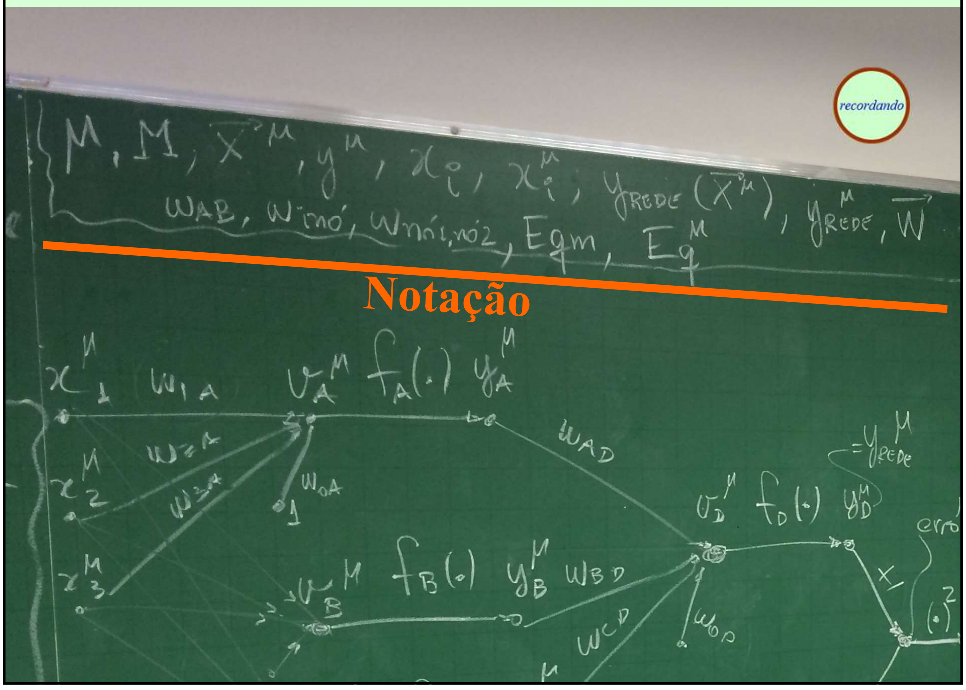
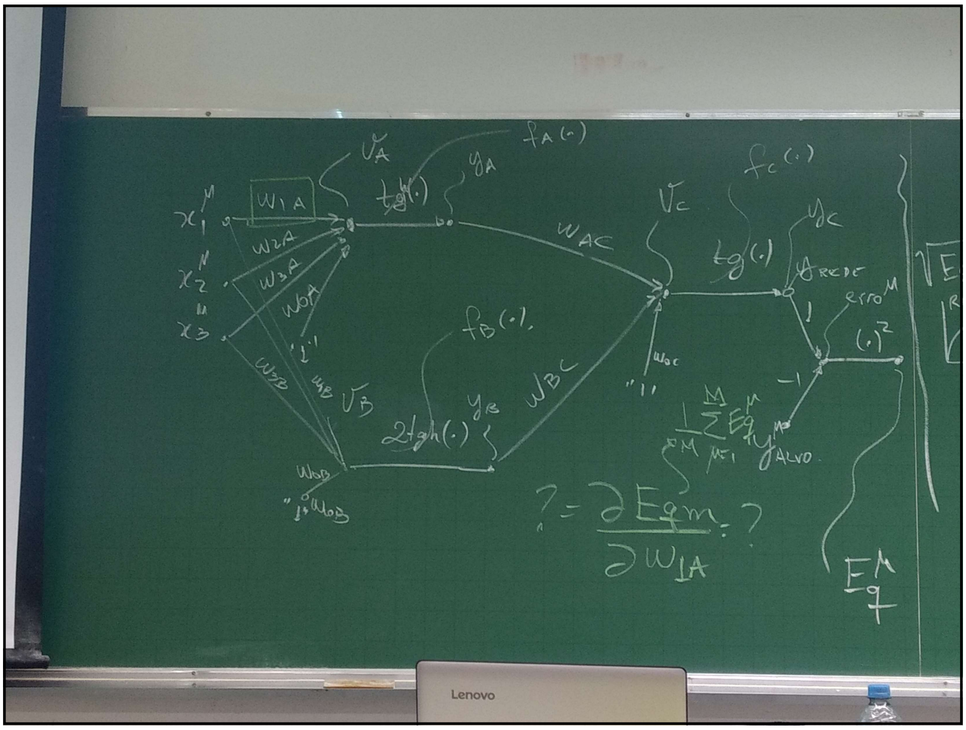
1

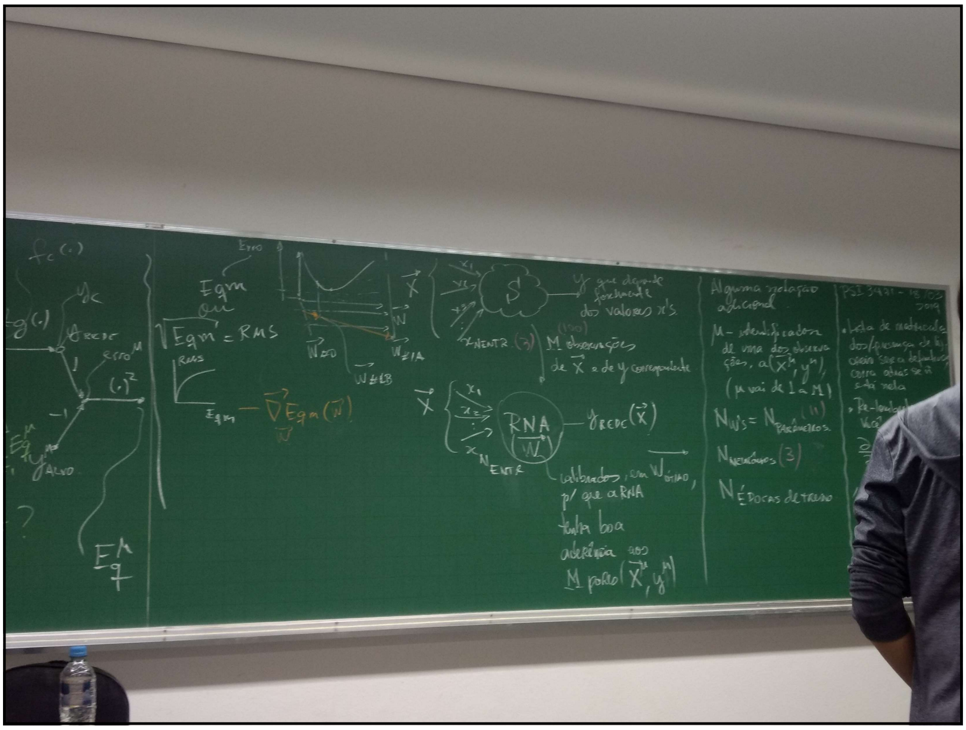
1



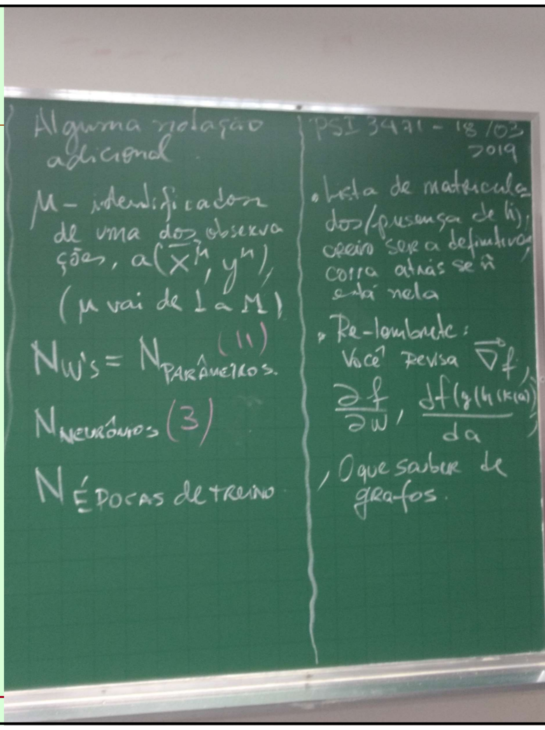
7



8



9



milio Del Moral Hernandez

10

Conexão de RNAs com elementos de "Aprendizado de Máquina" / "Machine Learning"

O "aprendizado do modelo" é feito de maneira automática a partir de casos / de exemplos concretos: a definição matemática do modelo é feita a partir de um conjunto rico de exemplos numéricos empíricos de pares (X,y)

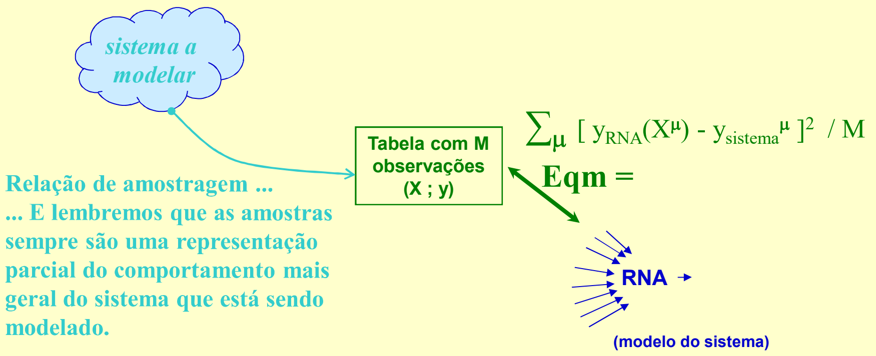
Conhecimento rico de exemplares / casos (X^μ ; y^μ): Temos M observações empíricas (X^μ ; y^μ), onde μ identifica cada observação, e varia entre 1 a M

y_{modelo} = cálculo neural (*ou seja, somas ponderadas e com tgh´s*) que opera sobre as componentes do vetor X (x_1, x_2, x_3 ...); Esse cálculo neural é calibrado (via escolha dos valores dos ponderadores w´s) a partir de M pares empíricos (X^μ ; y^μ)

EGF-006 – Métodos Numéricos e Reconhecimento de Padrões – Prof. Emilio Del Moral – EPUSP

11

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



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

12

12

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

13

13

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.

