REDES NEURAIS ARTIFICIAIS

MÉTODO DO GRADIENTE

Zenilton K. G. Patrocínio Jr
zenilton@pucminas.br

Gradiente da Função de Perda

O gradiente $\nabla_W L(W)$ representa o vetor de derivadas parciais

$$\nabla_W L(W) = \left[\frac{\partial L}{\partial W_{11}}, \frac{\partial L}{\partial W_{12}}, \cdots, \frac{\partial L}{\partial W_{21}}, \frac{\partial L}{\partial W_{22}}, \cdots\right]^T$$

Gradiente da Função de Perda

O gradiente $\nabla_W L(W)$ representa o vetor de derivadas parciais

$$\nabla_W L(W) = \left[\frac{\partial L}{\partial W_{11}}, \frac{\partial L}{\partial W_{12}}, \cdots, \frac{\partial L}{\partial W_{21}}, \frac{\partial L}{\partial W_{22}}, \cdots\right]^T$$

em que, por exemplo, $\frac{\partial L}{\partial W_{11}}$ mede o quão rápido varia a perda L em relação a uma variação do coeficiente W_{11} da matriz W

Gradiente da Função de Perda

O gradiente $\nabla_W L(W)$ representa o vetor de derivadas parciais

$$\nabla_W L(W) = \left[\frac{\partial L}{\partial W_{11}}, \frac{\partial L}{\partial W_{12}}, \cdots, \frac{\partial L}{\partial W_{21}}, \frac{\partial L}{\partial W_{22}}, \cdots\right]^T$$

em que, por exemplo, $\frac{\partial L}{\partial W_{11}}$ mede o quão rápido varia a perda L em relação a uma variação do coeficiente W_{11} da matriz W

Neste caso, para se minimizar o valor do risco empírico é necessário se igualar a zero o gradiente de L em relação à matriz W, isto é

$$\nabla_W L(W) = 0$$

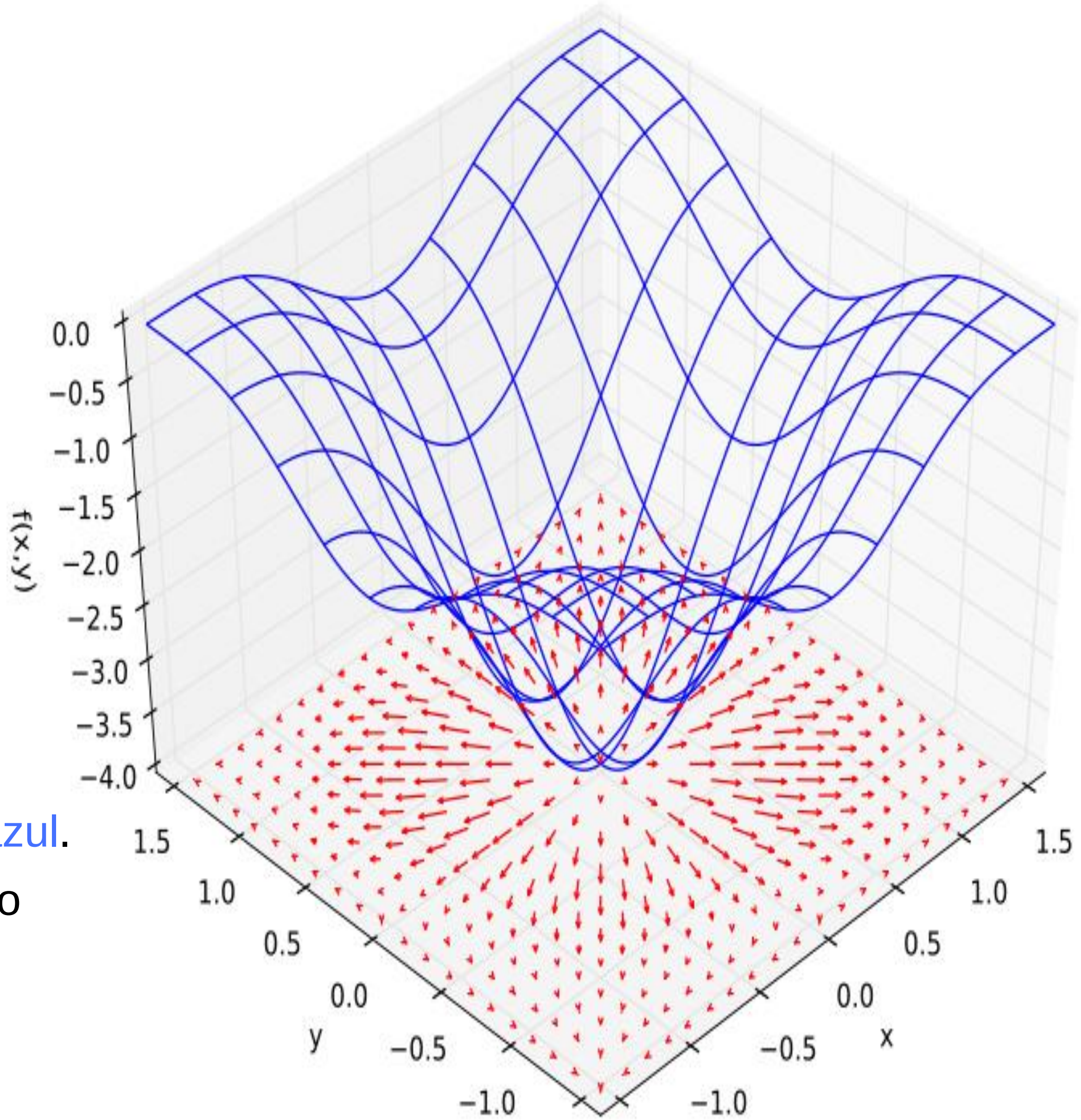
considerando, assim, que L é uma função da matriz W

