

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

* Chapter 1. Mathematical Preliminaries of Neurocomputing	1
* 1.1. Linear Matrix Algebra	1
* 1.1.1. Matrix Representations and Notations	1
* 1.1.2. Inner and Outer Product	3
1.1.3. Linear Independence of Vectors	4
1.1.4. Rank of a Matrix	5
* 1.1.5. Positive and Negative Definite Matrices	5
* 1.1.6. The Inverse and Pseudoinverse of Matrices	5
1.1.7. Orthogonality, Unitary Matrices, Conjugate Vectors	7
1.1.8. Eigenvalues and Eigenvectors	7
1.1.9. Vector and Matrix Norms	9
1.1.10. Singular Value Decomposition (SVD)	10
1.1.11. Condition Numbers	12
1.1.12. The Kronecker Product	14
1.2. Elements of Multivariable Analysis	15
1.2.1. Sets	15
* 1.2.2. Functions	16
* 1.2.3. Differentiation of a Scalar Function with Respect to a Vector	17
* 1.2.4. The Hessian Matrix	17
1.2.5. The Jacobian Matrix	18
* 1.2.6. Chain Rule	19
* 1.2.7. Taylor Series Expansion and Mean Value Theorems	20
1.3. Lyapunov's Direct Method	21
1.4. Unconstrained Optimization Algorithms	23
* 1.4.1. Necessary and Sufficient Conditions for an Extremum	23
* 1.4.2. Dynamic Gradient Systems	25
* 1.4.3. Newton's Methods	25
* 1.4.4. The Quasi-Newton Methods	27
* 1.4.5. The Conjugate Gradient Method	28
1.5. Constrained Nonlinear Programming Problems	25
1.5.1. Kuhn-Tucker Conditions	25
1.5.2. Lagrange Multipliers and Kuhn-Tucker Conditions for Constrained Minimization with Mixed Constraints	31

Chapter 3. Unconstrained Optimization and Learning Algorithms	88
* 3.1. The Use of Systems of Ordinary Differential Equations in Unconstrained Optimization Problems – Trajectory-Following Methods	89
3.1.1. Basic Iterative Gradient Descent Algorithms	89
3.1.2. Continuous-Time Realization of Iterative Algorithms	91
3.1.3. Basic Gradient Systems	93
3.1.4. Continuous-Time Algorithm with Prespecified Convergence Speed	96
* 3.2. Unconstrained Optimization by Applying a System of Second-Order Differential Equations	102
3.3. Branin's Method	104
○ 3.4. Optimization Networks Using a Combination of Deterministic and Random Search – Stochastic Gradient Algorithms	107
○ 3.5. Boltzmann Machine and Simulated Annealing	113
○ 3.6. Mean-Field Annealing Algorithm	117
3.7. Back-Propagation Learning Algorithms	122
3.7.1. Learning of the Single Layer Perceptron	122
3.7.2. Standard Back-Propagation Algorithm for the Multilayer Perceptron	127
* 3.7.3. Back-Propagation Algorithm with Momentum Updating	132
* 3.7.4. Batch Learning Algorithm	133
* 3.7.5. Comparison of the On-Line and the Batch Procedures	135
3.7.6. Back-Propagation Algorithm with a Variable Number of Neurons in the Hidden Layers	136
* 3.8. Back-Propagation Algorithms with Non-Euclidean Error Signals	137
* 3.9. Speeding up the Back-Propagation Learning Algorithms	142
3.9.1. Back-Propagation Learning Algorithm with an Adaptive Slope of the Activation Functions	143
* 3.9.2. Search-Then-Converge Strategy	144
3.9.3. Averaging Procedure	145
* 3.9.4. Global Adaptation of the Learning Rate and/or Momentum Rate	146
* 3.9.5. Local Adaptation of Learning Rates	148
* 3.9.6. Quickprop	151
3.10. Generalized Learning Algorithm for a Single Neuron	152
3.10.1. Generalized LMS Learning Rule	154
3.10.2. Potential Learning Rule	156
3.10.3. Correlation Learning Rule	156
3.10.4. Hebbian Learning Rule	158
3.10.5. Oja's Learning Rule	158
3.10.6. Standard Perceptron Learning Rule	159
3.10.7. Generalized Perceptron Learning Rule	159
Questions and Problems for Chapter 3	160
3.11. References and Sources for Further Reading	162

Métodos de Otimização Contínua “clássica”

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

$$\underline{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

Propriedades

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

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

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

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

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

Otimização de F_0

não linear

não convexa

w ordem elevada

irrestrita

contínua, derivável

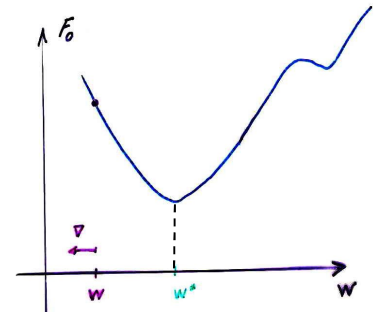
Como fazer ???

Noções Elementares de Otimização

I – Função objetivo – a minimizar

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

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



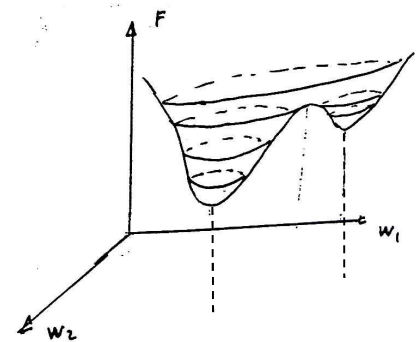
Mínimo, minimante $\underline{w}^*$

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

→ mínimo local

Mínimo global

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



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

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

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

considerando

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

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

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



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

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

Então

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

Aproximação linear ou de 1ª. ordem

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

Obs: Se $F(\underline{w})$ é linear a aproximação de 1ª. ordem é exata. $\underline{\nabla} F(\underline{w}) = \text{cte}(\underline{w})$ 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