14

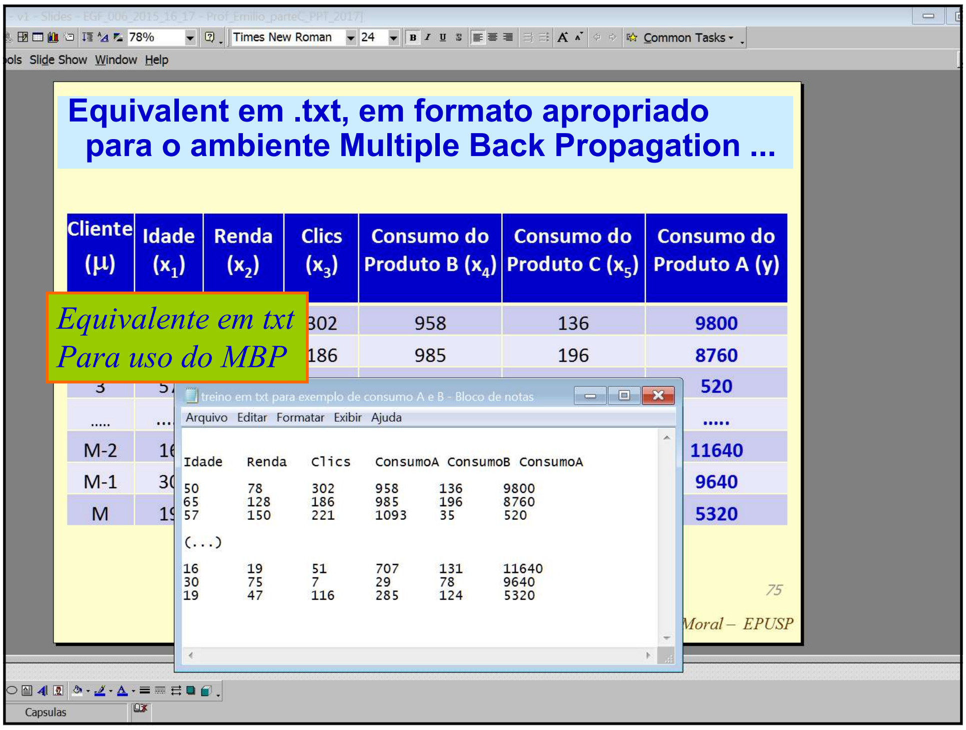
14

Exemplo de dados empíricos tabulados em Excel ...

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**

15

15



16

Como escolhemos os valores dos diversos w´s ?

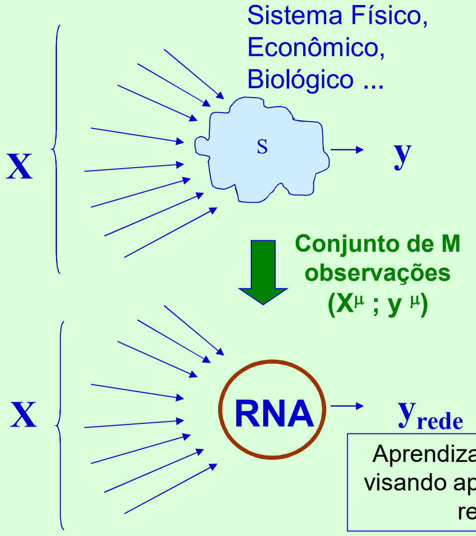
© Prof. Emilio Del Moral Hernandez

18

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

19

Conjunto de treino em arquiteturas supervisionadas (ex. clássico: MLP com Error Back Propagation)



A computação desejada da rede pode ser definida simplesmente através de observações / exemplares do comportamento requerido

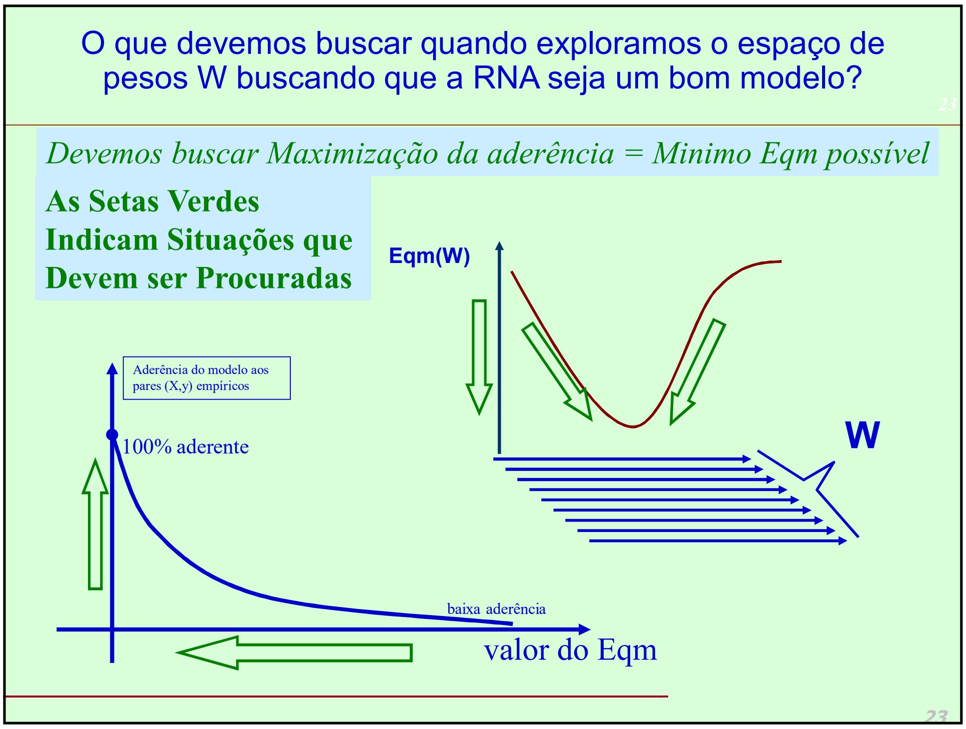
$$Eqm = \frac{1}{M}\sum_{\mu=1}^{M}\left(y_{rede}(\vec{X}^{\mu},\vec{W}) - y^{\mu}\right)^2$$

... em loop ...

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

Aprendizado: Espaço de pesos W é explorado visando aproximar ao máximo a computação da rede da computação desejada

22


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
Eqm(W)
W
Aderência do modelo aos pares (X,y) empíricos
100% aderente
baixa aderência
valor do Eqm
23

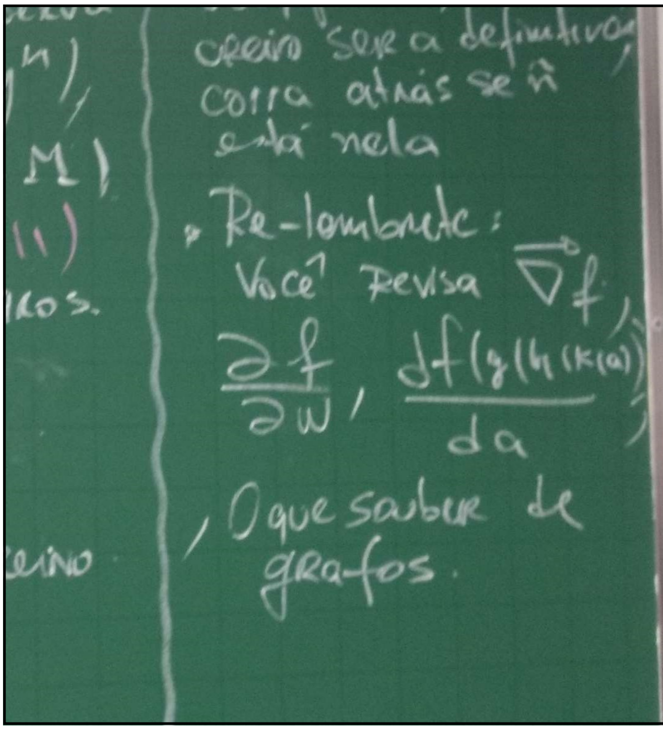
23


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
Será que temos apenas um mínimo??
Eqm(W)
W
Aderência do modelo aos pares (X,y) empíricos
100% aderente
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!
valor do Eqm
24
© Prof. Emilio Del Moral – EPUSP

24

Deduzindo as Equações do Aprendizado em RNAs do tipo MLP – Multi Layer Percetron – com o algoritmo Error Back Propagation (Gradiente Descendente)

