Noções de Otimização de Funções Contínuas

Aproximação local (Taylor), Gradiente ∇, Hessiana H

Extremos, mínimo, máximo

Propriedades ∇ e H

Métodos analíticos

Métodos numéricos recursivos

Gradiente descendente, passo, passo variável, ótimo

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

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.

Neural Networks for Optimization and Signal Processing

Andrzej Cichocki
Warsaw University of Technology
Poland

Rolf Unbehauen
Universität Erlangen-Nürnberg
Germany

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

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

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

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

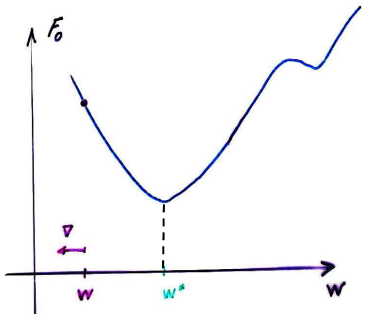
irrestrita

contínua, derivável

Como fazer ???

Noções Elementares de Otimização

I – Função objetivo – a minimizar



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

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

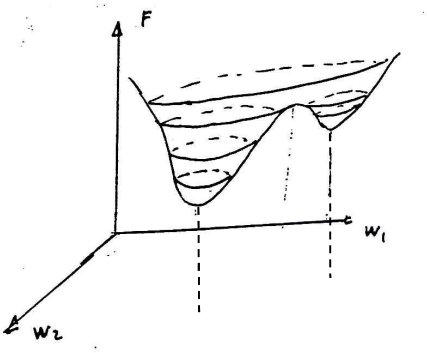
Mínimo, minimante $\underline{w}^*$

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

→ 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

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

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

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

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

Então

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

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(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 w.

Aproximação quadrática ou de 2ª. ordem

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

Obs: Se F(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

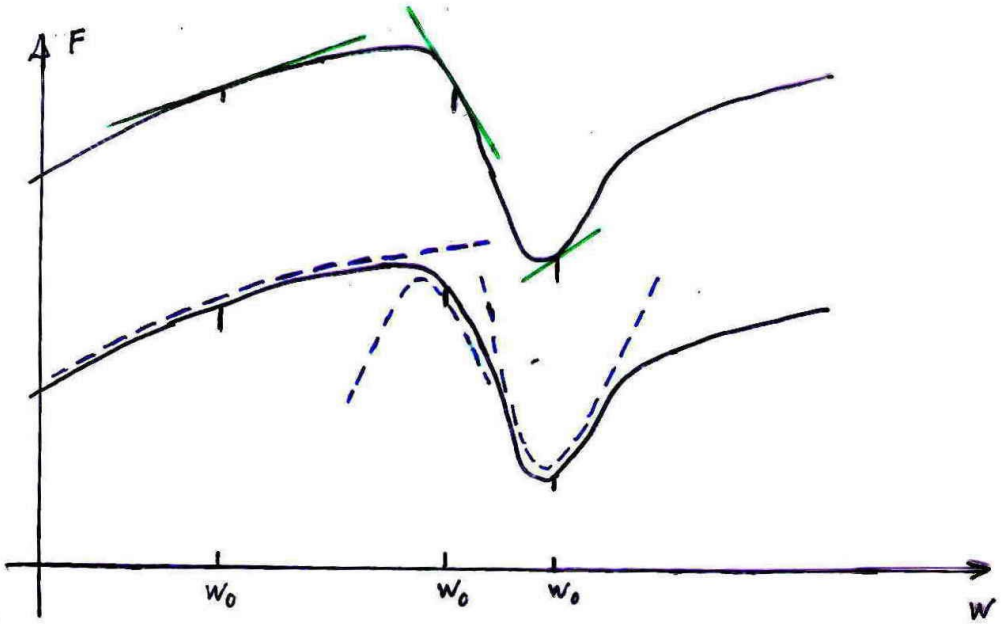
$$F(\underline{w}_0 + \Delta\underline{w}) \cong F(\underline{w}_0) + \sum_i \left.\frac{\partial F}{\partial w_i}\right|_{\underline{w}_0} \Delta w_i$$

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

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

Funções de ordem mais alta

validade das aproximações ?



A validade da aproximação

depende do acréscimo na função (nas variáveis dependentes),

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

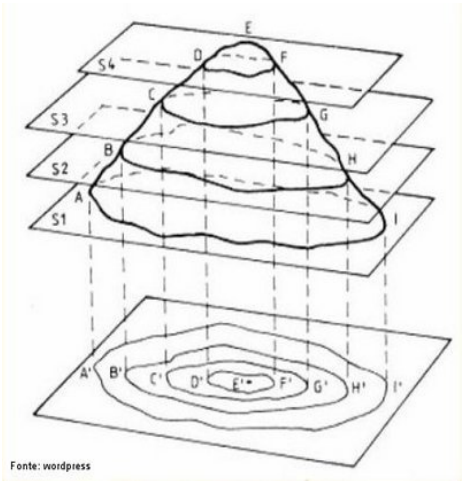
$$| \frac{1}{2} \underline{\Delta w}^t \; \underline{H}(\underline{w}_0) \underline{\Delta w} | \quad << \quad | \underline{\Delta w}^t \; \underline{\nabla} \; F(\underline{w}_0) |$$

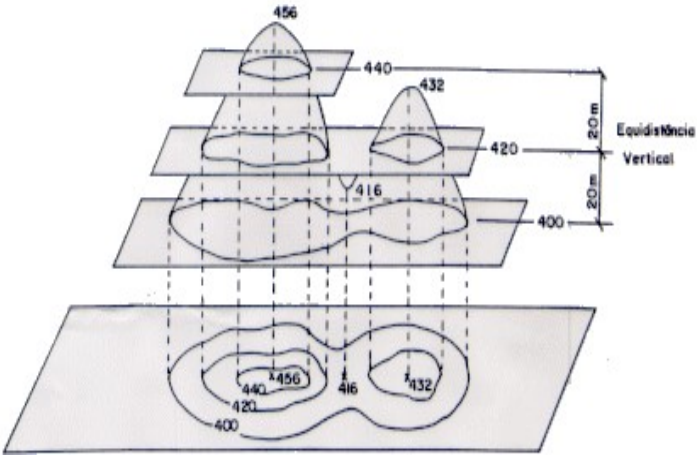
ou, alternativamente, enquanto

$$| \frac{1}{2} \underline{\Delta w}^t \; \underline{H}(\underline{w}_0) \underline{\Delta w} | \quad << \quad \varepsilon$$

