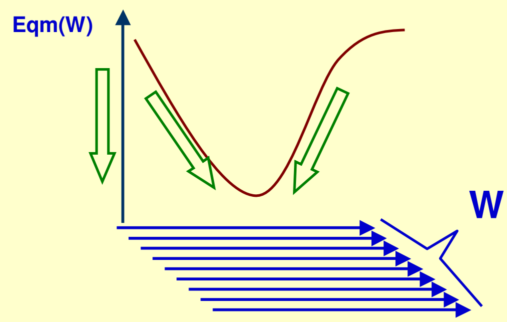
A aula de PSI3471 em
24-abril-2017 (#07)
continuou daqui,vinda de um conjunto de slides anterior

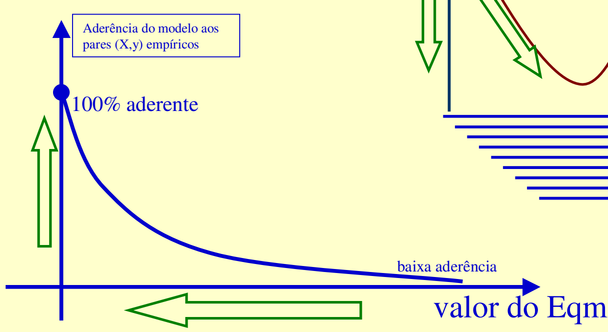
Aprendizado em RNAs do tipo MLP – Multi Layer Percetron – através do algoritmo Error Back Propagation

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

Devemos buscar Maximização da aderência = Minimo Eqm possível

As Setas Verdes Indicam Situações que Devem ser Procuradas



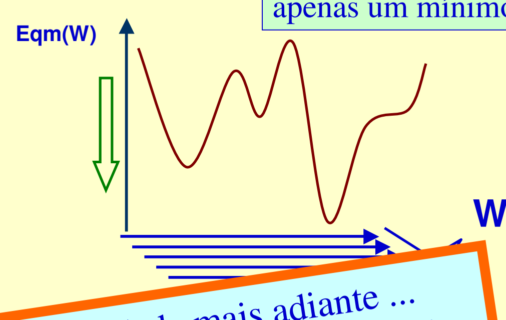


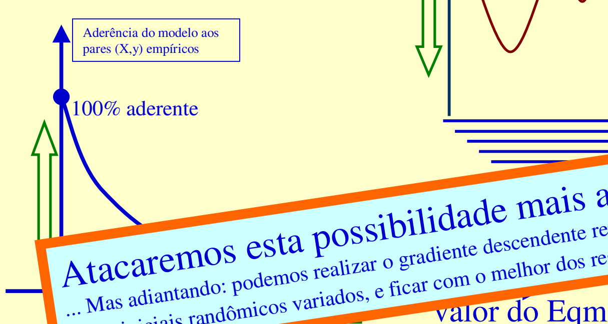
O que devemos buscar quando exploramos o espaço de pesos W buscando que a RNA seja um bom modelo?

Devemos buscar Maximização da aderência = Minimo Eqm possível

As Setas Verdes Indicam Situações que Devem ser Procuradas

Será que temos apenas um mínimo??





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!

O treinamento mira minimizar o Eqm das amostras (X ; y) de treino. (exclusivamente!)

sistema a modelar

Relação de amostragem ...
... E lembremos que as amostras sempre são uma representação parcial do comportamento mais geral do sistema que está sendo modelado.

Tabela de amostras (X ; y)

$$\mathbf{Eqm} = \sum_{\mu} [\, y_{RNA}(X^{\mu}) - y_{sistema}{}^{\mu} \,]^2 \; / \; M$$

RNA

(modelo do sistema)

Um Exemplo Ilustrativo para o Conceito de Conjunto de Treinamento e dos M pares (X,y)…

Exemplo de regressão multivariada para estimação contínua usando MLP

- O valor do y contínuo ... neste exemplo corresponde ao volume de consumo futuro num dado tipo de produto "A" a ser ofertado pela empresa a um cliente corrente já consumidor de outros produtos da empresa ("B" e "C"), volume esse previsto com base em várias medidas quantitativas que caracterizam tal indivíduo. ... Assim, y = Consumo do Produto A = F(x_1,x_2, x_3, x_4, x_5).

- Consideremos 4 variáveis de entrada no modelo preditivo neural, ou seja, temos 5 medidas em X:
  - x_1: Idade do indivíduo
  - x_2: Renda mensal do indivíduo
  - x_3: Volume de clicks do indivído no website de exibição de produtos oferecidos pela empresa
  - x_4: Volume de consumo desse cliente observado para outro Produto B da mesma empresa
  - x_5: Volume de consumo desse cliente Produto C da mesma empresa

- Problema: desenvolver uma MLP para regressão contínua multivariada que permita estimar esse volume de consumo futuro y com base no conhecimento dos X e numa base de dados de aprendizado com esses dados X e y para 350 já clientes de universo populacional similar ao do novo consumidor potencial.

Em termos de Excel, teriamos ...

Cliente (μ)	Idade (x_1)	Renda (x_2)	Clics (x_3)	Consumo do Produto B (x_4)	Consumo do Produto C (x_5)	Consumo do Produto A (y)
1	50	78	302	958	136	**9800**
2	65	128	186	985	196	**8760**
3	57	150	221	1093	35	**520**
.....	.....	.....	.....	.....	.....	**.....**
M-2	16	19	51	707	131	**11640**
M-1	30	75	7	29	78	**9640**
M	19	47	116	285	124	**5320**

Em termos de Excel, teriamos ...

Cliente (μ)	Idade (x_1)	Renda (x_2)	Clics (x_3)	Consumo do Produto B (x_4)	Consumo do Produto C (x_5)	Consumo do Produto A (y)
			302	958	136	**9800**
			186	985	196	**8760**
3	57					**520**
.....	...					**.....**
M-2	16					**11640**
M-1	30					**9640**
M	19					**5320**

Equivalente em txt
Para uso do MBP

treino em txt para exemplo de consumo A e B - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

```
Idade    Renda    Clics    ConsumoA ConsumoB ConsumoA

50       78       302      958      136      9800
65       128      186      985      196      8760
57       150      221      1093     35       520

(...)

16       19       51       707      131      11640
30       75       7        29       78       9640
19       47       116      285      124      5320
```


A estratégia de Aprendizado para o MLP mais conhecida:

Error Back Propagation (EBP)

= Propagação Reversa de Erro

= Método do Gradiente personalizado ao Eqm(W) do MLP

Mas entendamos PRIMEIRO

o que é o método numérico do

gradiente ascendente /

gradiente descendente

genérico,

que pode ser aplicado tanto para se chegar paulatinamente ao máximo de uma função quanto para se chegar ao mínimo de uma função

(ascendente / descendente)