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

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

Aproximação do Gradiente

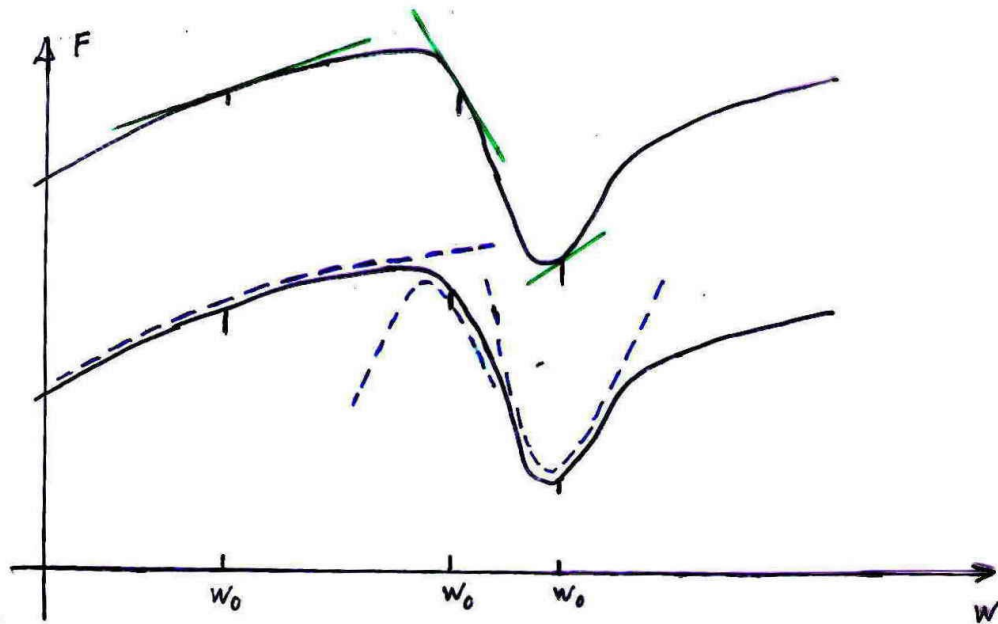
$$F(\underline{w}_0 + \underline{\Delta w}) \cong F(\underline{w}_0) + \sum_i \left. \frac{\partial F}{\partial w_i} \right|_{\underline{w}_0} \Delta w_i$$

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

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

Funções de ordem mais alta

validade das aproximações ?



A validade da aproximação

**depende do acréscimo na função (nas variáveis dependentes),
e não do acréscimo 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

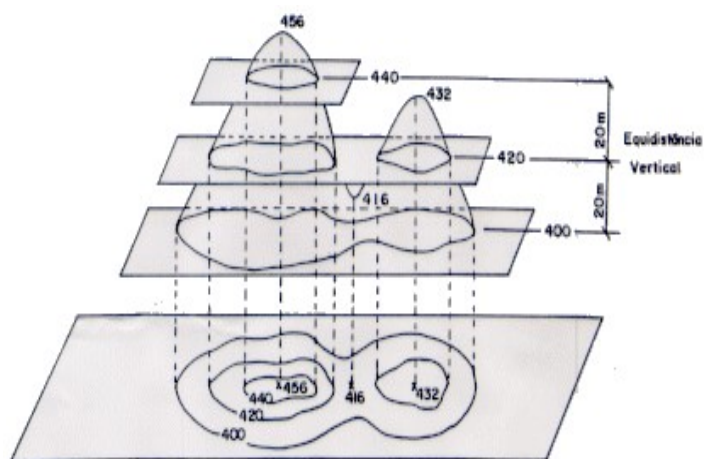
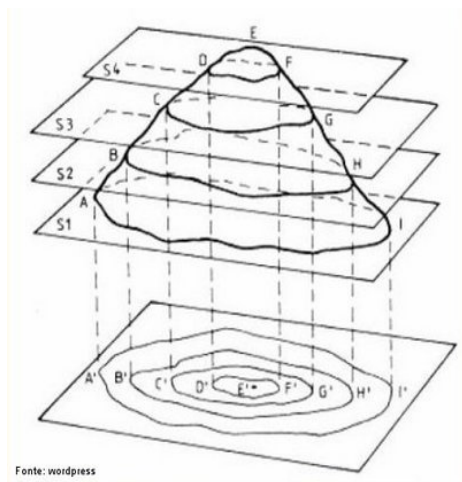
$$\left| \frac{1}{2} \underline{\Delta w}^t \underline{H}(\underline{w}_0) \underline{\Delta w} \right| \ll \varepsilon$$



Visualização de superfícies 3D em 2D

$h(x_1, x_2)$

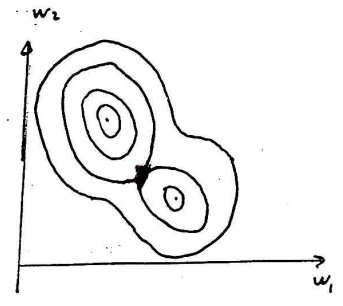
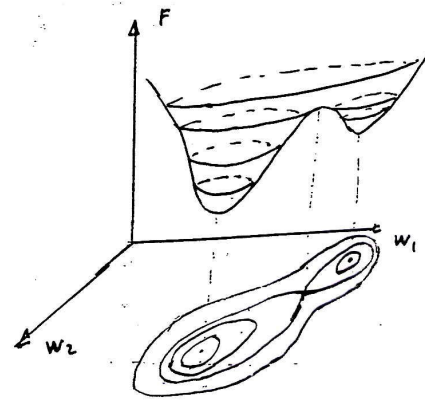
Curvas de nível



Funções de duas variáveis

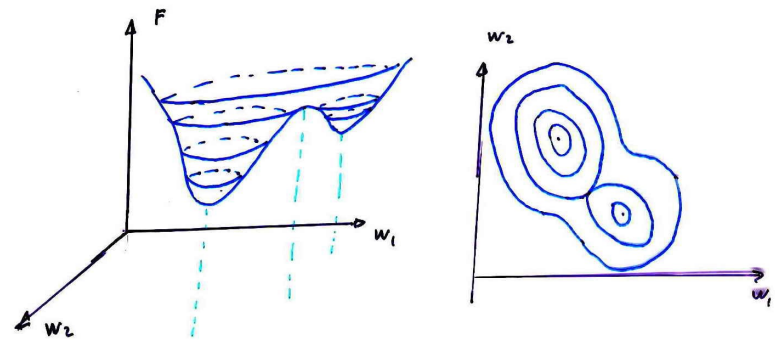
Visualização por

Curvas de Nível



Curvas /
Superfícies de nível

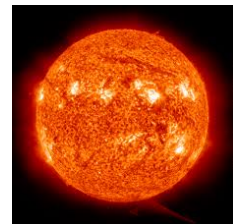
Tangentes



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

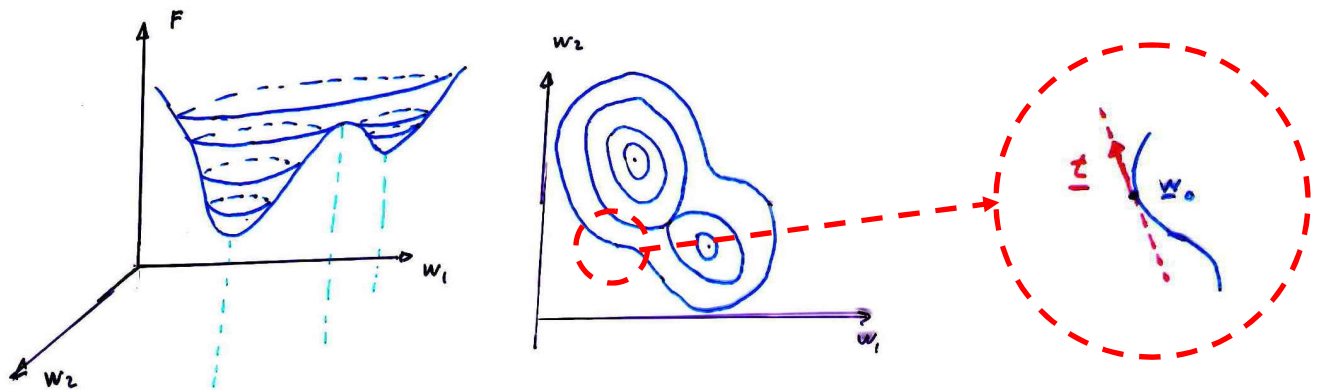


$\underline{w}$ (dim 3) $\rightarrow$ superfície de nível (dim 3) $\rightarrow$ plano t_g (dim 3)



