CPE 723 - OTIMIZAÇÃO NATURAL

3 módulos:

OTIMIZAÇÃO CONTÍNUA - Calôba

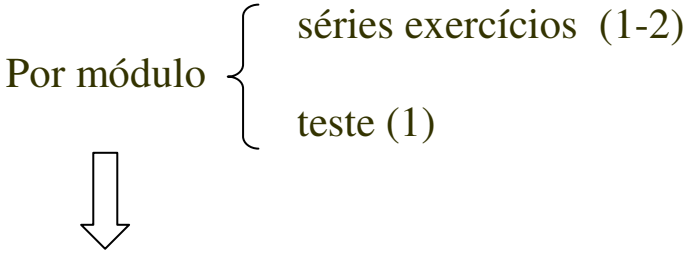
RECOZIMENTO SIMULADO - Gabriel

ALGORÍTMOS GENÉTICOS - Alexandre

Objetivos

Pré-requisitos

Avaliação



Módulo I - Otimização Contínua

Parte 1 – Métodos de Otimização

Aproximação local (Taylor), Gradiente ∇, Hessiana $\underline{H}$

Extremos, mínimo, máximo

Propriedades ∇ e $\underline{H}$

Métodos analíticos

Métodos numéricos recursivos

Gradiente descendente, passo, passo variável, ótimo

Métodos de 2ª. Ordem: Gradiente conjugado, Newton,
Newton amortecido, Quasi Newton

Módulo I - Otimização Contínua

Parte 2 – Treinamento como um Processo de Otimização

Erro

Treinamento

Batelada,
Lotes,
Regra Delta, Momento
Outros métodos

Parte 3 - Comentários Finais

Outras funções objetivo

Acompanhamento do treinamento

Crítica pós treinamento

Referências bibliográficas principais:

Cichocki, Unbehauen – "Neural Networks for Optimization and Signal Processing", Cap. 1 e 3, Wiley, 1993.

Chong, Zak – "An Introduction to Optimization", Part II, Wiley, 2001.

Addy, Dempster – "Introduction to optimization methods", Chapman and Hall, 1974.

www.lps.ufrj.br/~caloba/cpe723

cpe723_slides.pdf

cpe723_texto.pdf

cpe723_exercícios.pdf

cichocki&unbehauen.pdf

Neural Networks for Optimization and Signal Processing

Andrzej Cichocki
Warsaw University of Technology
Poland

Rolf Unbehauen
Universität Erlangen-Nürnberg
Germany

Parte 1 – Métodos de Otimização Contínua "clássica"

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

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

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

$$\frac{\partial^2 F}{\partial w_i \, \partial w_j} \quad \exists \quad \forall\, w_i, w_j, \ \forall\, \underline{w}$$

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

$$F_0(\underline{w}^*) \leq F_0(\underline{w}) \qquad \forall\ \underline{w} \neq \underline{w}^* \quad ?$$

Otimização de F_0

não linear

não convexa

$\underline{w}$ ordem elevada

irrestrita

contínua, derivável

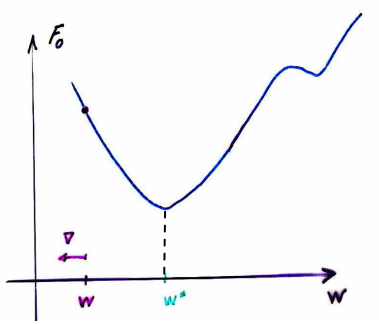
Como fazer ???

Noções Elementares de Otimização

I – Função objetivo – a minimizar

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

$\frac{\partial F}{\partial w_i}$ existe $\forall$ w_i



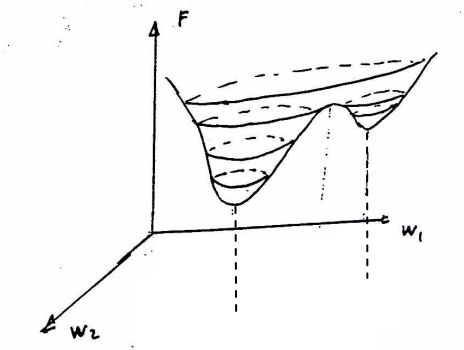
Mínimo, minimante $\underline{w}^*$

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

$\rightarrow$ mínimo local

Mínimo global

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



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

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

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

considerando

$$\underline{\Delta w} = \Delta w_i \,] \qquad \nabla F(\underline{w}_0) = \left.\frac{\partial F}{\partial w_i}\right|_{\underline{w}_0} \,] \qquad \textbf{gradiente}$$

$$\underline{H}(\underline{w}_0) = \left[\left.\frac{\partial^2 F}{\partial w_i \, \partial w_j}\right|_{\underline{w}_0}\right] \qquad \textbf{Hessiana} \text{ (simétrica)}$$

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

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

Então

$$F(\underline{w}_0 + \Delta\underline{w}) = F(\underline{w}_0) + \Delta\underline{w}^t \nabla F(\underline{w}_0) + \frac{1}{2} \Delta\underline{w}^t \underline{H}(\underline{w}_0) \Delta\underline{w} + \cdots$$

Aproximação linear ou de 1ª. ordem

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

Obs: Se F($\underline{w}$) é linear a aproximação de 1ª. ordem é exata. $\underline{\nabla} F(\underline{w}) = \text{ctte}(\underline{w})$ e as derivadas de ordem > 1 são nulas, i.e, a Hessiana $\underline{H}(\underline{w})$ é nula para todo $\underline{w}$.

Aproximação quadrática ou de 2ª. ordem

$$F(\underline{w}_0 + \Delta w) \simeq F(\underline{w}_0) + \Delta\underline{w}^{\,t} \; \underline{\nabla} \; F(\underline{w}_0) + \frac{1}{2} \Delta\underline{w}^t \; \underline{H}(\underline{w}_0) \Delta\underline{w}$$

Obs: Se F($\underline{w}$) é quadrática a aproximação de 2ª. ordem é exata. $\underline{H}(\underline{w}) = \text{ctte}(\underline{w})$ e as derivadas de ordem > 2 são nulas

Aproximação do Gradiente

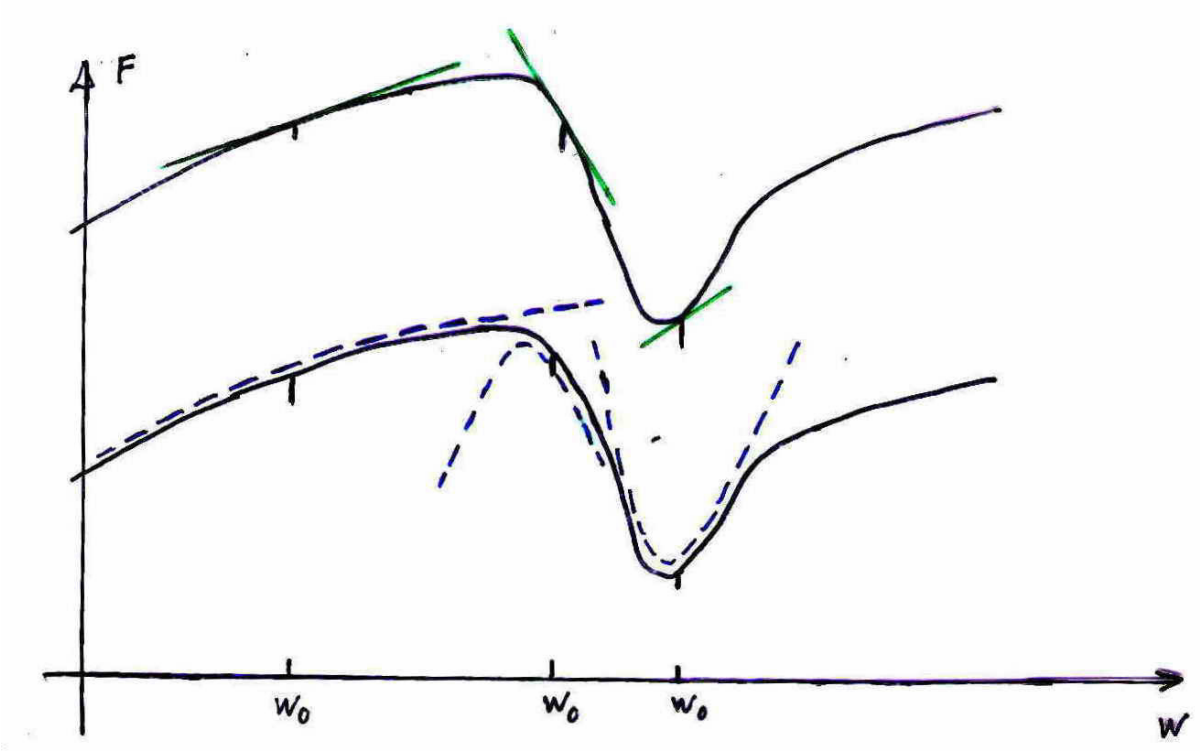
$$F(\underline{w}_0 + \Delta\underline{w}) \cong F(\underline{w}_0) + \sum_i \left.\frac{\partial F}{\partial w_i}\right|_{\underline{w}_0} \Delta w_i$$

$$\left.\frac{\partial F}{\partial w_i}\right|_{\underline{w}_0 + \Delta\underline{w}} \cong \left.\frac{\partial F}{\partial w_i}\right|_{\underline{w}_0} + \sum_j \frac{\partial}{\partial w_j} \left.\frac{\partial F}{\partial w_i}\right|_{\underline{w}_0} \Delta w_j$$

$$\underline{\nabla} F(\underline{w}_0 + \Delta\underline{w}) \cong \underline{\nabla} F(\underline{w}_0) + \underline{H}(\underline{w}_0)\,\Delta\underline{w}$$

Funções de ordem mais alta

validade das aproximações ?



A validade da aproximação

depende do acréscimo na função,

e não do acréscino nas variáveis independentes!

e.g., a aproximação de primeira ordem

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

é válida enquanto

$$\left| \frac{1}{2} \underline{\Delta w}^t \, \underline{H}(\underline{w}_0) \underline{\Delta w} \right| \ll \left| \underline{\Delta w}^t \, \underline{\nabla} F(\underline{w}_0) \right|$$

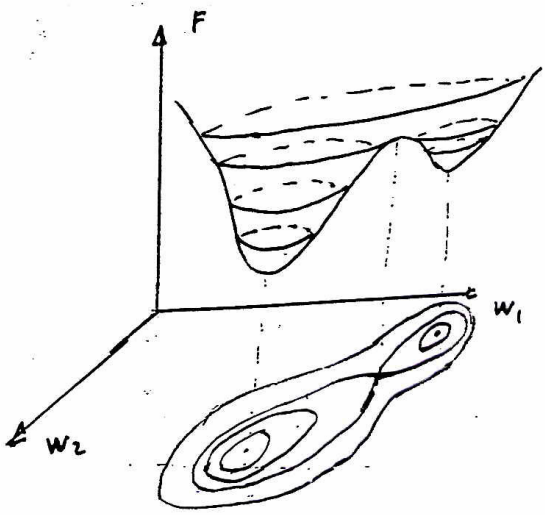
ou, alternativamente, enquanto

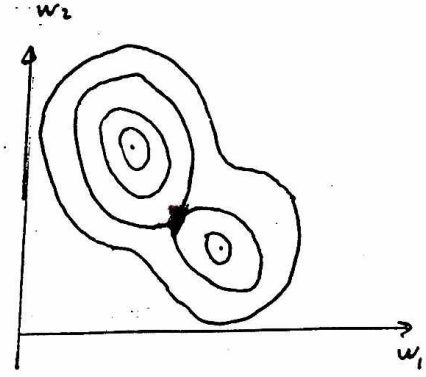
$$\left| \frac{1}{2} \underline{\Delta w}^t \, \underline{H}(\underline{w}_0) \underline{\Delta w} \right| \ll \varepsilon$$

