs (1975): Constrained minimization using recursive quadratic programming: some alternative subproblem formulation, em *Towards Global Optimization* (L. C. W. Dixon e G. P. Szegö, eds.), North-Holland, Amsterdam, 341-349.

[16] A. Björck (1987): Least squares methods, em *Handbook of Numerical Analysis, Vol 1* (P. G. Ciarlet e J. L. Lions, eds.), Elsevier-North Holland, Amsterdam.

[17] P. T. Boggs e J. W. Tolle (1984): A family of descent functions for constrained optimization, *SIAM Journal on Numerical Analysis* 21, 1146-1161.

[18] P. T. Boggs e J. W. Tolle (1985): An efficient strategy for utilizing a merit function in nonlinear programming algorithms, Report 85-5, Department of Operations Research and Systems Analysis, University of North Carolina.

[19] P. T. Boggs, J. W. Tolle e P. Wang (1982): On the local convergence of quasi-Newton methods for constrained optimization, *SIAM Journal on Control and Optimization* 20, 161-171.

[20] J. F. Bonnans, E. R. Panier, A. L. Tits e J. L. Zhou (1992): Avoiding the Maratos effect by means of a nonmonotone line search II. Inequality constrained problems - feasible iterates, *SIAM Journal on Numerical Analysis* 29, 1187-1202.

[21] M. J. Box, D. Davies e W. H. Swann (1970): *Nonlinear Optimization Techniques*, Chemical Industries Monograph 5, Oliver and Boyd, Edinburgh.

[22] P. N. Brown e Y. Saad (1990): Hybrid Krylov methods for nonlinear systems of equations, *SIAM Journal on Scientific and Statistical Computing* 11, 450-481.

[23] P. N. Brown e Y. Saad (1994): Convergence theory of nonlinear Newton-Krylov methods, *SIAM Journal on Optimization* 4, 297-330.

[24] C. G. Broyden (1965): A class of methods for solving nonlinear simultaneous equations, *Mathematics of Computation* 19, 577-593.

[25] C. G. Broyden, J. E. Dennis e J. J. Moré (1973): On the local and superlinear convergence of quasi-Newton methods, *Journal of the Institute of Mathematics and its Applications* 12, 223-245.

[26] J. R. Bunch e B. N. Parlett (1971): Direct methods for solving symmetric indefinite systems of linear equations, *SIAM Journal on Numerical Analysis* 8, 639-655.

[27] J. V. Burke (1989): A Sequential Quadratic Programming Method for Potentially Infeasible Mathematical Problems, *Journal of Mathematical Analysis and Applications* 139, 319-351.

[28] J. V. Burke e J. J. Moré, On the identification of active constraints, *SIAM Journal on Numerical Analysis* 25, (1988) 1197-1211.

[29] M. D. Canon, C. D. Culllum e E. Polak (1970): *Theory of Optimal Control and Mathematical Programming*, McGraw Hill, New York.

[30] M. R. Celis, J. E. Dennis e R. A. Tapia (1985): A trust region strategy for nonlinear equality constrained optimization, em *Numerical Optimization 1984*, (P. T. Boggs, R. H. Byrd e R. B. Schnabel eds.), SIAM, Philadelphia, 71-82.

[31] R. M. Chamberlain (1979): Some examples of cycling in variable metric methods for constrained minimization, *Mathematical Programming* 16, 378-383.

[32] R. M. Chamberlain, C. Lemaréchal, H. C. Pederson e M. J. D. Powell (1980): The watchdog technique for forcing convergence in algorithms for constrained optimization, Report DAMTP 80/NA1, University of Cambridge.

[33] I. E. Chambouleyron, J. M. Martínez, A. C. Moretti e M. Mulato (1997): The retrieval of the optical constants and the thickness of thin films from transmission spectra, *Applied Optics* 36, pp. 8238-8247.

[34] V. Chvatal (1980): *Linear programming*, W. H. Freeman and Company, New York / San Francisco.

[35] T. F. Coleman (1979): *A Superlinear Penalty Function Method to Solve the Nonlinear Programming Problem*, Ph. D. Thesis, University of Waterloo.

[36] T. F. Coleman e A. R. Conn (1982): Nonlinear programming via an exact penalty function method: asymptotic analysis, *Mathematical Programming* 24, 123-136.