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

Propriedade:



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

III – Extremos:

Mínimo, minimante $\mathbf{w}^*$ - local

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

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

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

+ + +/ -

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

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

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

Nos extremos

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

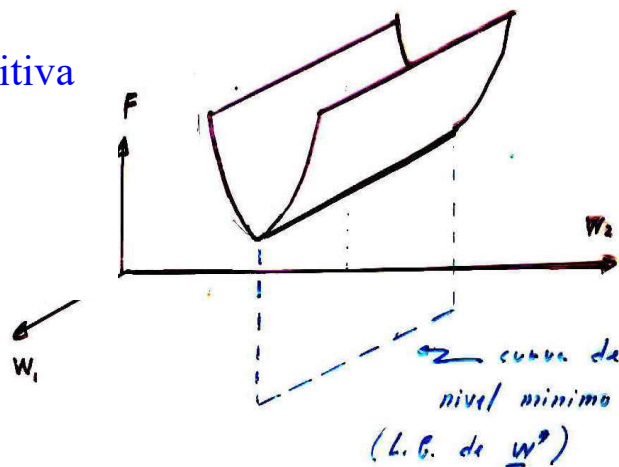
$$\begin{aligned} F(\underline{w}^* + \underline{\Delta w}) &= F(\underline{w}^*) + \Delta F \\ &= F(\underline{w}^*) + \frac{1}{2} \underline{\Delta w}^t \underline{H} \underline{\Delta w} \end{aligned}$$

$$\Delta F = \frac{1}{2} \underline{\Delta w}^t \underline{H} \underline{\Delta w} \quad \left\{ \begin{array}{ll} > 0 & \underline{H} \text{ definida positiva} \\ \geq 0 & \underline{H} \text{ semi-definida positiva} \\ \leq 0 & \underline{H} \text{ semi-definida negativa} \\ < 0 & \underline{H} \text{ definida negativa} \\ \pm & \underline{H} \text{ não definida} \end{array} \right.$$

H semidefinida positiva

$$\Delta F \geq 0$$

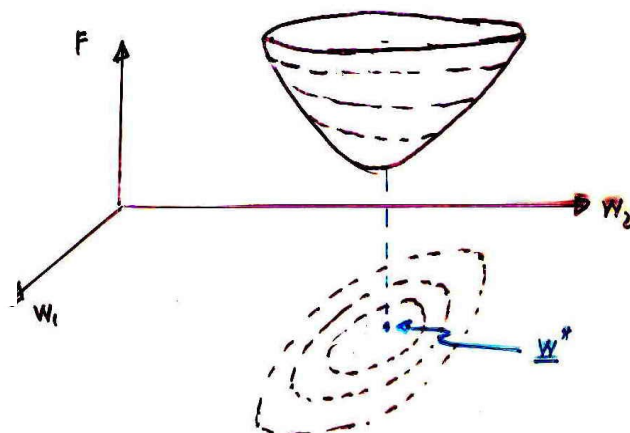
CALHA



H definida positiva

$$\Delta F > 0$$

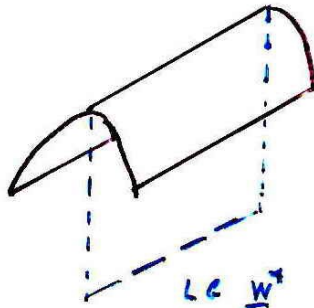
MÍNIMO
propriamente dito



H semidefinida
negativa

$$\Delta F \leq 0$$

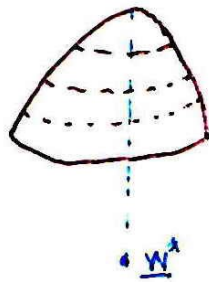
CUMEEIRA



H definida negativa

$$\Delta F < 0$$

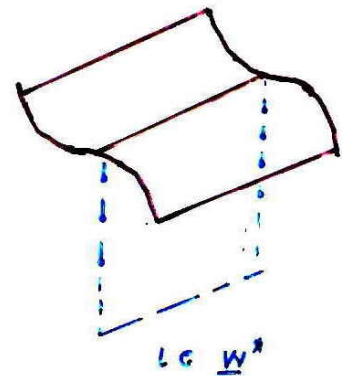
MÁXIMO



H não definida

$$\Delta F \geq 0 \text{ ou } \Delta F < 0$$

SELA



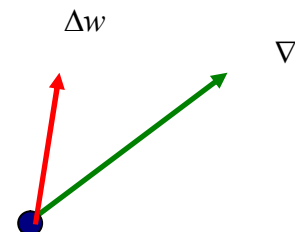
IV – Propriedades do Gradiente

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

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

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

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

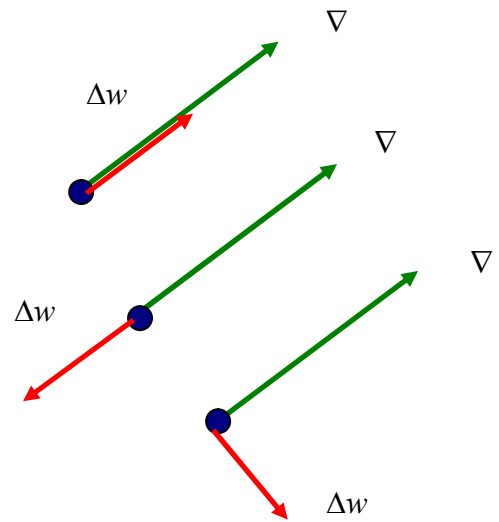


$$\Delta F \cong \underline{\Delta w}^t \nabla F = |\underline{\Delta w}| |\nabla F| \cos \theta$$

$$\theta = 0 \quad \longrightarrow \quad \Delta F \text{ max}$$

$$\theta = \pi \quad \longrightarrow \quad \Delta F \text{ min} \\ (-\Delta F \text{ max})$$