Funções de duas variáveis

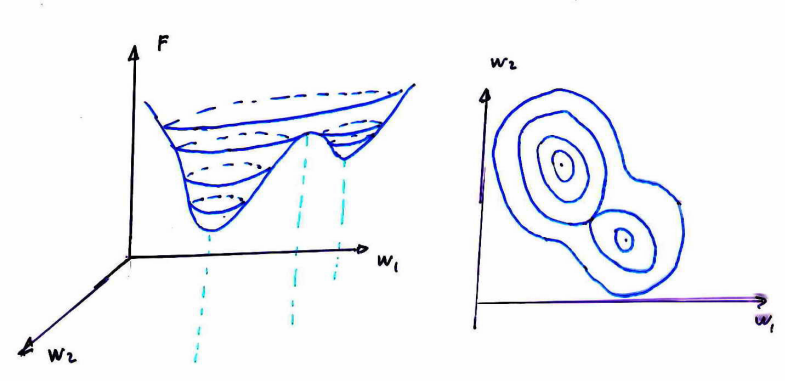
Visualização por

Curvas de Nível





Curvas / Superfícies de nível

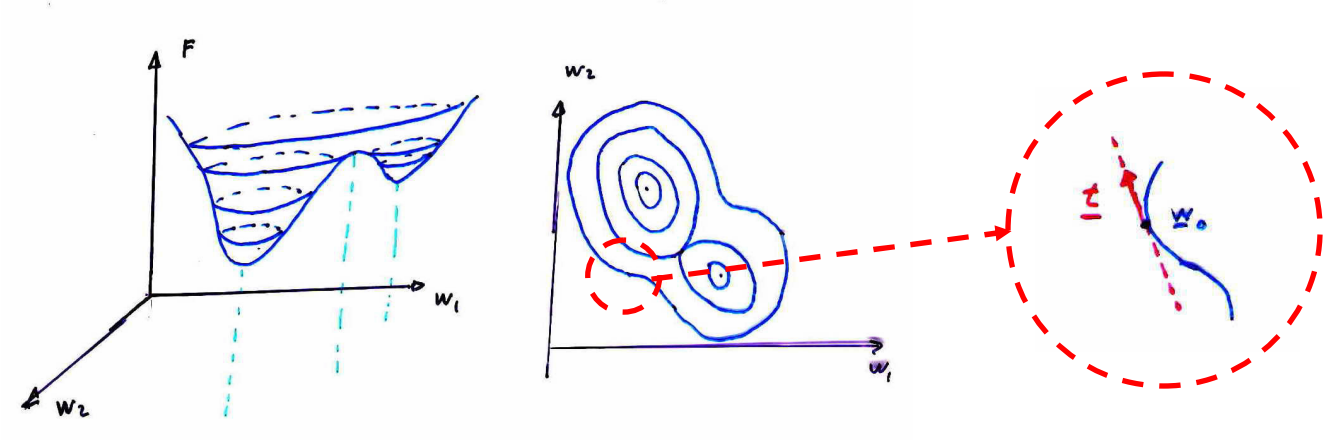


$\underline{w}$ $(\text{dim } 2)$ $\rightarrow$ curva de nível $(\text{dim } 2)$ $\rightarrow$ reta t_g $(\text{dim } 2)$

$\underline{w}$ $(\text{dim } 3)$ $\rightarrow$ superfície de nível $(\text{dim } 3)$ $\rightarrow$ plano t_g $(\text{dim } 3)$

$\underline{w}$ $(\text{dim } n)$ $\rightarrow$ superfície de nível $(\text{dim } n)$ $\rightarrow$ hiperplano t_g $(\text{dim } n)$

Propriedade:



$$\left.\frac{dF}{d\underline{w}}\right|_{\underline{t}} = 0 \qquad \left.\frac{d}{d\alpha} F(\underline{w}_0 + \alpha\, \underline{t})\right|_{\alpha=0} = 0$$

III – Extremos:

Mínimo, minimante $\underline{w}^*$

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

$$F(\underline{w}^* + \Delta\underline{w}) \cong F(\underline{w}^*) + \Delta w^t\, \nabla F + \frac{1}{2}\Delta\underline{w}^t\, \underline{H}\, \Delta\underline{w}$$

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

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

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

$$\Delta\underline{w}^t\ \underline{H}\ \Delta\underline{w} > 0$$

$$\Rightarrow \quad \underline{H}(\underline{w}^*) \quad \text{definida positiva}$$

Nos extremos

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

$$F(\underline{w}^{*} + \underline{\Delta w}) = F(\underline{w}^{*}) + \Delta F$$

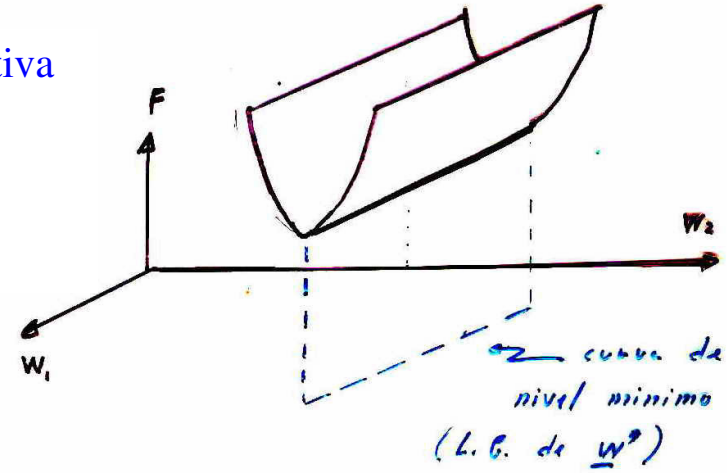
$$= F(\underline{w}^{*}) + \frac{1}{2}\underline{\Delta w}^{t}\,\underline{H}\,\underline{\Delta w}$$

$$\Delta F = \frac{1}{2}\underline{\Delta w}^{t}\,\underline{H}\,\underline{\Delta w} \begin{cases} >0 & \underline{H} \text{ definida positiva} \\ \geq 0 & \underline{H} \text{ semi-definida positiva} \\ \leq 0 & \underline{H} \text{ semi-definida negativa} \\ <0 & \underline{H} \text{ definida negativa} \\ \pm & \underline{H} \text{ não definida} \end{cases}$$

H semidefinida positiva

$\Delta F \geq 0$

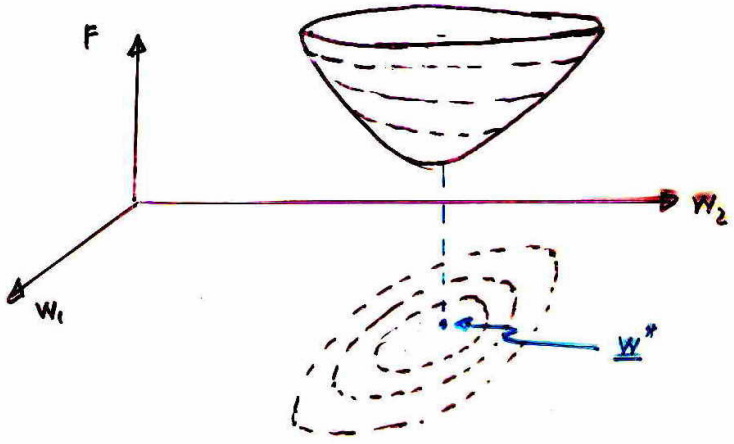
CALHA



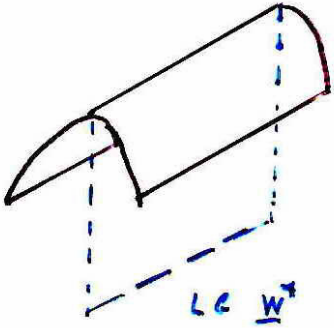
H definida positiva

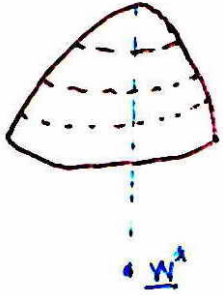
$\Delta F > 0$

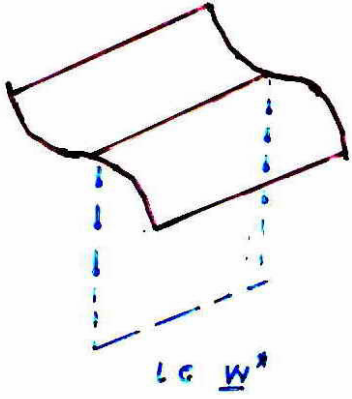
MÍNIMO
propriamente dito



H semidefinida negativa	H definida negativa	H não definida
$\Delta F \leq 0$	$\Delta F < 0$	$\Delta F \geq 0$ ou $\Delta F < 0$
CUMEEIRA	MÁXIMO	SELA







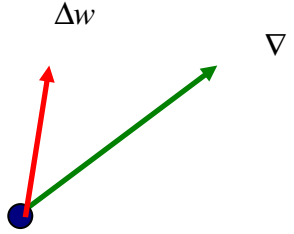
IV – Propriedades do Gradiente

1) Para $|\Delta \underline{w}| < \varepsilon$ e constante o gradiente indica a direção de máximo ΔF

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

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

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



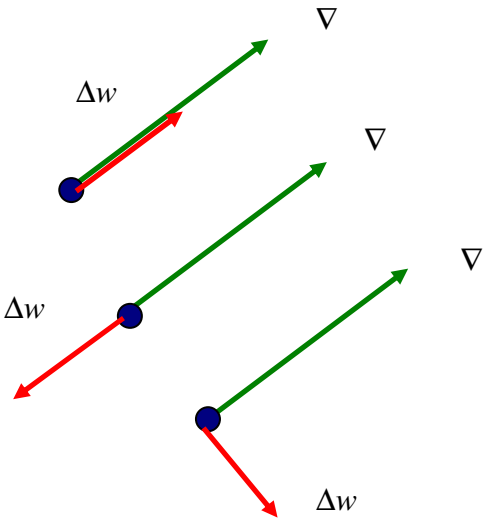
$$\Delta F \cong \underline{\Delta w}^t \nabla F = |\underline{\Delta w}| \; |\underline{\nabla} F| \cos\theta$$

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

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

$(-\Delta F$ max)

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



2) O gradiente é ortogonal ao plano tangente à superfície de nível no ponto

$$\underline{\Delta w} \in \pi_{tg} \qquad \underline{\Delta w} = \alpha \underline{t}$$

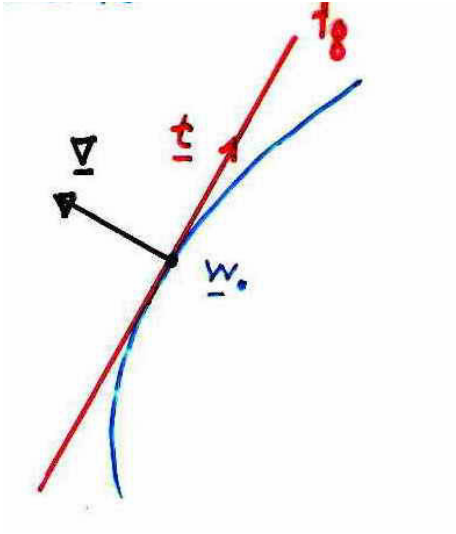
$$\frac{\mathrm{d}}{\mathrm{d}\alpha} F(\underline{w}_0 + \alpha \underline{t})\Big|_{\forall \underline{t} \in \pi_{\in}} = 0$$

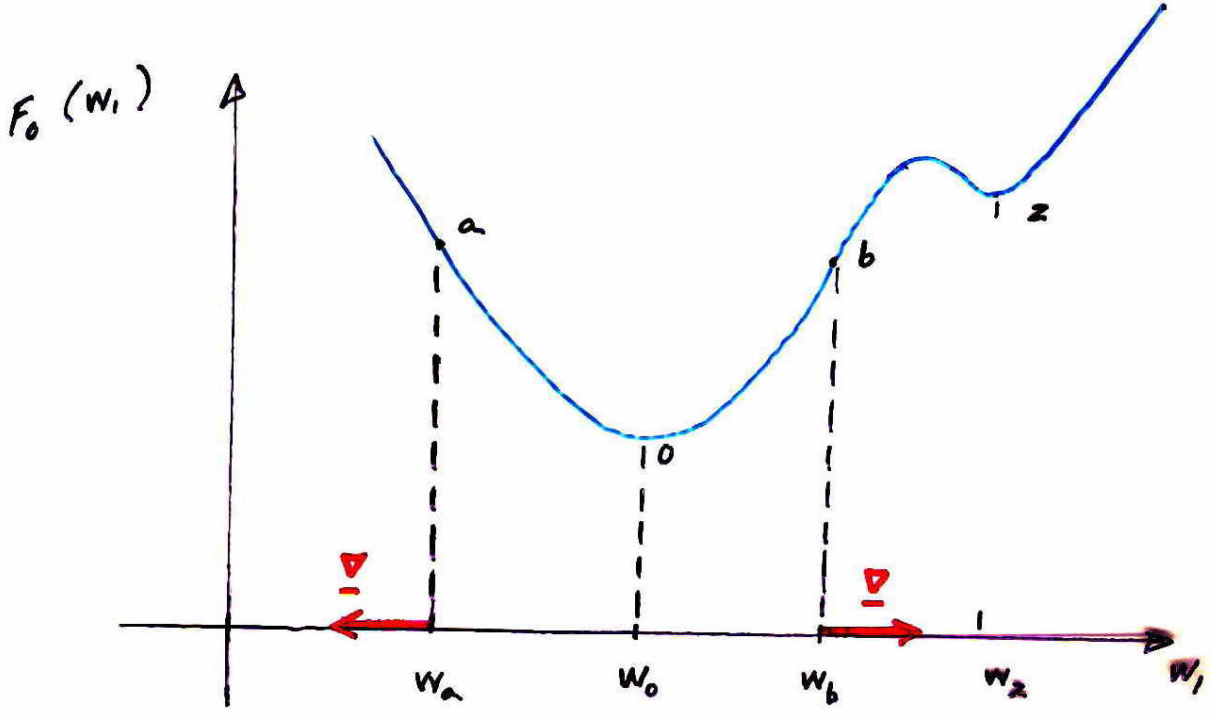
$$\Delta F \cong \underline{\Delta w}^t \, \underline{\nabla} F$$

