



Instituto Tecnológico de Aeronáutica

---

# **Introdução às Redes Neurais e aos Grandes Modelos de Linguagem**

**Prof. Mauri Aparecido de Oliveira**

---

**CURSO – Extensivo**

PO-249



# Introdução às Redes Neurais e aos Grandes Modelos de Linguagem

ai



## Revisão Matemática

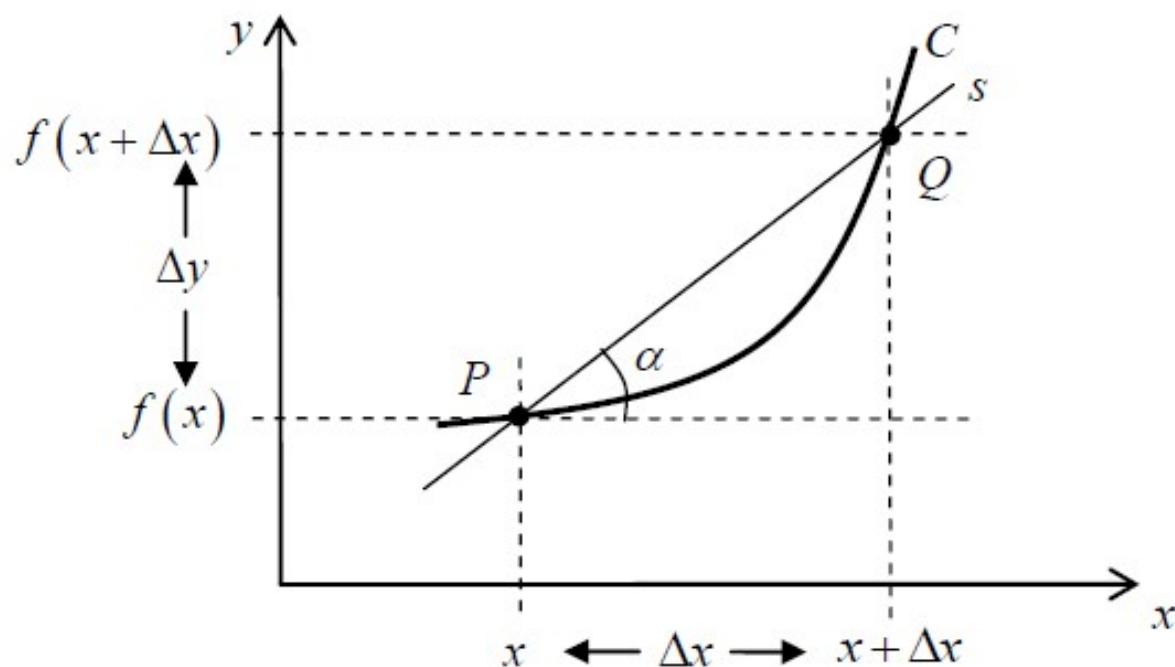
A construção do algoritmo de backpropagation e a modelagem de redes neurais artificiais (RNA) utilizam como uma de suas principais ferramentas, o cálculo diferencial.

O objetivo deste Aula é fornecer material de apoio para compreender o cálculo de derivadas, do gradiente e noções básicas de lógica proposicional.

## Derivada

Para apresentar o conceito de derivada, recorremos ao exemplo clássico de determinação da reta tangente a uma curva. Com o desenvolvimento da matemática do século XVII este problema começa a ser tratado usando o arcabouço do cálculo diferencial, sendo que Gilles Personne de Roberval, Pierre Fermat, Isaac Newton e Gottfried Wilhelm von Leibniz desempenharam papel fundamental para a solução completa do problema da forma como conhecemos hoje.

Assim, iniciamos com o problema de determinar a reta tangente a uma curva passando por um ponto  $P$ , tal como mostrado na Figura-1.



**Figura-1** Representação do ponto  $P$  na curva  $C$  por onde deve passar uma reta tangente  $t$ .

O coeficiente angular da reta secante  $s$  é dado por

$$\frac{f(x + \Delta x) - f(x)}{(x + \Delta x) - x},$$

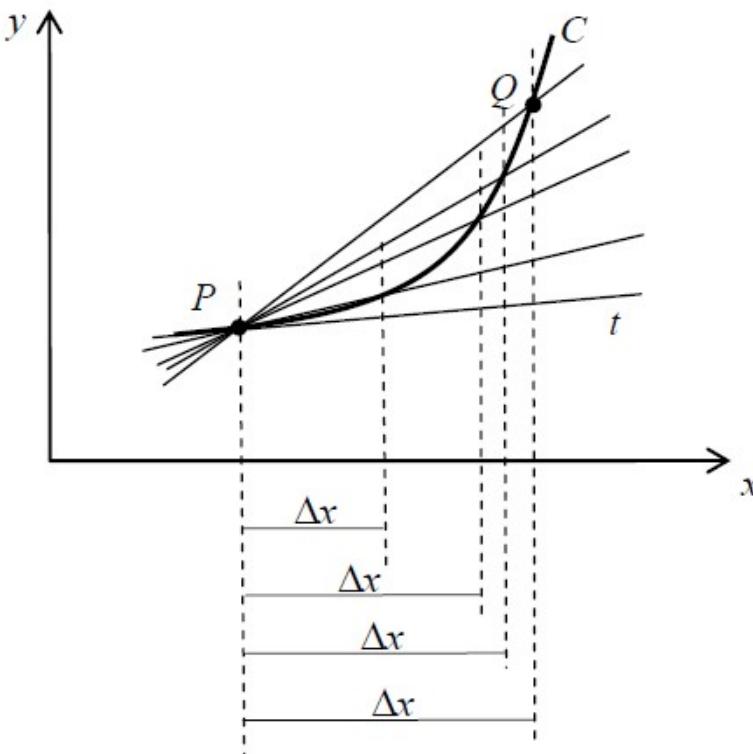
$$\frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Fazer  $\Delta x$  tender a zero implica em aproximar  $x + \Delta x$  de  $x$ . Dessa forma, mantendo o ponto  $P$  fixo, a reta secante tende a uma reta tangente  $t$  passando por  $P$ , tal como mostrado na Figura-2. No limite, fazendo  $\Delta x$  tender a zero, produz

$$\lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Esse limite recebe o nome de derivada de  $f$  com relação a  $x$  e uma das notações que recebe é dada por  $f'(x)$ . Dessa forma,

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$



**Figura–2** Representação da dinâmica de aproximação do ponto  $P$  ao ponto  $Q$ , a partir da condição de fazer  $\Delta x$  tender a zero.

**Definição.** Sejam  $f$  uma função contínua e  $x$  um ponto pertencente ao seu domínio. Se o limite

$$\lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

existe e é finito, então é denominado de derivada de  $f$  em relação a  $x$  e denotado por  $f'(x)$ . Assim, a derivada de  $f$  em relação a  $x$  é a função  $f'(x)$  tal que

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

## Derivada Parcial

A partir do século XVIII a cálculo de derivadas parciais passou a ser mais empregado na resolução de diversos problemas. Esse processo de diferenciação parcial era empregado por pesquisadores como Leonhard Euler e Gottfried Wilhelm Leibniz. Alguns dos primeiros trabalhos que influenciaram a utilização das derivadas parciais são de autoria de Alexis Fontaine des Bertins, sendo um deles contido no *Mémoires donnés à l'Académie Royale des Sciences, non Imprimés dans leur Temps*, publicado em 1764. À medida que os problemas envolviam cada vez mais variáveis nas equações, tornou-se aparente a necessidade de um método para obter derivadas relacionando essas diversas variáveis. Sendo importante entender a taxa de variação relacionando duas variáveis, mantendo as outras constantes.

Algumas notações para derivadas parciais são mostradas a seguir, para isso considere uma função  $z = f(x, y)$ , então podemos escrever que

$$\frac{\partial f(x, y)}{\partial x} = \frac{\partial f}{\partial x} = f_x(x, y) = \frac{\partial z}{\partial x} = f_1 = D_1 f = D_x f,$$

$$\frac{\partial f(x, y)}{\partial y} = \frac{\partial f}{\partial y} = f_y(x, y) = \frac{\partial z}{\partial y} = f_2 = D_2 f = D_y f.$$

**EXEMPLO**

Considere a função  $f(x, y) = x^3 + x^2y - xy^2$ . Determine a derivada parcial pela definição da função  $f$  com relação à variável  $x$ .

