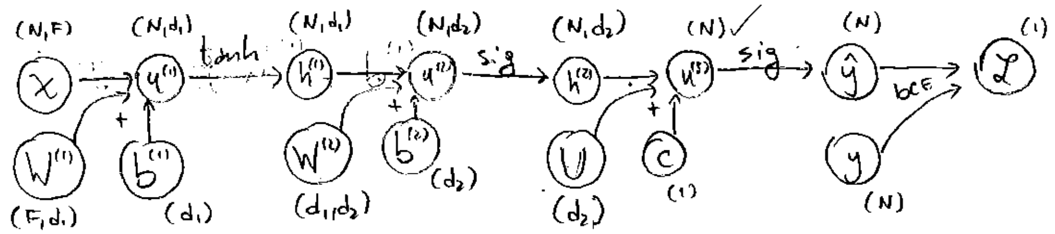
4) Ejemplo de Back Propagation completo a mano (1)



$$u^{(1)} = xW^{(1)} + b^{(1)} \qquad u^{(2)} = h^{(1)}W^{(2)} + b^{(2)} \qquad u^{(3)} = h^{(2)}U + c$$

$$h^{(1)} = \tanh(u^{(1)}) \qquad h^{(2)} = \mathrm{sig}(u^{(2)}) \qquad \hat{y} = \mathrm{sig}(u^{(3)})$$

$$\mathrm{error}(\hat{y}^{(i)}, y^{(i)}) = \underbrace{-\left(y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right)}_{H(P_{y^{(i)}},\ P_{\hat{y}^{(i)}})}$$

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{error}(\hat{y}^{(i)}, y^{(i)})$$

Interludio: Cross Entropy. Dados dos distribuciones de probabilidad P, Q sobre un dominio finito D la entropía cruzada entre P y Q (en ese orden) está dada por

$$-\sum_{v\in D} P(v)\log(Q(v))$$

← idea: cuantos bits "desperdicio" si asumo que mis datos distribuyen como Q cuando efectivamente distribuyen como P

En teo. información la entropía de una (variable aleatoria con) distribución de probabilidad P se calcula como

$$\sum_{v\in D} P(v)\underbrace{I(v)}_{\text{información que entrega el observar } v} = \sum_{v\in D} P(v)\log\left(\frac{1}{P(v)}\right) = -\sum_{v\in D} P(v)\log(P(v))$$

Promedio de la información obtenida al observar un valor al azar de la distribución. $\Rightarrow -\sum_{v\in D} P(v)\log(Q(v))$ promedio de la información obtenida al observar evento al azar según P si supongo (erroneamente) que los valores distribuyen según Q

Cantidad promedio de bits por evento que se necesitan (como mínimo) para codificar una fuente de datos según P. $\Rightarrow -\sum_{v\in D} P(v)\log(Q(v))$ es una medida de la cantidad de bits que se desperdician si se supone (erroneamente) que los datos distribuyen según Q

Volviendo a nuestro ejemplo: usemos entropía cruzada como error (por cada ejemplo)

- supongamos que sabemos que la respuesta de la red debe ser 1 ($y=1$) y nuestra red entrega el valor 0,89

$\Rightarrow$ entropía cruzada entre el valor empírico y la predicción es

$$-1 \cdot \log(0.89) = 0{,}117 \longleftarrow \text{error}$$

¿y si nuestra red entrega el valor 0,32?

$$-1 \cdot \log(0.32) = 1{,}139 \longleftarrow \text{error}$$

- supongamos que sabemos que la respuesta de la red debe ser 0 ($y=0$) y nuestra red entrega el valor 0,89

$$-1 \cdot \log(1-0{,}89) = 2{,}207 \longleftarrow \text{error}$$

¿y si entrega el valor 0,32?

$$-1 \cdot \log(1-0{,}32) = 0{,}386 \longleftarrow \text{error}$$

En general, si y es lo que esperamos ver ($y = 0$ ó 1) y la predicción de la red es $\hat{y}$ ($\hat{y}$ está entre $0, 1$)

$$\text{error}(\hat{y}, y) = \begin{cases} -\log(\hat{y}) & \text{si } y = 1 \\ -\log(1-\hat{y}) & \text{si } y = 0 \end{cases}$$

$$\text{error}(\hat{y}, y) = -\big(\underbrace{y}_{P_{empí}(y=1)} \log(\underbrace{\hat{y}}_{P_{NN}(y=1)}) + \underbrace{(1-y)}_{P_{empí}(y=0)} \log(\underbrace{1-\hat{y}}_{P_{NN}(y=0)})\big)$$

$\Rightarrow$ nuestra función de error es la entropía cruzada entre la distribución de probabilidad "empírica" y la generada por la red (para cada ejemplo)

4) Ejemplo Back Prop. (continuación)

Necesitamos $\frac{\partial \mathcal{L}}{\partial \theta}$ para cada parámetro θ. Lo hacemos con programación dinámica sobre el grafo de computación de la red.

