Notas de estudo

Método do Gradiente equipado com buscas lineares monótonas

Emanuel Mendes Queiroz
Samara Viriato Vilar Dias

Vitória da Conquista, 2023

Conteúdo

1 Resultados preliminares

Nesta seção, apresentamos uma seleção de teoremas e definições fundamentais que servirão como base para o desenvolvimento das ideias e conceitos abordados ao longo destas notas de estudo.

1.1 Resultados da Análise

Definição 1.1. *Seja $X \subset \mathbb{R}$. Um número $a \in \mathbb{R}$ chama-se ponto de acumulação do conjunto X quando todo intervalo aberto $(a-\epsilon, a+\epsilon)$, de centro a, contém algum ponto de X diferente de a.*

Definição 1.2. *Seja $f : \mathbb{R}^n \to \mathbb{R}$. Dizemos que x é um ponto estacionário de f se as derivadas parciais de f nesse ponto são nulas.*

Teorema 1.3. ***(Teorema do Valor Médio).*** *Considere $f : \mathbb{R}^n \to \mathbb{R}$ contínua, $\overline{x} \in \mathbb{R}^n$ e uma direção de descida $d \in \mathbb{R}^n$ a partir de $\overline{x}$. Se f é diferenciável no segmento $(\overline{x}, \overline{x}+d)$, então existe $\alpha \in (0,1)$ tal que*

$$f(\overline{x}+d) = f(\overline{x}) + \nabla f(\overline{x}+\alpha d)^T d \Longrightarrow f(\overline{x}+d) - f(\overline{x}) = \nabla f(\overline{x}+\alpha d)^T d.$$

Algumas outras versões que são chamadas de Teorema do Valor Médio:

Teorema 1.4. *Vale o seguinte*

(a) *Se para $x, y \in \mathbb{R}^n$ uma função $F : \mathbb{R}^n \to \mathbb{R}^l$ é continuamente diferenciável no intervalo $\{x + ty \mid t \in [0,1]\}$, então*

$$F(x+y) = F(x) + \int_0^1 F'(x+ty)y\,dt.$$

(b) *Se $f : \mathbb{R}^n \to \mathbb{R}$ é continuamente diferenciável num conjunto convexo e aberto $\Omega \subset \mathbb{R}^n$, então para todo $x, y \in \Omega$ existe $t \in [0,1]$ tal que*

$$f(y) - f(x) = \langle \nabla f(tx + (1-t)y), y - x \rangle.$$

Lema 1.5. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável no $\mathbb{R}^n$, com gradiente Lipschitz-contínuo no $\mathbb{R}^n$ com módulo $L > 0$. Então*

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

Demonstração: Pelo Teorema 1.4 (a), temos que $f(x+y) = f(x) + \int_0^1 \nabla f(x+ty)^T y\,dt$. Assim,

$$\begin{aligned} |f(x+y) - f(x) - \nabla f(x)^T y| &= |f(x) + \int_0^1 \nabla f(x+ty)^T y\,dt - f(x) - \nabla f(x)^T y| \\ &= |\int_0^1 \nabla f(x+ty)^T y\,dt - \nabla f(x)^T y| \\ &= |\int_0^1 (\nabla f(x+ty) - \nabla f(x))^T y\,dt| \\ &\leq \int_0^1 |(\nabla f(x+ty) - \nabla f(x))^T y|\,dt. \end{aligned}$$

Pela Desigualdade de Cauchy-Schawrz, temos que

$$\int_0^1 |(\nabla f(x+ty) - \nabla f(x))^T y|\, dt \leq \int_0^1 \|\nabla f(x+ty) - \nabla f(x)\| \|y\|\, dt.$$

Além disso, temos (por hipótese) que ∇f é Lipschitz-contínuo no $\mathbb{R}^n$, ou seja, $\|\nabla f(x+ty) - \nabla f(x)\| \leq L\|(x+ty) - x\| = L\|ty\| = Lt\|y\|$, logo

$$\begin{aligned}\int_0^1 \|\nabla f(x+ty) - \nabla f(x)\| \|y\|\, dt &\leq \int_0^1 Lt\|y\|^2\, dt \\ &= \left[L\frac{t^2}{2}\|y\|^2\right]_0^1 \\ &= \frac{L\|y\|^2}{2}.\end{aligned}$$

□

1.2 Resultados sobre convexidade

Quando uma função é diferenciável, a convexidade admite várias caracterizações que são muito úteis para determinar se uma função é convexa ou não.

Teorema 1.6. ***(Caracterizações de funções convexas diferenciáveis)*** *Sejam $\Omega \subset \mathbb{R}^n$ um conjunto convexo e aberto e $f : \Omega \to \mathbb{R}$ uma função diferenciável em Ω. Então, as propriedades seguintes são equivalentes:*

(a) *A função f é convexa em Ω;*

(b) *para quaisquer $x, y \in \Omega$,*

$$f(y) \geq f(x) + \langle \nabla f(x_k), y - x \rangle .$$

Demonstração: Veja a demonstração em [3, Teorema 3.4.30].

1.3 Ordem de convergência

Nesta seção, serão apresentadas definições importantes e exemplos no que diz respeito a alguns tipos de ordem de convergência de uma sequência.

Definição 1.7. *Considere uma sequência $\{x_k\}_{k\in\mathbb{N}} \in \mathbb{R}^n$. Dizemos que x_k converge linearmente para $\overline{x} \in \mathbb{R}^n$, quando existe uma constante $c \in [0, 1)$ e um número natural k_0, tais que*

$$\frac{\|x_{k+1} - \overline{x}\|}{\|x_k - \overline{x}\|} \leq c, \tag{1}$$

para todo $k \geq k_0$.

Se uma sequência cumpre a condição (1), então

$$\begin{aligned}\|x_{k_0+1}-\overline{x}\| &\leq c\|x_{k_0}-\overline{x}\| \\ \|x_{k_0+2}-\overline{x}\| &\leq c\|x_{k_0+1}-\overline{x}\| \leq c^2\|x_{k_0}-\overline{x}\|.\end{aligned}$$

Repetindo o processo acima, obtemos

$$\|x_{k_0+p}-\overline{x}\| \leq c^p\|x_{k_0}-\overline{x}\|, \tag{2}$$

para todo $p \in \mathbb{N}$. Como $c^p \to 0$, pois $c \in [0,1)$, segue de (2) que $x_k \to \overline{x}$.

Definição 1.8. *Uma sequência $\{x_k\}_{k\in\mathbb{N}} \in \mathbb{R}^n$ é dita superlinearmente convergente para $\overline{x} \in \mathbb{R}^n$ quando*

$$\lim_{k\to\infty} \frac{\|x_{k+1}-\overline{x}\|}{\|x_k-\overline{x}\|} = 0. \tag{3}$$

A condição (3) também implica que $x_k \to \overline{x}$, basta observar que dado $0 < c < 1$ existe $k_0 \in \mathbb{N}$ tal que $\|x_{k+1}-\overline{x}\| \leq \varepsilon\|x_k-\overline{x}\|$, para todo $k \geq k_0$.

Exemplo 1.9. *A sequência $x_k = \frac{1}{3^k}$ converge linearmente para 0, mas não superlinearmente. De fato, temos que*

$$\frac{|x_{k+1}-\overline{x}|}{|x_k-\overline{x}|} = \frac{|x_{k+1}-0|}{|x_k-0|} = \frac{|x_{k+1}|}{|x_k|} = \frac{1}{3^{k+1}}\cdot 3^k = \frac{1}{3^k\cdot 3}\cdot 3^k = \frac{1}{3} \leq c = \frac{1}{2}.$$

Entranto,

$$\lim_{k\to\infty}\frac{|x_{k+1}-\overline{x}|}{|x_k-\overline{x}|} = \lim_{k\to\infty}\frac{1}{3} = \frac{1}{3} \neq 0.$$

Exemplo 1.10. *A sequência $x_k = \frac{1}{2^{k^2}}$ converge superlinearmente para 0, pois*

$$\frac{|x_{k+1}-\overline{x}|}{|x_k-\overline{x}|} = \frac{|x_{k+1}-0|}{|x_k-0|} = \frac{1}{2^{(k+1)^2}}\cdot 2^{k^2} = \frac{2^{k^2}}{2^{(k+1)^2}} = \frac{2^{k^2}}{2^{k^2}\cdot 2^{2k}\cdot 2} = \frac{1}{2^{2k}\cdot 2} = \frac{1}{2}\cdot\frac{1}{2^{2k}}$$

e sabemos que

$$\lim_{k\to\infty}\frac{1}{2}\cdot\frac{1}{2^{2k}} = \lim_{k\to\infty}\frac{1}{2}\cdot\lim_{k\to\infty}\frac{1}{2^{2k}} = \frac{1}{2}\cdot 0 = 0.$$

Definição 1.11. *Seja $\{x_k\}_{k\in\mathbb{N}} \in \mathbb{R}^n$ que convergente para $\overline{x} \in \mathbb{R}^n$. Dizemos que $\{x_k\}_{k\in\mathbb{N}}$ converge com ordem quadrática para $\overline{x}$ se existe um número natural k_0 e uma constante $M > 0$ tal que*

$$\|x_{k+1}-\overline{x}\| \leq M\|x_k-\overline{x}\|^2, \tag{4}$$

para todo $k \geq k_0$.

Exemplo 1.12. *A sequência $x_k = \frac{1}{2^{2^k}}$ converge quadraticamente para 0. De fato,*

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

Logo, basta tomar $M \geq 1$.

Observação 1.13. *A convergência quadrática implica na convergência superlinear. No entanto, apenas a condição (4) não implica que $x_k \to \overline{x}$.*

2 Busca de Armijo

Uma busca linear inexata muito utilizada para a determinação do comprimento de passo em algoritmos de otimização é a busca denominada Regra de Armijo, veja [2, Pág. 67]. Essa regra consiste em determinar um $\alpha_k \in [0, \delta)$, com $\delta > 0$, tal que

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \eta_1 \alpha_k \nabla f(x_k)^T d_k, \tag{5}$$

onde $x_k \in \mathbb{R}^n$, $d_k \in \mathbb{R}^n \setminus \{0\}$ é uma direção de descida e $\eta_1 \in (0, 1)$.