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

- **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:**

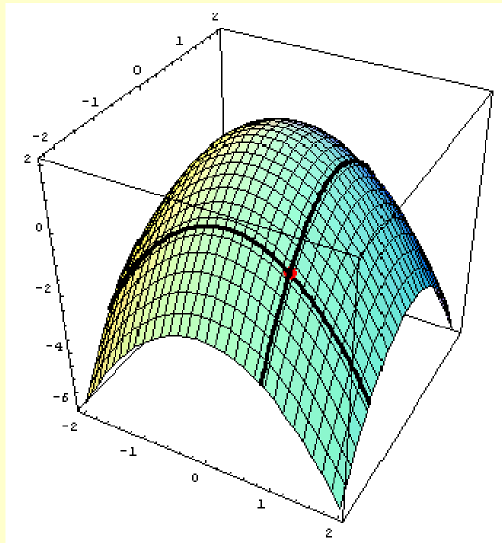
$(\partial Eqm(W)/\partial w_1\ ,\ \partial Eqm(W)/\partial w_2\ ,\ \partial Eqm(W)/\partial w_3\ , \ldots)$

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

- **Regra da cadeia, necessária ao cálculo de derivadas quando há encadeamento de funções:**

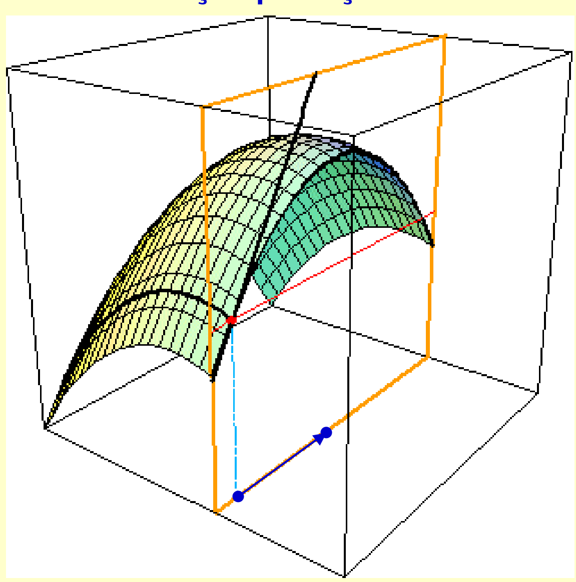
$\partial f(\ g(\ h(a)\ )\ )/\partial a = \partial f/\partial g\ .\ \partial g/\partial h\ .\ \partial h/\partial a$

Derivada parcial- ilustração p/ função de 2 variáveis apenas



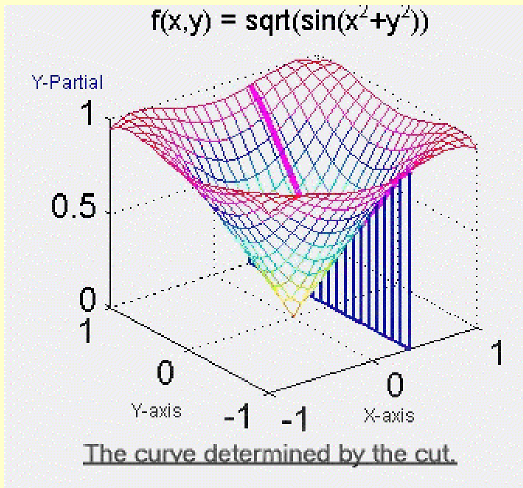
A visual model of the partial derivative

Derivada parcial- ilustração p/ função de 2 variáveis apenas



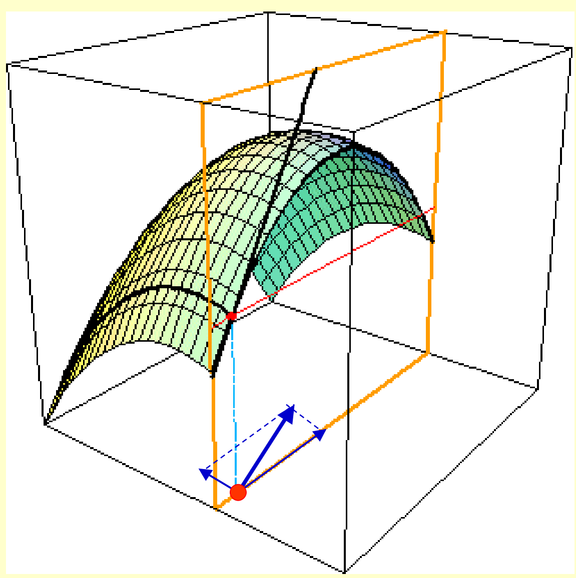
A visual model of the partial derivative

mais ilustrações p/ a derivada parcial em função de 2 variáveis



A visual model of the partial derivative with respect to y.

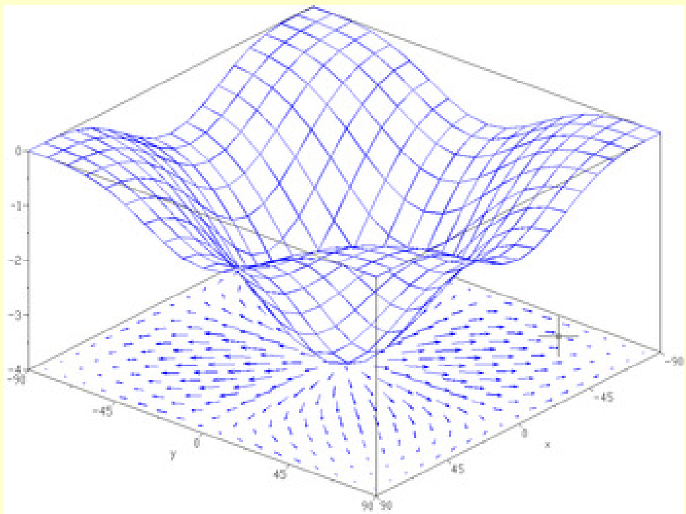
Formação do vetor gradiente a partir de duas derivadas parciais



A visual model of the partial derivative

http://en.wikipedia.org/wiki/Gradient

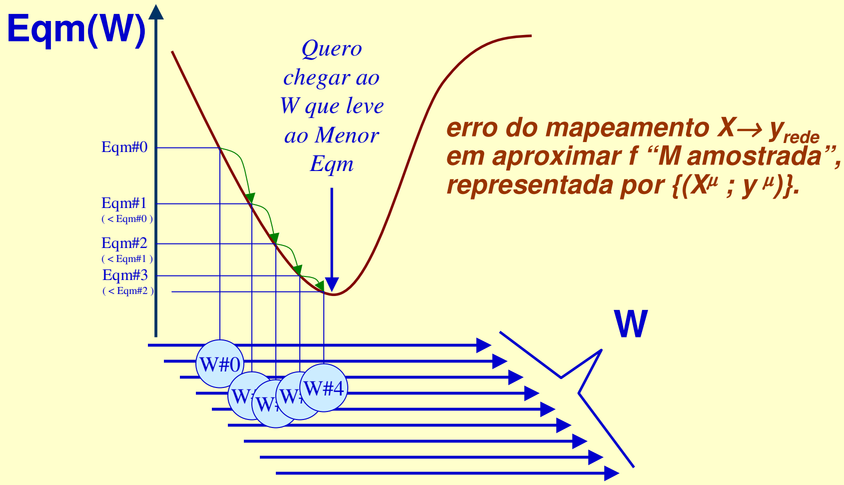
... O vetor gradiente indica a direção ascendente e seu módulo a magnitude de crescimento da função escalar – ilustração p/ função de 2 variáveis apenas



Processo de refinamentos graduais a cada iteração ...