Gradiente da Função de Perda

Quando $\nabla_W L(W) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção

Gradiente da Função de Perda

Quando $\nabla_W L(W) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção



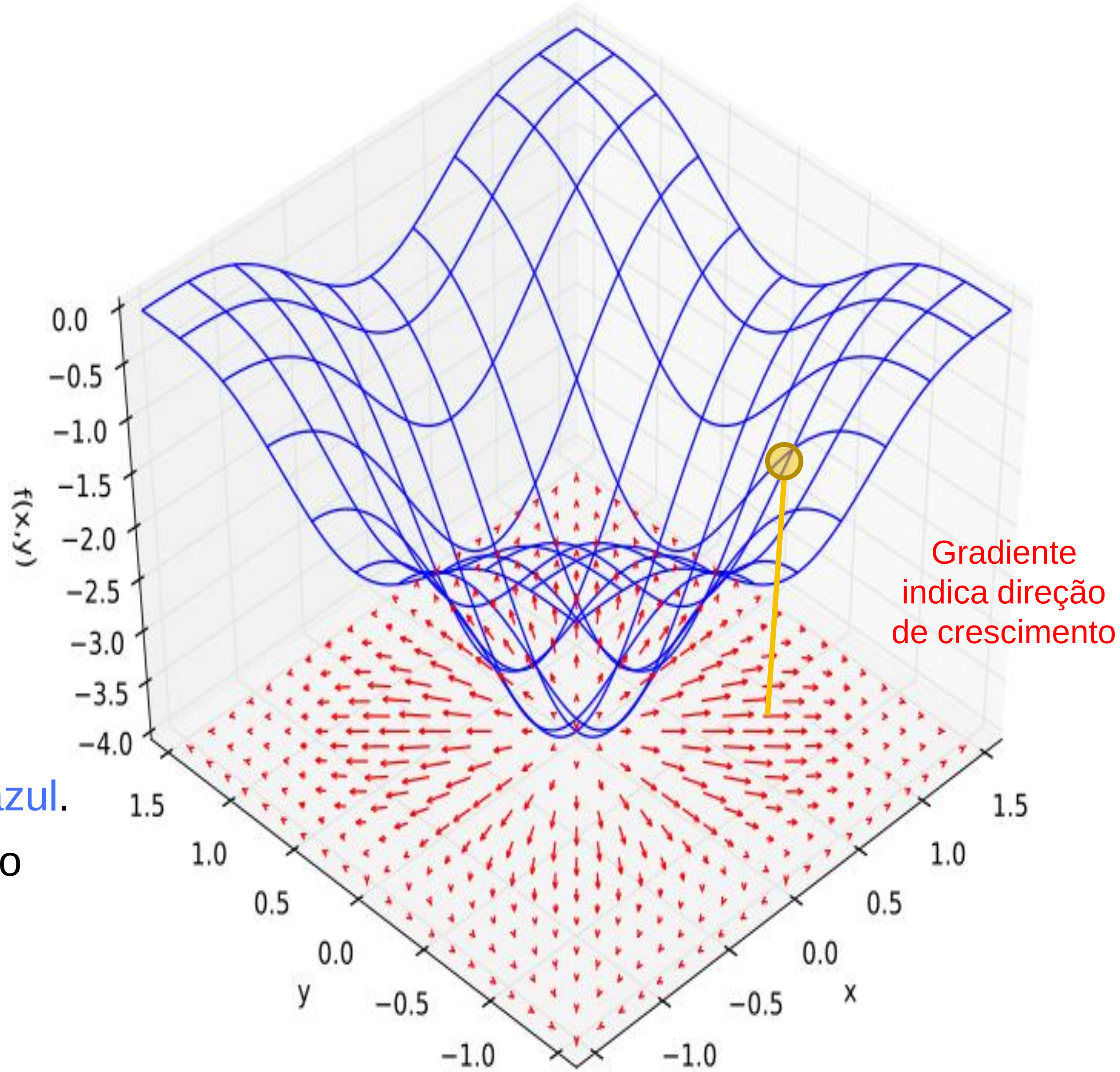
A superfície da perda aparece em azul.

Os gradientes são mostrados abaixo como setas vermelhas.

O gradiente é zero no mínimo.