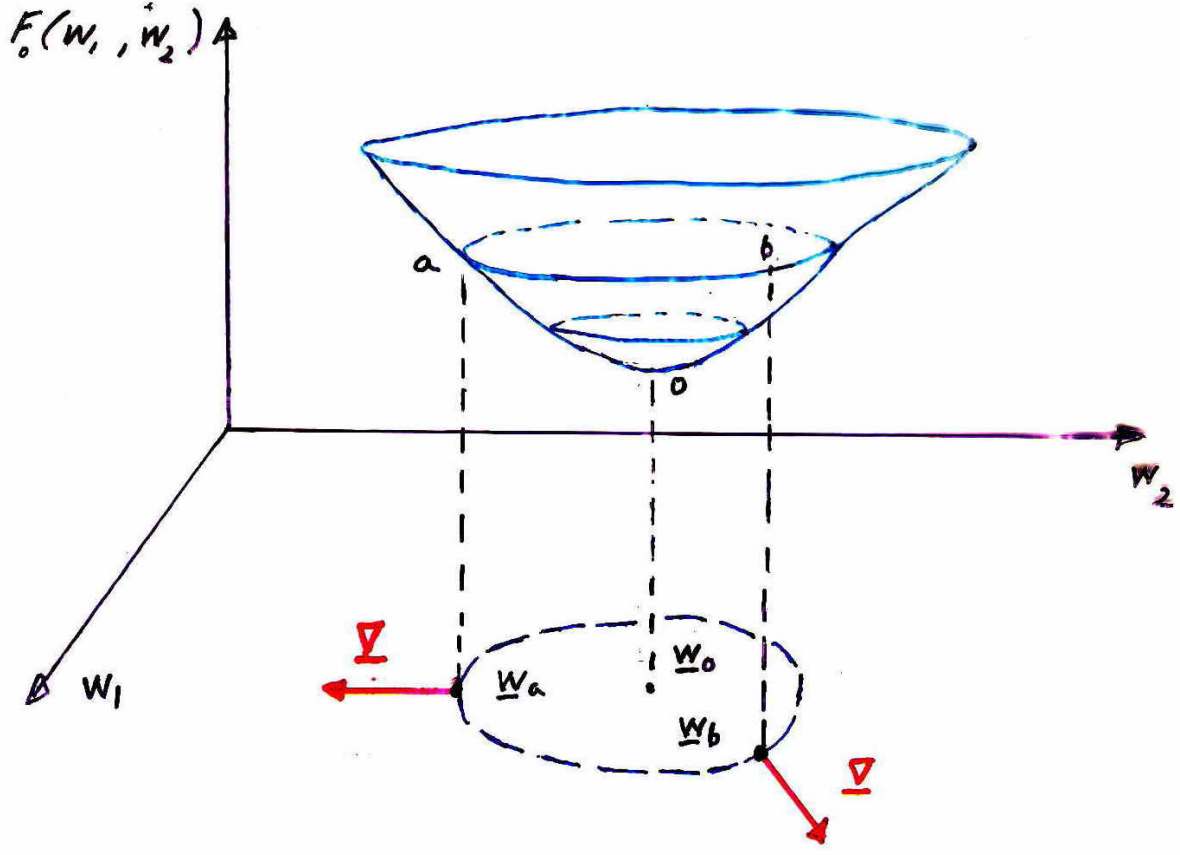
$$\Delta F = 0 \quad \rightarrow \quad \underline{\Delta w} \perp \underline{\nabla F}$$

$$\underline{\Delta w} \in \pi_{\perp \nabla}$$

$$\pi_t \equiv \pi_{\perp \nabla}$$




$F_0(w_1)$
a
b
z
0
1
w_a
w_0
w_b
w_2
w_1


$F_o(w_1, w_2)$
a
b
o
w_2
w_1
$\underline{v}$
$\underline{w}_a$
$\underline{w}_o$
$\underline{w}_b$
$\underline{v}$

V. Otimização

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

VI – Otimização: Solução Analítica

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

$$\underline{H}(\underline{w}^*) \text{ definida positiva}$$

A - Funções Lineares

$F(\underline{w}) = \underline{a}^t \underline{w} + b$ hiper-plano

$\nabla F = \underline{a} = \text{constante}$

$\nabla F = 0$? solução $\underline{w}^* \longrightarrow \infty$

B – Funções Quadráticas

$$F(\underline{w}) = \underline{w}^t \, \underline{A} \, \underline{w} \; + \; \underline{b}^t \, \underline{w} \; + \; c$$

$$F(\underline{w}_0 + \underline{\Delta w}) = F(\underline{w}_0) + \underline{\Delta w}^t \, \underline{\nabla}(\underline{w}_0) + \frac{1}{2} \underline{\Delta w}^t \, \underline{H}(\underline{w}_0) \, \underline{\Delta w}$$

$$\underline{\nabla}(\underline{w}_0 + \underline{\Delta w}) = \underline{\nabla}(\underline{w}_0) + \underline{H}(\underline{w}_0) \, \underline{\Delta w}$$

$$\underline{\nabla} = \underline{0} \qquad \Rightarrow \qquad \underline{\Delta w} = - \, \underline{H}^{-1}(\underline{w}_0) \, \underline{\nabla}(\underline{w}_0)$$

$$\mathbf{\underline{w}^* = \underline{w}_0 - H^{-1}(\underline{w}_0) \, \nabla(\underline{w}_0)}$$

Newton

Newton Raphson

Problema: $\underline{H}$, e principalmente $\underline{H}^{-1}$

Evitando $\underline{H}$

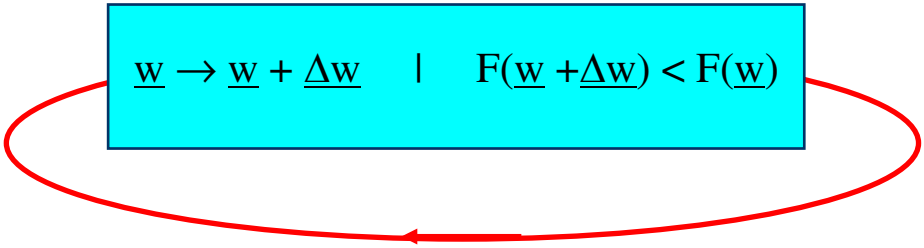
Pseudo Newton $\quad \underline{H} \approx \text{diag}\, \underline{H} \Longrightarrow \underline{H}^{-1}$

Quase Newton $\quad \sim \underline{H}^{-1}$

C – Funções de ordens mais elevadas e/ou transcendentes

muito complexo $\Longrightarrow$ Métodos Numéricos Recursivos

VII – Otimização: Métodos Numéricos Recursivos



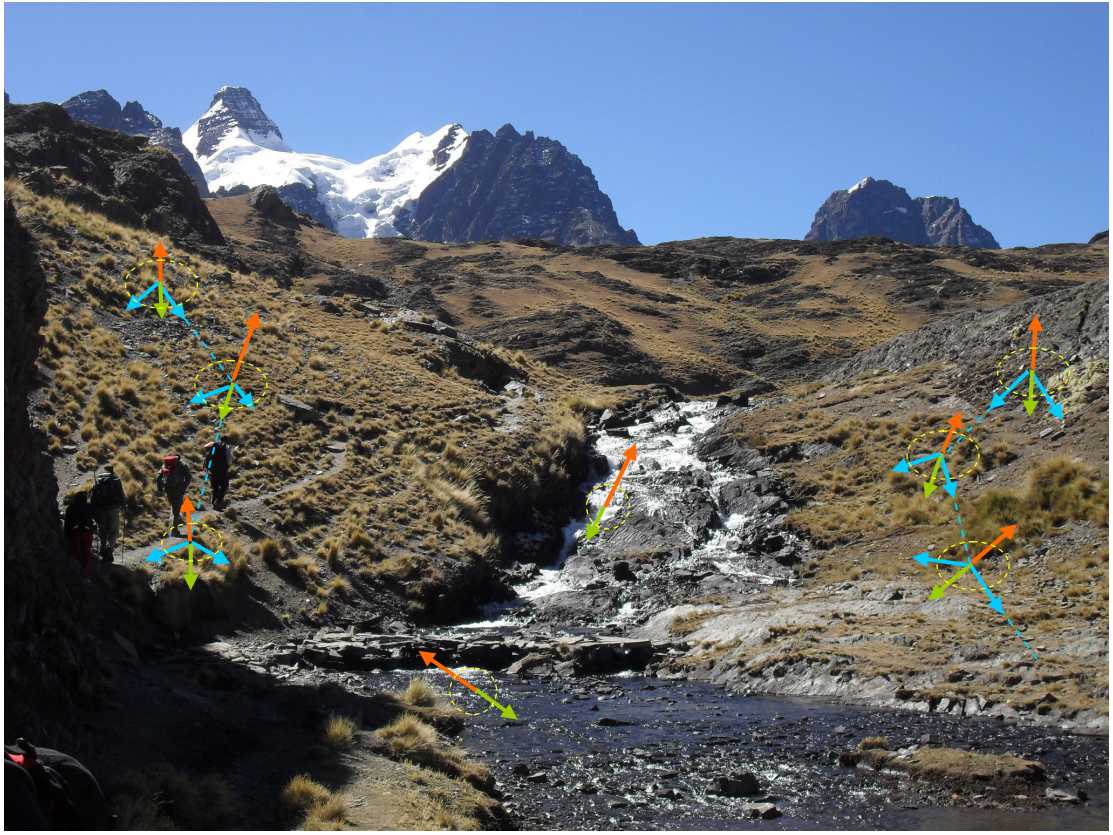
$\underline{\Delta w}$??

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

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

controla
direção $\underline{\Delta w}$

Descida continuada, permanente - decisão local



Condições necessárias

α pequeno, vale a aproximação de primeira ordem

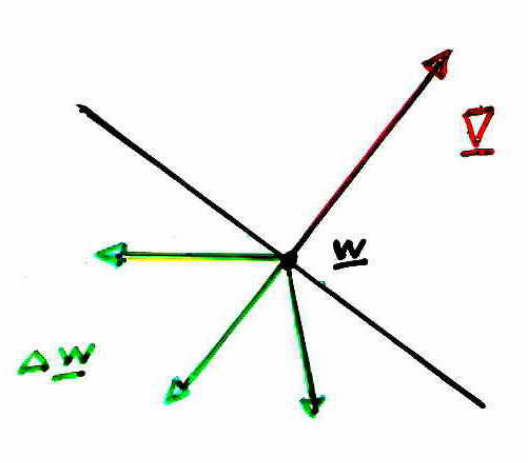
$$F(\underline{w} + \Delta\underline{w}) \approx F(\underline{w}) + \Delta\underline{w}^t \nabla(\underline{w})$$

$$\Delta F \approx \Delta\underline{w}^t \nabla(\underline{w})$$

$$= \alpha \, | \, \underline{d} \, | \, |\nabla(\underline{w})| \cos \angle \, \underline{d}, \nabla$$

direção $\underline{d}$

$$\angle \, \underline{d}, \nabla \quad \in \quad (90^o, 270^o)$$



Direção mais eficaz ?

Para $|\Delta \underline{w}|$ constante

qual **$\underline{d}$** para máximo $-\Delta F$?

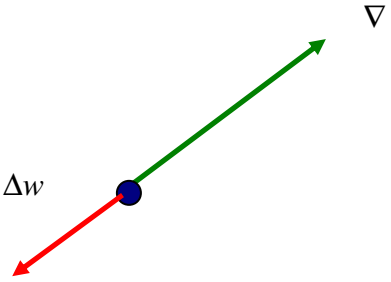
$$F(\underline{w} + \underline{\Delta w}) = F(\underline{w}) + \Delta \underline{w}^t \nabla$$

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

constante (-1, 1)

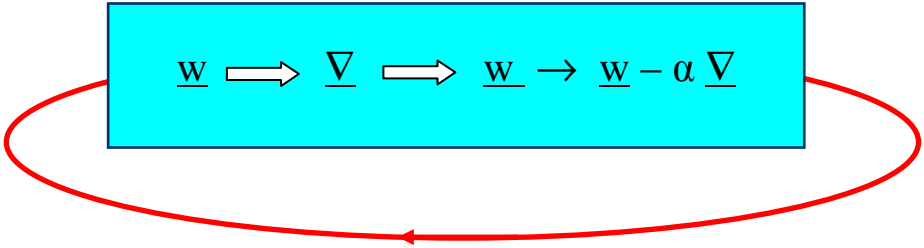
$$\underline{d} = - \underline{\nabla}$$

$$\Delta \underline{w} = - \alpha \underline{\nabla} \quad \Rightarrow \quad - \Delta F \text{ max}$$



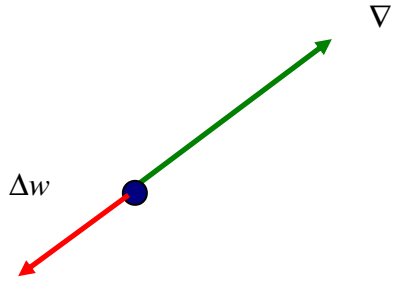
Método do Gradiente Descendente ou

Método da Descida pelo Gradiente

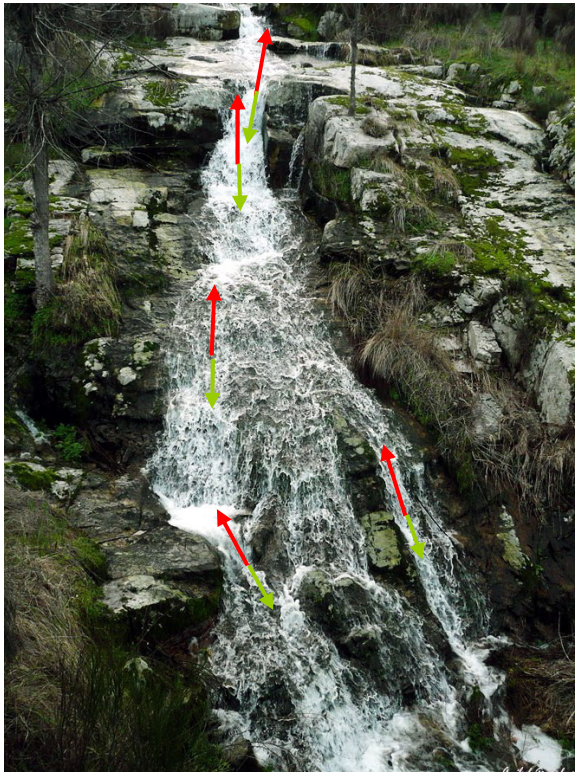


$$\Delta \underline{w} = - \alpha \, \underline{\nabla}$$

$$\Delta F = \Delta \underline{w}^t \, \nabla = - \alpha \, |\underline{\nabla}|^2$$

