PROGRAD
Pró-Reitoria de Graduação

PETIMAT
Programa de Educação Tutorial Institucional de Matemática

Notas de estudo

Estudos teóricos sobre o Método do Gradiente

Maria Clara Brito dos Reis

Vitória da Conquista, 2023

Conteúdo

1 Noções introdutórias

1.1 Definições preliminares

1.1.1 Conjunto Compacto

Dizemos que um conjunto X é *compacto* quando:
i) para todo $x \in X$, temos $a < x < b$, fixados $a, b \in \mathbb{R}$
ii) toda sequência de pontos $x_n \in X$ converge para um ponto $c \in X$.

1.1.2 Ponto de acumulação

O ponto $a \in \mathbb{R}$ é um *ponto de acumulação* do conjunto $X \subset \mathbb{R}$ quando, para todo $\epsilon > 0$, tem-se $X \cap (a - \epsilon, a + \epsilon) \neq \{a\}$.

1.1.3 Ponto estacionário

Seja $f : \mathbf{R}^n \to \mathbf{R}$. O ponto $x \in \mathbf{R}^n$ é *ponto estacionário* de f se as derivadas parciais de f nesse ponto são nulas ou não existem.

1.1.4 Produto interno

Um *produto interno* sobre o conjunto V é uma função associa cada par de vetores v_1, $v_2 \in V$ um número real, denotado $\langle v_1, v_2 \rangle$, que satisfaz as seguintes condições para todo $v_1, v_2, v_3 \in V$:
i) $\langle v_1, v_1 \rangle \geq 0$;
ii) $\langle v_1, v_1 \rangle = 0$ se, e somente se, $v_1 = 0$;
iii) $\langle \alpha v_1, v_2 \rangle = \alpha \langle v_1, v_2 \rangle$, $\alpha \in \mathbb{R}$;
iv) $\langle v_1 + v_2, v_3 \rangle = \langle v_1, v_3 \rangle + \langle v_2, v_3 \rangle$;
v) $\langle v_1, v_2 \rangle = \langle v_2, v_1 \rangle$;

1.1.5 Matriz não-singular

Dada uma matriz A de ordem $n \times n$. Dizemos que A é não-singular se e somente se $\det(A) \neq 0$, onde $\det(A)$ denota o determinante de A. Em outras palavras, uma matriz não-singular é aquela que possui uma inversa multiplicativa, ou seja, pode ser invertida.

1.2 Funções

1.2.1 Funções diferenciáveis

Uma função $f : U \to \mathbb{R}$, definida no aberto $U \subset \mathbb{R}^n$, é diferenciável em $a \in U$ quando para todo $v = (\alpha_1, ..., \alpha_n)$ tal que $a + v \in U$, tem-se

$$f(a + v) = f(a) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a) \cdot \alpha_i + r(v), \text{ em que } \lim_{|v| \to 0} \frac{r(v)}{|v|} = 0$$

1.2.2 Funções Lipschitziana

Uma função $f : X \to \mathbb{R}$ chama-se *lipschitziana* quando existe uma constante $k > 0$ (chamada constante de *Lipschitz* da função f) tal que

$$|f(x) - f(y)| \leq k|x - y|$$

sejam quais forem $x, y \in X$. A fim de que $f : X \to \mathbb{R}$ seja lipschitziana é necessário e suficiente que o quociente $\frac{|f(y)-f(x)|}{|y-x|} \leq k$.

1.2.3 Funções Convexas

Uma função $f : \mathbb{R}^n \to \mathbb{R}$ é considerada convexa se, para todos x, y no domínio da função e para todo t no intervalo $[0, 1]$, a seguinte desigualdade é satisfeita:

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$

Isso significa que a função está localizada abaixo do segmento de linha que conecta os pontos $f(x)$ e $f(y)$ para todos os pontos x e y no domínio da função e para qualquer valor t entre 0 e 1.

1.3 Resultados Importantes

1.3.1 Teorema de Bolzano - Weierstrass

Toda sequência limitada possui uma subsequência convergente.