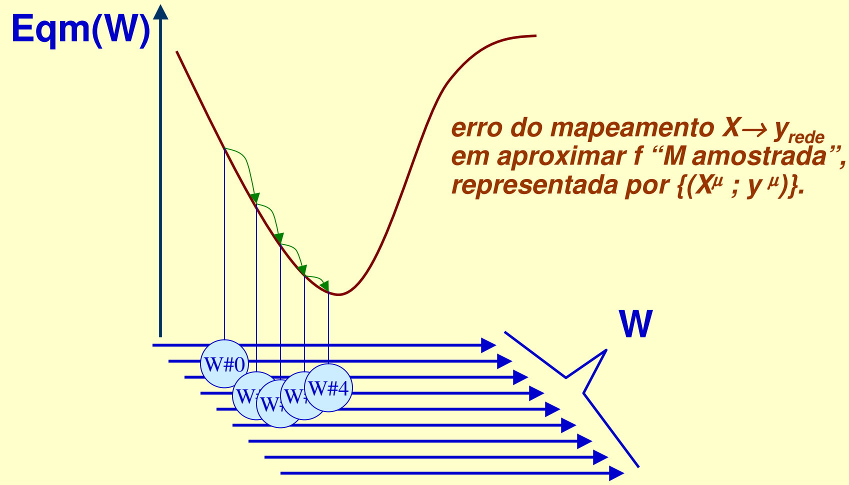
W#0	Eqm#0	GradEqm(W#0)	DeltaW#0 = - n.GradEqm(W#0)
W#1 (= W#0 + DeltaW#0)	Eqm#1 (< Eqm#0)	GradEqm(W#1)	DeltaW#1 = - n.GradEqm(W#1)
W#2 (= W#1 + DeltaW#1)	Eqm#2 (< Eqm#1)	GradEqm(W#2)	DeltaW#2 = - n.GradEqm(W#2)
W#3 (= W#2 + DeltaW#2)	Eqm#3 (< Eqm#2)	GradEqm(W#3)	DeltaW#3 = - n.GradEqm(W#3)
W#4 (= W#3 + DeltaW#3)	Eqm#4 (< Eqm#3)	GradEqm(W#4)	DeltaW#4 = - n.GradEqm(W#4)
...	...	...	...
W#k (= W#k-1 + DeltaW#k-1)	Eqm#k (< Eqm#k-1)	GradEqm(W#k)	DeltaW#k = - n.GradEqm(W#4)
...	...	...	...

A estratégia do EBP / Gradiente Descendente no aprendizado do MLP

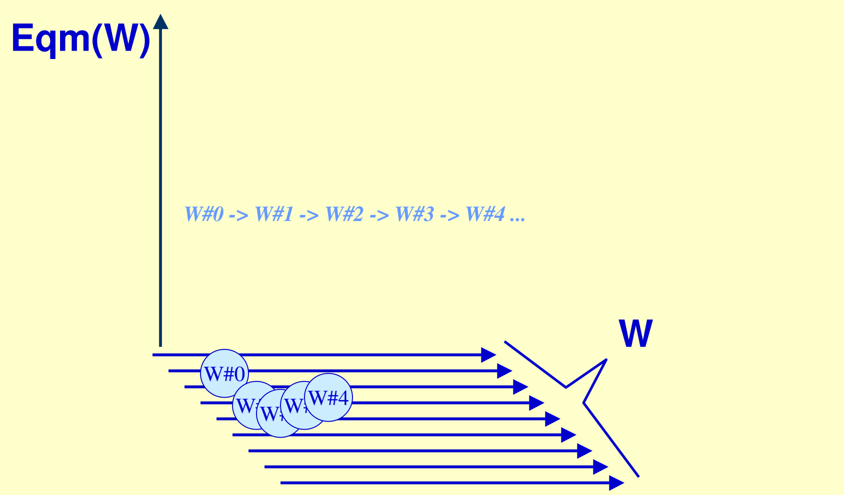


Este gráfico é denso e toca em muitos aspectos interrelacionados … revisitemos alguns desses aspectos isoladamente com focos específicos nessas revisitas, assim teremos gráficos algo mais simples de interpretar …

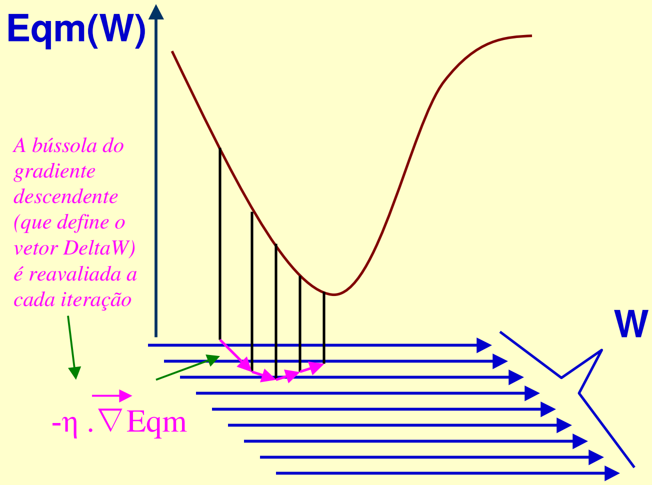
A estratégia do EBP / Gradiente Descendente no aprendizado do MLP



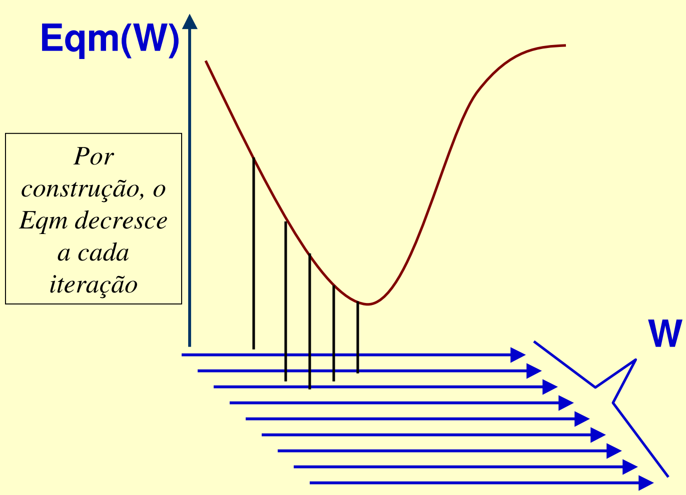
Foco na evolução dos w´s com as iterações ...