Gradiente da Função de Perda

Quando $\nabla_W L(W) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção



A superfície da perda aparece em azul.

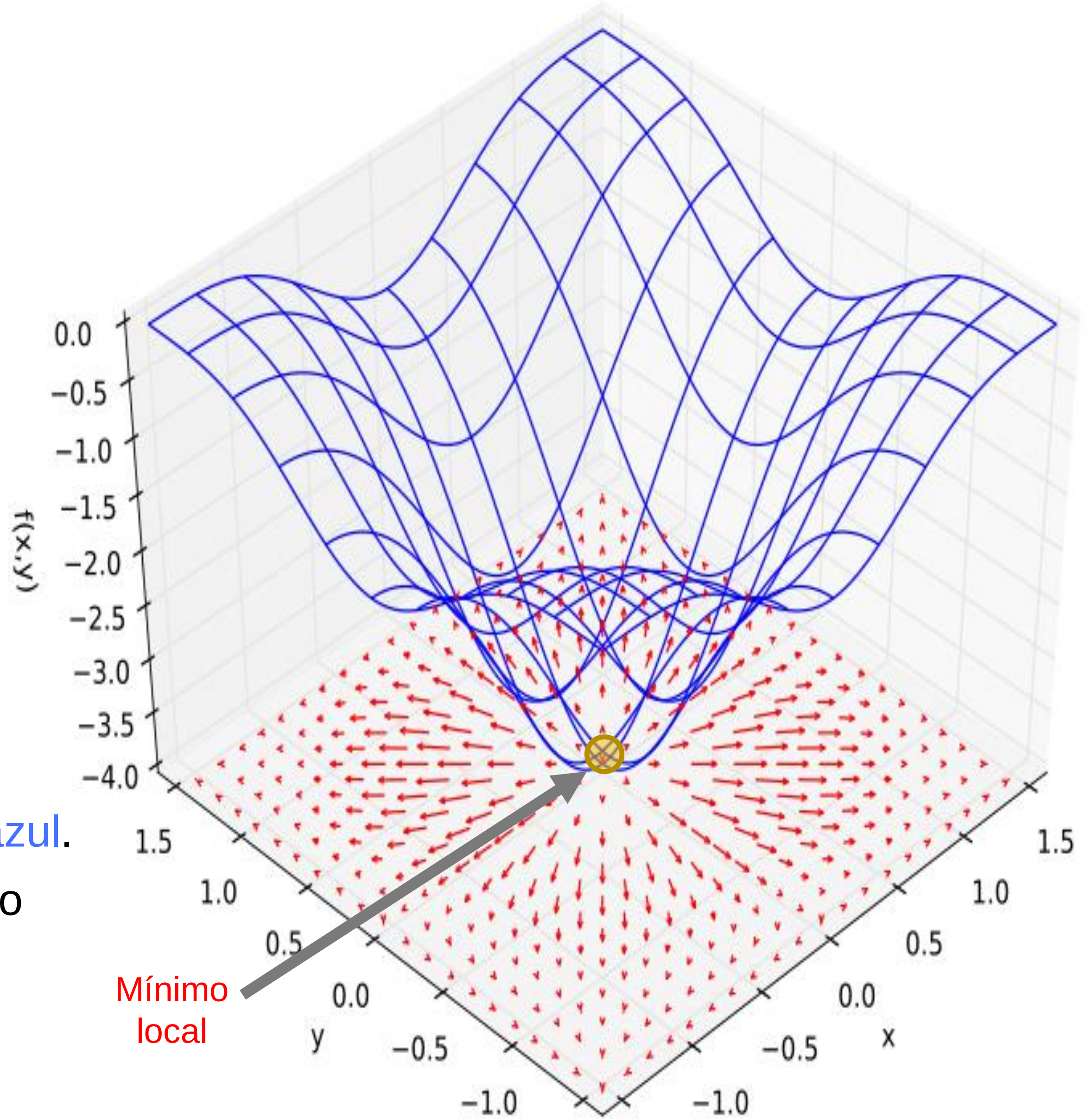
Os gradientes são mostrados abaixo como setas vermelhas.

O gradiente é zero no mínimo.

Gradiente da Função de Perda

Quando $\nabla_W L(W) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção

Portanto, obteve-se um ótimo local



A superfície da perda aparece em azul.