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



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

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

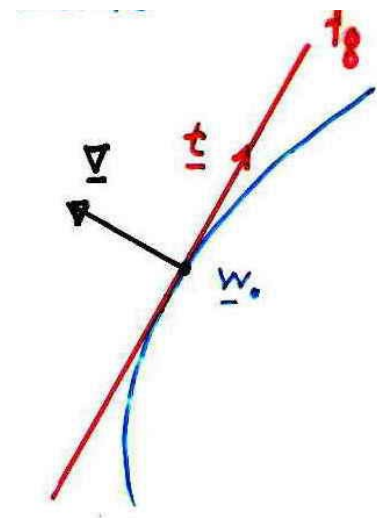
$$\left. \frac{d}{d\alpha} F(\underline{w}_0 + \alpha \underline{t}) \right|_{\forall \underline{t} \in \pi_{\epsilon}} = 0$$

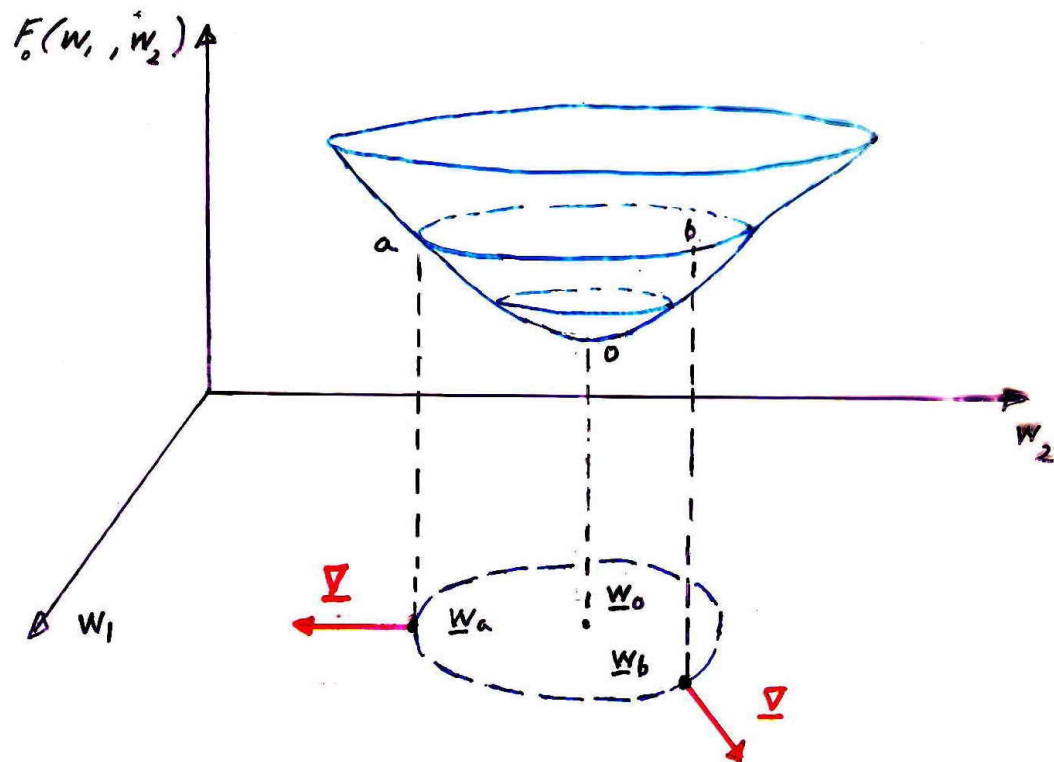
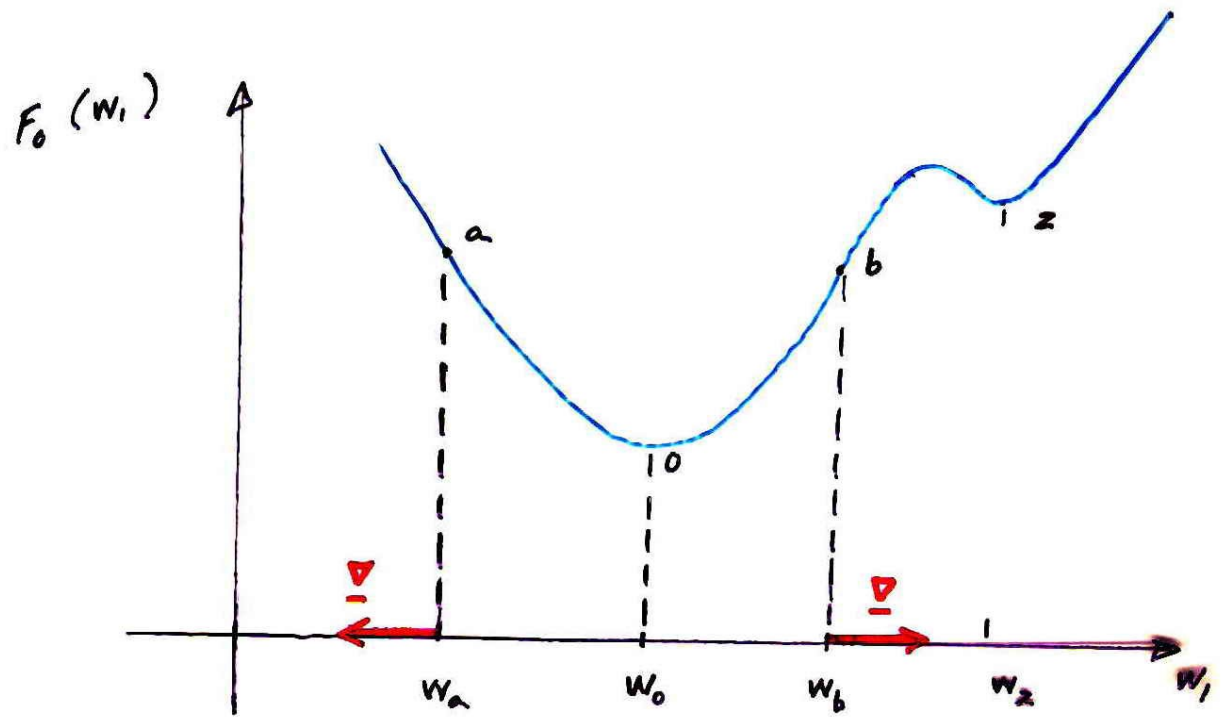
$$\Delta F \cong \underline{\Delta w}^t \nabla F$$

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

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

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





V. Otimização

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

VI – Otimização: Solução Analítica

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

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

A - Funções Lineares

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

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

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

B – Funções Quadráticas

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

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

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

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

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

Newton

Newton Raphson

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

Evitando $\underline{H}$

Pseudo Newton

$$\underline{H} \approx \text{diag } \underline{H} \quad \Rightarrow \quad \underline{H}^{-1}$$