Dada a função temos que:

$$\begin{aligned} f(x+h, y) &= (x+h)^3 + (x+h)^2 y - (x+h)y^2 \\ &= x^3 + 3x^2h + 3xh^2 + h^3 + (x^2 + 2xh + h^2)y - xy^2 - hy^2 \\ &= x^3 + 3x^2h + 3xh^2 + h^3 + x^2y + 2xhy + h^2y - xy^2 - hy^2, \end{aligned}$$

e

$$\begin{aligned} f(x+h, y) - f(x, y) &= x^3 + 3x^2h + 3xh^2 + h^3 + x^2y + 2xhy + h^2y - xy^2 - hy^2 - (x^3 + x^2y - xy^2) \\ &= x^3 + 3x^2h + 3xh^2 + h^3 + x^2y + 2xhy + h^2y - xy^2 - hy^2 - x^3 - x^2y + xy^2 \\ &= 3x^2h + 3xh^2 + h^3 + 2xhy + h^2y - hy^2. \end{aligned}$$

Então,

$$\begin{aligned} \frac{\partial f}{\partial x} &= \lim_{h \rightarrow 0} \frac{f(x+h, y) - f(x, y)}{h} \\ \frac{\partial f}{\partial x} &= \lim_{h \rightarrow 0} \frac{3x^2h + 3xh^2 + h^3 + 2xhy + h^2y - hy^2}{h} \\ \frac{\partial f}{\partial x} &= \lim_{h \rightarrow 0} (3x^2 + 3xh + h^2 + 2xy + hy - y^2) \\ \frac{\partial f}{\partial x} &= 3x^2 + 2xy - y^2. \end{aligned}$$



## Jacobiana

Seja  $\mathbf{x}$  um vetor n-dimensional dado por  $\mathbf{x} = x_1, x_2, \dots, x_n$  e  $f$  uma função de  $\mathbf{x}$ , podemos escrever que

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \text{ ou } \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

e

$$f(\mathbf{x}) = f(x_1, x_2, \dots, x_n) \text{ ou } f(\mathbf{x}) = f\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Jacobiana é dada por

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

**EXEMPLO**

Considere a função  $g(x_1, x_2, x_3) = 2x_1^3 - 5x_1x_2^2 - x_2x_3^4 + 7$ , sua matriz Jacobiana é dada por

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} 6x_1^2 - 5x_2^2 & -10x_1x_2 - x_3^4 & -4x_2x_3^3 \end{pmatrix}.$$

**EXEMPLO**

Considere a função  $h(x) = (3x^2, -x^3, x)$ , sua matriz Jacobiana é dada por

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} 6x \\ -3x^2 \\ 1 \end{pmatrix}.$$

A matriz Jacobiana, onde tanto o domínio quanto o contradomínio são  $\mathbb{R}$ , fornece a derivada da função de uma variável. Por exemplo, considere a função  $g(x) = x^3$  temos que

$$\mathbf{J}(x) = (3x^2),$$

que é uma matriz  $1 \times 1$ , ou ainda, nesse caso  $\mathbf{J}(x) = g'(x) = 3x^2$ .

## Hessiana

Seja uma função  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , a matriz Hessiana (introduzida por Ludwig Otto Hesse em 1884) de  $f$  é a matriz  $n \times n$ , denotada por  $\mathbf{H}(\mathbf{x})$  e definida como

$$\mathbf{H}_{ij}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

No caso da matriz Hessiana, temos que as derivadas parciais de segunda ordem,  $\frac{\partial^2 f}{\partial x_i^2}$ , medem a curvatura da  $f(\mathbf{x})$  na direção de  $x_i$ . A derivada parcial de segunda ordem  $\frac{\partial^2 f}{\partial x_i \partial x_j}$  mede a taxa na qual a inclinação na direção  $j$  muda à medida que nos movemos na direção  $i$ . Se a função  $f(\mathbf{x})$  é duas vezes diferenciável, então  $\mathbf{H}(\mathbf{x})$  é uma matriz simétrica tal que  $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$ .

Quando os modelos construídos envolvem várias variáveis, a importância do estudo das matrizes Jacobiana e Hessiana é evidente. Uma vez que o cálculo multivariado concentra-se principalmente no fato de que quando uma função  $f$  depende de várias variáveis, temos inclinações em infinitas direções e precisamos avaliar os comportamentos das inclinações para determinar as curvaturas e formas que a função assume próxima de certos pontos críticos. A matriz Hessiana também pode ser escrita como

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

A matriz Hessiana é uma matriz  $n \times n$  de segundas derivadas.



## EXEMPLO

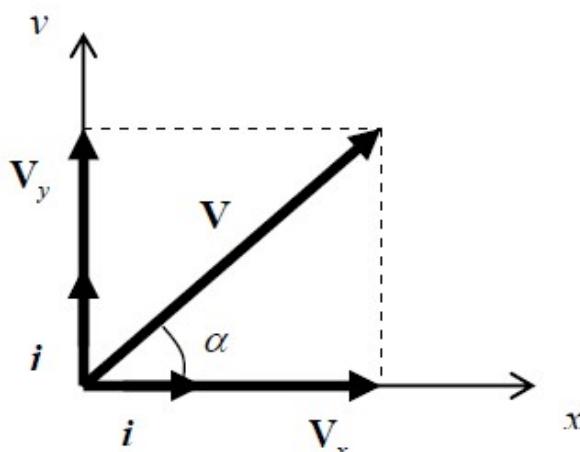
Para a função  $g: \mathbb{R}^3 \rightarrow \mathbb{R}$ , do Exemplo da seção anterior, dada por  $g(x_1, x_2, x_3) = 2x_1^3 - 5x_1x_2^2 - x_2x_3^4 + 7$ , temos que sua matriz Hessiana é igual a

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_3 \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \frac{\partial^2 f}{\partial x_3 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_3} & \frac{\partial^2 f}{\partial x_2 \partial x_3} & \frac{\partial^2 f}{\partial x_3^2} \end{pmatrix} = \begin{pmatrix} 12x_1 & -10x_2 & 0 \\ -10x_2 & -10x_1 & -4x_3^3 \\ 0 & -4x_3^3 & -12x_2x_3^2 \end{pmatrix}.$$

## Derivada Direcional

Podemos representar um vetor  $\mathbf{V}$  num sistema de coordenadas que considere vetores perpendiculares, tal como mostrado na Figura , onde verificamos que  $\mathbf{V} = \mathbf{V}_x + \mathbf{V}_y$  , sendo que  $V_x = V \cos \alpha$  e  $V_y = V \sin \alpha$  . Como os vetores unitários  $\mathbf{i}$  e  $\mathbf{j}$  estão definidos nas direções e sentidos dos eixos  $x$  e  $y$ , respectivamente, então  $\mathbf{V}_x = (V \cos \alpha) \mathbf{i}$  e  $\mathbf{V}_y = (V \sin \alpha) \mathbf{j}$  , dessa forma, o vetor  $\mathbf{V}$  pode ser representado por

$$\mathbf{V} = (V \cos \alpha) \mathbf{i} + (V \sin \alpha) \mathbf{j}.$$



**Figura** Componentes de um vetor  $\mathbf{V}$ .

$$D_{\mathbf{V}} f(x, y) = \lim_{h \rightarrow 0} \frac{f(x + ha, y + hb) - f(x, y)}{h}.$$

$$D_{\mathbf{V}} f(x, y) = f_x(x, y) \cos \alpha + f_y(x, y) \sin \alpha.$$

**EXEMPLO:**

Determine  $D_{\mathbf{V}}f(x, y)$  dado que  $f(x, y) = 2x^2 - 3y^2 + 3x + 2y$  e  $\mathbf{V}$  é o vetor unitário na direção  $\frac{\pi}{3}$ .

Dado que  $\mathbf{V} = (\cos \alpha)\mathbf{i} + (\sin \alpha)\mathbf{j}$  e  $\alpha = \frac{\pi}{3}$ , então  $\mathbf{V} = \left(\cos \frac{\pi}{3}\right)\mathbf{i} + \left(\sin \frac{\pi}{3}\right)\mathbf{j}$  ou  $\mathbf{V} = \frac{1}{2}\mathbf{i} + \frac{\sqrt{3}}{2}\mathbf{j}$ . A partir da definição da derivada direcional

$$D_{\mathbf{V}}f(x, y) = \lim_{h \rightarrow 0} \frac{f(x + h \cos \alpha, y + h \sin \alpha) - f(x, y)}{h},$$

temos

$$D_{\mathbf{V}}f(x, y) = \lim_{h \rightarrow 0} \frac{f\left(x + \frac{1}{2}h, y + \frac{\sqrt{3}}{2}h\right) - f(x, y)}{h},$$

$$-f(x, y) = -2x^2 + 3y^2 - 3x - 2y,$$

$$f\left(x + \frac{1}{2}h, y + \frac{\sqrt{3}}{2}h\right) = 2\left(x + \frac{1}{2}h\right)^2 - 3\left(y + \frac{\sqrt{3}}{2}h\right)^2 + 3\left(x + \frac{1}{2}h\right) + 2\left(y + \frac{\sqrt{3}}{2}h\right),$$

$$f\left(x + \frac{1}{2}h, y + \frac{\sqrt{3}}{2}h\right) = 2\left(x^2 + xh + \frac{1}{4}h^2\right) - 3\left(y^2 + \sqrt{3}yh + \frac{3}{4}h^2\right) + 3x + \frac{3}{2}h + 2y + \sqrt{3}h,$$

$$f\left(x + \frac{1}{2}h, y + \frac{\sqrt{3}}{2}h\right) = 2x^2 + 2xh + \frac{1}{2}h^2 - 3y^2 - 3\sqrt{3}yh - \frac{9}{4}h^2 + 3x + \frac{3}{2}h + 2y + \sqrt{3}h,$$

então,

$$f\left(x + \frac{1}{2}h, y + \frac{\sqrt{3}}{2}h\right) - f(x, y) = 2xh + \frac{1}{2}h^2 - 3\sqrt{3}yh - \frac{9}{4}h^2 + \frac{3}{2}h + \sqrt{3}h,$$

$$D_V f(x, y) = \lim_{h \rightarrow 0} \frac{2xh + \frac{1}{2}h^2 - 3\sqrt{3}yh - \frac{9}{4}h^2 + \frac{3}{2}h + \sqrt{3}h}{h},$$

$$D_V f(x, y) = \lim_{h \rightarrow 0} \left( 2x + \frac{1}{2}h - 3\sqrt{3}y - \frac{9}{4}h + \frac{3}{2} + \sqrt{3} \right),$$

$$D_V f(x, y) = 2x - 3\sqrt{3}y + \frac{3}{2} + \sqrt{3}.$$

## Gradiente

Podemos reescrever o resultado  $D_{\mathbf{V}}f$  em termos do produto escalar do vetor  $\mathbf{V} = (\cos \alpha) \mathbf{i} + (\sin \alpha) \mathbf{j}$  e do vetor dado por  $f_x(x, y) \mathbf{i} + f_y(x, y) \mathbf{j}$ , ou seja,

$$\begin{aligned} f_x(x, y) \cos \alpha + f_y(x, y) \sin \alpha &= [(\cos \alpha) \mathbf{i} + (\sin \alpha) \mathbf{j}] [f_x(x, y) \mathbf{i} + f_y(x, y) \mathbf{j}], \\ D_{\mathbf{V}}f(x, y) &= [(\cos \alpha) \mathbf{i} + (\sin \alpha) \mathbf{j}] [f_x(x, y) \mathbf{i} + f_y(x, y) \mathbf{j}]. \end{aligned}$$

