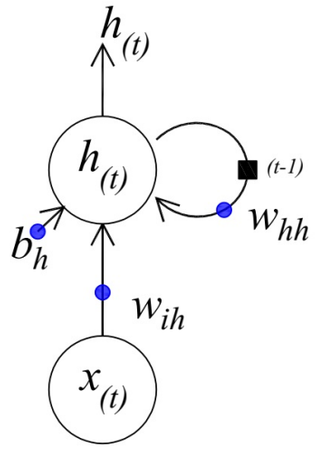
Back Propagation Through Time (BPTT)

Modelo a emplear:

$$h(t) = \tanh\Big(\underbrace{x_t \cdot w_{ih} + h_{t-1}\, w_{hh} + b_h}_{a(t)}\Big)$$



Parámetros a optimizar:

1° w_{hh} → Pesos hidden - hidden

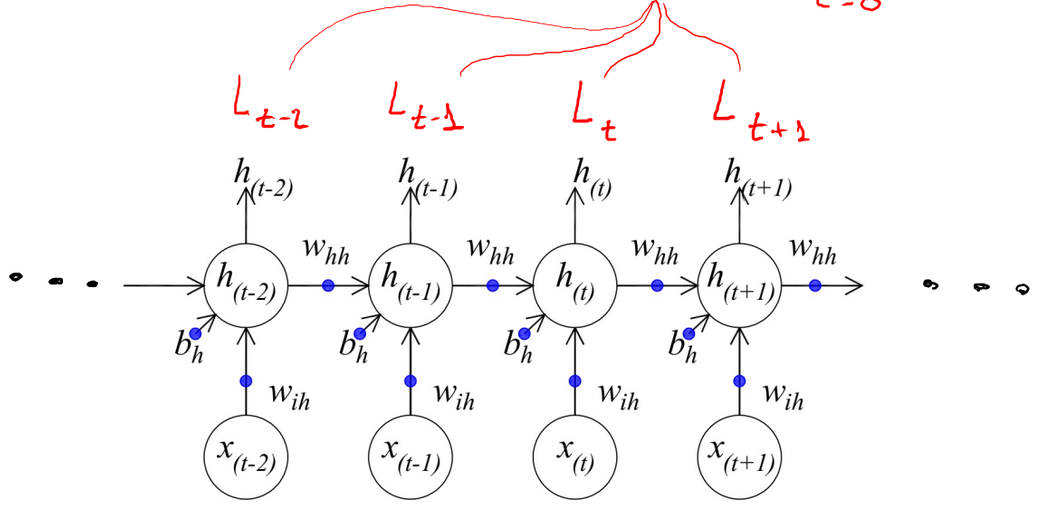
2° w_{ih} → Pesos input - hidden

3° b_h → bias hidden

¿como varia mi costo (L) respecto de ellos?

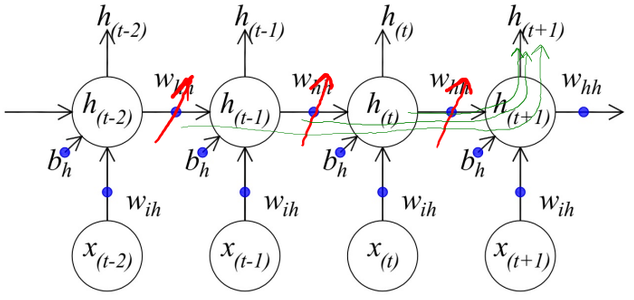
Si representamos la version "unfolded"

costo total → $L_T = \sum_{t=0}^{T} L_t$



1° w_{hh} → Pesos hidden – hidden

$$\frac{\partial L_T}{\partial w_{hh}} = ?$$



(A)

$$\frac{\partial L_{t+1}}{\partial w_{hh}} = \underbrace{\frac{\partial L_{t+1}}{\partial LF} \cdot \frac{\partial LF}{\partial h_{t+1}}}_{\text{derivada de Loss function respecto a } h_{t+1}} \cdot \underbrace{\frac{\partial h_{t-1}}{\partial w_{hh}}}_{\text{A RESOLVER!}}$$

$$h_{t+1} = \text{Tanh}\left(x_{t+1} \cdot w_{in} + h_{t-1} w_{hh} + b_h \right)$$

h_{t+1} es función de w_{hh} y $\underbrace{h_t}_{\text{que es función de } w_{hh} \text{ y } \underbrace{h_{t-1}}_{\ldots}}$

$$\frac{\partial h_{t+1}}{\partial w_{hh}} = \frac{\partial h_{t+1}}{\partial h_t} \cdot \frac{\partial h_t}{\partial w_{hh}}$$

se propone escribirlo como

$$\frac{\partial h_{t+1}}{\partial w_{hh}} = \sum_{k=0}^{t+1} \frac{\partial h_{t+1}}{\partial h_k} \cdot \frac{\partial h_k}{\partial w_{hh}}$$

