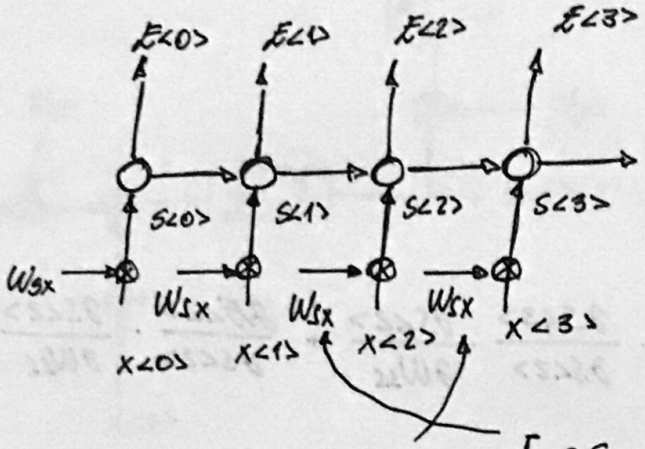
RNN Network Backprop through time



$$\frac{\partial E}{\partial W_{ax}} = \frac{\partial E^{<3>}}{\partial S^{<3>}} \cdot \frac{\partial S^{<3>}}{\partial W_{ax}} + \left[\frac{\partial E_3}{\partial S^{<3>}} \cdot \frac{\partial S^{<3>}}{\partial S^{<2>}} \cdot \frac{\partial S^{<2>}}{\partial W_{ax}} + \frac{\partial E^{<2>}}{\partial S^{<2>}} \cdot \frac{\partial S^{<2>}}{\partial W_{ax}} \right]$$

$$+ \left[\frac{\partial E^{<3>}}{\partial S^{<3>}} \cdot \frac{\partial S^{<3>}}{\partial S^{<2>}} \cdot \frac{\partial S^{<2>}}{\partial S^{<1>}} \cdot \frac{\partial S^{<1>}}{\partial W_{ax}} + \frac{\partial E^{<2>}}{\partial S^{<2>}} \frac{\partial S^{<2>}}{\partial S^{<1>}} \cdot \frac{\partial S^{<1>}}{\partial W_{ax}} + \right.$$

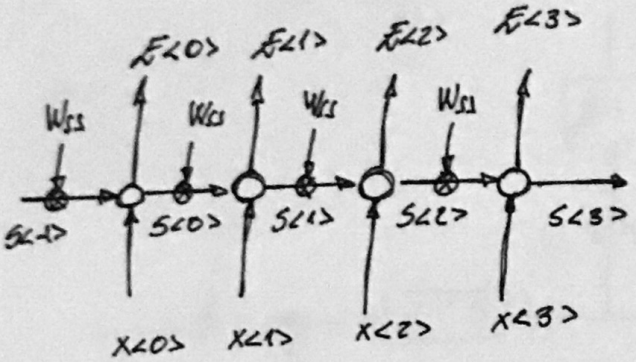
$$\left. + \frac{\partial E^{<1>}}{\partial S^{<1>}} \cdot \frac{\partial S^{<1>}}{\partial W_{ax}} \right] + \left[\frac{\partial E^{<3>}}{\partial S^{<3>}} \cdot \frac{\partial S^{<3>}}{\partial S^{<2>}} \cdot \frac{\partial S^{<2>}}{\partial S^{<1>}} \cdot \frac{\partial S^{<1>}}{\partial S^{<0>}} \frac{\partial S^{<0>}}{\partial W_{ax}} + \right.$$

$$+ \frac{\partial E^{<2>}}{\partial S^{<2>}} \cdot \frac{\partial S^{<2>}}{\partial S^{<1>}} \cdot \frac{\partial S^{<1>}}{\partial S^{<0>}} \cdot \frac{\partial S^{<0>}}{\partial W_{ax}} + \frac{\partial E^{<1>}}{\partial S^{<1>}} \cdot \frac{\partial S^{<1>}}{\partial S^{<0>}} \cdot \frac{\partial S^{<0>}}{\partial W_{ax}} +$$

$$\left. + \frac{\partial E^{<0>}}{\partial S^{<0>}} \frac{\partial S^{<0>}}{\partial W_{ax}} \right] = \sum_{K=3}^{3} \frac{\partial E^{<K>}}{\partial S^{<K>}} \cdot \prod_{j=3+1}^{K} \left(\frac{\partial S^{<j>}}{\partial S^{<j-1>}} \right) \cdot \frac{S^{<3>}}{\partial W_{ax}}$$

