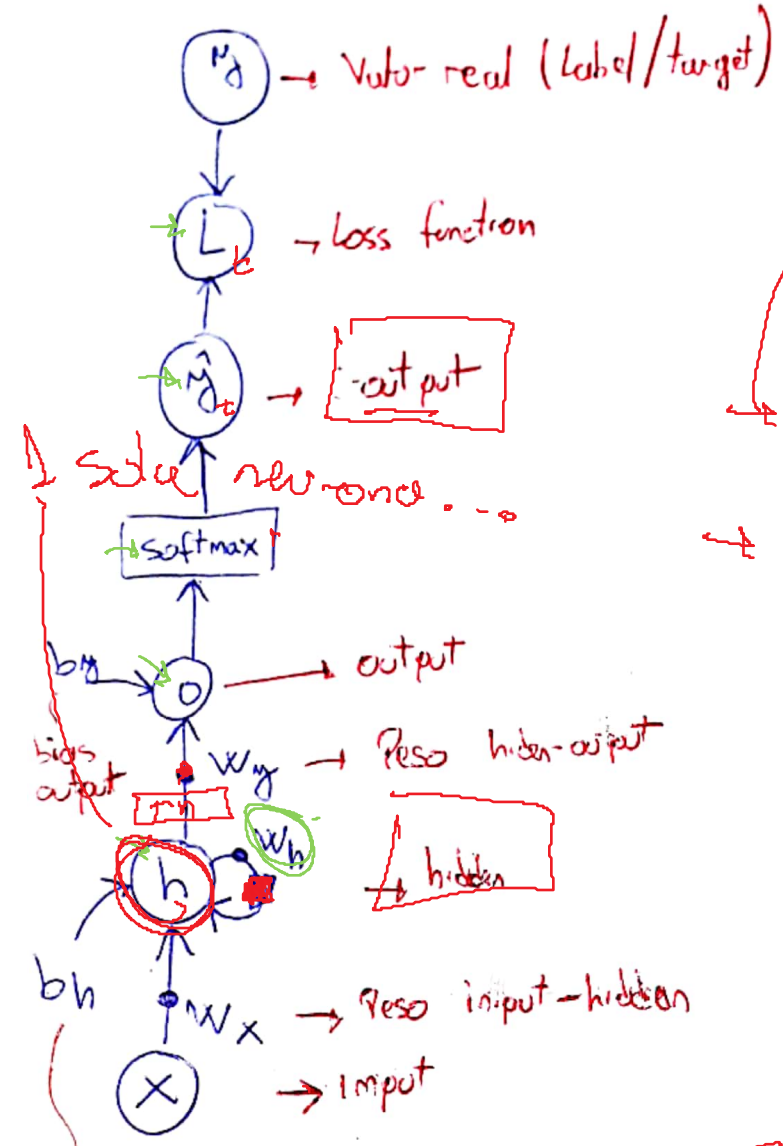
# Back propagation Through Time (BPTT)