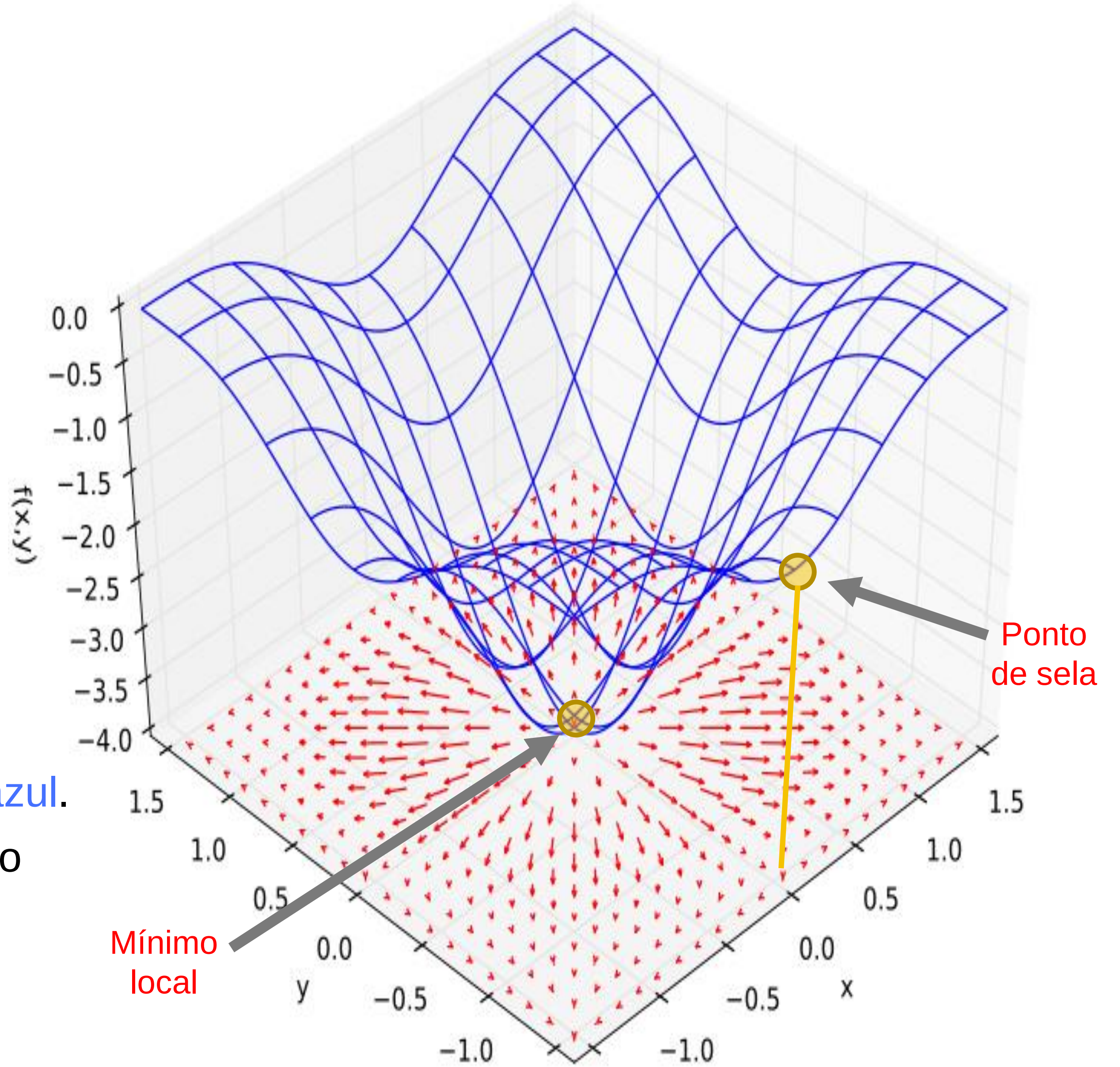
Os gradientes são mostrados abaixo como setas vermelhas.

O gradiente é zero no mínimo.

Gradiente da Função de Perda

Quando $\nabla_W L(W) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção

Portanto, obteve-se um ótimo local (ou, pelo menos, um ponto de sela)