Levenberg-Marquardt

Quase Newton

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

C – Funções de ordens mais elevadas e/ou transcendentais

muito complexo $\Rightarrow$ **Métodos Numéricos Recursivos**

VII – Otimização: Métodos Numéricos Recursivos

$$\underline{w} \rightarrow \underline{w} + \underline{\Delta w} \quad | \quad F(\underline{w} + \underline{\Delta w}) < F(\underline{w})$$

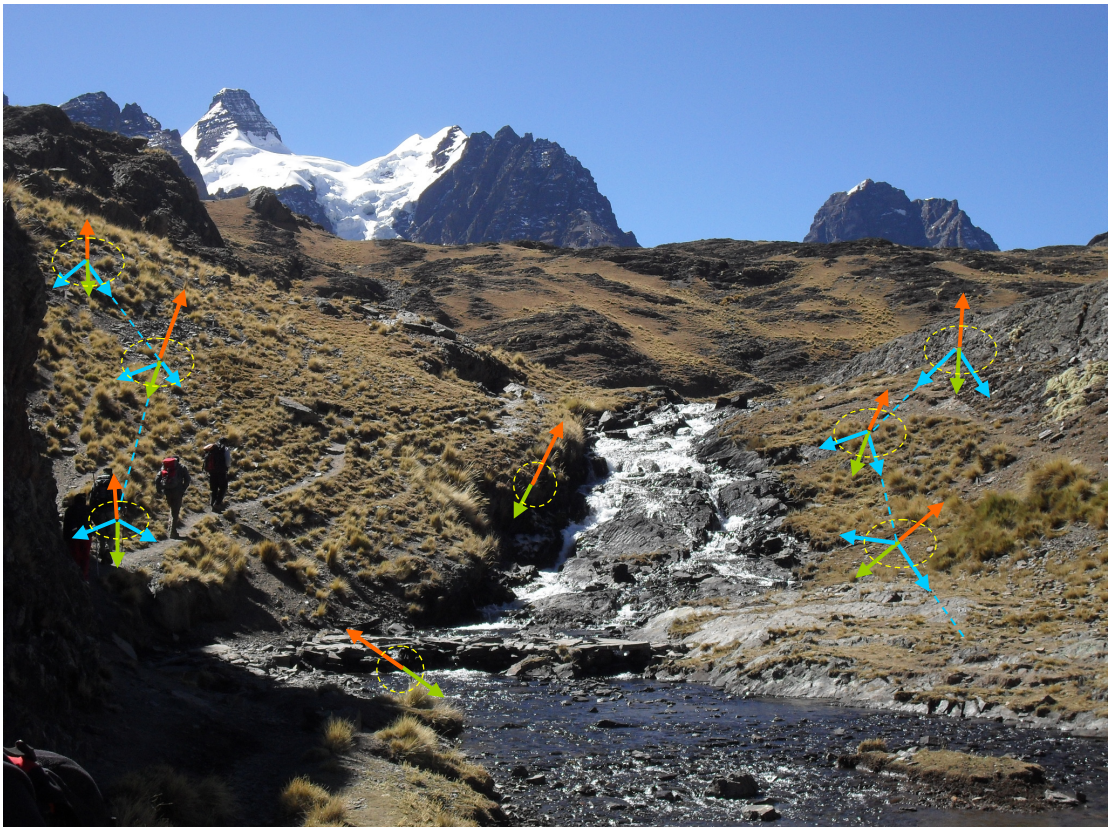
$\underline{\Delta w}$??

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

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

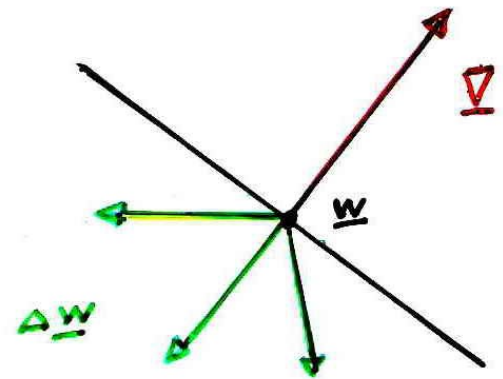
$$F(\underline{w} + \Delta \underline{w}) \approx F(\underline{w}) + \Delta \underline{w}^t \nabla(\underline{w})$$

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

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

direção $\underline{d}$

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



Direção mais eficaz ?

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

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

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

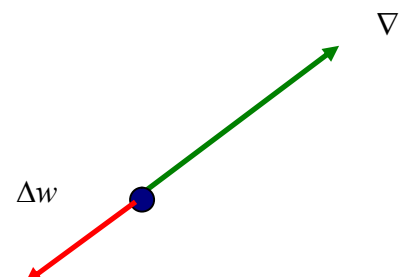
$$\Delta F = \Delta \underline{w}^t \nabla = \|\Delta \underline{w}\| \|\nabla\| \cos \angle \underline{d}, \nabla$$