[37] T. F. Coleman e A. R. Conn (1982): Nonlinear programming via an exact penalty function method: global analysis, *Mathematical Programming* 24, 137-161.

[38] T. F. Coleman e A. R. Conn (1984): On the local convergence of a quasi-Newton method for the nonlinear programming problem, *SIAM Journal on Numerical Analysis* 21, 755-769.

[39] T. F. Coleman e L. A. Hulbert (1989): A direct active set algorithm for large sparse quadratic programs with simple bounds, *Mathematical Programming* 45, 373-406.

[40] A. R. Conn, N. I. M. Gould, A. Sartenaer e Ph. L. Toint (1993): Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints, *SIAM Journal on Optimization* 3, 164-221.

[41] A. R. Conn, N. I. M. Gould e Ph. L. Toint (1988): Global convergence of a class of trust region algorithms for optimization with simple bounds, *SIAM Journal on Numerical Analysis* 25 433 - 460. Ver, também, *SIAM Journal on Numerical Analysis* 26 (1989) 764 - 767.

[42] A. R. Conn, N. I. M. Gould e Ph. L. Toint (1989): Testing a class of methods for solving minimization problems with simple bounds on the variables, *Mathematics of Computation* 50, 399 - 430.

[43] A. R. Conn, N. I. M. Gould e Ph. L. Toint (1992): *LANCELOT: a Fortran package for large-scale nonlinear optimization (release A)*, Springer Verlag, Berlin and New York.

[44] A. R. Conn, N. I. M. Gould e Ph. L. Toint (1991): A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds, *SIAM Journal on Numerical Analysis* 28, 545 - 572.

[45] L. Contesse e J. Villavicencio (1982): Resolución de un modelo económico de despacho de carga eléctrica mediante el método de penalización Lagrangeana con cotas, *Revista del Instituto Chileno de Investigación Operativa* 1982, 80-112.

[46] G. B. Dantzig (1963): *Linear programming and extensions*, Princeton University Press, Princeton, NJ.

[47] W. C. Davidon (1959): Variable metric method for minimization, AEC Research and Development Report ANL (Argonne National Laboratory) - 5990.

[48] R. S. Dembo, S. C. Eisenstat e T. Steihaug (1982): Inexact Newton methods, *SIAM Journal on Numerical Analysis* 19, 400–408.

[49] R. S. Dembo e U. Tulowitzki (1985): Sequential Truncated Quadratic Programming Methods, em *Numerical Optimization 1984* (P. T. Boggs, R. H. Byrd e R. B. Schnabel eds.), SIAM, Philadelphia, 83-101.

[50] J. E. Dennis, N. Echebest, M. Guardarucci, J. M. Martínez, H. D. Scolnik e C. Vacino (1991): A Curvilinear Search Using Tridiagonal Secant Updates for Unconstrained Optimization, *SIAM Journal on Optimization* 1, 352-372.

[51] J. E. Dennis, M. El-Alem e M. C. Maciel (1995): A global convergence theory for general trust-region-based algorithms for equality constrained optimization, por aparecer em *SIAM Journal on Optimization.*

[52] J. E. Dennis e J. J. Moré (1974): A Characterization of Superlinear Convergence and its Application to Quasi-Newton Methods, *Mathematics of Computation* 28, 546-560.

[53] J. E. Dennis e R.B. Schnabel (1979): Least change secant updates for quasi-Newton methods, *SIAM Review* 21, 443-459.

[54] J. E. Dennis e Schnabel (1983): *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs.

[55] J. E. Dennis e H. F. Walker (1981): Convergence theorems for least-change secant update methods, *SIAM Journal on Numerical Analysis* 18, 949-987.

[56] P. Deuflhard (1991): Global inexact Newton methods for very large scale nonlinear problems, *Impact of Computing in Science and Engineering* 3, 366–393.

[57] P. Deuflhard, R. Freund A. Walter (1990): Fast secant methods for the iterative solution of large nonsymmetric linear systems, *Impact of Computing in Science and Engineering* 2, 244-276.

[58] I. I. Dikin (1967): Iterative solution of problems of linear and quadratic programming, *Soviet Math. Dokl.* 8, 674-675.

[59] M. A. Diniz - Ehrhardt e J. M. Martínez (1993): A parallel projection method for overdetermined nonlinear systems of equations, *Numerical Algorithms* 4, 241-262.

