**Treinamento da Rede Neural =**

**Minimização de $F_{in}$**

**Otimização (minimização) de Funções Contínuas**

$$F(w_1, w_2, \ldots, w_n) = F(\underline{w}) \qquad \underline{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

Propriedades

$$F(\underline{w}) \geq 0$$

$$\frac{\partial F}{\partial w_i} \quad \exists \quad \forall\, w_i, \ \forall\, \underline{w}$$

Como encontrar $\underline{w}^*$ tal que

$$F(\underline{w}^*) \leq F(\underline{w}) \qquad \forall\ \underline{w} \neq \underline{w}^* \quad ?$$

## Otimização de F

**não linear**

**não convexa**

**$\underline{w}$ ordem elevada**

**irrestrita**

**contínua, derivável**

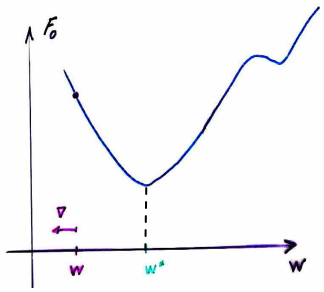
**Como fazer ???**

# Noções Elementares de Otimização

## 1 – Função objetivo à minimizar

(encontrar o ponto de mínimo ou minimante)

$$F(w_1, w_2, \ldots) = F(\underline{w})$$



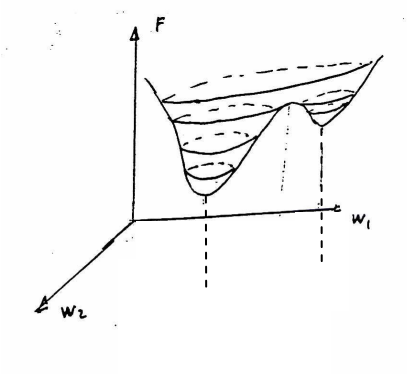

Mínimo, minimante $\underline{w}^*$

$$F(\underline{w}^*) < F(\underline{w}^* + \underline{\Delta w}) \quad \forall\ \underline{\Delta w} \ \mid\ |\underline{\Delta w}| < \varepsilon$$

$\rightarrow$ mínimo local

Mínimo global

$$F(\underline{w}^*) \leq F(w) \quad \forall\ w \neq \underline{w}^*$$




## 2 – Aproximações no entorno de um ponto qualquer $\underline{w}_0$ (Taylor)

$$\underline{w} = \underline{w}_0 + \underline{\Delta w} \qquad |\underline{\Delta w}| \leq \varepsilon$$

$$F(\underline{w}_0 + \underline{\Delta w}) = F(\underline{w}_0) + \sum_i \left.\frac{\partial F}{\partial w_i}\right|_{\underline{w}_0} \Delta w_i + \frac{1}{2}\sum_i \sum_j \left.\frac{\partial^2 F}{\partial w_i\, \partial w_j}\right]_{\underline{w}_0} \Delta w_i \Delta w_j + \ldots$$

Aproximação linear ou de 1ª. ordem

$$F(\underline{w}_0 + \underline{\Delta w}) \approx F(\underline{w}_0) + \sum_i \left.\frac{\partial F}{\partial w_i}\right|_{\underline{w}_0} \Delta w_i$$

a aproximação é válida em um pequeno entorno de $\underline{w}_0$

Notação usando vetores

vetor **variáveis livres** vetor **acréscimos** nas variáveis livres

$$\underline{w} = \begin{matrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{matrix} \Bigg] \qquad \underline{\Delta w} = \Delta w_i \Bigg] = \begin{matrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_n \end{matrix} \Bigg]$$

vetor **gradiente**

$$\nabla F(\underline{w}_0) = \left. \frac{\partial F}{\partial w_i} \right|_{\underline{w}_0} \Bigg] = \begin{matrix} \left. \frac{\partial F}{\partial w_1} \right|_{\underline{w}_0} \\ \left. \frac{\partial F}{\partial w_2} \right|_{\underline{w}_0} \\ \vdots \\ \left. \frac{\partial F}{\partial w_n} \right|_{\underline{w}_0} \end{matrix} \Bigg]$$

$$F(\underline{w}_0 + \Delta\underline{w}) = F(\underline{w}_0) + \sum_i \left. \frac{\partial F}{\partial w_i} \right|_{\underline{w}_0} \Delta w_i + \ldots$$

$$F(\underline{w}_0 + \Delta\underline{w}) = F(\underline{w}_0) + \Delta\underline{w}^t \nabla F(\underline{w}_0) + \cdots$$