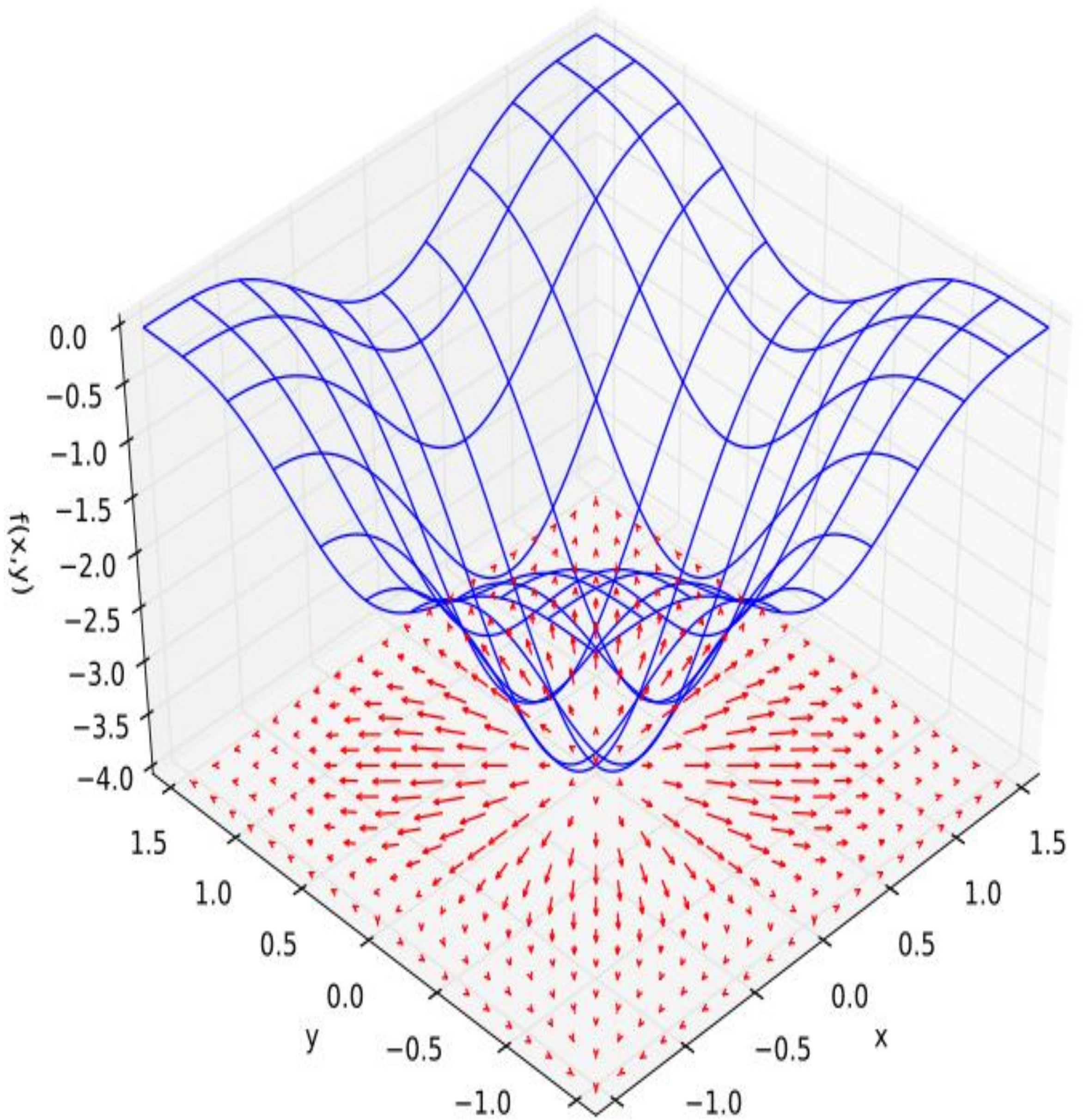
A superfície da perda aparece em azul.

Os gradientes são mostrados abaixo como setas vermelhas.

O gradiente é zero no mínimo.

Método do Gradiente (ou Descida Mais Íngreme)

Para se alcançar um ponto de mínimo deve-se seguir a direção contrária ao gradiente,