## Ecuaciones

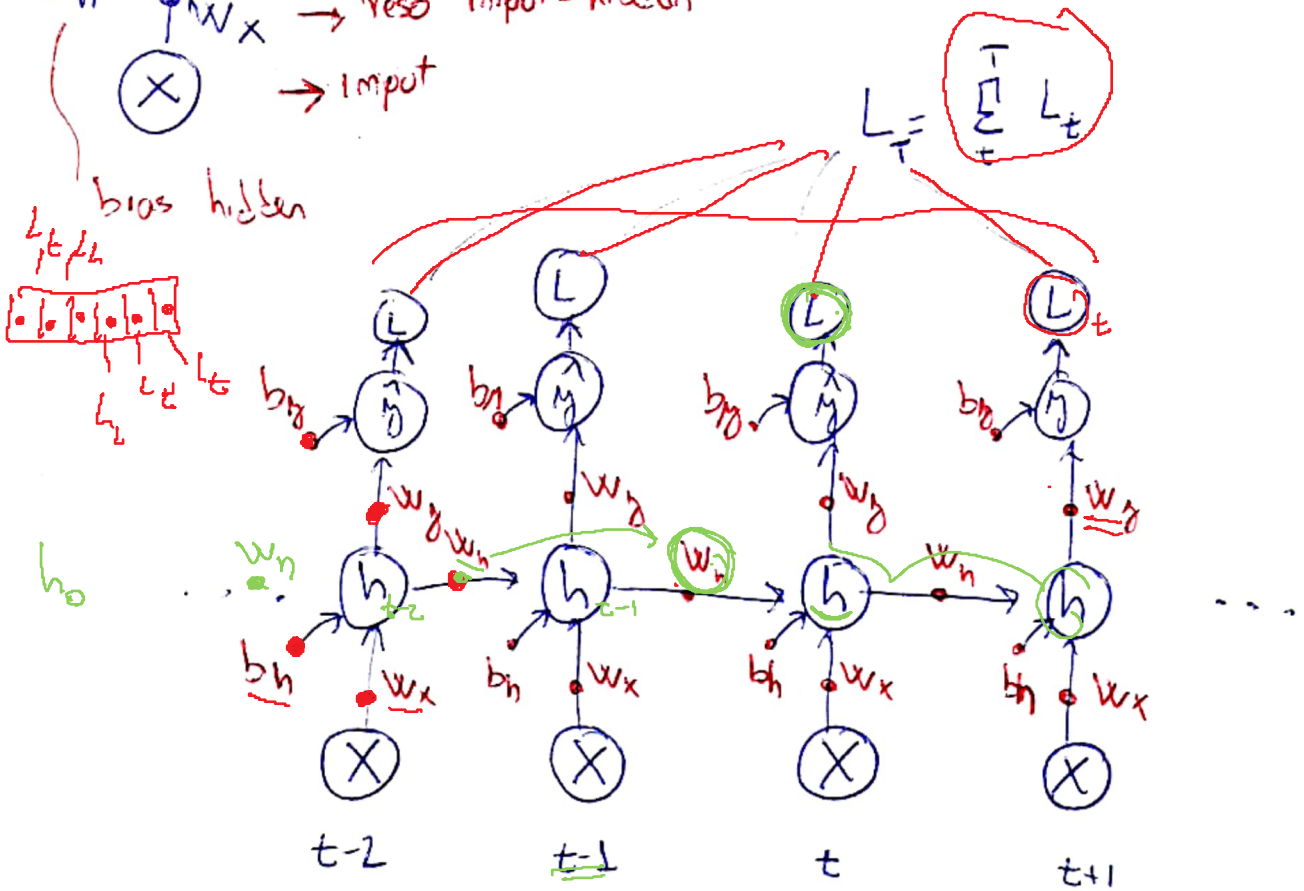
$$L_T = \sum_t^T L_t \quad (1)$$

$$L_t = -y_t \log \hat{y}_t \quad (2)$$

$$\hat{y}_t = \text{softmax}(o_t) = \frac{e^{o_t}}{\sum_k e^{o_k}} \quad (3)$$

$$o_t = h_t \cdot W_y + b_y \quad (4)$$

$$h_t = \tanh\left[X_t W_x + h_{t-1} W_h + b_h\right] \quad (5)$$



(II)

Parámetros a optimizar.

1° $W_y$ → Peso hidden - output

2° $b_y$ → bios output

3° $W_h$ → Peso hiden - hidden

4° $W_x$ → Peso input - hidden

5° $b_h$ → bios hiden

¿Como vara ni $L_T$ respecto de ellos?

1° $\frac{\partial L_T}{\partial W_y} = \frac{\partial}{\partial W_y} \sum_t^T L_t = \sum_t^T \frac{\partial L_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial W_y}$

Los pesos no varian con los $\Delta t$

sumo todos lo $t$ para tener una $L$ global

derivada de Loss function respecto de una softmax

Ya fué demostrado ← $(\hat{y}_t - y_t)$

de ecuacion (4) en hoja (I)

$o_t = W_y \cdot h_t + b_y$

$\frac{\partial o_t}{\partial W_y} = h_t$

- entonces llegamos a que $\frac{\partial L_t}{\partial W_y} = (\hat{y}_t - y_t)\, h_t \Rightarrow$

$$\frac{\partial L_T}{\partial W_y} = \sum_t^T (\hat{y}_t - y_t)\, h_t \quad (6)$$

2° $\frac{\partial L_T}{\partial b_y}$ se calcula igual q 1° reemplazando la última $\frac{\partial o_t}{\partial w_y}$ por $\frac{\partial' o_t}{\partial b_y}$

$o_t = w_y \cdot h_t + b_y$

$$\frac{\partial L_T}{\partial b_y} = \sum_t^T (\hat{y}_t - y_t) \cdot 1 \quad (7)$$

$\rightarrow \frac{\partial o_t}{\partial b_y}$

3° $\frac{\partial L_T}{\partial w_h}$ ¿Cómo varía $L_T$ cuando varío $w_h$?

$o_{t+1} = w_y h_{t+1} + b_y$

ojo! $h_t = \tanh(x w_x + h w_h + b_h)$

$$\frac{\partial L_{t+1}}{\partial w_h} = \frac{\partial L_{t+1}}{\partial \hat{y}_{t+1}} \cdot \frac{\partial \hat{y}_{t+1}}{\partial o_{t+1}} \cdot \frac{\partial o_{t+1}}{\partial h_{t+1}} \cdot \frac{\partial h_{t+1}}{\partial w_h} \quad (8)$$

Ya lo vieron $\leftarrow (\hat{y}_{t+1} - y_{t+1})$

de ecuación (4) en hoja (1) $= w_y$

a resolver...