constante

$(-1, 1)$

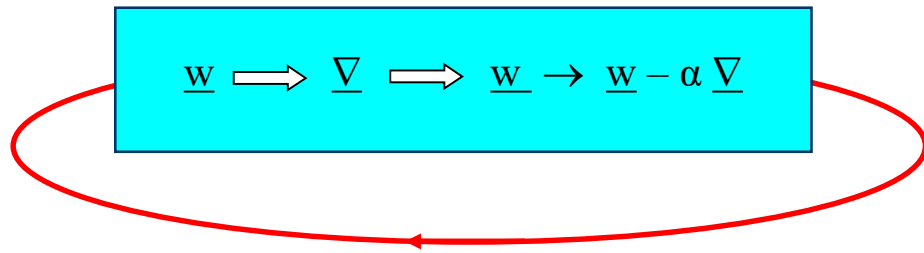
$$\underline{d} = -\nabla$$

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



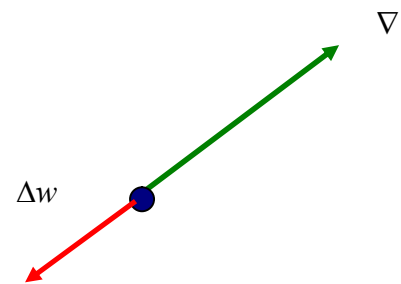
Método do Gradiente Descendente ou

Método da Descida pelo Gradiente

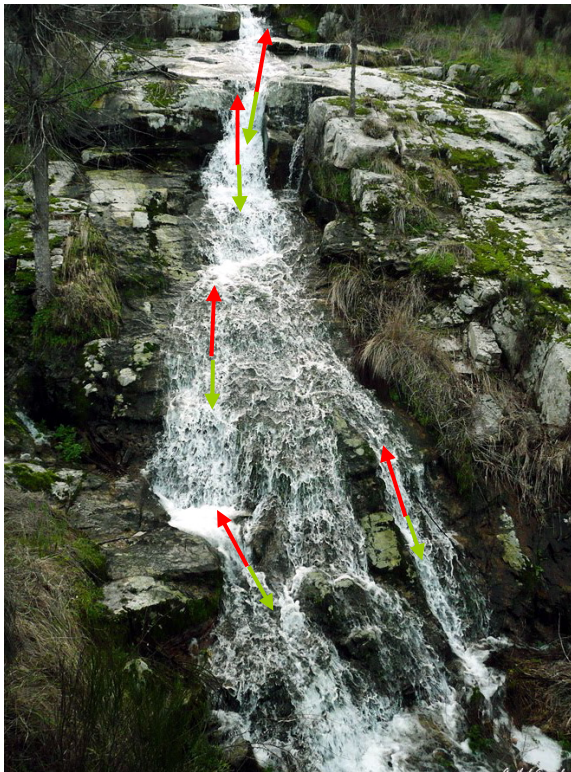


$$\Delta \underline{w} = -\alpha \underline{\nabla}$$

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



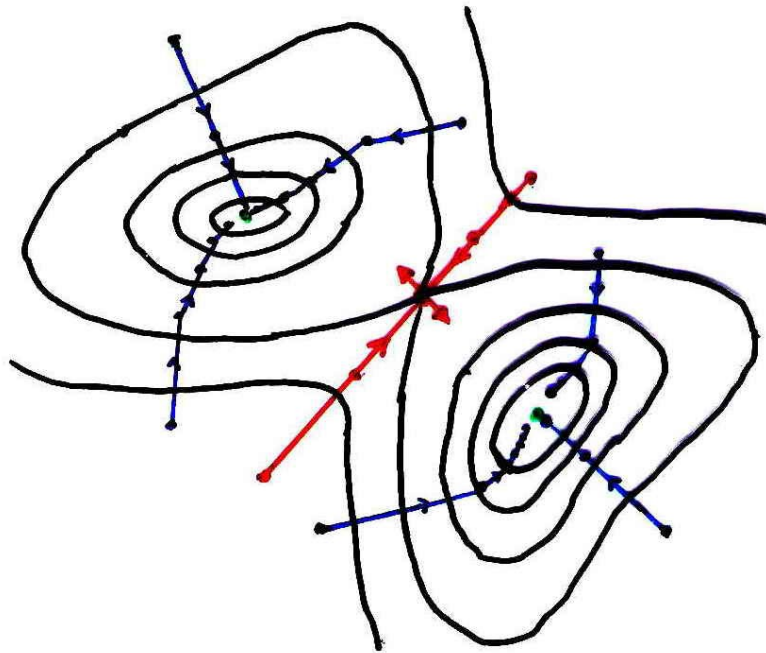
Gradiente Descendente



Método do

Gradiente Descendente

$$\underline{w} \rightarrow \underline{\nabla} \rightarrow \underline{w} \rightarrow \underline{w} - \alpha \underline{\nabla}$$



Algoritmo

Até que o critério de parada seja satisfeito

Para cada variável w_i

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

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

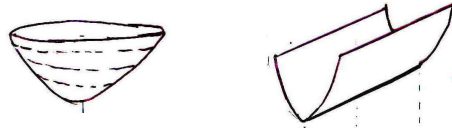
Fim (natural) do processo

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

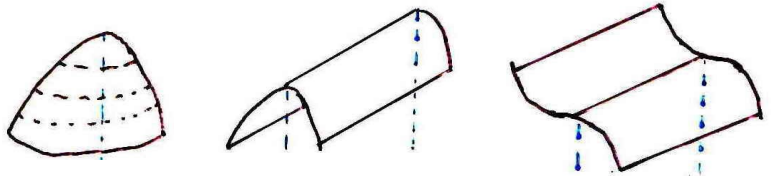
$$E(\Delta \underline{w}) = \underline{0} \quad \Rightarrow \quad \nabla(\underline{w}_0) = \underline{0}$$

O que é $\underline{w}_0$??