Lema 2.1. ***(Cota inferior para o valor do comprimento de passo dado pela Busca de Armijo)*** *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável no $\mathbb{R}^n$, com derivada Lipschitz-contínua no $\mathbb{R}^n$ com módulo $L > 0$. Se $x_k, d_k \in \mathbb{R}^n$ satisfazem a condição $\nabla f(x_k)^T d_k < 0$, então a desigualdade (5) é válida para todo $\alpha \in (0, \overline{\alpha}_k]$, onde*

$$\overline{\alpha}_k = \frac{2(\eta_1 - 1)(\nabla f(x_k)^T d_k)}{L\|d_k\|^2} > 0. \tag{6}$$

Demonstração: Pelo Lema 1.5, para todo $\alpha \in \mathbb{R}$, tem-se que

$$f(x_k + \alpha d_k) - f(x_k) - \nabla f(x_k)^T(\alpha d_k) < |f(x_k + \alpha d_k) - f(x_k) - \nabla f(x_k)^T(\alpha d_k)| \leq \frac{L}{2}\|\alpha d_k\|^2.$$

Daí, segue que

$$\begin{aligned}
&f(x_k + \alpha d_k) - f(x_k) - \nabla f(x_k)^T(\alpha d_k) \leq \frac{L}{2}\alpha^2\|d_k\|^2 \\
\Longrightarrow &f(x_k + \alpha d_k) - f(x_k) \leq \nabla f(x_k)^T(\alpha d_k) + \frac{L}{2}\alpha^2\|d_k\|^2 \\
\Longrightarrow &f(x_k + \alpha d_k) - f(x_k) \leq \alpha(\nabla f(x_k)^T d_k + \frac{L}{2}\alpha\|d_k\|^2).
\end{aligned}$$

Logo, para todo $\alpha \in (0, \overline{\alpha}_k]$,

$$\begin{aligned}
&f(x_k + \alpha d_k) - f(x_k) \leq \alpha(\nabla f(x_k)^T d_k + \frac{L}{2}\overline{\alpha}_k\|d_k\|^2) \\
\Longrightarrow &f(x_k + \alpha d_k) - f(x_k) \leq \alpha(\nabla f(x_k)^T d_k + \frac{L}{2}\|d_k\|^2 \cdot \frac{2(\eta_1 - 1)(\nabla f(x_k)^T d_k)}{L\|d_k\|^2}) \\
\Longrightarrow &f(x_k + \alpha d_k) - f(x_k) \leq \alpha(\nabla f(x_k)^T d_k + (\eta_1 - 1)(\nabla f(x_k)^T d_k)) \\
\Longrightarrow &f(x_k + \alpha d_k) - f(x_k) \leq \alpha(\nabla f(x_k)^T d_k + \eta_1 \nabla f(x_k)^T d_k - \nabla f(x_k)^T d_k) \\
\Longrightarrow &f(x_k + \alpha d_k) - f(x_k) \leq \alpha\eta_1 \nabla f(x_k)^T d_k,
\end{aligned}$$

onde a segunda desigualdade segue de (6). □

A interpretação geométrica do Lema 2.1 é dada pela Figura 1: no caso quando o gradiente de f é Lipschitz-contínuo, o valor de $f(x_k + \alpha d_k) - f(x_k)$ é menor ou igual ao valor da função quadrática $\alpha \nabla f(x_k)^T d_k + \frac{L\alpha^2\|d_k\|^2}{2}$, cujo minimizador é o ponto $\frac{-\nabla f(x_k)^T d_k}{L\|d_k\|^2}$ (vértice da parábola).

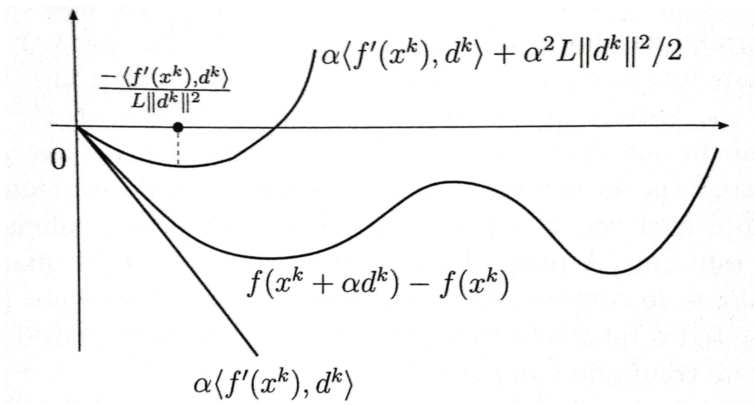


Figura 1: Interpretação geométrica do Lema 2.1.

Sob as hipóteses do Lema 2.1, se

$$\frac{\nabla f(x_k)^T d_k}{\|d_k\|^2} \leq \delta < 0, \tag{7}$$

onde δ é uma constante que não depende de k, e se os parâmetros $\hat{\alpha}$, η_1 e θ são os mesmos para cada iteração, segue-se que a desigualdade (5) é satisfeita para todo $\alpha \in (0, \overline{\alpha}]$, onde

$$\overline{\alpha} = \frac{-2\delta(1-\eta_1)}{L} > 0.$$

Concluímos então que

$$\alpha_k \geq min\{\hat{\alpha}, \theta\overline{\alpha}\} = \check{\alpha} > 0. \tag{8}$$

3 Busca de Goldstein

Ao se utilizar a Busca de Armijo podem ser admitidos passos muito pequenos, aumentando o esforço computacional, tornando assim a resolução do problema mais lenta. Para contornar essa dificuldade, é proposto na literatura uma outra busca linear inexata que é conhecida como Regra de Goldstein. Dados um ponto x_k, uma direção de descida $d_k \in \mathbb{R}^n \setminus \{0\}$ e os parâmetros $0 < \eta_1 < \eta_2 < 1$, a Busca de Goldstein consiste em obter um comprimento de passo α_k que satisfaça simultaneamente as seguintes desigualdades:

$$f(x_k) + \eta_2\alpha_k\nabla f(x_k)^T d_k \leq f(x_k + t_k d_k) \leq f(x_k) + \eta_1\alpha_k\nabla f(x_k)^T d_k \tag{9}$$

A desigualdade à direita é a de Armijo, que garante um descrescimento da função objetivo, já a desigualdade à esquerda busca eliminar a aceitação de comprimentos de passo muito pequenos.

Seja $l_1(\alpha) = f(x_k) + \eta_1\alpha_k\nabla f(x_k)^T d_k$ e $l_2(\alpha) = f(x_k) + \eta_2\alpha_k\nabla f(x_k)^T d_k$. O resultado a seguir garante que a busca de Goldstein está bem definida.

Teorema 3.1. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função continuamente diferenciável e limitada inferiormente ao longo do conjunto $\{x_k + \alpha d_k; \alpha > 0\}$ e d_k uma direção de descida a partir de x_k. Se $0 < \eta_1 < \eta_2 < 1$, então existe um intervalo de comprimentos de passo que satisfazem simultaneamente as condições de Goldstein.*

Demonstração: Por hipótese, temos que f é contínua e limitada inferiormente, então a restrição de f na direção d_k a partir de x_k dada por $\varphi(\alpha) = f(x_k + \alpha d_k)$ é contínua e limitada inferiormente para todo $\alpha > 0$. Além disso, temos que d_k é direção de descida, ou seja, $\nabla f(x_k)^T d_k < 0$, o que nos fornece $\lim\limits_{t\to\infty} l_1(\alpha) = \lim\limits_{t\to\infty} l_2(\alpha) = -\infty$. Note que a função $l_1(\cdot)$ tem a inclinação negativa dada por $\eta_1 \alpha_k \nabla f(x_k)^T d_k$, mas como $\eta_1 \in (0, 1)$, a reta l_1 fica acima do gráfico de φ para pequenos valores positivos de α (ou para α próximo de 0), e o mesmo ocorre para a reta l_2. Então, existe α' próximo de zero tal que $\varphi(\alpha') = c' < l_2(\alpha')$. Por outro lado, pela definição de limite infinito no infinito, temos que dado $c'' < 0$, existe $N > 0$ tal que $\alpha'' > N$ implica $\varphi(\alpha'') \geq c'' > l_2(\alpha'')$. Pelo Teorema do Valor Intermediário, existe $\alpha''' \in (\alpha', \alpha'')$ tal que $\varphi(\alpha''') = c''' = l_2(\alpha''')$, ou seja, φ intersecta l_2. Utilizando um raciocínio análogo, pode-se mostrar que φ também intersecta l_1. Note também que $\eta_1 < \eta_2$, então o coeficiente angular de l_2 é menor do que o coeficiente angular de l_1, ou seja, l_2 descresce mais rapidamente do que l_1, o que implica que $l_2(\alpha) < l_1(\alpha)$, para todo $\alpha > 0$. Disso decorre que $\varphi(\alpha)$ intersecta $l_2(\alpha)$ antes de intersectar $l_1(\alpha)$. Além disso, $\varphi(\alpha)$ intersecta l_1 e l_2 em no máximo um número finito de vezes, já que $\varphi(\alpha)$ é limitada inferiormente, enquanto l_1 e l_2 não são. Seja α_1 o comprimento do passo que gera a primeira intersecção de φ com l_1 e α_2 o maior comprimento de passo menor do que α_1 onde φ intersecta l_2. O intervalo (α_2, α_1) é formado por comprimentos de passo que satisfazem as condições de Goldstein. O mesmo critério pode ser usado para os outros pontos de intersecção de φ com l_1 e l_2.

3.1 Convergência global do Método do Gradiente equipado com a Busca de Goldstein

Para que um algoritmo de busca linear seja confiável, ele deve ser globalmente convergente, ou seja, as normas de gradiente, $\|\nabla f(x_k)\|$, devem convergir para zero a cada iteração, ou seja, $\lim\limits_{k\to\infty} \|\nabla f(x_k)\| = 0$.

Segundo [4], essa propriedade pode ser mostrada a partir do **Teorema de Zoutendijk** que diz que, se o algoritmo de busca linear satisfizer as condições de Goldstein e tiver uma direção de descida que faz um ângulo com a direção de descida mais íngreme que é limitado a partir de 90° ($d_k = -\nabla f(x_k)$), o algoritmo é globalmente convergente.

O teorema de Zoutendijk afirma que, dada uma iteração onde d_k é a direção de descida e α_k é o passo que satisfaz as condições de Goldstein, se a função f é limitada inferiormente em $\mathbb{R}^n$ e é continuamente diferenciável em um conjunto aberto U contendo o conjunto de nível definido $\Omega = \{x | f(x) < f(x_0)\}$, onde x_0 é o ponto inicial da iteração, e o gradiente ∇f Lipschitz é contínuo em U, então $\sum_{k=0}^{\infty} cos^2\theta_k \|\nabla f_k\|^2 < \infty$, onde θ_k é o ângulo entre $d_k = -\nabla f(x_k)$ (direção de descida mais íngreme) e ∇f_k.

A condição Zoutendijk acima implica que $\lim_{k\to\infty} cos^2\theta_k \|\nabla f_k\|^2 = 0$, pelo teste de divergência do n-ésimo termo. Portanto, se o algoritmo escolher uma direção de descida que seja limitada a partir de 90° em relação ao gradiente, ou seja, dado $\epsilon > 0$, $cos\ \theta_k \geq \epsilon > 0$, $\forall k$, segue que $\lim_{k\to\infty} \|\nabla f_k\| = 0$.