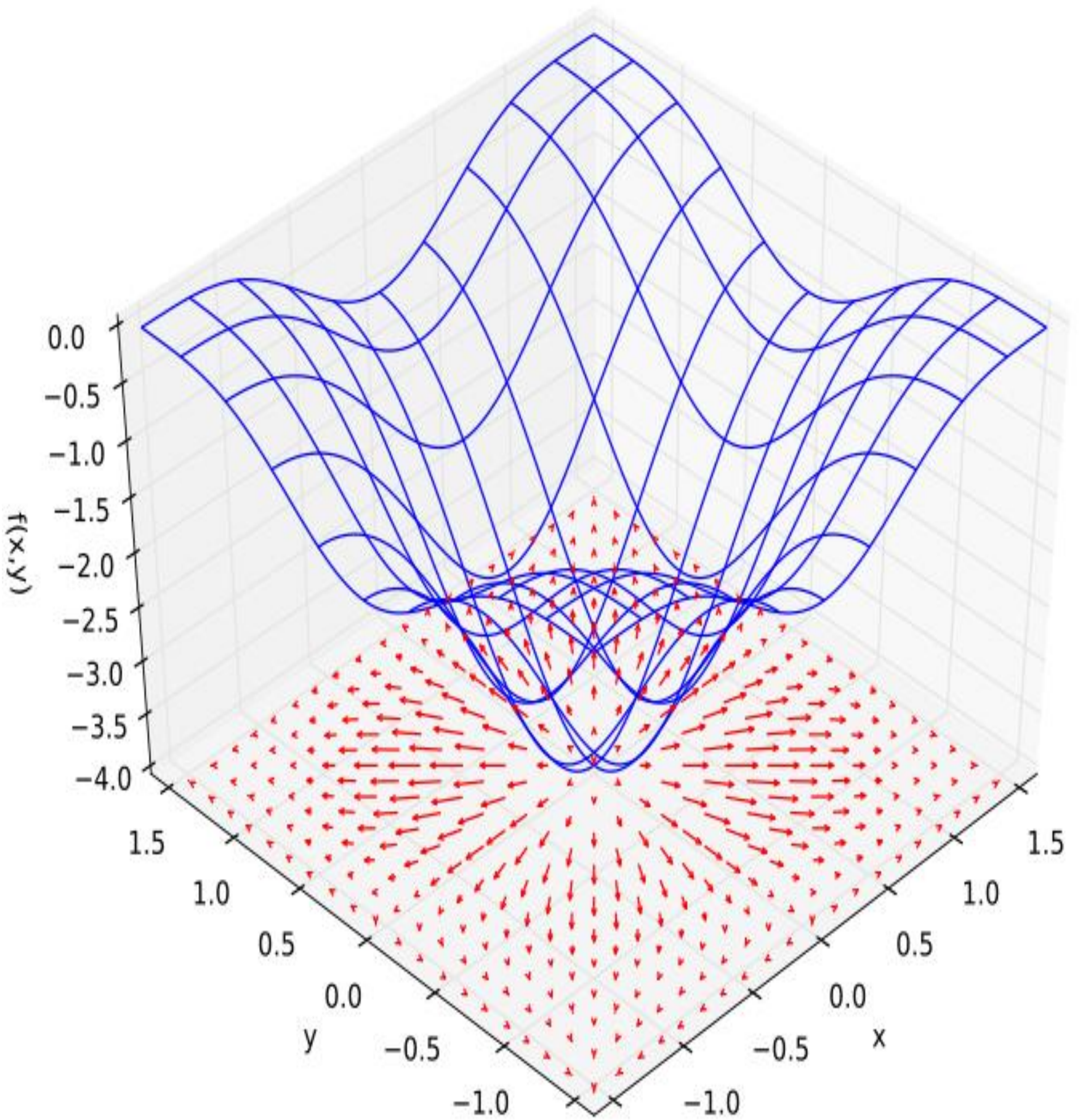


Método do Gradiente (ou Descida Mais Íngreme)

Para se alcançar um ponto de mínimo deve-se seguir a direção contrária ao gradiente,

isto é, deve-se dar passos na direção

$$-\nabla_W L(W)$$



Método do Gradiente (ou Descida Mais Íngreme)

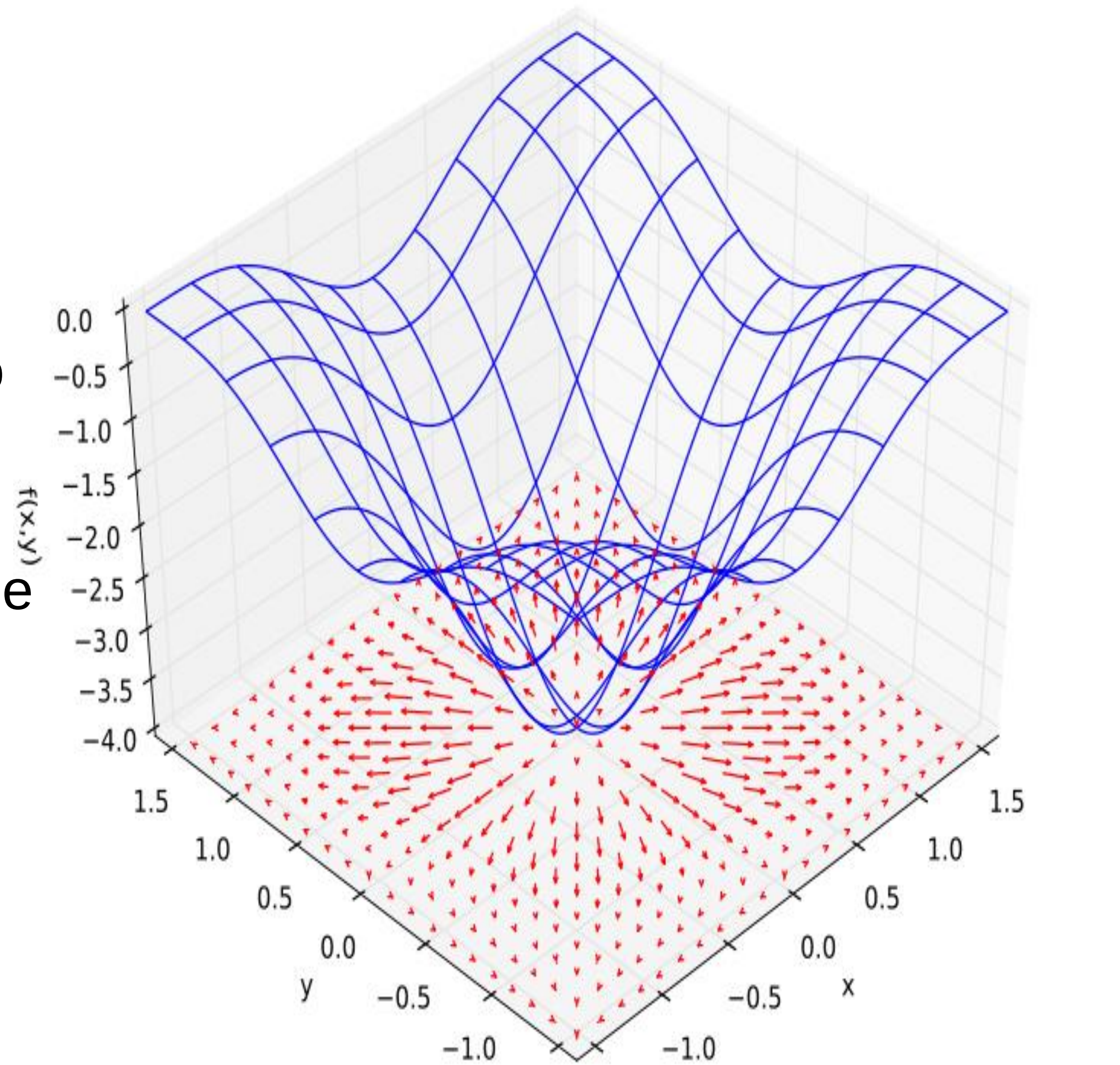
Para se alcançar um ponto de mínimo deve-se seguir a direção contrária ao gradiente,

