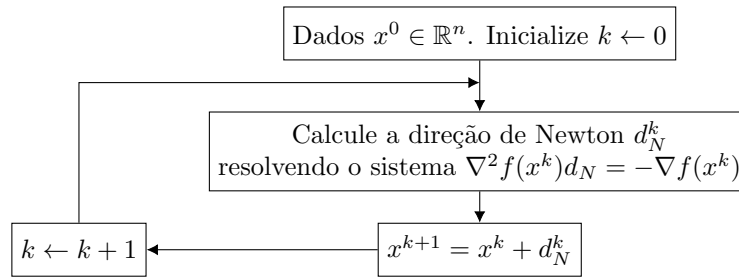


Capítulo 12

Método de Newton



Algoritmo 12.1: Método de Newton.

12.1 Convergência local

Dada uma matriz quadrada A , lembre-se que a *norma de A* é dada por

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$

onde $\|Ax\|$ é norma usual do vetor Ax . Neste caso dizemos que a norma matricial $\|\cdot\|$ é induzida pela norma correspondente $\|\cdot\|$ de vetores. Nesta seção, $\|A\|$ indicará a norma matricial induzida pela norma euclidiana de vetores.

Exercício: Mostre que

1. $\|\cdot\|$ é de fato uma norma no espaço das matrizes, isto é, $\|A\| = 0 \Leftrightarrow A = 0$, $\|\lambda A\| = |\lambda|\|A\|$ e $\|A + B\| \leq \|A\| + \|B\|$;
2. $\|Ax\| \leq \|A\|\|x\|$ (A matriz, x vetor) e $\|AB\| \leq \|A\|\|B\|$ (A, B matrizes);
3. $\|I\| = 1$, onde I é matriz identidade.

Lema 12.1 (Perturbação da identidade). *Se F é uma matriz quadrada tal que $\|F\| < 1$, então $I - F$ é não singular e $\|(I - F)^{-1}\| \leq \frac{1}{1 - \|F\|}$.*

Demonstração. Veja o Lema 2.3.3 de [2]. □

O próximo lema, retirado de [3, Lema 5.4.1], diz que, dada uma matriz não singular A_* , qualquer matriz suficientemente próxima é também não singular, e a norma da sua inversa pode ser majorada pela norma da inversa de A_* . Ele será útil na prova de convergência do método de Newton, dado que temos que lidar com a inversa da hessiana de f próximo à solução.

Lema 12.2. *Seja A_* matriz quadrada não singular. Se A é tal que $\|A - A_*\| \leq \frac{1}{2\|A_*^{-1}\|}$ então A é não singular e $\|A^{-1}\| \leq 2\|A_*^{-1}\|$.*

Demonstração. Note que

$$\|I - AA_*^{-1}\| = \|(A - A_*)A_*^{-1}\| \leq \|A - A_*\| \|A_*^{-1}\| \leq 1/2.$$

Aplicando o Lema 12.1 com $F = I - AA_*^{-1}$ concluímos que $AA_*^{-1} = I - F$ é não singular e

$$\|(AA_*^{-1})^{-1}\| = \|[I - (I - AA_*^{-1})]^{-1}\| \leq \frac{1}{1 - \|I - AA_*^{-1}\|} \leq 2.$$

Como o produto AA_*^{-1} é não singular, A é não singular. Da desigualdade acima segue que

$$\|A^{-1}\| = \|A_*^{-1}A_*A^{-1}\| = \|A_*^{-1}(AA_*^{-1})^{-1}\| \leq \|A_*^{-1}\| \|(AA_*^{-1})^{-1}\| \leq 2\|A_*^{-1}\|,$$

como queríamos demonstrar. \square

Teorema 12.1 (Convergência local do método de Newton). *Suponha que f tenha derivadas de segunda ordem contínuas. Seja x^* tal que $\nabla f(x^*) = 0$ e suponha que $\nabla^2 f(x^*)$ seja não singular. Então existe $\varepsilon > 0$ tal que, se $\|x^0 - x^*\| \leq \varepsilon$, vale:*