$$+ \sum_{K=2}^{3} \frac{\partial E^{<K>}}{\partial S^{<K>}} \cdot \prod_{j=3+1}^{K} \left(\frac{\partial S^{<j>}}{\partial S^{<j-1>}} \right) \cdot \frac{\partial S^{<2>}}{\partial W_{ax}} + \dots \longrightarrow$$

$$\frac{\partial E}{\partial W_{ax}} = \sum_{l=0}^{T_x} \sum_{K=l}^{T_y} \frac{\partial E^{<K>}}{\partial S^{<K>}} \prod_{j=l+1}^{K} \left(\frac{\partial S^{<j>}}{\partial S^{<j-1>}} \right) \cdot \frac{\partial S^{<l>}}{\partial W_{ax}}$$



$$\frac{\partial E}{\partial W_{ss}} = \left[\frac{\partial E\langle 3\rangle}{\partial S\langle 3\rangle}\,\frac{\partial S\langle 3\rangle}{\partial W_{ss}}\right] + \left[\frac{\partial E\langle 3\rangle}{\partial S\langle 3\rangle}\cdot\frac{\partial S\langle 3\rangle}{\partial S\langle 2\rangle}\cdot\frac{\partial S\langle 2\rangle}{\partial W_{ss}} + \frac{\partial E\langle 2\rangle}{\partial S\langle 2\rangle}\cdot\frac{\partial S\langle 2\rangle}{\partial W_{ss}}\right] +$$

$$+\left[\frac{\partial E\langle 3\rangle}{\partial S\langle 3\rangle}\cdot\frac{\partial S\langle 3\rangle}{\partial S\langle 2\rangle}\cdot\frac{\partial S\langle 2\rangle}{\partial S\langle 1\rangle}\cdot\frac{\partial S\langle 1\rangle}{\partial W_{ss}} + \frac{\partial E\langle 2\rangle}{\partial S\langle 2\rangle}\,\frac{\partial S\langle 2\rangle}{\partial S\langle 1\rangle}\cdot\frac{\partial S\langle 1\rangle}{\partial W_{ss}} +\right.$$

$$\left. + \frac{\partial E\langle 1\rangle}{\partial S\langle 1\rangle}\cdot\frac{\partial S\langle 1\rangle}{\partial W_{ss}}\right] + \left[\frac{\partial E\langle 0\rangle}{\partial S\langle 0\rangle}\cdot\frac{\partial S\langle 0\rangle}{\partial W_{ss}} + \frac{\partial E\langle 3\rangle}{\partial S\langle 3\rangle}\cdot\frac{\partial S\langle 3\rangle}{\partial S\langle 2\rangle}\cdot\frac{\partial S\langle 2\rangle}{\partial S\langle 1\rangle}\cdot\frac{\partial S\langle 1\rangle}{\partial S\langle 0\rangle}\right.$$

$$\left. + \frac{\partial E\langle 0\rangle}{\partial W_{ss}} + \frac{\partial E\langle 2\rangle}{\partial S\langle 2\rangle}\cdot\frac{\partial S\langle 2\rangle}{\partial S\langle 1\rangle}\cdot\frac{X\langle 1\rangle}{\partial S\langle 0\rangle}\cdot\frac{\partial S\langle 0\rangle}{\partial W_{ss}} + \frac{\partial E\langle 1\rangle}{\partial S\langle 1\rangle}\cdot\frac{\partial S\langle 1\rangle}{\partial S\langle 0\rangle}\cdot\frac{\partial S\langle 0\rangle}{\partial W_{ss}}\right] \Rightarrow$$

$$\frac{\partial E}{\partial W_{ss}} = \sum_{l=0}^{T_x}\sum_{k=l}^{T_y}\frac{\partial E_k}{\partial S\langle k\rangle}\prod_{j=l+1}^{k}\left(\frac{\partial S\langle j\rangle}{\partial S\langle j-1\rangle}\right)\frac{\partial S\langle l\rangle}{\partial W_{ss}}$$

$$\frac{\partial E}{\partial X\langle l\rangle} = \sum_{k=l}^{T_y}\frac{\partial E_k}{\partial S\langle k\rangle}\prod_{j=l+1}^{k}\left(\frac{\partial S\langle j\rangle}{\partial S\langle j-1\rangle}\right)\frac{\partial S\langle l\rangle}{\partial X\langle l\rangle}$$

