4 - Métodos mais Eficientes

Otimização por 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 a
direção de $\underline{\Delta w}$

Aproximação de uma função (multivariável) por Série de Taylor

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

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

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

considerando

acréscimo **gradiente** **Hessiana**

$\underline{\Delta w} = \left.\Delta w_i\right]$ $\nabla F(\underline{w_0}) = \left[\left.\frac{\partial F}{\partial w_i}\right|_{\underline{w_0}}\right]$ $\underline{H}(\underline{w}_0) = \left[\left.\frac{\partial^2 F}{\partial w_i\, \partial w_j}\right|_{\underline{w_0}}\right]$

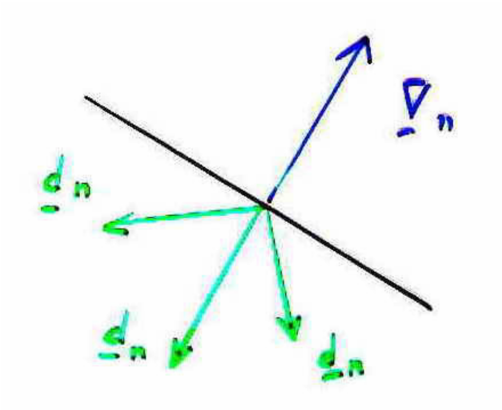
$$F(\underline{w}_0 + \underline{\Delta w}) \approx F(\underline{w}_0) + \underline{\Delta w}^t \underline{\nabla} F(\underline{w}_0) + \frac{1}{2} \underline{\Delta w}^t\, \underline{H}(\underline{w}_0)\, \underline{\Delta w}$$

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

Métodos de 1ª Ordem (redução da função)

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

$\Delta F \approx \underline{\Delta w}^t \underline{\nabla} = \alpha\, \underline{d}^t \underline{\nabla}$



Para $\Delta F < 0$

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

Máximo decréscimo por passo

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

BP Resiliente (melhor)

Métodos de Segunda ordem (redução de gradiente):

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

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

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

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

$$\alpha = \underset{\alpha > 0}{\mathrm{Arg}} \left[\mathrm{Min} \quad F(\underline{w}_0 + \alpha\,\underline{d})\right] \quad \text{(otimização em linha)}$$

Problemas: $\underline{H}(\underline{w}_0) = \left[\left. \dfrac{\partial^2 F}{\partial w_i\, \partial w_j} \right|_{\underline{w}_0} \right]$ e $\underline{H}^{-1}$