1. *A sequência $x^{k+1} = x^k + d_N^k = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$, $k \geq 0$, está bem definida (isto é, o passo de Newton está bem definido para todo $k \geq 0$);*
2. *$\{x^k\}$ converge a x^* com ordem superlinear;*
3. *Se a função $\nabla^2 f$ é Lipschitz contínua, isto é, se existe $L > 0$ tal que*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n,$$

então $\{x^k\}$ converge a x^ com ordem quadrática.*

Demonstração. Como $\nabla^2 f$ é contínua e não singular em x^* , temos $\|\nabla^2 f(x^*)^{-1}\| > 0$ e a existência de um $\varepsilon > 0$ tal que

$$\|x - x^*\| \leq \varepsilon \implies \|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq \frac{1}{4\|\nabla^2 f(x^*)^{-1}\|}. \quad (12.1)$$

Assim, tomando $\|x^0 - x^*\| \leq \varepsilon$, o Lema 12.2 garante que $x^1 = x^0 - (\nabla^2 f(x^0))^{-1} \nabla f(x^0)$ está bem definido. Por hipótese temos $\nabla f(x^*) = 0$, e logo

$$\begin{aligned} \|x^1 - x^*\| &= \|x^0 - x^* - \nabla^2 f(x^0)^{-1} \nabla f(x^0)\| \\ &= \|\nabla^2 f(x^0)^{-1} [-\nabla^2 f(x^0)(x^0 - x^*) + \nabla f(x^0) - \nabla f(x^*)]\| \\ &\leq \|\nabla^2 f(x^0)^{-1}\| \|\nabla f(x^0) - \nabla f(x^*) - \nabla^2 f(x^0)(x^0 - x^*)\|. \end{aligned} \quad (12.2)$$

Consideremos a função φ de uma variável definida por

$$\varphi(t) = \nabla f(tx^0 + (1-t)x^*), \quad t \in [0, 1].$$

Temos

$$\varphi'(t) = \nabla^2 f(tx^0 + (1-t)x^*)(x^0 - x^*).$$

Pelo Teorema do Valor Médio aplicado à φ , existe $\bar{t} \in (0, 1)$ tal que

$$\varphi'(\bar{t}) = \frac{\varphi(1) - \varphi(0)}{1 - 0},$$

o que implica

$$\nabla^2 f(\bar{t}x^0 + (1-\bar{t})x^*)(x^0 - x^*) = \nabla f(x^0) - \nabla f(x^*).$$

Usando essa expressão em (12.2), chegamos a

$$\begin{aligned} \|x^1 - x^*\| &\leq \|\nabla^2 f(x^0)^{-1}\| \|\nabla^2 f(\bar{t}x^0 + (1-\bar{t})x^*)(x^0 - x^*) - \nabla^2 f(x^0)(x^0 - x^*)\| \\ &\leq r_0 \|x^0 - x^*\|, \end{aligned} \quad (12.3)$$

onde

$$r_0 = \|\nabla^2 f(x^0)^{-1}\| \max_{t \in [0,1]} \|\nabla^2 f(tx^0 + (1-t)x^*) - \nabla^2 f(x^0)\|.$$

Note que para $t \in [0, 1]$, temos

$$\|tx^0 + (1-t)x^*\| = (1-t)\|x^0 - x^*\| \leq \|x^0 - x^*\| \leq \varepsilon.$$

Assim, (12.1) e o Lema 12.2 garantem que

$$r_0 \leq 2\|\nabla^2 f(x^*)^{-1}\| \frac{1}{4\|\nabla^2 f(x^*)^{-1}\|} = \frac{1}{2}.$$