Visualização de superfícies 3D em 2D

$h(x_1, x_2)$ Curvas de nível

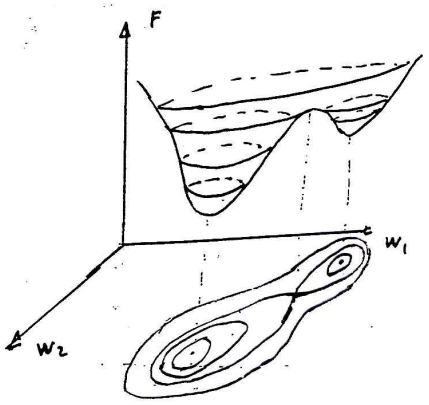


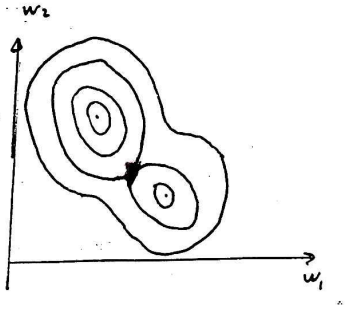


Funções de duas variáveis

Visualização por

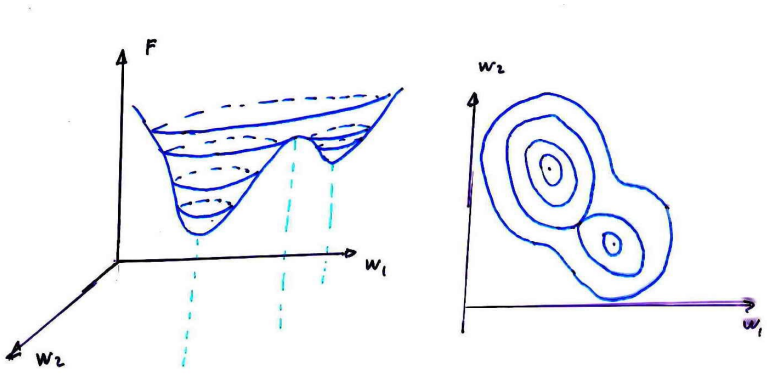
Curvas de Nível





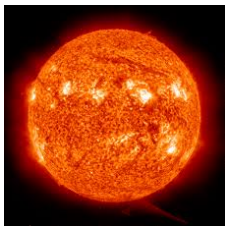
Curvas / Superfícies de nível

Tangentes



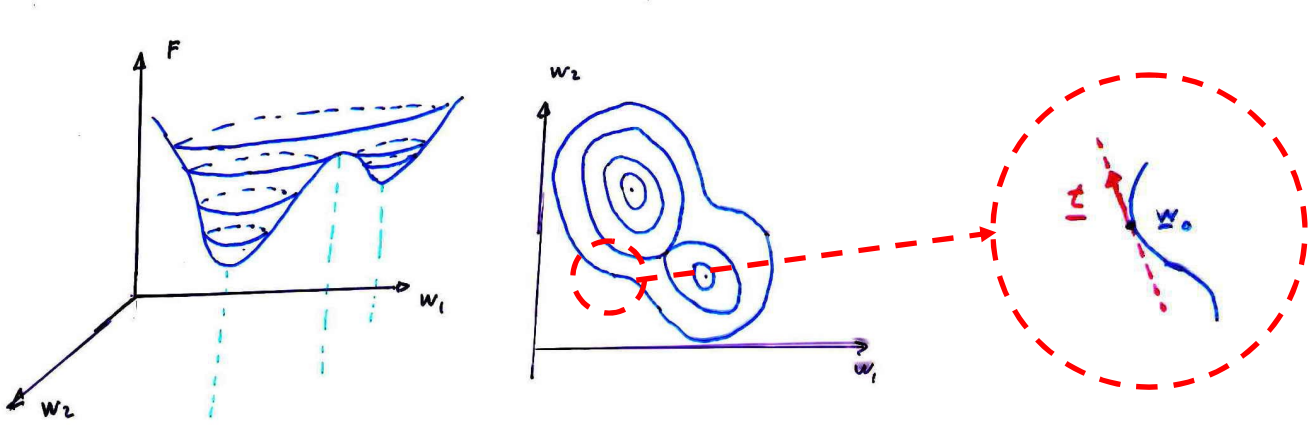
$\underline{w}$ $(\dim 2)$ $\rightarrow$ curva de nível $(\dim 2)$ $\rightarrow$ reta t_g $(\dim 2)$

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



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

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 $\underline{w}^*$ - local

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

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

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

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

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

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

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

Nos extremos

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

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

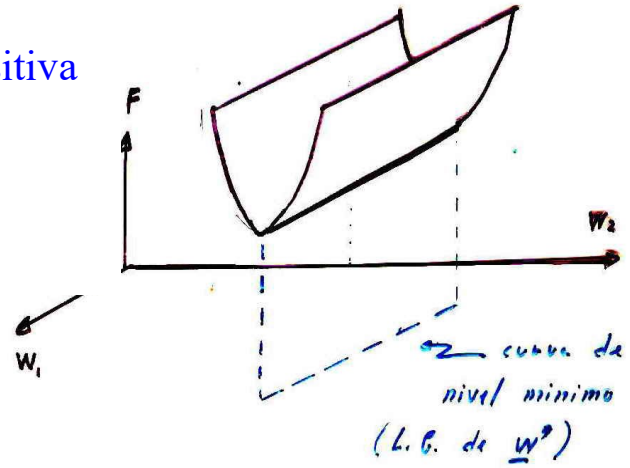
$$= F(\underline{w}^*) + \frac{1}{2}\Delta\underline{w}^t\, \underline{H}\, \Delta\underline{w}$$