Métodos de 2ª Ordem $\underline{d} = -\underline{H}^{-1}\underline{\nabla}$

Newton

Newton Amortecido

Levemberg-Marquadt

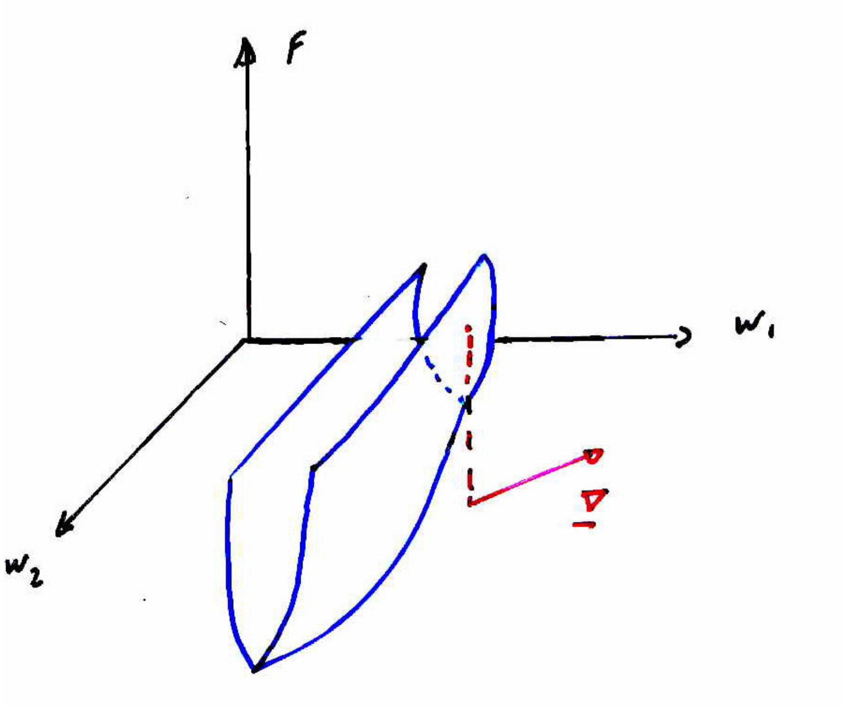
Quasi Newton

Gradiente Conjugado

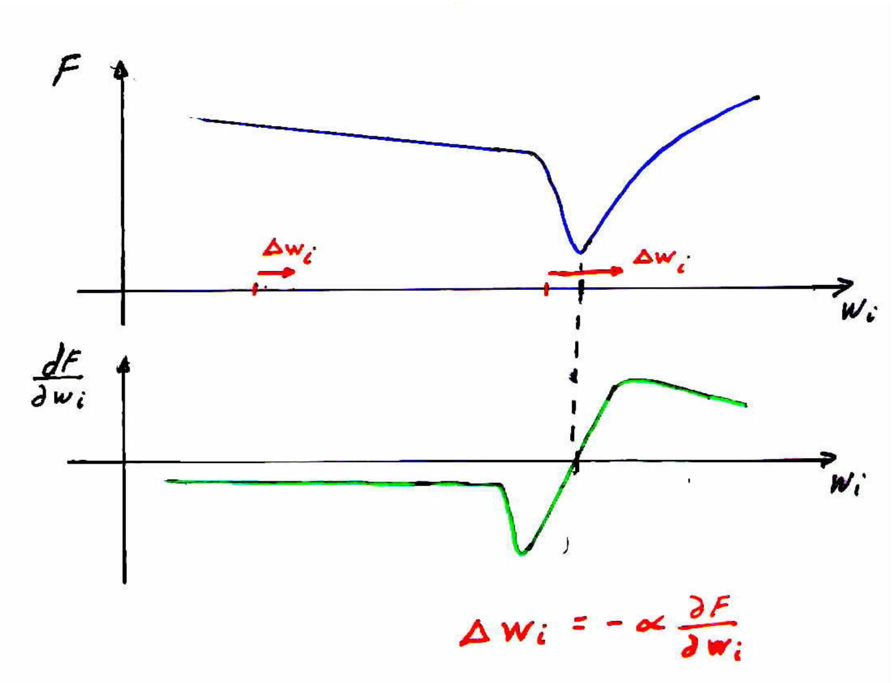
Métodos de Primeira Ordem Aprimorado

Resilient Backpropagation

Acelerando o processo - passo variável por sinapse



Acelerando o processo – passo variável no tempo



Resilient Backpropagation (modificado)

(Silva e Almeida, Super SAB, etc.)

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

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

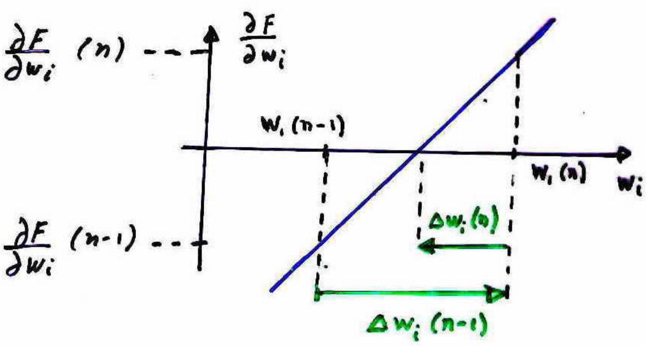
$$\alpha_i(n) = \begin{cases} a\,\alpha_i(n-1) & \text{se } \operatorname{sign}\frac{\partial F}{\partial w_i}(n) = \operatorname{sign}\frac{\partial F}{\partial w_i}(n-1) \\ b\,\alpha_i(n-1) & \text{se } \operatorname{sign}\frac{\partial F}{\partial w_i}(n) \neq \operatorname{sign}\frac{\partial F}{\partial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

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

$\alpha_i(0) \approx .05$ $\alpha^* \approx .3$ $a \approx 1.05$ $b \approx .9$

Outra alternativa

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



BP Resiliente - Algorítmo

Inicializar: $n = 1; \quad \alpha_i(1) = .1; \quad \alpha_{\max} \approx .5; \quad a \approx 1.05; \quad b \approx .9; \quad \frac{\partial F}{\partial w_i}(0) = 0^+;$