Daí, segue de (12.3) que

$$\|x^1 - x^*\| \leq r_0 \|x^0 - x^*\| \leq \varepsilon/2 \leq \varepsilon,$$

e usando novamente (12.1) e o Lema 12.2, concluímos que x^2 está bem definido. Repetindo indutivamente o argumento, provamos que, para todo $k \geq 0$, o iterando $x^{k+1} = x^k - (\nabla^2 f(x^k)^{-1})\nabla f(x^k)$ está bem definido (item 1) e que

$$\|x^{k+1} - x^*\| \leq r_0 \cdot r_1 \cdots r_k \|x^k - x^*\| \leq \frac{1}{2^{k+1}} \|x^k - x^*\|, \quad (12.4)$$

onde

$$r_k = \|\nabla^2 f(x^k)^{-1}\| \max_{t \in [0,1]} \|\nabla^2 f(tx^k + (1-t)x^*) - \nabla^2 f(x^k)\| \leq 1/2.$$

Isso mostra que $\lim_k x^k = x^*$ com ordem superlinear (item 2).

Vamos mostrar a convergência quadrática (item 3). Suponha que $\nabla^2 f$ seja Lipschitz contínua com constante $L > 0$, como no enunciado. O passo de indução realizado na prova dos itens 1 e 2 fornece uma expressão análoga à (12.3), com x^{k+1} no lugar de x^1 e x^k no lugar de x^0 . Assim, usando o Lema 12.2, temos

$$\begin{aligned} \|x^{k+1} - x^*\| &\leq r_k \|x^k - x^*\| \\ &= \left(\|\nabla^2 f(x^k)^{-1}\| \max_{t \in [0,1]} \|\nabla^2 f(tx^k + (1-t)x^*) - \nabla^2 f(x^k)\| \right) \|x^k - x^*\| \\ &\leq \left(2\|\nabla^2 f(x^*)^{-1}\| \max_{t \in [0,1]} L \underbrace{\|tx^k + (1-t)x^* - x^k\|}_{(1-t)\|x^k - x^*\|} \right) \|x^k - x^*\| \\ &\leq 2\|\nabla^2 f(x^*)^{-1}\| \|x^k - x^*\|^2 \max_{t \in [0,1]} (1-t), \end{aligned}$$