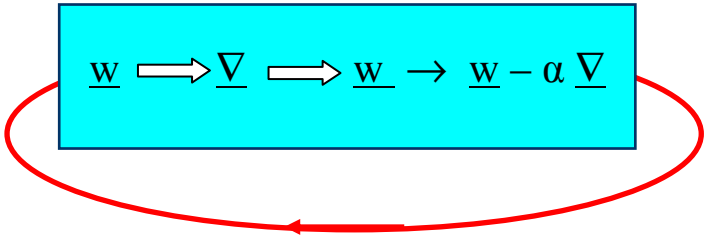


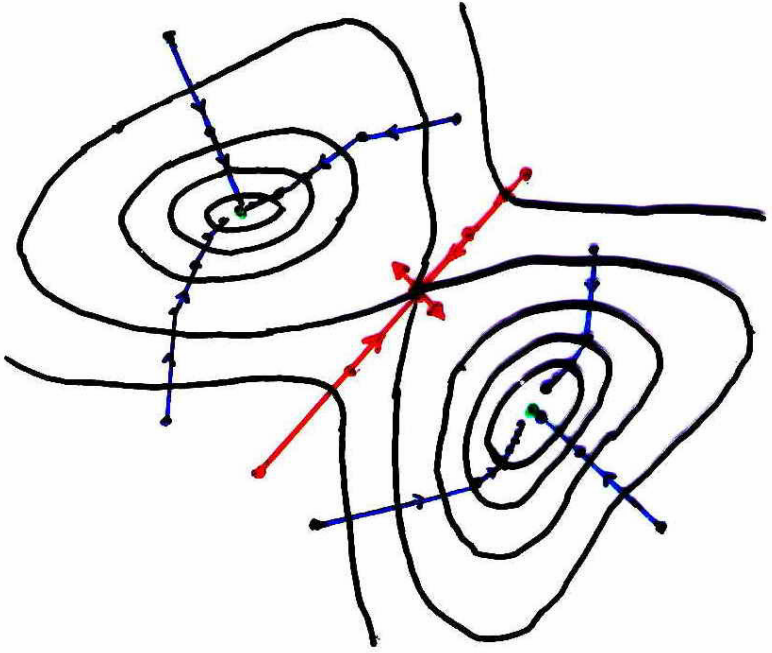
Gradiente Descendente



Método do

Gradiente Descendente





Algorítmo

Até que o critério de parada seja satisfeito

Para cada variável w_i

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

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

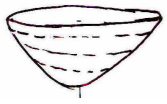
Fim (natural) do processo

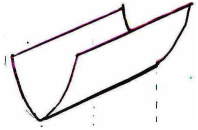
$\underline{w} = \underline{w}_0$

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

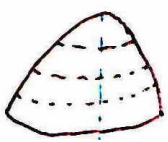
O que é $\underline{w}_0$??

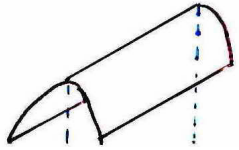
pontos estáveis: mínimo, calha

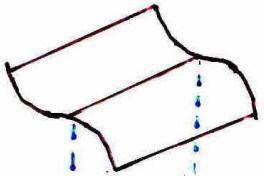




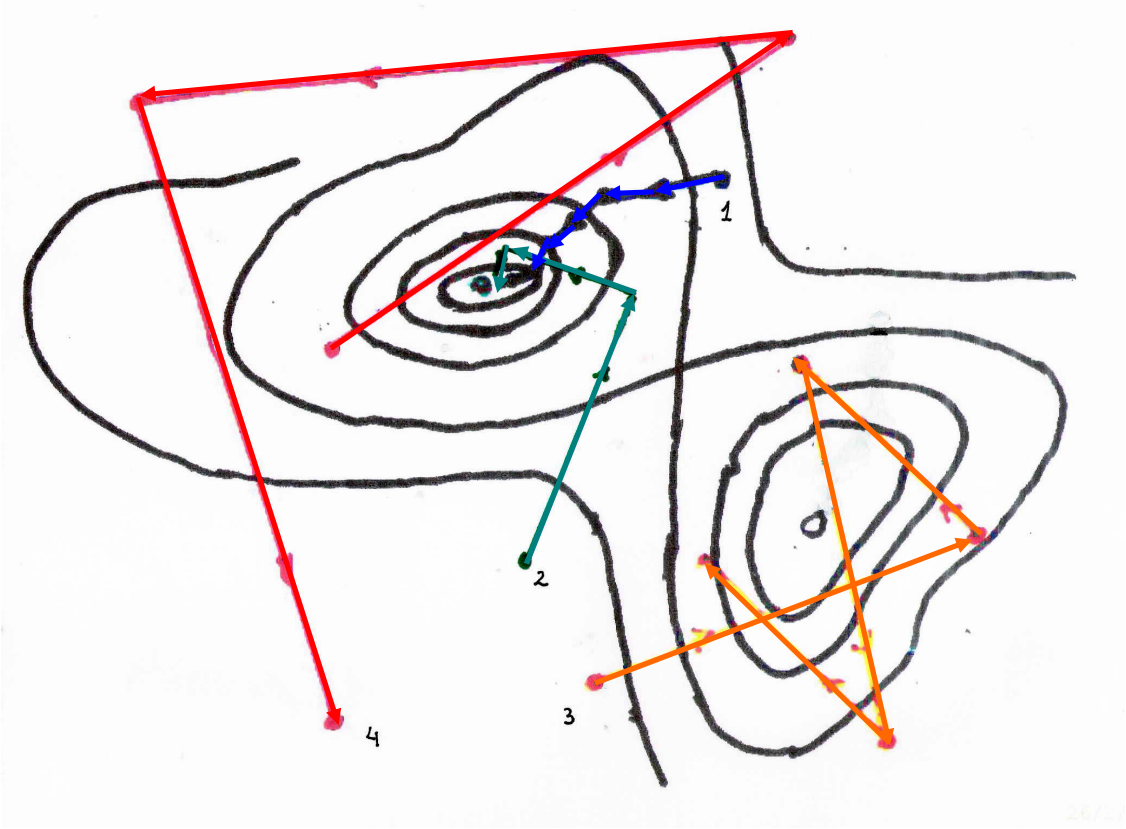
pontos instáveis: máximo, cumeeira, sela





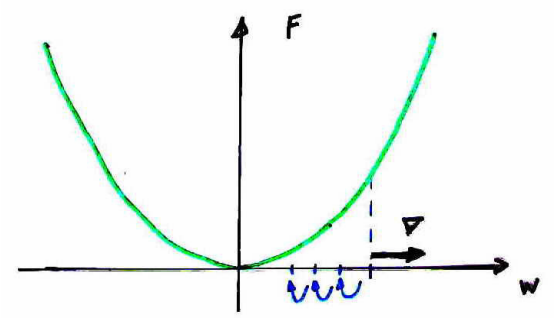


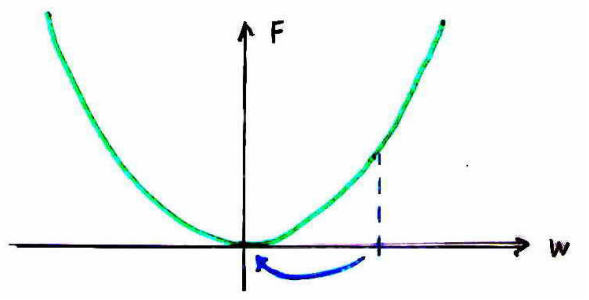
Passo de Treinamento α

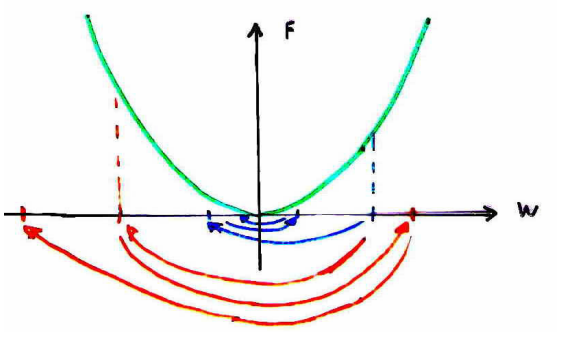


Efeito do passo de treinamento α

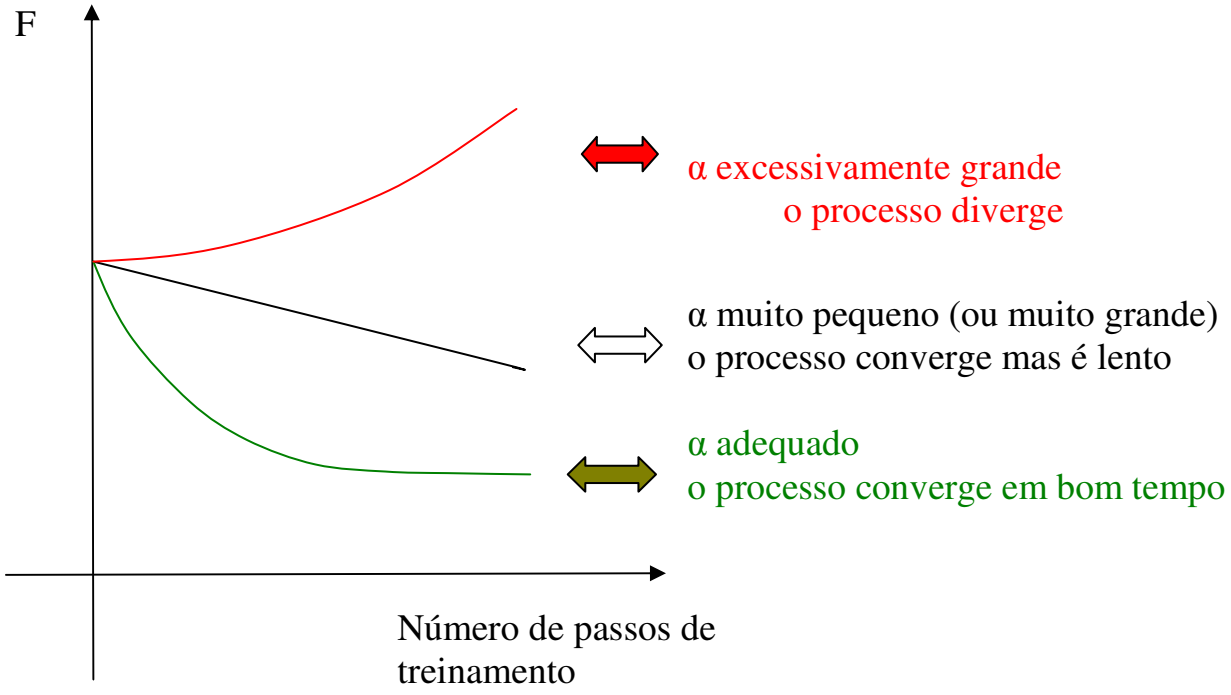
Exemplo simples: $\mathbf{F = w^2}$







Evolução do erro ao longo do treinamento:



Passo de Treinamento α

α << convergência lenta

α >> F_0 oscila muito, convergência lenta, **pode divergir**

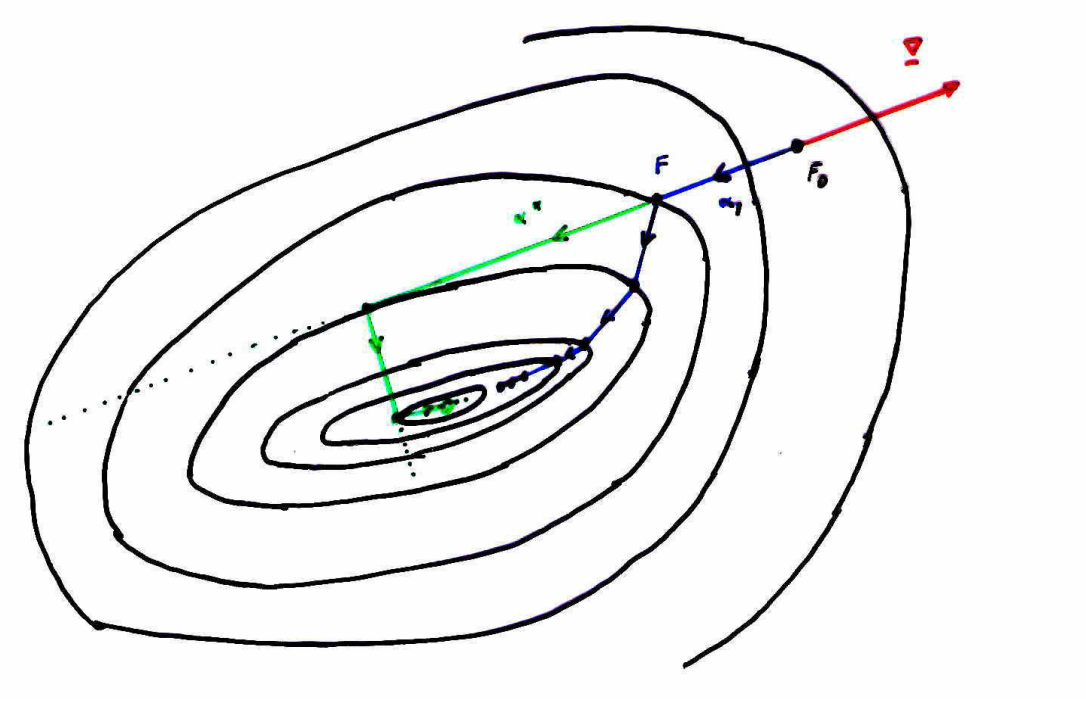
$\Rightarrow \alpha_{\text{ótimo}}$ **??**