No entanto, a condição de Zoutendijk não garante a convergência para um mínimo local, mas apenas para pontos estacionários. Antes de apresentar e demonstrar o Teorema que garante a convergência, relembramos alguns fatos relevantes:

- Seja θ_k o ângulo entre uma direção de descida d_k e $\nabla f(x_k)$, então

$$\cos\theta_k = \frac{\nabla f(x_k)^T d_k}{\|f(x_k)\|\|d_k\|};$$

- O conjunto $C = \{x|f(x) = c\}$ é o conjunto de nível de f para o valor c. Da mesma forma, $C' = \{x|f(x) \leq c\}$ é seu conjunto de subnível correspondente;

- O gradiente de uma função $f(x_k)$ ser Lipschitz contínuo em U significa que existe uma constante $L > 0$ tal que para todo $x, \overline{x} \in U$, temos

$$\|\nabla f(x) - \nabla f(\overline{x})\| \leq L\|x - \overline{x}\|.$$

Teorema 3.2. *Suponha que f é limitada inferiormente e é continuamente diferenciável em um conjunto aberto U contendo o nível definido $\Omega = \{x|f(x) \leq f(x_0)\}$, onde x_0 é o ponto inicial da iteração. Assuma também que o gradiente de f é Lipschitz contínuo em U. Seja $R > 1$ uma constante positiva. Em cada iterada k, suponha que d_k seja escolhido como uma direção de descida, então $\nabla {f_k}^T d_k < 0$, e a_k, b_k são escolhidos com $0 < b_k < Ra_k$ tal que o passo a_k satisfaz $f(x_k + a_k d_k) \leq f(x_k) + m_1 a_k \nabla f(x_k)^T d_k$ e o passo b_k satisfaz $f(x_k + b_k d_k) \geq f(x_k) + m_2 b_k \nabla f(x_k)^T d_k$, com $0 < m_1 < m_2 < 1$. Defina $x_{k+1} = x_k + a_k d_k$, com $k = 0, 1, 2, \ldots$. Então $\nabla f_k = 0$, para algum k, ou então $\sum\limits_{k=0}^{\infty} cos^2\theta_k \|\nabla f_k\|^2 < \infty$, onde θ_k é o ângulo entre d_k e ∇f_k.*

Demonstração: Tome $m_1 < 1/2$ e $m_2 = 1 - m_1$. Por $f(x_k + b_k d_k) \geq f(x_k) + m_2 b_k \nabla f(x_k)^T d_k$ temos

$$f(x_k + b_k d_k) \geq f(x_k) + (1 - m_1) b_k \nabla f(x_k)^T d_k \tag{10}$$

$$\Longrightarrow f(x_k + b_k d_k) - f(x_k) \geq (1 - m_1) b_k \nabla f(x_k)^T d_k. \tag{11}$$

Aplicando o Teorema 1.3 e considerando $d = b_k d_k$, temos

$$f(x_k + b_k d_k) - f(x_k) = \nabla f(x_k + \alpha b_k d_k)^T (b_k d_k). \tag{12}$$

De (10) e (12), temos

$$\nabla f(x_k + \alpha b_k d_k)^T (b_k d_k) = f(x_k + b_k d_k) - f(x_k) \geq (1 - m_1) b_k \nabla f(x_k)^T d_k.$$

Dividindo ambos os lados da desiguladade por b_k, obtemos

$$\nabla f(x_k + \alpha b_k d_k)^T d_k \geq (1 - m_1) \nabla f(x_k)^T d_k.$$

Considerando $c = \alpha b_k$, temos que $0 < c < b_k$, pois $\alpha \in (0, 1)$. Assim,

$$\nabla f(x_k + c d_k)^T d_k \geq (1 - m_1) \nabla f(x_k)^T d_k = \nabla f(x_k)^T d_k - m_1 \nabla f(x_k)^T d_k.$$

Subtraindo $\nabla f(x_k)^T d_k$ de ambos os lados e pela definição de produto interno, obtemos

$$\left[\nabla f(x_k + c d_k) - \nabla f(x_k)\right]^T d_k \geq -m_1 \|\nabla f(x_k)\| \|d_k\| \cos\theta_k.$$

Observamos que ambos os lados são positivos, já que d_k é uma direção de descida, então $cos\theta_k = \dfrac{\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \|d_k\|} < 0$. Pela Desigualdade de Schwarz, temos

$[\nabla f(x_k + cd_k) - \nabla f(x_k)]^T d_k = |\, [\nabla f(x_k + cd_k) - \nabla f(x_k)]^T d_k| \leq \|\nabla f(x_k{+}cd_k){-}\nabla f(x_k)\| \|d_k\|$

Seja L uma constante de Lipschitz para ∇f. Então:

$$\begin{aligned}\|\nabla f(x_k + cd_k) - \nabla f(x_k)\| \leq L\|(x_k + cd_k) - x_k\| &\Longrightarrow \|\nabla f(x_k + cd_k) - \nabla f(x_k)\| \leq L\|cd_k\| \\ &\Longrightarrow \|\nabla f(x_k + cd_k) - \nabla f(x_k)\| \leq Lc\|d_k\|.\end{aligned}$$

Multiplicando ambos dos lados por $\|d_k\|$, temos

$$\begin{aligned}\|\nabla f(x_k + cd_k) - \nabla f(x_k)\| \|d_k\| \leq Lc\|d_k\|^2 &\Longrightarrow [\nabla f(x_k + cd_k) - \nabla f(x_k)]^T d_k \leq Lc\|d_k\|^2 \\ &\Longrightarrow -m_1\|\nabla f(x_k)\| \|d_k\| cos\theta_k \leq Lc\|d_k\|^2.\end{aligned}$$

Além disso, temos que $c \leq b_k \leq Ra_k$, então

$$\begin{aligned}\frac{-m_1\|\nabla f(x_k)\| \|d_k\| cos\theta_k}{\|d_k\|} \leq \frac{Lc\|d_k\|^2}{\|d_k\|} &\Longrightarrow -m_1\|\nabla f(x_k)\| cos\theta_k\| \leq Lc\|d_k\| \leq Ra_k L\|d_k\| \\ &\Longrightarrow -m_1\|\nabla f(x_k)\| cos\theta_k\| \leq Ra_k L\|d_k\| \\ &\Longrightarrow \frac{-m_1\|\nabla f(x_k)\| cos\theta_k\|}{RL} \leq a_k\|d_k\| \\ &\Longrightarrow -\left(\frac{m_1}{RL}\right)\|\nabla f(x_k)\| cos\theta_k \leq a_k\|d_k\|.\end{aligned}$$

Enquanto isso, de $f(x_k + a_k d_k) \leq f(x_k) + m_1 a_k \nabla f(x_k)^T d_k$ (da condição de Armijo), temos:

$$\begin{aligned}-m_1 a_k \nabla f(x_k)^T d_k \leq f(x_k) - f(x_k + a_k d_k) &\Longrightarrow -m_1 a_k\|\nabla f(x_k)\| \|d_k\| cos\theta_k \leq f(x_k) - f(x_k + a_k d_k) \\ &\Longrightarrow -m_1 a_k\|d_k\| \|\nabla f(x_k)\| cos\theta_k \leq f(x_k) - f(x_k + a_k d_k)\end{aligned}$$

De $-\left(\dfrac{m_1}{RL}\right)\|\nabla f(x_k)\| cos\theta_k \leq a_k\|d_k\|$, temos que

$$\begin{aligned}&- m_1\left[-\frac{m_1}{RL}\|\nabla f(x_k)\| cos\theta_k\right]\|\nabla f(x_k)\| cos\theta_k \leq f(x_k) - f(x_k + a_k d_k) \\ &\Longrightarrow \frac{m_1^2}{RL}\|\nabla f(x_k)\|^2 cos^2\theta_k \leq f(x_k) - f(x_k + a_k d_k) \\ &\Longrightarrow f(x_k + a_k d_k) - f(x_k) \leq -\frac{m_1^2}{RL}\|\nabla f(x_k)\|^2 cos^2\theta_k.\end{aligned}$$

Agora, defina $x_k + a_k d_k = x_{k+1}$. Daí: $f(x_{k+1}) \leq f(x_k) - \dfrac{m_1^2}{RL}\|\nabla f(x_k)\|^2 cos^2\theta_k$. Somando essa expressão sobre todos os índices menores ou iguais a k, temos:

$$\begin{aligned}f(x_1) &\leq f(x_0) - \frac{m_1^2}{RL} cos^2\theta_0\|\nabla f(x_0)\|^2 \\ f(x_2) &\leq f(x_1) - \frac{m_1^2}{RL} cos^2\theta_1\|\nabla f(x_1)\|^2 \\ &\vdots \\ f(x_{k+1}) &\leq f(x_k) - \frac{m_1^2}{RL} cos^2\theta_k\|\nabla f(x_k)\|^2,\end{aligned}$$

obtemos:

$$\begin{aligned}
&\sum_{j=1}^{k+1} f(x_j) \leq \sum_{j=0}^{k} f(x_j) - c\sum_{j=0}^{k} cos^2\theta_j \|\nabla f(x_j)\|^2 \\
\Longrightarrow &\sum_{j=1}^{k} f(x_j) + f(x_{k+1}) \leq f(x_0) + \sum_{j=1}^{k} f(x_j) - \frac{m_1^2}{RL}\sum_{j=0}^{k} cos^2\theta_j \|\nabla f(x_j)\|^2 \\
\Longrightarrow &f(x_{k+1}) \leq f(x_0) - \frac{m_1^2}{RL}\sum_{j=0}^{k} cos^2\theta_j \|\nabla f(x_j)\|^2 \\
\Longrightarrow &\frac{m_1^2}{RL}\sum_{j=0}^{k} cos^2\theta_j \|\nabla f(x_j)\|^2 \leq f(x_0) - f(x_{k+1}).
\end{aligned}$$

Como f é limitada inferiormente, digamos $f(x) > -M$ para alguma constante $M > 0$, então: $f(x_{k+1}) > -M \Longrightarrow -f(x_{k+1}) < M \Longrightarrow f(x_0) - f(x_{k+1}) < f(x_0) + M$ para todo $k \geq 0$. Pelo fato de $f(x_0) > f(x_{k+1})$ para $k \geq 0$, temos $\frac{m_1^2}{RL}\sum_{j=0}^{k} cos^2\theta_j \|\nabla f(x_j)\|^2 \leq f(x_0) - f(x_{k+1}) \leq f(x_0) + M$.

Aplicando o limite quando k tende a ∞ no primeiro e no último membro da desigualdade, temos

$$\begin{aligned}
&\lim_{k\to\infty}\left[\frac{m_1^2}{RL}\sum_{j=0}^{k} cos^2\theta_j \|\nabla f(x_j)\|^2\right] \leq \lim_{k\to\infty}[f(x_0) + M] \\
\Longrightarrow &\frac{m_1^2}{RL}\sum_{j=0}^{\infty} cos^2\theta_j \|\nabla f(x_j)\|^2 \leq f(x_0) + M < \infty \\
\Longrightarrow &\frac{m_1^2}{RL}\sum_{j=0}^{\infty} cos^2\theta_j \|\nabla f(x_j)\|^2 < \infty \\
\Longrightarrow &\sum_{j=0}^{\infty} cos^2\theta_j \|\nabla f(x_j)\|^2 < \infty.
\end{aligned}$$

Pela condição necessária de de convergência de séries, isso implica que $cos^2\theta_k \|\nabla f\|^2 \to 0$.

4 Método do Gradiente

O Método do Gradiente é um dos mais antigos e conhecidos na minimização de funções. Trata-se de um método iterativo que é linearmente convergente e que visa minimizar uma função utilizando a direção oposta à do vetor gradiente, que nos fornece o decrescimento mais acentuado da função objetivo. O esquema de iteração é dado por

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),\ k = 0, 1, ... \tag{13}$$

A seguir, apresentamos formalmente o algoritmo do Método do Gradiente equipado com a busca de Armijo ou de Goldstein.

Algoritmo 1: Método do Gradiente

1 Tome um ponto inicial $x_0 \in \mathbb{R}^n$, $\varepsilon > 0$ e faça $k := 0$
2 $d_k := -\nabla f(x_k)$
3 Se $\|d_k\| > \varepsilon$, calcule t_k por (5) ou (9)
4 $x_{k+1} := x_k + t_k d_k$
5 $k \leftarrow k + 1$ e retorne para o passo 2

4.1 Propriedades de convergência global

A desigualdade de Armijo (5) neste caso tem a seguinte forma:

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - \eta_1 \alpha \|\nabla f(x_k)\|^2.$$

Observe que quando $\nabla f(x_k) \neq 0$, a condição (7) é satisfeita automaticamente (com $\delta = -1$), e a estimativa (6) de passo mais longo é dada por

$$\overline{\alpha}_k = \frac{2(1 - \eta_1)}{L}. \tag{14}$$

Como esta estimativa não depende de k, quando o gradiente da função f é Lipschitz-contínuo no $\mathbb{R}^n$ e está sendo uttilizada a Busca de Armijo, pelo Lema 2.1 e por (8),

$$\alpha_k \geq \breve{\alpha} > 0, \tag{15}$$

onde $\breve{\alpha}$ não depende de k.