Com relação à essa expressão podemos escrevê-la de forma mais compacta utilizando uma nova notação para o vetor  $f_x(x, y) \mathbf{i} + f_y(x, y) \mathbf{j}$ . Em particular esse vetor tem interpretação geométrica bastante importante e é denominado de vetor gradiente ou gradiente da função  $f$ . Assim temos a definição do vetor gradiente apresentada no quadro em destaque a seguir.

A partir da definição do gradiente podemos reescrever a derivada direcional  $D_{\mathbf{V}}f(x, y)$  como

$$D_{\mathbf{V}}f(x, y) = \nabla f(x, y) \cdot \mathbf{V}.$$

$$D_{\mathbf{V}}f(x, y) = f_x(x, y) \cos \alpha + f_y(x, y) \sin \alpha.$$

## Definição – Vetor Gradiente

O vetor gradiente de uma função diferenciável  $f(x, y)$  é denotado por  $\nabla f(x, y)$  e definido como

$$\nabla f(x, y) = f_x(x, y)\mathbf{i} + f_y(x, y)\mathbf{j},$$

onde  $\nabla$  é a letra grega delta maiúscula invertida e lida como “del”. Muitas vezes a abreviação  $\text{grad } f$  também é utilizada.

## Teorema

Suponha que  $f$  é uma função diferenciável de duas ou três variáveis. O valor máximo da derivada direcional  $D_{\mathbf{V}}f(x, y)$  é  $\|\nabla f(x, y)\|$  e ocorre quando  $\mathbf{V}$  tem a mesma direção do vetor gradiente  $\nabla f(x, y)$ .

## Prova

Dos resultados anteriores temos que

$$D_{\mathbf{V}}f(x, y) = \nabla f(x, y) \cdot \mathbf{V} = \|\nabla f(x, y)\| \|\mathbf{V}\| \cos \alpha = \|\nabla f(x, y)\| \cos \alpha,$$

onde  $\alpha$  é o ângulo entre  $\nabla f(x, y)$  e  $\mathbf{V}$ . O valor máximo de  $\alpha$  é 1 e ocorre quando  $\alpha = 0$ . Portanto, o valor máximo de  $D_{\mathbf{V}}f(x, y)$  é dado por  $\|\nabla f(x, y)\|$  e ocorre quando o ângulo  $\alpha$  é igual a 0, ou seja, quando o vetor  $\mathbf{V}$  tem a mesma direção de  $\nabla f(x, y)$ .

$$\nabla f(x, y, z) = \text{grad}f = f_x(x, y, z)\mathbf{i} + f_y(x, y, z)\mathbf{j} + f_z(x, y, z)\mathbf{k}.$$

Ou ainda, de forma mais abreviada

$$\nabla f = \langle f_x, f_y, f_z \rangle = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}.$$

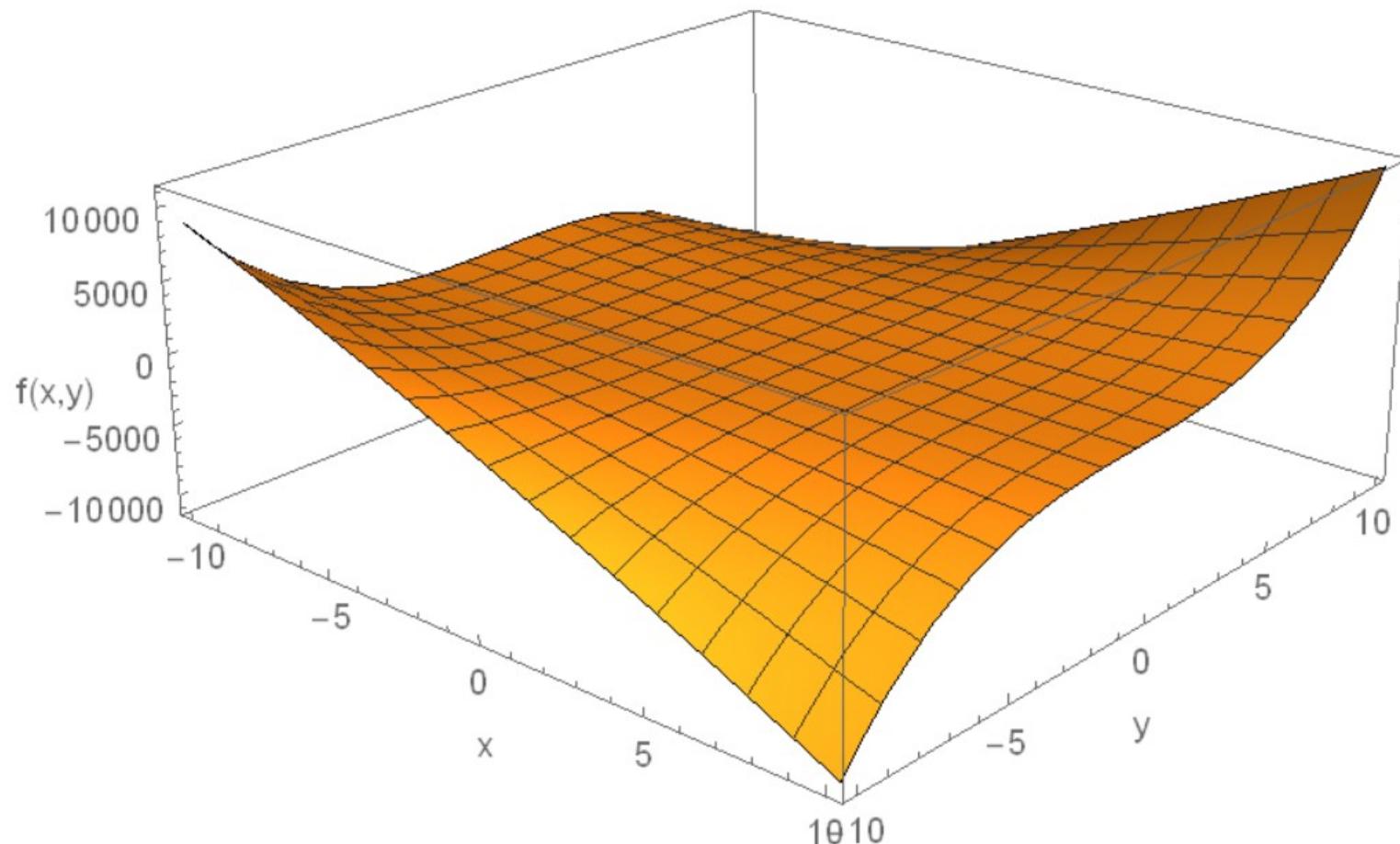
Sendo  $f(x, y)$  uma função a valores reais continuamente diferenciável, com  $\nabla f \neq \mathbf{0}$ . Então temos que:

- i. O gradiente  $\nabla f$  é normal a qualquer curva  $f(x, y) = c$ .
- ii. O valor de  $f(x, y)$  aumenta mais rápido na direção de  $\nabla f$ .
- iii. O valor de  $f(x, y)$  diminui mais rápido na direção de  $-\nabla f$ .

## EXEMPLO

Em qual direção a função  $f(x, y) = x^2y + xy^3$  aumenta mais rápido a partir do ponto  $(2,1)$ ?

Em qual direção ela decresce mais rápido?