$w_i(1) \in [+.2, -.2] \quad randômicos;$

Até que o critério de parada esteja satisfeito

Passo de treinamento n

Cálculo das derivadas por sinapse

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

calcular $\underline{\tilde{y}}^p = \varphi(\underline{x}^p) \qquad \varepsilon_m^p = y_m^p - \tilde{y}_m^p$ e $\frac{\partial \tilde{y}_m^p}{\partial w_i} \quad \forall\, m, i$

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

outro par

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

Cálculo do acréscimo

para i = 1, 2, …

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

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

Atualização das sinapses

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

n = n + 1

fim do algorítmo

Método de Segunda Ordem

Levenberg – Marquadt (Newton Amortecido)

Aproximação para H

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

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

$$\underline{H} = \left[\frac{\partial^2 F}{\partial w_i \, \partial w_j}\right] \cong \underline{\tilde{H}} = \left[\sum_k \frac{\partial \varepsilon_k}{\partial w_i} \cdot \frac{\partial \varepsilon_k}{\partial w_j}\right]$$

Nova aproximação para garantir a inversão

$$\underline{H} \cong \underline{\tilde{H}} \cong \underline{\tilde{\tilde{H}}} = \left[\underline{\tilde{H}} + \lambda \underline{I}\right] \quad \Longrightarrow \quad \underline{H}^{-1} \approx \underline{\tilde{\tilde{H}}}^{-1}$$

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

λ deve ser pequeno para garantir a aproximação, mas grande o suficiente para garantir a inversão de $\underline{\tilde{\tilde{H}}}$. Por exemplo, deve ser tal que $\underline{\tilde{\tilde{H}}}$ seja diagonal dominante, ou então que tenha todos os seus autovalores com a mesma polaridade. λ deve ir sendo reduzido ao longo do tempo, a medida em que o ponto de operação se move para regiões onde a aproximação quadrática da função no entorno do mínimo é mais precisa.

Levenberg- Marquadt - Algorítmo

Inicializar: $n = 1; \quad w_i(1) \in [+.2, -.2]$ *randômicos*;

Até que o critério de parada esteja satisfeito

Passo de treinamento n

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

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

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

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

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

outro par

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

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

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

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

n = n + 1

fim do algorítmo

Obs sobre a otimização em linha

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

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

$$\Delta F \approx \alpha\, \underline{d}^t\, \underline{\nabla} + \frac{1}{2}\, \alpha^2\, \underline{d}^t\, \underline{H}\, \underline{d} \qquad \text{eq.(1)}$$

$$\alpha^* = \underset{\alpha^*>0}{\mathrm{Arg}} \left[\mathrm{Min}\ F(\underline{w}_0 + \alpha^* \underline{d})\right]$$

$$\frac{\partial F}{\partial \alpha} = 0 \qquad \alpha^* = -\frac{\underline{d}^t\, \underline{\nabla}}{\underline{d}^t\, \underline{H}\, \underline{d}} \qquad \text{eq. (2)}$$

necessita de $\underline{H}$!

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

α_1 arbitário

$$F(\underline{w}_0 + \alpha_1\, \underline{d}) \approx F(\underline{w}_0) + \alpha_1\, \underline{d}^t\, \underline{\nabla} + \frac{1}{2}\, \alpha_1^{\,2}\, \underline{d}^t\, \underline{H}\, \underline{d}$$

$$F_0 = F(\underline{w}_0) \qquad F_1 = F(\underline{w}_0 + \alpha_1\, \underline{d})$$

$$\underline{d}^t\, \underline{H}\, \underline{d} = \frac{2}{\alpha_1^2}\left(F_1 - F_0 - \alpha_1 \underline{d}^t\, \underline{\nabla}\right)$$

aplicando nas eq. (2) e (1)

$$\alpha^* = -\frac{\alpha_1^2\, \underline{d}^t\, \underline{\nabla}}{2\left(F_1 - F_0 - \alpha_1 \underline{d}^t\, \underline{\nabla}\right)} \qquad e \qquad \Delta F = \frac{1}{2}\underline{\alpha^* d}^t\, \underline{\nabla}$$

Obs:

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

2 - É suposto que a aproximação quadrática vale para α_1 e α^*.
Se $\alpha^* >> \alpha_1$ possivelmente o α^* encontrado não é válido.

3 - Se o valor de F medido no ponto $\underline{w}_0 + \alpha^* \underline{d}$ confirmar o valor predito pela fórmula $F(\underline{w}_0 + \alpha^* \mathrm{d}) = F(\underline{w}_0) + \frac{1}{2} \underline{\alpha^* \mathrm{d}}^t \underline{\nabla}$ então o valor encontrado para α^* provavelmente é válido.