# Aprendizado do MLP por Error Back Propagation ...

$$\Delta\vec{W} = -\eta.\vec{\nabla}Eqm\_$$

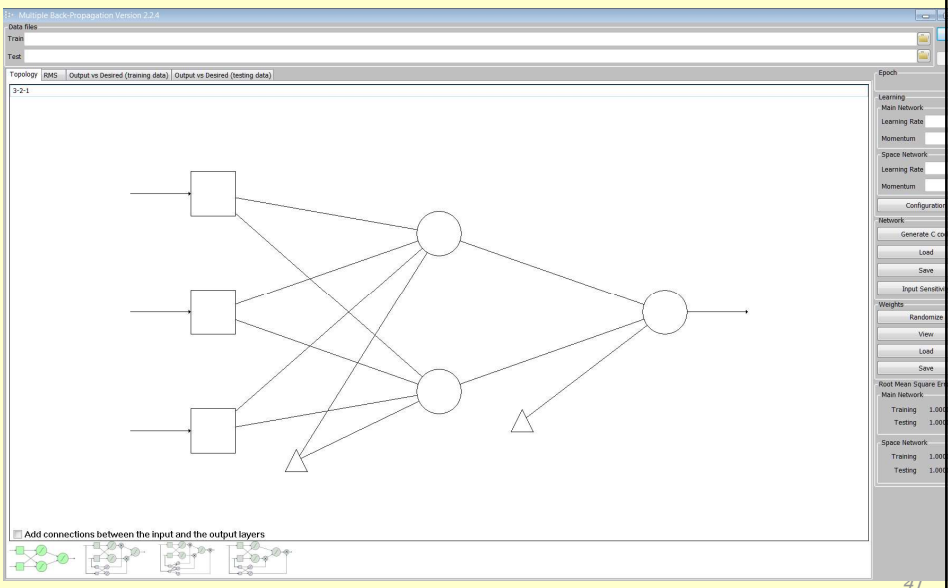
Gradiente de Eqm no espaço de pesos = ($\partial Eqm(W)/\partial w_1$ , $\partial Eqm(W)/\partial w_2$, $\partial Eqm(W)/\partial w_3$, ... )