**Aproximação linear ou de 1ª. ordem**

$$F(\underline{w}_0 + \Delta\underline{w}) \simeq F(\underline{w}_0) + \Delta\underline{w}^t \nabla F(\underline{w}_0)$$

válida em um pequeno entorno de $\underline{w}_0$

## 3 - Mínimo, minimante $\underline{w}^*$

$$F(\underline{w}^* + \Delta\underline{w}) = F(\underline{w}^*) + \Delta F(\underline{w}^*) > F(\underline{w}^*) \quad \forall\ \Delta\underline{w} \quad | \quad |\Delta\underline{w}| \leq \varepsilon$$

$$F(\underline{w}^* + \Delta\underline{w}) \cong F(\underline{w}^*) + \Delta w^t \nabla F$$

$$\Delta F = \Delta\underline{w}^t\ \nabla\underline{F} = |\Delta\underline{w}|\ |\nabla F|\ \cos\angle\Delta\underline{w}, \nabla\underline{F} > 0$$

$$+ \quad + \quad +/-$$

$$\Rightarrow \quad \nabla\underline{F}(\underline{w}^*) = \underline{0}$$

Nos pontos de mínimo (minimantes) o gradiente é nulo

## 4 - Propriedades do Gradiente

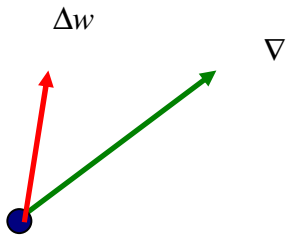
4.1 - Nos pontos de mínimo (minimantes) o gradiente é nulo

$$\nabla\underline{F}(\underline{w}^*) = \underline{0}$$

4.2 - Para $\Delta\underline{w}$ pequeno e constante o gradiente indica a direção de máximo $\Delta F$ (e $-\nabla\underline{F}$ indica a direção de maior decréscimo de F )

vale a aproximação de primeira ordem $|\Delta\underline{w}| \leq \varepsilon|$

$$F(\underline{w}_0 + \Delta\underline{w}) = F(\underline{w}_0) + \Delta F \cong F(\underline{w}_0) + \Delta\underline{w}^t\ \nabla\underline{F}$$




$$\Delta F \cong \Delta\underline{w}^t\ \nabla F = |\Delta\underline{w}|\ |\nabla\underline{F}| \cos\theta$$

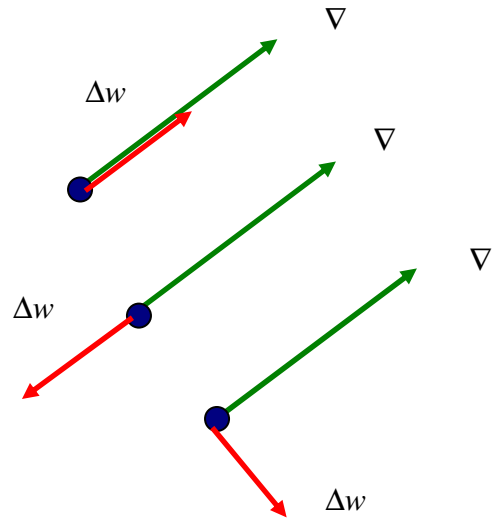
$$\Delta F \cong \underline{\Delta w}^t \nabla F = |\underline{\Delta w}| \; |\underline{\nabla} F| \cos\theta$$