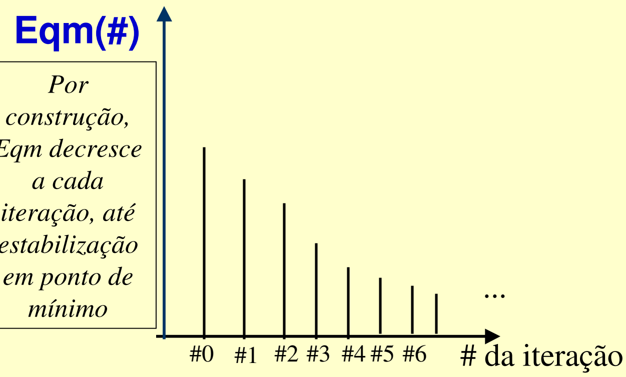
Foco nos diferentes DeltaW de cada iteração ...



Foco na evolução do Eqm com as iterações ...

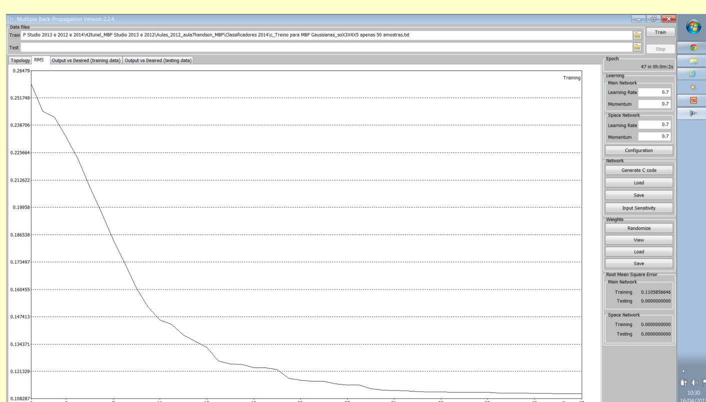


Plotando a evolução do Eqm com as iterações ...



Por construção, Eqm decresce a cada iteração, até estabilização em ponto de mínimo

Gráfico fornecido pelo ambiente MBP da evolução do Eqm com o número de repetidos usos da "bússola do gradiente descendente": isto conecta o MBP com o gráfico apresentado no slide anterior



Nota: o RMS do eixo vertical deste gráfico significa Root Mean Square, e é a raiz quadrada do nosso conhecido Eqm

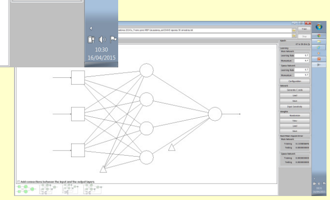


Gráfico mostrando as primeiras 47 iterações do processo de refinamentos sucessivos do modelo ...

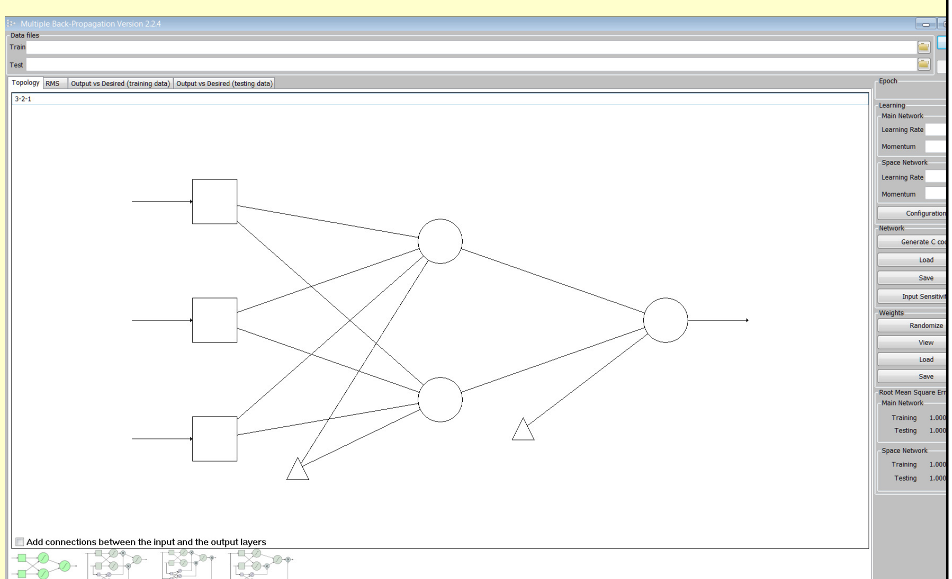
Aprendizado do MLP por Error Back Propagation ...

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

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

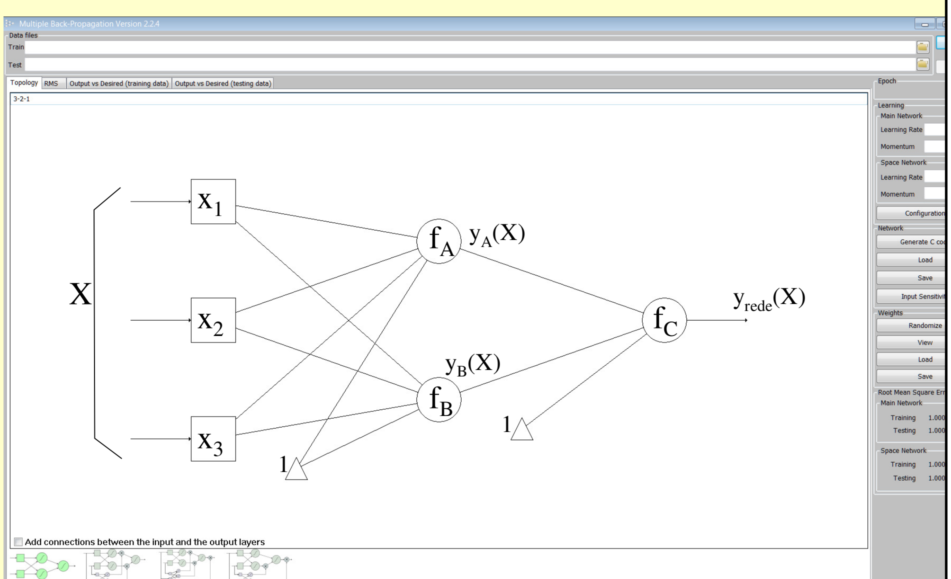
Relembrando o que está por trás de um desenho como o que segue ...