* These consecutive multiplications may cause vanishing gradient or exploding gradient:

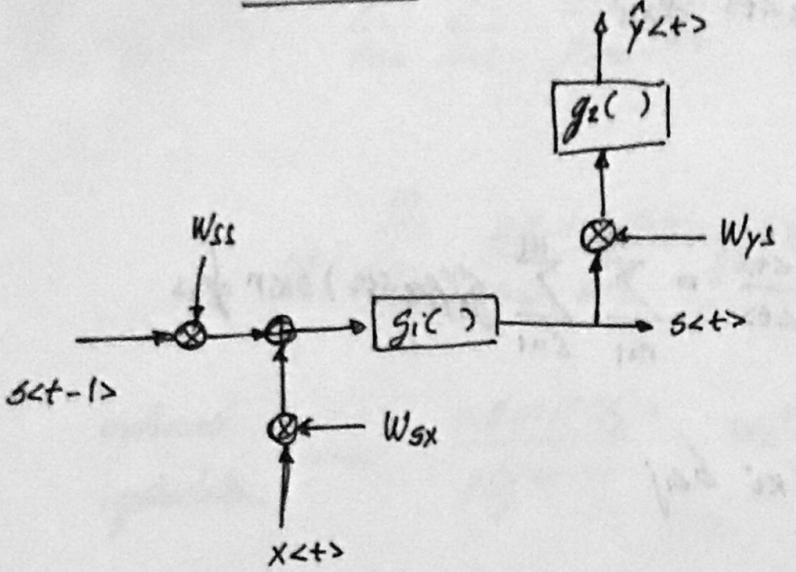
- Exploding gradient $\longrightarrow \frac{\partial E}{\partial \theta} \rightarrow \left\|\frac{\partial E}{\partial \theta}\right\| > \text{Threshold} \rightarrow \frac{\partial E}{\partial \theta} = \frac{\text{thres}}{\left\|\frac{\partial E}{\partial \theta}\right\|}\cdot\frac{\partial E}{\partial \theta}$

- Vanishing gradient $\longrightarrow$ LSTM, GRU

$$\frac{\partial S_{re}\langle t\rangle}{\partial S_{ij}\langle t-1\rangle} = \tanh'(z_{re}\langle t\rangle)\cdot W_{esi}\,\delta_{ej}$$

$0.X \leftarrow z \le -1$ or $z \ge 1$ $\qquad$ $\begin{cases} < 1 ; 0.X^T \rightarrow \text{Vanishing} \\ > 1 ; 1.X^T \rightarrow \text{Exploding} \end{cases}$

RNN Cell



* Forward →

$$s^{\langle t \rangle} = g_1(z^{\langle t \rangle})$$

$$z^{\langle t \rangle} = W_{ss}\, s^{\langle t-1 \rangle} + W_{sx}\, x^{\langle t \rangle} + b_s$$

$$\hat{y}^{\langle t \rangle} = g_2(p^{\langle t \rangle})$$

$$p^{\langle t \rangle} = W_{ys}\, s^{\langle t \rangle} + b_y$$

Dimensions → $x^{\langle t \rangle} \in \mathbb{R}^{(n_x, m)}$; $s^{\langle t \rangle} \in \mathbb{R}^{(n_s, m)}$; $y^{\langle t \rangle} \in \mathbb{R}^{(n_y, m)}$

$W_{ss} \in \mathbb{R}^{(n_s, n_c)}$; $W_{sx} \in \mathbb{R}^{(n_s, n_x)}$; $W_{ys} \in \mathbb{R}^{(n_y, n_s)}$

$$\frac{\partial s_{ke}^{\langle t \rangle}}{\partial z_{ij}^{\langle t \rangle}} = g_1'(z_{ke}^{\langle t \rangle})\, \delta_{ki}\, \delta_{ej} \quad ; \quad \frac{\partial \hat{y}_{ke}^{\langle t \rangle}}{\partial p_{ij}^{\langle t \rangle}} = g_2'(s_{ke}^{\langle t \rangle})\, \delta_{ki}\, \delta_{ej}$$