[60] G. Di Pillo e L. Grippo (1979): A new class of augmented Lagrangians in nonlinear programming, *SIAM Journal on Control and Optimization* 17, 618-628.

[61] L. C. W. Dixon (1979): Exact penalty functions in nonlinear programming, Report 103, Numerical Optimisation Centre, Hatfield Polytechnique.

[62] I. S. Duff, A. M. Erisman e J. K. Reid (1986): *Direct methods for sparse matrices*, Clarendon Press, Oxford.

[63] S. C. Eisenstat e H. F. Walker (1994): Globally convergent inexact Newton methods, por aparecer em *SIAM Journal on Optimization*.

[64] M. El-Alem (1992): A robust trust region algorithm with a non-monotone penalty parameter scheme for constrained optimization, Technical Report 92-30, Department of Computational and Applied Mathematics, Rice University, Houston.

[65] A. V. Fiacco (1983): *Introduction to sensitivity and stability analysis in nonlinear programming*, Academic Press, New York.

[66] A. V. Fiacco e G. P. McCormick (1990): *Nonlinear programming: sequential unconstrained minimization techniques*, SIAM Publications, Philadelphia.

[67] R. Fletcher (1970): A class of methods for nonlinear programming with termination and convergence properties, em *Integer and Nonlinear Programming* (J. Abadie, ed.), North Holland, Amsterdam, 157-175.

[68] R. Fletcher (1974): Methods Related to Lagrangian Functions, em *Numerical Methods for Constrained Optimization* (P.E.Gill e W.Murray eds.), 235-239.

[69] R. Fletcher (1987): *Practical methods for optimization*, John Wiley & Sons, Chichester.

[70] R. Fletcher e M. J. D. Powell (1963): A rapidly convergent descent method for minimization *Computer Journal* 6, 163-168.

[71] R. Fontecilla (1983): *A general convergence theory for quasi-Newton methods for constrained optimization*, Ph. D. Dissertation, Mathematical Sciences Department, Rice University, Houston.

[72] R. Fontecilla (1985): On inexact quasi-Newton methods for constrained optimization, em *Numerical Optimization 1984* (P. T. Boggs, R. H. Byrd e R. B. Schnabel eds.), 102-118, SIAM, Philadelphia.

[73] R. Fontecilla (1990): Inexact secant methods for nonlinear constrained optimization, *SIAM Journal on Numerical Analysis* 27, 154-165.

[74] A. Forsgren e W. Murray (1993): Newton methods for large-scale linear equality constrained minimization, *SIAM Journal on Matrix Analysis and applications* 14, 560-587.

[75] A. Forsgren e W. Murray (1995): Newton methods for large-scale linear inequality constrained minimization, Technical Report, Systems Optimization Laboratory, Stanford University.

[76] A. Friedlander, M. A. Gomes-Ruggiero, D. N. Kozakevich, J.M. Martínez e S.A. Santos (1995): Solving nonlinear systems of equations by means of quasi-Newton methods with a nonmonotone strategy, Relatório técnico, Instituto de Matemática, Universidade Estadual de Campinas.

[77] A. Friedlander, C. Lyra, H. M. Tavares e E. L. Medina (1989): Optimization with staircase structure – An application to generation scheduling, *Computers and Operations Research* 17, 143-152.

[78] A. Friedlander e J.M. Martínez (1989): On the numerical solution of bound constrained optimization problems, *RAIRO Operations Research* 23, 319-341.

[79] A. Friedlander e J.M. Martínez (1994): On the maximization of a concave quadratic function with box constraints, *SIAM Journal on Optimization* 4, 177-192.

[80] A. Friedlander, J.M. Martínez, B. Molina e M. Raydan (1994): "Gradient methods with retards", por aparecer em *SIAM Journal on Numerical Analysis.*

[81] A. Friedlander, J.M. Martínez e M. Raydan (1995): A new method for large-scale box constrained quadratic minimization problems, *Optimization Methods and Software* 5, pp. 57-74.

[82] A. Friedlander, J.M. Martínez e S.A. Santos (1994): A new trust region algorithm for bound constrained minimization, *Applied Mathematics and Optimization* 30, 235-266.

[83] A. Friedlander, J. M. Martínez e S. A. Santos (1994): On the resolution of linearly constrained convex minimization problems, *SIAM Journal on Optimization* 4, 331-339.