"a escolha de α é uma arte" – Bernard Widrow

No treinamento de redes neurais

$0 < \alpha \leq 1$ $\quad \alpha_{\text{tipico}} = .05, \ .1$

Passo Ótimo de Treinamento α^*



Que $\alpha = \alpha^*$ maximiza $-\Delta F$?? **Otimização em linha.**

Passo ótimo α*

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

$$F(\underline{w}_0 + \Delta \underline{w}) \approx F(\underline{w}_0) + \nabla^t \Delta \underline{w} + \frac{1}{2} \Delta \underline{w}^t \underline{H} \Delta \underline{w}$$

$$F(\underline{w}_0 - \alpha \underline{\nabla}) \approx F(\underline{w}_0) - \alpha \underline{\nabla}^t \underline{\nabla} + \frac{1}{2} \alpha^2 \underline{\nabla}^t \underline{H} \underline{\nabla} \qquad (1)$$

$$\Delta F \approx - \alpha \underline{\nabla}^t \underline{\nabla} + \frac{1}{2} \alpha^2 \underline{\nabla}^t \underline{H} \underline{\Delta}$$

Que $\alpha = \alpha^*$ maximiza $- \Delta F$?? **Otimização em linha.**

$\alpha = \alpha^*$ $\Rightarrow$ **– ΔF** é máximo $\frac{\partial F}{\partial \alpha} = \frac{\partial \Delta F}{\partial \alpha} = 0$

$$\alpha^* = \underset{\alpha^*>0}{\text{Arg}} \left[\text{Min } F(\underline{w}_0 - \alpha^* \underline{\nabla})\right]$$

$$\frac{\partial}{\partial \alpha} F = - |\underline{\nabla}|^2 + \alpha \, \underline{\nabla}^t \, \underline{H} \, \underline{\Delta} = 0$$

$$\alpha^* = \frac{|\underline{\nabla}|^2}{\underline{\nabla}^t \, \underline{H} \, \underline{\nabla}} \qquad (2)$$

necessita de $\underline{H}$!

Contornando o problema de $\underline{H}$:

Aplicar $\alpha = \alpha_1 \quad \underline{w}_0 \rightarrow \underline{w}_0 - \alpha_1 \underline{\nabla} \quad F_0 \rightarrow F_1$

De (1) $\qquad F_1 = F_0 - \alpha_1 \underline{\nabla}^t \underline{\nabla} + \frac{1}{2} \alpha_1^2 \underline{\nabla}^t \underline{H} \underline{\nabla}$

$$\therefore \qquad \underline{\nabla}^t \underline{H} \underline{\nabla} = \frac{2}{\alpha_1^2} (F_1 - F_0 + \alpha_1 |\underline{\nabla}|^2) \qquad (3)$$

aplicando (3) em (2)

$$\alpha^* = \frac{|\underline{\nabla}|^2}{\underline{\nabla}^t \underline{H} \underline{\nabla}} \qquad (2)$$

$$\alpha^* = \frac{\alpha_1^2 \left|\underline{\nabla}\right|^2}{2\left(F_1 - F_0 + \alpha_1 \left|\underline{\nabla}\right|^2\right)} \qquad \text{e}$$

$$\Delta F = \frac{1}{2}\, \alpha^* |\underline{\nabla}|^2$$

Obs:

1 – α^* leva a extremos. Se $\alpha^* < 0$ o extremo é um máximo.

2 - o passo de ensaio, $-\alpha_1 \underline{\nabla}$, levanta as características da função no entorno de $\underline{w}_0$ com raio $|\alpha_1 \underline{\nabla}|$. Se $\alpha^* >> \alpha_1$ possivelmente o α^* encontrado não é válido.

Passo variável controlando o decréscimo de F

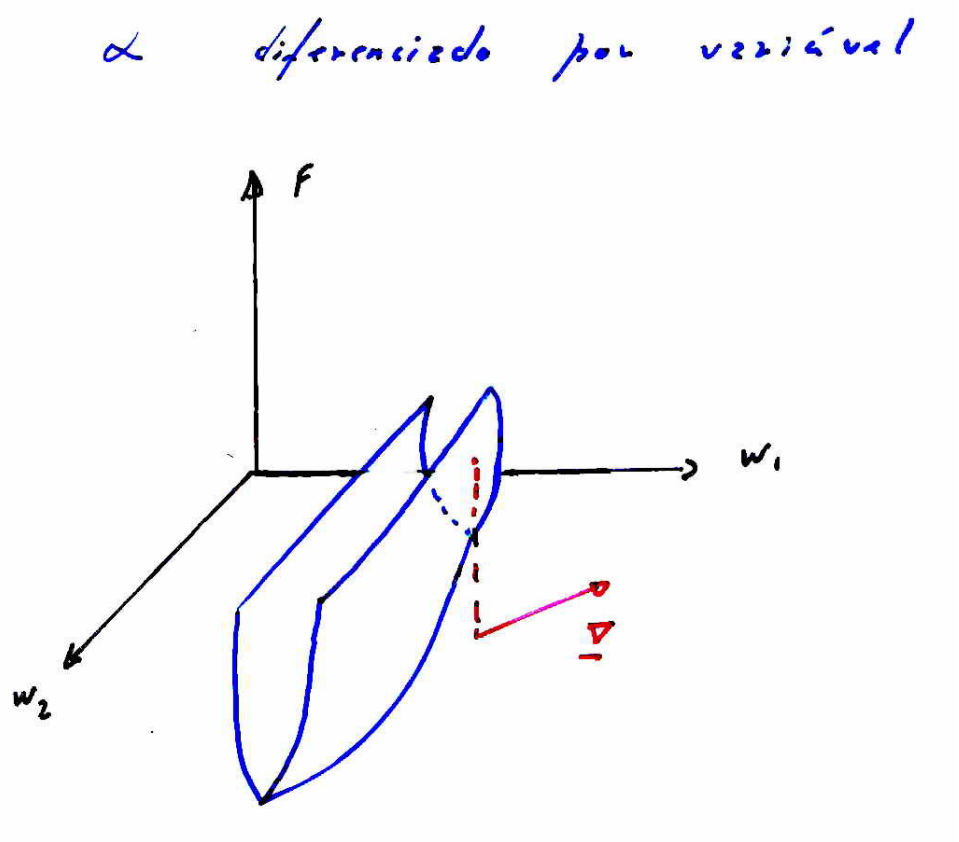
$$\Delta F \approx - \Delta \underline{w}^t \, \underline{\nabla} = - \alpha \left\| \underline{\nabla} \right\|^2$$

$$\alpha = \alpha_0 \quad \text{ctte} \quad \Rightarrow \qquad \Delta F = - \alpha \left\| \underline{\nabla} \right\|^2$$

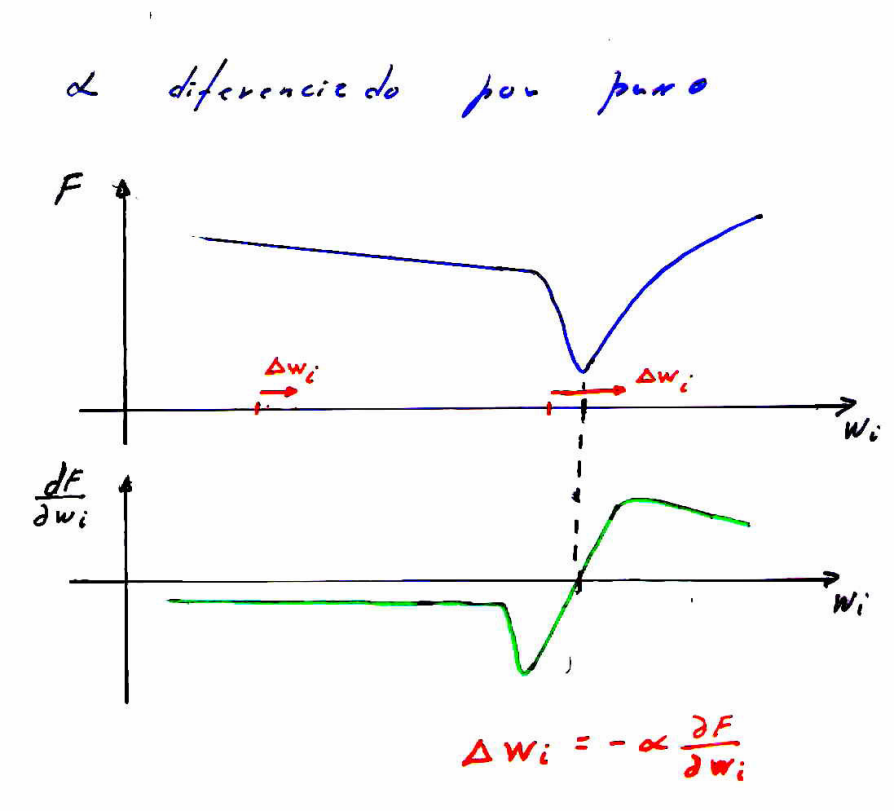
$$\alpha = \alpha_0 \frac{1}{\left\| \underline{\nabla} \right\|^2} \quad \Rightarrow \qquad \Delta F = - \alpha_0 \quad \text{ctte}$$

$$\alpha = \alpha_0 \frac{F}{\left\| \underline{\nabla} \right\|^2} \quad \Rightarrow \qquad \frac{\Delta F}{F} = - \alpha_0 \quad \text{ctte}$$

Acelerando o processo – passo variável



Acelerando o processo – passo variável



Resilient Backpropagation (modificado)

(Silva e Almeida, Super SAB, etc.)

α variável $\begin{cases} \text{por variável } w_i \\ \text{por passo de treinamento } n \end{cases}$

$$\Delta w_i(n) = - \alpha_i(n) \frac{\partial F}{\partial w_i}(n)$$

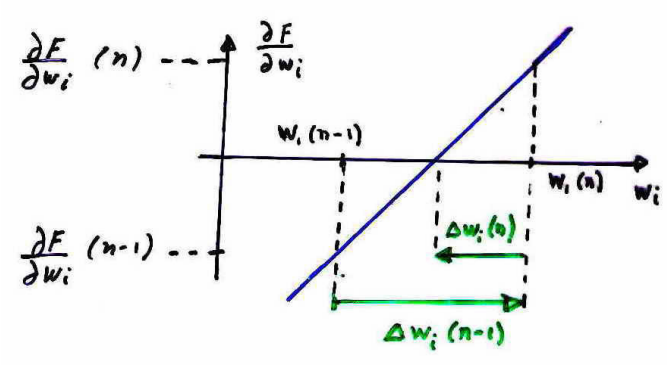
$$\alpha_i(n) = \begin{cases} a\, \alpha_i(n-1) & \text{se } \operatorname{sign}\frac{\partial F}{\partial w_i}(n) = \operatorname{sign}\frac{\partial F}{\partial w_i}(n-1) \\ b\, \alpha_i(n-1) & \text{se } \operatorname{sign}\frac{\partial F}{\partial w_i}(n) \neq \operatorname{sign}\frac{\partial F}{\partial w_i}(n-1) \end{cases}$$

se $\alpha_i(n) > \alpha^*$ faça $\alpha = \alpha^*$

$\alpha_i(0) \approx .05 \quad \alpha^* \approx .2 \quad a \approx 1.05 \quad b \approx .9$

Outra alternativa

$$b = \frac{\frac{\partial F}{\partial w_i}(n-1)}{\frac{\partial F}{\partial w_i}(n-1) - \frac{\partial F}{\partial w_i}(n)}$$

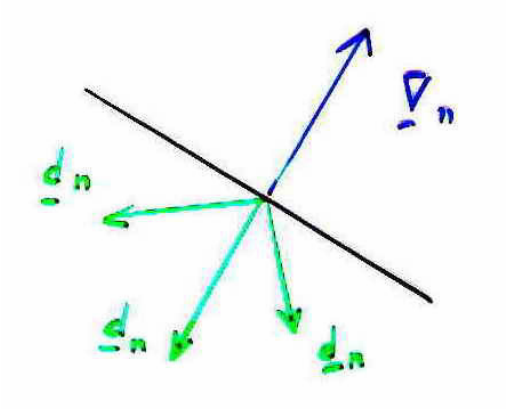


Métodos de 1ª. Ordem (Resumo)

$$\underline{w}_{n+1} = \underline{w}_n + \alpha_n \underline{d}_n$$

Para $F(\underline{w}_{n+1}) < F(\underline{w}_n)$

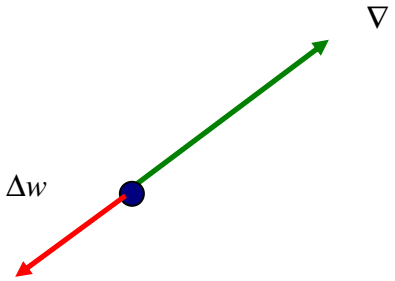
$$\underline{d}_n^t \ \underline{\nabla}_n < 0$$



Gradiente Descendente

$$\underline{w}_{n+1} = \underline{w}_n + \alpha_n \underline{d}_n$$

$$\underline{d}_n = -\underline{\nabla}_n \qquad \alpha_n = \alpha_0$$



Passo ótimo

$$\alpha_n = \underset{\alpha_n > 0}{\text{Arg}} \left[\text{Min } F(\underline{w}_n + \alpha_n \underline{d})\right]$$

BP Resiliente: α variável por sinapse e por passo

MÉTODOS DE SEGUNDA ORDEM:
MAIS SOFISTICADOS / MAIS EFICIENTES (?)