Nesse caso temos que  $\nabla f(x,y) = (2xy + y^3, x^2 + 3xy^2)$ , o que resulta no ponto dado em  $\nabla f(2,1) = (5,10)$ , ou seja, o vetor gradiente é diferente do vetor nulo  $\mathbf{0}$ , o que significa que o ponto de máximo da função não foi alcançado. O valor de  $f$  aumenta mais rápido na direção de  $\nabla f$ . Um vetor unitário nessa direção é obtido fazendo  $\mathbf{n} = \frac{\nabla f}{\|\nabla f\|}$ , que resulta em

$$\mathbf{n} = \left( \frac{5}{\sqrt{5^2 + 10^2}}, \frac{10}{\sqrt{5^2 + 10^2}} \right) = \left( \frac{5}{\sqrt{125}}, \frac{10}{\sqrt{125}} \right),$$

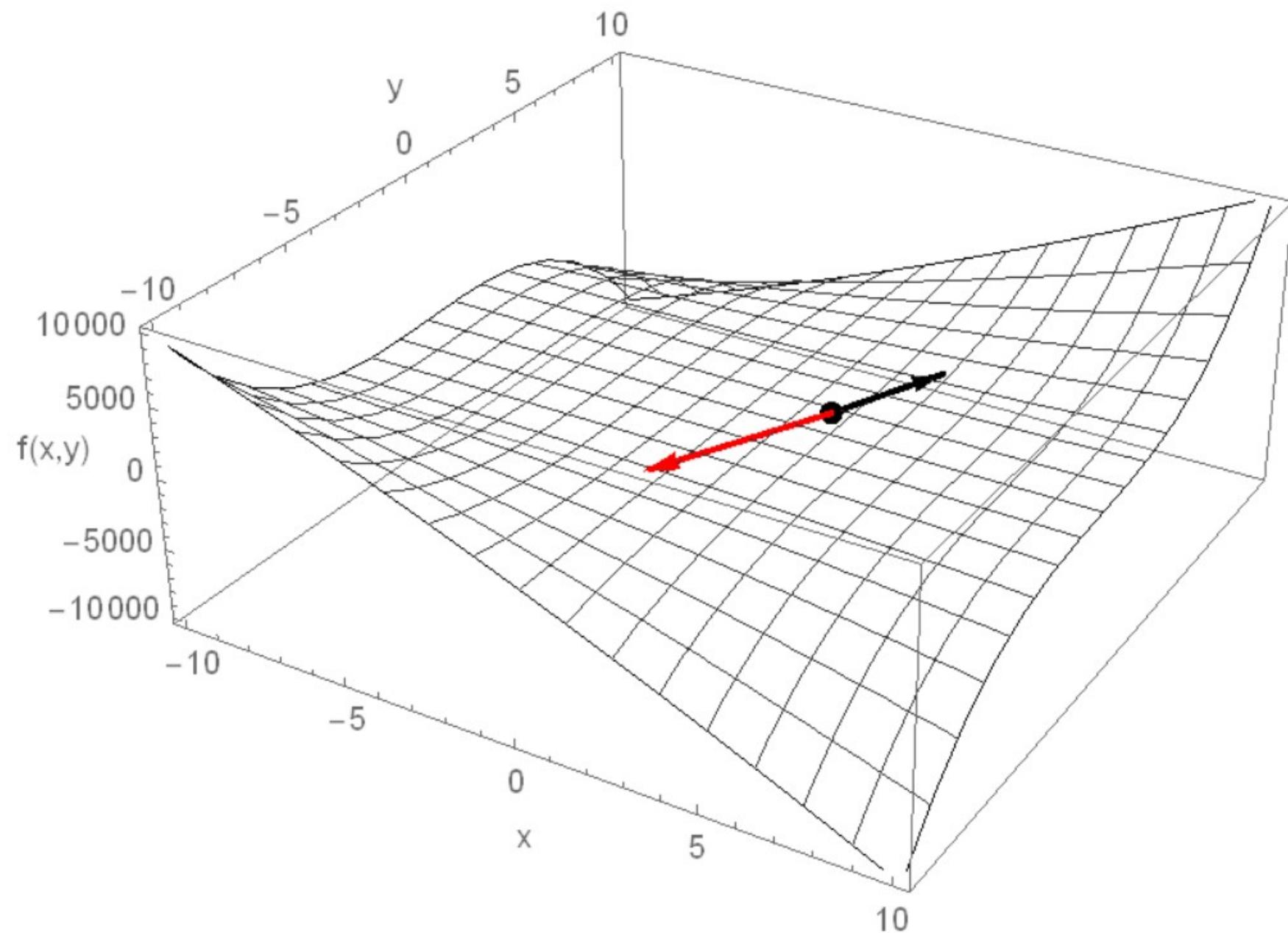
$$\mathbf{n} = \left( \frac{5}{\sqrt{5 \times 25}}, \frac{10}{\sqrt{5 \times 25}} \right) = \left( \frac{5}{5\sqrt{5}}, \frac{10}{5\sqrt{5}} \right),$$

$$\mathbf{n} = \left( \frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}} \right).$$

Sendo que

$$\|\mathbf{n}\| = \sqrt{\left(\frac{1}{\sqrt{5}}\right)^2 + \left(\frac{2}{\sqrt{5}}\right)^2} = \sqrt{\frac{1}{5} + \frac{4}{5}} = 1.$$

O valor de  $f$  diminui mais rápido na direção de  $-\nabla f$ , e isso resulta em  $\mathbf{n} = \left(-\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}\right)$ .



## Método do Gradiente Descendente

O Método do Gradiente Descendente ou Método da Descida Mais Íngreme é uma técnica bastante utilizada para determinar mínimos locais de uma função. Entre os pesquisadores que trabalharam e contribuíram para o desenvolvimento desta técnica pode-se destacar no fim do século XIX o matemático russo Pavel Alekseevich Nekrasov e no começo do século XX o físico-químico Peter Joseph William Debye que foi laureado com o prêmio Nobel de Química em 1936.

O primeiro passo na execução do algoritmo constitui-se em impor uma estimativa inicial da solução e determinar o gradiente da função nesse ponto. O próximo passo consiste em buscar a solução considerando a direção negativa do gradiente e repetimos o processo. O algoritmo irá convergir, eventualmente, para a situação em que o gradiente é zero, o que corresponderá a um mínimo local. No caso contrário, tem-se o Método do Gradiente Ascendente onde obtém o máximo local, determinado a partir da direção positiva do gradiente, tal como descrito anteriormente. Esses dois métodos são considerados de primeira ordem, isso porque utilizam apenas a primeira derivada da função.



## Algoritmo da Descida Mais Íngreme

Definir a função  $f(x)$

Calcular o gradiente  $\nabla f(x)$

Escolher um ponto inicial  $\mathbf{x}_0$

Determinar um valor para a precisão da aproximação  $\varepsilon$

Atribuir  $k=0$

### **Repetir**

Determinar o tamanho do passo  $\lambda_k$  que minimiza a função na direção do menor gradiente:  $\min_{\lambda_k} f(\mathbf{x}_k - \lambda_k \mathbf{n}_k)$

Calcular o novo ponto:  $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \lambda_k \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}$

Atualizar  $k \leftarrow k + 1$

### **Até**

$$|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| < \varepsilon$$

Outras opções para critério de parada são  $\|\nabla f(\mathbf{x})\| < \varepsilon$  ou  $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$ .

A construção do algoritmo baseado no Método do Gradiente Descendente começa considerando o problema de determinar a solução para o mínimo de uma função  $f(\mathbf{x})$ . Dada uma solução inicial  $\mathbf{x}_0$  para ser utilizada como aproximação da solução final  $\mathbf{x}$ , é possível alterar esse valor em várias direções, tal como apresentado na seção anterior. Para descrever a maneira como definimos a direção que minimiza a função  $f$ , utilizamos o conceito de gradiente  $\nabla f$ . O gradiente fornece a inclinação da reta tangente à curva em  $\mathbf{x}$  e sua direção aponta para o máximo crescimento da função. Um arranjo  $\mathbf{x}$  de  $n$  números reais dados por  $x_1, x_2, \dots, x_n$  é chamado de vetor e é escrito como

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

ou

$$\mathbf{x}' = [x_1, x_2, \dots, x_n],$$

onde o apóstrofo denota a operação de transposição de uma coluna para uma linha.



A busca começa em um ponto arbitrário  $\mathbf{x}_0$  e, em seguida, o movimento ocorre na direção negativa do gradiente, até estarmos perto o suficiente da solução. Ou seja, o pesquisador precisa estipular um erro que considere aceitável no processo de aproximação. Podemos escolher arbitrariamente o tamanho do avanço (ou passo) na direção do mínimo. O objetivo do algoritmo é detectar um ponto mínimo local  $\mathbf{x}^*$  dado um ponto inicial  $\mathbf{x}_0$ . Nesse caso, vamos denotar o tamanho do passo como  $\lambda$ . Em outras palavras, o processo iterativo pode ser descrito como:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \lambda \nabla f(\mathbf{x}_k).$$

Onde  $\lambda > 0$  é um número pequeno que força o algoritmo a realizar pequenos saltos. Precisamos escolher  $\lambda$  de tal forma que ele seja pequeno o suficiente para termos  $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$  e não tão pequeno de tal forma que a aproximação ao mínimo seja muito lenta. Uma maneira usual de implementação do Método do Gradiente Descendente consiste em utilizar o *gradiente normalizado* ao invés de  $\nabla f(\mathbf{x}_k)$ , o que permite reescrever (28) como

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \lambda \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}.$$



Vamos denotar  $\frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}$  por  $\mathbf{n}_k$ . A determinação do tamanho do passo  $\lambda$  é realizada muitas vezes utilizando um algoritmo para minimizar a função de uma dimensão  $f\left(\mathbf{x}_k - \lambda \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}\right)$ . Isso é chamado de minimização em linha ou busca em linha, o que significa

que a função  $f$  é minimizada ao longo da linha  $\mathbf{x}_k - \lambda \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}$ .