Teorema 4.1. ***(Convergência global do Método do Gradiente I)*** *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável no $\mathbb{R}^n$ com derivada Lipschitz-contínua no $\mathbb{R}^n$ módulo $L > 0$. Então se uma sequência $\{x_k\}$ gerada pelo Algortimo 1 equipado com a Regra de Armijo possui um ponto de acumulação, ou se a função f é limitada inferiormente no $\mathbb{R}^n$, tem-se que*

$$\{\nabla f(x_k)\} \to 0 \ (k \to \infty). \tag{16}$$

Em particular, cada ponto de acumulação de qualquer sequência $\{x_k\}$ gerada pelo Algoritmo 1 é um ponto estacionário do problema $\min f(x)$*, com $x \in \mathbb{R}^n$.*

Demonstração: Se $\nabla f(x_k) \neq 0$ para todo k, a sequência $\{f(x_k)\}$ é decrescente. Suponhamos que a sequência $\{x_k\}$ tenha um ponto de acumulação, denotado por a. Isso significa que toda bola aberta de centro a contém algum ponto de $\{x_k\}$ diferente do ponto a. Noutros termos, $\forall \varepsilon_1 > 0$, deve existir $x \in \{x_k\}$ tal que $0 < |x - a| < \varepsilon_1$. E isso também é equivalente a dizer que toda bola aberta de centro a contém uma infinidade de pontos de $\{x_k\}$.

Agora, considerando a continuidade de f, sabemos que $\forall \varepsilon > 0, \exists \delta > 0; x \in \{x_k\}, |x - x_0| < \delta \implies |f(x) - f(x_0)| < \varepsilon$, onde x_0 é um ponto fixo. Como a é ponto de acumulação da sequência $\{x_k\}$, podemos escolher um $\varepsilon > 0$ arbitrariamente pequeno. Assim, $\exists \delta > 0$ correspondente, de acordo com a continuidade da função. Observe que pela definição do ponto de acumulação basta tomarmos $\delta = \varepsilon_1$.

Agora, consideremos uma bola aberta contendo a com raio δ. Pela definição de ponto de acumulação da sequência $\{x_k\}$, sabemos que existem infinitos termos da sequência nessa bola. Consideremos então todos esses termos em uma subsequência de $\{x_k\}$ que será dada por $\{x_{k_j}\}_{j\in\mathbb{N}} = \{x_{k_1}, x_{k_2}, x_{k_3}, ...\}$. Como f é contínua, temos que $|f(x_{k_j}) - f(a)| < \varepsilon$, para todo $x_{k_j} \in \{x_{k_j}\}_{j\in\mathbb{N}}$. Portanto, $\{f(x_k)\}$ também possui um ponto de acumulação, que corresponde ao valor de $f(a)$, uma vez que existem infinitos termos dentro de qualquer intervalo aberto contendo $f(a)$.

Como $\{f(x_k)\}$ é monótona descrescente e tem um ponto de acumulação a, então existe $\{f(x_{k_j})\} \to a$. Suponha por absurdo que $\{f(x_k)\}$ não seja limitada inferiormente, ou seja, existe $c < a$ tal que $f(x_{k'}) < c < a$. Como $\{f(x_k)\}$ é decrescente, tem-se $... < f(x_{k'+2}) < f(x_{k'+1}) < f(x_{k'}) < c < a$, ou seja, $f(x_n) < c < a$ para todo $n \geq k'$, o que contradiz $f(x_{k_j}) \to a$. Isso prova que $\{f(x_k)\}$ é limitada inferiormente e, como é monótona decrescente, é também limitada superiormente (basta tomar $f(x_1)$ como limitante superior), logo, converge (pelo Teorema da Monotonicidade). (se a função f é limitada inferiormente no $\mathbb{R}^n$, a conclusão vale mesmo quando $\{x_k\}$ não tem pontos de acumulação).

Pela desigualdade de Armijo e (15), para todo k temos que

$$f(x_k) - f(x_{k+1}) \geq \eta\alpha_k\|\nabla f(x_k)\|^2 \geq \eta\breve{\alpha}_k\|\nabla f(x_k)\|^2 \tag{17}$$

Como $f(x_k) - f(x_{k+1}) \to 0 \ \ (k \to \infty)$, obtemos $\{\nabla f(x_k)\} \to 0 \ \ (k \to \infty)$ da desigualdade anterior. O fato de que cada ponto de acumulação de $\{x_k\}$ é um ponto estacionário do problema $\min f(x)$ segue de (16) e da continuidade do gradiente.

$\square$

No caso da minimização pela Regra de Armijo, é possível substituir a hipótese de que o gradiente de f seja Lipschitz-contínuo pela hipótese mais fraca de que ele seja apenas contínuo. A dificuldade neste caso tem a ver com a possível inexistência de $\breve{\alpha} > 0$ que satisfaz (15), isto é, com a possibilidade de α_k se aproximar de zero.

Teorema 4.2. ***(Convergência global do Método do Gradiente II)*** *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável no $\mathbb{R}^n$ com gradiente contínuo. Suponhamos que o Algortimo 1 utiliza a Regra de Armijo. Então cada ponto de acumulação de qualquer sequência $\{x_k\}$ gerada pelo Algoritmo 1 é um ponto estacionário do problema $\min f(x)$, com $x \in \mathbb{R}^n$. Se a sequência $\{x_k\}$ é limitada, então vale (16).*

Demonstração: Suponhamos que $\nabla f(x_k) \neq 0$ para todo k, a sequência $\{x_k\}$ tenha um ponto de acumulação $\overline{x} \in \mathbb{R}^n$, e seja $\{x_{k_j}\} \to \overline{x} \ \ (j \to \infty)$.

O caso que existe $\breve{\alpha} > 0$ tal que $\alpha_{k_j} \geq \breve{\alpha}$ para todo j, pode ser analisado de maneira análoga ao Teorema 4.1. Em particular, a sequência monótona $f(x_k)$ possui o ponto de acumulação $f(\overline{x})$. Portanto, ela converge. Além disso, por $f(x_{k+1}) < f(x_k)$ e (17), temos que

$$\begin{aligned} f(x_{k_{j+1}}) \leq f(x_{k_{j+1}-1}) \leq ... &\leq f(x_{k_j+1}) \\ &\leq f(x_{k_j}) - \eta\alpha_{k_j}\|\nabla f(x_{k_j})\|^2 \\ &\leq f(x_{k_j}) - \eta\breve{\alpha}\|\nabla f(x_{k_j})\|^2. \end{aligned}$$

Passando o limite quando $j \to \infty$, considerando a convergência de $\{f(x_k)\}$ e que η não depende de j, temos:

$$f(\overline{x}) \leq f(\overline{x}) - \breve{\alpha}\|\nabla f(\overline{x})\|^2 \Longrightarrow \breve{\alpha}\|\nabla f(\overline{x})\|^2 \leq 0 \Longrightarrow \|\nabla f(\overline{x})\|^2 \leq 0 \Longrightarrow \nabla f(\overline{x}) = 0.$$

Considerando o caso de não existir $\breve{\alpha} > 0$ tal que $\alpha_{k_j} \geq \breve{\alpha}$ para todo j, tomando uma subsequência se for necessário, podemos admitir que

$$\{\alpha_{k_j}\} \to 0 \quad (j \to \infty).$$

Neste caso, para todo j suficientemente grande, o valor inicial do comprimento de passo $\breve{\alpha}$ foi reduzido pelo menos uma vez, ou seja, o valor $\alpha = \theta^{-1}\alpha_{k_j} > \alpha_{k_j}$ não satisfaz a desigualdade de Armijo, ou seja:

$$\begin{aligned}
&f(x_{k_j} - \theta^{-1}\alpha_{k_j}\nabla f(x_{k_j})) > f(x_{k_j}) - \eta\theta^{-1}\alpha_{k_j}\|\nabla f(x_{k_j})\|^2 \\
\Longrightarrow &f(x_{k_j} - \theta^{-1}\alpha_{k_j}\nabla f(x_{k_j})) - f(x_{k_j}) > -\eta\theta^{-1}\alpha_{k_j}\|\nabla f(x_{k_j})\|^2 \\
\Longrightarrow &\frac{f(x_{k_j} - \theta^{-1}\alpha_{k_j}\nabla f(x_{k_j})) - f(x_{k_j})}{\theta^{-1}\alpha_{k_j}} > -\eta\|\nabla f(x_{k_j})\|^2.
\end{aligned}$$

Passando o limite quando $j \to \infty$ na desigualdade anterior e escrevendo $D_{\nabla f} f(x) = \nabla f(x)^T \nabla f(x) = \|\nabla f(x)\|^2$, obtemos

$$\begin{aligned}
&\lim_{j\to\infty} \frac{f(x_{k_j} - \theta^{-1}\alpha_{k_j}\nabla f(x_{k_j})) - f(x_{k_j})}{\theta^{-1}\alpha_{k_j}} > \lim_{j\to\infty} -\eta\|\nabla f(x_{k_j})\|^2 \\
\Longrightarrow &-\|\nabla f(\overline{x})\|^2 \geq -\eta\|\nabla f(\overline{x})\|^2.
\end{aligned}$$

Como $\eta \in (0,1)$, note que isso só é possível quando $\nabla f(\overline{x}) = 0$, pois caso tenhamos $\eta \in (0,1)$ e $\nabla f(\overline{x}) \neq 0$ sempre ocorrerá $-\|\nabla f(\overline{x})\|^2 < -\eta\|\nabla f(\overline{x})\|^2$.
Finalmente, se $\{x_k\}_{k\in\mathbb{N}}$ é limitada, pelo Teorema de Bolzano-Weierstrass temos que existe uma subsequência $\{x_{k_j}\}_{j\in\mathbb{N}}$ que é convergente e, consequentemente, $\{x_k\}_{k\in\mathbb{N}}$ tem um ponto de acumulação. Como já provamos que cada ponto de acumulação de qualquer sequência $\{x_k\}$ é um ponto estacionário do problema $min\ f(x), x \in \mathbb{R}^n$, então esse ponto de acumulação é um ponto estacionário do problema e concluímos que $\{\nabla f(x_k)\} \to 0$ $(k \to \infty)$.