29



36

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:**

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

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

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

37

37

A estratégia de Aprendizado para o MLP mais conhecida:

Error Back Propagation (EBP)

= Propagação Reversa de Erro

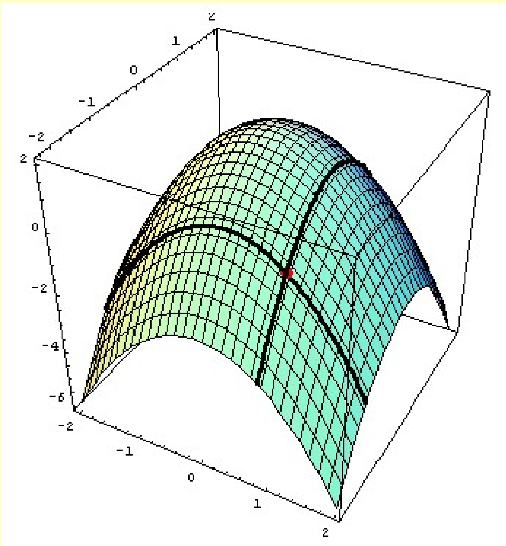
= Método do Gradiente personalizado ao Eqm(W) do MLP

38

38

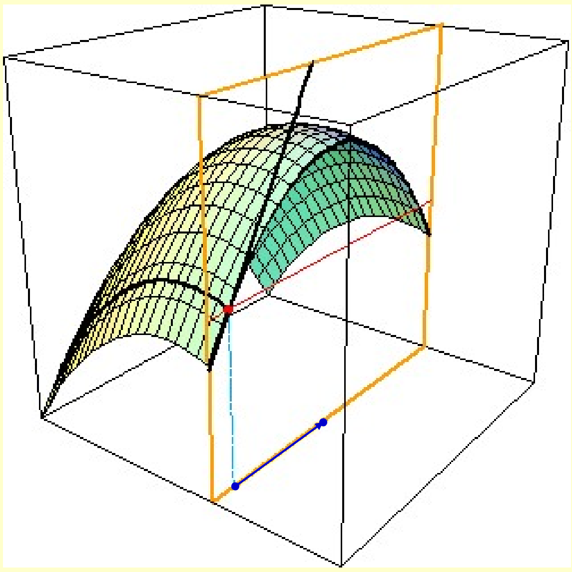
Entendamos PRIMEIRO
o que é o método numérico do
gradiente ascendente /
gradiente descendente
genérico (o EBP é um caso particular),
que pode ser aplicado tanto para se chegar paulatinamente ao máximo de uma função quanto para se chegar ao mínimo de uma função
(ascendente / descendente)

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

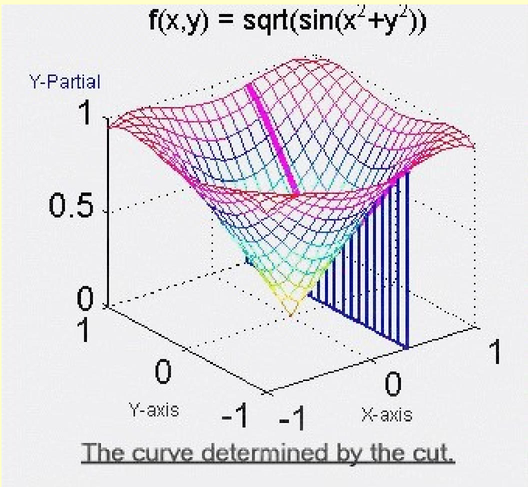


A visual model of the partial derivative

41

41

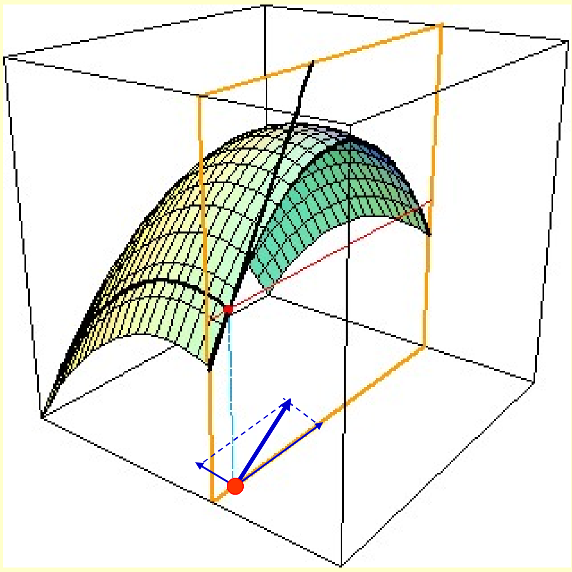
mais ilustrações p/ a derivada parcial em função de 2 variáveis



A visual model of the partial derivative with respect to y. 42

42

Formação do vetor gradiente a partir de duas derivadas parciais



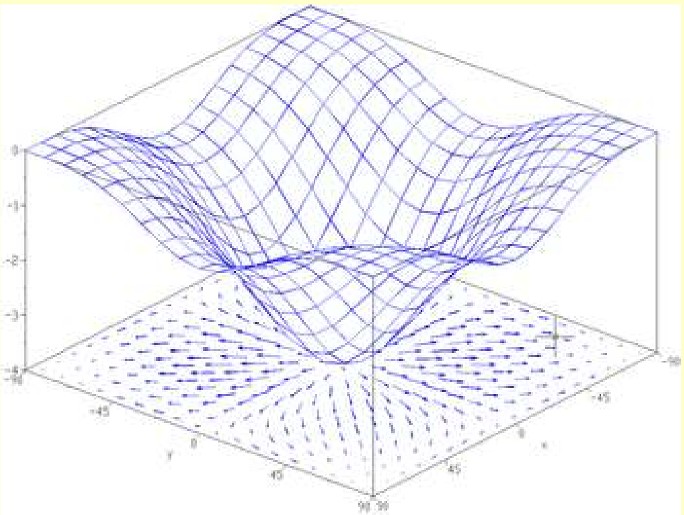
A visual model of the partial derivative

43

43

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

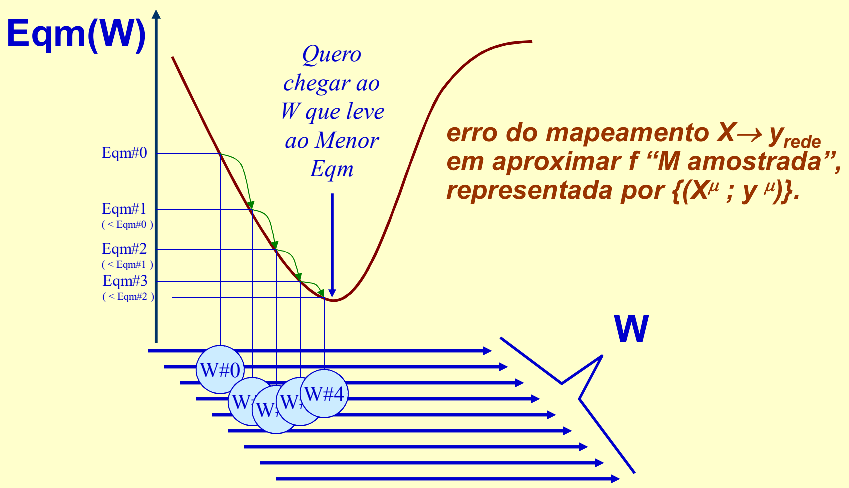
... O vetor gradiente indica a direção ascendente e seu módulo a magnitude de crescimento da função escalar – ilustração p/ função de 2 variáveis apenas