isto é, deve-se dar passos na direção

$$-\nabla_W L(W)$$

Mais precisamente, seja W^t a matriz de pesos no passo t então

$$W^{t+1} = W^t - \alpha \ \nabla_W L(W)$$



Método do Gradiente (ou Descida Mais Íngreme)

Para se alcançar um ponto de mínimo deve-se seguir a direção contrária ao gradiente,

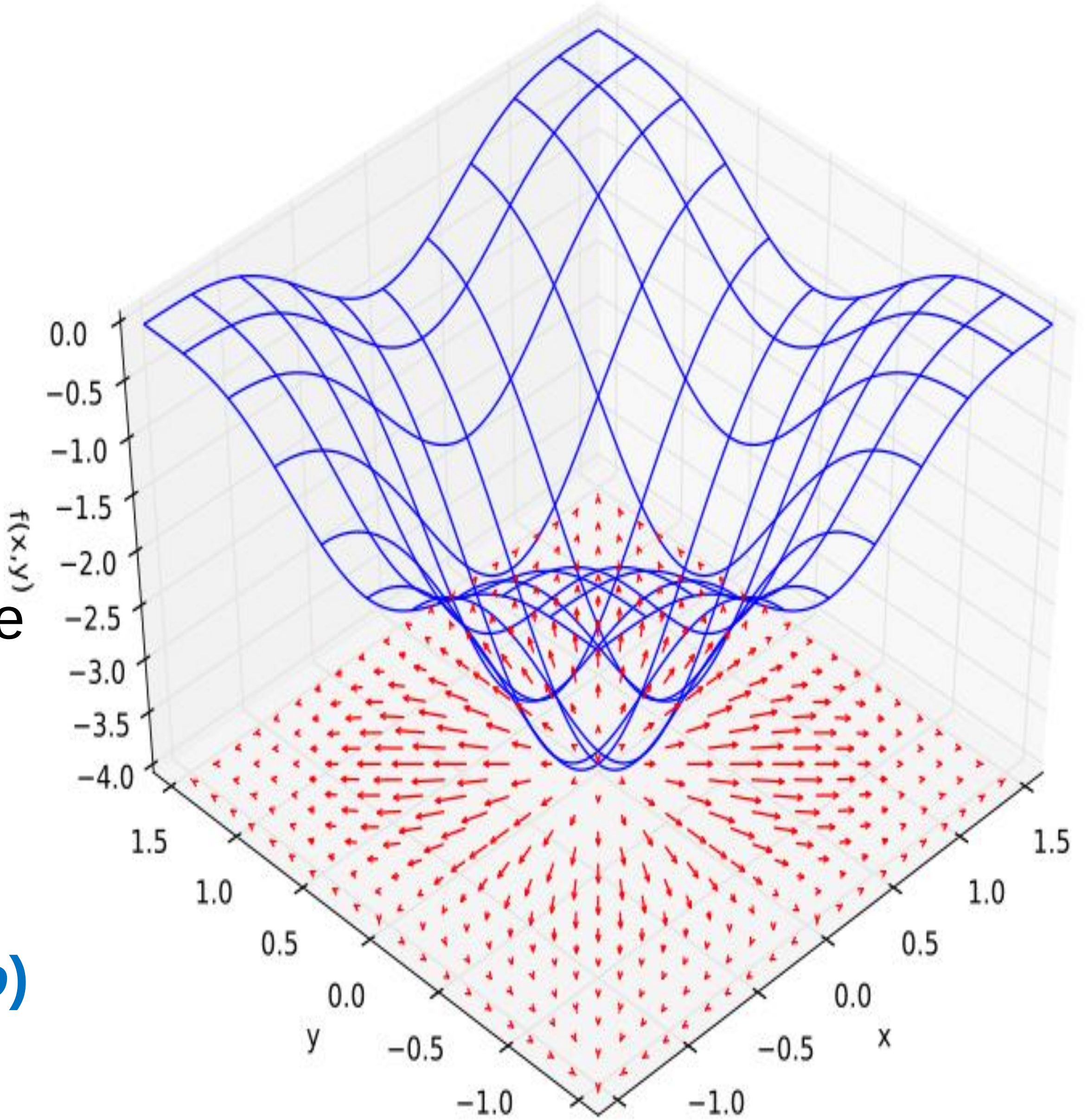
isto é, deve-se dar passos na direção

$$-\nabla_W L(W)$$

Mais precisamente, seja W^t a matriz de pesos no passo t então

$$W^{t+1} = W^t - \alpha \ \nabla_W L(W)$$

em que α é chamado de ***taxa de aprendizado (ou tamanho do passo)***



Método do Gradiente (ou Descida Mais Íngreme)