$\frac{\partial \mathcal{L}}{\partial \hat{y}}$ $\quad \mathcal{L} = \frac{-1}{N} \sum_{i=1}^{N} \left(y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i) \right)$ $\quad \hat{y}_i = \hat{y}^{(i)}$, $y_i = y^{(i)}$ ← vector

$$\frac{\partial \mathcal{L}}{\partial \hat{y}}_i = \frac{-1}{N}\left(\frac{y_i}{\hat{y}_i} + \frac{(1-y_i)}{(1-\hat{y}_i)}\cdot -1\right) = \frac{1}{N}\left(\frac{(1-y_i)}{(1-\hat{y}_i)} - \frac{y_i}{\hat{y}_i}\right)$$

$$\frac{\partial \mathcal{L}}{\partial u^{(3)}_i} = \frac{\partial \mathcal{L}}{\partial \hat{y}_j} \cdot \underbrace{\frac{\partial \hat{y}_j}{\partial u^{(3)}_i}}_{=0 \text{ si } j \neq i} \qquad \hat{y}_j = \mathrm{sig}(u^{(3)}) \implies \hat{y}_i = \mathrm{sig}(u^{(3)}_i)$$

suma de Einstein !!!

$$\implies \frac{\partial \mathcal{L}}{\partial u^{(3)}_i} = \frac{\partial \mathcal{L}}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial u^{(3)}_i}$$

$$\boxed{\begin{aligned} \frac{\partial\, \mathrm{sig}(u)}{\partial u} &= \frac{\partial \frac{1}{1+e^{-u}}}{\partial u} \\ &= \frac{-1}{(1+e^{-u})^2} \cdot e^{-u} \cdot -1 \\ &= \frac{e^{-u}}{(1+e^{-u})} \cdot \frac{1}{(1+e^{-u})} \\ &= \frac{1+e^{-u}-1}{(1+e^{-u})} \cdot \frac{1}{(1+e^{-u})} \\ &= \left(1 - \frac{1}{1+e^{-u}}\right) \frac{1}{1+e^{-u}} \\ &= (1-\mathrm{sig}(u))\,\mathrm{sig}(u) \end{aligned}}$$

$$\frac{\partial \hat{y}_i}{\partial u^{(3)}_i} = \frac{\partial\, \mathrm{sig}(u^{(3)}_i)}{\partial u^{(3)}_i} = \underbrace{\mathrm{sig}(u^{(3)}_i)}_{\hat{y}_i} \underbrace{\left(1 - \mathrm{sig}(u^{(3)}_i)\right)}_{(1-\hat{y}_i)}$$

$$\frac{\partial \mathcal{L}}{\partial u^{(3)}_i} = \frac{1}{N}\left(\frac{(1-y_i)}{(1-\hat{y}_i)} - \frac{y_i}{\hat{y}_i}\right) \cdot \hat{y}_i(1-\hat{y}_i)$$

$$= \frac{1}{N}\left((1-y_i)\hat{y}_i - y_i(1-\hat{y}_i)\right) = \frac{1}{N}(\hat{y}_i - y_i)$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial u^{(3)}} = \frac{1}{N}(\hat{y} - y)} \;\star$$

$$\frac{\partial \mathcal{L}}{\partial U} \implies \frac{\partial \mathcal{L}}{\partial U_i} = \frac{\partial \mathcal{L}}{\partial u^{(3)}_j} \cdot \frac{\partial u^{(3)}_j}{\partial U_i} \qquad u^{(3)}_j = h^{(2)}_{jk} U_k + c$$

solo me importa cuando $k=i$!!

$$\implies \frac{\partial u^{(3)}_j}{\partial U_i} = \frac{\partial\, h^{(2)}_{jk} U_k + c}{\partial U_i} = \frac{\partial\, h^{(2)}_{ji} U_i}{\partial U_i} = h^{(2)}_{ji}$$

$$\frac{\partial \mathcal{L}}{\partial U_i} = \frac{\partial \mathcal{L}}{\partial u^{(3)}_j} \cdot h^{(3)}_{ji} \implies \boxed{\frac{\partial \mathcal{L}}{\partial U} = \frac{\partial \mathcal{L}}{\partial u^{(3)}} \cdot h^{(2)}} \;\star$$