44

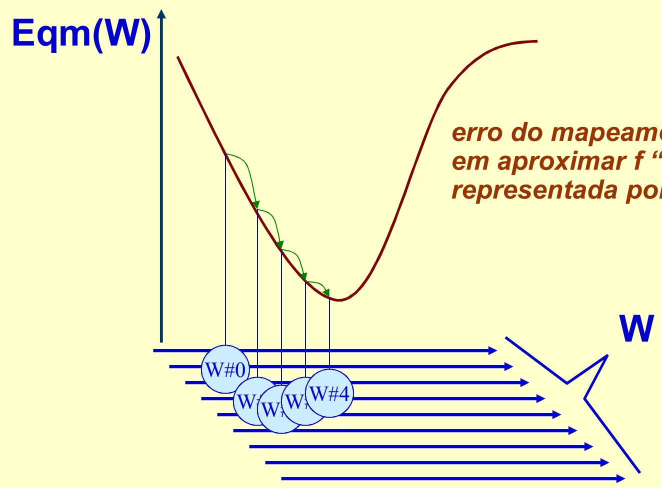
44

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

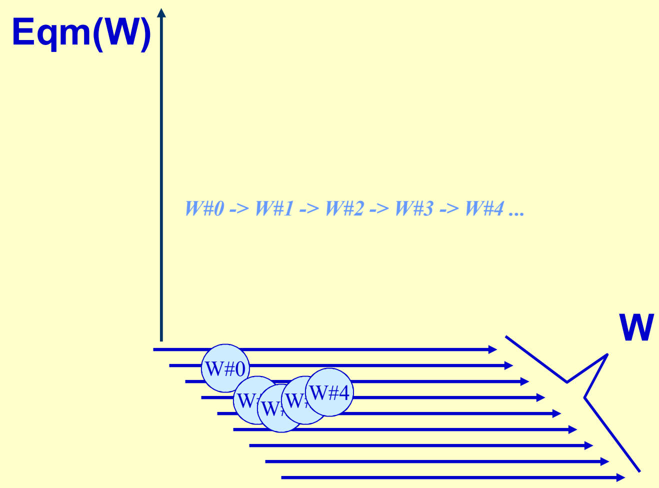


48

© Prof. Emilio Del Moral – EPUSP

48

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

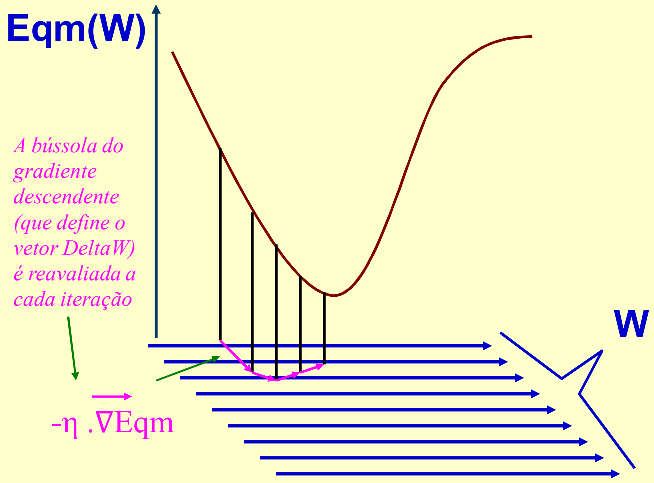


49

© Prof. Emilio Del Moral – EPUSP

49

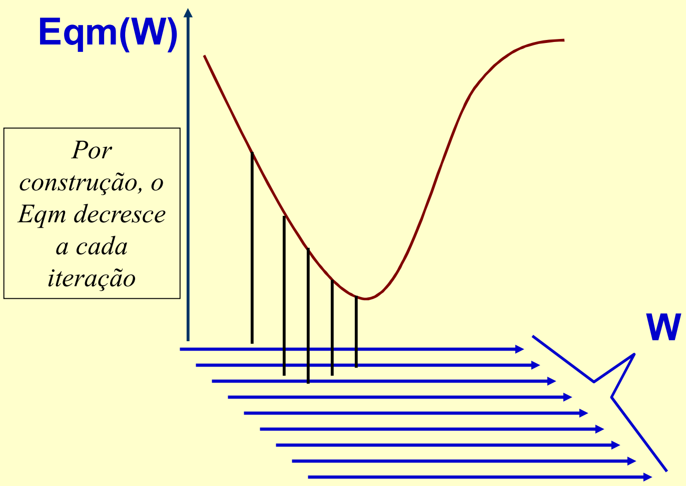
Foco nos diferentes DeltaW de cada iteração ...



50

50

Foco na evolução do Eqm com as iterações ...



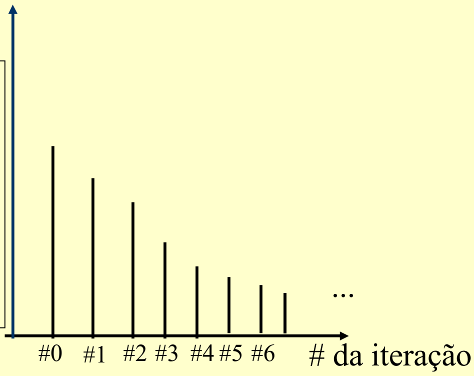
51

51

Plotando a evolução do Eqm com as iterações ...

Eqm(#)

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



Como o que estamos aprendendo no domínio de equações analíticas se relaciona com o exercitado no ambiente MBP?

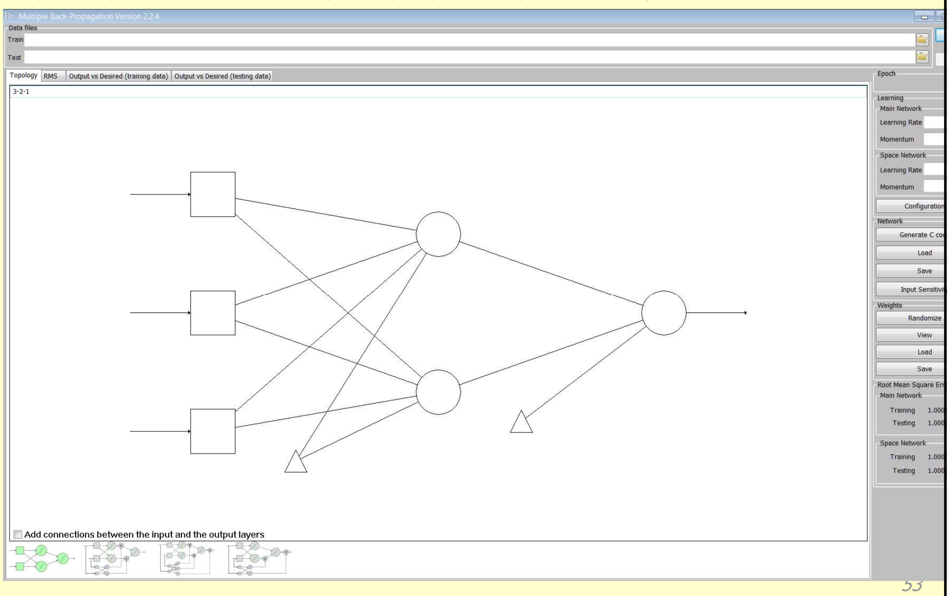
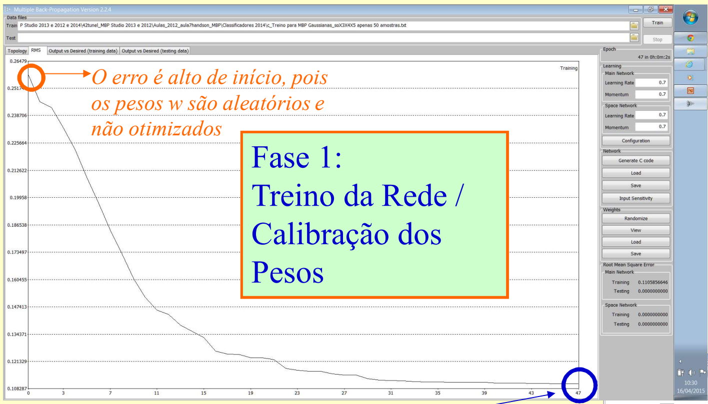


Gráfico fornecido pelo ambiente MBP da evolução do Eqm com o número de repetidos usos da "bússola do gradiente descendente": isto conecta o MBP com o gráfico apresentado no slide anterior



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

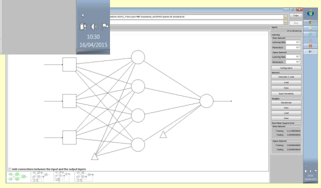


Gráfico mostrando as primeiras 47 iterações do processo de refinamentos sucessivos do modelo ... O erro é baixo ao final, pois os pesos vão sendo paulatinamente melhorados

55

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 novos 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.

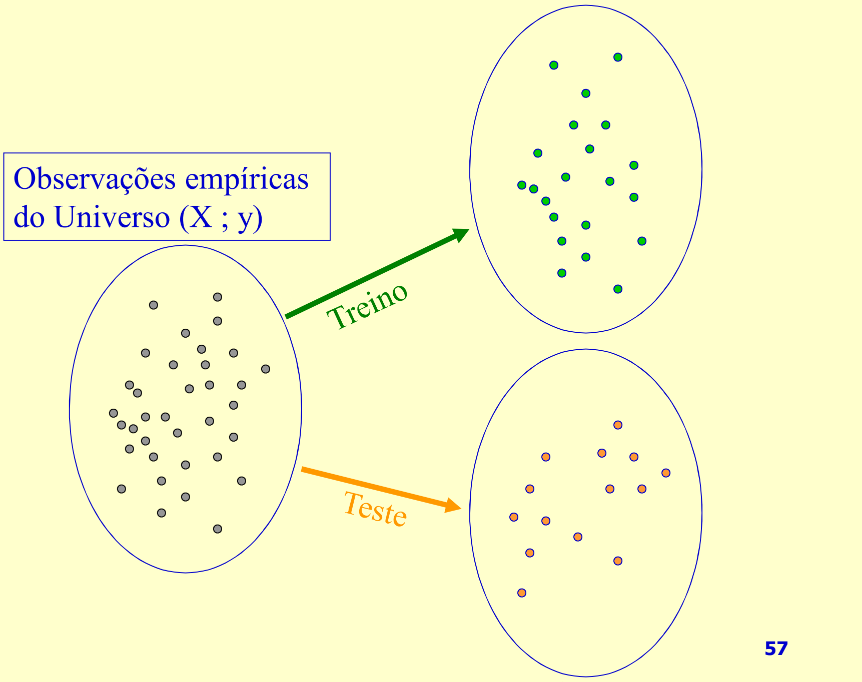
[Fase de refinamentos da RNA, dados e modelo, em ciclos, desde 0]