$$\frac{\partial z_{ke}^{\langle t \rangle}}{\partial W_{ss_{ij}}} = \delta_{ki}\, s_{je}^{\langle t-1 \rangle} \quad ; \quad \frac{\partial z_{ke}^{\langle t \rangle}}{\partial W_{sx_{ij}}} = \delta_{ki}\, x_{je}^{\langle t \rangle} \quad ; \quad \frac{\partial z_{ke}^{\langle t \rangle}}{\partial s_{ij}^{\langle t-1 \rangle}} = W_{ss_{ki}}\, \delta_{ej}$$

$$\frac{\partial z_{ke}^{\langle t \rangle}}{\partial x_{ij}^{\langle t \rangle}} = W_{sx_{ki}}\, \delta_{ej} \quad ; \quad \frac{\partial p_{ke}^{\langle t \rangle}}{\partial W_{ys_{ij}}} = \delta_{ki}\, s_{je}^{\langle t \rangle} \quad ; \quad \frac{\partial p_{ke}^{\langle t \rangle}}{\partial s_{ij}^{\langle t \rangle}} = W_{ys_{ki}}\, \delta_{ej}$$

$$\frac{\partial s_{ke}^{\langle t \rangle}}{\partial x_{ij}^{\langle t \rangle}} = \sum_{r=1}^{n_a} \sum_{s=1}^{m} \frac{\partial s_{ke}^{\langle t \rangle}}{\partial z_{rs}^{\langle t \rangle}} \frac{\partial z_{rs}^{\langle t \rangle}}{\partial x_{ij}^{\langle t \rangle}} = \sum_{r=1}^{n_c} \sum_{s=1}^{m} g_1'(z_{ke}^{\langle t \rangle})\, \delta_{ki}\, \delta_{ej}$$

$$\cdot W_{sx_{ki}}\, \delta_{sj} = g_1'(z_{ke}^{\langle t \rangle}) \cdot W_{ki}\, \delta_{ej} \quad ; \quad \frac{\partial s_{ke}^{\langle t \rangle}}{\partial s_{ij}^{\langle t-1 \rangle}} = g_1'(z_{ke}^{\langle t \rangle}) \cdot W_{ss_{ki}}\, \delta_{ej}$$

$W_{sx_{ki}}$

$$\frac{\partial s_{ke}^{\langle t \rangle}}{\partial W_{ss_{ij}}} = \sum_{r=1}^{n_a} \sum_{s=1}^{m} \frac{\partial s_{ke}^{\langle t \rangle}}{\partial z_{rs}^{\langle t \rangle}} \cdot \frac{\partial z_{rs}^{\langle t \rangle}}{\partial W_{ss_{ij}}} = \sum_{r=1}^{n_c} \sum_{s=1}^{m} g_1'(z_{ke}^{\langle t \rangle})\, \delta_{kr}\, \delta_{es}\, \delta_{ri}\, s_{js}^{\langle t-1 \rangle}$$

$$= g_1'(z_{ke}^{\langle t \rangle})\, \delta_{je}\, \delta_{ki} \rightarrow s_{je}^{\langle t-1 \rangle}$$

$$\frac{\partial s_{rt}^{\langle t\rangle}}{\partial W_{x1,ij}} = g_1'(z_{rt}^{\langle t\rangle}) \cdot x_{jt}^{\langle t\rangle} \delta_{ki}$$

$$\frac{\partial \hat{y}_{rt}^{\langle t\rangle}}{\partial s_{ij}^{\langle t\rangle}} = \sum_{r=1}^{n_y} \sum_{s=1}^{m} \frac{\partial \hat{y}_{rt}^{\langle t\rangle}}{\partial p_{rt}^{\langle t\rangle}} \cdot \frac{\partial p_{rs}^{\langle t\rangle}}{\partial s_{ij}^{\langle t\rangle}} = \sum_{r=1}^{n_y} \sum_{s=1}^{m} g_2'(p_{rt}^{\langle t\rangle}) \delta_{ki} \delta_{js}$$

$$W_{ys_{ri}} \delta_{sij} = g_2'(p_{rt}^{\langle t\rangle}) \cdot W_{ysrki} \delta_{ej}$$