$(d_2) \quad (N) \quad (N, d_2)$ ✓

$$\frac{\partial \mathcal{L}}{\partial c} = \frac{\partial \mathcal{L}}{\partial u^{(3)}_i} \frac{\partial u^{(3)}_i}{\partial c} = \frac{\partial \mathcal{L}}{\partial u^{(3)}_i} \mathbb{1}_i = \mathrm{sum}\left(\frac{\partial \mathcal{L}}{\partial u^{(3)}}\right) \implies \boxed{\frac{\partial \mathcal{L}}{\partial c} = \mathrm{sum}\left(\frac{\partial \mathcal{L}}{\partial u^{(3)}}\right)} \;\star$$

$u^{(3)}_i = h^{(2)}_{ik} U_k + c$

tensor de una dimensión con solo 1's

$$\frac{\partial \mathcal{L}}{\partial h^{(2)}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial h^{(2)}_{ij}} = \frac{\partial \mathcal{L}}{\partial u^{(3)}_k} \frac{\partial u^{(3)}_k}{\partial h^{(2)}_{ij}} \Longrightarrow \quad u^{(3)}_k = h^{(2)}_{kl} U_l + c$$

$$\frac{\partial u^{(3)}_k}{\partial h^{(2)}_{ij}} = \frac{\partial h^{(2)}_{kl} U_l}{\partial h^{(2)}_{ij}}$$

solo me q cuando $k=$ y $j=l$

$$= \frac{\partial \mathcal{L}}{\partial u^{(3)}_i} \frac{\partial u^{(3)}_i}{\partial h^{(2)}_{ij}} = \frac{\partial \mathcal{L}}{\partial u^{(3)}_i} \cdot U_j \Longrightarrow \frac{\partial \mathcal{L}}{\partial h^{(2)}_{ij}} = \frac{\partial \mathcal{L}}{\partial u^{(3)}_i} U_j$$

$$\Longrightarrow \boxed{\frac{\partial \mathcal{L}}{\partial h^{(2)}} = \frac{\partial \mathcal{L}}{\partial u^{(3)}} \otimes U} \;\star$$

$(N,d_2) \quad (N) \quad (d_2)$ ✓

(← ojo! en Pytorch no hay producto tensorial, pero acá se puede simular con producto de matrices $\underset{(N,1)}{\frac{\partial \mathcal{L}}{\partial u^{(3)}}} \cdot U_{(1,d_2)}$

$$\frac{\partial \mathcal{L}}{\partial u^{(2)}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial u^{(2)}_{ij}} = \frac{\partial \mathcal{L}}{\partial h^{(2)}_{kl}} \cdot \frac{\partial h^{(2)}_{kl}}{\partial u^{(2)}_{ij}}$$

← seguimos usando notación de Einstein

$$h^{(2)}_{kl} = sig\left(u^{(2)}_{kl}\right)$$

$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial u^{(2)}_{ij}} = \frac{\partial \mathcal{L}}{\partial h^{(2)}_{ij}} \cdot \frac{\partial h^{(2)}_{ij}}{\partial u^{(2)}_{ij}} \Longrightarrow \frac{\partial h_{kl}}{\partial u^{(2)}_{ij}} \neq 0 \iff k=i \text{ y } j=l$$

$$= \frac{\partial \mathcal{L}}{\partial h^{(2)}_{ij}} \cdot \frac{\partial\, sig\left(u^{(2)}_{ij}\right)}{\partial u^{(2)}_{ij}} = \frac{\partial \mathcal{L}}{\partial h^{(2)}_{ij}} \cdot \left(sig\left(u^{(2)}_{ij}\right)\left(1 - sig\left(u^{(2)}_{ij}\right)\right)\right)$$

$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial u^{(2)}} = \frac{\partial \mathcal{L}}{\partial h^{(2)}} * sig\left(u^{(2)}\right) * \left(1 - sig\left(u^{(2)}\right)\right)$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial u^{(2)}} = \frac{\partial \mathcal{L}}{\partial h^{(2)}} * h^{(2)} * \left(1 - h^{(2)}\right)} \;\star$$