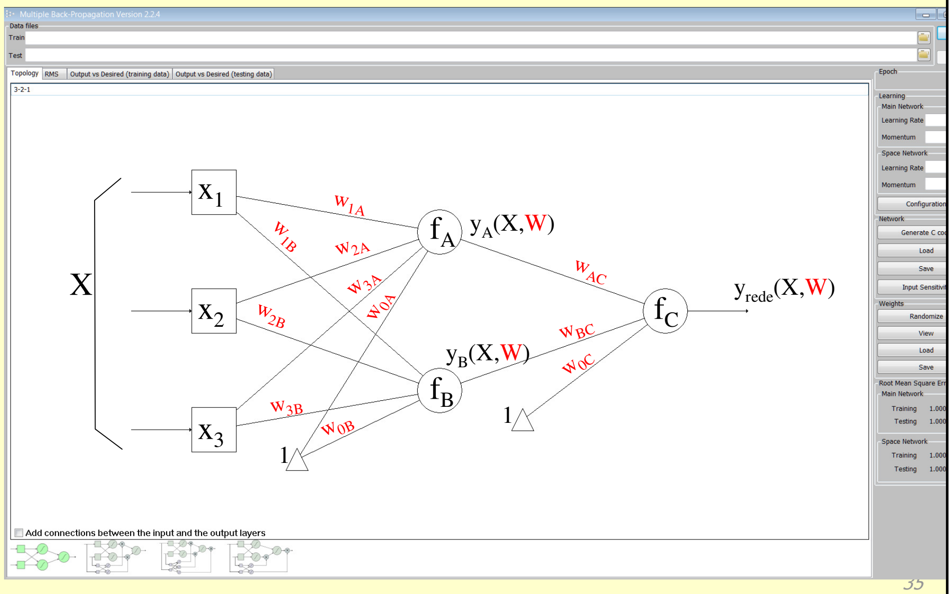


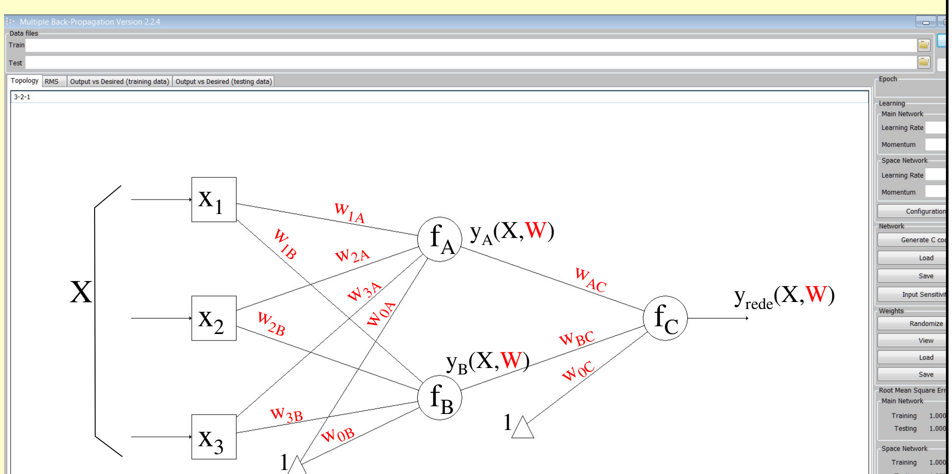
Importante ... O que está implícito nas imagens de redes apresentadas em lousa e nas telas do MBP ...
(relembrando premissas colocadas nas primeiras aulas)

... O fato de que cada nó neural (cada "círculo" nos diagramas da RNA) que compõe a rede gera na saída um cálculo do tipo tgh (ou sigmoidal de sua escolha) que opera sobre a somatória ponderada (ponderação pelos seus w's) das diversas entradas que lhe chegam (e incluindo também um viés nessa somatória)

- **Isto ocorre para todos os nós neurais da RNA: se tivermos por exemplo 2+1 (3 no total) nós neurais, como na RNA exemplo do slide anterior, cada um deles realizará um cálculo desse tipo, ou seja, tgh(soma ponderada), empregando os valores específicos de seus pesos ponderadores e seus vieses exclusivos, valores esses que podem ser distintos para cada nó**
- **Em particular, os nós das camadas mais adiante na RNA têm como suas entradas as variáveis de saída dos nós da camada anterior (ou seja, operam sobre as saidas de tangentes hiperbólicas das camadas nteriores)**
- **Isto que digo estar implícito no MBP já faz parte de qualquer rede neural** *(revise os slides das primeiras aulas)*, **por isso não precisa ser detalhado em cada figura**
- **De qualquer forma, para ajudar a visualizar o que está implícito e que é premissa em RNAs, nos slides que seguem eu adiciono os significados assumidos na rede exemplo 3-2-1 do slide anterior.**