$$\frac{\partial \hat{y}_{rt}^{\langle t\rangle}}{\partial W_{yx,ij}} = \sum_{r=1}^{n_y} \sum_{s=1}^{m} \frac{\partial \hat{y}_{rt}^{\langle t\rangle}}{\partial p_{rs}^{\langle t\rangle}} \frac{\partial p_{rs}^{\langle t\rangle}}{\partial W_{ys,ij}} = \sum_{r=1}^{n_y} \sum_{s=1}^{m} g_2'(p_{rt}^{\langle t\rangle}) \delta_{ki} \delta_{ej}$$

$$\cdot s_{js}^{\langle t\rangle} \delta_{ri} = g_2'(p_{rt}^{\langle t\rangle}) \cdot s_{jt}^{\langle t\rangle} \delta_{ki}$$

$$\frac{\partial \hat{y}_{rt}^{\langle t\rangle}}{\partial W_{xs,ij}} = \sum_{r=1}^{n_s} \sum_{s=1}^{m} \frac{\partial \hat{y}_{rt}^{\langle t\rangle}}{\partial s_{rs}^{\langle t\rangle}} \frac{\partial s_{rs}^{\langle t\rangle}}{\partial W_{xs,ij}} = \sum_{r=1}^{n_s} \sum_{s=1}^{m} g_2'(p_{rt}^{\langle t\rangle}) \cdot W_{ysrk} \delta_{ej}$$

$$\cdot g_1'(z_{rs}^{\langle t\rangle}) \cdot x_{js}^{\langle t\rangle} \delta_{ri} = g_2'(p_{rt}^{\langle t\rangle}) \cdot W_{ysrki} g'(z_{ie}^{\langle t\rangle}) x_{jt}^{\langle t\rangle}$$

$$\frac{\partial \hat{y}_{rt}^{\langle t\rangle}}{\partial s_{ij}^{\langle t-1\rangle}} = \sum_{r=1}^{n_s} \sum_{s=1}^{m} \frac{\partial p_{rt}^{\langle t\rangle}}{\partial s_{ij}^{\langle t\rangle}} \cdot \frac{\partial s_{rs}^{\langle t\rangle}}{\partial s_{ij}^{\langle t-1\rangle}} = \sum_{r=1}^{n_a} \sum_{s=1}^{m} g_2'(p_{rt}^{\langle t\rangle}) \cdot W_{ysrk} \delta_{ls}$$

$$\cdot g_1'(z_{rs}^{\langle t\rangle}) \cdot W_{ssrki} \cdot \delta_{sj} =$$

$$= \sum_{r=1}^{n_s} g_2'(p_{rt}^{\langle t\rangle}) \cdot W_{ysrk} \cdot g_1'(z_{rj}^{\langle t\rangle}) W_{ssrki} \delta_{ej}$$

$$\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S_{ij}^{\langle t-1\rangle}} = \sum_{k=1}^{n_s}\sum_{c=1}^{m} \frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S_{kc}^{\langle t\rangle}} \cdot \frac{\partial S_{kc}^{\langle t\rangle}}{\partial S_{ij}^{\langle t-1\rangle}} = \sum_{k=1}^{n_s}\sum_{c=1}^{m} \frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S_{kc}^{\langle t\rangle}} \cdot g_1'(z_{kc}^{\langle t\rangle}) \cdot W_{ss_{ki}} s_{cj}$$

$$= \sum_{k=1}^{n_s} \frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S_{kj}^{\langle t\rangle}} \cdot g_1'(z_{kj}^{\langle t\rangle}) \cdot W_{sski} \longrightarrow$$

Vectorized implementation $\longrightarrow$

$$\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S_{ij}^{\langle t-1\rangle}} = W_{ss}^{T} \cdot \left(\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S^{\langle t\rangle}} * g_1'(z^{\langle t\rangle}) \right)$$

$$\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial X_{ij}^{\langle t\rangle}} = W_{sx}^{T} \left(\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S^{\langle t\rangle}} * g_1'(z^{\langle t\rangle}) \right)$$

$$\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial W_{ss_{ij}}} = \sum_{k=1}^{n_s}\sum_{c=1}^{m} \frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S_{kc}^{\langle t\rangle}} \cdot \frac{\partial S_{kc}^{\langle t\rangle}}{\partial W_{ss_{ij}}} = \sum_{k=1}^{n_s}\sum_{c=1}^{m} \frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S_{kc}^{\langle t\rangle}} \cdot g_1'(z_{kc}^{\langle t\rangle}) \cdot s_{jc}^{\langle t-1\rangle} \cdot s_{ki}$$

$$= \sum_{c=1}^{m} \frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S_{ic}^{\langle t\rangle}} \cdot g_1'(z_{ic}^{\langle t\rangle}) \cdot s_{jc}^{\langle t-1\rangle}$$