# Chegando às fórmulas das derivadas parciais, necessárias à Bússola do Gradiente



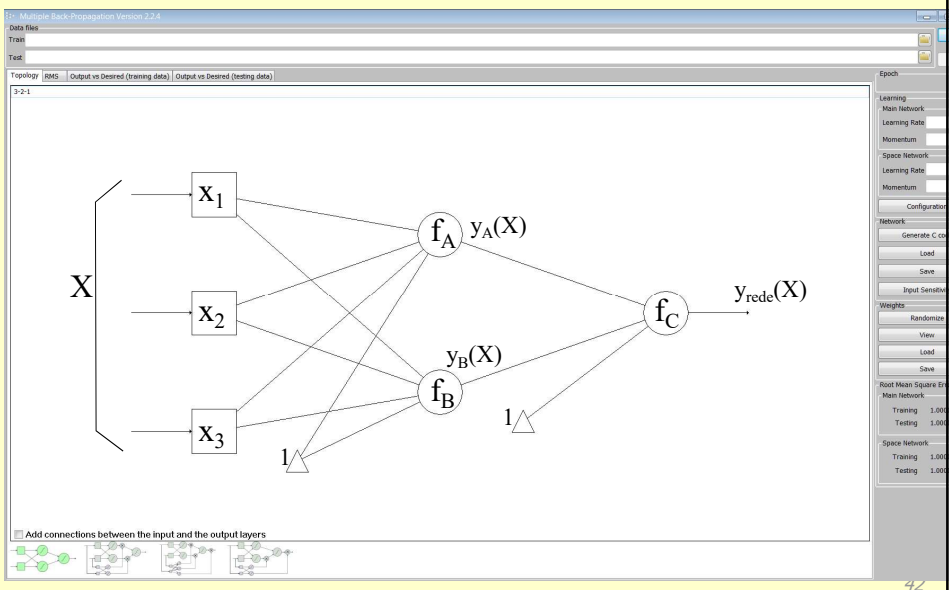
© Prof. Emilio Del Moral – EPUSP

*Relembrando o que está por trás de um desenho como o que segue ...*



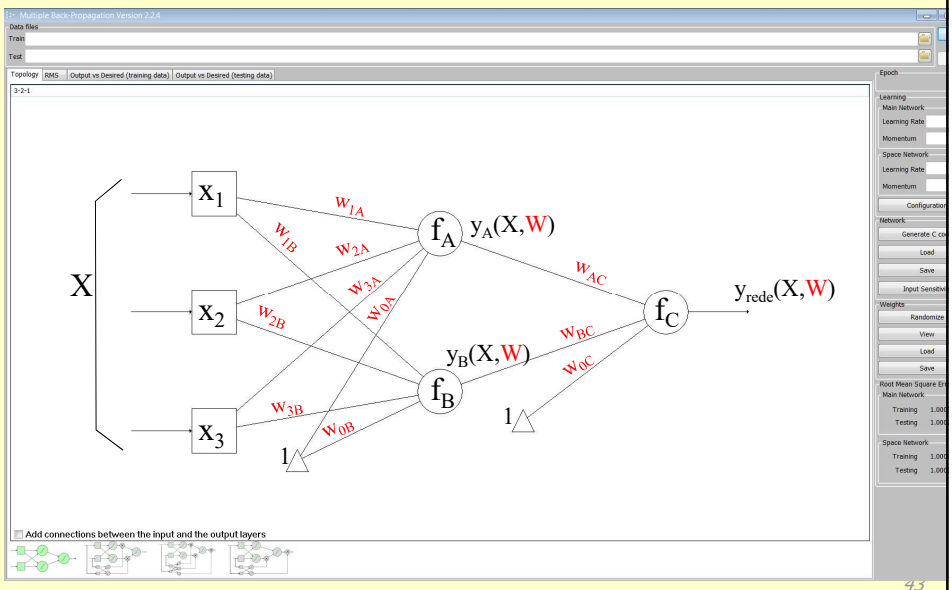



© Prof. Emilio Del Moral – EPUSP



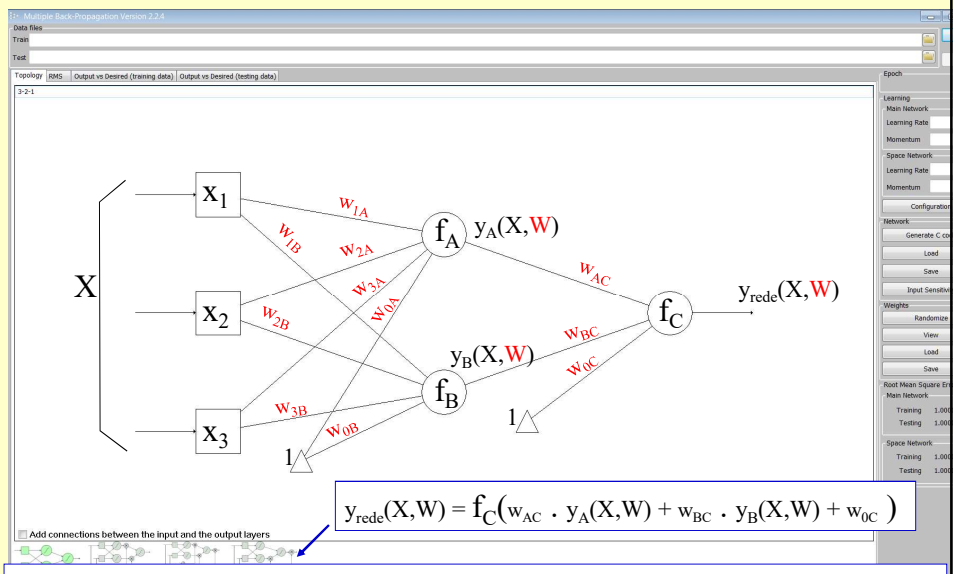



© Prof. Emilio Del Moral – EPUSP





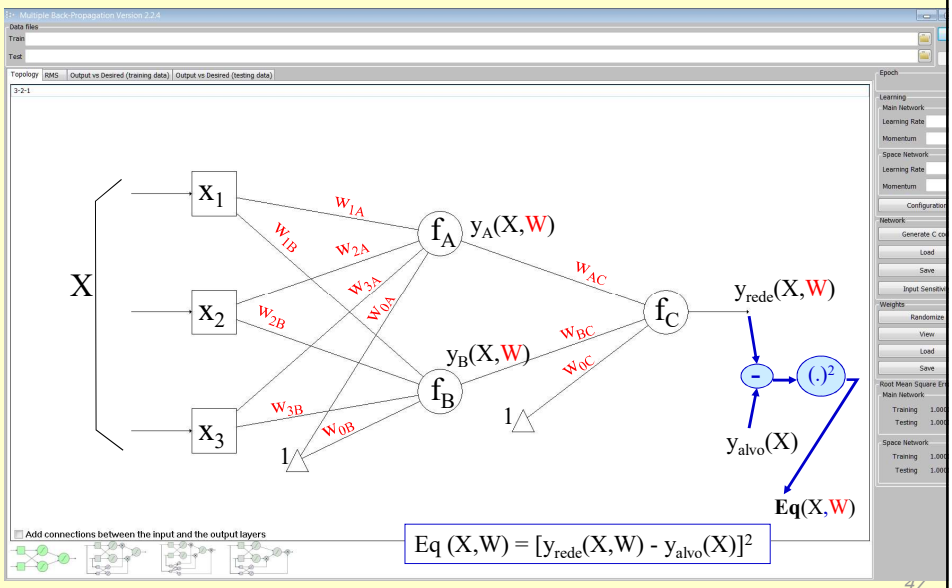

© Prof. Emilio Del Moral – EPUSP




$$y_{rede}(X,W) = f_C\left(w_{AC} \cdot y_A(X,W) + w_{BC} \cdot y_B(X,W) + w_{0C}\right)$$

$$y_{rede}(X,W) = f_C\left(w_{AC} \cdot f_A(w_{1A}.x_1 + w_{2A}.x_2 + \ldots + w_{0A}) + w_{BC} \cdot f_B(w_{1B}.x_1 + w_{2B}.x_2 + \ldots + w_{0B}) + w_{0C}\right)$$

© Prof. Emilio Del Moral – EPUSP




$$Eq(X,W) = [y_{rede}(X,W) - y_{alvo}(X)]^2$$

© Prof. Emilio Del Moral – EPUSP

**Chamada oral sobre a lição de casa: estudar / reestudar os conceitos e a parte operacional de derivadas parciais, do vetor Gradiente ...**

- **Derivadas parciais (que são as componentes do gradiente):**

$\partial f(a,b,c)/\partial a \qquad \partial f(a,b,c)/\partial b \qquad \partial f(a,b,c)/\partial c$

- **Vetor Gradiente, útil ao método do máximo declive:**

$(\partial Eqm(W)/\partial w_1\ ,\ \partial Eqm(W)/\partial w_2,\ \partial Eqm(W)/\partial w_3,\ \ldots)$ $\qquad \Delta\vec{W} = -\eta.\vec{\nabla}Eqm\_$