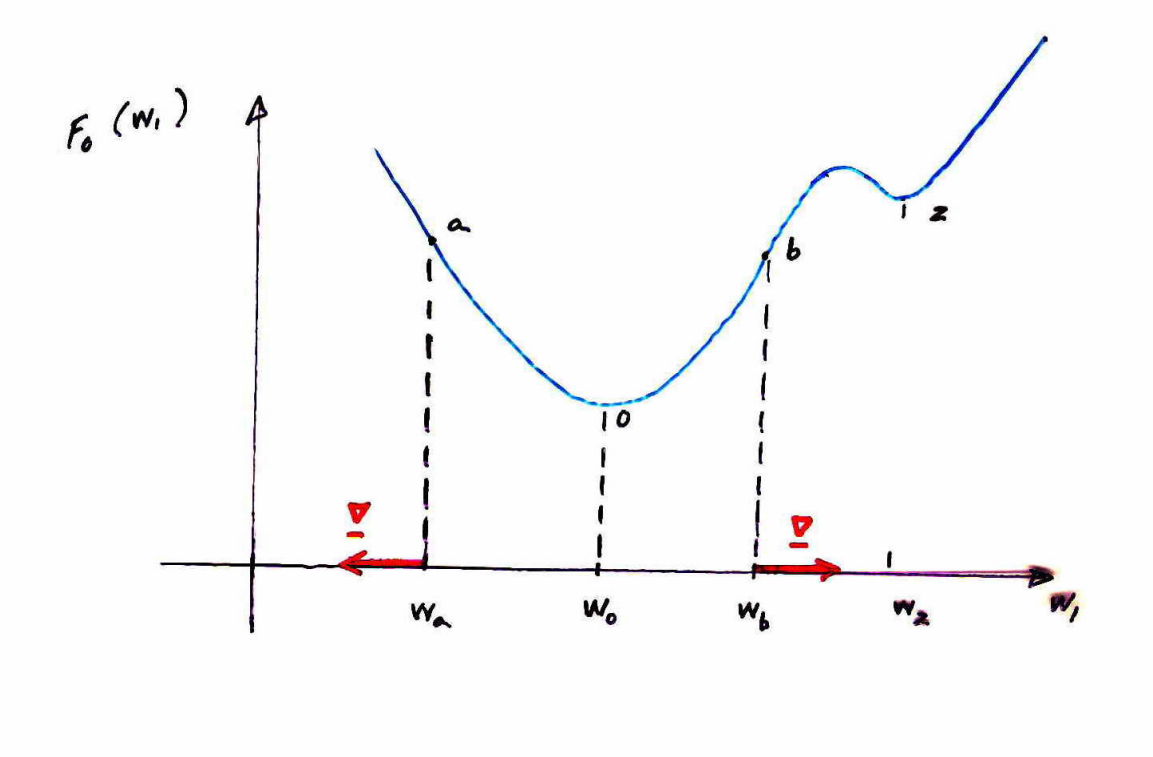
$\theta = 0 \quad \Rightarrow \quad \Delta F$ max

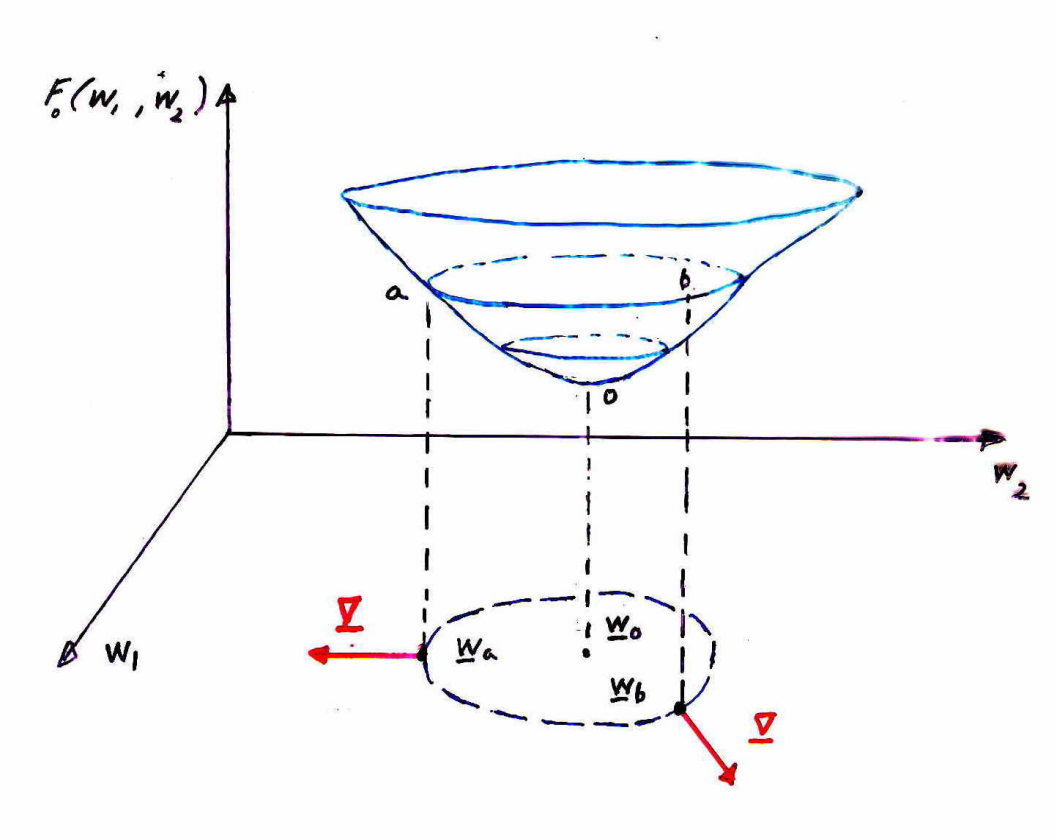
$\theta = \pi \quad \Rightarrow \quad \Delta F$ min