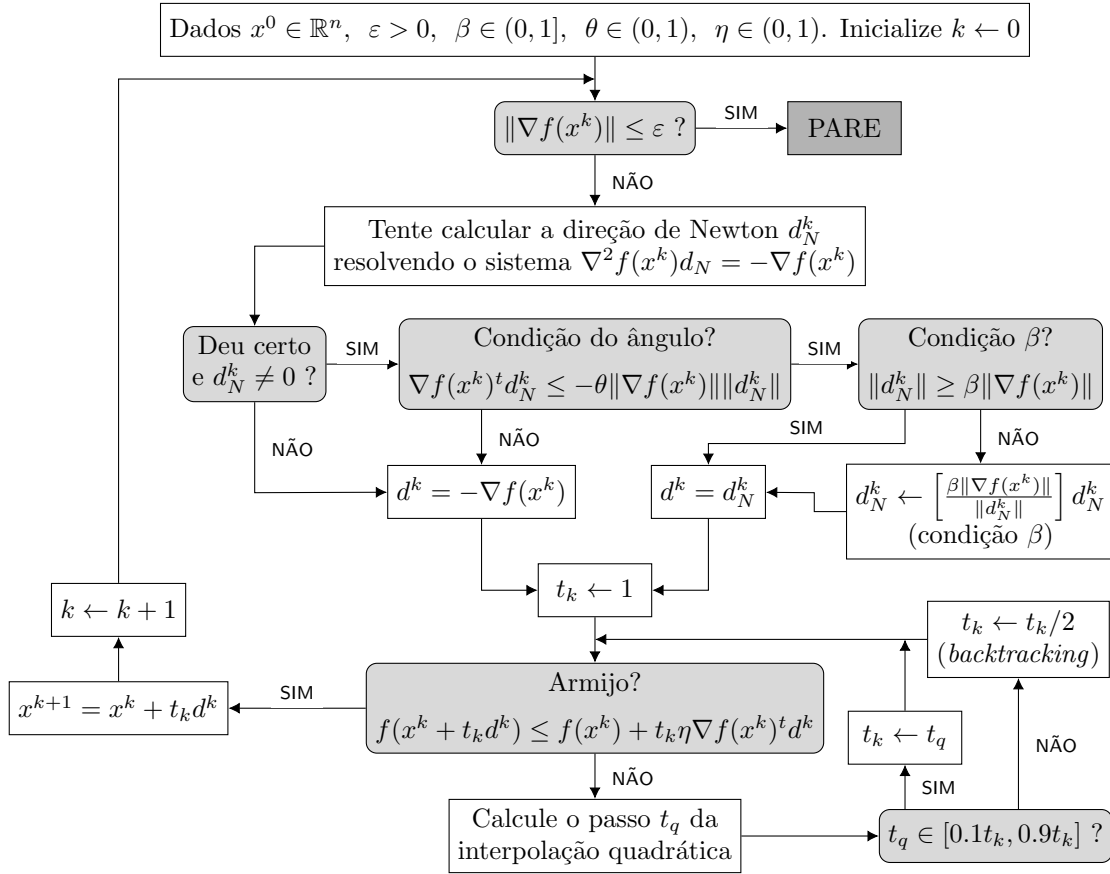
o que implica

$$\|x^{k+1} - x^*\| \leq C \|x^k - x^*\|^2, \quad (12.5)$$

onde $C = 2\|\nabla^2 f(x^*)^{-1}\| > 0$. Ou seja, $\lim_k x^k = x^*$ com ordem quadrática. \square

12.2 Newton globalizado com busca linear

Como sabemos, o método de Newton não possui convergência global. Apesar disso, sua ordem convergência é rápida quando próximo à solução. Uma ideia é “mergulhar” o método de Newton em um esquema com convergência global, tentando assim aproveitar os benefícios de Newton quando estivermos próximo à solução. Existem diferentes maneiras de fazer isso. Nesta seção apresentamos uma hibridização de Newton com o método do gradiente (Algoritmo 11.2, página 39), mergulhados no esquema geral de descida (Algoritmo 11.1, página 36). Vamos nos referir ao método resultante como “Newton globalizado”.



Algoritmo 12.2: Método de Newton globalizado.

O Algoritmo 12.2 possui convergência global, pois se enquadra no esquema geral de descida (Algoritmo 11.1), ao mesmo tempo que combina a convergência global rápida de Newton próximo à solução. Observe que a direção $-\nabla f(x^k)$ é considerada caso a direção de Newton não passe nos testes exigidos no esquema geral de descida. Como $-\nabla f(x^k)$ satisfaz a condição β para todo $\beta \in (0, 1]$ (Lema 11.1, página 38), restringimos o parâmetro $\beta \in (0, 1]$ no Algoritmo 12.2. Assim, a condição β só precisa ser verificada para as direções de Newton.

O método de Newton “puro” (Algoritmo 12.1) a sua convergência local (Teorema 12.1) não dependem da direção d_N^k ser de descida. Ou seja, Newton não tem preferência por minimizadores ou maximizadores. Ele de fato pode convergir à maximizadores (veja o Exercício 6.11 de [1]). Caso a direção de Newton não seja de descida, ela será rejeitada pela condição do ângulo no Algoritmo 12.2. Portanto é importante saber quando ela será de descida. Uma condição suficiente para tal é $\nabla^2 f(x^k)$ ser definida positiva pois, neste caso, $\nabla^2 f(x^k)^{-1}$ o será e logo

$$\nabla f(x^k)^t d_N^k = -\nabla f(x^k)^t \nabla^2 f(x^k)^{-1} \nabla f(x^k) < 0.$$

Na verdade, como a análise acerca de Newton é apenas ao redor de um ponto de acumulação x^* , basta que $\nabla^2 f(x^*)$ seja definida positiva.

O resultado a seguir resume a convergência do Algoritmo 12.2. Em particular, ele diz que a partir de uma certa iteração, o método se comporta como Newton puro para uma escolha adequada dos parâmetros.