© Prof. Emilio Del Moral – EPUSP

# *Invertamos o operador gradiente e a somatória*

*.. afinal, gradiente é uma derivada, e a derivada de um soma de várias funções é igual à soma das derivadas individuais de cada componente da soma:*

$$Grad(Eqm) =$$

$$Grad(\ \Sigma_\mu\ Eq^\mu) / M$$

$$\Sigma_\mu\ Grad\ (Eq^\mu) / M$$



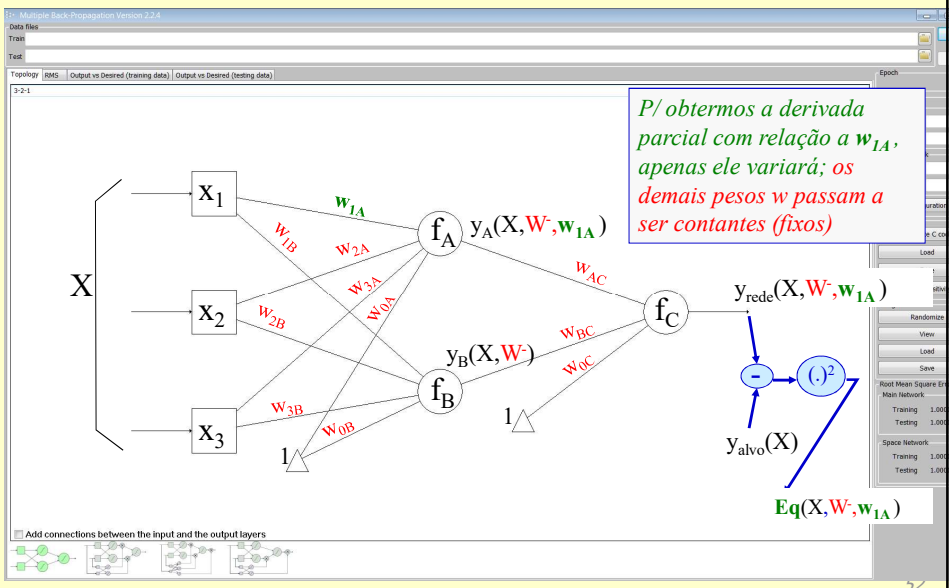
© Prof. Emilio Del Moral – EPUSP

Note que a inversão do gradiente com a somatória nada mais é que usar de forma repetida – e em separado para cada dimensão do vetor $\boldsymbol{Grad(\Sigma_\mu Eq^\mu)}$ – a seguinte propriedade simples e sua velha conhecida ...