$(-\Delta F$ max)

$\theta = \pm \frac{\pi}{2} \quad \Rightarrow \quad \Delta F = 0$






## Otimização

## Determinação do minimante de F(w)

$$\underline{w}^* \quad | \quad F(\underline{w}^*) \text{ é mínima}$$

ou

$$\underline{w}^* \quad | \quad \nabla F(\underline{w}^*) = \underline{0}$$

$$\frac{\partial F}{\partial w_i} = 0 \quad \forall\ w_i$$

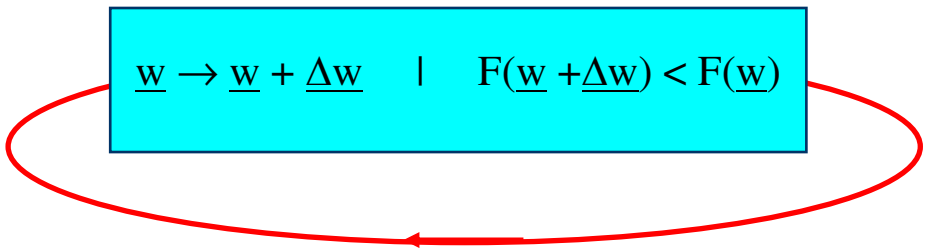
## Otimização: Solução Analítica

$$\underline{w}^* \quad | \quad \nabla F(\underline{w}^*) = \underline{0}$$

**usualmente muito complexo**

## Otimização: Métodos Numéricos Recursivos

### Descida continuada



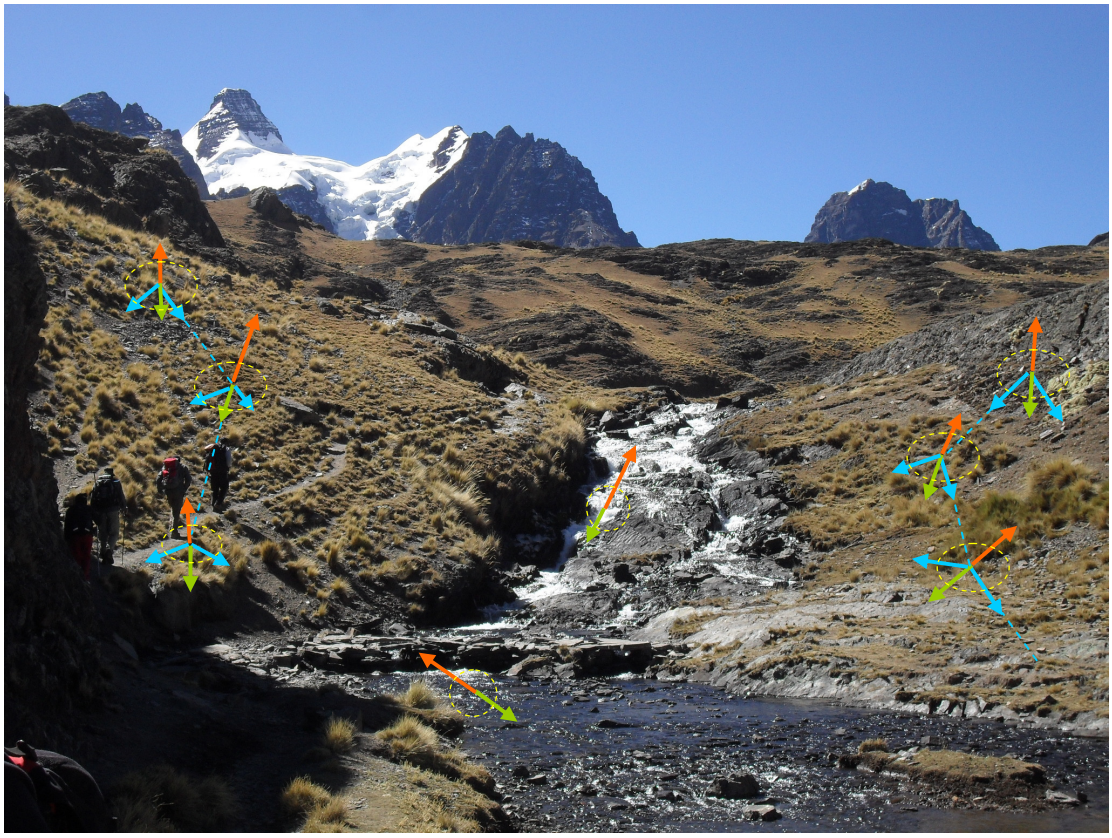

$\Delta\underline{w}$ ??

$$\Delta\underline{w} = \alpha\, \underline{d}$$

Passo
controla $|\Delta\underline{w}|$