1.3.2 Teorema do Valor Médio

Se para $x, y \in \mathbf{R}^n$ uma função $F : \mathbf{R}^n \to \mathbf{R}^l$ é contínua no intervalo $\{x + ty | t \in [0, 1]\}$ e diferenciável em $\{x + ty | t \in (0, 1)\}$, então:

$$||F(x + y) - F(x)|| \leq \ sup_{t\in(0,1)} \ ||F'(x + ty)|| \ ||y||.$$

2 Introdução á Otimização Irrestrita

O conceito de Otimização refere-se ao processo de encontrar os pontos mínimos e/ou máximos globais de funções. Em termos formais, a Otimização Irrestrita consiste em resolver problemas do tipo

$$\min f(x), \text{ sujeito a } x \in \mathbf{R}^n, \text{ em que } f : \mathbf{R}^n \to \mathbf{R} \tag{1}$$

2.0.1 Minimizador global

Dizemos que o ponto $\bar{x} \in \Omega$ é **minimizador global** de (1), se

$$f(\bar{x}) \leq f(x), \forall\ x \in \Omega.$$

Para encontrar o minimizador global de f, recorremos a diferentes métodos com o objetivo de gerar uma sequência que convirja para $\bar{x}$. Esses métodos começam com um ponto inicial x_0, frequentemente chamado de "**chute**", e, a partir de diferentes técnicas, obtêm um ponto melhor x_1. As informações derivadas dos pontos anteriores definem uma sequência que gradualmente se aproxima da solução ótima do problema, isto é, do minimizador $\bar{x}$.

2.1 Condições de otimalidade para casos irrestritos

2.1.1 Teorema (Weierstrass):

Seja a função contínua $f : D \to \mathbf{R}$, em que $D \subset \mathbf{R}^n$ é um conjunto compacto não-vazio. Então o problema (1) tem uma solução global. A implicação fundamental do teorema de Weierstrass é que, sob as condições estabelecidas, sempre podemos encontrar uma solução global para problemas de minimização.

2.1.2 Teorema

Se $f : \mathbf{R}^n \to \mathbf{R}$ é diferenciável em $\bar{x} \in \mathbf{R}^n$, em que $\bar{x}$ é solução do problema (1). Então $\bar{x}$ é ponto estacionário de f.

Definição: Um algorítmo é dito globalmente convergente quando para qualquer sequência (x^k) gerada pelo algoritmo e qualquer ponto de acumulação $\bar{x}$ de (x^k) temos que $\bar{x}$ é estacionário.

2.2 Métodos de descida

Para encontrar o minimizador de uma função f a partir de uma aproximação x^k da solução do problema, buscamos um ponto x^{k+1} tal que $f(x^{k+1}) < f(x^k)$. Para isso, tomamos uma direção d^k e determinamos um comprimento de passo $\alpha_k > 0$ de modo que

$$f(x^k + \alpha_k d^k) < f(x^k).$$

Essa estratégia de atualização dos pontos permite que nos movamos em direção às regiões onde a função decresça, utilizando uma combinação apropriada de direção e comprimento de passo para guiar o processo de iterativo em direção à solução desejada.

2.2.1 Direção de descida

Dizemos que $d \in \mathbf{R}^n$ é uma *direção de descida da função* $f : \mathbf{R}^n \to \mathbf{R}$ *no ponto* $x \in \mathbf{R}^n$, se existe $\epsilon > 0$ tal que

$$f(x + td) < f(x) \forall t \in (0, \epsilon] \tag{2}$$

Denotamos por $D_f(x)$ o conjunto de todas as direções de descida da função f no ponto x.

2.2.2 Teorema (Direções de descida)

Seja $f : \mathbf{R}^n \to \mathbf{R}$ uma *função diferenciável no ponto* $x \in \mathbf{R}^n$. Então:
a) Para todo $d \in D_f(x)$, tem-se $\langle f'(x), d\rangle \leq 0$.
b) Se $d \in \mathbf{R}^n$ satisfaz $\langle f'(x), d\rangle < 0$, tem-se que $d \in D_f(x)$.