Teorema 12.2 (Convergência de Newton globalizado). *Suponha que f seja continuamente diferenciável. Seja x^* um ponto de acumulação da sequência $\{x^k\}$ gerada pelo Algoritmo 12.2, desconsiderando o critério de parada. Então*

$$\nabla f(x^*) = 0.$$

Se f possuir derivadas de segunda ordem contínuas e $\nabla^2 f(x^*)$ for definida positiva então existe $\bar{k} \in \mathbb{N}$ tal que

1. a direção de Newton $d_N^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$ está bem definida para todo $k \geq \bar{k}$, ou seja, é calculada com sucesso a partir da iteração \bar{k} ;

2. para $\theta \in (0, \frac{1}{4\kappa})$, d_N^k satisfaz a condição do ângulo para todo $k \geq \bar{k}$, onde

$$\kappa = \|\nabla^2 f(x^*)^{-1}\| \|\nabla^2 f(x^*)\| \geq 1$$

é o número de condição da matriz $\nabla^2 f(x^*)$;

3. para $\beta \in (0, \frac{1}{2\|\nabla^2 f(x^*)\|})$, $\|d_N^k\| \geq \beta \|\nabla f(x^k)\|$ para todo $k \geq \bar{k}$, ou seja, a direção de Newton nunca é ajustada a partir da iteração \bar{k} ;

4. para $\eta \in (0, \frac{1}{2})$, a condição de Armijo é satisfeita com $d^k = d_N^k$ e $t_k = 1$ para todo $k \geq \bar{k}$;

5. $\{x^k\}$ converge a x^* com ordem pelo menos superlinear, e quadrática se $\nabla^2 f$ é Lipschitz contínua.

Demonstração. Seja $K \subseteq \mathbb{N}$ o conjunto de índices para o qual a subsequência de $\{x^k\}$ converge, isto é, $\lim_{k \in K} x^k = x^*$. A condição de primeira ordem $\nabla f(x^*) = 0$ segue do Teorema 11.1 (página 36). A matriz $\nabla^2 f(x^*)$ é não singular pois é definida positiva. Ademais, da continuidade de $\nabla^2 f$ em x^* , existe $\varepsilon_1 > 0$ tal que $\nabla^2 f(x)$ é não singular para todo x satisfazendo $\|x - x^*\| \leq \varepsilon_1$. Por sua vez, o Teorema 12.1 garante que existe $\varepsilon_2 > 0$ para o qual a direção de Newton será calculada com sucesso caso $\|x^k - x^*\| \leq \varepsilon_2$. Em particular, d_N^k será calculada com sucesso na iteração k caso

$$k \in S = \{k \in K \mid \|x^k - x^*\| \leq \varepsilon\},$$

onde $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$. Mais ainda, para estes índices, $\nabla^2 f(x^k)$ é não singular, e logo $d_N^k = 0$ se, e somente se, $\nabla f(x^k) = 0$. Neste caso o algoritmo prossegue com $d^k = 0$, e nada há o que provar. Logo, podemos supor sem perda de generalidade que $d_N^k \neq 0$ para todo $k \in S$. Note ainda que S é infinito e que $\lim_{k \in S} x^k = x^*$.

Vamos mostrar primeiro que cada um dos itens 1 a 4 ocorre não para um único \bar{k} , mas para todo k suficientemente grande e *restrito* à S . Em seguida seremos capazes de argumentar que os itens 1 a 4 ocorrem para todo $k \geq \bar{k}$, onde k é *qualquer* índice de \mathbb{N} .

Pelo visto acima, a boa definição de d_N^k vale para todo $k \in S$.