$\alpha > 0$, pequeno

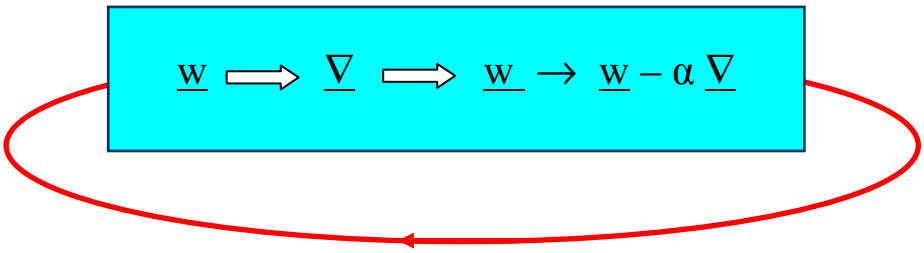
controla
direção $\Delta\underline{w}$

**Descida continuada, permanente - decisão local**

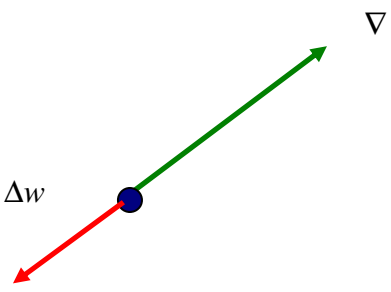


## Método do Gradiente Descendente ou

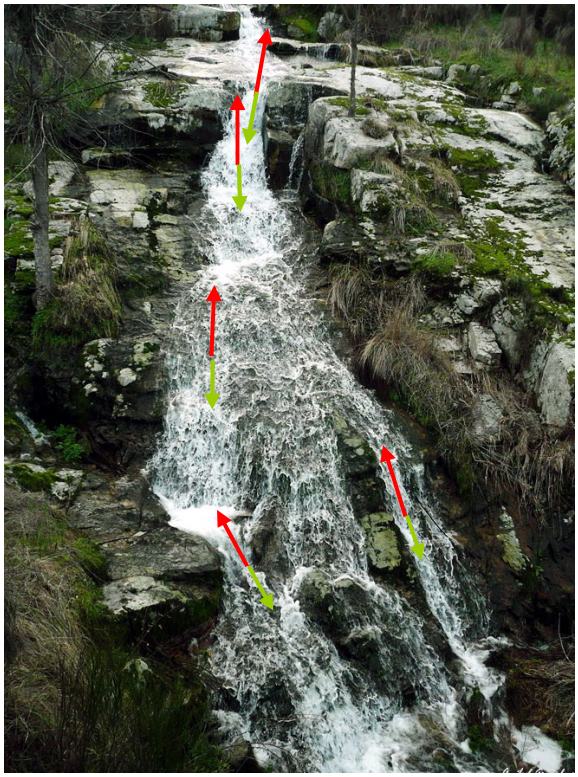
## Método da Descida pelo Gradiente




$$\Delta \underline{w} = - \alpha \underline{\nabla}$$



## Gradiente Descendente



## Algorítmo

Até que o critério de parada seja satisfeito

Para cada variável $w_i$

Calcular $\frac{\partial F}{\partial w_i}$

Fazer $w_{i\,novo} = w_{i\,antigo} - \alpha \frac{\partial F}{\partial w_i}$

Fim (natural) do processo

$\underline{w} = \underline{w}_0$

$E(\Delta\underline{w}) = \underline{0} \quad \Rightarrow \quad \nabla(\underline{w}_0) = \underline{0}$