□

Teorema 4.3. ***(Convergência global do Método do Gradiente no caso convexo)*** *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função convexa, diferenciável no $\mathbb{R}^n$, com gradiente contínuo. Suponhamos que o Algoritmo 1 utiliza a Regra de Armijo com $\hat{\alpha} \leq 1$. Se o conjunto de minimizadores irrestritos de f é não-vazio, então qualquer sequência $\{x_k\}$ gerada pelo Algoritmo 1 converge a uma solução do problema* $\min f(x)$*, com $x \in \mathbb{R}^n$.*

Demonstração: Seja $\overline{x} \in \mathbb{R}^n$ uma solução do problema. Como para todo k vale $f(\overline{x}) \leq f(x_k)$ e

$$f(x_{k+1}) \leq f(x_k) - \eta\alpha_k\|\nabla f(x_k)\|^2 \Longrightarrow f(x_k) - f(x_{k+1}) \geq \eta\alpha_k\|\nabla f(x_k)\|^2,$$

temos que

$$\begin{aligned}
f(x_0) - f(\overline{x}) &\geq f(x_0) - f(x_k) \\
&= f(x_0) + 0 - f(x_k) \\
&= f(x_0) - f(x_1) + f(x_1) - f(x_2) + ... + f(x_{k-2}) - f(x_{k-1}) + f(x_{k-1}) - f(x_k) \\
&= \sum_{i=0}^{k-1} (f(x_i) - f(x_{i+1})) \\
&\geq \eta \sum_{i=0}^{k-1} \alpha_i \|\nabla f(x_i)\|^2.
\end{aligned}$$

Passando o limite quando $k \to \infty$, obtemos que

$$\eta \sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 \leq \sum_{i=0}^{\infty} (f(x_i) - f(x_{i+1})) \leq f(x_0) - f(\overline{x}).$$

Daí, temos que

$$\sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 \leq \frac{f(x_0) - f(\overline{x})}{\eta} \Longrightarrow \sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 < +\infty. \tag{18}$$

Pela convexidade de f (veja Teorema 1.6) e pela otimalidade de $\overline{x}$, tem-se que

$$f(\overline{x}) \geq f(x_k) + \langle \nabla f(x_k), \overline{x} - x_k \rangle \Longrightarrow \langle \nabla f(x_k), \overline{x} - x_k \rangle \leq f(\overline{x}) - f(x_k) \leq 0.$$

Donde, usando também a iteração do método do gradiente ($x_{k+1} = x_k - \alpha_k \nabla f(x_k)$), obtemos

$$\begin{aligned}
\|x_{k+1} - \overline{x}\|^2 &= \|x_{k+1} - \overline{x} + x_k - x_k\|^2 \\
&= \|(x_k - \overline{x}) + (x_{k+1} - x_k)\|^2 \\
&= \langle (x_k - \overline{x}) + (x_{k+1} - x_k), (x_k - \overline{x}) + (x_{k+1} - x_k) \rangle \\
&= \langle (x_k - \overline{x}), (x_k - \overline{x}) \rangle + 2 \langle (x_k - \overline{x}), (x_{k+1} - x_k) \rangle + \langle (x_{k+1} - x_k), (x_{k+1} - x_k) \rangle \\
&= \|x_k - \overline{x}\|^2 + 2 \langle x_{k+1} - x_k, x_k - \overline{x} \rangle + \|x_{k+1} - x_k\|^2 \\
&= \|x_k - \overline{x}\|^2 + 2 \langle x_k - \alpha_k \nabla f(x_k) - x_k, x_k - \overline{x} \rangle + \|x_k - \alpha_k \nabla f(x_k) - x_k\|^2 \\
&= \|x_k - \overline{x}\|^2 + 2 \langle -\alpha_k \nabla f(x_k), x_k - \overline{x} \rangle + \| - \alpha_k \nabla f(x_k)\|^2 \\
&= \|x_k - \overline{x}\|^2 - 2\alpha_k \langle \nabla f(x_k), x_k - \overline{x} \rangle + \alpha_k^2 \|\nabla f(x_k)\|^2.
\end{aligned}$$

Como $\langle \nabla f(x_k), \overline{x} - x_k \rangle = -(\langle \nabla f(x_k), x_k - \overline{x} \rangle) \leq 0 \Longrightarrow \langle \nabla f(x_k), x_k - \overline{x} \rangle \geq 0$, temos que o termo $-2\alpha_k \langle \nabla f(x_k), x_k - \overline{x} \rangle$ sempre será negativo. Além disso, o parâmetro $\hat{\alpha}$ é escolhido da forma $0 < \alpha_k \leq \hat{\alpha} \leq 1 \Longrightarrow \alpha_k^2 \leq \alpha_k$, daí

$$\begin{aligned}
\|x_{k+1} - \overline{x}\|^2 &\leq \|x_k - \overline{x}\|^2 + \alpha_k^2 \|\nabla f(x_k)\|^2 \\
&\leq \|x_k - \overline{x}\|^2 + \alpha_k \|\nabla f(x_k)\|^2.
\end{aligned}$$

Seja k arbitrário porém fixo. Utilizando a última desigualdade em sequência, para qualquer $j \geq k+1$, obtemos

$$\begin{aligned}
\|x_j - \overline{x}\|^2 &\leq \|x_{j-1} - \overline{x}\|^2 + \alpha_{j-1}\|\nabla f(x_{j-1})\|^2 \\
&\leq \|x_{j-2} - \overline{x}\|^2 + \alpha_{j-2}\|\nabla f(x_{j-2})\|^2 + \alpha_{j-1}\|\nabla f(x_{j-1})\|^2 \\
&\vdots \\
&\leq \|x_k - \overline{x}\|^2 + \sum_{i=k}^{j-1} \alpha_i \|\nabla f(x_i)\|^2 \\
&\leq \|x_k - \overline{x}\|^2 + \sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2.
\end{aligned} \tag{19}$$

Por (18), obtemos que

$$\|x_j - \overline{x}\|^2 \leq \|x_k - \overline{x}\|^2 + \sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 < +\infty. \tag{20}$$

Concluímos que a sequência $\{x_k\}$ é limitada. Portanto, $\{x_k\}$ tem um ponto de acumualação $\hat{x}$, pelo Teorema de Bolzano-Weierstress. Pelo Teorema 4.2, temos que todo ponto de acumualação é um ponto estacionário, ou seja, $\nabla f(\hat{x}) = 0$ o que, no caso convexo, significa que $\hat{x}$ é uma solução do problema. Podemos então tomar $\overline{x} = \hat{x}$ na análise acima, para concluir de (19), que para todo $j \geq k+1$, temos

$$\|x_j - \hat{x}\|^2 \leq \|x_k - \hat{x}\|^2 + \sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2. \tag{21}$$

Além disso, temos que

$$\begin{aligned}
&\sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 = \sum_{i=0}^{k-1} \alpha_i \|\nabla f(x_i)\|^2 + \sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 \\
\Longrightarrow &\sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 = \sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 - \sum_{i=0}^{k-1} \alpha_i \|\nabla f(x_i)\|^2 \\
\Longrightarrow &\lim_{k\to\infty} \left(\sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 \right) = \lim_{k\to\infty} \left(\sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 \right) - \lim_{k\to\infty} \left(\sum_{i=0}^{k-1} \alpha_i \|\nabla f(x_i)\|^2 \right) \\
\Longrightarrow &\lim_{k\to\infty} \left(\sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 \right) = \sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 - \sum_{i=0}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 \\
\Longrightarrow &\lim_{k\to\infty} \left(\sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 \right) = 0.
\end{aligned}$$

Em particular, para todo $\delta > 0$ arbitrariamente pequeno, podemos escolher k suficientemente grande, tal que

$$\frac{\delta}{2} > \sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2.$$

Como $\hat{x}$ é um ponto de acumulação da sequência $\{x_k\}$, podemos também escolher um k tal que
$$\frac{\delta}{2} > \|x_k - \hat{x}\|^2.$$
De (21), concluímos que para todo $\delta > 0$, existe k tal que

$$\|x_j - \hat{x}\|^2 \leq \|x_k - \hat{x}\|^2 + \sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2 < \frac{\delta}{2} + \frac{\delta}{2} = \delta \Longrightarrow \|x_j - \hat{x}\|^2 < \delta, \ \ \forall j \geq k+1.$$

Isso significa que $\{x_k\}$ converge a $\hat{x}$.

$\square$

A seguir, analisaremos o comportamento local do Método do Gradiente, isto é, o comportamento numa vizinhança de um ponto estacionário $\overline{x} \in \mathbb{R}^n$ do problema $\min f(x)$, com $x \in \mathbb{R}^n$. Para provar a convergência local do método do gradiente, precisaremos do resultado a seguir, cuja demonstração pode ser vista em [3][Teoremas 1.3.1, 1.3.3 e 1.3.4].

Teorema 4.4. ***(Condições de otimalidade no caso irrestrito)***

(a) *Suponhamos que a função $f : \mathbb{R}^n \to \mathbb{R}$ seja diferenciável no ponto $\overline{x} \in \mathbb{R}^n$. Suponhamos também que $\overline{x}$ seja um minimizador local do problema* $\min f(x)$*, com $x \in \mathbb{R}^n$. Então*
$$\nabla f(\overline{x}) = 0. \tag{22}$$
Se f é duas vezes diferenciável em $\overline{x}$, então além de (22) tem-se que a matriz Hessiana de f no ponto $\overline{x}$ é semidefinida positiva, isto é,
$$\left\langle \nabla^2 f(\overline{x}) d, d \right\rangle \geq 0, \ \forall d \in \mathbb{R}^n. \tag{23}$$

(b) *Suponhamos que $f : \mathbb{R}^n \to \mathbb{R}$ seja duas vezes diferenciável no ponto $\overline{x} \in \mathbb{R}^n$. Se $\overline{x}$ satisfaz (22) e se a matriz Hessiana de f em $\overline{x}$ é semidefinida positiva, isto é, se existe $\gamma > 0$ tal que*
$$\left\langle \nabla^2 f(\overline{x}) d, d \right\rangle \geq \gamma \|d\|^2, \ \forall d \in \mathbb{R}^n, \tag{24}$$
então $\overline{x}$ é minimizador local estrito do problema $\min f(x)$*, com $x \in \mathbb{R}^n$.*

Suponhamos que a função f seja duas vezes diferenciável em $\overline{x}$, e que $\overline{x}$ satisfaça a condição suficiente de segunda ordem (veja Teorema 4.4). Em particular, $\overline{x}$ é um minimizador local estrito. Mais ainda, neste caso, f cresce com ordem pelo menos quadrática numa vizinhança de $\overline{x}$.
Com efeito, existe uma vizinhança U de $\overline{x}$ tal que, aplicando o Teorema 1.4 (b) e considerando que $o(\|x - \overline{x}\|^2) = f(\|x - \overline{x}\|^2)$ significa que a função $f : \mathbb{R}_+ \to \mathbb{R}$ possui a propriedade de que $f(\|x - \overline{x}\|^2)/\|x - \overline{x}\|^2 \to 0$ quando $\|x - \overline{x}\|^2 \to 0+$, temos
$$f(x) - f(\overline{x}) = \langle \nabla f(\overline{x}), x - \overline{x} \rangle + \frac{1}{2} \left\langle \nabla^2 f(\overline{x})(x - \overline{x}), x - \overline{x} \right\rangle + o(\|x - \overline{x}\|^2).$$
Por (22), temos
$$f(x) - f(\overline{x}) = \frac{1}{2} \left\langle \nabla^2 f(\overline{x})(x - \overline{x}), x - \overline{x} \right\rangle + o(\|x - \overline{x}\|^2).$$

Agora, aplicando (24), temos que existe um número $\beta > 0$ tal que

$$f(x) - f(\overline{x}) \geq \beta\|x - \overline{x}\|^2, \ \forall x \in U. \tag{25}$$

Esta propriedade (25) terá um papel importante em nossa análise local.