Newton	$\underline{\mathbf{H}}$
Newton Amortecido	$\approx \underline{\mathbf{H}}$
Quasi Newton	$\approx \underline{\mathbf{H}}^{-1}$
Gradiente Conjugado	

Newton

$$\underline{w}^{*} = \underline{w}_1 - \underline{H}_1^{-1}\ \underline{\nabla}_1$$

$$\underline{w}_{n+1} = \underline{w}_n + \alpha_n\ \underline{d}_n$$

$$\underline{d}_n = -\ \underline{H}_n^{-1}\ \underline{\nabla}_n$$

$$\alpha_n = \underset{\alpha_n>0}{\text{Arg}}\left[\text{Min } F(\underline{w}_n + \alpha_n\ \underline{d})\right]$$

$\underline{H}_n^{-1}$ existe ? é d.p. ?

Newton Amortecido

Marquadt – Levenberg

para garantir a inversão de $\underline{H}$

$$\underline{H}_n \rightarrow \underline{H}_n + \lambda \underline{I}$$

$$\lambda > 0 \qquad\qquad \lambda \rightarrow 0$$

Levenberg – Marquadt

$$F(\underline{w}) = \frac{1}{2}\sum_k \varepsilon_k^2$$

$$\underline{H}_n = [h_{ij}]$$

$$\Rightarrow \quad h_{ij} = \frac{\partial^2 F}{\partial w_i \, \partial w_j} \approx \sum_k \frac{\partial \varepsilon_k}{\partial w_i} \cdot \frac{\partial \varepsilon_k}{\partial w_j}$$

$$\Rightarrow \quad \underline{H}_n \rightarrow \underline{H}_n + \lambda \, \underline{I}$$

$$\lambda > 0 \qquad \lambda \rightarrow 0$$

$$\underline{w}_{n+1} = \underline{w}_n + \alpha_n \underline{d}_n$$

$$\underline{d}_n = - \left[\left[h_{ij}\right] + \lambda \underline{I}\right]^{-1} \underline{\nabla}_n$$

$$\alpha_n = \underset{\alpha_n > 0}{\text{Arg}} \left[\text{Min } F(\underline{w}_n + \alpha_n \underline{d})\right]$$

Quasi Newton

Fórmulas básicas:

$$\underline{\nabla}_{n+1} = \underline{\nabla}_n + \underline{H}_n \, (\underline{w}_{n+1} - \underline{w}_n)$$

onde $\underline{w}_{n+1} - \underline{w}_n = \Delta \underline{w}$

$$\underline{w}_{n+1} - \underline{w}_n = \underline{G}_n \, (\underline{\nabla}_{n+1} - \underline{\nabla}_n)$$

onde $\underline{G}_n \approx \underline{H}_n^{-1}$

BFGS – Broyden – Fletcher – Goldfarb – Shanno

$$\underline{w}_{n+1} = \underline{w}_n + \alpha_n \underline{d}_n$$

$$\underline{d}_n \approx \underline{w}_{n+1} - \underline{w}_n = - \underline{G}_n \underline{\nabla}_n$$

$$\alpha_n = \underset{\alpha_n>0}{\text{Arg}} \left[\text{Min } F(\underline{w}_n + \alpha_n \underline{d})\right]$$

$$\underline{y} = \underline{\nabla}_{n+1} - \underline{\nabla}_n$$

$$\underline{G}_{n+1} = \left[\underline{I} - \frac{\underline{d}_n \underline{y}_n^t}{\underline{d}^t \underline{y}_n}\right] \underline{G}_n \left[\underline{I} - \frac{\underline{y}_n \underline{d}_n^t}{\underline{d}_n^t \underline{y}_n}\right] + \frac{\underline{d}_n \underline{d}_n^t}{\underline{d}_n^t \underline{y}_n}$$

$$\underline{G}_n \approx \underline{H}_n^{-1}$$

Gradiente Conjugado - não usa H

Fletcher – Reeves

$\underline{d}_i$ – uma só vez

$$\underline{w}_{n+1} = \underline{w}_n + \alpha_n \underline{d}_n$$

$$\underline{d}_n = \beta_n \underline{d}_{n-1} - \underline{\nabla}_n$$

$$\alpha_n = \underset{\alpha_n>0}{\text{Arg}} \left[\text{Min } F(\underline{w}_n + \alpha_n \underline{d})\right]$$

Leonard & Kramer

$$\beta_n = \frac{|\underline{\nabla}_n|^2}{|\nabla_{n-1}|^2}$$

Polak – Ribière

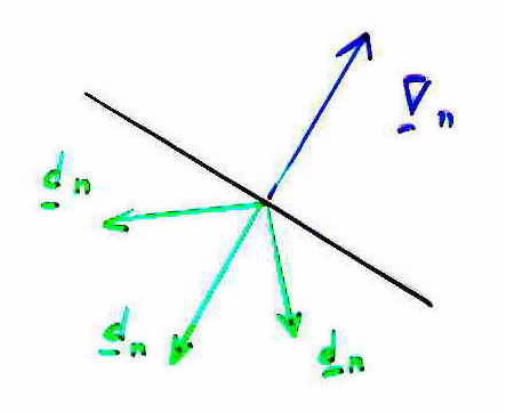
$$\beta_n = \frac{(\underline{\nabla}_n - \underline{\nabla}_{n-1})^t \underline{\nabla}_n}{|\underline{\nabla}_{n-1}|^2}$$

Métodos de 2ª. Ordem (Resumo)

$$\underline{w}_{n+1} = \underline{w}_n + \alpha_n \underline{d}_n$$

$$\alpha_n = \underset{\alpha_n > 0}{\text{Arg}} \left[\text{Min } F(\underline{w}_n + \alpha_n \underline{d}) \right]$$

$\underline{d}_n$???



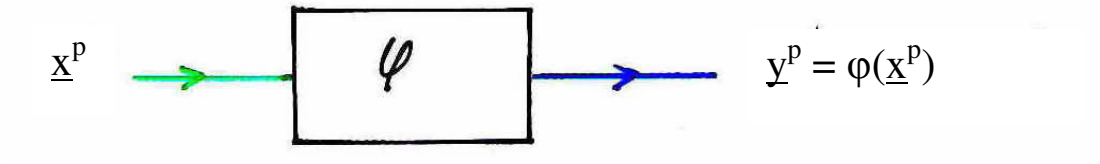
Parte II

Aprendizado

(ou treinamento, ou ajuste de parâmetros)

como um processo de otimização

Sistema à modelar / simular



Conhecimento do Sistema

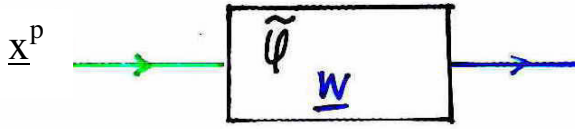
Fenomenológico – equações – caixa branca

Funcional – relação entrada/saída – caixa preta

Pares entrada-saída

$(\underline{x}^p, \underline{y}^p)$ $\quad p = 1, \ldots P$

Modelo Matemático Fenomenológico / Numérico - Simulador



$$\underline{\tilde{y}}^{p} = \tilde{\varphi}(\underline{x}^{p}) \approx \underline{y}^{p} \quad \forall \; p$$

Modelo Fenomenológico – Caixa branca

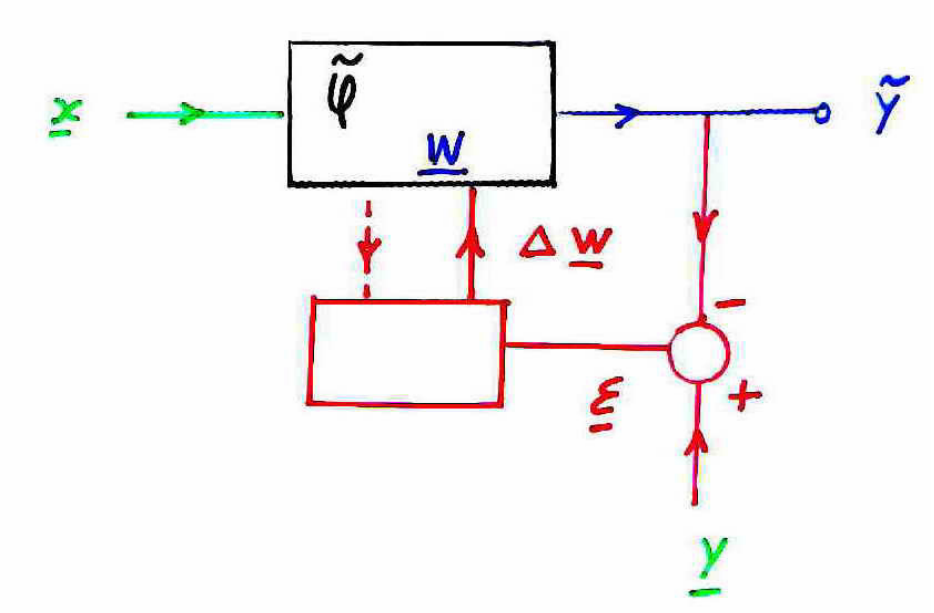
Modelo Numérico Geral (e.g. rede neural) – Caixa Preta

Ajuste dos parâmetros $\underline{w}$?

Treinamento,

Aprendizado ou

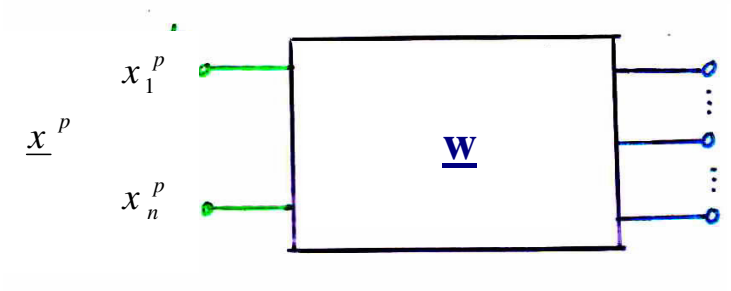
Ajuste dos Parâmetros



como um processo de otimização

Erro na saída ⟶ **Função objetivo à minimizar**

A saída é multidimensional:



Erro quadrático na saída para o k-ésimo par entrada-saída k

$$\left(\varepsilon^p\right)^2 = \left|\underline{y}^p - \underline{\tilde{y}}^p\right|^2 = \sum_{m=1}^{M} \left(y_m^p - \tilde{y}_m^p\right)^2 = \sum_{m=1}^{M} \left(\varepsilon_m^p\right)^2$$

onde

$$\varepsilon_m^p = y_m^p - \tilde{y}_m^p \qquad e \qquad \tilde{y}_m^p = \tilde{y}_m^p\left(\underline{w}\right)$$

Erro quadrático **médio** para todos os P pares entrada-saída

$(\underline{x}^p, \underline{y}^p)$ $\quad p = 1, \ldots P$

Erro médio quadrático:

$$F(\underline{w}) = \mathrm{E}_p(\varepsilon^p)^2 = \frac{1}{P}\sum_{p=1}^{P}(\varepsilon^p)^2$$

Como calcular

$$\frac{\partial}{\partial w_i}\mathrm{E}_p(\varepsilon^p)^2$$???

$$\frac{\partial}{\partial w_i} \mathop{E}_{p} (\varepsilon^p)^2 = \frac{\partial}{\partial w_i} \frac{1}{P} \sum_{p=1}^{P} (\varepsilon^p)^2 = \frac{1}{P} \sum_{k=1}^{P} \frac{\partial}{\partial w_i} (\varepsilon^p)^2 = E \mathop{\frac{\partial}{\partial w_i}}_{p} (\varepsilon^p)^2$$

$$\nabla \mathop{E}_{p} (\varepsilon^p)^2 = \mathop{E}_{p} \nabla (\varepsilon^p)^2$$

Propriedade importante: o gradiente do valor esperado do erro quadrático é igual ao valor esperado do gradiente do erro quadrático

Consequência:

Calcularemos separadamente para cada par

$$\frac{\partial}{\partial w_i} (\varepsilon^p)^2 \qquad \forall\ p$$

e tomaremos a média

$$\frac{\partial}{\partial w_i} \mathop{E}_{p} (\varepsilon^p)^2 = \mathop{E}_{p} \frac{\partial}{\partial w_i} (\varepsilon^p)^2$$