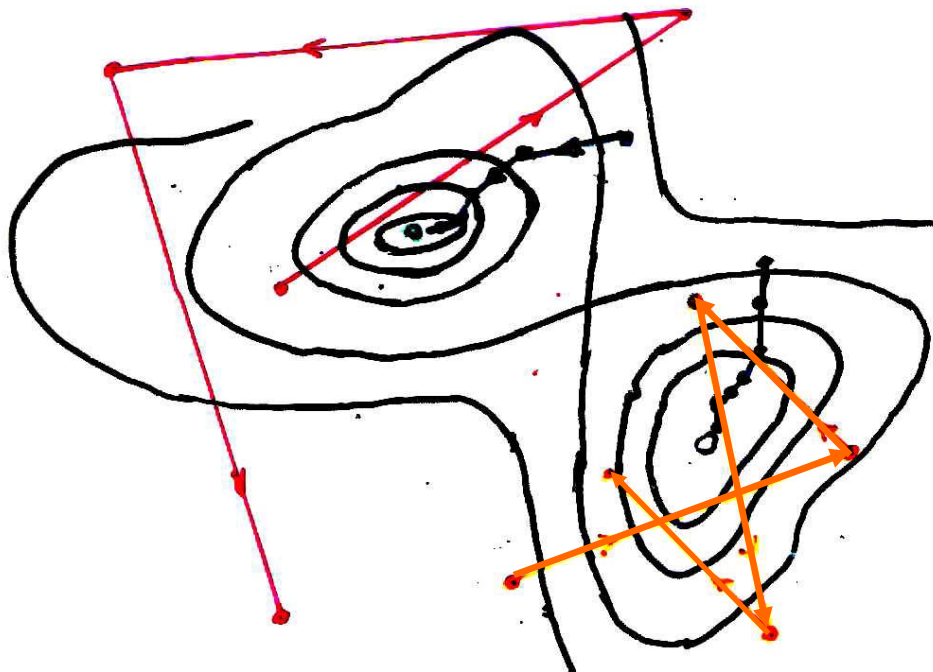
pontos estáveis: mínimo, calha



pontos instáveis: máximo, cumeeira, sela

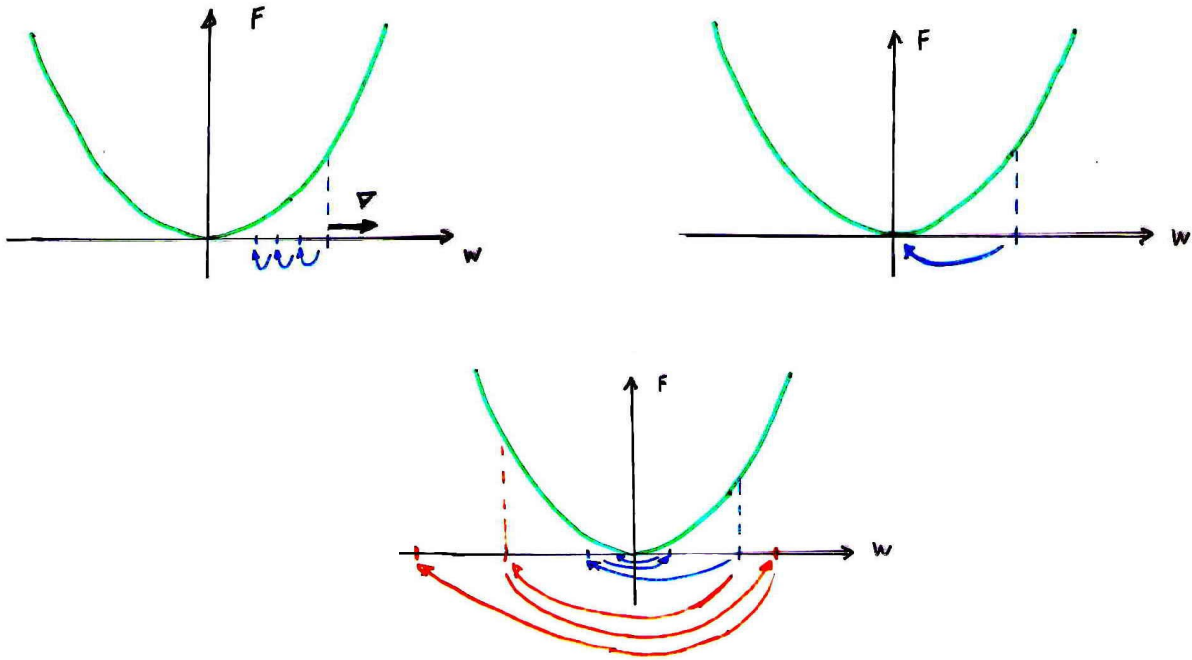


Passo de Treinamento α

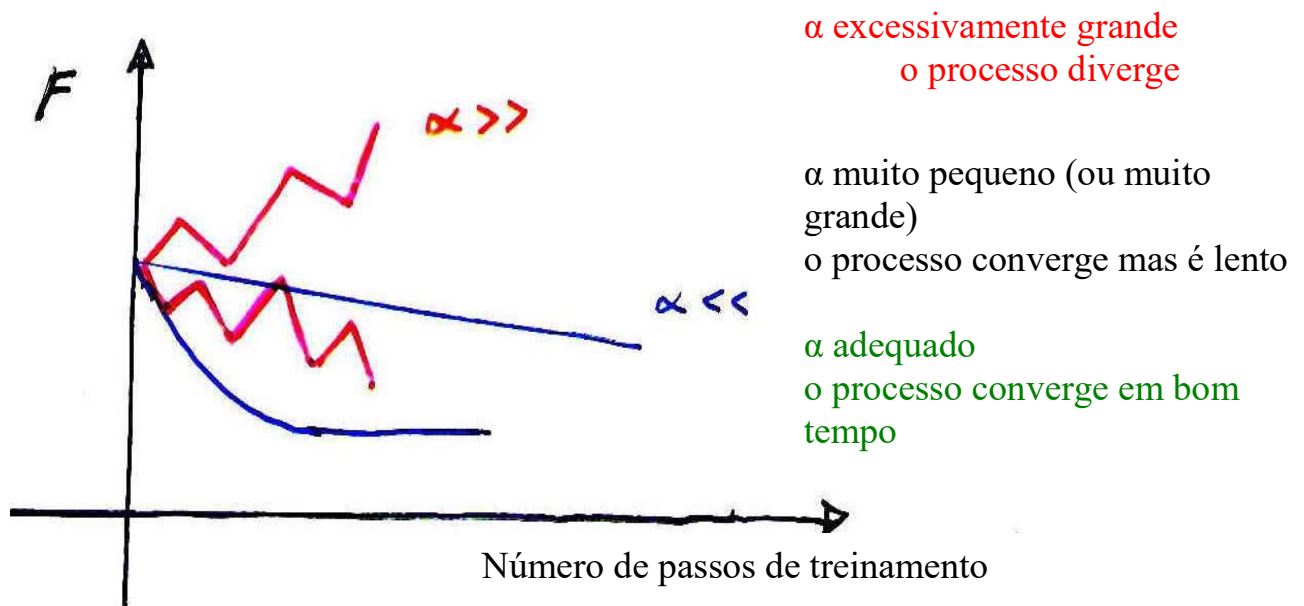


Efeito do passo de treinamento α

Exemplo simples: $F = w^2$



Evolução do erro ao longo do treinamento:



Passo de Treinamento α

$\alpha \ll$ convergência lenta

$\alpha \gg$ F_0 oscila muito, convergência lenta, **pode divergir**

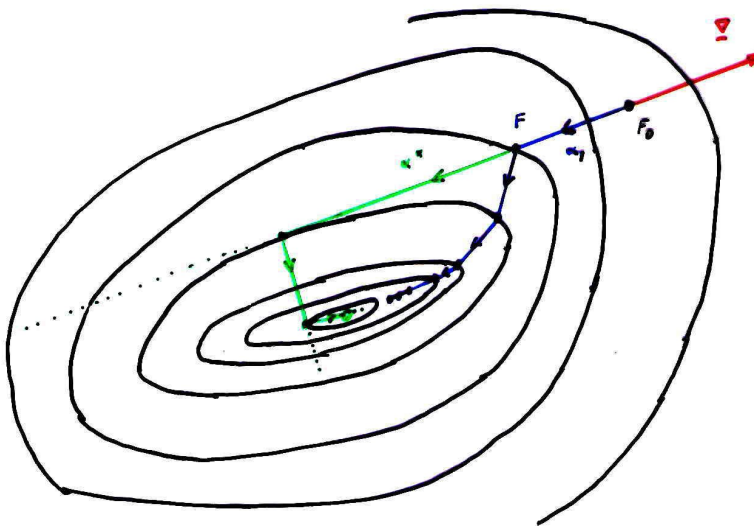
$\Rightarrow \alpha_{\text{ótimo}} ??$

“a escolha de α é uma arte” – Bernard Widrow

No treinamento de redes neurais

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

*Passo Ótimo de Treinamento α^**



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

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

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

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

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

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

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

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

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

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

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

necessita de $\underline{H}$!

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

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

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

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

aplicando (3) em (2)

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

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

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

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 $\alpha^* \gg \alpha_1$ possivelmente o α^* encontrado não é válido.