Lema 4.5. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável numa vizinhança do ponto $\overline{x} \in \mathbb{R}^n$ e duas vezes diferenciável em $\overline{x}$. Suponhamos que $\overline{x}$ seja um ponto estacionário do problema min $f(x)$ que satisfaz a condição suficiente de segunda ordem dada no Teorema 4.4 (isto é, $\overline{x}$ satisfaz (22) e (24)). Então para qualquer número $\nu \in (0, 4)$ existe uma vizinhança U de $\overline{x}$ tal que, além de (25), vale*

$$\|\nabla f(x)\|^2 \geq \nu\beta(f(x) - f(\overline{x})), \ \forall x \in U. \tag{26}$$

Demonstração: Escrevendo a expansão de Taylor de segunda ordem de $f(x)$ em torno de $\overline{x}$, temos

$$f(x) \approx f(\overline{x}) + \nabla f(\overline{x})(x - \overline{x}) + \frac{\nabla^2 f(\overline{x})}{2}(x - \overline{x})^2.$$

Fazendo a derivada em relação a x, temos

$$\nabla f(x) \approx 0 + \nabla f(\overline{x}) + \nabla^2 f(\overline{x})(x - \overline{x}).$$

Para $x \in \mathbb{R}^n$ sufientemente próximo a $\overline{x}$ e usando $\nabla f(\overline{x}) = 0$, temos que

$$\begin{aligned}\nabla f(x) = \nabla f(x) - 0 = \nabla f(x) - \nabla f(\overline{x}) &= \nabla f(\overline{x}) + \nabla^2 f(\overline{x})(x - \overline{x}) - \nabla f(\overline{x}) + o(\|x - \overline{x}\|^2) \\ &= \nabla^2 f(\overline{x})(x - \overline{x}) + o(\|x - \overline{x}\|^2).\end{aligned} \tag{27}$$

Portanto, por (25), temos que

$$\begin{aligned}f(x) - f(\overline{x}) &= \frac{1}{2}\left\langle \nabla^2 f(\overline{x})(x - \overline{x}), x - \overline{x}\right\rangle + o(\|x - \overline{x}\|^2) \\ &= \frac{1}{2}\left\langle \nabla f(x), x - \overline{x}\right\rangle + o(\|x - \overline{x}\|^2),\end{aligned}$$

isto é,

$$\langle \nabla f(x), x - \overline{x}\rangle = 2(f(x) - f(\overline{x})) + o(\|x - \overline{x}\|^2).$$

Subtraindo $\sqrt{\nu}(f(x) - f(\overline{x}))$ em ambos os lados da igualdade anterior e usando (25), obtemos

$$\begin{aligned}&\langle \nabla f(x), x - \overline{x}\rangle - \sqrt{\nu}(f(x) - f(\overline{x})) = 2(f(x) - f(\overline{x})) - \sqrt{\nu}(f(x) - f(\overline{x})) + o(\|x - \overline{x}\|^2) \\ \Longrightarrow &\langle \nabla f(x), x - \overline{x}\rangle - \sqrt{\nu}(f(x) - f(\overline{x})) = (2 - \sqrt{\nu})(f(x) - f(\overline{x})) + o(\|x - \overline{x}\|^2) \\ \Longrightarrow &\langle \nabla f(x), x - \overline{x}\rangle - \sqrt{\nu}(f(x) - f(\overline{x})) \geq (2 - \sqrt{\nu})\beta\|x - \overline{x}\|^2 + o(\|x - \overline{x}\|^2) > 0,\end{aligned}$$

onde a desigualdade vale para todo x suficientemente próximo de $\overline{x}$. Portanto,

$$\begin{aligned}&\langle \nabla f(x), x - \overline{x}\rangle - \sqrt{\nu}(f(x) - f(\overline{x})) > 0 \\ \Longrightarrow &\langle \nabla f(x), x - \overline{x}\rangle \geq \sqrt{\nu}(f(x) - f(\overline{x})).\end{aligned} \tag{28}$$

Sabemos que $|\langle \nabla f(x), x-\overline{x}\rangle| \geq \langle \nabla f(x), x-\overline{x}\rangle$. Assim, pela desigualdade de Schwarz, temos que

$$\|\nabla f(x)\|\|x-\overline{x}\| \geq |\langle \nabla f(x), x-\overline{x}\rangle| \geq \langle \nabla f(x), x-\overline{x}\rangle . \tag{29}$$

De (28), temos que

$$\|\nabla f(x)\|\|x-\overline{x}\| \geq \sqrt{\nu}(f(x)-f(\overline{x})).$$

Donde, usando também (25), segue-se que

$$\|\nabla f(x)\|\|x-\overline{x}\| \geq \sqrt{\nu}\beta\|x-\overline{x}\|^2,$$

e dividindo ambos os lados por $\|x-\overline{x}\|$ e multiplicando por $\|\nabla f(x)\|$, temos

$$\begin{aligned} &\|\nabla f(x)\| \geq \sqrt{v}\beta\|x-\overline{x}\| \\ \Longrightarrow &\|\nabla f(x)\|^2 \geq \sqrt{v}\beta\|x-\overline{x}\|\|\nabla f(x)\|. \end{aligned} \tag{30}$$

E por (29), temos

$$\|\nabla f(x)\|^2 \geq \sqrt{v}\beta \langle \nabla f(x), x-\overline{x}\rangle .$$

Portanto, usando (28),

$$\|\nabla f(x)\|^2 \geq v\beta(f(x)-f(\overline{x})),$$

o que é o resultado desejado.

Apresentamos agora o resultado principal sobre convergência local do Método do Gradiente.

Teorema 4.6. ***(Convergência local do método do gradiente I)*** *Além das hipóteses do Teorema 4.1, suponhamos que f seja duas vezes diferenciável no ponto $\overline{x} \in \mathbb{R}^n$. Suponhamos que $\overline{x}$ seja um ponto estacionário do problema min $f(x)$ que satisfaz a condição suficiente de segunda ordem dada no Teorema 4.4 (isto é, $\overline{x}$ satisfaz (22) e (24)). Além disso, consideremos o Algoritmo 1 equipado com a regra de Armijo. Então*

(a) *Para qualquer $x_0 \in \mathbb{R}^n$ suficientemente próximo a $\overline{x}$, a sequência gerada pelo Algoritmo 1 converge a $\overline{x}$; a taxa de convergência em relação à função objetivo é linear, e em relação às variáveis é geométrica: $\forall k = 0, 1, ...$, tem-se que*

$$f(x_{k+1}) - f(\overline{x}) \leq q(f(x_k) - f(\overline{x})), \tag{31}$$

$$\begin{aligned} \|x_k-\overline{x}\| &\leq \Gamma\sqrt{(f(x_k)-f(\overline{x}))} \\ &\leq (\sqrt{q})^k\Gamma\sqrt{(f(x_0)-f(\overline{x}))}, \end{aligned} \tag{32}$$

onde os números $q \in [0, 1)$ e $\Gamma > 0$ não dependem nem de k, nem de x_0.

(b) *Se $\hat{\alpha}$ satisfaz $\hat{\alpha} \geq 1/L$ e θ satisfaz a condição $\theta \leq 1/2$ tem-se que*

$$\lim_{k\to\infty} \sup \frac{f(x_{k+1})-f(\overline{x})}{f(x_k)-f(\overline{x})} \leq 1 - 2\eta(1-\eta)\rho_1/\rho_n \tag{33}$$

onde $\rho_1 > 0$ e $\rho_n > 0$ são o menor e o maior auto-valor da matriz $\nabla^2 f(\overline{x})$, respectivamente. Se no lugar da condição sobre θ supõe-se que vale $\eta \leq 1/2$, tem-se que

$$\lim_{k\to\infty} \sup \frac{f(x_{k+1})-f(\overline{x})}{f(x_k)-f(\overline{x})} \leq 1 - \theta\rho_1/\rho_n. \tag{34}$$

Demonstração: Fixemos uma vizinhança U de $\overline{x}$ tal que as propriedades (25) e (26) valem com algumas constantes $\beta > 0$ e $\nu \in (0, 4)$ (a condição suficiente de segunda ordem e o Lema 4.5 garantem a existência desta vizinhança). Seja $x_k \in U$. Escolhendo o maior número $\hat{\alpha}$, que garante que (8) se satisfaz, temos que

$$\begin{aligned} & f(x_{k+1}) \leq f(x_k) - \eta\alpha_k \|\nabla f(x_k)\|^2 \\ \Longrightarrow & f(x_{k+1}) - f(\overline{x}) \leq f(x_k) - f(\overline{x}) - \eta\alpha_k \|\nabla f(x_k)\|^2 \leq f(x_k) - f(\overline{x}) - \eta\hat{\alpha}\|\nabla f(x_k)\|^2 \end{aligned}$$

Por (26), temos

$$f(x_k) - f(\overline{x}) - \eta\hat{\alpha}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(\overline{x}) - \eta\hat{\alpha}\nu\beta(f(x_k) - f(\overline{x})) = (1 - \eta\hat{\alpha}\nu\beta)(f(x_k) - f(\overline{x})),$$

então

$$f(x_{k+1}) - f(\overline{x}) \leq (1 - \eta\hat{\alpha}\nu\beta)(f(x_k) - f(\overline{x})). \tag{35}$$

Concluímos que vale a desigualdade (31), com algum $q < 1$. Vamos provar que se x_0 está suficientemente próximo a $\overline{x}$, então a sequência $\{x_k\}$ não sai da vizinhança U. Para isso, precisaremos da seguinte desigualdade, que segue de (25):

$$\begin{aligned} & f(x) - f(\overline{x}) \geq \beta\|x - \overline{x}\|^2 \\ \Longrightarrow & \frac{f(x) - f(\overline{x})}{\beta} \geq \|x - \overline{x}\|^2 \\ \Longrightarrow & \sqrt{\frac{f(x) - f(\overline{x})}{\beta}} \geq \|x - \overline{x}\|. \end{aligned} \tag{36}$$

Agora, usando a hipótese de que o gradiente é Lipschitz-contínuo no $\mathbb{R}^n$ módulo $L > 0$ e que $\overline{x}$ é um ponto estacionário, temos que

$$\begin{aligned} & \|\nabla f(x) - \nabla f(\overline{x})\| \leq L\|x - \overline{x}\| \\ \Longrightarrow & \|\nabla f(x)\| \leq L\|x - \overline{x}\| \end{aligned}$$

Usando a desigualdade anterior e (36), temos

$$\|\nabla f(x)\| \leq L\|x - \overline{x}\| \leq L\sqrt{\frac{f(x) - f(\overline{x})}{\beta}}, \ \forall x \in U. \tag{37}$$

Fixemos $r > 0$ tal que $B(\overline{x}, r) \subset U$, e definamos $\delta > 0$ satisfazendo a condição

$$\delta + \frac{\hat{\alpha}L\sqrt{\frac{f(x) - f(\overline{x})}{\beta}}}{1 - \sqrt{|q|}} \leq r, \ \forall x \in B(\overline{x}, \delta). \tag{38}$$

Notamos que $\delta \leq r$.
Agora, vamos provar que $x_k \in U$, $\forall k \in \mathbb{N}$, por indução sob k.
Seja $x_0 \in B(\overline{x}, \delta)$, temos por (13) que $\|x_1 - x_0\| = \hat{\alpha}\|\nabla f(x_0)\|$. Pela desigualdade triangular, temos

$$\begin{aligned} \|x_1 - \overline{x}\| = \|x_1 - \overline{x} + x_0 - x_0\| &\leq \|x_0 - \overline{x}\| + \|x_1 - x_0\| \\ &\leq \delta + \hat{\alpha}\|\nabla f(x_0)\|. \end{aligned}$$