Demonstração: Seja $d \in D_f(x)$. Para todo $t > 0$ suficientemente pequeno, pela diferenciabilidade de f em x,

$$0 > f(x + td) - f(x) = t(\langle f'(x), d\rangle + o(t)/t.$$

Dividindo ambos lados da desigualdade por $t > 0$ e passando o limite quando $t \to 0+$, obtemos $0 \geq \langle f'(x), d\rangle$. Supondo agora que $\langle f'(x), d\rangle < 0$. Pela diferenciabilidade de f em x, temos

$$f(x + td) - f(x) = t(\langle f'(x), d\rangle + o(t)/t).$$

Em particular, para $t > 0$ suficientemente pequeno, temos

$$\langle f'(x), d\rangle + o(t)/t \leq \tfrac{1}{2}\langle f'(x), d\rangle < 0$$

o que implica que $d \in D_f(x)$.

3 Taxas de convergência

A taxa de convergência é essencial para avaliar a velocidade com que uma sequência de aproximações converge para a solução do problema.

A **convergência linear** ocorre quando a distância entre as aproximações sucessivas diminui linearmente a cada iteração. Isto é,

$$|x^{k+1} - \bar{x}| \leq C|x^k - \bar{x}|,$$

onde C é uma constante positiva menor que 1 e $\bar{x}$ é a solução do problema. Em outras palavras, a cada iteração, a aproximação melhora em uma proporção constante em relação à sua distância até a solução. Embora a convergência linear represente uma melhoria gradual, ela pode ser relativamente lenta quando comparada a outras taxas.

4 Buscas Lineares

4.1 Armijo

A *regra de Armijo* consiste em computar um comprimento de passo que resulta em um decréscimo suficiente da função f em relação ao valor $f(x_k)$, que garanta convergência, isto é

$$f(x^k + \alpha d^k) \leq f(x^k) + \sigma\alpha\langle f'(x^k), d^k\rangle$$

em que f é diferenciável no ponto x^k, $\alpha > 0$ e $\sigma \in (0, 1)$.

4.1.1 Lema (A regra de Armijo está bem definida)

Seja $f : \mathbf{R}^n \to \mathbf{R}$ uma função diferenciável no ponto $x^k \in \mathbf{R}^n$. Suponhamos que $d^k \in \mathbf{R}^n$ satisfaça

$$\langle f'(x^k), d^k\rangle < 0.$$

Então a desigualdade

$$f(x^k + \alpha d^k) \leq f(x^k) + \sigma\alpha\langle f'(x^k), d^k\rangle$$

é satisfeita para todo α suficientemente pequeno. Em particular, a regra de Armijo está bem definida e termina com um $\alpha_k > 0$.

Demonstração: Como a função f é diverenciável no ponto x^k, temos

$$f(x^k + \alpha d^k) - f(x^k) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^k) \cdot \alpha d^k + r(\alpha) \text{ com } \lim_{|\alpha|\to 0} \frac{r(\alpha)}{|v|} = 0$$

Além disso,

$$\langle f'(x^k), \alpha d^k\rangle < 0.$$

Segue que

$$\begin{aligned} f(x^k + \alpha d^k) - f(x^k) &= \langle f'(x^k), \alpha d^k\rangle + r(\alpha) \\ &= \sigma\alpha\langle f'(x^k), d^k\rangle + (1-\sigma)\alpha\langle f'(x^k), d^k\rangle + r(\alpha) \\ &= \sigma\alpha\langle f'(x^k), d^k\rangle + \alpha((1-\sigma)\langle f'(x^k), d^k\rangle + r(\alpha)/\alpha) \end{aligned}$$

Note que

$$\langle f'(x^k), \alpha d^k\rangle \leq \sigma\alpha\langle f'(x^k), d^k\rangle,\ \sigma \in (0, 1).$$

Logo,

$$f(x^k + \alpha d^k) - f(x^k) \leq \sigma\alpha\langle f'(x^k), d^k\rangle.$$

Interpretação do Lema: A *regra de Armijo* produz uma valor $\alpha_k > 0$ aceitável, após um número finito de reduções do valor inicial α.