[84] A. Friedlander, J. M. Martínez e S. A. Santos (1995): Resolution of linear complementarity problems using minimization with simple bounds, *Journal of Global Optimization* 6, pp. 1-15.

[85] A. Friedlander, J. M. Martínez e S. A. Santos (1995): A new strategy for solving variational inequalities in bounded polytopes, *Numerical Functional Analysis and Optimization* 16, pp. 653-668.

[86] U. M. García-Palomares e O. L. Mangasarian (1976): Superlinearly convergent quasi-Newton algorithms for nonlinearly constrained optimization problems, *Mathematical Programming* 11, 1-13.

[87] D. M. Gay (1979): Some convergence properties of Broyden's method, *SIAM Journal on Numerical Analysis* 16, 623 - 630.

[88] D. M. Gay (1981): Computing optimal locally constrained steps, *SIAM Journal on Scientific and Statistical Computing* 2, 186-197.

[89] D. M. Gay (1984): A trust-region approach to linearly constrained optimization, em *Numerical Analysis* (D. F. Griffiths, ed.), Lecture Notes in Mathematics 1066, Springer-Verlag, Berlin, Heidelberg, New York, 72-105.

[90] A. George e E. Ng (1987): Symbolic factorization for sparse Gaussian elimination with partial pivoting, *SIAM Journal on Scientific and Statistical Computing* 8, 877-898.

[91] F. Giannessi (1994): General optimality conditions via a separation scheme, em *Algorithms for continuous optimization* (E. Spedicato, ed.), Kluwer Academic Publishers, The Netherlands, 1-23.

[92] P. E. Gill, W. Murray (1974): Newton-type methods for unconstrained and linearly constrained optimization, *Mathematical Programming* 7, 311-350.

[93] P. E. Gill, W. Murray, M. A. Saunders e M. H. Wright (1992): Some theoretical properties of an augmented Lagrangian function, em *Advances in Optimization and Parallel Computing* (P. E. Pardalos ed.), 127-143, Elsevier, Amsterdam.

[94] P. E. Gill, W. Murray, M. A. Saunders e M. H. Wright (1991): Inertia-controlling methods for general quadratic programming, *SIAM Review* 33, 1-36.

[95] P. E. Gill, W. Murray e M. H. Wright (1981): *Practical Optimization*, Academic Press, London e New York.

[96] G. H. Golub e Ch. F. Van Loan (1989): *Matrix Computations*, The Johns Hopkins University Press, Baltimore and London.

[97] D. Goldfarb (1969): Extensions of Davidon's variable metric method to maximization under linear inequality and equality constraints, *SIAM Journal on Applied Mathematics* 17, 739-764.

[98] F. M. Gomes, M. C. Maciel e J. M. Martínez (1995): Successive quadratic programming for minimization with equality and inequality constraints using trust regions, augmented Lagrangians and nonmonotone penalty parameters, por aparecer em *Mathematical Programming*.

[99] H. S. Gomes e J. M. Martínez (1991): "A Numerically Stable Reduced-Gradient Type Algorithm for Solving Large-Scale Linearly Constrained Minimization Problems", *Computers and Operations Research*, 18, 17–31.

[100] M. A. Gomes–Ruggiero e J. M. Martínez (1992): The Column–Updating Method for solving nonlinear equations in Hilbert space, *RAIRO Mathematical Modelling and Numerical Analysis* 26, 309-330.

[101] M. A. Gomes–Ruggiero, J. M. Martínez e A. C. Moretti (1992): Comparing algorithms for solving sparse nonlinear systems of equations, *SIAM Journal on Scientific and Statistical Computing* 13, 459 - 483.

[102] C. C. Gonzaga (1989): *Algoritmos de pontos interiores para programação linear*, 17° Colóquio Brasileiro de Matemática, IMPA, CNPq, Rio de Janeiro.

[103] C. C. Gonzaga (1992): Path-following methods for linear programming, *SIAM Review* 34, 167-224.

[104] N. I. M. Gould (1986): On the accurate determination of search directions for simple differentiable penalty functions, *IMA Journal of Numerical Analysis* 6, 357-372.

[105] A. Griewank (1992): Achieving Logarithmic Growth of Temporal and Spacial Complexity in Reverse Automatic Differentiation, *Optimization Methods and Software* 1, 35 - 54.