A condição do ângulo é satisfeita para todo $k \in S$ suficientemente grande. O Teorema da Função Inversa aplicado à ∇f garante que $\nabla^2 f^{-1}$ é contínua em x^* . A continuidade de $\nabla^2 f$ e $\nabla^2 f^{-1}$ em x^* implica

$$0 < \|\nabla^2 f(x^k)\| \leq 2\|\nabla^2 f(x^*)\|, \quad 0 < \|\nabla^2 f(x^k)^{-1}\| \leq 2\|\nabla^2 f(x^*)^{-1}\| \quad (12.6)$$

para todo $k \in S$ suficientemente grande, digamos, para todo $k \in S$, $k \geq k_0$. Assim,

$$\theta \leq \frac{1}{4\kappa} = \frac{1}{4\|\nabla^2 f(x^*)\|\|\nabla^2 f(x^*)^{-1}\|} \leq \frac{1}{\|\nabla^2 f(x^k)\|\|\nabla^2 f(x^k)^{-1}\|}. \quad (12.7)$$

Como $\nabla f(x^k)$ é simétrica, sua inversa $\nabla f(x^k)^{-1}$ também é. Ademais, $\nabla f(x^k)^t d_N^k = \nabla f(x^k)^t \nabla^2 f(x^k)^{-1} \nabla f(x^k)$, e logo segue de (12.7) e das propriedades da norma matricial que

$$\begin{aligned} \theta \|\nabla f(x^k)\| \|d_N^k\| &\leq \frac{\|\nabla f\| \|d_N^k\|}{\|\nabla^2 f\| \|\nabla^2 f^{-1}\|} \frac{|\nabla f^t d_N^k|}{|\nabla f^t \nabla^2 f^{-1} \nabla f|} \leq \frac{\|\nabla f\| \|d_N^k\| |\nabla f^t d_N^k|}{\|\nabla^2 f\| \|\nabla^2 f^{-1} \nabla f\|^2} \\ &= \frac{\|\nabla f\| |\nabla f^t d_N^k|}{\|\nabla^2 f\| \|\nabla^2 f^{-1} \nabla f\|} \leq \frac{\|\nabla f\| |\nabla f^t d_N^k|}{\|\nabla^2 f \nabla^2 f^{-1} \nabla f\|} \\ &= |\nabla f^t d_N^k| \end{aligned}$$

para todo $k \in S$, $k \geq k_0$ (omitimos “ (x^k) ” por simplicidade). Isso implica que

$$\nabla f(x^k)^t d_N^k = -|\nabla f(x^k)^t d_N^k| \leq -\theta \|\nabla f(x^k)\| \|d_N^k\|$$

para todo $k \in S$, $k \geq k_0$.

A condição β é satisfeita para todo $k \in S$ suficientemente grande. De (12.6) vem

$$\nabla^2 f(x^k) d_N^k = -\nabla f(x^k) \Rightarrow \|d_N^k\| \geq \frac{1}{\|\nabla^2 f(x^k)\|} \|\nabla f(x^k)\| \geq \beta \|\nabla f(x^k)\|$$

para todo $k \in S$, $k \geq k_0$.

A condição de Armijo é satisfeita com $t_k = 1$ para todo $k \in S$ suficientemente grande.

A aproximação quadrática de f em torno de x^* (Teorema 1.2, página 8) ao longo da direção de Newton fornece

$$f(x^k + d_N^k) = f(x^k) + \nabla f(x^k)^t d_N^k + \frac{1}{2} (d_N^k)^t \nabla^2 f(x^k) d_N^k + r(d_N^k)$$

onde

$$\frac{r(d_N^k)}{\|d_N^k\|} \rightarrow 0 \quad \text{quando} \quad d_N^k \rightarrow 0. \quad (12.8)$$

Substituindo $\nabla f(x^k) = -\nabla^2 f(x^k) d_N^k$ no termo linear da expressão acima, obtemos

$$f(x^k + d_N^k) = f(x^k) - \frac{1}{2} (d_N^k)^t \nabla^2 f(x^k) d_N^k + r(d_N^k). \quad (12.9)$$