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

$\underline{H}$ semidefinida positiva

$\Delta F \geq 0$

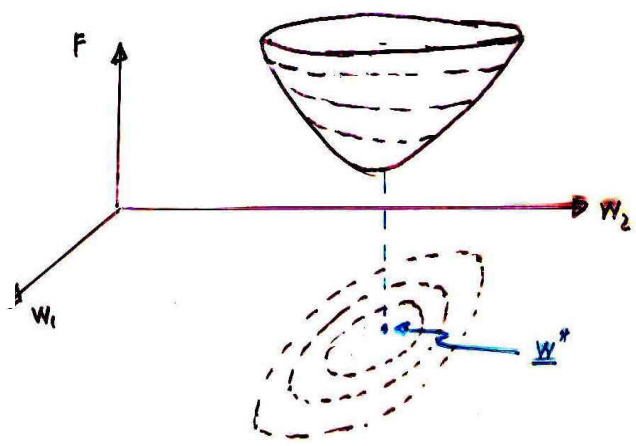
CALHA



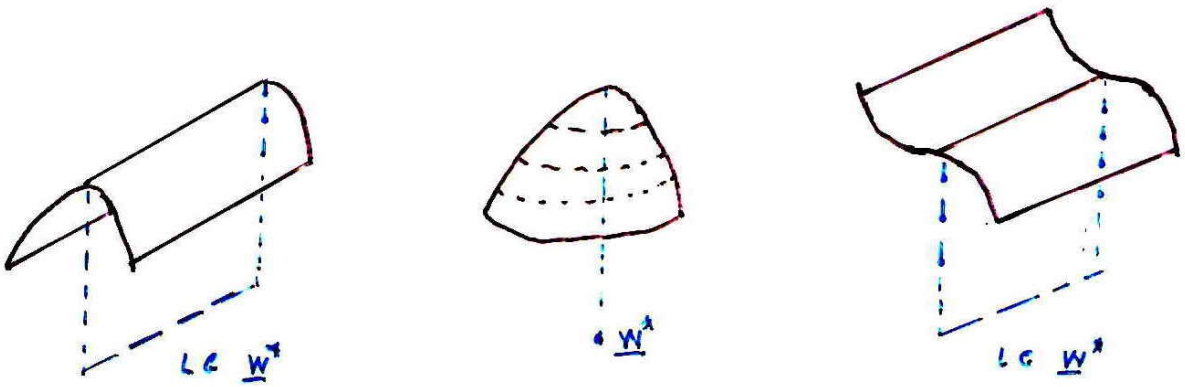
$\underline{H}$ definida positiva

$\Delta F > 0$

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



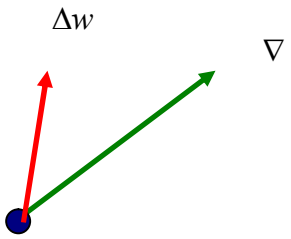
IV – Propriedades do Gradiente

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

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

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

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



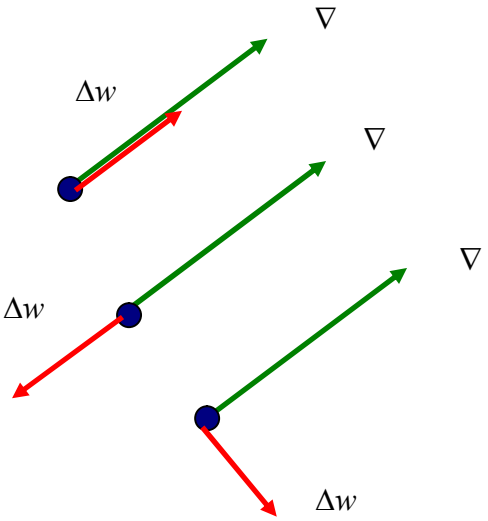
$$\Delta F \cong \underline{\Delta w}^t \, \nabla F = |\underline{\Delta w}| \; |\underline{\nabla} F| \cos\theta$$

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

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

$(-\Delta F \quad \max)$

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



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

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

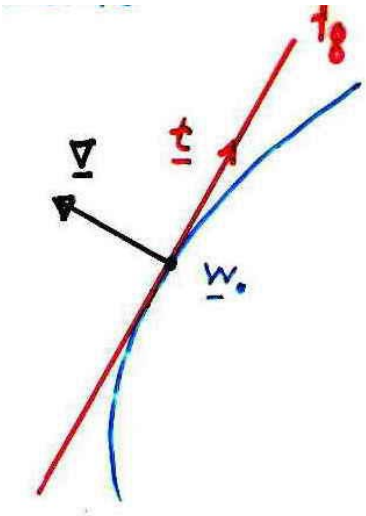
$$\frac{d}{d\alpha} F(\underline{w}_0 + \alpha \, \underline{t}) \Big|_{\forall \underline{t} \in \pi_{\in}} = 0$$

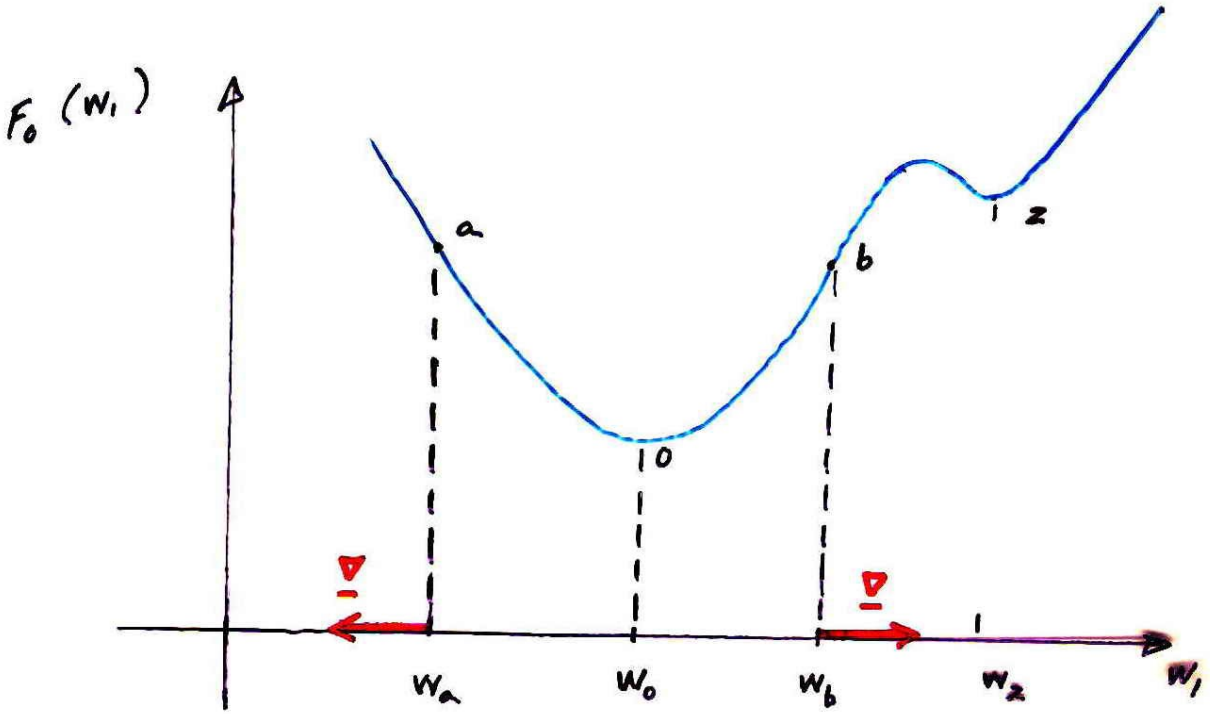
$\Delta F \cong \underline{\Delta w}^t \, \underline{\nabla} F$