[106] A. Griewank e Ph. L. Toint (1982): On the unconstrained optimization of partially separable functions, in *Nonlinear Optimization 1981* (M.J.D. Powell, ed.), Academic Press, New York.

[107] A. Griewank e Ph. L. Toint (1982): Partitioned variable metric for large structured optimization problems, *Numerische Mathematik* 39, 119 - 137.

[108] A. Griewank e Ph. L. Toint (1982): Local convergence analysis for partitioned quasi-Newton updates, *Numerische Mathematik* 39, 429-448.

[109] A. Griewank e Ph. L. Toint (1984): Numerical experiments with partially separable optimization problems, in *Numerical Analysis Proceedings Dundee 1983*, edited by D.F. Griffiths, Lecture Notes in Mathematics vol. 1066, Springer - Verlag, Berlin, 203-220.

[110] L. Grippo, F. Lampariello e S. Lucidi (1986): A nonmonotone line search technique for Newton's method, *SIAM Journal on Numerical Analysis* 23, 707 - 716.

[111] S-P. Han (1976): Superlinearly convergent variable metric algorithms for general nonlinear programming problems, *Mathematical Programming* 11, 263-282.

[112] S-P. Han (1977): Dual variable metric algorithms for constrained optimization, *SIAM Journal on Control and Optimization* 15, 546-565.

[113] S-P. Han (1977): A globally convergent method for nonlinear programming, *Journal of Optimization Theory and Applications* 22, 297-310.

[114] E. R. Hansen (1979): Global optimization using interval analysis: the one-dimensional case, *Journal of Optimization Theory and Applications* 29, 331-344.

[115] M. D. Hebden (1973): An algorithm for minimization using exact second derivatives, Atomic Energy Research Establishment Report TP 515, Harwell, Inglaterra.

[116] M. Heinkenschloss (1993): Mesh independence for nonlinear least squares problems with norm constraints, *SIAM Journal on Optimization* 3, 81-117.

[117] J. Herskovits (1986): A two-stage feasible directions algorithm for nonlinearly constrained optimization, *Mathematical Programming* 36, 19-38.

[118] M. R. Hestenes (1969): Multiplier and gradient methods, *Journal of Optimization Theory and Applications* 4, 303-320.

[119] M.R. Hestenes e E. Stiefel (1952): Methods of conjugate gradients for solving linear systems, *Journal of Research of the National Bureau of Standards* B 49, 409-436.

[120] D. M. Himmelblau (1972): *Applied Nonlinear Programming*, Mc Graw-Hill, New York.

[121] W. Hock e K. Schittkowski (1981): *Test examples for nonlinear programming codes*, Lecture Notes in Economics and Mathematical Systems 187, Springer Verlag, Berlin.

[122] H. Y. Huang (1970): Unified approach to quadratically convergent algorithms for function minimization, *Journal of Optimization Theory and Applications* 5, 405-423.

[123] G. W. Johnson e N. H. Austria (1983): A quasi-Newton method employing direct secant updates of matrix factorizations, *SIAM Journal on Numerical Analysis* 20, 315-325.

[124] N. Karmarkar (1984): A new polynomial-time algorithm for linear programming, *Combinatorica* 4, 373-395.

[125] C. T. Kelley e E. W. Sachs (1987): A quasi-Newton method for elliptic boundary value problems, *SIAM Journal on Numerical Analysis* 24, 516 - 531.

[126] L. S. Lasdon (1982), Reduced gradient methods, in *Nonlinear Optimization 1981* (M. J. D. Powell, ed.), Academic Press, New York, 235-242.

[127] L. Lasdon (1985): Nonlinear programming algorithms - applications, software and comparisons, em *Numerical Optimization 1984* (P. T. Boggs, R. H. Byrd e R.B.Schnabel eds.), SIAM Publications, Philadelphia, 41-70.

[128] L. S. Lasdon, A. D. Warren, A. Jain e M. Ratner (1978): Design and testing of a generalized reduced gradient code for nonlinear programming, *ACM Transactions on Mathematical Software* 4, 34-50.

[129] D. Luenberger (1986): *Linear and nonlinear programming*, Addison-Wesley, New York.

[130] L. Lukšan (1994): Inexact trust region method for large sparse systems of nonlinear equations, por aparecer em *Journal of Optimization Theory and Applications.*

[131] N. Maratos (1978): *Exact penalty function algorithms for finite-dimensional and control optimization problems*, Ph. D. Thesis, University of London, England.