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

56

56



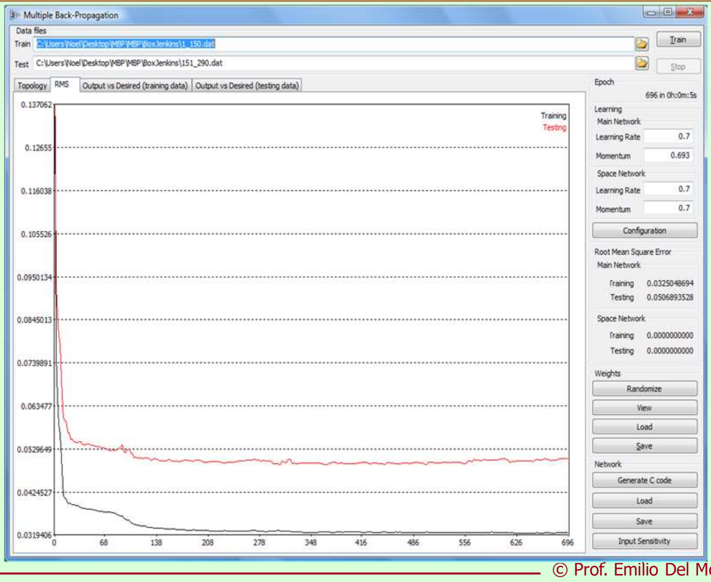
57

© Prof. Emilio Del Moral Hernandez

57

Algumas Telas do MBP: acompanhando a queda do erro (RMS) da rede neural enquanto está aprendendo / adaptando seus pesos

(nota: O Eqm é um primo direto do RMS ... RMS = $Eqm^{1/2}$)



© Prof. Emilio Del Moral Hernandez

58

Aprendizado do MLP por Error Back Propagation ...

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

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

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

83

83

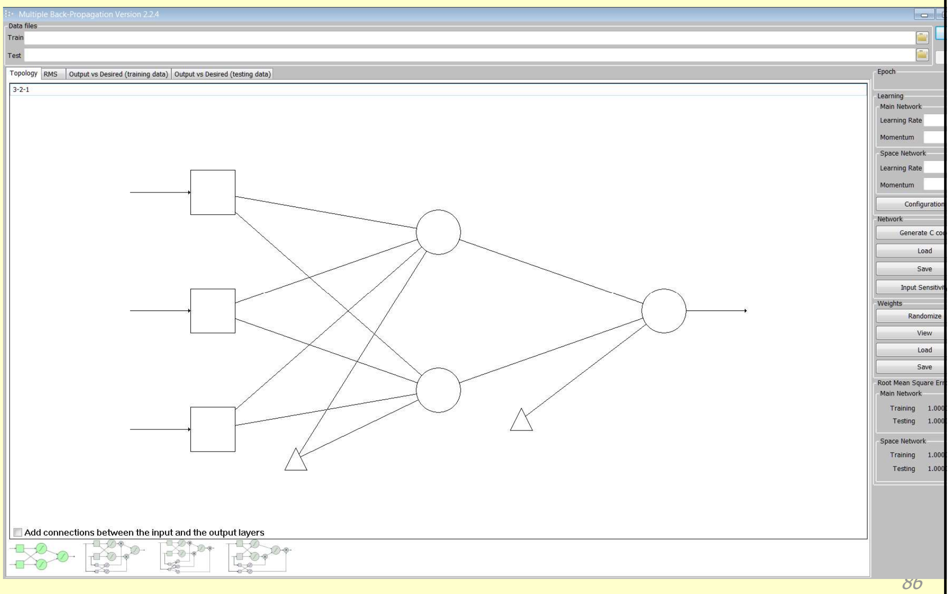
Processo de refinamentos graduais a cada iteração ...

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)
...	...	...	...