No caso em que

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

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

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

Passo variável controlando o decréscimo de F

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

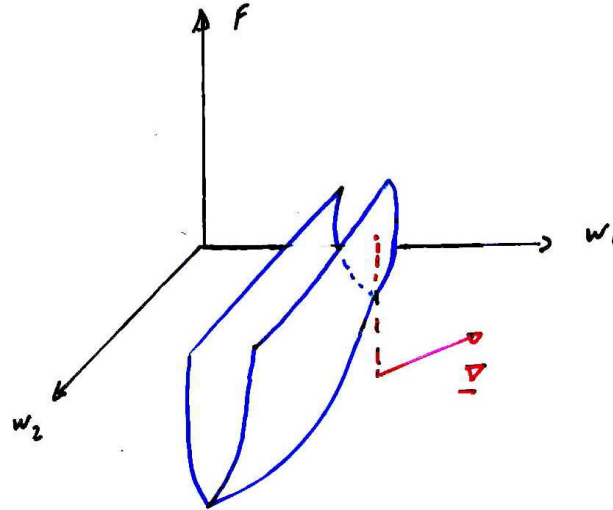
$$\alpha = \alpha_0 \quad \text{ctte} \quad \Rightarrow \quad \Delta F = - \alpha \|\underline{\nabla}\|^2$$

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

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

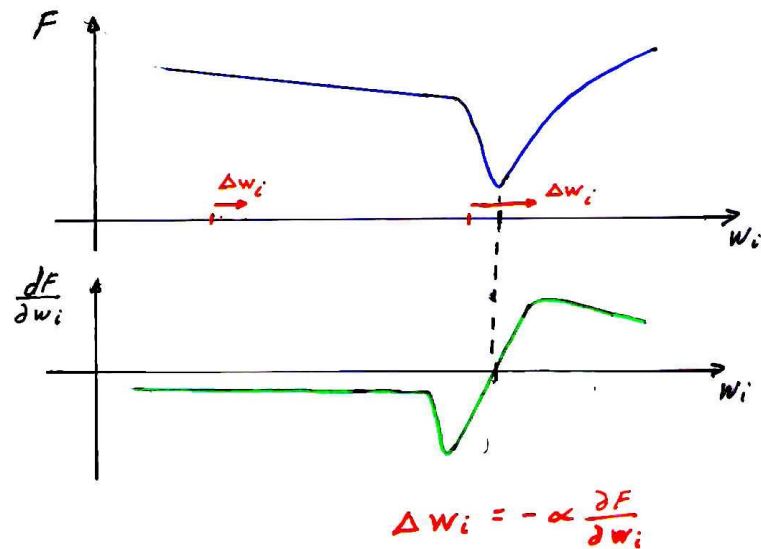
Acelerando o processo – passo variável

α diferenciado por variável



Acelerando o processo – passo variável

α diferenciado por passo



Resilient Backpropagation (modificado)

(Silva e Almeida, Super SAB, etc.)

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

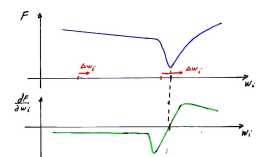
$$\Delta w_i(n) = -\alpha_i(n) \frac{\partial F}{\partial w_i}(n)$$

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

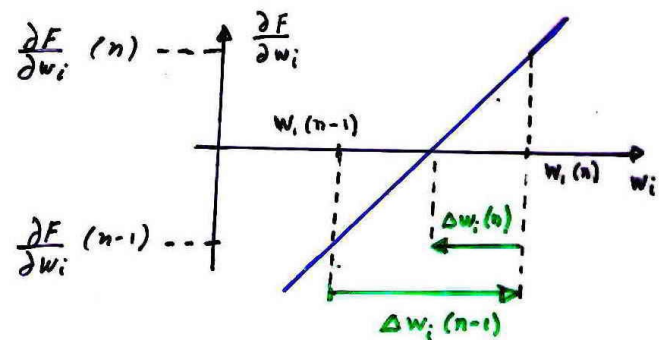
se $\alpha_i(n) > \alpha^*$ faça $\alpha = \alpha^*$

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

Outra alternativa (passo ótimo)



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



Algoritmo Resiliente

Inicialização

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

inicializar as $i = 1, 2, \dots, V$ variáveis

Até que o critério de parada esteja satisfeito

Aplicar um passo de treinamento

Para cada parâmetro $w_i \quad i = 1, 2, \dots, V$

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

Calcular o passo α a ser usado

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

Calcular e aplicar os acréscimos

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

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

$$n = n + 1$$

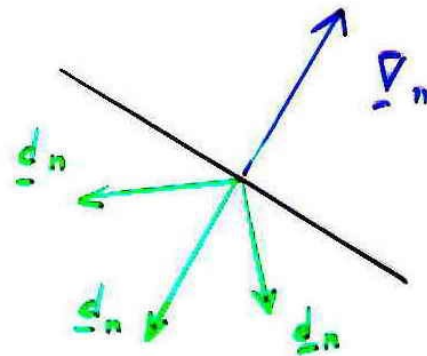
Próximo passo

Métodos de 1ª. Ordem (Resumo)

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

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

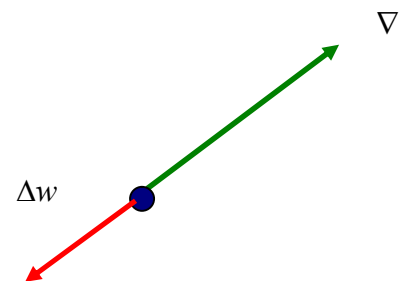
$$\underline{d}_n^t \nabla_n < 0$$



Gradiente Descendente

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

$$\underline{d}_n = - \nabla_n \quad \alpha_n = \alpha_0$$



Passo ótimo

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

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

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

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

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

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

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

Newton Amortecido

Levenberg-Marquardt

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

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

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

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

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

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

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

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

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

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

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

Uma solução

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

$$\lambda > 0 \quad \lambda \rightarrow 0$$

Uma solução melhor

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

$$0 \leq \lambda \leq \frac{|h_{ii}|}{\sum_{\substack{j=1 \\ j \neq i}}^n |h_{ij}|} - 1 \quad \forall i = 1, \dots, n \quad \lambda \rightarrow 0$$

Levenberg – Marquardt

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

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

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

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

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

Levenberg Marquadt - Algoritmo

Inicialização

$n = 1$; inicializar as w_i $i = 1, 2, \dots, V$ variáveis

Até que o critério de parada esteja satisfeito

Passo de treinamento n

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

$$\text{calcular} \quad \frac{\partial F}{\partial w_i}(n) = g_i(n) = E_p \frac{\partial \varepsilon^p}{\partial w_i}$$

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

Cálculo e aplicação do passo Δw

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

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

realize a otimização em linha

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

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

$$n = n + 1$$

Quasi Newton

Fórmulas básicas:

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

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

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

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

BFGS – Broyden – Fletcher – Goldfarb – Shanno

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

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

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

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

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

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

Gradiente Conjugado - não usa H

Fletcher – Reeves

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

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

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

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

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)

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

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

$\underline{d}_n$ depende do método