[132] J. M. Martínez (1984): A quasi–Newton method with modification of one column per iteration, *Computing* 33, 353–362.

[133] J. M. Martínez (1990): A family of quasi-Newton methods for nonlinear equations with direct secant updates of matrix factorizations, *SIAM Journal on Numerical Analysis* 27, 1034-1049.

[134] J. M. Martínez (1990): Local convergence theory of inexact Newton methods based on structured least change updates, *Mathematics of Computation* 55, 143-168.

[135] J. M. Martínez (1992): On the relation between two local convergence theories of least change secant update methods, *Mathematics of Computation* 59, 457–481.

[136] J. M. Martínez (1992): Fixed-Point Quasi-Newton methods, *SIAM Journal on Numerical Analysis* 29, 1413–1434.

[137] J. M. Martínez (1993): A theory of secant preconditioners, *Mathematics of Computation* 60, 681–698.

[138] J. M. Martínez (1995): An extension of the theory of secant preconditioners, *Journal of Computational and Applied Mathematics* 60, pp. 115-125.

[139] J. M. Martínez (1993): On the Convergence of the Column-Updating Method, *Matemática Aplicada e Computacional* 12, 83-94.

[140] J. M. Martínez (1994): Local minimizers of quadratic functions on Euclidean balls and spheres, *SIAM Journal on Optimization* 4, 159 -176.

[141] J. M. Martínez e L. Qi (1995): Inexact Newton methods for solving nonsmooth equations, *Journal of Computational and Applied Mathematics* 60, pp. 127-145.

[142] J. M. Martínez e S. A. Santos (1995): A trust region strategy for minimization on arbitrary domains, *Mathematical Programming* 68, pp. 267-302.

[143] J. M. Martínez e L. T. Santos (1998): Some new theoretical results on recursive quadratic programming algorithms, *Journal of Optimization Theory and Applications* 97, pp. 435-454.

[144] J. M. Martínez e S. A. Santos (1997): Convergence results on an algorithm for norm constrained regularization and related problems, *RAIRO Operations Research* 31, pp. 269-294.

[145] G. P. McCormick (1983): *Nonlinear programming; theory, algorithms and applications*, John Wiley & Sons, New York.

[146] R. E. Moore (1991): Global optimization to prescribed accuracy, *Computers and Mathematics with Applications* 21, 25-39.

[147] J. J. Moré (1983): Recent developments in algorithms and software for trust region methods, in *Mathematical Programming Bonn 1982 - The state of art*, (A. Bachem, M. Grötschel e B. Korte, eds.), 258-287, Mathematical Programming Society, Bonn.

[148] J. J. Moré e D. C. Sorensen (1983): Computing a trust region step, *SIAM Journal on Scientific and Statistical Computing* 4, 553-572.

[149] J.J. Moré e G. Toraldo (1991): On the solution of large quadratic programming problems with bound constraints, *SIAM Journal on Optimization* 1, 93-113.

[150] W. Murray (1969): *Constrained Optimization*, Ph. D. Thesis, University of London.

[151] W. Murray (1969): An algorithm for constrained minimization, em *Optimization* (R. Fletcher, ed.), Academic Press, London e New York, 247-258.

[152] W. Murray e M. W. Wright (1978): Projected Lagrangian methods based on the trajectories of penalty and barrier functions, Report SOL 78-23, Department of Operations Research, Stanford University.

[153] W. Murray e M. H. Wright (1980): Computation of the search direction in constrained optimization algorithms, Report SOL 80-2, Department of Operations Research, Stanford University.

[154] R. B. Murtagh e M. A. Saunders (1977): MINOS User's Guide, Report SOL 77-9, Department of Operations Research, Stanford University.

[155] R. B. Murtagh e M. A. Saunders (1978): Large-scale linearly constrained optimization, *Mathematical Programming* 14, 41-72.

[156] R. H. Nickel e J. W. Tolle (1989): A Sparse Sequential Quadratic Programming Algorithm, *Journal of Optimization Theory and Applications* 60, 453-473.

[157] J. Nocedal (1993): Theory of algorithms for unconstrained optimization, *Acta Numerica* 1, 199-242.