85

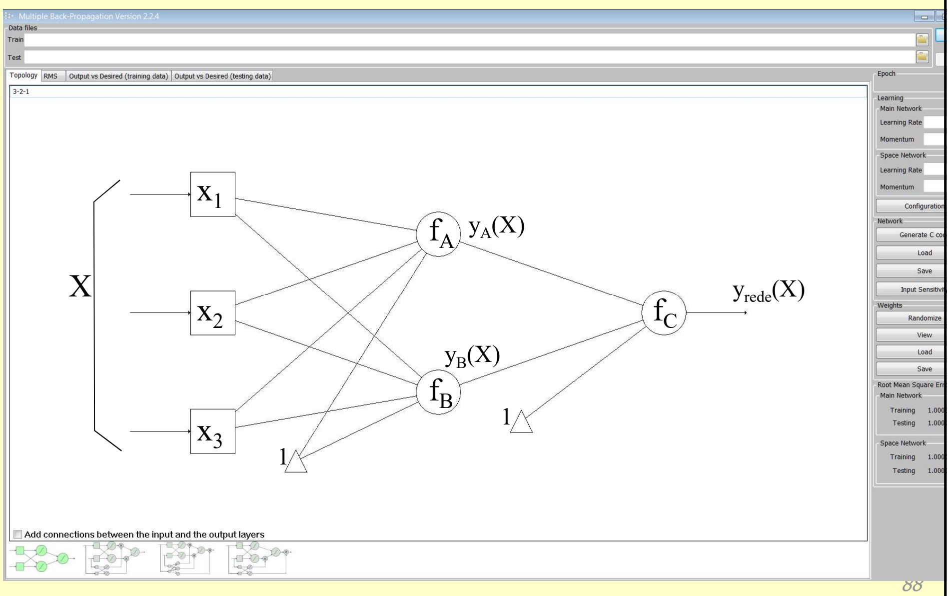
85

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



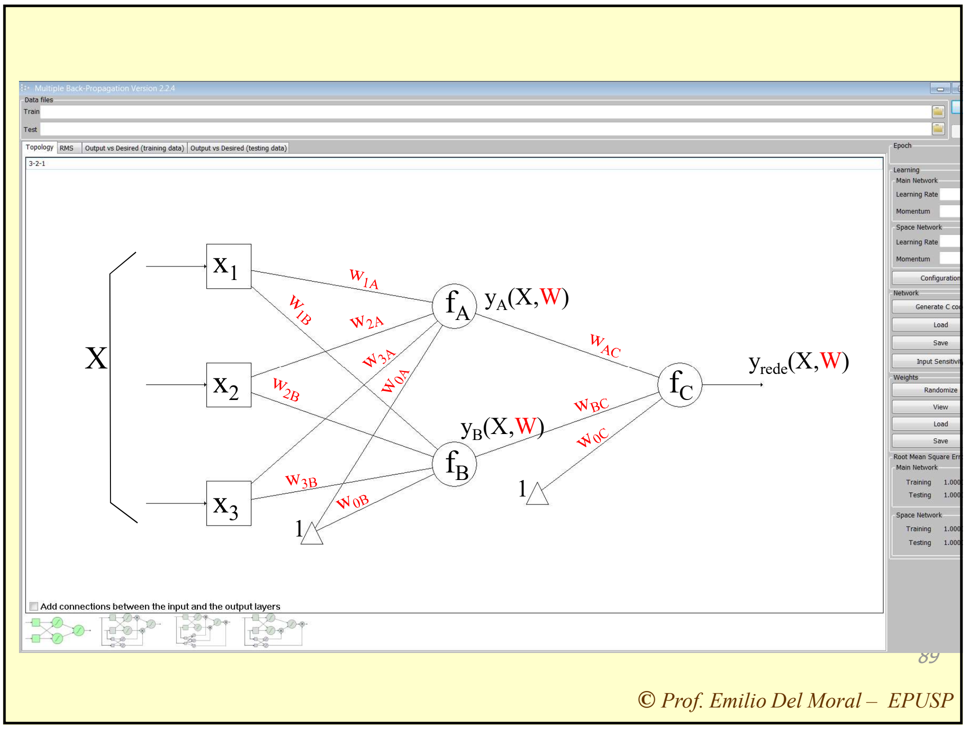
© Prof. Emilio Del Moral – EPUSP

86

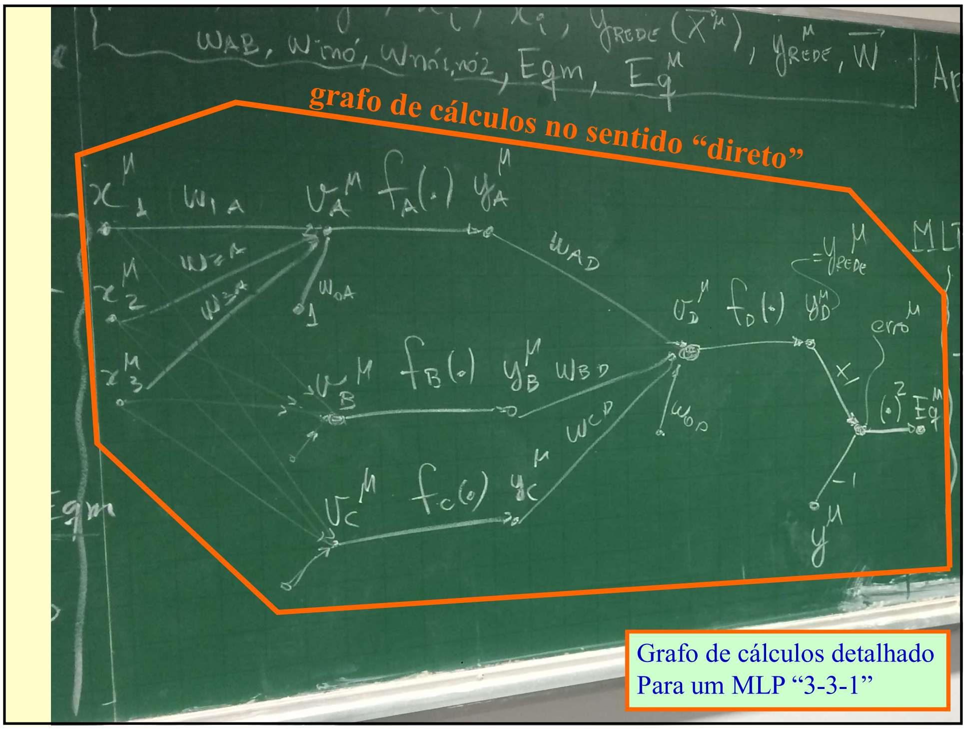


© Prof. Emilio Del Moral – EPUSP

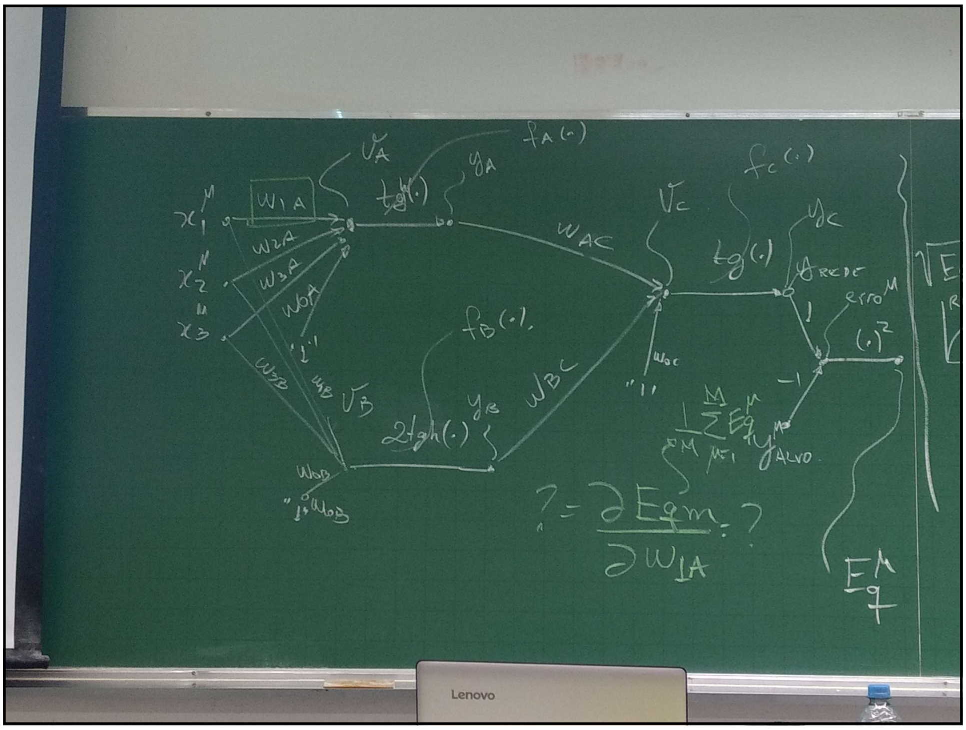
88


X
x1
x2
x3
w1A
w1B
w2A
w3A
w0A
w2B
w3B
w0B
fA
yA(X,W)
fB
yB(X,W)
wAC
wBC
w0C
fC
yrede(X,W)
1
1
Add connections between the input and the output layers
© Prof. Emilio Del Moral – EPUSP

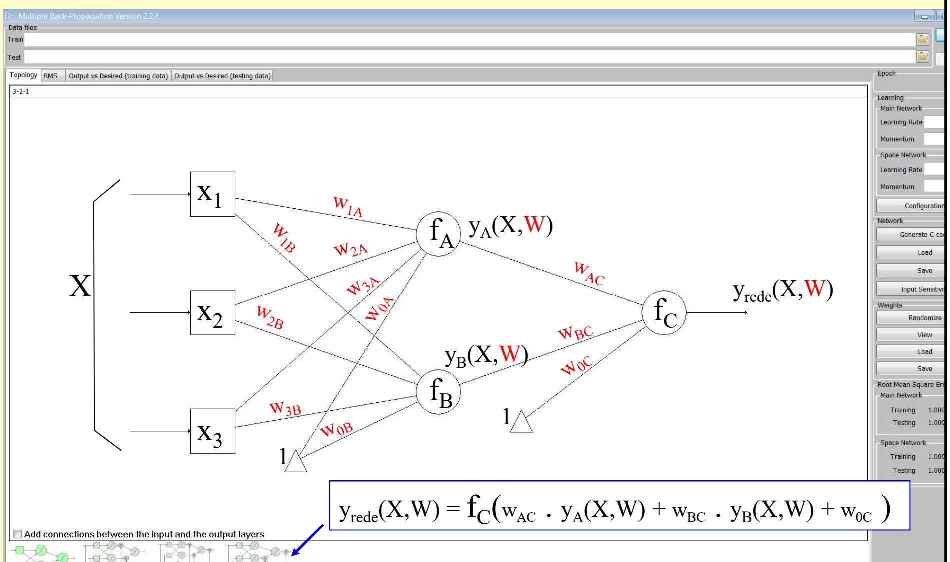
89


grafo de cálculos no sentido “direto”
Grafo de cálculos detalhado
Para um MLP “3-3-1”

90



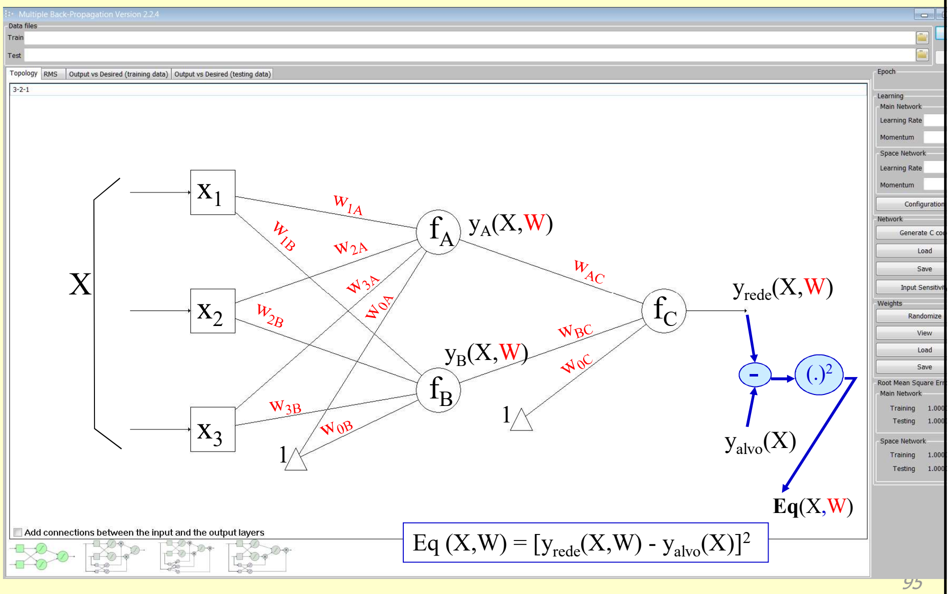
91



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

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

94



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:**

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

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

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) =

$$\textbf{Grad}(\Sigma_\mu \textbf{Eq}^\mu) / \textbf{M}$$

$$\Sigma_\mu \textbf{Grad} (\textbf{Eq}^\mu) / \textbf{M}$$

97

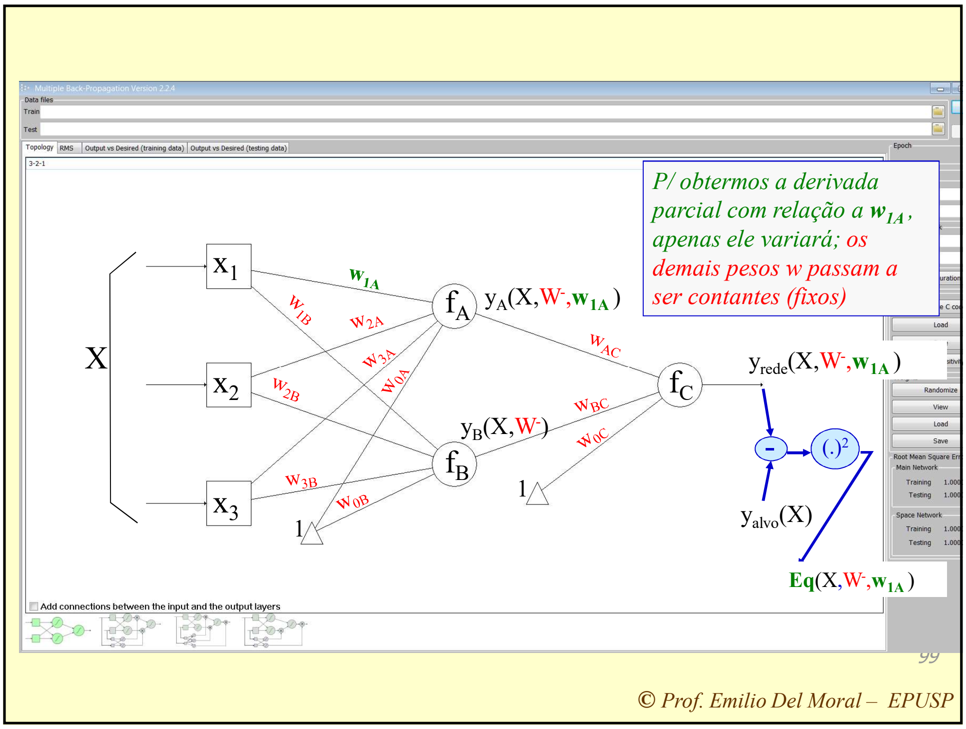
97

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 ***Grad(*** Σ_μ ***Eq***$^\mu$***)*** – a seguinte propriedade simples e sua velha conhecida ...