$$\mathbf{d(f_1(x)+f_2(x)) / dx \;\; = \;\; df_1(x)/dx + df_2(x)/dx}$$

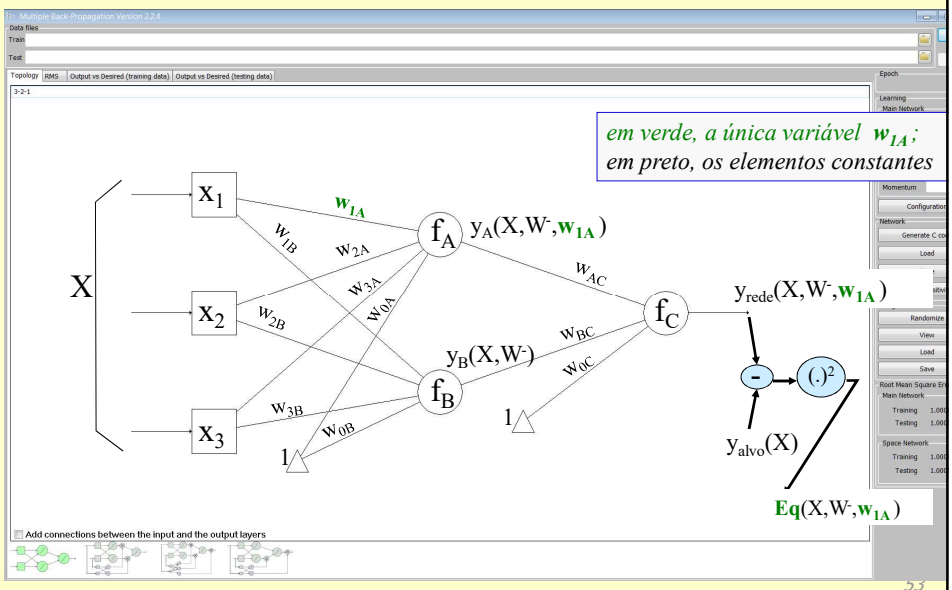


© Prof. Emilio Del Moral – EPUSP

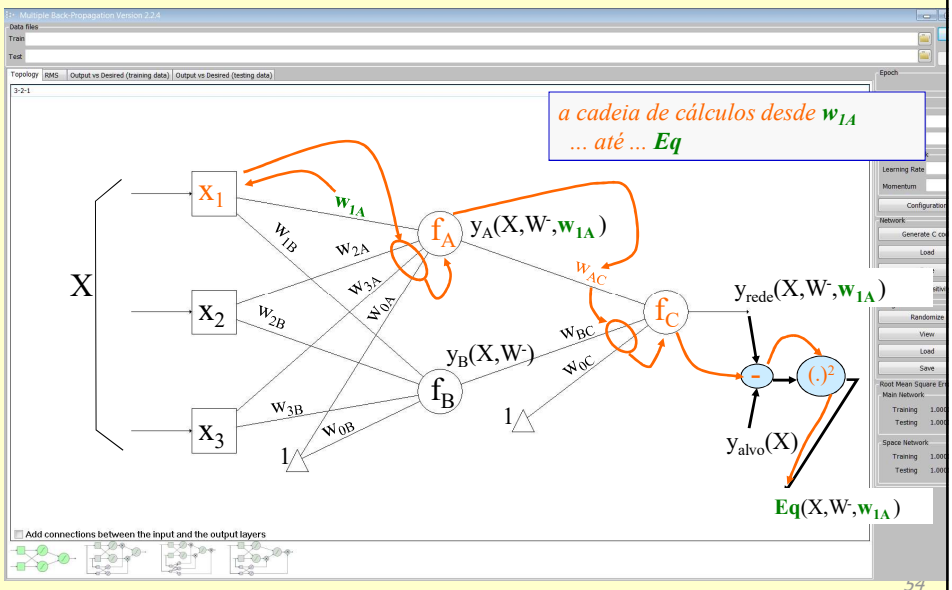




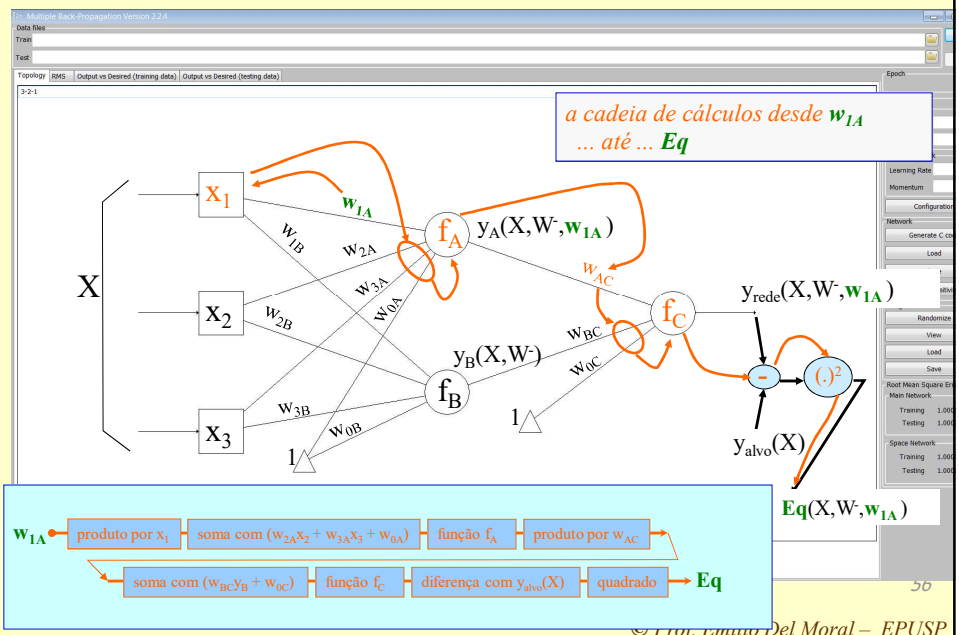

© Prof. Emilio Del Moral – EPUSP




© Prof. Emilio Del Moral – EPUSP




© Prof. Emilio Del Moral – EPUSP






