Demonstrații: - pe scurt (schițe)

1) $\boxed{\nabla_z L = \hat{y} - y}$

• **Cazul binar:** Pornim de la $L = -1 \cdot \log \frac{1}{1+e^{-z}}$ (Analog, cazul $y=0$ prin complementaritate) $L = -\log\left(\frac{e^{-z}}{1+e^{-z}}\right)$, $1 - \frac{1}{e^{-z}+1} = \frac{e^{-z}}{1+e^{-z}}$

$$\hat{y} = \frac{1}{1+e^{-z}} \Rightarrow \hat{y} - y = \frac{1}{1+e^{-z}} - 1 = \frac{-e^{-z}}{1+e^{-z}}$$

$$\nabla_z L = -(1+e^{-z}) \cdot \frac{-1}{(1+e^{-z})^2} \cdot (-e^{-z}) = \frac{-e^{-z}}{1+e^{-z}} = \hat{y} - y$$

$$\nabla_z L = \frac{1+e^{-z}}{e^{-z}} \cdot \frac{e^{-z}}{(1+e^{-z})^2}$$

• **Cazul multiclass:** Fie i indicele pt. care $y[i]=1$, not. $x = z[i]$

Pornim de la $L = -\log \frac{e^x}{\sum_t z[t]}$ $\quad \nabla_{z[i]} L$

$$\nabla_{z[i]} L = -\frac{\sum_t z[t]}{e^x} \cdot \frac{e^x \cdot \sum_t z[t] - (e^x)^2}{(\sum_t z[t])^2} = \frac{e^x + e^x \cdot \sum_{k \neq i} e^{z[k]} - e^x}{e^x \cdot \sum_t e^{z[t]}}$$

$$\Rightarrow \nabla_z L = \left[\frac{e^{z[j]}}{\sum_t z[t]} \cdots \frac{-\sum_{k\neq i} e^{z[k]}}{\sum_t z[t]} \cdots \right]$$

$$\nabla_z L = \begin{cases} \dfrac{e^{z[j]}}{\sum_t z[t]}, & j \neq i \rightsquigarrow \text{exact } softmax(z[j]) \\ -\dfrac{\sum_{k \neq i} e^{z[k]}}{\sum_t z[t]}, & j = i \rightarrow \text{exact } softmax(z[i]) - 1 \end{cases}$$

(pt. derivare în raport cu fiecare componentă a vectorului z, m-am folosit de (*))

$$\hat{y} - y = (softmax(z_1) \ldots softmax(z_i) \ldots softmax(z_n)) - (0 \ldots \underset{i}{1} \ldots 0)$$
$$= (softmax(z_1) \ldots softmax(z_i) - 1 \ldots softmax(z_n))$$

$\boxed{\nabla_W L = \nabla_z L \cdot x^T}$ (Tratăm direct cazul multiclass) — i rămâne indicele de mai sus

$$L = -\log \frac{e^{W^T[i] \cdot x + b}}{\sum_k e^{W^T[k] \cdot x + b}} = -\log \frac{e^{x^T \cdot W[i] + b[i]}}{\sum_k e^{x^T \cdot W[k] + b[k]}}$$

$i \Rightarrow 1, len(z)$

$$\underbrace{\nabla_{W[i]} L}_{y[i]=1} = -\frac{\sum_k e^{x^T \cdot W[k] + b[k]}}{e^{x^T \cdot W[i] + b[i]}} \cdot \frac{x^T e^{x^T \cdot W[i] + b[i]} \cdot \sum_k e^{x^T \cdot W[k] + b[k]} - x^T \left(e^{x^T \cdot W[i] + b[i]}\right)^2}{\left(\sum_k e^{x^T \cdot W[k] + b[k]}\right)^2}$$

Subject :

Date : ....../....../......

$$\frac{-\sum_{k \neq i} e^{x^T \cdot w[k] + b[k]}}{\sum_k e^{x^T \cdot w[k] + b[k]}} \cdot x^T = \underbrace{\nabla_{z[i]} L \cdot x^T}_{\text{dimensiune } |X| \times 1}$$

Pt. acest punct am considerat derivarea în raport cu fiecare dintre liniile lui W

$$\nabla_{\underset{k \neq i}{w[k]}} L = \frac{-\cancel{\sum_k e^{x^T \cdot w[k]}}}{\cancel{e^{x^T \cdot w[i]}}} \cdot - \frac{x^T \cdot e^{x^T \cdot w[k]} \cdot e^{x^T \cdot w[i]}}{\left(\sum_k e^{x^T \cdot w[k]}\right)^2}$$

$$= \underbrace{\nabla_{z[k]} L \cdot x^T}_{|X| \times 1}$$

$z[k]$

$\text{softmax}(z[k])$

3) $\boxed{\nabla_b L = \nabla_z L}$

$$L = -\log \frac{e^{x^T \cdot w[i] + b[i]}}{\sum_k e^{x^T \cdot w[k] + b[k]}}$$

$$\sum_{k \neq i} e^{x^T \cdot w[k] + b[k]} + e^{x^T \cdot w[i] + b[i]}$$

$$\nabla_{b[i]} L = \frac{-\cancel{\sum_k e^{x^T \cdot w[k] + b[k]}}}{\cancel{e^{x^T \cdot w[i] + b[i]}}} \cdot \frac{e^{x^T w[i] + b[i]} \cdot \sum_k e^{x^T \cdot w[k] + b[k]} - e^{x^T w[i] + b[i]} \cdot e^{x^T w[i] + b[i]}}{\left(\sum_k e^{x^T \cdot w[k] + b[k]}\right)^2}$$

$$= \frac{-\sum_{k \neq i} e^{x^T \cdot w[k] + b[k]}}{\sum_k e^{x^T \cdot w[k] + b[k]}} = \nabla_{z[i]} L = \nabla_z L[i]$$

$$\nabla_{\underset{k \neq i}{b[k]}} L = -\frac{\cancel{\sum_k e^{x^T \cdot w[k] + b[k]}}}{\cancel{e^{x^T \cdot w[i] + b[i]}}} \cdot - \frac{e^{x^T \cdot w[k] + b[k]} \cdot \cancel{e^{x^T \cdot w[i] + b[i]}}}{\left(\sum_k e^{x^T \cdot w[k] + b[k]}\right)^2}$$

$$= + \frac{e^{x^T \cdot w[k] + b[k]}}{\sum_k e^{x^T \cdot w[k] + b[k]}} = \nabla_{z[k]} L = \nabla_z L[k]$$

k fiind orice indice $\neq i$ pe cazul acesta

Punând elementele în ordine, $\nabla_b L = \nabla_z L$.