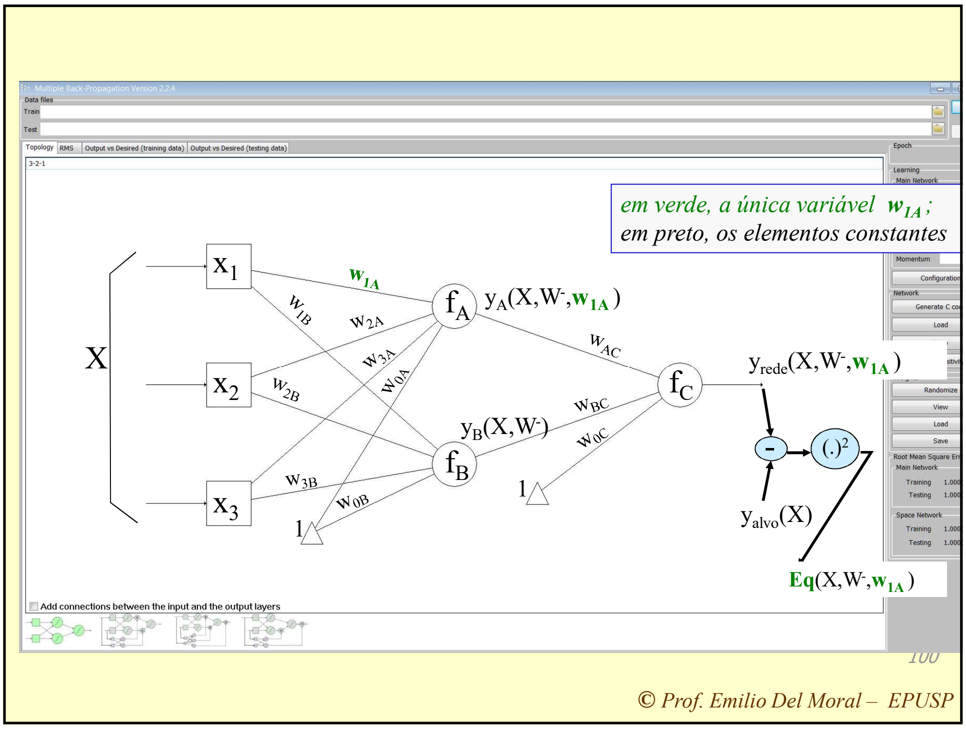
$$\mathbf{d(f_1(x)+f_2(x)) / dx \;=\; df_1(x)/dx + df_2(x)/dx}$$

98

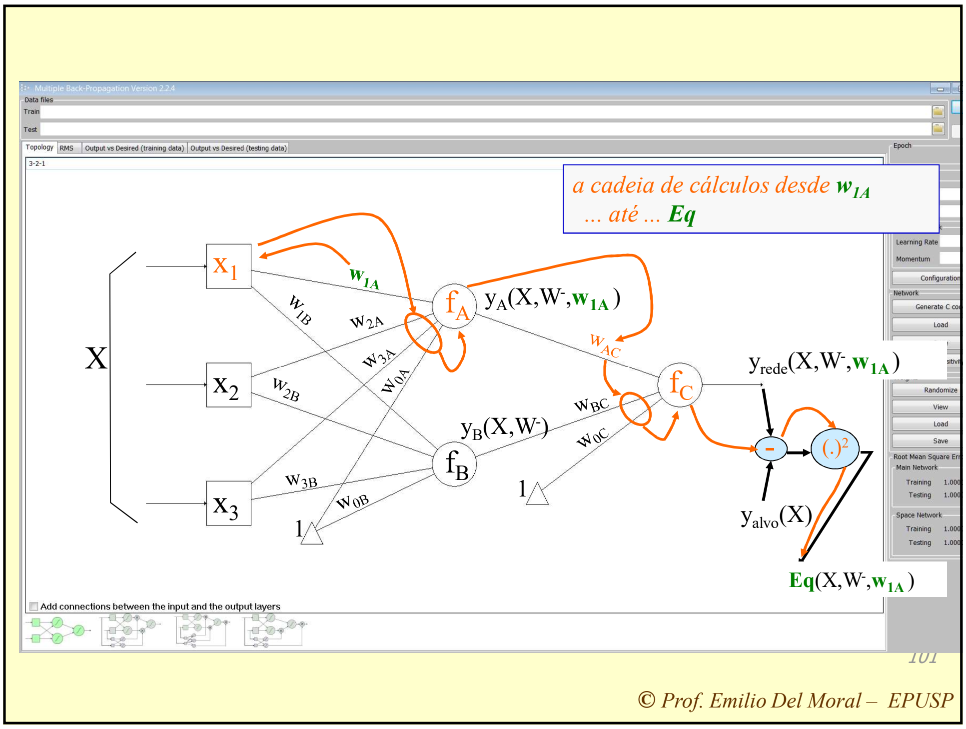
98


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_{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}
$y_{rede}(X,W^-,w_{1A})$
$(.)^2$
$y_{alvo}(X)$
$Eq(X,W^-,w_{1A})$
Add connections between the input and the output layers
© Prof. Emilio Del Moral – EPUSP

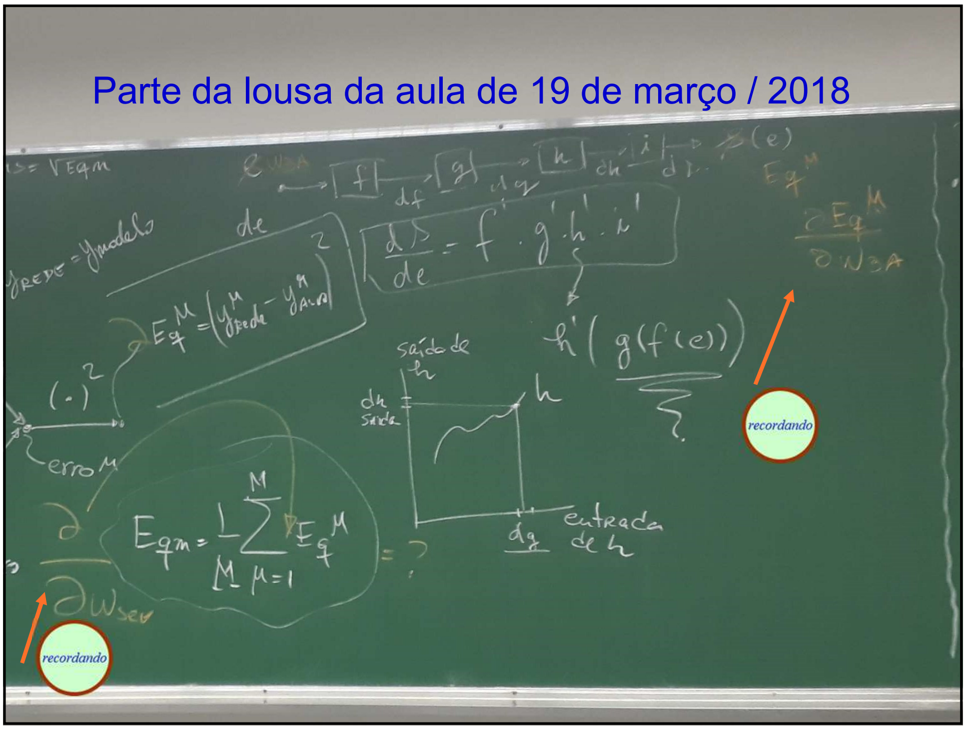
99


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_{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}
$y_{rede}(X,W^-,w_{1A})$
$(.)^2$
$y_{alvo}(X)$
$Eq(X,W^-,w_{1A})$
Add connections between the input and the output layers
© Prof. Emilio Del Moral – EPUSP