Continuando

$$\frac{\partial}{\partial w_i}\left(\varepsilon^p\right)^2 = \frac{\partial}{\partial w_i}\sum_{m=1}^{M}\left(\varepsilon_m^p\right)^2 = \sum_{m=1}^{M}\frac{\partial}{\partial w_i}\left(\varepsilon_m^p\right)^2 =$$

$$= 2\sum_{m=1}^{M}\varepsilon_m^p\frac{\partial}{\partial w_i}\varepsilon_m^p \qquad \textit{onde} \qquad \varepsilon_m^p = y_m^p - \tilde{y}_m^p$$

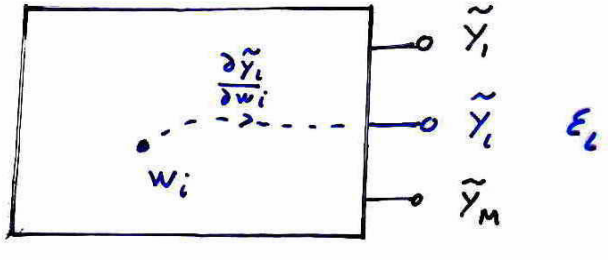
$$= -2\sum_{m=1}^{M}\varepsilon_m^p\frac{\partial\,\tilde{y}_m^p}{\partial w_i}$$

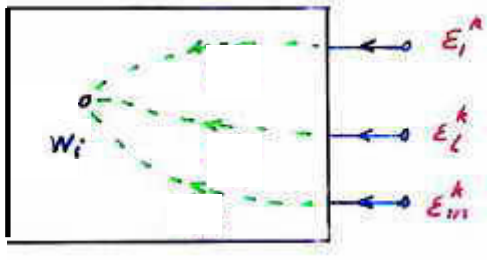
onde o termo $-2\varepsilon_m^p$ deriva da **formula do erro** usada

e o termo $\dfrac{\partial\, y_m^p}{\partial w_i}$ depende das **equações do modelo** usado

Para um único par entrada-saída o acréscimo a ser aplicado em w_i seria

$$\Delta w_i^p = -\alpha \frac{\partial}{\partial w_i} E_{p=1}\left(\varepsilon^p\right)^2 = -\alpha \frac{\partial}{\partial w_i}(\varepsilon)^2 = 2\alpha \sum_{m=1}^{M} \varepsilon_m \frac{\partial \tilde{y}_m}{\partial w_i}$$





$$\Delta w_i^p = 2\alpha \delta_i^p \quad onde \quad \delta_i^p = \sum_{m=1}^{M} \varepsilon_m \frac{\partial \tilde{y}_m}{\partial w_i}$$ - erro retropropagado

error backpropagation

Para os P pares:

$$\Delta w_i = \alpha \frac{\partial}{\partial w_i} \underset{k}{E}\left(\varepsilon^k\right)^2 = -2\alpha \frac{1}{P} \sum_{k=1}^{P} \sum_{l=1}^{m} \varepsilon_l^k \frac{\partial \tilde{y}_l^k}{\partial w_i}$$

Formas de Aplicação

Processo em Batelada – Batch

Batelada - Algorítmo

Até que o critério de parada seja satisfeito

para cada par $(\underline{x}^p, \underline{y}^p)$ $\quad p = 1, \ldots, P$

calcular $\quad \underline{\tilde{y}}^p = \varphi\left(\underline{x}^p\right)$

$$\varepsilon_m^p = y_m^p - \tilde{y}_m^p \quad \text{e} \quad \frac{\partial \tilde{y}_m^p}{\partial w_i} \quad \forall l, i$$

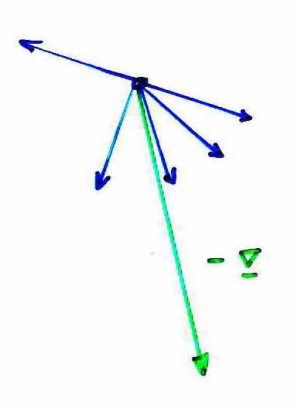
$$\frac{\partial \varepsilon^{p2}}{\partial w_i} = -2 \sum_{m=1}^{M} \varepsilon_m^p \frac{\partial y_m^p}{\partial w_i} \quad \forall i$$

outro par

calcular $\quad \Delta w_i = -\alpha \mathop{E}_{p} \frac{\partial \varepsilon^{p2}}{\partial w_i}$

$$w_i(n+1) = w_i(n) + \Delta w_i$$

reiniciar



Se P >> (P > ~ 200)

⇒ **muito lento**

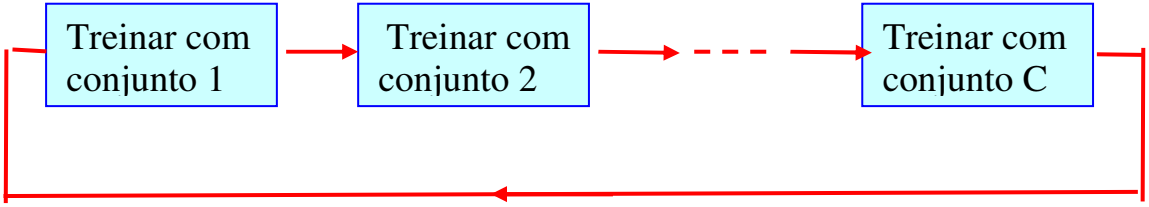
Processo por Lotes

Se P > ~200

Alocar os pares entrada-saída aleatoriamente em C conjuntos,

$$C \approx P / 200$$

Algorítmo



Lotes - Algorítmo

Até que o critério de parada seja satisfeito

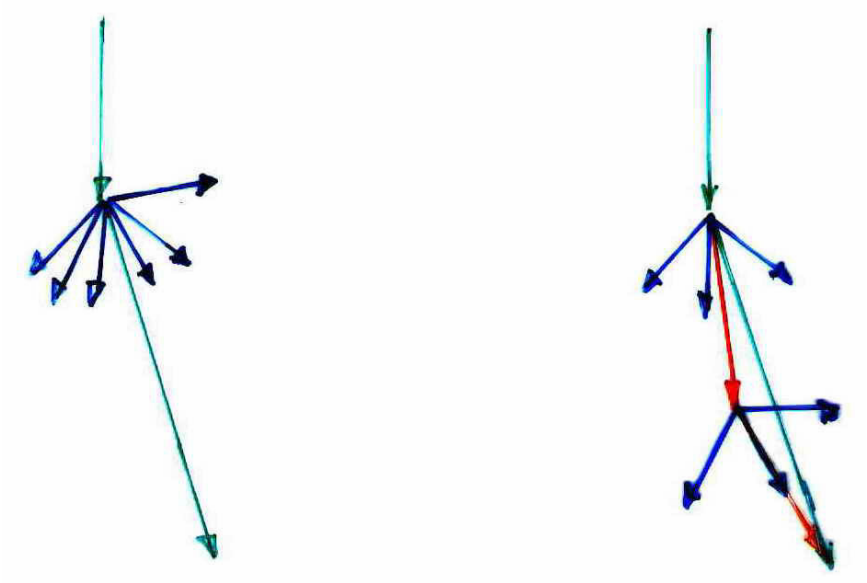
Treinar (em batelada) usando os pares do conjunto 1

Treinar (em batelada) usando os pares do conjunto 2

...

Treinar (em batelada) usando os pares do conjunto C

retornar



Reduz o tempo

Preserva a generalização mas

Diminui (pouco) a estabilidade do treinamento

Processo em “Regra Delta” -

uso “on line”

Atualiza os parâmetros a cada par entrada-saída apresentado

Regra Delta - Algorítmo

Até que o critério de parada seja satisfeito

para cada par ($\underline{x}^p$, $\underline{y}^p$) $\quad p = 1, \ldots, P$

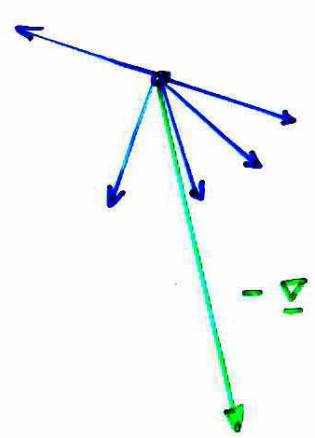
calcular $\quad \underline{\tilde{y}}^p = \varphi\left(\underline{x}^p\right)$

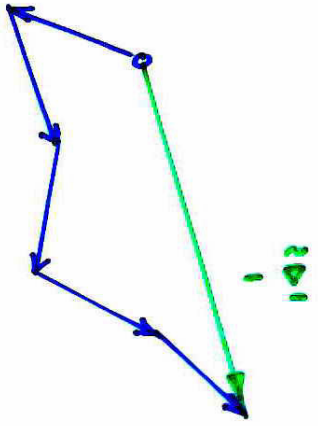
$$\varepsilon_m^p = y_m^p - \tilde{y}_m^p \quad \text{e} \quad \frac{\partial \tilde{y}_m^p}{\partial w_i} \qquad \forall\, l, i$$

$$\frac{\partial \varepsilon^{p2}}{\partial w_i} = -2 \sum_{m=1}^{M} \varepsilon_m^p \frac{\partial y_m^p}{\partial w_i} \qquad \forall\, i$$

$$w_i(n+1) = w_i(n) - \alpha \frac{\partial \varepsilon^{p2}}{\partial w_i} \qquad \forall i$$

outro par





Rápido mas F oscila !

Normalmente usado com momento

Treinamento com Momento

Regra delta: $$\Delta w_{i_D}(n) = 2\alpha \sum_{l=1}^{m} \varepsilon_l \frac{\partial \tilde{y}_l}{\partial w_i}$$

Momento $$\Delta w_i(n) = \beta \, \Delta w_i(n-1) + (1-\beta) \, \Delta w_{i_D}(n)$$

$$\beta_{\text{típico}} \approx .9$$

intermediário entre regra delta e batelada (lote)

$$\text{lote} \quad N_0 \approx \frac{4}{1-\beta}$$

$n = 1$ $\Delta w_D(1)$

$$\Delta w(1) = \beta \Delta w(0) + (1-\beta) \Delta w_D(1)$$

$$= (1-\beta) \Delta w_D(1)$$

$n = 2$ $\Delta w_D(2)$

$$\Delta w(2) = \beta \Delta w(1) + (1-\beta) \Delta w_D(2) =$$

$$= \beta(1-\beta) \Delta w_D(1) + (1-\beta) \Delta w_D(2)$$

$n = 3$ $\Delta w_D(3)$

$$\Delta w(3) = \beta \Delta w(2) + (1-\beta) \Delta w_D(3) =$$

$$= \beta^2(1-\beta) \Delta w_D(1) + \beta(1-\beta) \Delta w_D(2) + (1-\beta) \Delta w_D(3)$$

$n = 4 \quad \Delta w_D(4)$

$$\Delta w(4) = \beta^3(1-\beta)\,\Delta w_D(1) + \beta^2(1-\beta)\,\Delta w_D(2) + \beta(1-\beta)\,\Delta w_D(3) + (1-\beta)\,\Delta w_D(4)$$

…

$n \quad \Delta w_D(n)$

$$\Delta w(n) = \beta^{n-1}(1-\beta)\,\Delta w_D(1) + \beta^{n-2}(1-\beta)\,\Delta w_D(2) + \ldots$$

$$\ldots + \beta^{n-1}(1-\beta)\,\Delta w_D(i) + \ldots + (1-\beta)\,\Delta w_D(n)$$

$$\Delta w(n) = (1-\beta) \sum_{i=1}^{n} \Delta w_D(i)\, \beta^{n-i}$$

Contribuições em cada passo:

Média ponderada exponencialmente pelo atrazo

n – i = atrazo da i-ésima entrada em relação ao instante atual n

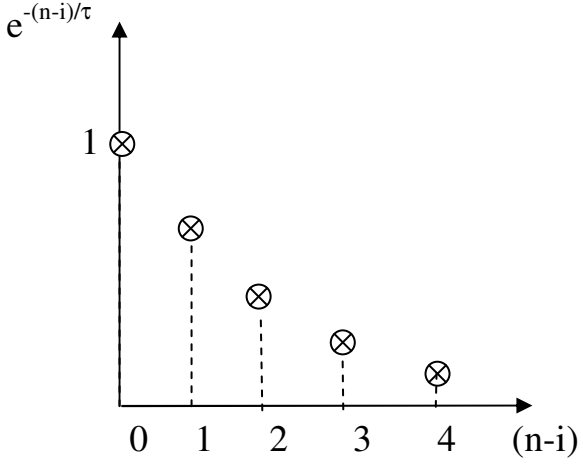
No passo n

$$\Delta w(n) = (1-\beta) \sum_{i=1}^{n} \Delta w_D(i)\ \beta^{n-i}$$

$$\beta^{n-i} = e^{(n-i)\ln\beta} = e^{-\frac{(n-i)}{\tau}}$$

$$\tau = -\frac{1}{\ln\beta} \approx \frac{1}{1-\beta} \qquad \text{se } \beta \approx 1$$

$$\Delta w(n) = (1-\beta) \sum_{i=1}^{n} \Delta w_D(i)\ e^{-(n-i)/\tau}$$

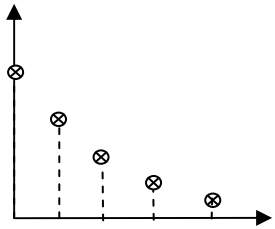


média ponderada exponencialmente pelo atraso

Contribuições significativas: peso > .02 $\Longrightarrow$ n-i < 4 τ

Máximo atrazo significativo (lote significativo):