Por (37) e (38),

$$\begin{aligned}
\|x_1 - \overline{x}\| &\leq \delta + \hat{\alpha}\|\nabla f(x_0)\| \\
&\leq \delta + \hat{\alpha}L\sqrt{\frac{(f(x_0) - f(\overline{x}))}{\beta}} \\
&\leq \delta + \frac{\hat{\alpha}L\sqrt{\frac{f(x_0)-f(\overline{x})}{\beta}}}{1-\sqrt{|q|}} \\
&\leq r,
\end{aligned}$$

isto é, $x_1 \in B(\overline{x}, r)$. Assim, por (25), segue-se que $q \geq 0$, pois caso contrário (31) seria inválida para $k = 0$. Suponhamos que $x_k \in B(\overline{x}, r)$, $\forall k = 1, ..., n$, temos que $\|x_k - \overline{x}\| \leq r$. Neste caso, aplicando (31) sucessivas vezes, obtemos

$$\begin{aligned}
f(x_k) - f(\overline{x}) &\leq q(f(x_{k-1}) - f(\overline{x})) \\
&\leq q^2(f(x_{k-2}) - f(\overline{x})) \\
&\leq q^3(f(x_{k-3}) - f(\overline{x})) \\
&\quad\vdots \\
&\leq q^k(f(x_0) - f(\overline{x})),\ \forall k = 0, ..., n.
\end{aligned}$$

Assim, temos que

$$\sqrt{f(x_k) - f(\overline{x})} \leq \sqrt{q^k(f(x_0) - f(\overline{x}))} = \sqrt{q^k}\sqrt{f(x_0) - f(\overline{x})} = (\sqrt{q})^k\sqrt{f(x_0) - f(\overline{x})}. \quad (39)$$

Portanto, usando também (13), (37) e como estamos escolhendo o maior número $\hat{\alpha}$, temos

$$\begin{aligned}
\|x_{k+1} - x_k\| &\leq \hat{\alpha}\|\nabla f(x_k)\| \\
&\leq \hat{\alpha}L\sqrt{\frac{(f(x_k) - f(\overline{x}))}{\beta}},\ \forall k = 0, ..., n.
\end{aligned}$$

Usando (39), temos

$$\begin{aligned}
\|x_{k+1} - x_k\| &\leq \hat{\alpha}L\frac{\sqrt{f(x_k) - f(\overline{x})}}{\sqrt{\beta}} \\
&\leq \hat{\alpha}L\frac{(\sqrt{q})^k\sqrt{f(x_k) - f(\overline{x})}}{\sqrt{\beta}} \\
&= (\sqrt{q})^k\hat{\alpha}L\sqrt{\frac{(f(x_0) - f(\overline{x}))}{\beta}},\ \forall k = 0, ..., n.
\end{aligned}$$

Daí, pela Desigualdade Triangular e por (38), segue-se que

$$\begin{aligned}
\|x_{n+1}-\overline{x}\| \leq \|x_n-\overline{x}\|+\|x_{n+1}-x_n\| \leq ... &\leq \|x_0-\overline{x}\|+\sum_{k=0}^{n}\|x_{k+1}-x_k\| \\
&\leq \delta+\sum_{k=0}^{n}(\sqrt{q})^k\hat{\alpha}L\sqrt{\frac{(f(x_0)-f(\overline{x}))}{\beta}} \\
&\leq \delta+\sum_{k=0}^{\infty}(\sqrt{q})^k\hat{\alpha}L\sqrt{\frac{(f(x_0)-f(\overline{x}))}{\beta}} \qquad (40)\\
&= \delta+\left(\frac{1}{1-\sqrt{q}}\right)\cdot\hat{\alpha}L\sqrt{\frac{(f(x_0)-f(\overline{x}))}{\beta}} \\
&= \delta+\frac{\hat{\alpha}L\sqrt{\frac{(f(x_0)-f(\overline{x}))}{\beta}}}{1-\sqrt{q}} \leq r,
\end{aligned}$$

isto é, $x_{n+1}\in B(\overline{x},r)$. Assim, acabamos de mostrar $\{x_k\}\subset U$. Em particular, para todo k vale a estimativa (31).
Daí, usando também (25) e (39), com $\Gamma=1/\sqrt{\beta}$, obtemos (32):

$$\begin{aligned}
\|x_k-\overline{x}\| &\leq \frac{1}{\sqrt{\beta}}\sqrt{f(x_k)-f(\overline{x})} \\
&\leq (\sqrt{q})^k\Gamma\sqrt{f(x_0)-f(\overline{x})}
\end{aligned}$$

Passando o limite na desigualdade acima, obtemos:

$$\begin{aligned}
&\lim_{k\to\infty}(\|x_k-\overline{x}\|) \leq \lim_{k\to\infty}(\sqrt{q})^k\Gamma\sqrt{f(x_0)-f(\overline{x})} \\
\Longrightarrow &\lim_{k\to\infty}(\|x_k-\overline{x}\|) \leq 0 \\
\Longrightarrow &\lim_{k\to\infty}(\|x_k-\overline{x}\|) = 0 \\
\Longrightarrow &x_k\to\overline{x}.
\end{aligned}$$

Isso conclui a prova do item (a).
Para provar (b) vamos utilizar (a) e a propriedade (35), estimativas assintóticas do número β e da constante de Lipschitz da derivada de f em U em termos de ρ_1 e ρ_n, respectivamente, reduzindo a vizinhança U de $\overline{x}$. Além disso, deve ser utilizado o fato de que no Lema 4.5 o número ν pode ser escolhido arbitrariamente próximo a 4. Na demonstração do item (b), será necessário escolher o maior número $\hat{\alpha}$, que garante que (8) se satisfaz. □

Teorema 4.7. ***(Convergência local do Método do Gradiente II)*** *Além das hipóteses do Teorema 4.1, suponhamos que o valor ótimo do problema* $\min f(x)$*, com* $x\in\mathbb{R}^n$ *seja finito, e que (3.44) seja satisfeita para* $U=L_{\|\nabla f(\cdot)\|,\mathbb{R}^n}(\delta)$ *e alguns* $\delta>0$ *e* $\gamma>0$*. Seja satisfeita também a condição de separação de superfícies de nível críticas de* f*. Suponhamos que o*

Algoritmo 1 usa a regra de Armijo. Então qualquel sequência $\{x_k\}$ do Algoritmo 1 converge a um ponto estacionário do problema $\min f(x)$*, com* $x \in \mathbb{R}^n$ *. A taxa de convergência é linear em relação à função e é geométrica em relação às variáveis.*

Demonstração: Veja a demonstração em [3, Teorema 3.1.23].

5 Função Quociente de Rayleigh

O sistema do Google utiliza um método chamado PageRank que ordena os resultados de uma busca de acordo com a importância. Em [1] é argumentado que para definir a importância de uma página pode-se modelar matematicamente o problema de encontrar o autovetor associado ao maior autovalor da chamada matriz do Google. Nesse caso, a função Quociente de Rayleigh, definida a seguir, é utilizada no problema envolvido.

5.1 Definição e gradiente

A função Quociente de Rayleigh é definida por $f : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$, tal que

$$f(x) = \frac{x^T A x}{x^T x}, \tag{41}$$

onde A é uma matriz simétrica de ordem $n \times n$ e $(\cdot)^T$ denota a matriz transposta. O valor mínimo da função definida em (41) é o menor autovalor da matriz simétrica A que define essa função. Vamos calcular $\nabla f(x)$.
Considere $f_1(x) = x^T A x$ e $f_2(x) = x^T x$ e considere a condição de diferenciabilidade:

$$f(x+v) = f(x) + \nabla f(x)[v] + r(v), \quad \text{com} \lim_{\|v\| \to 0} \frac{r(v)}{\|v\|} = 0.$$

Para $f_1(x)$, temos que:

$$\begin{aligned} f_1(x+v) &= (x+v)^T A(x+v) \\ &= (x^T + v^T)A(x+v) \\ &= x^T A x + x^T A v + v^T A x + v^T A v \\ &= \underbrace{x^T A x}_{f_1(x)} + \underbrace{2x^T A v}_{\nabla f_1(x)[v]} + \underbrace{v^T A v}_{r(v)}. \end{aligned}$$

Vamos mostrar que $\lim_{\|v\| \to 0} \frac{v^T A v}{\|v\|} = 0$. Pela desigualdade de Cauchy-Schwarz, temos:

$$
\begin{aligned}
&0 \leq v^T A v \leq |v^T A v| \leq \|v^T\| \|Av\| \\
\Longrightarrow &0 \leq \frac{v^T A v}{\|v\|} \leq \frac{|v^T A v|}{\|v\|} \leq \frac{\|v^T\| \|Av\|}{\|v\|} \\
\Longrightarrow &0 \leq \frac{v^T A v}{\|v\|} \leq \|Av\| \leq \|A\| \|v\| \quad \text{(norma matricial induzida)} \\
\Longrightarrow &\lim_{\|v\| \to 0} 0 \leq \lim_{\|v\| \to 0} \frac{v^T A v}{\|v\|} \leq \lim_{\|v\| \to 0} \|A\| \|v\| \\
\Longrightarrow &0 \leq \lim_{\|v\| \to 0} \frac{v^T A v}{\|v\|} \leq 0 \\
\Longrightarrow &\lim_{\|v\| \to 0} \frac{v^T A v}{\|v\|} = 0.
\end{aligned}
$$

Para $f_2(x)$, temos que:

$$
\begin{aligned}
f_2(x+v) &= (x+v)^T (x+v) \\
&= (x^T + v^T)(x+v) \\
&= x^T x + x^T v + v^T x + v^T v \\
&= \underbrace{x^T x}_{f_2(x)} + \underbrace{2x^T v}_{\nabla f_2(x)[v]} + \underbrace{v^T v}_{r(v)}.
\end{aligned}
$$

Vamos mostrar que $\lim_{\|v\| \to 0} \frac{v^T v}{\|v\|} = 0$. Pela desigualdade de Cauchy-Schwarz, temos:

$$
\begin{aligned}
&0 \leq v^T v \leq |v^T v| \leq \|v\| \|v\| \\
\Longrightarrow &0 \leq \frac{v^T v}{\|v\|} \leq \frac{|v^T v|}{\|v\|} \leq \frac{\|v\| \|v\|}{\|v\|} \\
\Longrightarrow &0 \leq \frac{v^T v}{\|v\|} \leq \|v\| \\
\Longrightarrow &\lim_{\|v\| \to 0} 0 \leq \lim_{\|v\| \to 0} \frac{v^T v}{\|v\|} \leq \lim_{\|v\| \to 0} \|v\| \\
\Longrightarrow &0 \leq \lim_{\|v\| \to 0} \frac{v^T v}{\|v\|} \leq 0 \\
\Longrightarrow &\lim_{\|v\| \to 0} \frac{v^T v}{\|v\|} = 0.
\end{aligned}
$$

Agora, utilizamos a Regra do Quociente para calcular $\nabla f(x)$:

$$
\begin{aligned}
\nabla f(x) &= \frac{f_1'(x)f_2(x) - f_1(x)f_2'(x)}{[f_2(x)^2]} \\
&= \frac{2Axx^Tx - x^TAx2x}{(x^Tx)^2} \\
&= \frac{1}{(x^Tx)^2}\left(2Axx^Tx - \left(\frac{x^Tx}{x^Tx}\right)x^TAx2x\right) \\
&= \frac{2x^Tx}{(x^Tx)^2}\left(Ax - \frac{x^TAx}{x^Tx}x\right) \\
&= \frac{2}{x^Tx}(Ax - f(x)Ix) \\
&= \frac{2(A - f(x)I)x}{x^Tx}.
\end{aligned}
$$

6 Experimentos Numéricos

Nesta seção, apresentamos os resultados de alguns experimentos numéricos que realizamos com base na teoria discutida anteriormente.

Os experimentos numéricos foram realizados no software VSCode com a linguagem de programação Julia em uma máquina com processador Intel Core i5 2.11 GHZ com 8.00 GB de memória RAM e sistema operacional Windows 11 Home 64 bits.