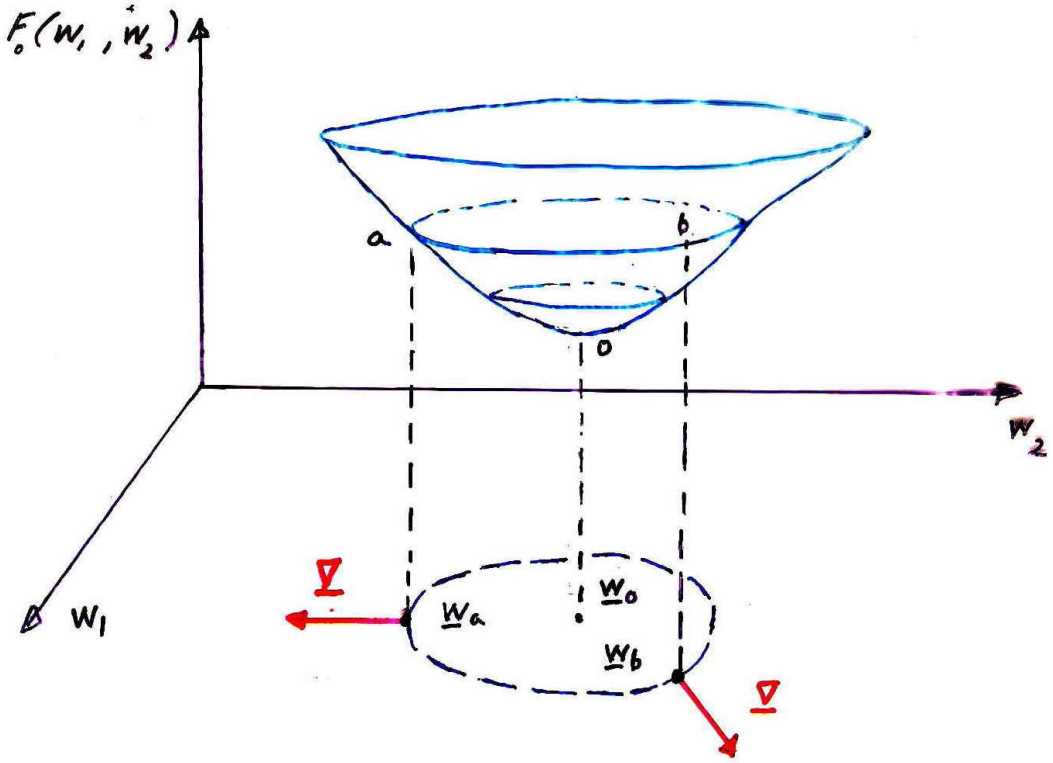
$\Delta F = 0 \quad \rightarrow \quad \underline{\Delta w} \perp \underline{\nabla F}$

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

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




$F_0(w_1)$
a
b
z
0
$-\nabla$
$-\nabla$
w_a
w_0
w_b
w_2
w_1


$F_0(w_1, w_2)$
a
b
0
w_2
w_1
$-\nabla$
$\underline{w}_a$
$\underline{w}_0$
$\underline{w}_b$
$-\nabla$

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

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

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

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

B – Funções Quadráticas

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

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

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

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

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

Newton Newton Raphson

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

Evitando $\underline{H}$

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

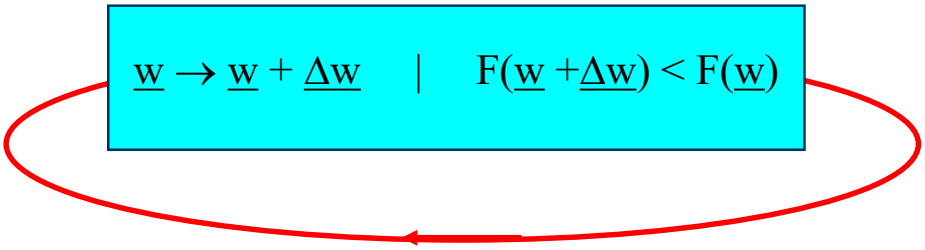
Levemberg-Marquardt

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

C – Funções de ordens mais elevadas e/ou transcendentes

muito complexo $\Longrightarrow$ Métodos Numéricos Recursivos

VII – Otimização: Métodos Numéricos Recursivos



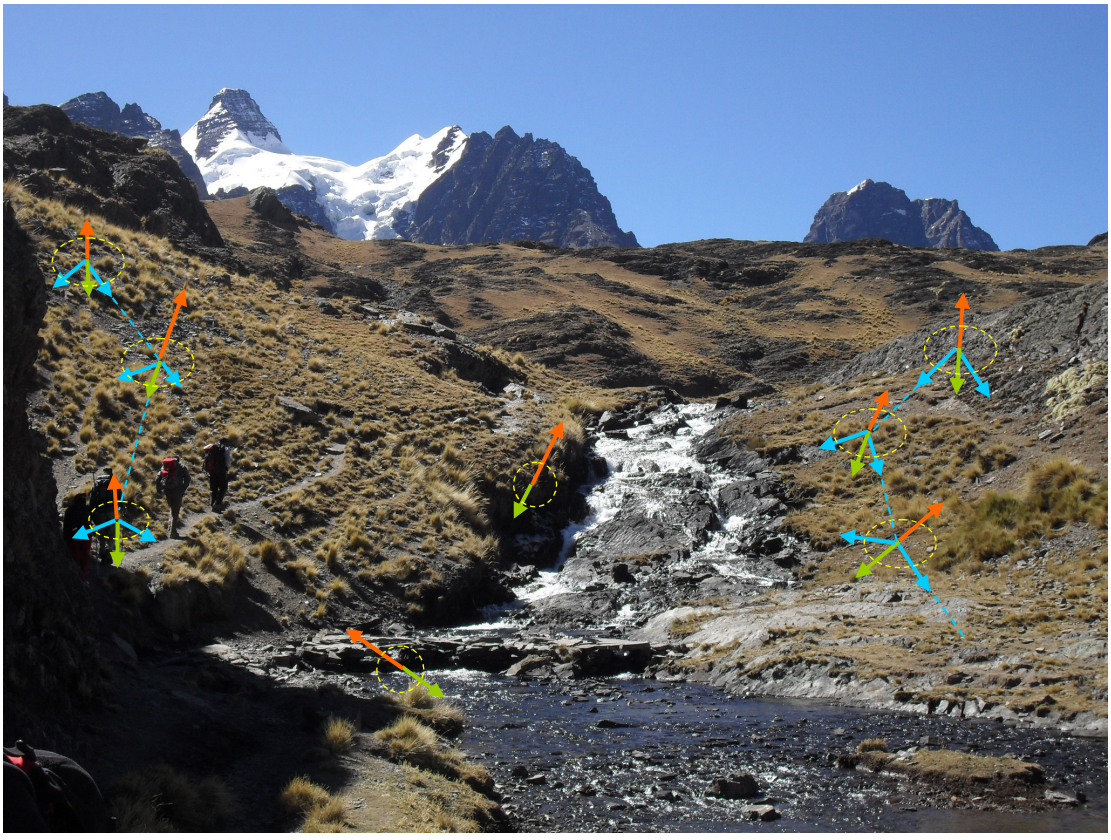
$\underline{\Delta w}$??

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

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

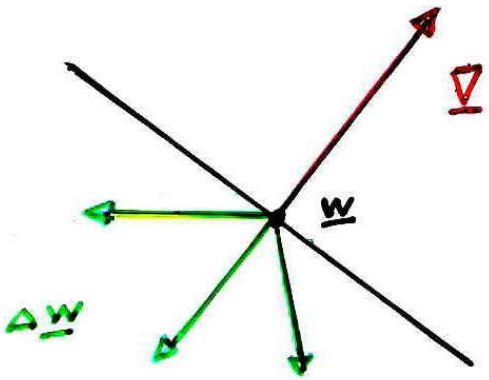
$$F(\underline{w} + \Delta\underline{w}) \approx F(\underline{w}) + \Delta\underline{w}^t \nabla(\underline{w})$$

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

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

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 $-\Delta F$?