Vectorized Implementation $\longrightarrow$

$$\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial W_{ss}} = \left(\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S^{\langle t\rangle}} * g_1'(z^{\langle t\rangle}) \right) \cdot S^{T\langle t-1\rangle}$$

$$\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial W_{sx}} = \left(\frac{\partial E^{\langle t+1:T_y\rangle}}{\partial S^{\langle t\rangle}} * g_1'(z^{\langle t\rangle}) \right) \cdot X^{T\langle t\rangle}$$

$$\frac{\partial E^{<t>}}{\partial b_{ij}^{<t>}} = \sum_{k=1}^{n_y} \sum_{c=1}^{m} \frac{\partial E^{<t>}}{\partial \hat{y}_{kc}^{<t>}} \cdot \frac{\partial \hat{y}_{kc}^{<t>}}{\partial b_{ij}^{<t>}} = \sum_{k=1}^{n_y} \sum_{c=1}^{m} \frac{\partial E^{<t>}}{\partial \hat{y}_{kc}^{<t>}} \cdot g_2'(p_{kc}^{<t>}) \cdot W_{ys_{ki}} \cdot \delta_{cj}$$

$$= \sum_{k=1}^{n_y} \frac{\partial E^{<t>}}{\partial \hat{y}_{kj}^{<t>}} \cdot g_2'(p_{kj}^{<t>}) \cdot W_{ys_{ki}}$$

$$\frac{\partial E^{<t>}}{\partial W_{ys_{ij}}} = \sum_{k=1}^{n_y} \sum_{c=1}^{m} \frac{\partial E^{<t>}}{\partial \hat{y}_{kc}^{<t>}} \cdot \frac{\partial \hat{y}_{kc}^{<t>}}{\partial W_{ys_{ij}}} = \sum_{k=1}^{n_y} \sum_{c=1}^{m} \frac{\partial E^{<t>}}{\partial \hat{y}_{kc}^{<t>}} \cdot g_2'(p_{kc}^{<t+1>}) \cdot s_{jc}^{<t>} \delta_{ki}$$

$$= \sum_{c=1}^{m} \frac{\partial E^{<t>}}{\partial \hat{y}_{ic}^{<t>}} \cdot g_2'(p_{ic}^{<t>}) \cdot s_{jc}^{<t>}$$

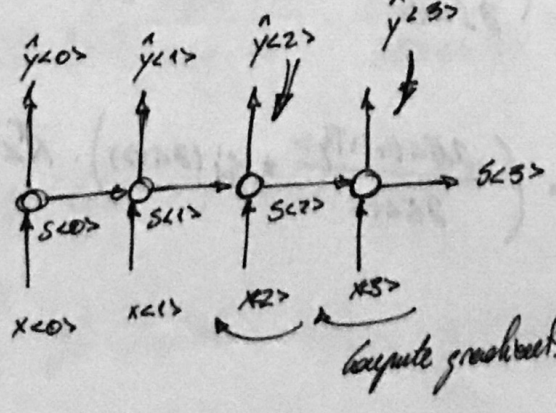
Vectorized Implementation $\longrightarrow$

$$\frac{\partial E^{<t>}}{\partial s^{<t>}} = W_{ys}^{T} \left(\frac{\partial E^{<t>}}{\partial \hat{y}^{<t>}} * g_2'(p^{<t>}) \right)$$

$$\frac{\partial E^{<t>}}{\partial W_{ys}} = \left(\frac{\partial E^{<y>}}{\partial \hat{y}^{<t>}} * g_2'(p^{<t>}) \right) \cdot s^{T<t>}$$

$$\frac{\partial E}{\partial s^{<t+1>}} = \left[W_{yc}^{T} \left(\frac{\partial E^{<t>}}{\partial \hat{y}^{<t>}} * g_2'(p^{<t>}) \right) * f_1'(z^{<t>}) \right] + W_{ss}^{T} \left(\frac{\partial E^{<t+1:T_x>}}{\partial s^{<t>}} * g_1'z^{<t>} \right)$$

W_{ss}^{T}



Implementation.

Compute gradients, Iterate across time & accumulate grad.