Nesse caso, vamos denotar essa função por  $\varphi(\lambda)$ , assim temos que:

$$\varphi(\lambda) = f(\mathbf{x}_k - \lambda \mathbf{n}_k).$$

A função  $\varphi(\lambda)$  pode ser derivada e igualada a zero para se procurar o valor de  $\lambda$  que minimiza e acelera a aproximação para o mínimo de  $f$ .

A partir da determinação do tamanho do passo  $\lambda_k$  temos a fórmula principal que descreve o esquema iterativo do Método do Gradiente Descendente, dada por:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \lambda_k \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}.$$

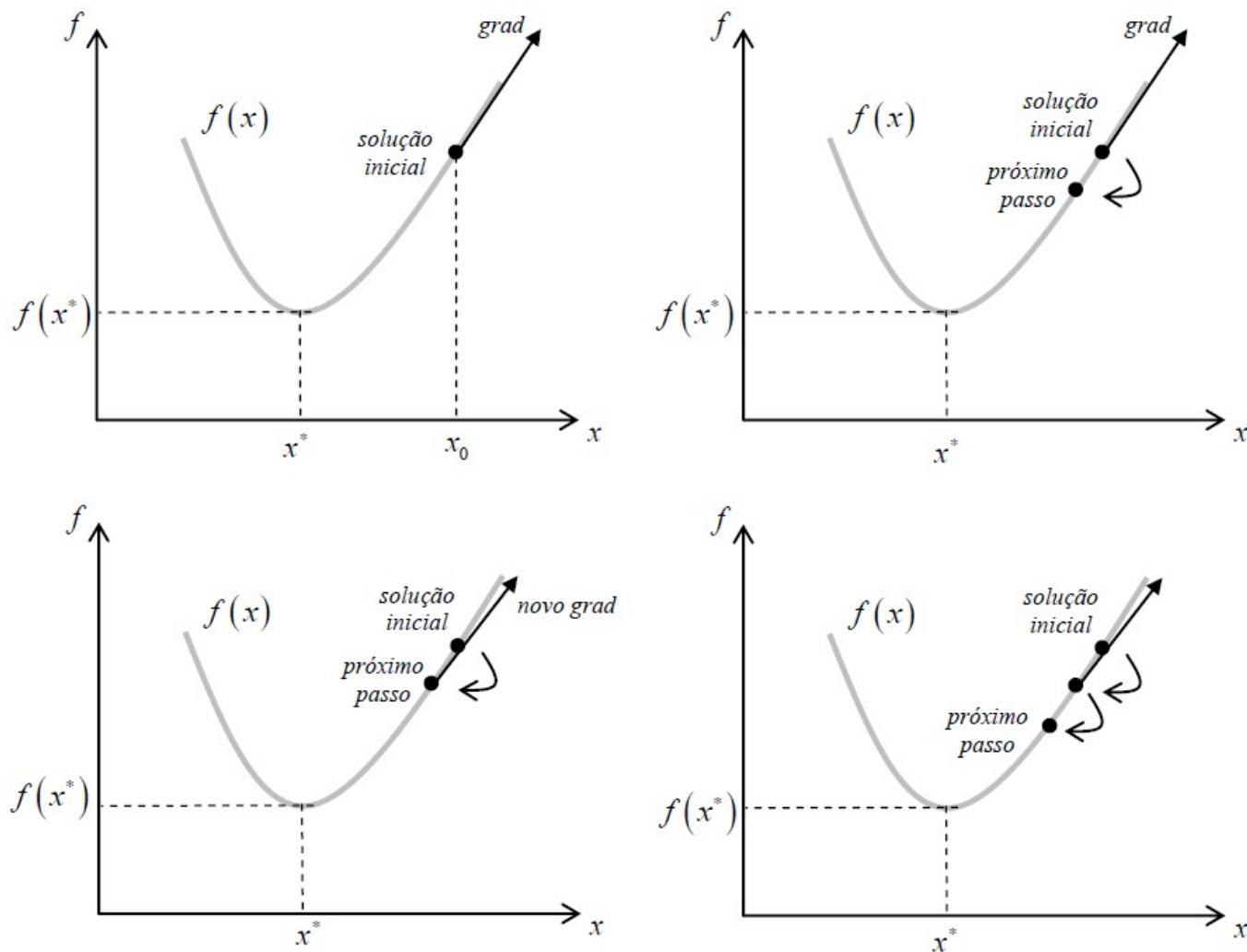
Dessa forma, a distância entre  $\mathbf{x}_{k+1}$  e  $\mathbf{x}_k$  é exatamente o tamanho do passo  $\lambda_k$ . Sendo que as iterações são realizadas até que um valor tolerado de acurácia tenha sido alcançado, nesse caso vamos denotar esse parâmetro de convergência por  $\varepsilon$ .

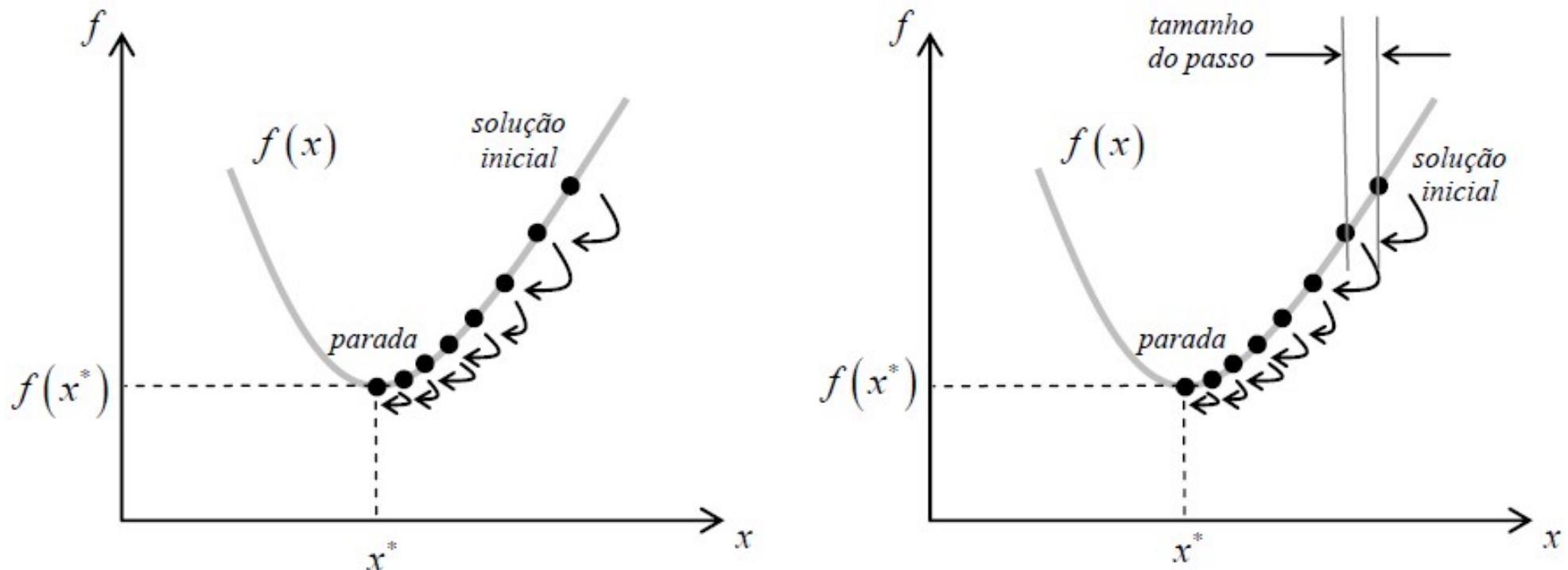
A máxima taxa de mudança da função  $f(\mathbf{x})$  em qualquer ponto  $\mathbf{x}_k$  é a magnitude do vetor gradiente, que é dada por  $\|\nabla f(\mathbf{x}_k)\|$ .



A Figura ilustra como o mínimo  $x^*$  de uma função  $f(x)$  é encontrado iniciando com  $x_0$  e a informação sobre a inclinação da função nesse ponto.

Quando não é possível determinar analiticamente o mínimo de uma função, métodos iterativos são empregados para obter uma solução aproximada, entre esses métodos o de Descida Mais Íngreme é bastante utilizado.