$$N_0 \cong 4\tau \cong \frac{4}{1-\beta}$$



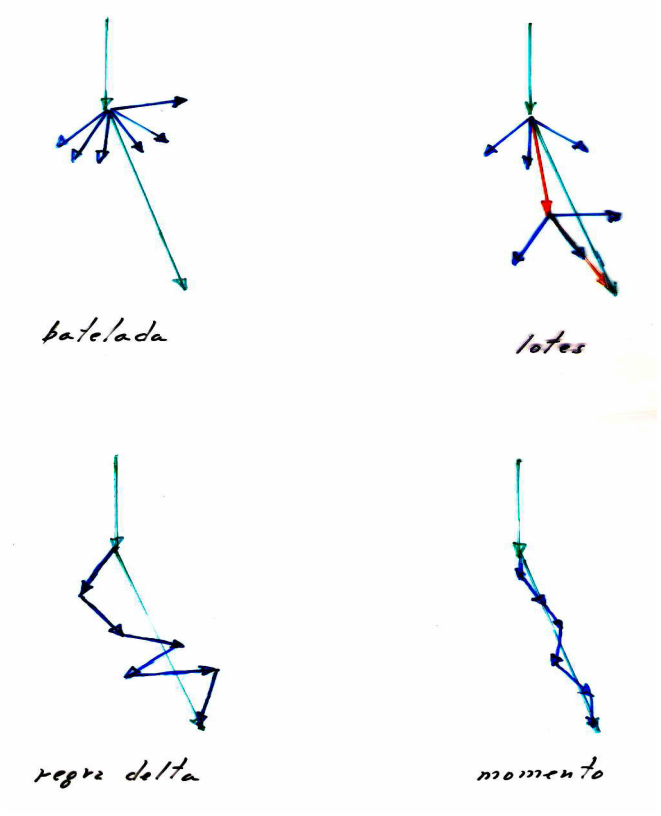
Contribuição total de $\Delta w_D(n)$:

$$\Delta w_D(n)\,(1-\beta)\,(1+\beta+\beta^2+\ldots) =$$

$$= \quad \Delta w_D(n)\,(1-\beta)\,\frac{1}{1-\beta} \quad = \quad \mathbf{\Delta w_D(n)}$$

Compensa a “instabilidade” do

processo de regra delta



Ajuste de Modelo por Regra Delta com Momento - Algorítmo

Inicialização n=1; $\alpha = .1$ $\beta = .9$ $\Delta w_i^{RD}(0) = 0 \quad \forall\, i$

para as i = 1, 2, …, V variáveis $w_i(1) \in [+.2, -.2]$ randômicos;

Até que o critério de parada seja satisfeito

para cada par ($\underline{x}^p$, $\underline{y}^p$) p = 1, …, P

calcular

$$\underline{\tilde{y}}^p = \varphi\left(\underline{x}^p\right)$$

$$\varepsilon_m^p = y_m^p - \tilde{y}_m^p \quad \text{e} \quad \frac{\partial \tilde{y}_m^p}{\partial w_i} \quad \forall\, m, i$$

Regra Delta com Momento - Algorítmo (continuação)

$$\Delta w_i^{RD}(n) = 2\alpha \sum_{m=1}^{M} \varepsilon_m^p \frac{\partial y_m^p}{\partial w_i} \qquad \forall \, i$$

$$\Delta w_i(n) = \beta \Delta w_i(n-1) + (1-\beta) \Delta w_i^{RD}(n) \qquad \forall \, i$$

$$w_i(n+1) = w_i(n) + \Delta w_i(n) \qquad \forall i$$

$$n = n + 1$$

outro par

Ajuste de modelo por BP Resiliente - Algorítmo

Inicialização

n = 1 $\alpha_{\max} \approx .5$ $a \approx 1.05$

para as i = 1, 2, …, V variáveis

$w_i(1) \in [+.2, -.2]$ randômicos; $\alpha_i(1) = .1$ $\frac{\partial F}{\partial w_i}(0) = 0$

Até que o critério de parada esteja satisfeito

Passo de treinamento n

BP Resiliente - Algorítmo (continuação)

Cálculo do gradiente

para cada par $(\underline{x}^p, \underline{y}^p)$ $\quad p = 1, \ldots, P$

calcular $\quad \underline{\tilde{y}}^p = \varphi\left(\underline{x}^p\right)$

$$\varepsilon_m^p = y_m^p - \tilde{y}_m^p \quad \text{e} \quad \frac{\partial \tilde{y}_m^p}{\partial w_i} \quad \forall\, m, i$$

$$\frac{\partial \varepsilon^{p2}}{\partial w_i} = -2 \sum_{m=1}^{M} \varepsilon_m^p \frac{\partial y_m^p}{\partial w_i} \quad \forall\, i$$

outro par

calcular $\quad \frac{\partial F}{\partial w_i}(n) = \underset{p}{E} \frac{\partial \varepsilon^{p2}}{\partial w_i}$

BP Resiliente - Algorítmo (continuação)

Cálculo do acréscimo

para i = 1, 2, …, V

$$\alpha_i(n) = \begin{cases} a\,\alpha_i(n-1) & se\ \ sign\dfrac{\partial F}{\partial w_i}(n) = sign\dfrac{\partial F}{\partial w_i}(n-1) \\ & mas\ se\ \alpha_i(n) > \alpha_{max}\ \ faça\ \alpha_i(n) = \alpha_{max} \\ \dfrac{\dfrac{\partial F}{\partial w_i}(n-1)}{\dfrac{\partial F}{\partial w_i}(n-1) - \dfrac{\partial F}{\partial w_i}(n)}\alpha_i(n-1) & se\ \ sign\dfrac{\partial F}{\partial w_i}(n) \neq sign\dfrac{\partial F}{\partial w_i}(n-1) \end{cases}$$

$$\Delta w_i(n) = -\,\alpha_i(n)\,\frac{\partial F}{\partial w_i}(n)$$

$$w_i(n+1) = w_i(n) + \Delta w_i(n)$$

n = n + 1

Ajuste de Modelo por Levenberg- Marquadt - Algorítmo

Inicialização

n = 1; para as i = 1, 2, …, V variáveis $w_i(1) \in [+.2, -.2]$ randômicos

Até que o critério de parada esteja satisfeito

Passo de treinamento n

Levenberg- Marquadt - Algorítmo (continuação)

Cálculo do gradiente g_i] e da aproximação da Hessiana [$\tilde{h}_{ij}$]

para cada par ($\underline{x}^p$, $\underline{y}^p$) p = 1, …, P

calcular $\underline{\tilde{y}}^p = \varphi\left(\underline{x}^p\right)$

$$\varepsilon_m^p = y_m^p - \tilde{y}_m^p \quad \text{e} \quad \frac{\partial \tilde{y}_m^p}{\partial w_i} \quad \forall\, m, i$$

$$\frac{\partial \varepsilon^p}{\partial w_i} = -\sum_{m=1}^{M} \frac{\partial y_m^p}{\partial w_i} \qquad \frac{\partial \varepsilon^{p2}}{\partial w_i} = -2\sum_{m=1}^{M} \varepsilon_m^p \frac{\partial y_m^p}{\partial w_i} \quad \forall\, i$$

outro par

calcular $$\frac{\partial F}{\partial w_i}(n) = g_{i\,(n)} = \underset{p}{E} \frac{\partial \varepsilon^{p2}}{\partial w_i}$$

$$\frac{\partial^2 F}{\partial w_i \partial w_j}(n) \cong \tilde{h}_{ij\,(n)} = 2\,\underset{p}{E} \frac{\partial \varepsilon^p}{\partial w_i} \frac{\partial \varepsilon^p}{\partial w_j}$$

Levenberg- Marquadt - Algorítmo (continuação)

Cálculo e aplicação do passo $\Delta \underline{w}$

escolha λ_n adequado e inverta $\tilde{\tilde{\underline{H}}}_n = [\tilde{\underline{H}}_n + \lambda_n \underline{I}]$

calcule $\underline{d}_n = -\tilde{\tilde{\underline{H}}}_n^{-1} \underline{g}_n$

realize a otimização em linha

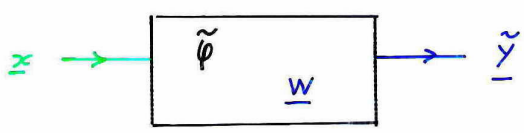
calcule $\alpha^* = \underset{\alpha>0}{\text{Arg}} [\text{Min } F(\underline{w}_n + \alpha_n \underline{d}_n)]$

$\underline{w}_{n+1} = \underline{w}_n + \alpha^* \underline{d}_n$

n = n + 1

Comentários finais

Composição de F como erro



$F\left(\underline{y}, \underline{\tilde{y}}\right)$ $\quad$ $\underline{\tilde{y}}\left(\underline{w}\right)$ $\quad$ $\dfrac{\partial F}{\partial w_i} = \dfrac{\partial F}{\partial \tilde{y}} \dfrac{\partial \tilde{y}}{\partial w_i}$

onde $\dfrac{\partial F}{\partial \tilde{y}}$ depende da função erro utilizada e

$\dfrac{\partial \tilde{y}}{\partial w_i}$ depende das equações do modelo utilizado

no caso de saídas múltiplas

$$\frac{\partial F}{\partial w_i} = \sum_{m=1}^{M} \frac{\partial F}{\partial \tilde{y}_m} \frac{\partial \tilde{y}_m}{\partial w_i}$$

Outros critérios de erro

Erro quadrático médio ponderado

$$F = \underset{p}{E} \sum_{m} a(p, m)\, \varepsilon_m^2$$

a(m) escalamento de y_m

a(p) população

Saídas com relevâncias diferentes

Erro relativo médio quadrático

$$F = \underset{p}{E}\left(\frac{Y_m - \tilde{Y}_m}{Y_m} \right)^2$$

Y – variáveis não normalizadas

Domínios com baixa população

$$a(p,m) = \frac{P_{\max}}{Pop(p)}$$

Crítica (acompanhamento) durante o treinamento

Evolução do Erro durante o treinamento

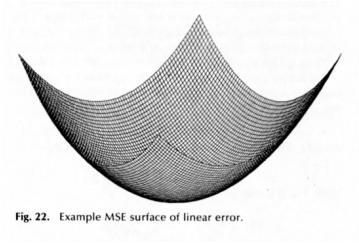


Fig. 22. Example MSE surface of linear error.

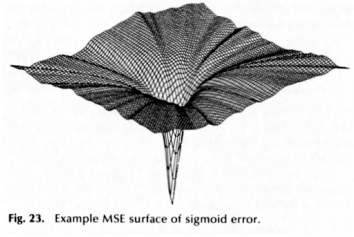


Fig. 23. Example MSE surface of sigmoid error.

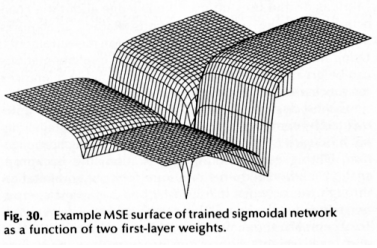
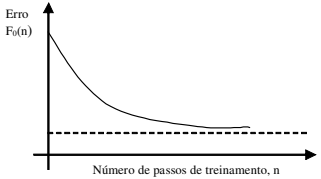
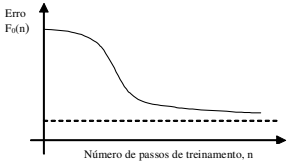
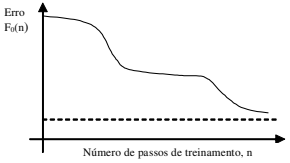


Fig. 30. Example MSE surface of trained sigmoidal network as a function of two first-layer weights.

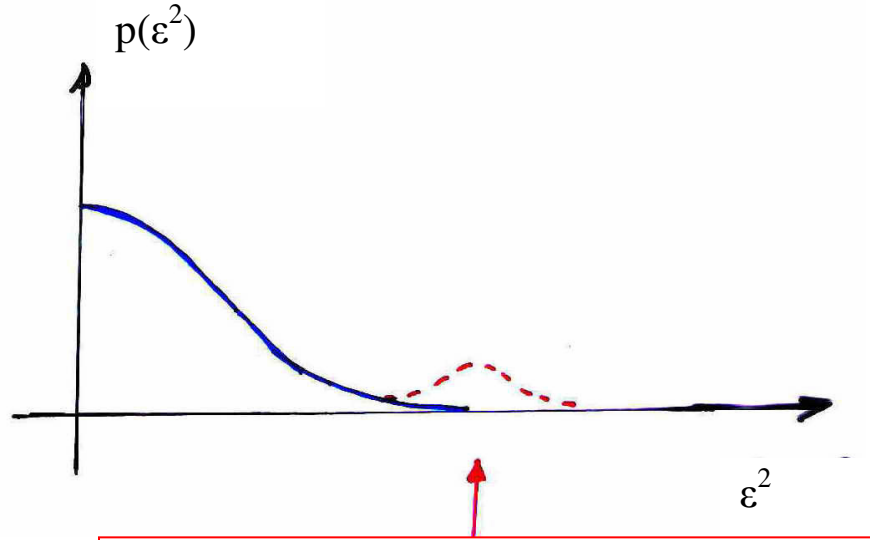






Paralisia (o ponto de partida é crítico)

Crítica pós treinamento



possívelmente devido à

elementos de região com baixa população ou

intruso (outlayer) e seus vizinhos

Por enquanto

FIM

Mas mais detalhes (importantes) serão vistos para redes neurais em

CPE 721

no próximo período.