Passo de Treinamento $\alpha$

pequeno (decisão local)

valores típicos para RNs $.01 < \alpha < .1$

**Em nosso caso, a função à minimizar é**

**Dispersão total intra classe para todas as classes:**

$$F_{in} = \sum_{\forall C_j} F_j$$

$$F_j = \sum_{\forall \vec{x} \in C_j} \left| \vec{x} - \vec{m}_j \right|^2 = \sum_{\forall \vec{x} \in C_j} \sum_{l=1}^{n} \left( x_l - m_{jl} \right)^2$$

$$\frac{\partial F_{in}}{\partial m_{jl}} = -2 \sum_{\forall \vec{x} \in C_j} \left( x_l - m_{jl} \right)$$

## Solução analítica

$$\nabla F_{in} = \frac{\partial F_{in}}{\partial m_{jl}} \Bigg] = -2 \sum_{\forall \vec{x} \in C_j} \left( x_l - m_{jl} \right) \Bigg] = 0]$$

$$\sum_{\forall \vec{x} \in C_j} \left( x_l - m_{jl} \right) = 0 \qquad \sum_{\forall \vec{x} \in C_j} x_l - n_j m_{jl} = 0 \qquad m_{jl} = \frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} x_l = \underset{\forall \vec{x} \in C_j}{E} x_l$$

$$\vec{m}_j = \frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} \vec{x} = \underset{\forall \vec{x} \in C_j}{E} \vec{x}$$

## Solução Numérica Recursiva

$$\Delta m_{ij} = -\alpha \frac{\partial F_{in}}{\partial m_{jl}} = 2\alpha \sum_{\forall \vec{x} \in C_j} \left( x_l - m_{jl} \right)$$

$$\Delta \vec{m}_j = 2\alpha \sum_{\forall \vec{x} \in C_j} \left( \vec{x} - \vec{m}_j \right)$$

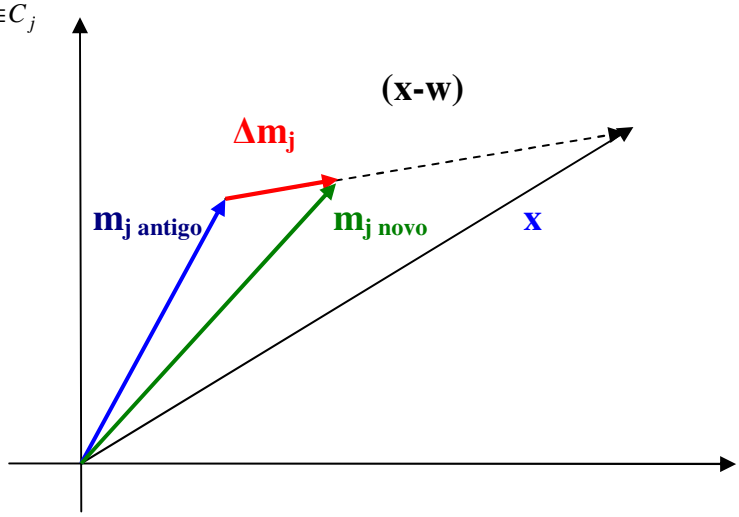
Este é um processo por batelada.

Alternativa: Calcular o acréscimo devido à cada entrada e aplicar imediatamente (processo por “regra delta”). Plástico e reduz a quantidade de memória necessária.

## Processo por "regra delta"

$$\Delta m_{ij} = -\alpha \frac{\partial F_{in}}{\partial m_{jl}} = 2\alpha \left( x_l - m_{jl} \right) \Big|_{\forall \vec{x} \in C_j}$$

$$\Delta \vec{m}_j = 2\alpha \left( \vec{x} - \vec{m}_j \right) \Big|_{\forall \vec{x} \in C_j}$$




## Fim do processo recursivo:

$$E(\Delta \vec{m}_j) = 2\alpha E\left( \vec{x} - \vec{m}_j \right) \Big|_{\forall \vec{x} \in C_j} = \vec{0}$$

$$\vec{m}_j = \frac{1}{n_j} \sum_{\forall \vec{x} \in C_j} \vec{x} = \underset{\forall \vec{x} \in C_j}{E} \vec{x}$$