Para se alcançar um ponto de mínimo deve-se seguir a direção contrária ao gradiente,

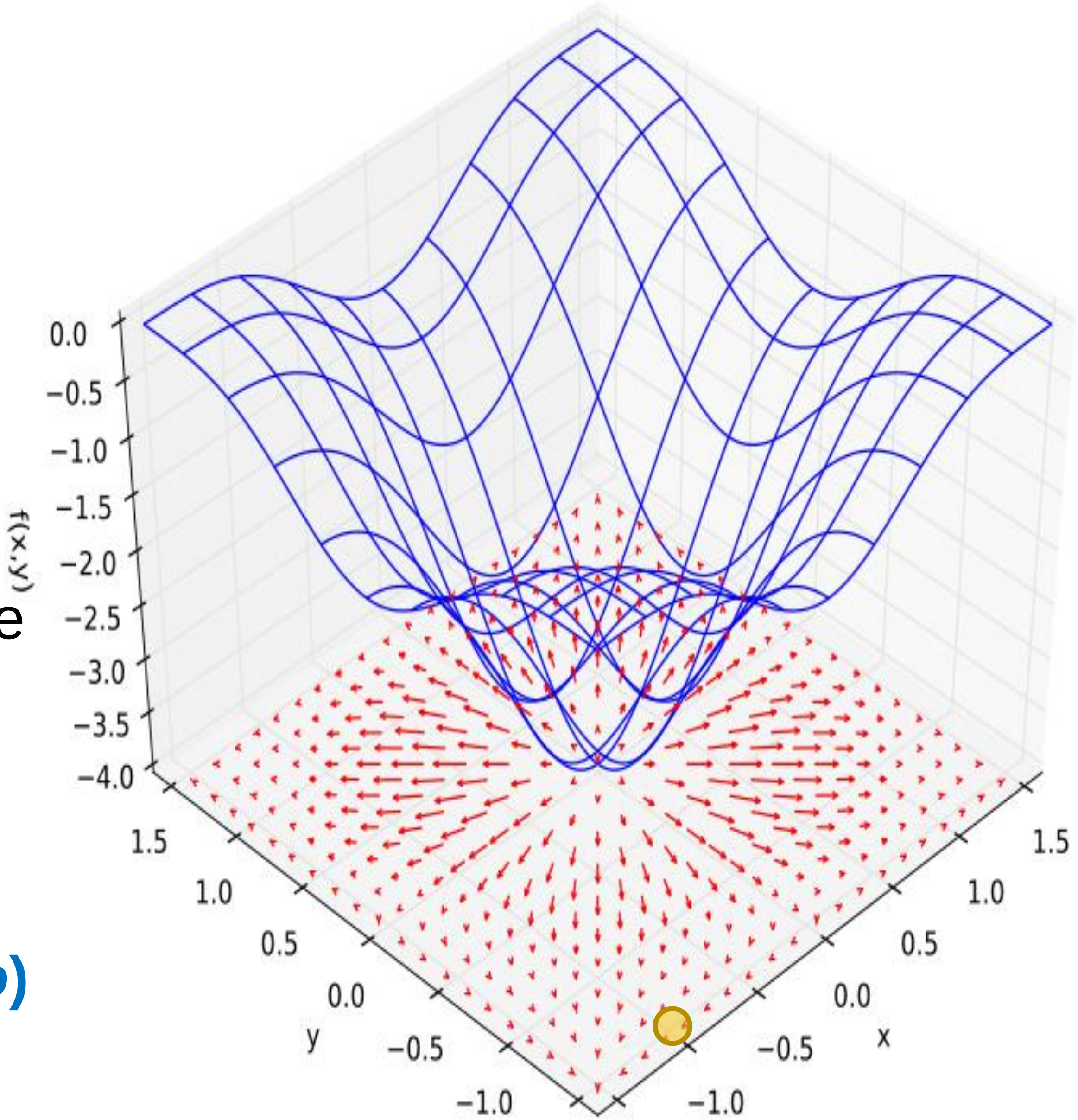
isto é, deve-se dar passos na direção

$$-\nabla_W L(W)$$

Mais precisamente, seja W^t a matriz de pesos no passo t então

$$W^{t+1} = W^t - \alpha \ \nabla_W L(W)$$

em que α é chamado de ***taxa de aprendizado (ou tamanho do passo)***



Método do Gradiente (ou Descida Mais Íngreme)

```
# Vanilla Gradient Descent

while True:
  weights_grad = evaluate_gradient(loss_fun, data, weights)
  weights += - step_size * weights_grad # perform parameter update
```

Método do Gradiente (ou Descida Mais Íngreme)

```
# Vanilla Gradient Descent

while True:
  weights_grad = evaluate_gradient(loss_fun, data, weights)
  weights += - step_size * weights_grad # perform parameter update
```