© Prof. Emilio Del Moral – EPUSP

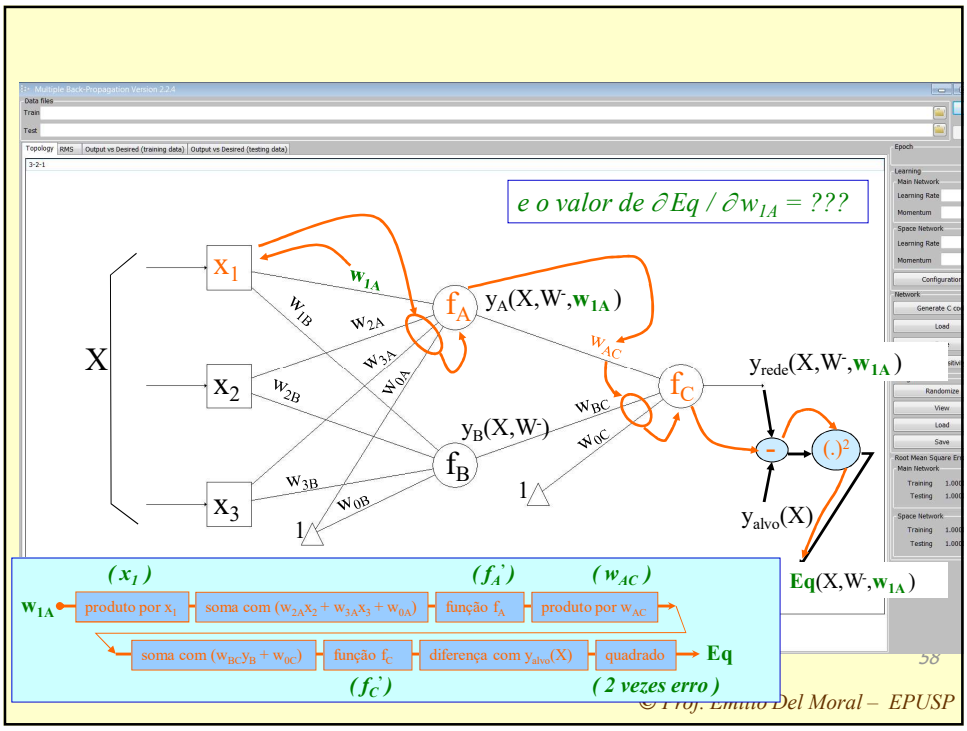
Note que aqui temos uma cadeia com muitos estágios que levam da váriável $w_{1A}$, à variável $Eq^{\mu}$, e para a qual podemos calcular a derivada da saída ($Eq^{\mu}$) com relação à entrada ($w_{1A}$) aplicando de forma repetida a seguinte propriedade simples e sua velha conhecida ...