$$h_{t+1} = \tanh(X_t w_x + h_t w_h + b_h)$$

$\frac{\partial h_{t+1}}{\partial w_h} = 0$

$h_{t+1}$ es función de $w_h$ y $h_t$

q es función de $w_h$ y $h_{t-1}$

... $w_h$ $h_{t-2}$

$$\frac{\partial h_{t+1}}{\partial w_h} = \frac{\partial h_{t+1}}{\partial h_t} \frac{\partial h_t}{\partial w_h} \Rightarrow \text{Podemos escribir:}$$

$$\frac{\partial h_{t+1}}{\partial w_h} = \sum_{k=1}^{t+1} \frac{\partial h_{t+1}}{\partial h_k} \frac{\partial h_k}{\partial w_h}$$

← veamos que da eso

$$\frac{\partial h_3}{\partial w_h} = \sum_{k=1}^{3} \frac{\partial h_3}{\partial h_k} \frac{\partial h_k}{\partial w_h}$$

$$= \frac{\partial h_3}{\partial h_1}\frac{\partial h_1}{\partial w_h} + \frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial w_h} + \frac{\partial h_3}{\partial h_3}\frac{\partial h_3}{\partial w_h}$$

$h_3$ varia porque $h_1$ varia por que $w_h$ varia

$h_3$ varia xq $h_2$ varia xq $w_h$ varia

$h_3$ varia xq $w_h$ varia

IV

$$h_3 = x_3 w_x + w_h h_2 + b_h$$

Recapitulamos ecuacion (8) de hoja III

$$\frac{\partial L_{t+1}}{\partial w_h} = \frac{\partial L_{t+1}}{\partial \hat{y}_{t+1}} \cdot \frac{\partial \hat{y}_{t+1}}{\partial o_{t+1}} \cdot \frac{\partial o_{t+1}}{\partial h_{t+1}} \cdot \sum_{k=1}^{t+1} \frac{\partial h_{t+1}}{\partial h_k} \cdot \frac{\partial h_k}{\partial w_h}$$

$w_h$

$(\hat{y}_{t+1} - y_{t+1})$

$w_y$

Esto es otra regla de la cadena mas!!

$$\frac{\partial h_{t+1}}{\partial h_k} = \prod_{j=k}^{t} \frac{\partial h_{j+1}}{\partial h_j} \rightarrow \frac{\partial h_{t+1}}{\partial h_k} \cdot \frac{\partial h_k}{\partial h_{k-1}} \cdots \frac{\partial h_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t+1}}{\partial h_t}$$

algunos terminos

el 1º

el último

Ahora si unimos todo

$$\frac{\partial L_{t+1}}{\partial w_h} = \frac{\partial L_{t+1}}{\partial \hat{y}_{t+1}} \cdot \frac{\partial \hat{y}_{t+1}}{\partial o_{t+1}} \cdot \frac{\partial o_{t+1}}{\partial h_{t+1}} \sum_{k=1}^{t+1} \left[ \prod_{j=k}^{t} \left( \frac{\partial h_{j+1}}{\partial h_j} \right) \cdot \frac{\partial h_k}{\partial w_h} \right]$$

$w_y$

obs! es para 1 solo tiempo!

$$\frac{\partial L_T}{\partial W_h} = \sum_{t}^{T} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial O_t} \cdot \frac{\partial O_t}{\partial h_t} \sum_{k=1}^{t} \left[ \prod_{j=k}^{t} \left( \frac{\partial h_j}{\partial h_{j-1}} \right) \frac{\partial h_k}{\partial W_h} \right]$$

(V)

1 x rec.


(9)

4° Para $\frac{\partial L_T}{\partial W_x}$ se computa igual que 3° $\left(\frac{\partial L_T}{\partial W_h}\right)$


[illegible] cada hidd. [illegible] función de $W_x$ y $h_{t-1}$

$$\frac{\partial L_T}{\partial W_x} = \sum_{t}^{T} \frac{\partial L_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial O_t} \cdot \frac{\partial O_t}{\partial h_t} \cdot \sum_{k=1}^{t} \left[ \prod_{j=k}^{t} \left( \frac{\partial h_j}{\partial h_{j-1}} \right) \cdot \frac{\partial h_k}{\partial W_x} \right]$$

$\hat{y} - y$ ; $W_y$

$$h_t = x_t W_x + h_{t-1} W_h + b_h$$

$$h_{t-1} = x_{t-1} W_x + h_{t-2} W_h + b_h$$

Obs. acá cambia

5° Igual procedimiento

$$\frac{\partial L_T}{\partial b_h} = \sum_{t}^{T} \frac{\partial L_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial O_t} \cdot \frac{\partial O_t}{\partial h_t} \sum_{k=1}^{t} \left[ \prod_{j=k}^{t} \left( \frac{\partial h_j}{\partial h_{j-1}} \right) \cdot \frac{\partial h_k}{\partial b_x} \right]$$

$= 1$ de ecuación (5) hasta (I)