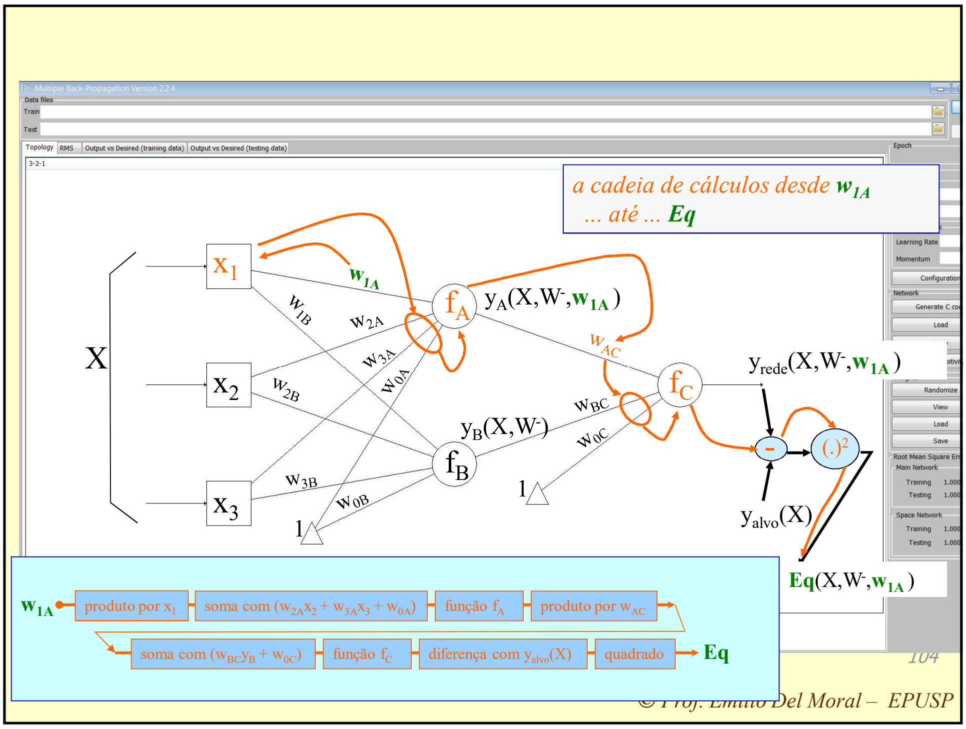
100


a cadeia de cálculos desde w1A
... até ... Eq
X
x1
x2
x3
w1A
w1B
w2A
w3A
w0A
w2B
w3B
w0B
fA
yA(X,W-,w1A)
fB
yB(X,W-)
wAC
wBC
w0C
fC
yrede(X,W-,w1A)
yalvo(X)
(.)2
Eq(X,W-,w1A)
Add connections between the input and the output layers
© Prof. Emilio Del Moral – EPUSP

101


Parte da lousa da aula de 19 de março / 2018
recordando
recordando

103



104

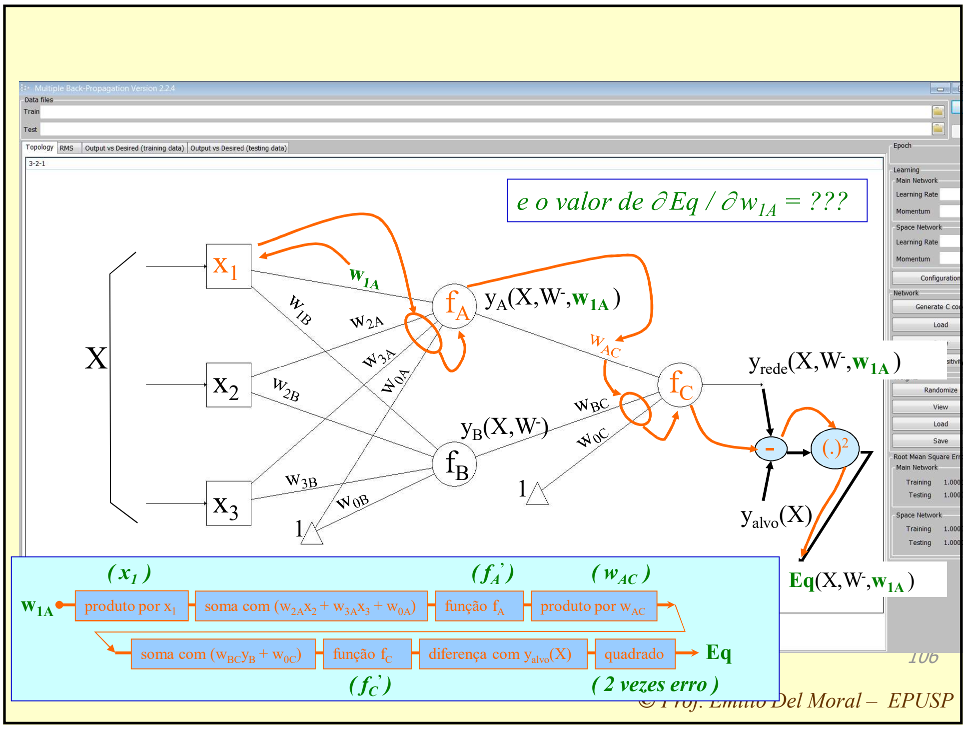
Note que aqui temos uma cadeia com muitos estágios que levam da váriável w_{1A}, à variável Eq^{μ}, e para a qual podemos calcular a derivada da saída (Eq^{μ}) com relação à entrada (w_{1A}) aplicando de forma repetida a seguinte propriedade simples e sua velha conhecida ...

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

..., 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.

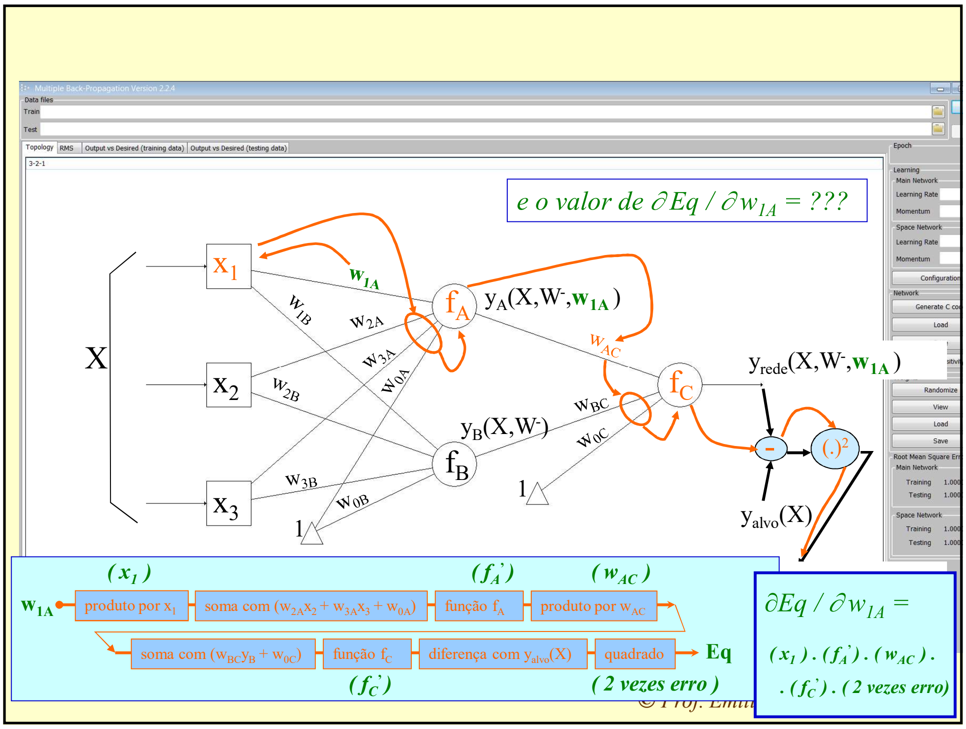
105

105



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

106



e o valor de ∂Eq / ∂w1A = ???
X
x1
x2
x3
w1A
fA
yA(X,W-,w1A)
wAC
fC
yrede(X,W-,w1A)
yB(X,W-)
fB
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)

107

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: $f_C(w_{AC} \cdot f_A(w_{1A} \cdot x_1 + w_{2A} \cdot x_2 + ... + w_{0A}) + w_{BC} \cdot f_B(w_{1B} \cdot x_1 + w_{2B} \cdot x_2 + ... + w_{0B}) + w_{0C})$
  - para f_A: $f_A(w_{1A} \cdot x_1 + w_{2A} \cdot x_2 + ... + w_{0A})$
- Similarmente, para o bloco "quadrado", cuja derivada é a função "2 vezes erro", o argumento é $[y_{rede}(X,W) - y_{alvo}(X)]$

112

112

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^{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!

113

113

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**

114

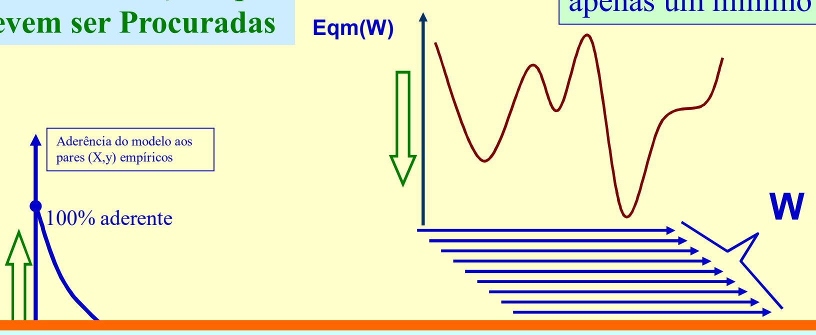
114

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

Será que temos apenas um mínimo??



Atacando a possibilidade de múltiplos mínimos locais:

... Executamos o gradiente descendente repetidamente, com a mesma arquitetura neural e os mesmos dados empíricos de treino, mas com os pesos iniciais randômicos sendo distintos em cada rodada; anotamos todos os pesos otimizados e valores de Eqm final de todas as rodadas; mantemos os w´s do melhor dos resultados de treino entre eles (aquele ensaio com menor valor do Eqm)!

115