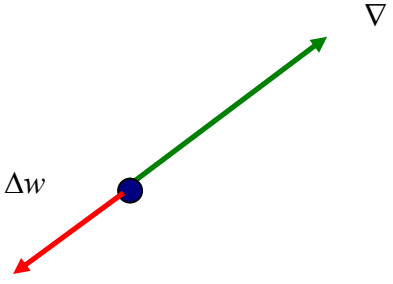
$$F(\underline{w} + \Delta\underline{w}) = F(\underline{w}) + \Delta\underline{w}^t \nabla$$

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

constante (-1, 1)

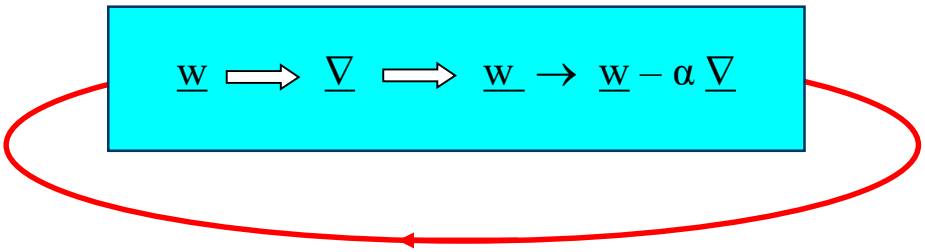
$$\underline{d} = - \nabla$$

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



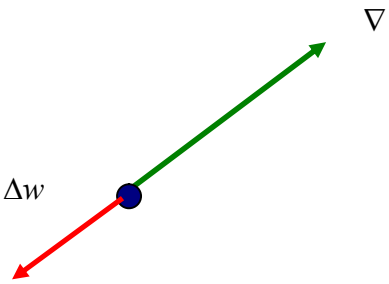
Método do Gradiente Descendente ou

Método da Descida pelo Gradiente

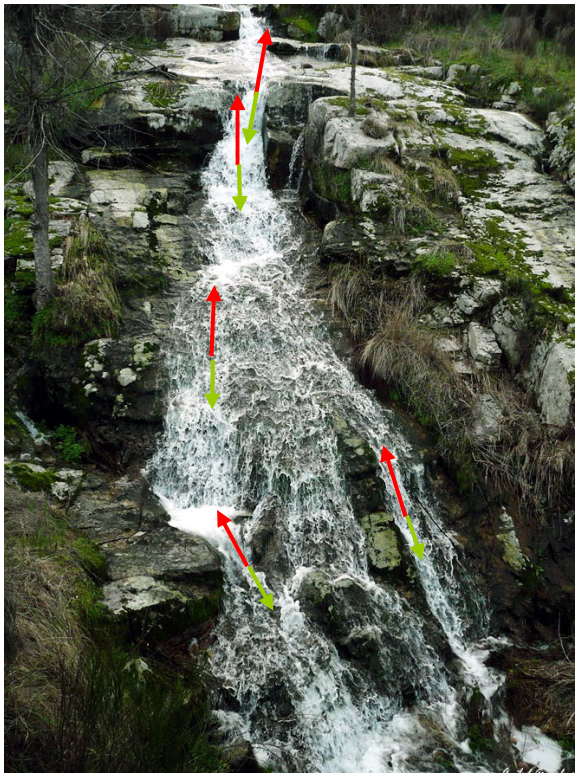


$$\Delta \underline{w} = - \alpha \, \underline{\nabla}$$

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

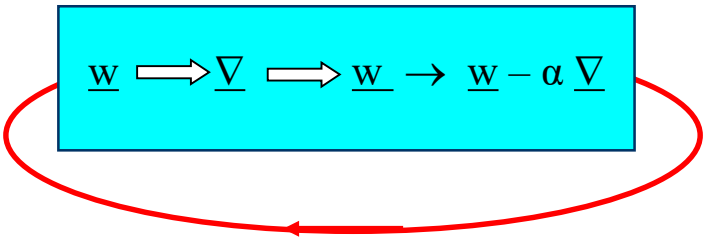


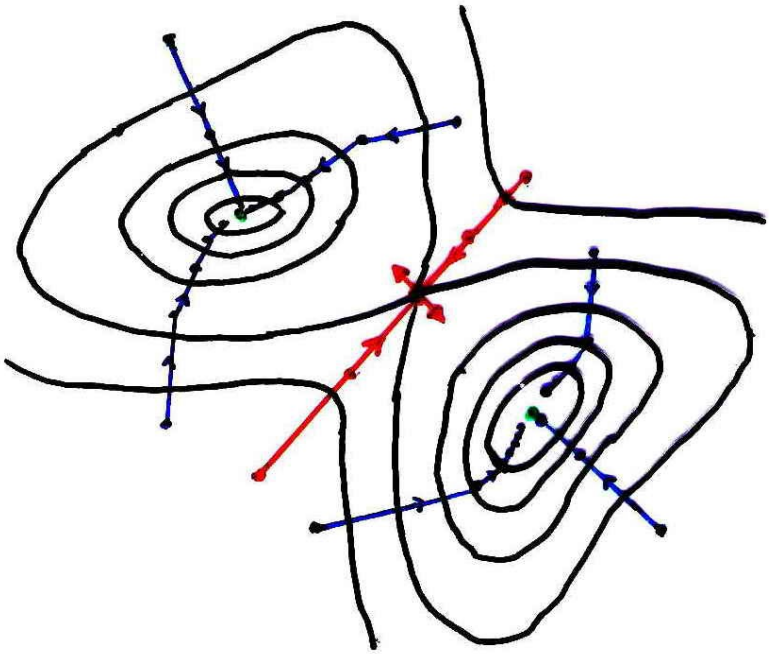
Gradiente Descendente



Método do

Gradiente Descendente





Algorítmo

Até que o critério de parada seja satisfeito

Para cada variável w_i

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

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

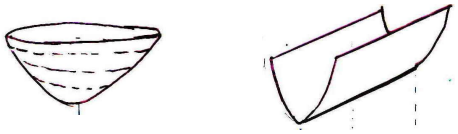
Fim (natural) do processo

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

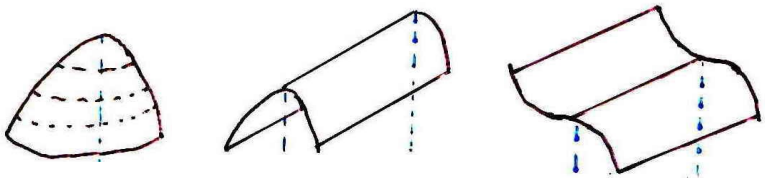
$E(\Delta \underline{w}) = \underline{0} \quad \Rightarrow \quad \nabla(\underline{w}_0) = \underline{0}$

O que é $\underline{w}_0$??