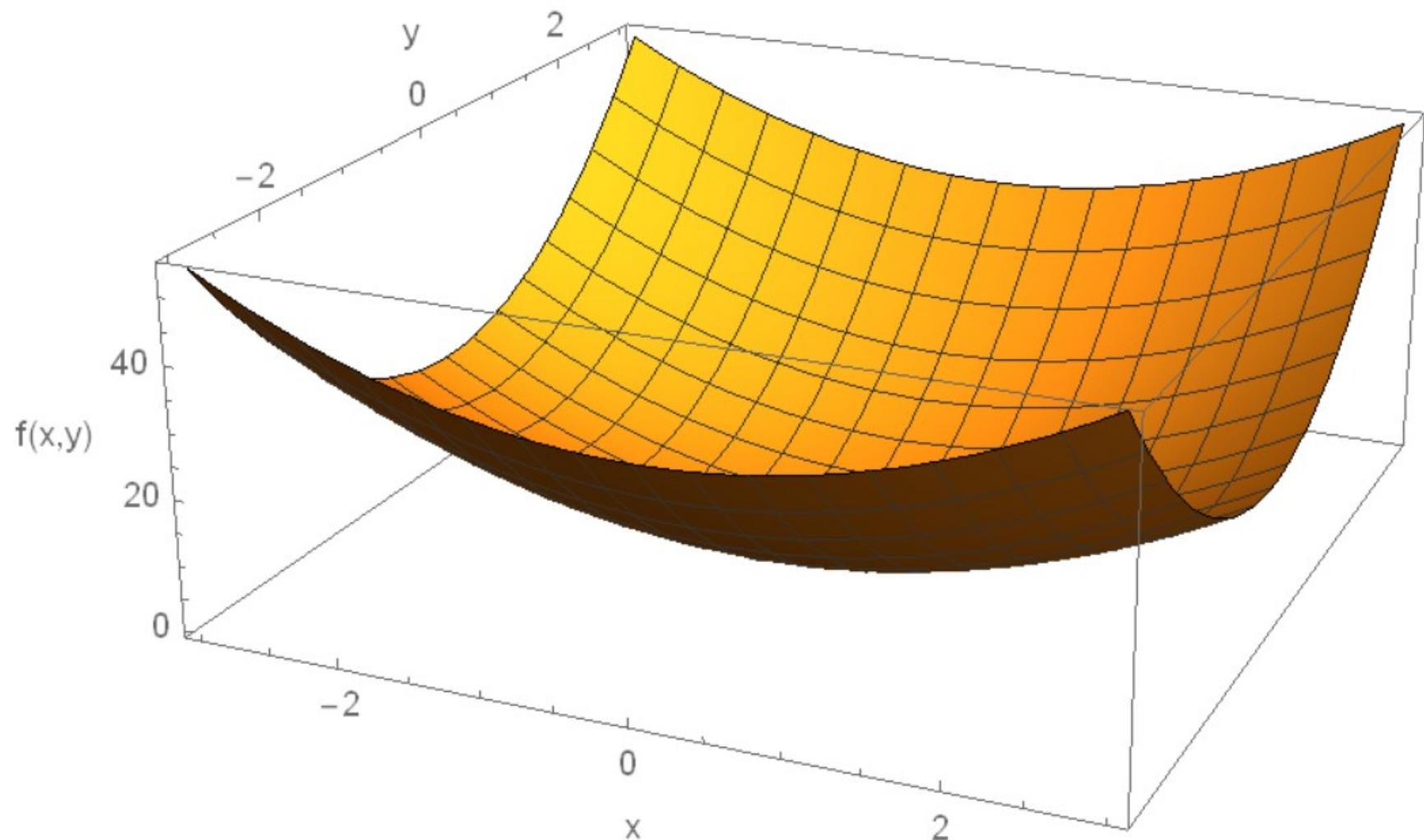
O Método do Gradiente Descendente é uma maneira de encontrar um mínimo local de uma função. O procedimento começa com uma suposição inicial da solução e a partir daí fazemos exames da inclinação da função nesse ponto. Avançamos para a solução considerando a direção negativa do gradiente e repetimos o processo. O algoritmo irá eventualmente convergir para um ponto que corresponde a um mínimo local onde o gradiente é zero. Esse método não é indicado para realizar ajustes não lineares, mas serve como base de outro método mais útil, o método de Marquardt que será descrito no Capítulo 7.

Um tamanho de passo,  $\lambda$ , errado pode implicar em não alcançar convergência, portanto uma seleção cuidadosa do tamanho do passo é importante. Se for demasiado grande ele divergirá, demasiado pequeno levará muito tempo para convergir. Uma opção é escolher um tamanho de passo fixo que assegure a convergência onde quer que você comece a descida do gradiente. Outra opção é escolher um tamanho de passo diferente em cada iteração (tamanho de passo adaptativo).



## EXEMPLO

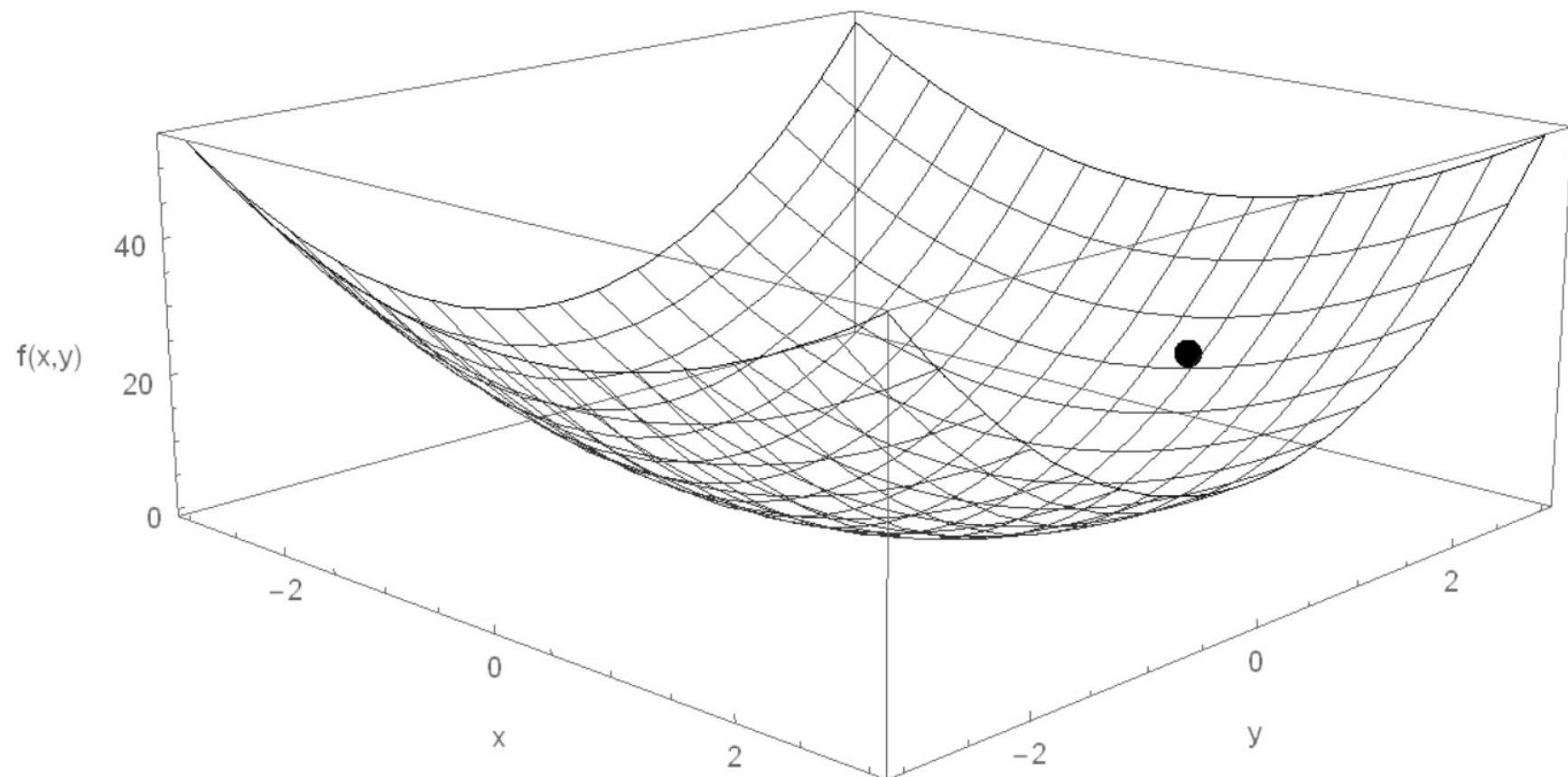
Aplique o algoritmo de descida mais íngreme à função  $f(x,y) = 2x^2 + 4y^2$ , utilizando tamanho do passo  $\lambda = 0,5$  e com ponto inicial dado por  $x_0 = 1$  e  $y_0 = 2$ .



Podemos observar que as variáveis  $x$  e  $y$  estão elevadas ao quadrado e da forma como a função está escrita é possível concluir que o ponto de mínimo é  $(0,0)$  o que implica que  $f(x,y)=0$ . Além disso, o valor da função no ponto inicial é dado por

$$f(x_0, y_0) = 2x_0^2 + 4y_0^2 = 2(1)^2 + 4(2)^2 = 18.$$

Nesse exemplo é possível, rapidamente, ter um ideia de afastamento do ponto inicial  $\mathbf{x}_0 = (1, 2)$  para o ponto de mínimo  $\mathbf{x}^* = (0, 0)$ .





## Passo 1

Cálculo do gradiente, do gradiente normalizado e verificação se  $\|\nabla f(x, y)\| < \varepsilon$ .