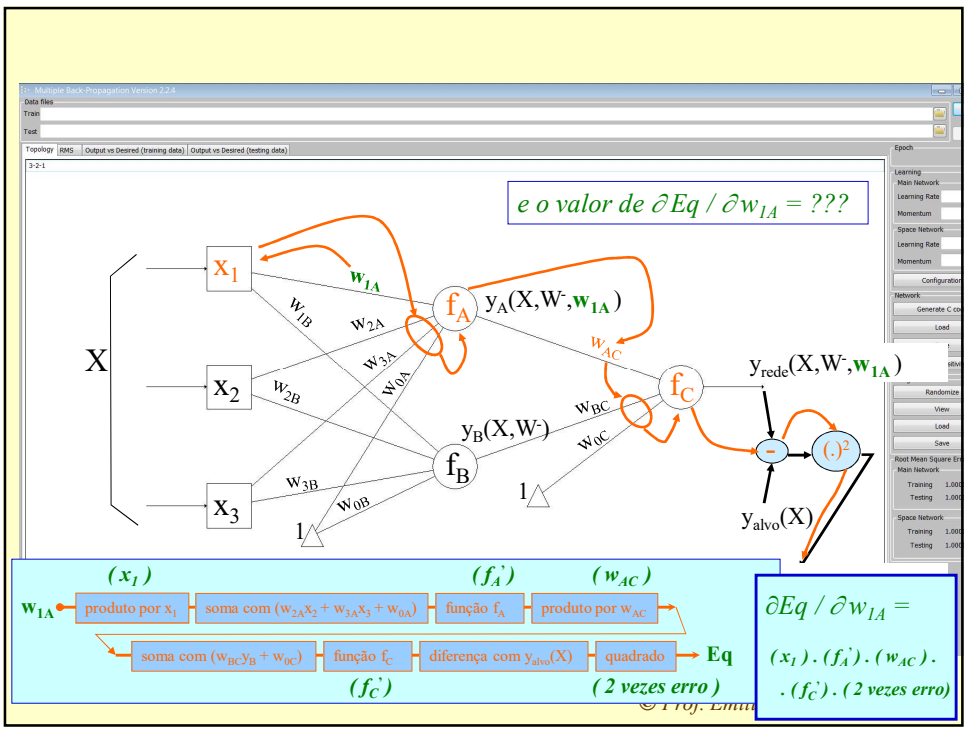
$$\mathbf{d(f_1(f_2(x)) / dx \;=\; df_1(x)/df_2 \,.\, df_2(x)/dx}$$

..., ou seja, calculando isoladamente o valor da derivada para cada estágio da cadeia, e finalizando o cálculo de derivada de ponta a ponta nessa cadeia toda através do produto dos diversos valores de cada estágio.



© Prof. Emilio Del Moral – EPUSP


e o valor de ∂Eq / ∂w1A = ???
X
x1
x2
x3
1
w1A
w1B
w2A
w3A
w0A
w2B
w3B
w0B
fA
yA(X,W-,w1A)
wAC
fB
yB(X,W-)
fC
wBC
w0C
yrede(X,W-,w1A)
-
(.)2
yalvo(X)
Eq(X,W-,w1A)
(x1)
(fA')
(wAC)
w1A
produto por x1
soma com (w2Ax2 + w3Ax3 + w0A)
função fA
produto por wAC
soma com (wBCyB + w0C)
função fC
diferença com yalvo(X)
quadrado
Eq
(fC')
(2 vezes erro)
58
© Prof. Emilio Del Moral – EPUSP


e o valor de ∂Eq / ∂w1A = ???
X
x1
x2
x3
1
w1A
w1B
w2A
w3A
w0A
w2B
w3B
w0B
fA
yA(X,W-,w1A)
wAC
fB
yB(X,W-)
fC
wBC
w0C
yrede(X,W-,w1A)
-
(.)2
yalvo(X)
(x1)
(fA')
(wAC)
w1A
produto por x1
soma com (w2Ax2 + w3Ax3 + w0A)
função fA
produto por wAC
soma com (wBCyB + w0C)
função fC
diferença com yalvo(X)
quadrado
Eq
(fC')
(2 vezes erro)
∂Eq / ∂w1A =
(x1) . (fA') . (wAC) .
. (fC') . (2 vezes erro)
© Prof. Emilio