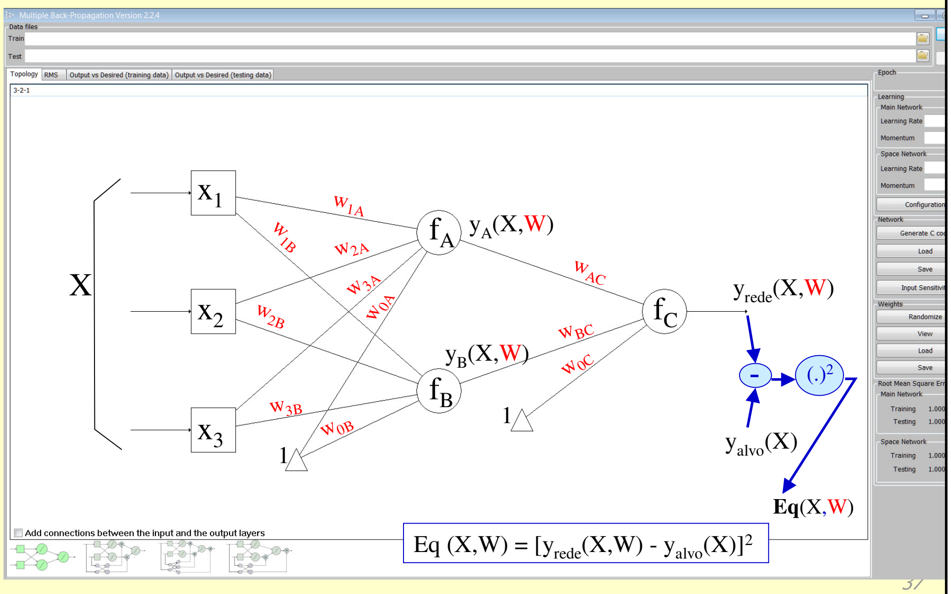




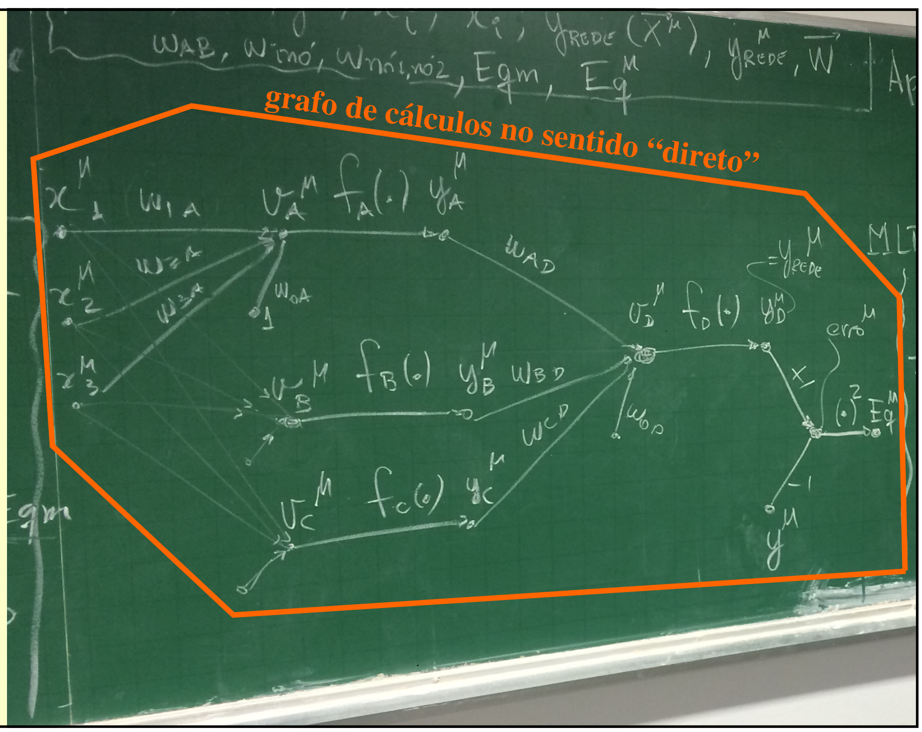


$$y_{rede}(X,W) = f_C\left(w_{AC} \cdot y_A(X,W) + w_{BC} \cdot y_B(X,W) + w_{0C}\right)$$

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



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



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:**

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

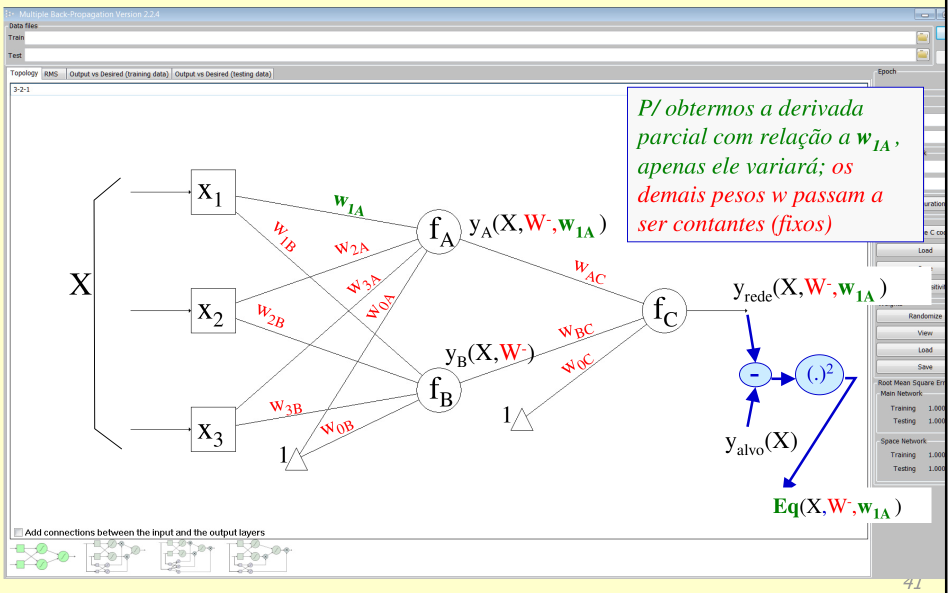
Invertamos o operador gradiente e a somatória

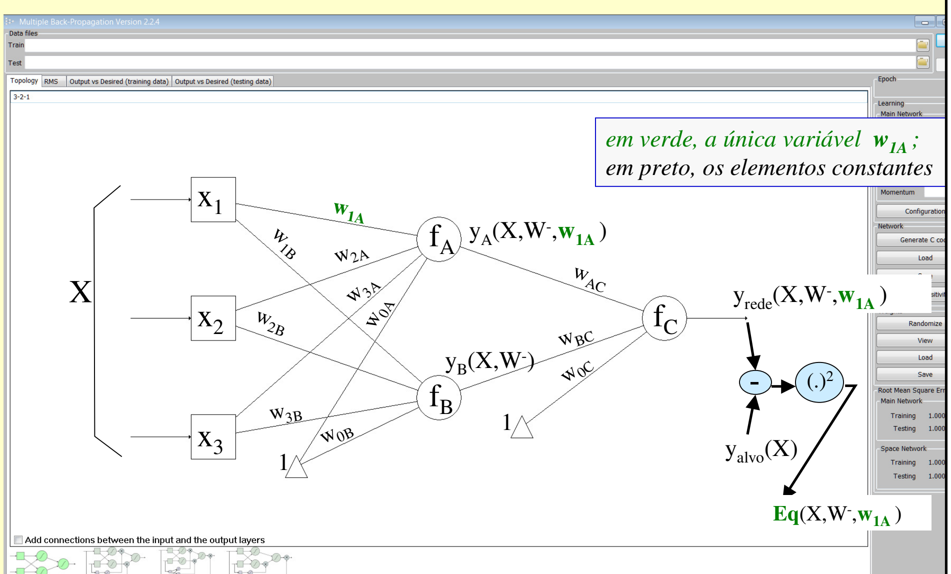
.
.. afinal, gradiente é uma derivada, e a derivada de um soma de várias funções é igual à soma das derivadas individuais de cada componente da soma:

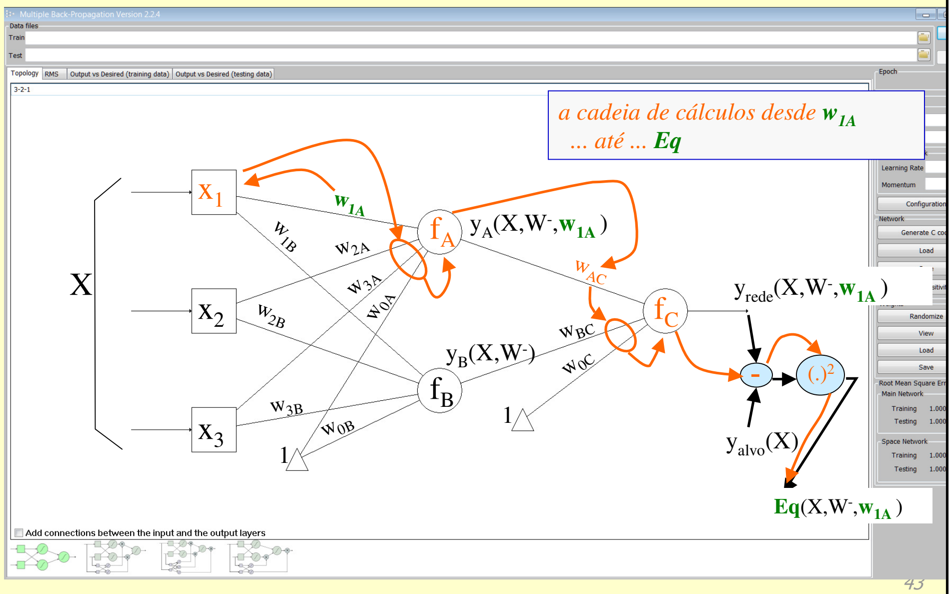
$$Grad(Eqm) =$$

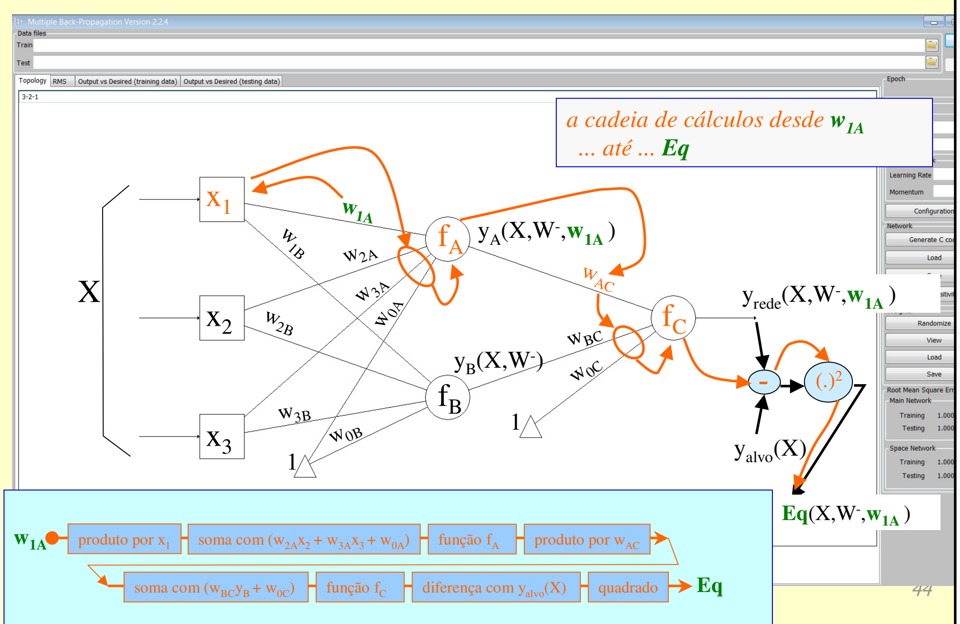
$$Grad(\Sigma_\mu Eq^\mu) / M$$

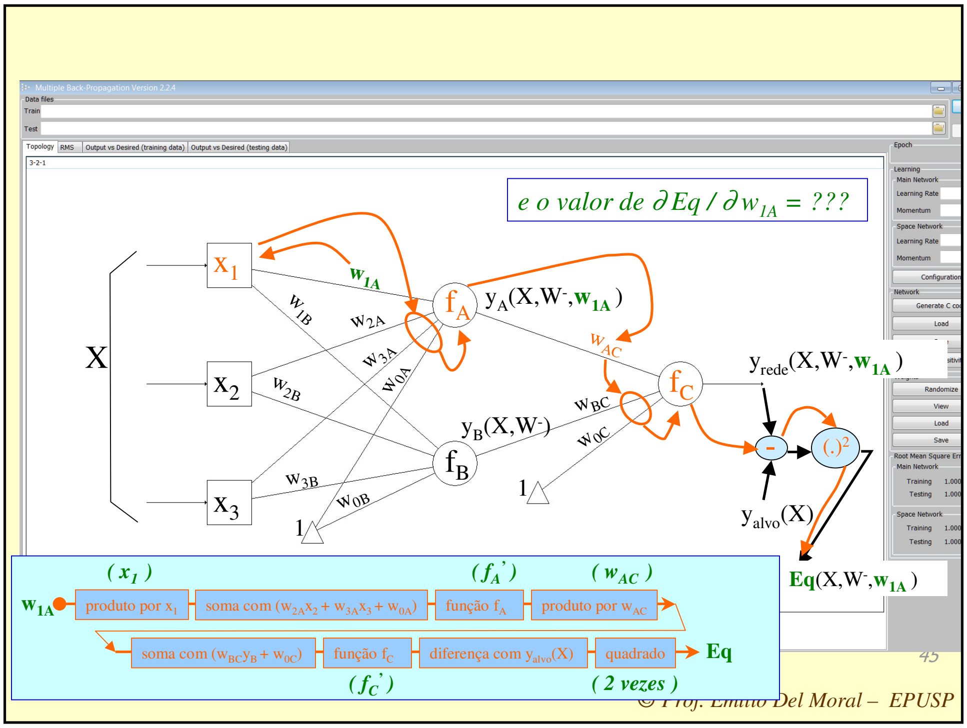
$$\Sigma_\mu Grad(Eq^\mu) / M$$


P/ obtermos a derivada parcial com relação a w_{1A}, apenas ele variará; os demais pesos w passam a ser contantes (fixos)
X
x_1
x_2
x_3
w_{1A}
w_{1B}
w_{2A}
w_{2B}
w_{3A}
w_{3B}
w_{0A}
w_{0B}
f_A
f_B
f_C
$y_A(X,W^-,w_{1A})$
$y_B(X,W^-)$
w_{AC}
w_{BC}
w_{0C}
1
$y_{rede}(X,W^-,w_{1A})$
-
$(.)^2$
$y_{alvo}(X)$
$Eq(X,W^-,w_{1A})$

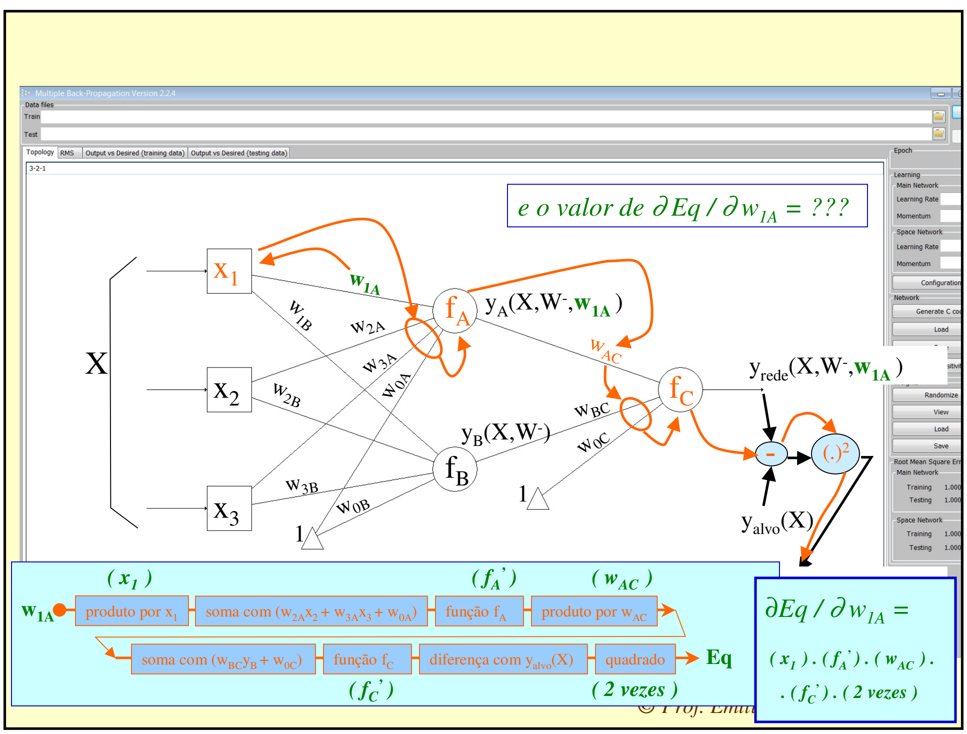

em verde, a única variável w_{1A};
em preto, os elementos constantes
X
x_1
x_2
x_3
w_{1A}
w_{1B}
w_{2A}
w_{2B}
w_{3A}
w_{3B}
w_{0A}
w_{0B}
f_A
f_B
f_C
$y_A(X,W^-,w_{1A})$
$y_B(X,W^-)$
w_{AC}
w_{BC}
w_{0C}
1
$y_{rede}(X,W^-,w_{1A})$
-
$(.)^2$
$y_{alvo}(X)$
$Eq(X,W^-,w_{1A})$