[158] J. Nocedal e M. L. Overton (1985): Projected Hessian Updating Algorithms for Nonlinearly Constrained Optimization, *SIAM Journal on Numerical Analysis* 22, 821-850.

[159] J. M. Ortega e W. G. Rheinboldt (1970): *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

[160] E. R. Panier e A. L. Tits (1991): Avoiding the Maratos effect by means of a nonmonotone line search I. General constrained problems, *SIAM Journal on Numerical Analysis* 28, 1183-1195.

[161] M. J. D. Powell (1969): A method for nonlinear constraints in minimization problems, em *Optimization*, (R. Fletcher, ed.), Academic Press, London and New York, 283-298.

[162] M. J. D. Powell (1970): A hybrid method for nonlinear equations, em *Numerical methods for nonlinear algebraic equations* (P. Rabinovitz, ed.), Gordon and Breach, New York, 87-114.

[163] M. J. D. Powell (1977): A fast algorithm for nonlinearly constrained optimization calculations, em Numerical Analysis (Proceedings, Dundee 1977) (G. A. Watson ed.), *Lecture Notes in Mathematics* 630, Springer-Verlag, Berlin, Heidelberg and New York, 144-157.

[164] M. J. D. Powell (1978): The convergence of variable metric methods for nonlinearly constrained optimization calculations, em *Nonlinear*

Programming 3 (O. L. Mangasarian, R. R. Meyer e S. M. Robinson, eds.), Academic Press, London and New York, 27-63.

[165] M. J. D. Powell (1985): How bad are the BFGS and the DFP method when the objective function is quadratic?, University of Cambridge, DAMTP Report 85/NA4.

[166] M. J. D. Powell (1978): Algorithms for nonlinear constraints that use Lagrangian functions, *Mathematical Programming* 14, 224-248.

[167] M. J. D. Powell (1985): The performance of two subroutines for constrained optimization on some difficult test problems, em *Numerical Optimization 1984* (P. T. Boggs, R. H. Byrd e R. B. Schnabel, eds.), SIAM, Philadelphia, 160-177.

[168] M. J. D. Powell e Y. Yuan (1986): A recursive quadratic programming algorithm that uses differentiable exact penalty functions, *Mathematical Programming* 35, 265-278.

[169] L. Qi (1995): Superlinearly convergent approximate Newton methods for LC^1 optimization problems, por aparecer em *Mathematical Programming.*

[170] M. Raydan (1993): On the Barzilai and Borwein choice of steplength for the gradient method, *IMA Journal of Numerical Analysis* 13, 321-326.

[171] C. H. Reinsch (1971): Smoothing by spline functions II, *Numerische Mathematik* 16, 451-454.

[172] K. Ritter (1973): A superlinearly convergent method for minimization problems with linear inequality constraints, *Mathematical Programming* 4, 44-71.

[173] K. Ritter (1975): A method of conjugate direction for linearly constrained nonlinear programming problems, *SIAM Journal on Numerical Analysis* 12, 274-303.

[174] K. Ritter (1980): Convergence and superlinear convergence of algorithms for linearly constrained minimization problems, em *Nonlinear Optimization: Theory and Algorithms, Part II* (L. C. W. Dixon, E. Spedicato e G. P. Szegö, editores), Birkhäuser, Boston, Massachusetts, 221-251.

[175] R. T. Rockafellar (1974): Augmented Lagrange multiplier functions and duality in nonconvex programming, *Siam Journal on Control* 12, 268-285.

[176] R. T. Rockafellar (1990): Computational schemes for solving large scale problems in extended linear-quadratic programming, *Mathematical Programming* 48, 447-474.

[177] R. T. Rockafellar e R. J. Wets (1990): Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time, *SIAM Journal on Control and Optimization* 28, 810-822.

[178] J. B. Rosen (1960): The gradient projection method for nonlinear programming, part I. Linear constraints, *SIAM Journal on Applied Mathematics* 9, 181-217.

[179] Y. Saad e M. H. Schultz (1986): GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM Journal on Numerical Analysis* 7, 856-869.

[180] S. A. Santos e D. C. Sorensen (1995): A new matrix-free algorithm for the large-scale trust-region subproblem, em preparação.

[181] R. W. H. Sargent (1974): Reduced-gradient and Projection Methods for Nonlinear Programming, em *Numerical Methods for Constrained Optimization* (P.E.Gill and W.Murray eds.), Academic Press, New York, 149-174.