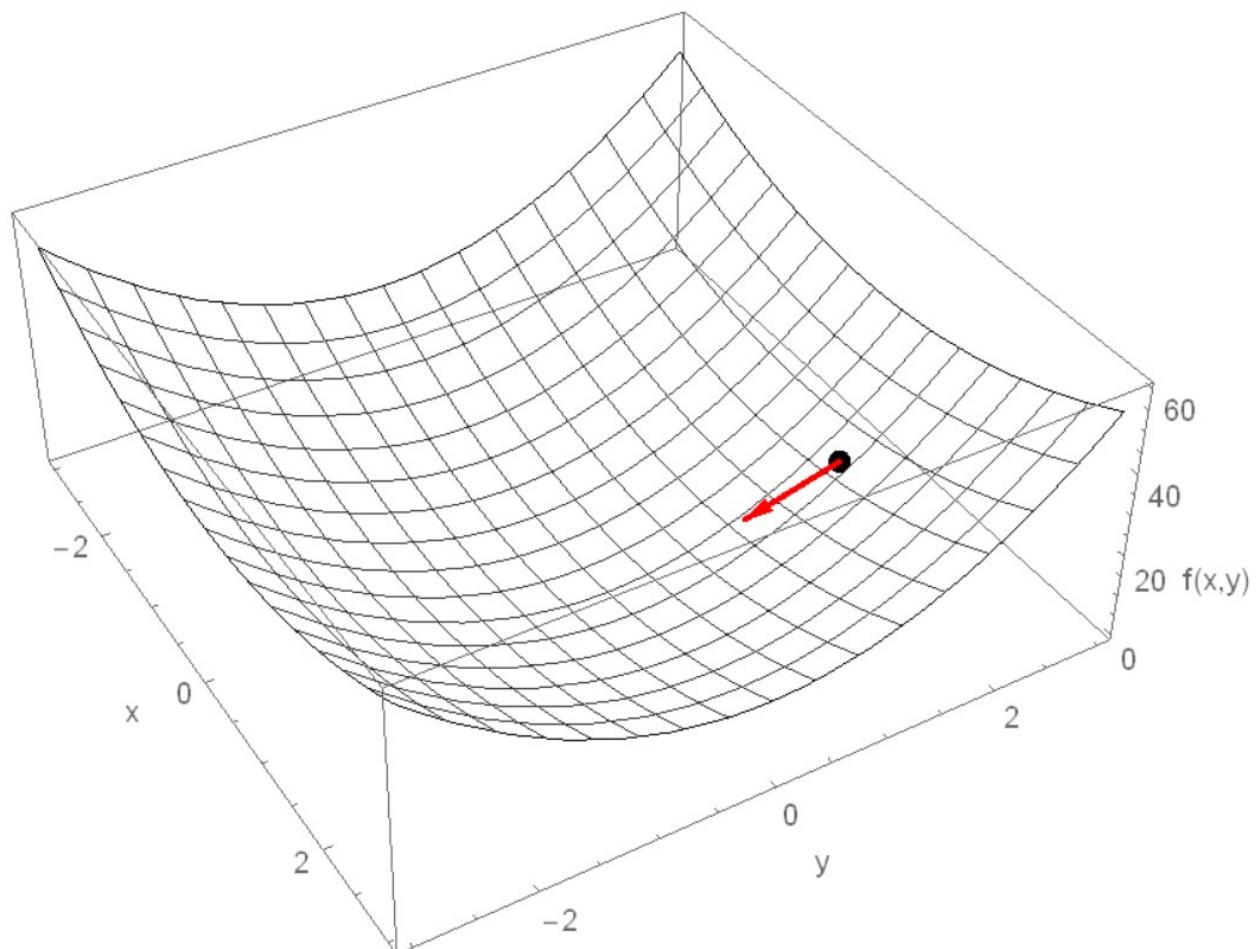
$\nabla f(x, y) = (4x, 8y)$  ou na forma matricial  $\nabla f(x, y) = \begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} 4x \\ 8y \end{bmatrix}$ . No ponto inicial o valor do gradiente é

$$\nabla f(x_0, y_0) = \begin{bmatrix} 4x_0 \\ 8y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 16 \end{bmatrix}.$$

O valor do gradiente normalizado é dado por

$$\frac{\nabla f(x, y)}{\|\nabla f(x, y)\|} = \frac{1}{\sqrt{4^2 + 16^2}} \begin{bmatrix} 4 \\ 16 \end{bmatrix} = \begin{bmatrix} 0,2425 \\ 0,9701 \end{bmatrix}.$$

Ou seja, o vetor  $\mathbf{n}$  é dado por  $(n_1, n_2) = (0,2425; 0,9701)$ . Nesse exemplo, não vamos fazer a verificação se  $\|\nabla f(x, y)\| < \varepsilon$ . A princípio estamos interessados na mecânica da condução do algoritmo, por isso vamos proceder apenas algumas iterações sem verificar a precisão da solução obtida. Cabe ao pesquisador determinar inicialmente o valor do  $\varepsilon$  que julga conveniente no processo de aproximação ao mínimo.





## Passo 2

Cálculo de  $(x_1, y_1)$  e de  $f(x_1, y_1)$ .

$$(x_1, y_1) = (x_0, y_0) - \lambda_0 \frac{\nabla f(x_0, y_0)}{\|\nabla f(x_0, y_0)\|},$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \lambda_0 \frac{\nabla f(x_0, y_0)}{\|\nabla f(x_0, y_0)\|},$$

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - 0,5 \begin{bmatrix} 0,2425 \\ 0,9701 \end{bmatrix} = \begin{bmatrix} 0,8788 \\ 1,5150 \end{bmatrix}.$$

A partir de  $\mathbf{x}^{(1)}$  podemos calcular  $f(\mathbf{x}^{(1)})$  que é dado por

$$f(\mathbf{x}_1) = 2x_1^2 + 4y_1^2 = 2(0,8788)^2 + 4(1,5150)^2 = 10,7255.$$

Ao invés de considerar o tamanho do passo como sendo um valor fixo, vamos expressar  $f(\mathbf{x}^{(1)})$  como função de  $\lambda_0$ , o que permite escrever

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \lambda_0 \frac{\nabla f(x_0, y_0)}{\|\nabla f(x_0, y_0)\|} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \lambda_0 \begin{bmatrix} 0,2425 \\ 0,9701 \end{bmatrix} = \begin{bmatrix} 1 - 0,2425\lambda_0 \\ 2 - 0,9701\lambda_0 \end{bmatrix},$$

$$\varphi(\lambda_0) = 2(1 - 0,2425\lambda_0)^2 + 4(2 - 0,9701\lambda_0)^2.$$



Como  $\varphi(\lambda_0)$  é uma função de  $\lambda_0$ , podemos utilizar a seguinte abordagem analítica

$$\frac{d\varphi(\lambda_0)}{d\lambda_0} = \frac{d}{d\lambda_0} \left[ 2(1 - 0,2425\lambda_0)^2 + 4(2 - 0,9701\lambda_0)^2 \right],$$

$$\frac{d\varphi(\lambda_0)}{d\lambda_0} = 2(2)(1 - 0,2425\lambda_0)(-0,2425) + 2(4)(2 - 0,9701\lambda_0)(-0,9701).$$

Igualando essa primeira derivada a zero temos que

$$2(2)(1 - 0,2425\lambda_0)(-0,2425) + 2(4)(2 - 0,9701\lambda_0)(-0,9701) = 0,$$

$$-0,9700(1 - 0,2425\lambda_0) - 7,7608(2 - 0,9701\lambda_0) = 0,$$

$$-0,9700 + 0,2352\lambda_0 - 15,5216 + 7,5288\lambda_0 = 0,$$

$$\lambda_0 = \frac{16,4916}{7,7640},$$

$$\lambda_0 = 2,1241.$$

Podemos expressar a derivada de  $\varphi(\lambda_0)$  em termos de  $\mathbf{n}$  e  $\mathbf{x}$ . Assim temos que

$$\frac{d\varphi(\lambda_0)}{d\lambda_0} = 2(2)(x_1 - n_1 \lambda_0)(-n_1) + 2(4)(x_2 - n_2 \lambda_0)(-n_2).$$

Igualando  $\frac{d\varphi(\lambda_0)}{d\lambda_0}$  a zero, produz

$$-x_1 n_1 + n_1^2 \lambda_0 - 2x_2 n_2 + 2n_2^2 \lambda_0 = 0,$$

$$(n_1^2 + 2n_2^2) \lambda_0 = x_1 n_1 + 2x_2 n_2,$$

$$\lambda_0 = \frac{x_1 n_1 + 2x_2 n_2}{n_1^2 + 2n_2^2}.$$

A partir do valor de  $\lambda_0$  determinamos o novo valor de  $\mathbf{x}^{(1)}$ .

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - 2,1241 \begin{bmatrix} 0,2425 \\ 0,9701 \end{bmatrix} = \begin{bmatrix} 0,4849 \\ -0,0606 \end{bmatrix}.$$

Assim  $f(\mathbf{x}^{(1)})$  é então dada por

$$f(\mathbf{x}^{(1)}) = 2x_1^2 + 4y_1^2 = 2(0,4849)^2 + 4(-0,0606)^2 = 0,4850.$$

O valor 0,4850 é o valor mínimo que conseguimos encontrar nesta iteração.



## Passo 3

Fazemos a atualização  $k = k + 1$  e determinamos  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \lambda^{(k)} \frac{\nabla f(\mathbf{x}^{(k)})}{\|\nabla f(\mathbf{x}^{(k)})\|}$ .