a cadeia de cálculos desde w_{1A} ... até ... Eq
X
x_1
x_2
x_3
w_{1A}
w_{1B}
w_{2A}
w_{3A}
w_{0A}
w_{2B}
w_{3B}
w_{0B}
f_A
f_B
f_C
$y_A(X,W^-,w_{1A})$
$y_B(X,W^-)$
w_{AC}
w_{BC}
w_{0C}
1
1
$y_{rede}(X,W^-,w_{1A})$
$y_{alvo}(X)$
$(.)^2$
$Eq(X,W^-,w_{1A})$
Add connections between the input and the output layers


a cadeia de cálculos desde w_{1A} ... até ... Eq
X
x_1
x_2
x_3
w_{1A}
w_{1B}
w_{2A}
w_{3A}
w_{0A}
w_{2B}
w_{3B}
w_{0B}
f_A
f_B
f_C
$y_A(X,W^-,w_{1A})$
$y_B(X,W^-)$
w_{AC}
w_{BC}
w_{0C}
1
1
$y_{rede}(X,W^-,w_{1A})$
$y_{alvo}(X)$
$(.)^2$
$Eq(X,W^-,w_{1A})$
w_{1A} → produto por x_1 → soma com $(w_{2A}x_2 + w_{3A}x_3 + w_{0A})$ → função f_A → produto por w_{AC}
→ soma com $(w_{BC}y_B + w_{0C})$ → função f_C → diferença com $y_{alvo}(X)$ → quadrado → Eq



e o valor de ∂Eq / ∂w1A = ???
X
x1
x2
x3
w1A
w1B
w2A
w3A
w0A
w2B
w3B
w0B
1
fA
yA(X,W-,w1A)
fB
yB(X,W-)
wAC
wBC
w0C
1
fC
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)
45
© Prof. Emilio Del Moral – EPUSP



e o valor de ∂Eq / ∂w1A = ???
X
x1
x2
x3
w1A
w1B
w2A
w3A
w0A
w2B
w3B
w0B
1
fA
yA(X,W-,w1A)
fB
yB(X,W-)
wAC
wBC
w0C
1
fC
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)
∂Eq / ∂w1A =
(x1) . (fA') . (wAC) .
. (fC') . (2 vezes)

Lembretes ....

- Na maioria dos slides anteriores, onde aparece X, leia-se X^μ, 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^μ
- 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", o argumento é $[y_{rede}(X,W) - y_{alvo}(X)]$

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 um gradiente de 11 dimensões, com as derivadas de Eq^μ com relação aos 11 diferentes pesos w: $Grad_W(Eq^\mu)$
- Essas 11 fórmulas devem ser aplicadas repetidamente aos M exemplares numéricos de X^μ 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 <u>médio</u> é a Bussola do Gradiente!

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^{μ};y^{μ}),**
    - **Calcule o gradiente de Eq^{μ} associado a um exemplo μ, e vá varrendo μ e somando os gradientes de cada Eq^{μ}, para compor o vetor gradiente de Eqm, assim que sair deste loop em μ ;**
    - **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 ΔW nesse espaço, com direção e magnitude dados por $-\eta$*vetor gradiente <u>médio</u> para os M Exemplos (X^{μ};y^{μ}) de treino**

O Ciclo completo da modelagem:

0) Formalização do problema, mapeamento quantitativo em um modelo neural inicial e ... 0b) coleta de pares empíricos (X,y)

1) Fase de TREINO da RNA (MLP): com conhecimento dos X e dos y, que são ambos usados na calibração do modelo

2) Fase de TESTE / Caracterização da qualidade da RNA para generalizar: temos <u>novos</u> pares X e y, com y guardado "na gaveta", usado apenas para avaliação, não para re-calibração. É como um ensaio de uso final do modelo, com possibilidade de medir a sua qualidade com o y que foi guardado na gaveta.

[Fase de refinamentos da RNA, dados e modelo, em ciclos, desde 0]

3) Fase de USO FINAL da RNA, com y efetivamente não conhecido, e estimado com conhecimento dos X + uso do modelo calibrado.

.... Diferenças e semelhanças entre 1, 2 e 3

... + exercícios de treino já disponíveis no STOA ...

Ex 4) Defina dois problemas de regressão multivariada (y(X) é analógico) no contexto de sistemas eletrônicos, um deles em imagens e outro não (por exemplo sistemas multisensores) que possa se beneficiar de aprendizado de máquina e de treinamento de uma rede neural. Defina claramente as variáveis de entrada, de saída, suas unidades e significados, explique porque o regressor é algo útil, porque precisamos do aprendizado de máquina na modelagem, e defina como coletará os pares epíricos (X,y).

Ex 4b) Ensaie esse problema no MBP.

Ex 5) Defina dois problemas de detecção de padrões (y(X) é binário) no contexto de sistemas eletrônicos, um deles em imagens e outro não que possa se beneficiar de aprendizado de máquina e de treinamento de uma rede neural. Defina claramente as variáveis de entrada, de saída, suas unidades e significados, explique porque o regressor é algo útil, porque precisamos do aprendizado de máquina na modelagem, e defina como coletará os pares epíricos (X,y).

Ex 5b) Ensaie esse problema no MBP.

Ex 6) Para um MLP de 2 camadas com topologia 2-3-1 (três entradas, 3 neurônios intermediários e 1 neurônio de saída), deduza a fórmula da derivada parcial do Eqm com relação a um peso de primeira camada (um peso que liga uma das entradas da rede a um dos 3 neurônios da primeira camada) e a fórmula da derivada parcial do Eqm com relação a um viés de neurônio da primeira camada.

Ex 6b) Deduza agora a fórmula da derivada de Eqm com relação a um peso que interliga a primeira camada de neurônios com a segunda. Deduza também a fórmula da derivada de Eqm com relação a um viés do neurônio de saída.

Ex 6c) Escreva agora com base nos seus resultados em 6 e 6b todas as 13 fórmulas de derivadas parciais do Eqm com relação aos 13 pesos sinápticos da rede (isto inclui os 4 pesos de viés).

Ex 6d) Como o item 6c pode ser usado por você para escrever um código de programa de aprendizado de pesos de uma rede neural? Além de explicar