

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 sequência e, além disso, numa vizinhança de x_* não existe nenhum outro ponto onde se anule o gradiente, então a sequência 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 sequência 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 consequência, 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 $x_k \in \mathbb{R}^n$,

- (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 $\mu \rightarrow \infty$ 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 $f \in C^3(\mathbb{R}^n)$ (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 $\|x_k - x_*\| \leq \varepsilon$, então

$$f(x_k + d_k) \leq f(x_k) + \alpha g(x_k)^T d_k, \quad (6.2.1)$$

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 $0 < \|x - x_*\| \leq \varepsilon_0$. Vejamos primeiro que

$$\lim_{k \rightarrow \infty} x_k = x_*, \quad (6.2.2)$$

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 $\|\nabla^2 f(x)^{-1}\| \leq M$ sempre que $\|x - x_*\| \leq \varepsilon_1$. Portanto, quando $\|x_k - x_*\| \leq \varepsilon_1$, temos $\|B_k^{-1}\| \leq M$ e

$$\|x_{k+1} - x_k\| \leq \|d_k\| \leq \|B_k^{-1}\| \|g(x_k)\| \leq M \|g(x_k)\|. \quad (6.2.3)$$

Portanto, pela continuidade de $g(x)$, existe $\varepsilon_2 \leq \frac{\varepsilon_1}{2}$ tal que

$$\|x_{k+1} - x_k\| \leq \frac{\varepsilon_1}{2} \text{ sempre que } \|x_k - x_*\| \leq \varepsilon_2. \quad (6.2.4)$$

Agora, f é contínua na coroa $\varepsilon_2 \leq \|x - x_*\| \leq \varepsilon_1$. Portanto, atinge um valor mínimo m em algum ponto dessa região. Pela suposição feita sobre ε_0 , temos que $m > f(x_*)$. Definimos

$$V = \{x \in \mathbb{R}^n \mid \|x - x_*\| < \varepsilon_2 \text{ e } f(x) < m\}. \quad (6.2.5)$$

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),

$$\|x_{k_0+1} - x_*\| \leq \|x_{k_0} - x_*\| + \|x_{k_0+1} - x_{k_0}\| \leq \varepsilon_2 + \frac{\varepsilon_1}{2} \leq \varepsilon_1. \quad (6.2.6)$$

Ao mesmo tempo, exceto no caso trivial em que $x_{k_0} = x_*$, que podemos analisar por separado,

$$f(x_{k_0+1}) < f(x_{k_0}) < m. \quad (6.2.7)$$

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 ,

$$f(x_k + d_k) = f(x_k) + g(x_k)^T d_k + \frac{1}{2}(d_k)^T \nabla^2 f(x_k) d_k + r_2(d_k) \quad (6.2.8)$$

onde $\lim_{d_k \rightarrow 0} \frac{r_2(d_k)}{\|d_k\|^2} = 0$.

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,

$$\frac{r_2(d_k)}{\|d_k\|^2} > \left(\frac{1}{2} - \alpha\right) \frac{(d_k)^T \nabla^2 f(x_k) d_k}{(d_k)^T d_k} \geq \left(\frac{1}{2} - \alpha\right) \lambda_1(k) \quad (6.2.9)$$

onde $\lambda_1(k)$ é o menor autovalor de $\nabla^2 f(x_k)$.

Quando $x_k \rightarrow x_*$, $d_k \rightarrow 0$ 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 $\alpha \in (0, \frac{1}{2})$, 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 $x_k \in \mathbb{R}^n$ 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 $f \in C^3(\mathbb{R}^n)$, 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

$$\nabla^2 f(x_k) d_k = -g(x_k) \quad (6.3.1)$$