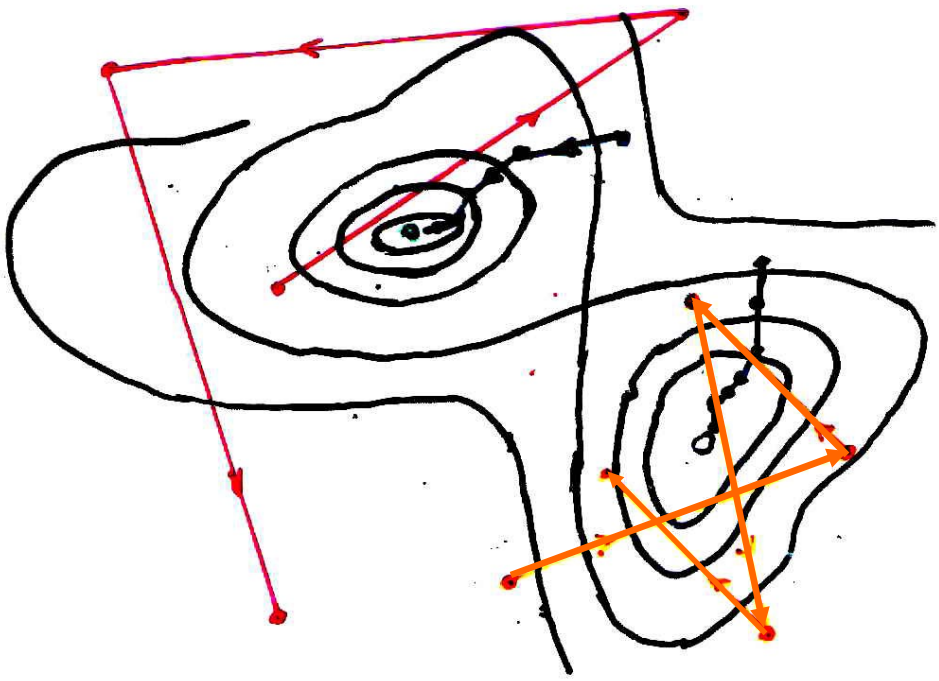
pontos estáveis: mínimo, calha



pontos instáveis: máximo, cumeeira, sela

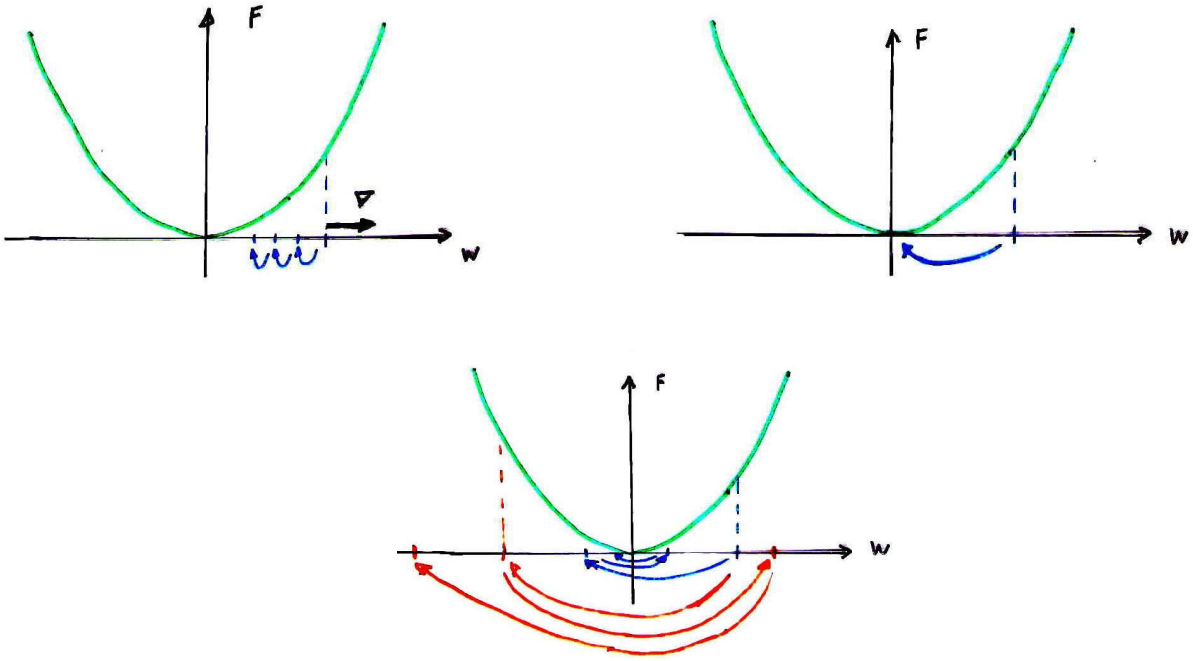


Passo de Treinamento α

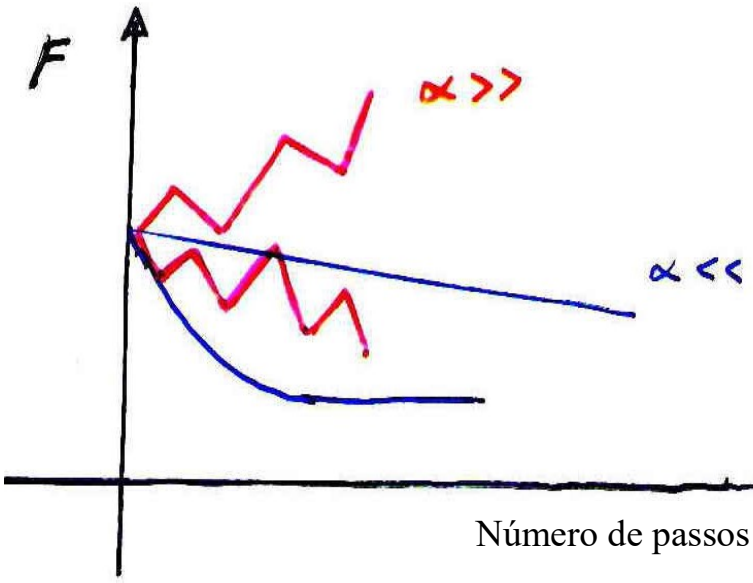


Efeito do passo de treinamento α

Exemplo simples: $F = w^2$



Evolução do erro ao longo do treinamento:



α excessivamente grande
o processo diverge

α muito pequeno (ou muito grande)
o processo converge mas é lento

α adequado
o processo converge em bom tempo

Passo de Treinamento α

α << convergência lenta

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

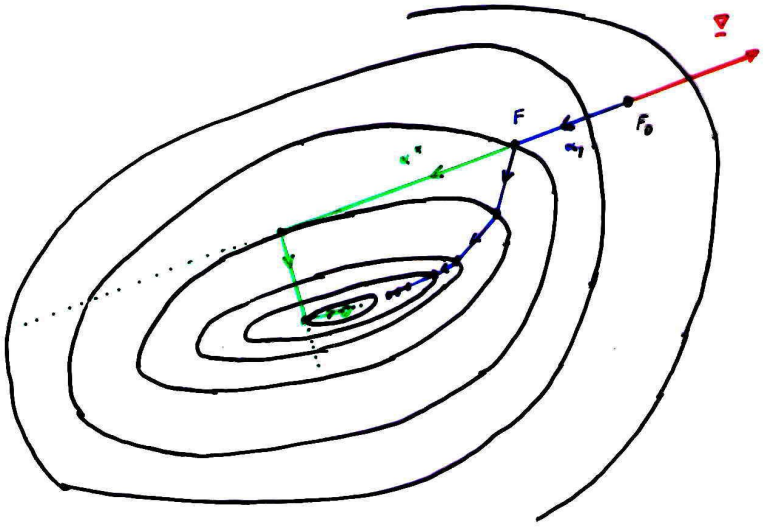
⇒ **$\alpha_{ótimo}$??**

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

No treinamento de redes neurais

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

Passo Ótimo de Treinamento α^*



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

Passo ótimo α* em uma direção $\underline{d}$ qualquer

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

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

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

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

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

$\alpha = \alpha^* \quad \Rightarrow \quad$ **$-\Delta F$** é máximo $\qquad \frac{\partial F}{\partial \alpha} = \frac{\partial \Delta F}{\partial \alpha} = 0$

$$\alpha^* = \text{Arg} \left[\text{Min } F(\underline{w}_0 + \alpha^* \, \underline{d}) \right]$$

$$\frac{\partial}{\partial \alpha} F = \underline{\nabla}^t \underline{d} + \alpha \, \underline{d}^t \, \underline{H} \, \underline{d} = 0$$

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

necessita de $\underline{H}$!

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

Aplicar $\alpha = \alpha_1 \quad \underline{w}_0 \rightarrow \underline{w}_0 + \alpha_1 \underline{d} \quad F_0 \rightarrow F_1$

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

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

aplicando (3) em (2)

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

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

$$\Delta F = \frac{1}{2} \alpha^* \underline{\nabla}^t \underline{d}$$

Obs:

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

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

No caso em que

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

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

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

Passo variável controlando o decréscimo de F

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

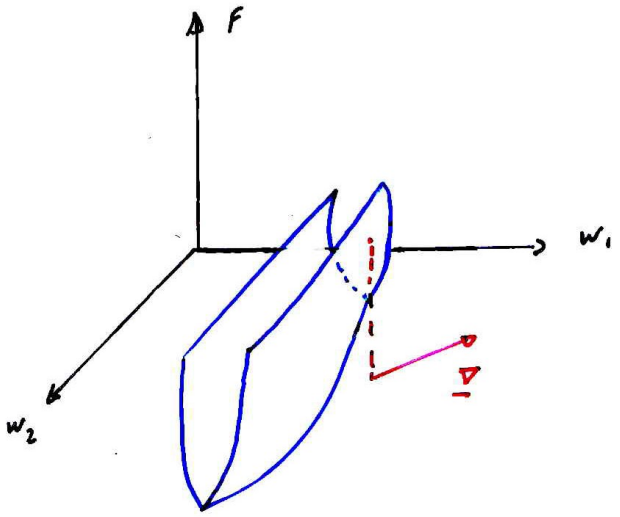
$$\alpha = \alpha_0 \quad \text{ctte} \quad \Rightarrow \qquad \Delta F = - \alpha \ \|\underline{\nabla}\|^2$$

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

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

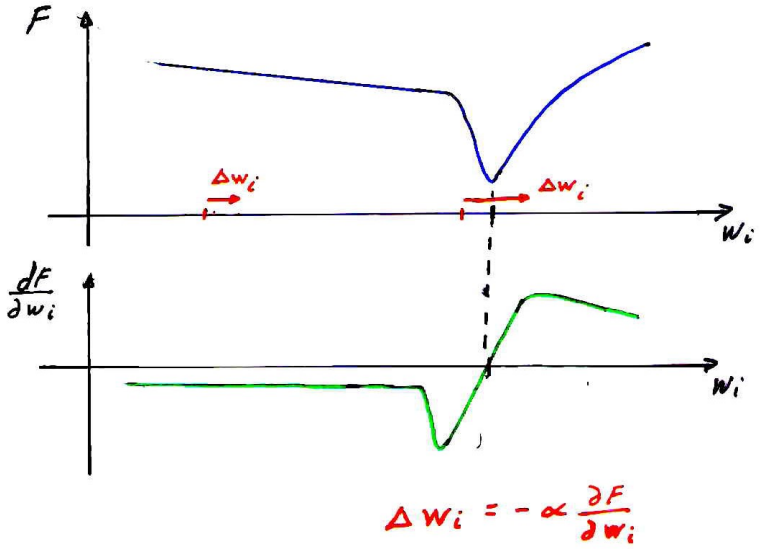
Acelerando o processo – passo variável

α diferenciado por variável



Acelerando o processo – passo variável

α diferenciado por passo



Resilient Backpropagation (modificado)

(Silva e Almeida, Super SAB, etc.)

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

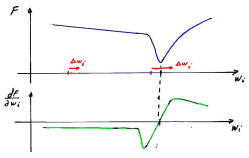
$$\Delta w_i(n) = - \alpha_i(n) \frac{\partial F}{\partial w_i}(n)$$

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

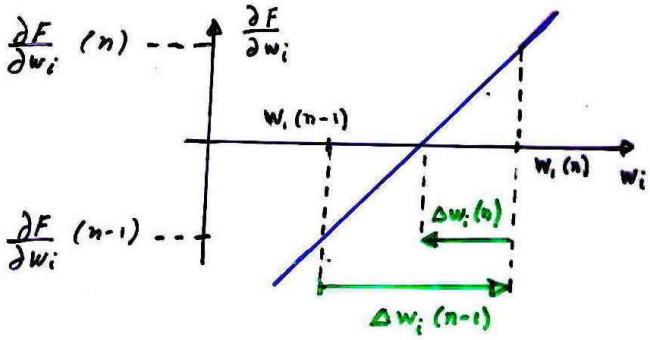
se $\alpha_i(n) > \alpha^*$ faça $\alpha = \alpha^*$

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

Outra alternativa (passo ótimo)



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



Algorítmo Resiliente

Inicialização

n = 1 $\alpha_{max} \approx .5$ $a \approx 1.05$ $\alpha_i(1) = .1$ $\frac{\partial F}{\partial w_i}(0) = 0$

inicializar as i = 1, 2, …, V variáveis

Até que o critério de parada esteja satisfeito

Aplicar um passo de treinamento

Para cada parâmetro w_i i = 1, 2, …, V

Calcular a componente do gradiente $\frac{\partial F}{\partial w_i}(n)$

Calcular o passo α a ser usado

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

Calcular e aplicar os acréscimos

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

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

n = n + 1

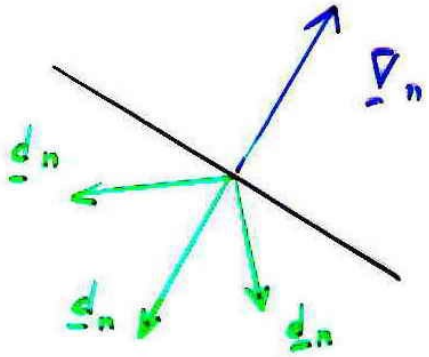
Próximo passo

Métodos de 1ª. Ordem (Resumo)