4.1.2 Lema (Cota inferior para o valor do comprimento de passo dado pela regra de Armijo)

Seja $f : \mathbf{R}^n \to \mathbf{R}$ uma função diferenciável no $\mathbf{R}^n$, com derivada *Lipschitz-contínua* no $\mathbf{R}^n$ com módulo $L > 0$. Se $x^k, d^k \in \mathbf{R}^n$ satisfazem

$$\langle f'(x^k), d^k \rangle < 0,$$

então a desigualdade

$$f(x^k + \alpha d^k) \leq f(x^k) + \sigma\alpha\langle f'(x^k), d^k\rangle$$

é válida para todo $\alpha \in (0, \bar{\alpha}_k)$, onde

$$\bar{\alpha}_k = \tfrac{2(\sigma-1)\langle f'(x^k), d^k\rangle}{L||d^k||^2} > 0.$$

Demonstração: *Lema (1.5.7): Seja $f : \mathbf{R}^n \to \mathbf{R}$ uma função diferenciável no $\mathbf{R}^n$, com derivada Lipschitz-contínua no $\mathbf{R}^n$ com módulo $L > 0$.*
Então

$$|f(x + y) - f(x) - \langle f'(x), y\rangle| \leq L||y||^2/2 \forall x, y \in \mathbf{R}^n.$$

Pelo Lema acima, para todo $\alpha \in \mathbf{R}$, tem-se que

$$\begin{aligned} f(x^k + \alpha d^k) - f(x^k) &\leq \langle f'(x^k), \alpha d^k\rangle + \frac{L}{2}\alpha^2||d^k||^2 \\ &= \alpha(\langle f'(x^k), d^k + \frac{L}{2}\alpha||d^k||^2). \end{aligned}$$

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

$$\begin{aligned} & f(x^k + \alpha d^k) - f(x^k) \leq \alpha(\langle f'(x^k), d^k\rangle + \frac{L}{2}\bar{\alpha_k}||d^k||^2) \\ \Rightarrow\ & f(x^k + \alpha d^k) - f(x^k) \leq \alpha(\langle f'(x^k), d^k\rangle + \frac{L}{2}||d^k||^2 \cdot \frac{2(\sigma-1)\langle f'(x^k), d^k\rangle}{L||d^k||^2}) \\ \Rightarrow\ & f(x^k + \alpha d^k) - f(x^k) \leq \alpha(\langle f'(x^k), d^k\rangle + (\sigma-1)\langle f'(x^k), d^k\rangle) \\ \Rightarrow\ & f(x^k + \alpha d^k) - f(x^k) \leq \alpha(\langle f'(x^k), d^k\rangle + \sigma\langle f'(x^k), d^k\rangle - \langle f'(x^k), d^k\rangle) \\ \Rightarrow\ & f(x^k + \alpha d^k) - f(x^k) \leq \alpha\sigma\langle f'(x^k), d^k\rangle \end{aligned}$$

onde a segunda igualdade segue de

$$\bar{\alpha}_k = \tfrac{2(\sigma-1)\langle f'(x^k), d^k\rangle}{L||d^k||^2} > 0.$$

Interpretação do lema: O comprimento de passo α_k pode ser limitado inferiormente por um $\bar{\alpha}_k > 0$, quando a função f tem derivada Lipschitz-contínua.

4.2 Goldstein

Dados um ponto x^k, uma direção de descida $d^k \in \mathbf{R}^n\{0\}$ 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:

$$f(x^k) + \sigma_1 \alpha \langle f'(x^k), d^k \rangle \geq f(x^k + \alpha d^k) \tag{3}$$

$$f(x^k) + \sigma_2 \alpha \langle f'(x^k), d^k \rangle \leq f(x^k + \alpha d^k) \tag{4}$$

Na Regra de Goldstein é adicionada à condição de Armijo, a desigualdade (4) com o intuito de descartar comprimentos de passo excessivamente pequenos.