No ponto 1 o valor do gradiente é

$$\nabla f(x_1, y_1) = \begin{bmatrix} 4(0,4849) \\ 8(-0,0606) \end{bmatrix} = \begin{bmatrix} 1,9396 \\ -0,4848 \end{bmatrix}.$$

O valor do gradiente normalizado é dado por

$$\frac{\nabla f(x, y)}{\|\nabla f(x, y)\|} = \frac{1}{\sqrt{(1,9396)^2 + (-0,4848)^2}} \begin{bmatrix} 1,9396 \\ -0,4848 \end{bmatrix} = \begin{bmatrix} 0,9702 \\ -0,2425 \end{bmatrix}.$$

Próxima etapa é calcular  $\mathbf{x}^{(2)}$  em função de  $\lambda_1$ , que resulta em

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - \lambda_1 \frac{\nabla f(x_1, y_1)}{\|\nabla f(x_1, y_1)\|} = \begin{bmatrix} 0,4849 \\ -0,0606 \end{bmatrix} - \lambda_1 \begin{bmatrix} 0,9702 \\ -0,2425 \end{bmatrix} = \begin{bmatrix} 0,4849 - 0,9702\lambda_1 \\ -0,0606 + 0,2425\lambda_1 \end{bmatrix}.$$

A função  $\varphi(\lambda_1)$  que é uma função de  $\lambda_1$ , pode ser escrita como

$$\varphi(\lambda_1) = 2(0,4849 - 0,9702\lambda_1)^2 + 4(-0,0606 + 0,2425\lambda_1)^2.$$

Igualando a primeira derivada de  $\varphi(\lambda_1)$  a zero permite determinar o valor de  $\lambda_1$

$$\frac{d\varphi(\lambda_1)}{d\lambda_1} = \frac{d}{d\lambda_1} \left[ 2(0,4849 - 0,9702\lambda_1)^2 + 4(-0,0606 + 0,2425\lambda_1)^2 \right] = 0,$$

$$2(2)(0,4849 - 0,9702\lambda_1)(-0,9702) + 2(4)(-0,0606 + 0,2425\lambda_1)(0,2425) = 0,$$

$$-3,8808(0,4849 - 0,9702\lambda_1) + 1,9400(-0,0606 + 0,2425\lambda_1) = 0,$$

$$-1,8818 + 3,7652\lambda_1 - 0,1176 + 0,4705\lambda_1 = 0,$$

$$\lambda_1 = \frac{1,8936}{4,2357},$$

$$\lambda_1 = 0,4471.$$

A partir do valor de  $\lambda_1$  determinamos o novo valor  $\mathbf{x}^{(2)}$ .

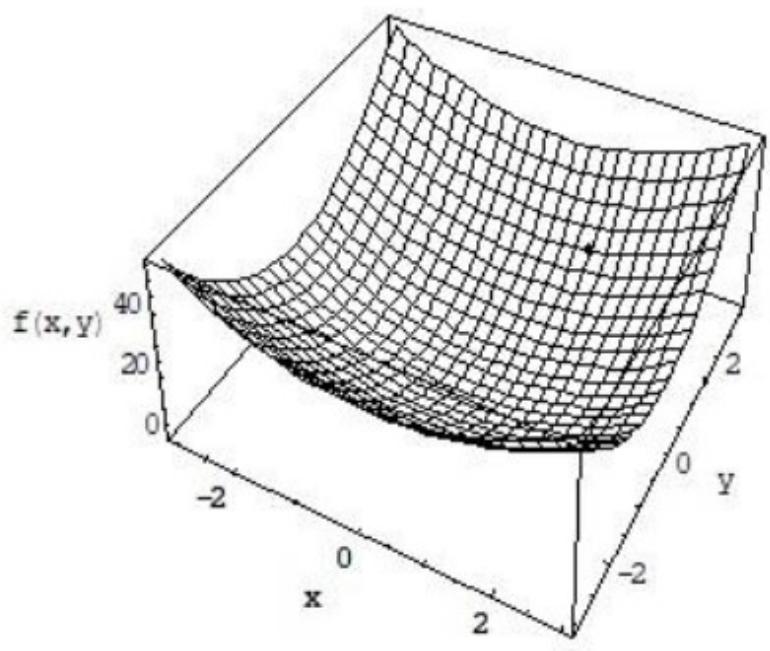
$$\mathbf{x}^{(2)} = \begin{bmatrix} 0,4849 \\ -0,0606 \end{bmatrix} - 0,4471 \begin{bmatrix} 0,9702 \\ -0,2425 \end{bmatrix} = \begin{bmatrix} 0,0511 \\ 0,0478 \end{bmatrix}.$$



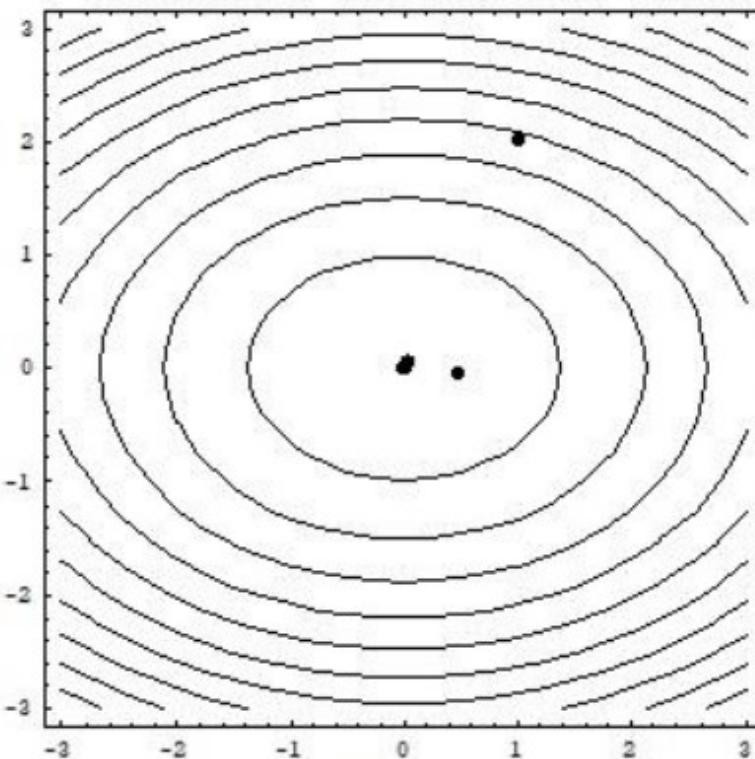
Assim  $f(\mathbf{x}^{(2)})$  é então dada por

$$f(\mathbf{x}^{(2)}) = 2x_1^2 + 4y_1^2 = 2(0,0511)^2 + 4(0,0478)^2 = 0,0143.$$

O valor 0,0143 é o valor mínimo que conseguimos encontrar nesta iteração. Verifique que este valor é menor que o obtido na iteração anterior, ou seja, estamos nos aproximando do  $\mathbf{x}^*$ . Esse processo iterativo pode ser programado no MATLAB, utilizando a seguinte sequência de comandos.



(a)



(b)

**Figura** Gráficos de (a)  $f(x,y) = 2x^2 + 4y^2$  e (b) vista de topo de  $f(x,y)$ , em ambos são mostrados o ponto inicial e os pontos obtidos pelo algoritmo do gradiente descendente.

## MATLAB - Algoritmo da descida mais íngreme utilizado no Exemplo

```
k=0; %Contador de iteração
x=[1;2]; %Ponto inicial
eps=1; %Valor inicial para o erro, denotado anteriormente como epsilon

%Looping utilizando o comando while
while eps>0.00001&k<10
    f=2*x(1)^2+4*x(2)^2 %Função a ser analisada
    gradf=[4*x(1);8*x(2)] %Gradiente da função
    normgradf=(gradf(1)^2+gradf(2)^2)^0.5 %Norma do gradiente
    n=gradf/normgradf %Gradiente normalizado

    %Expressão do lambda, obtida a partir da função phi(lambda)
    lambda=(n(1)*x(1)+2*n(2)*x(2))/(n(1)^2+2*n(2)^2)

    xnovo=x-lambda*n
    x=xnovo
    fnova=2*x(1)^2+4*x(2)^2
    eps=abs(fnova-f)
    k=k+1
end
k,x,f,fnova,eps
```



python™



ChatGPT

```

import numpy as np

# Inicialização das variáveis
k = 0 # Contador de iteração
x = np.array([1.0, 2.0]) # Ponto inicial
eps = 1.0 # Valor inicial para o erro

# Loop enquanto eps > 0.00001 e k < 10
while eps > 0.00001 and k < 10:
    # Função a ser analisada: f = 2*x1^2 + 4*x2^2
    f = 2 * x[0]**2 + 4 * x[1]**2

    # Gradiente da função: gradf = [4*x1; 8*x2]
    gradf = np.array([4 * x[0], 8 * x[1]])

    # Norma do gradiente
    normgradf = np.sqrt(gradf[0]**2 + gradf[1]**2)

    # Gradiente normalizado
    n = gradf / normgradf

    # Cálculo de lambda
    lambda_val = (n[0] * x[0] + 2 * n[1] * x[1]) / (n[0]**2 + 2 * n[1]**2)

    # Atualização do ponto x
    xnovo = x - lambda_val * n
    x = xnovo

    # Nova avaliação da função
    fnova = 2 * x[0]**2 + 4 * x[1]**2

    # Cálculo do erro
    eps = abs(fnova - f)

    # Incremento do contador
    k += 1

# Exibir resultados
print(f"k = {k}")
print(f"x = {x}")
print(f"f = {f}")
print(f"fnova = {fnova}")
print(f"eps = {eps}")

```



k = 5  
x = [ 3.51781599e-04 -4.39726998e-05]  
f = 9.475598615303712e-06  
fnova = 2.552349795368005e-07  
eps = 9.220363635766912e-06