$(N,d_2) \quad (N,d_2) \quad (N,d_2) \quad (N,d_2)$ ✓

ojo! "broadcast" hace la operación con todos los elementos del tensor

---

$$\frac{\partial \mathcal{L}}{\partial W^{(2)}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial W^{(2)}_{ij}} = \frac{\partial \mathcal{L}}{\partial u^{(2)}_{kl}} \cdot \frac{\partial u^{(2)}_{kl}}{\partial W^{(2)}_{ij}} \Longrightarrow \quad u^{(2)}_{kl} = h^{(1)}_{kr} W^{(2)}_{rl} + b^{(2)}_l$$

$$\frac{\partial h^{(1)}_{kr} W^{(2)}_{rl} + b^{(2)}_l}{\partial W^{(2)}_{ij}} \neq 0 \iff \begin{matrix} r=i \\ l=j \end{matrix}$$

en cuyo caso es simplemente $h^{(2)}_{ki}$

$$\Longrightarrow \frac{\partial u^{(2)}_{kl}}{\partial W^{(2)}_{ij}} = \begin{cases} h^{(1)}_{ki} & \text{si } l=j \\ 0 & \text{si } l \neq j \end{cases}$$

$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial W^{(2)}_{ij}} = \frac{\partial \mathcal{L}}{\partial u^{(2)}_{kj}} \cdot h^{(1)}_{ki} \Longrightarrow \boxed{\frac{\partial \mathcal{L}}{\partial W^{(2)}} = \left(h^{(1)}\right)^T \cdot \frac{\partial \mathcal{L}}{\partial u^{(2)}}} \;\star$$

$(d_1,d_2) \quad (d_1,N) \quad (N,d_2)$ ✓

4) Ejemplo Back Prop (continuación)

$$\frac{\partial \mathcal{L}}{\partial b^{(2)}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial b^{(2)}_i} = \frac{\partial \mathcal{L}}{\partial u^{(2)}_{kl}} \cdot \frac{\partial u^{(2)}_{kl}}{\partial b^{(2)}_i}$$

$u^{(2)}_{kl} = h^{(1)}_{kr} W^{(2)}_{rl} + b^{(2)}_l$ ← sólo esto importa

$$\frac{\partial u^{(2)}_{kl}}{\partial b^{(1)}_i} = \begin{cases} 1 & \text{si } l = i \ (\forall k) \\ 0 & \text{si } l \neq i \end{cases}$$

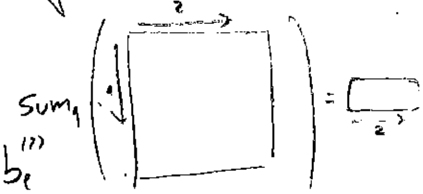
$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial b^{(2)}_i} = \frac{\partial \mathcal{L}}{\partial u^{(2)}_{ki}} \cdot \mathbb{1}_{ki}$$

$$\frac{\partial \mathcal{L}}{\partial b^{(2)}_i} = \sum_k \frac{\partial \mathcal{L}}{\partial u^{(2)}_{ki}}$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial b^{(2)}} = \mathrm{sum}_1\left(\frac{\partial \mathcal{L}}{\partial u^{(2)}}\right)} \ \star$$

(d_2) $\quad$ (N, d_2) $\quad$ sum $\checkmark$

como hay varias dimensiones debo especificar sobre cuál sumar



$$\frac{\partial \mathcal{L}}{\partial h^{(1)}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial h^{(1)}_{ij}} = \frac{\partial \mathcal{L}}{\partial u^{(2)}_{kl}} \cdot \frac{\partial u^{(2)}_{kl}}{\partial h^{(1)}_{ij}}$$

$u^{(2)}_{kl} = h^{(1)}_{kr} W^{(2)}_{rl} + b^{(2)}_l$

$$\frac{\partial u^{(2)}_{kl}}{\partial h^{(1)}_{ij}} = \begin{cases} W^{(2)}_{jl} & i = k \\ 0 & i \neq k \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial h^{(1)}_{ij}} = \frac{\partial \mathcal{L}}{\partial u^{(2)}_{il}} \cdot W^{(2)}_{jl} \Longrightarrow \boxed{\frac{\partial \mathcal{L}}{\partial h^{(1)}} = \frac{\partial \mathcal{L}}{\partial u^{(2)}} \cdot (W^{(2)})^T} \ \star$$