[182] K. Schittkowski (1980): Nonlinear Programming Codes, *Lecture Notes in Economics and Mathematical Systems* 183, Springer-Verlag, Berlin, Heidelberg and New York.

[183] K. Schittkowski (1981): The nonlinear programming method of Wilson, Han and Powell with an augmented Lagrangian type line search function, *Numerische Mathematik* 38, 83-114.

[184] V. E. Shamanski (1967): A modification of Newton's method, *Ukrain Mat. Z.*, 19, 133-138.

[185] D. F. Shanno e K. H. Phua (1989): Numerical Experience with Sequential Quadratic Programming Algorithms for Equality Constrained Nonlinear Programming, *ACM Transactions on Mathematical Software* 15, 49-63.

[186] S. Smith e L. Lasdon (1992): Solving large sparse nonlinear programs using GRG, *ORSA Journal on Computing* 4, 1-15.

[187] D. C. Sorensen (1982): Newton's method with a model trust region modification, *SIAM Journal on Numerical Analysis* 19, 409-426.

[188] D. C. Sorensen (1994): Minimization of a large scale quadratic function subject to an ellipsoidal constraint, Technical Report, Department of Computational and Applied Mathematics, Rice University.

[189] R. Swanepoel (1983): Determination of the thickness and optical constants of amorphous silicon, *J. Phys. E: Sci. Instrum.* 16, 1214-1222.

[190] R. A. Tapia (1977): Diagonalized multiplier methods and quasi-Newton methods for constrained optimization, *Journal of Optimization Theory and Applications* 22, 135-194.

[191] A. N. Tikhonov e V. Y. Arsenin (1977): *Solutions of ill-posed problems*, John Wiley and Sons, New York and Toronto, 1977.

[192] Ph. L. Toint (1986): Numerical solution of large sets of algebraic nonlinear equations, *Mathematics of Computation* 16, 175 - 189.

[193] R. J. Vanderbei, M. J. Meketon e B. A. Freedman (1986): A modification of Karmarkar's linear programming algorithm, *Algorithmica* 1, 395-407.

[194] S. A. Vavasis (1991): *Nonlinear optimization*, Oxford University Press, Oxford.

[195] C. R. Vogel (1990): A constrained least-squares regularization method for nonlinear ill-posed problems, *SIAM Journal on Control and Optimization* 28, 34-49.

[196] H. F. Walker (1992): A GMRES-backtracking Newton iterative method, *Proceedings of the Copper Mountain Conference on Iterative Methods, April 1992.*

[197] D. J. Wilde e C. S. Beightler (1967): *Foundations of Optimization*, Prentice Hall, Englewood Cliffs, N.J.

[198] R. B. Wilson (1963): *A Simplicial Algorithm for Concave Programming*, Ph. D. Dissertation, Harvard University, Graduate School of Business Administration.

[199] P. Wolfe (1961): A duality theorem for non-linear programming, *Quarterly of Applied Mathematics* 19, 239-244.

[200] P. Wolfe (1963): Methods of nonlinear programming, em *Recent Advances in Mathematical Programming* (R.L.Graves e P. Wolfe eds.), McGraw Hill, New York, 67-86.

[201] H. Wolkowicz (1994): A semidefinite framework for trust region subproblems with applications to large scale minimization, Technical Report, University of Waterloo.

[202] M. H. Wright (1976): *Numerical Methods for Nonlinearly Constrained Optimization*, Ph. D. Thesis, Stanford University, California.

[203] A. Xavier (1993): *Penalização hiperbólica*, Tese de Doutorado, COPPE, Universidade Federal do Rio de Janeiro.

[204] H. Yabe, N. Yamaki e S. Takahashi (1991): Global convergence of sequential inexact QP method for constrained optimization, *SUT Journal of Mathematics* 27, 127-138.

[205] T. J. Ypma (1984): Local convergence of inexact Newton methods, *SIAM Journal on Numerical Analysis* 21, 583-590.

[206] T. J. Ypma (1993): On the history of Newton's method, Contributed paper at the Workshop on linear and nonlinear iterative methods and verification of solution, Matsuyama, Japão, julho 1993.

[207] C. Zhu e R. T. Rockafellar (1995): Primal-dual projected gradient algorithm for extended linear-quadratic programming, por aparecer em *SIAM Journal on Optimization*.