4.3 Wolfe

Adicionalmente, a Regra de Wolfe complementa a desigualdade de Armijo com uma condição de curvatura, estipulando que, além da redução suficiente em f, para $0 < \sigma_1 < \sigma_2 < 1$, o comprimento de passo $\alpha^k > 0$ deve cumprir

$$\begin{gathered} f(x^k + \alpha^k d^k) \leq f(x^k) + \sigma_1 \alpha^k \langle f'(x^k), d^k \rangle \\ \langle f'(x^k + \alpha^k d^k), d^k \rangle \geq \sigma_2 \langle f'(x^k), d^k \rangle. \end{gathered} \tag{5}$$

5 O Método do Gradiente

O Método do Gradiente (MG) é um método de primeira ordem que utiliza a direção do anti-gradiente iterativamente, com o intuito de minimizar $f(x)$. A atualização do ponto x^k para um novo ponto x^{k+1} é descrita no seguinte esquema

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

onde α_k representa o tamanho do passo e $\nabla f(x^k)$ é o gradiente de f no ponto x^k.

5.1 Algoritmo

Algoritmo 1: Método do Gradiente (MG)

1 Tome um ponto inicial $x^0 \in \mathbb{R}^n$ e $\epsilon > 0$.
2 Defina $k = 0$.
3 **Repita** enquanto $||\nabla f(x^k)|| > \epsilon$.
4 Calcule $d^k = -\nabla f(x^k)$
5 Calcule o comprimento de passo α_k.
6 Defina $x^k = x^k + \alpha_k d^k$
7 Defina $k \leftarrow k + 1$.

5.2 Convergência global do método do gradiente I

Seja $f : \mathbf{R}^n \to \mathbf{R}$ uma função diferenciável e com derivada Lipschitz-contínua no $\mathbf{R}^n$ $L > 0$. Então se a sequência (x^k) gerada pelo Algortimo 1 equipado com a **regra de Armijo** possui um ponto de acumulação, tem-se que

$$(\nabla f(x^k)) \to 0 \quad (k \to \infty).$$

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 a. Isso é, toda bola aberta de centro a contém uma infinidade de pontos de (x^k). Por definição, $\forall \epsilon_1 > 0$, deve existir $x \in (x^k)$ tal que $0 < |x - a| < \epsilon_1$. Pela continuidade de f, a sequência $(f(x^k))$ também tem um ponto de acumulação.

Por definição, temos que $\forall \epsilon > 0, \exists \delta > 0; x \in (x^k), |x - x^0| < \delta \Rightarrow |f(x) - f(x^0)| < \epsilon$. Como a é ponto de acumulação da sequência (x^k) e f é contínua, podemos escolher um $\epsilon > 0$ arbitrário, de modo que exista um $\delta > 0$ tal que
$|x - a| < \delta \Rightarrow |f(x) - f(a)| < \epsilon$. Observe que, por definição, a sequência $(f(x^k))$ também tem um ponto de acumulação que corresponde ao valor de $f(a)$. 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 a. Isso é, toda bola aberta de centro a contém uma infinidade de pontos de (x^k). Por definição, $\forall \epsilon_1 > 0$, deve existir $x \in (x^k)$ tal que $0 < |x - a| < \epsilon_1$. Pela continuidade de f, a sequência $(f(x^k))$ também tem um ponto de acumulação.

Sabendo que $\alpha_k \geq \bar{\alpha} > 0$, pela desigualdade de Armijo, para todo k temos que

$$f(x^k) - f(x^{k+1}) \geq \sigma \alpha_k ||\nabla f(x^k)||^2 \geq \sigma \bar{\alpha} ||\nabla f(x^k)||^2.$$

Como $f(x^k)$ é monótona decrescente e tem um ponto de acumulação $f(a)$, então existe $f(x^{k_j}) \to f(a)$.

Suponha por absurdo que $(f(x^k))$ não seja limitada inferiormente, ou seja, existe $c < f(a)$ tal que $f(x^{k'}) < c < f(a)$. Como $(f(x^k))$ é decrescente, tem-se

$$... < f(x^{k'+2}) < f(x^{k'+1}) < f(x^{k'}) < c < f(a),$$

ou seja, $f(x^n) < c < f(a)$ para todo $n \geq k'$, o que contradiz $f(x^{k_j}) \to f(a)$. Isso prova que $f(x^k)$ é limitada inferiormente e, como é monótona decrescente, é também limitada superiormente, logo converge.