(N, d_1) $\quad$ (N, d_2) $\quad$ (d_2, d_1) $\checkmark$

$$\frac{\partial \tanh(u)}{\partial u} = \frac{\partial \frac{e^u - e^{-u}}{e^u + e^{-u}}}{\partial u} = 1 - \frac{(e^u - e^{-u})^2}{(e^u + e^{-u})^2} = 1 - (\tanh(u))^2$$

$$\frac{\partial \mathcal{L}}{\partial u^{(1)}} \Longrightarrow \frac{\partial \mathcal{L}}{\partial u^{(1)}_{ij}} = \frac{\partial \mathcal{L}}{\partial h^{(1)}_{kl}} \cdot \frac{\partial h^{(1)}_{kl}}{\partial u^{(1)}_{ij}}$$

$h^{(1)}_{kl} = \tanh(u^{(1)}_{kl})$

$$\Longrightarrow \frac{\partial h^{(1)}_{kl}}{\partial u^{(1)}_{ij}} \neq 0 \Longleftrightarrow k = i, \ l = j$$

$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial u^{(1)}_{ij}} = \frac{\partial \mathcal{L}}{\partial h^{(1)}_{ij}} \cdot \frac{\partial h^{(1)}_{ij}}{\partial u^{(1)}_{ij}} = \frac{\partial \mathcal{L}}{\partial h^{(1)}_{ij}} \cdot \left(1 - \tanh(u^{(1)}_{ij})^2\right)$$

$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial u^{(1)}} = \frac{\partial \mathcal{L}}{\partial h^{(1)}} * \left(1 - \tanh(u^{(1)}) * \tanh(u^{(1)})\right)$$

$$\Longrightarrow \boxed{\frac{\partial \mathcal{L}}{\partial u^{(1)}} = \frac{\partial \mathcal{L}}{\partial h^{(1)}} * \left(1 - h^{(1)} * h^{(1)}\right)} \ \star$$

(N, d_1) $\quad$ (N, d_1) $\quad$ (N, d_1) $\checkmark$

$$\boxed{\frac{\partial \mathcal{L}}{\partial W^{(1)}} = (\underbrace{h^{(0)}}_{x})^T \cdot \frac{\partial \mathcal{L}}{\partial u^{(1)}}} \ \star \qquad \boxed{\frac{\partial \mathcal{L}}{\partial b^{(1)}} = \mathrm{sum}_1\left(\frac{\partial \mathcal{L}}{\partial u^{(1)}}\right)} \ \star$$

se obtiene de manera análoga a $\frac{\partial \mathcal{L}}{\partial W^{(2)}}$

se obtiene de manera análoga a $\frac{\partial \mathcal{L}}{\partial b^{(2)}}$

Resumiendo:

backward

$$\frac{\partial \mathcal{L}}{\partial u^{(3)}} := \frac{1}{N}(\hat{y} - y)$$

$$\frac{\partial \mathcal{L}}{\partial U} := \frac{\partial \mathcal{L}}{\partial u^{(3)}} \cdot h^{(2)} \qquad \frac{\partial \mathcal{L}}{\partial c} := \text{sum}\left(\frac{\partial \mathcal{L}}{\partial u^{(3)}}\right) \qquad \frac{\partial \mathcal{L}}{\partial h^{(2)}} := \frac{\partial \mathcal{L}}{\partial u^{(3)}} \otimes U$$

$$\frac{\partial \mathcal{L}}{\partial u^{(2)}} := \frac{\partial \mathcal{L}}{\partial h^{(2)}} * h^{(2)} * (1 - h^{(2)})$$

$$\frac{\partial \mathcal{L}}{\partial W^{(2)}} := (h^{(1)})^T \cdot \frac{\partial \mathcal{L}}{\partial u^{(2)}} \qquad \frac{\partial \mathcal{L}}{\partial b^{(2)}} := \text{sum}_1\left(\frac{\partial \mathcal{L}}{\partial u^{(2)}}\right) \qquad \frac{\partial \mathcal{L}}{\partial h^{(1)}} := \frac{\partial \mathcal{L}}{\partial u^{(2)}} \cdot (W^{(2)})^T$$

$$\frac{\partial \mathcal{L}}{\partial u^{(1)}} := \frac{\partial \mathcal{L}}{\partial h^{(1)}} * (1 - h^{(1)} * h^{(1)})$$

$$\frac{\partial \mathcal{L}}{\partial W^{(1)}} = X^T \cdot \frac{\partial \mathcal{L}}{\partial u^{(1)}} \qquad \frac{\partial \mathcal{L}}{\partial b^{(1)}} = \text{sum}_1\left(\frac{\partial \mathcal{L}}{\partial u^{(1)}}\right)$$

Se pueden calcular eficientemente usando resultados intermedios y los valores ya calculados en la pasada forward !!