ejemplo

$$\frac{\partial h_2}{\partial w_{hh}} = \sum_{k=0}^{2} \frac{\partial h_2}{\partial h_k} \cdot \frac{\partial h_k}{\partial w_{hh}} =$$

$$= \underbrace{\frac{\partial h_2}{\partial h_0} \frac{\partial h_0}{\partial w_{hh}}}_{h_2 \text{ varia porque } h_0 \text{ varia porque } w_{hh} \text{ varia}} + \underbrace{\frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial w_{hh}}}_{h_2 \text{ varia porque } h_1 \text{ varia porque } w_{hh} \text{ varia}} + \underbrace{\frac{\partial h_2}{\cancel{\partial h_2}} \frac{\cancel{\partial h_2}}{\partial w_{hh}}}_{h_2 \text{ varia porque } w_{hh} \text{ varia}}$$

Retomamos (A)

$$\frac{\partial L_{t+1}}{\partial w_{hh}} = \frac{\partial L_{t+1}}{\partial L_F} \cdot \frac{\partial L_F}{\partial h_{t+1}} \cdot \sum_{k=0}^{t+1} \frac{\partial h_{t+1}}{\partial h_k} \frac{\partial h_k}{\partial w_{hh}}$$

$\frac{\partial h_{t+1}}{\partial h_k}$: ¡ES OTRA REGLA DE LA CADENA!

$$\frac{\partial h_{t+1}}{\partial h_k} = \prod_{j=k}^{t} \frac{\partial h_{j+1}}{\partial h_j} \longrightarrow \frac{\partial h_{k+1}}{\partial h_k} \cdot \ldots \cdot \frac{\partial h_{t+1}}{\partial h_t}$$

algunos términos

$\frac{\partial h_{k+1}}{\partial h_k}$: el 1º término

$\frac{\partial h_{t+1}}{\partial h_t}$: el último término

Unimos todo...

$$\frac{\partial L_{t+1}}{\partial w_{hh}} = \frac{\partial L}{\partial L_F} \cdot \frac{\partial L_F}{\partial h_{t+1}} \cdot \sum_{k=0}^{t+1} \left[\prod_{j=k}^{t+1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \cdot \frac{\partial h_k}{\partial w_{hh}} \right]$$

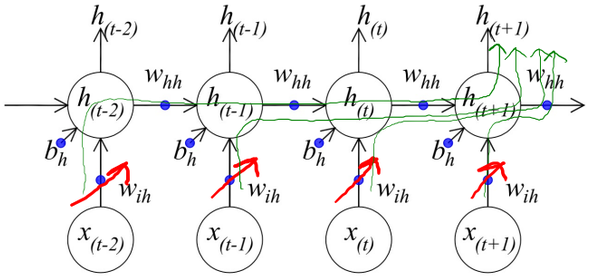
Solo para $t+1$!!

$$\frac{\partial L_T}{\partial w_{hh}} = \sum_{t=0}^{T} \frac{\partial L_t}{\partial L_F} \cdot \frac{\partial L_F}{\partial h_t} \sum_{k=0}^{t} \left[\prod_{j=k}^{t} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \cdot \frac{\partial h_k}{\partial w_{hh}} \right]$$

2º W_{ih} → Pesos input- hidden

Procedimiento semejante ya que

h_t es funcion de W_{ih} y h_{t-1}

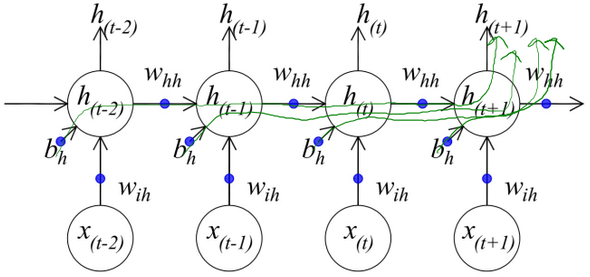


$$\frac{\partial L_T}{\partial W_{ih}} = \sum_{t=0}^{T} \frac{\partial L_t}{\partial L_F} \cdot \frac{\partial L_F}{\partial h_t} \sum_{k=0}^{t} \left[\prod_{j=k}^{t} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \cdot \frac{\partial h_k}{\partial W_{ih}} \right]$$

acá cambió

3º b_h → bias hidden

idem



$$\frac{\partial L_T}{\partial b_h} = \sum_{t=0}^{T} \frac{\partial L_t}{\partial L_F} \cdot \frac{\partial L_F}{\partial h_t} \sum_{k=0}^{t} \left[\prod_{j=k}^{t} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \cdot \frac{\partial h_k}{\partial b_h} \right]$$

acá cambió