Suponha por absurdo que a condição de Armijo falhe infinitas vezes para $t_k = 1$, digamos para $k \in K_1 \subseteq S$, isto é,

$$f(x^k + d_N^k) > f(x^k) + \alpha \nabla f(x^k)^t d_N^k = f(x^k) - \alpha (d_N^k)^t \nabla^2 f(x^k) d_N^k, \quad (12.10)$$

para todo $k \in K_1$. Combinando (12.9) e (12.10) e dividindo por $\|d_N^k\|^2$, obtemos, para todo $k \in K_1$,

$$\frac{r(d_N^k)}{\|d_N^k\|^2} \geq \left(\frac{1}{2} - \alpha\right) \frac{(d_N^k)^t \nabla^2 f(x^k) d_N^k}{(d_N^k)^t d_N^k} \geq \left(\frac{1}{2} - \alpha\right) \lambda_{\min}(\nabla^2 f(x^k)), \quad (12.11)$$

onde $\lambda_{\min}(\nabla^2 f(x^k))$ é o menor autovalor da matriz $\nabla^2 f(x^k)$.

Como $\nabla f(x^*) = 0$, temos $\lim_{k \in K} \|\nabla^2 f(x^k) d_N^k\| = \lim_{k \in K} \|\nabla f(x^k)\| = 0$, e como $\nabla^2 f$ é contínua e definida positiva em x^* , segue que $\lim_{k \in K} d_N^k = 0$. Mais ainda, os autovalores são funções contínuas da matriz, e logo $\lambda_{\min}(\nabla^2 f(x^k)) \geq 1/2 \lambda_{\min}(\nabla^2 f(x^*)) > 0$ para todo $k \in K_1$ suficientemente grande. Assim, levando $k \in K_1$ ao limite em (12.11) obtemos

$$\lim_{k \in K_1} \frac{r(d_N^k)}{\|d_N^k\|^2} \neq 0,$$

contrariando (12.8). Concluimos portanto que (12.10) ocorre apenas finitas vezes, isto é, existe $k_1 \in S$ tal que a condição de Armijo é satisfeita para a direção de Newton d_N^k com $t_k = 1$, para todo $k \in S$, $k \geq k_1$.

Validade para todo $k \in \mathbb{N}$, $k \geq \bar{k}$. De acordo com os itens anteriores, é natural definirmos $\bar{k} = \max\{k_0, k_1\}$. Pela prova de convergência local do método de Newton (Teorema 12.1), temos

$$\|(x^k + d_N^k) - x^*\| \leq \|x^k - x^*\| \leq \varepsilon \quad (12.12)$$

sempre que $k \in S$ (veja (12.4)). Ou seja, a direção de Newton mantém o iterando próximo à x^* , dentro da vizinhança de raio ε . Isso nos permite utilizar o seguinte argumento: para $k = \bar{k}$, a direção de Newton é calculada com sucesso e não é descartada pois, pelo visto acima, as condições do ângulo, β e de Armijo (com $t_k = 1$) são satisfeitas. O novo iterando $x^{k+1} = x^k + d_N^k$ satisfaz (12.12), e logo $k+1 \in S$. Repetindo esse argumento sucessivamente, concluimos que $k \in S$ para todo $k \geq \bar{k}$, $k \in \mathbb{N}$. Ou seja, os itens 1, 2, 3 e 4 valem com este \bar{k} .

Finalmente, o item 5 segue diretamente dos itens anteriores e do Teorema 12.1. \square

Observação 12.1. A fim de aproveitar os benefícios do método de Newton, o teorema acima sugere escolhas para os parâmetros do Algoritmo 12.2. Em particular, o parâmetro η da busca de Armijo deve ser menor que $1/2$; na prática é comum escolher $\eta = 10^{-4}$. Os outros parâmetros são menos diretos pois dependem da hessiana de f . Mesmo assim, o item 2 sugere que quanto mais mal condicionada é a hessiana de f em x^* ($\kappa \gg 1$), mais próximo de zero θ deve ser. Na dúvida, podemos escolher $\theta \approx 0$ (usualmente, $\theta = 10^{-6}$). Uma análise similar vale para β .