## Lembretes ....

- Na maioria dos slides anteriores, onde aparece X, leia-se $X^{\mu}$, não incluído para não complicar demais os desenhos
- ... similarmente, onde aparece $y_{alvo}$, leia-se $y_{alvo}{}^{\mu}$ . Idem para os Eq, leia-se $Eq^{\mu}$
- Nos itens de cadeia de derivadas $(f_A')$ e $(f_C')$ , atenção para os valores dos argumentos, que devem ser os mesmos de $f_A$ e $f_C$ na cadeia original que leva $w_{1A}$ a $Eq$.
- ... lembrando ... na cadeia original tinhamos ...
  - para $f_C$: $f_C(w_{AC} \cdot f_A(w_{1A}.x_1 + w_{2A}.x_2 + w_{0A}) + w_{BC} \cdot f_B(w_{1B}.x_1 + w_{2B}.x_2 + w_{0B}) + w_{0C})$
  - para $f_A$: $f_A(w_{1A}.x_1 + w_{2A}.x_2 + w_{0A})$
- Similarmente, para o bloco "quadrado", cuja derivada é a função "2 vezes erro", o argumento é $[y_{rede}(X,W) - y_{alvo}(X)]$



© Prof. Emilio Del Moral – EPUSP

## Lembretes ....

- O mesmo que foi feito para $w_{1A}$ deve ser feito agora para os demais 10 pesos: $w_{2A}$, $w_{3A}$, $w_{0A}$, $w_{1B}$, $w_{2B}$, $w_{3B}$, $w_{0B}$, $w_{AC}$, $w_{BC}$, e $w_{0C}$ !
- Assim compomos um gradiente de 11 dimensões, com as derivadas de $Eq^{\mu}$ com relação aos 11 diferentes pesos w: $Grad_W(Eq^{\mu})$
- Essas 11 fórmulas devem ser aplicadas repetidamente aos M exemplares numéricos de $X^{\mu}$ e $y_{alvo}{}^{\mu}$, calculando M gradientes!
- Com eles, se obtém o gradiente médio dos M pares empíricos: $Grad_W(Eqm) = [\Sigma_{\mu} Grad_W(Eq^{\mu})] / M$
- Esse gradiente médio é a Bussola do Gradiente!



© Prof. Emilio Del Moral – EPUSP

**Método do Gradiente Aplicado aos nossos MLPs: a partir de um W#0, temos aproximações sucessivas ao Eqm mínimo, por repetidos pequenos passos DeltaW, sempre contrários ao gradiente ...**

- "Chute" um W inicial para o "Wcorrente", ou "W melhor até agora"
- Em loop até obter Eqm zero, ou baixo o suficiente, ou estável:
  - Determine o vetor gradiente do Eqm, nesse espaço de Ws
  - Em loop varrendo todos os M exemplos ($X^{\mu};y^{\mu}$),
    - Calcule o gradiente de $Eq^{\mu}$ associado a um exemplo $\mu$, e vá varrendo $\mu$ e somando os gradientes de cada $Eq^{\mu}$, para compor o vetor gradiente de Eqm, assim que sair deste loop em $\mu$ ;
    - Cada cálculo como esse, envolve primeiro calcular os argumentos de cada tangente hiperbólica e depois usar esses argumentos na regra da cadeia das derivadas necessárias
  - Dê um passo Delta $\Delta W$ nesse espaço, com direção e magnitude dados por $-\eta$*vetor gradiente <u>médio</u> para os M Exemplos ($X^{\mu};y^{\mu}$) de treino

63

© Prof. Emilio Del Moral – EPUSP

## O que devemos mirar quando exploramos o espaço de pesos W buscando que a RNA seja um bom modelo?

*Devemos mirar Maximização da aderência = Minimo Eqm possível*

**As Setas Verdes Indicam Situações que Devem ser Procuradas**


**Atacaremos esta possibilidade mais adiante ...**
... Mas adiantando: podemos realizar o gradiente descendente repetidamente, com pesos iniciais randômicos variados, e ficar com o melhor dos resultados de treino!

64

© Prof. Emilio Del Moral – EPUSP