Como $f(x^k) - f(x^{k+1}) \to 0 \quad (k \to \infty)$, obtemos

$$(\nabla f(x^k)) \to 0 \quad (k \to \infty)$$

da desigualdade anterior. Logo, considerando a continuidade do gradiente, concluímos que cada ponto de acumulação de (x^k) é um ponto estacionário do problema de minimizar f.

5.3 Convergência global do método do gradiente II

Seja $f : \mathbf{R}^n \to \mathbf{R}$ uma função diferenciável no $\mathbf{R}^n$, com derivada contínua. 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 de minimizar

f.

Demonstração: Suponhamos que a sequência (x^k) tenha um ponto de acumulação $\bar{x} \in \mathbf{R}^n$, e que $\nabla f(x^k) \neq 0$ para todo k. Suponhamos também que $(x^{k_j}) \to \bar{x}$ $(j \to \infty)$. O caso que existe $\bar{\alpha} > 0$ tal que $\alpha_{k_j} \geq \bar{\alpha}$ para todo j, pode ser analisado de maneira análoga ao teorema anterior. Em particular, a sequência monótona $f(x^k)$ possui o ponto de acumulação $f(\bar{x})$. Portanto, a sequência $f(x^{k_j})$ converge. Além disso, pela desigualdade de Armijo temos

$$f(x^{k_{j+1}}) \leq f(x^{k_j}) - \sigma\alpha_{k_j}||\nabla f(x^{k_j})||^2 \leq f(x^{k_j}) - \sigma\bar{\alpha}||\nabla f(x^{k_j})||^2$$

Passando o limite quando $j \to \infty$, temos

$$f(\bar{x}) \leq f(\bar{x}) - \bar{\alpha}||\nabla f(\bar{x})||^2 \Rightarrow \bar{\alpha}||\nabla f(\bar{x})||^2 \leq 0 \Rightarrow ||\nabla f(\bar{x})||^2 \leq 0 \Rightarrow \nabla f(\bar{x}) = 0.$$

Considerando o caso de não existir $\bar{\alpha} > 0$ tal que α para todo j, tomando uma subsequência se for necessário, podemos admitir que $\alpha_{k_j} \to 0$ $(j \to \infty)$. Neste caso, para todo j suficientemente grande, o valor inicial do comprimento de passo $\bar{\alpha}$ foi reduzido pelo menos uma vez, ou seja, o valor $\alpha = \theta^{-1}\alpha_{k_j}$ não satisfaz a desigualdade de Armijo. Isto é

$$\begin{aligned}
f(x^{k_j} - \theta^{-1}\alpha_{k_j}\nabla f(x^{k_j})) &> f(x^{k_j}) - \sigma\theta^{-1}\alpha_{k_j}||\nabla f(x^{k_j})||^2 \\
f(x^{k_j} - \theta^{-1}\alpha_{k_j}\nabla f(x^{k_j})) - f(x^{k_j}) &> -\sigma\theta^{-1}\alpha_{k_j}||\nabla f(x^{k_j})||^2 \\
\frac{f(x^{k_j} - \theta^{-1}\alpha_{k_j}\nabla f(x^{k_j})) - f(x^{k_j})}{\theta^{-1}\alpha_{k_j}} &> -\sigma||\nabla f(x^{k_j})||^2
\end{aligned}$$

Passando o limite quando $j \to \infty$, obtemos

$$\begin{aligned}
&\lim_{j\to\infty} \frac{f(x^{k_j} - \theta^{-1}\alpha_{k_j}\nabla f(x^{k_j})) - f(x^{k_j})}{\theta^{-1}\alpha_{k_j}} > \lim_{j\to\infty} -\sigma||\nabla f(x^{k_j})||^2 \\
&\Rightarrow \frac{f(\bar{x}) - \theta^{-1}\alpha_{k_j}||\nabla f(\bar{x})||^2 - f(\bar{x})}{\theta^{-1}\alpha_{k_j}} \geq -\sigma||\nabla f(\bar{x})||^2 \\
&\Rightarrow -||\nabla f(\bar{x})||^2 \geq -\sigma||\nabla f(\bar{x})||^2
\end{aligned}$$

Como $\sigma \in (0, 1)$, a desigualdade acima só é possível quando $\nabla f(\bar{x}) = 0$. Se $(x^k)_{k\in\mathbb{N}}$ é limitada, existe uma subsequência $(x^{k_j})_{j\in\mathbb{N}}$ que é convergente e, consequentemente,(x^k) tem um ponto de acumulação.