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

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

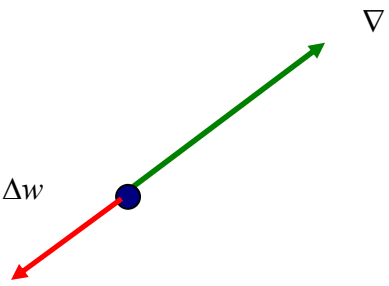
$$\underline{d}_n^t \, \underline{\nabla}_n < 0$$



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

$$\alpha_n = \text{Arg} \left[\text{Min } F(\underline{w}_n - \alpha_n \underline{\nabla}_n) \right]$$

BP Resiliente: α variável por sinapse e por passo

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

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

Condições necessárias

α pequeno, vale a aproximação de segunda ordem

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

Newton

$\underline{w}^* = \underline{w}_1 - \underline{H}_1^{-1}\ \underline{\nabla}_1$ para F($\underline{w}$) de segunda ordem

$\underline{w}_{n+1} = \underline{w}_n + \alpha_n\ \underline{d}_n$ para F($\underline{w}$) qualquer

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

$\alpha_n = \text{Arg}\ [\text{Min}\ F(\underline{w}_n + \alpha_n\ \underline{d})]$

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

Newton Amortecido

Levenberg-Marquardt

Caso $F(\underline{w}) = \underset{\forall k}{E}\ \varepsilon_k^2$

$$\nabla = g_i \Big] = \frac{\partial F}{\partial w_i} \Big] = 2 \underset{\forall k}{E}\ \varepsilon_k \frac{\partial \varepsilon_k}{\partial w_i} \Big]$$

$$H = \Big[h_{ij} \Big] = \left[\frac{\partial^2 F}{\partial w_i \partial w_j} \right] = \left[\underset{\forall k}{E} \frac{\partial \varepsilon_k^2}{\partial w_i \partial w_j} \right]$$

Duas aproximações de $\underline{H}$ são utilizadas

1 - Aproximação de $\underline{H}$ para evitar a derivada segunda

para um par k $\quad F_k(\underline{w}) = \varepsilon_k^2$

$$g_i = \frac{\partial(\varepsilon_k^2)}{\partial w_i} = 2\varepsilon_k \frac{\partial \varepsilon_k}{\partial w_i}$$

$$h_{ij} = \frac{\partial^2(\varepsilon_k^2)}{\partial w_i \partial w_j} = \frac{\partial}{\partial w_i}\left(\frac{\partial}{\partial w_j}\varepsilon_k^2\right) = 2\frac{\partial}{\partial w_i}\left(\varepsilon_k \frac{\partial \varepsilon_k}{\partial w_j}\right) = 2\left(\frac{\partial \varepsilon_k}{\partial w_i}\frac{\partial \varepsilon_k}{\partial w_j} + \varepsilon_k \frac{\partial^2 \varepsilon_k}{\partial w_i \partial w_j}\right)$$

$$h_{ij} \approx 2\frac{\partial \varepsilon_k}{\partial w_i}\frac{\partial \varepsilon_k}{\partial w_j} \approx \frac{1}{2\varepsilon_k^2} g_i g_j \qquad \underline{H}_k \approx \frac{1}{2\varepsilon_k^2}\nabla_k \nabla_k^t$$

a aproximação é válida no entorno do mínimo

para todos os pares

$$\nabla = 2\underset{\forall k}{E}\left[\varepsilon_k \frac{\partial \varepsilon_k}{\partial w_i}\right]$$

$$H \approx \underset{\forall k}{E}\left[\frac{1}{2\varepsilon_k^2}\nabla_k \nabla_k^t\right]$$

2- Aproximação de $\underline{H}$ para garantir a inversão da matriz

fazer $\underline{H}$ diagonal dominante

$$|h_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |h_{ij}| \qquad \forall\ i = 1,\ldots,n$$

Uma solução

$$\underline{\mathrm{H}}^{-1} \approx \left(\underline{\mathrm{H}} + \lambda \underline{I}\right)^{-1}$$

$$\lambda > 0 \qquad \lambda \to 0$$

Uma solução melhor

$$\underline{\mathrm{H}}^{-1} \approx \left(\underline{\mathrm{H}} + \lambda \, diag \underline{H}\right)^{-1}$$

$$0 \le \lambda \le \frac{|h_{ii}|}{\sum_{\substack{j=1 \\ j \neq i}}^{n} |h_{ij}|} - 1 \qquad \forall \; i = 1, \ldots, n \qquad \lambda \to 0$$

Levenberg – Marquardt

$$\underline{\nabla}_n = 2 \underset{\forall k}{\underline{E}} \left[\varepsilon_{\mathrm{k}} \frac{\partial \varepsilon_{\mathrm{k}}}{\partial w_i} \right]$$

$$\underline{H}_n \approx \underset{\forall k}{\underline{E}} \left[\frac{1}{2\varepsilon_{\mathrm{k}}^2} \nabla_k \nabla_k^t \right]$$

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

$$\underline{d}_n = \left[\underline{H}_n + \lambda \, \underline{\mathrm{I}}\right]^{-1} \underline{\nabla}_{\mathrm{n}}$$

$$\alpha_n = \mathrm{Arg} \left[\mathrm{Min} \, \mathrm{F}(\underline{\mathrm{w}}_{\mathrm{n}} + \alpha_n \, \underline{d}_n)\right]$$

Levenberg Marquadt - Algorítmo

Inicialização

n = 1; inicializar as w_i i = 1, 2, …, V variáveis

Até que o critério de parada esteja satisfeito

Passo de treinamento n

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

calcular $$\frac{\partial F}{\partial w_i}(n) = g_{i\ (n)} = E_p \frac{\partial \varepsilon^{p2}}{\partial w_i}$$

$$\frac{\partial^2 F}{\partial w_i\, \partial w_j}(n) \cong \tilde{h}_{ij\ (n)} = E_p \frac{1}{2\varepsilon^{p2}} \frac{\partial \varepsilon^p}{\partial w_i} \frac{\partial \varepsilon^p}{\partial w_j}$$

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

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

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

realize a otimização em linha

calcule α^* = Arg $\left[\text{Min } F(\underline{w}_n + \alpha_n\, \underline{d}_n)\right]$

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

n = n + 1

Quasi Newton

Fórmulas básicas:

$$\nabla_{n+1} = \nabla_n + H_n(w_{n+1} - w_n)$$

$$(\nabla_{n+1} - \nabla_n) = H_n(w_{n+1} - w_n)$$

$$(w_{n+1} - w_n) = G_n(\nabla_{n+1} - \nabla_n)$$

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

BFGS – Broyden – Fletcher – Goldfarb – Shanno

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

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

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

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

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

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

Gradiente Conjugado - não usa H

Fletcher – Reeves

$\underline{d}_i$ – uma só vez por dimensão

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

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

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

Leonard & Kramer

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

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)

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

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

$\underline{d}_n$ depende do método