Em todos os experimentos utilizamos o critério de parada $\|\nabla f(\overline{x})\| \leq 10^{-6}$, onde $\| \cdot \|$ é a norma Euclidiana. No caso de uma matriz $A \in \mathbb{R}^{n^2}$ temos que $\|A\| = \sqrt{tr(A^TA)}$. Utilizamos também como critério de parada o máximo de 10000 iteradas.

Como vimos na seção anterior, o Algoritmo 1 representa o Método do Gradiente equipado com as buscas de Armijo ou Goldstein para minimizar a função Quociente de Rayleigh. Para a apresentação dos testes numéricos chamaremos esse método equipado com a Busca de Armijo de MGA e equipado com a Busca de Goldstein de MGG. Nós aplicamos os algoritmos em uma coleção de problemas dados por várias dimensões diferentes para a matriz A definida no Quociente de Rayleigh. No estudo em que fizemos, consideramos matrizes A simétricas de dimensões $n = 2, 3, 5, 10, 50, 80, 100, 150, 200, 300$. As matrizes foram geradas da forma $A = 0,5^*B^*B'$ onde $B = rand(snd, n, n)$ e $snd = MersenneTwister(12345)$. Além disso, escolhemos 5 pontos iniciais para cada dimensão, totalizando 50 problemas.

Os pontos inicias x_0 foram escolhidos de maneira aleatória usando $x_0 = rand(snd, n, 1)$. Em geral, foi possível observar que o MGG é superior ao MGA em tempo de processamento. Veja na Figura 2 que o MGA consegue resolver pouco mais de 25 por cento dos problemas inicialmente, enquanto o MGG resolve 80 por cento e chega a resolver 100 por cento dos problemas antes do MGA.

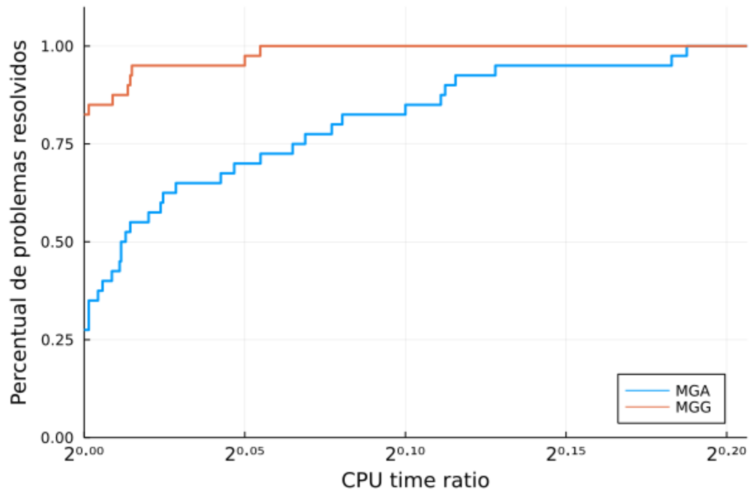


Figura 2: Perfomance profile relativo ao tempo de CPU para as dimensões $2, 3, 5, 10, 50, 80, 100, 150, 200, 300$.

7 Aplicação da Busca linear de Goldstein

Nesta última seção, apresentamos um exemplo de aplicação da busca de Goldstein com o auxílio de ferramentas de visualização gráfica.

Com o objetivo de obter uma aproximação do valor mínimo global da função f, vamos aplicar as condições de Goldstein para encontrarmos comprimentos de passos t_k que serão aplicados para decrescer a função f em uma determinada direção.

Exemplo 7.1. *Seja $f : \mathbb{R}^2 \to \mathbb{R}$ definida por $f(x;y) = \frac{1}{2}(x-2)^2 + (y-1)^2$. Utilizando os resultados clássicos do Cálculo Diferencial encontramos que a função f admite um mínimo global no ponto $x^* = (2;1)$. Assim, dado $x_0 = (0;-2)$ vamos aplicar a condição de Goldstein em uma direção de descida que será obtida pelo método do gradiente a partir de x_0. Como o gradiente da função f é dado por $\nabla f(x;y) = (x-2; 2(y-1))$, temos a direção $d_0 = -\nabla f(0;-2) = (2;6)$. A direção d_0 é de descida, pois*

$$\nabla f(x_0)^T d_0 = \nabla f \begin{pmatrix} 0 \\ -2 \end{pmatrix}^T \begin{pmatrix} 2 \\ 6 \end{pmatrix} = \begin{pmatrix} -2 & -6 \end{pmatrix} \begin{pmatrix} 2 \\ 6 \end{pmatrix} = -4 - 36 = -40 < 0.$$

Daí, a relação $f(x_0) + (1-c)t_0 \nabla f(x_0)^T d_0 \leq f(x_0 + t_0 d_0) \leq f(x_0) + c t_0 \nabla f(x_0)^T d_0$, com $0 < c < 1/2$, pode ser escrita como

$$f\begin{pmatrix} 0 \\ -2 \end{pmatrix} + (1-c)t_0 \begin{pmatrix} -2 & -6 \end{pmatrix} \begin{pmatrix} 2 \\ 6 \end{pmatrix} \leq f\begin{pmatrix} 2t_0 \\ -2+6t_0 \end{pmatrix} \leq f\begin{pmatrix} 0 \\ -2 \end{pmatrix} + c t_0 \begin{pmatrix} -2 & -6 \end{pmatrix} \begin{pmatrix} 2 \\ 6 \end{pmatrix}.$$

Donde obtemos

$$\frac{20c}{19} \leq t_0 \leq \frac{20(1-c)}{19}. \tag{42}$$

Por exemplo, se tomarmos $c = 1/4$, então qualquer

$$0,26 \leq t_0 \leq 0,78 \tag{43}$$

assegura a desigualdade (42). Assim, começando com $t_0 = 0,6$, *o passo será aceito, pois satisfaz (43) e obteremos* $x_0 + t_0 d_0 = (1,2; 1,6) = x_1$. *Donde segue que* $f(x_1) = 0,68 < 11 = f(x_0)$, *isto é, o comprimento de passo* $t_0 = 0,6$ *fornece um decrescimento da função* f *na direção de descida* d_0 *a partir de* x_0. *Agora, vamos continuar decrescendo a função* f *a partir de* $x_1 = (1,2; 1,6)$, *na direção* $d_1 = -\nabla f(1,2; 1,6) = (0,8; -1,2)$. *Temos que* d_1 *é uma direção de descida, pois*

$$\nabla f(x_1)^T d_1 = \begin{pmatrix} -0,8 & 1,2 \end{pmatrix} \begin{pmatrix} 0,8 \\ -1,2 \end{pmatrix} = -2,08 < 0.$$

Assim, vamos escrever a relação $f(x_1) + (1-c)t_1 \nabla f(x_1)^T d_1 \leq f(x_1 + t_1 d_1) \leq f(x_1) + ct_1 \nabla f(x_1)^T d_1$, *com* $0 < c < 1/2$, *como*

$$\begin{aligned} f\begin{pmatrix} 1,2 \\ 1,6 \end{pmatrix} + (1-c)t_1 \begin{pmatrix} -0,8 & 1,2 \end{pmatrix} \begin{pmatrix} 0,8 \\ -1,2 \end{pmatrix} &\leq f\begin{pmatrix} 1,2 + 0,8t_1 \\ 1,6 - 1,2t_1 \end{pmatrix} \\ &\leq f\begin{pmatrix} 1,2 \\ 1,6 \end{pmatrix} + ct_1 \begin{pmatrix} -0,8 & 1,2 \end{pmatrix} \begin{pmatrix} 0,8 \\ -1,2 \end{pmatrix}, \end{aligned}$$

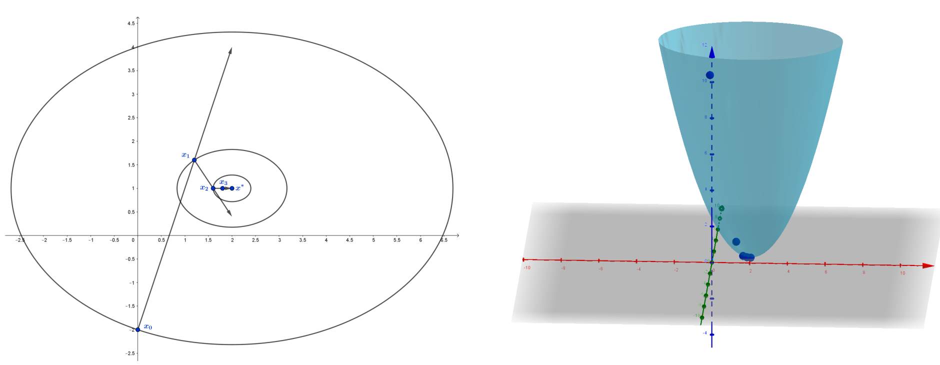
donde obtemos

$$\frac{0,13c}{0,11} \leq t_1 \leq \frac{0,13(1-c)}{0,11}. \tag{44}$$

Tomando $c = 1/3$, *então qualquer*

$$0,39 \leq t_1 \leq 0,78 \tag{45}$$

assegura a desigualdade (44). Assim, basta fazermos $t_1 = 0,5$ *e teremos* $x_1 + t_1 d_1 = (1,6; 1) = x_2$, *donde segue que* $f(x_2) = 0,08 < 0,68 = f(x_1)$. *Analogamente ao procedimento anterior, façamos uma última aplicação das condições de Goldstein na direção de descida* $d_2 = -\nabla f(1,6; 1) = (0,4; 0)$ *a partir de* $x_2 = (1,6; 1)$. *Tomando* $c = 1/5$, *basta fazermos* $t_2 = 0,5$ *e obteremos* $x_2 + t_2 d_2 = (1,8; 1) = x_3$. *Na Figura 3(a) podem ser observadas as curvas de nível de* f. *Além disso, na Figura 3(b) podemos observar as imagens da função* f *em* x_0, x_1, x_2 *e* x_3.



(a) Curvas de nível da função f. (b) Valores de f em x_0, x_1, x_2 e x_3.

Figura 3: Curvas de nível e gráfico da função f.

Note que $f(x_{k+1}) < f(x_k)$, *com* $k = 0, 1, 2, 3$. *A Tabela 1 apresenta a sequência dos pontos obtidos usando as condições de Goldstein. Nesta tabela,* k *é o índice do termo da sequência,* x_k *é o termo da sequência,* d_k *é a direção de descida a partir de* x_k, t_k *é o comprimento de passo e* $f(x_k)$ *é o valor da função.*

k	x_k	d_k	t_k	$f(x_k)$
0	(0; -2)	(2; 6)	0.6	11
1	(1.2; 1.6)	(0.8; -1.2)	0.5	0.68
2	(1.6; 1)	(0.4; 0)	0.9	0.08
3	(1.8; 1)	—	—	0.02
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Tabela 1: Iteradas obtidas pelas condições de Goldstein para se aproximar do valor mínimo global da função f.

Observe que na última coluna $f(x_k)$ *está diminuindo a cada iterada, indicando que* $f(x_k)$ *está cada vez mais próximo de 0, que corresponde ao valor de mínimo global da função* f.

Referências

[1] K. Bryan and T. Leise. The $25,000,000,000 eigenvector: The linear algebra behind google. *SIAM review*, 48(3):569–581, 2006.

[2] A. Izmailov and M. Solodov. *Otimização, Volume 2: métodos computacionais*. IMPA, 2018.

[3] A. Izmailov and M. Solodov. *Otimização, Volume 1*. Rio de Janeiro: IMPA, 2020.

[4] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.