Como cada ponto de acumulação de qualquer sequência (x^k) é um ponto estacionário do problema de minimizar $f(x), x \in \mathbf{R}^n$, concluímos que

$$(\nabla f(x^k)) \to 0 \quad (k \to \infty).$$

5.4 Convergência global do método do gradiente no caso convexo

Seja $f :^n \to$ uma função convexa, diferenciável no $\mathbf{R}^n$, com derivada contínua. Suponhamos que o Algoritmo 1 utiliza a regra de Armijo com $\bar{\alpha} \leq 1$. 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 $x \in^n$.
Demonstração: Seja $\overline{x} \in^n$ uma solução do problema. Como para todo k vale

$$f(\overline{x}) \leq f(x_k) \text{ e}$$

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

temos que

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

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

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

Pela convexidade de f 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 estratégia de atualização de pontos usada no método do gradiente ($x_{k+1} = x_k - \alpha_k \nabla f(x_k)$), obtemos

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

Veja que $\langle \nabla f(x_k), x_k - \overline{x} \rangle \geq 0$. Logo o termo $-2\alpha_k \langle \nabla f(x_k), x_k - \overline{x} \rangle$ sempre será negativo. Além disso, $\alpha_k^2 \leq \alpha_k$, 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

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

Por

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

obtemos

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

Concluímos que a sequência (x_k) é limitada. Portanto, (x_k) tem um ponto de acumualação a. Pelas condições de otimalidade de primeira ordem no caso de conjunto viável convexo, temos que todo ponto de acumualação é um ponto estacionário, ou seja, $\nabla f(a) = 0$ o que, no caso convexo, significa que a é uma solução do problema. Podemos então tomar $\overline{x} = a$ na análise acima, para concluir de

$$\|x_j - \overline{x}\|^2 \leq \|x_k - \overline{x}\|^2 + \sum_{i=k}^{j-1} \alpha_i \|\nabla f(x_i)\|^2,$$

que para todo $j \geq k+1$, temos

$$\|x_j - a\|^2 \leq \|x_k - a\|^2 + \sum_{i=k}^{\infty} \alpha_i \|\nabla f(x_i)\|^2. \tag{6}$$

Além disso, temos que

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

$$\implies \lim_{k\to\infty}\left(\sum_{i=k}^{\infty}\alpha_i\|\nabla f(x_i)\|^2\right) = \lim_{k\to\infty}\left(\sum_{i=0}^{\infty}\alpha_i\|\nabla f(x_i)\|^2\right) - \lim_{k\to\infty}\left(\sum_{i=0}^{k-1}\alpha_i\|\nabla f(x_i)\|^2\right)$$

$$\implies \lim_{k\to\infty}\left(\sum_{i=k}^{\infty}\alpha_i\|\nabla f(x_i)\|^2\right) = \sum_{i=0}^{\infty}\alpha_i\|\nabla f(x_i)\|^2 - \sum_{i=0}^{\infty}\alpha_i\|\nabla f(x_i)\|^2$$

$$\implies \lim_{k\to\infty}\left(\sum_{i=k}^{\infty}\alpha_i\|\nabla f(x_i)\|^2\right) = 0.$$

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 a é um ponto de acumulação da sequência (x_k), podemos também escolher um k tal que

$$\frac{\delta}{2} > \|x_k - a\|^2.$$

De

$$\|x_j - a\|^2 \leq \|x_k - a\|^2 + \sum_{i=k}^{\infty}\alpha_i\|\nabla f(x_i)\|^2,$$

concluímos que para todo $\delta > 0$, existe k tal que

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

Isso significa que (x_k) converge a a.