BPN

CNNs

$$\mathcal{L}(\theta) = \| y - f_K(\theta, x) \| \qquad \theta = w_1, w_2, b_1, b_2 \ldots$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{K-1}} = \frac{\partial L}{\partial f_K} \cdot \frac{\partial f_K}{\partial \theta_{K-1}}$$

(chain rule)

θ_{K-2} $\rightarrow$

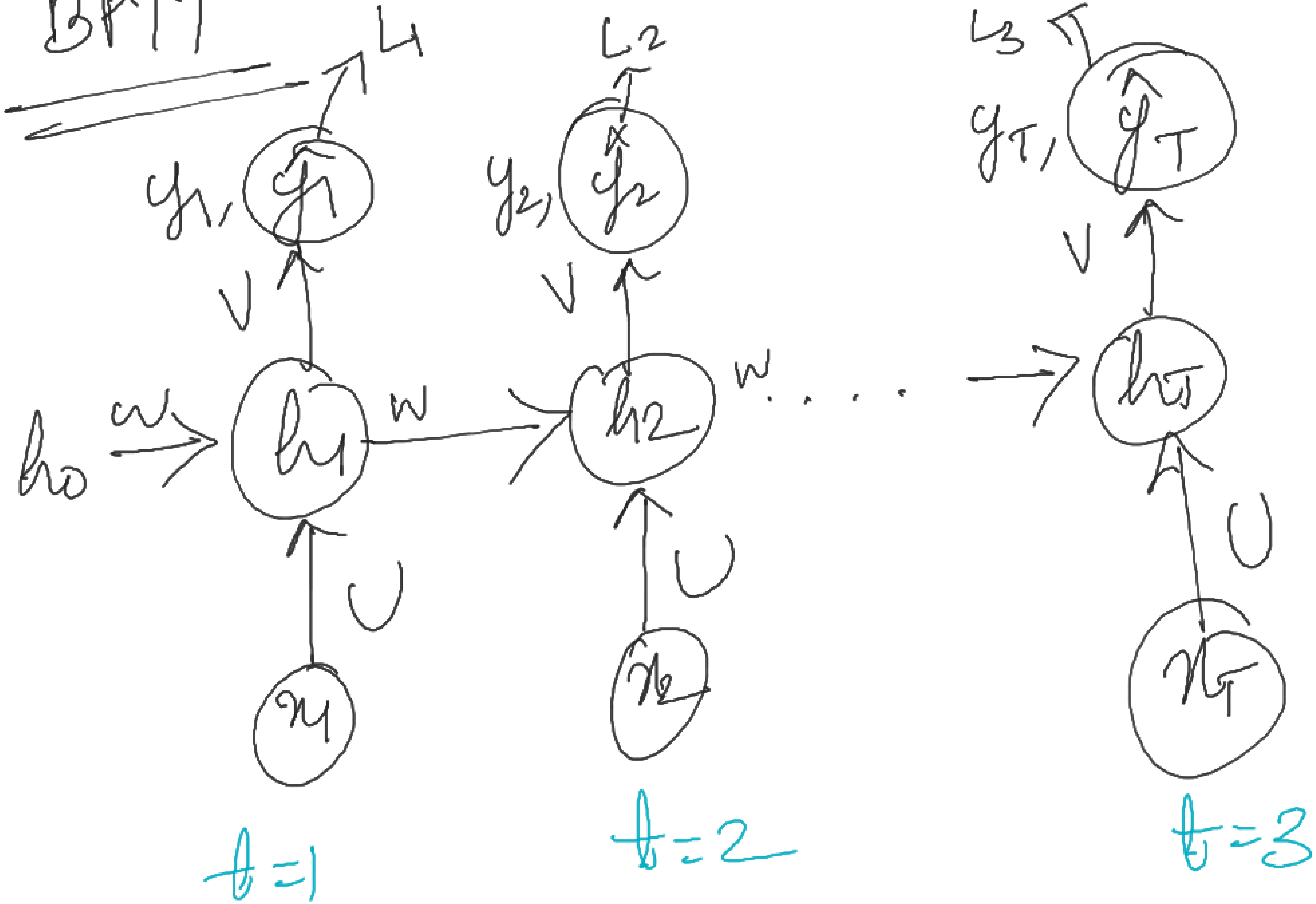
K-th

K-th

$$\frac{\partial L}{\partial \theta_{K-2}} = \frac{\partial L}{\partial f_K} \cdot \frac{\partial f_K}{\partial f_{K-1}} \cdot \frac{\partial f_{K-1}}{\partial \theta_{K-2}} \cdots$$

⋮

BPTT



$$L = \sum_{t=1}^{T} L_t$$

T = # of timesteps

BPTT

$$L_t = -y_t \log \hat{y}_t$$

actual output

predicted output

<u>Cross-entropy loss</u>

Forward Prop. Eq^{ns}:-

$$h_t = \tanh(U \cdot x_t + W \cdot h_{t-1})$$

$$\hat{y}_t = \text{softmax}(V \cdot h_t)$$

$$L_t = -y_t \log \hat{y}_t$$

$$L = \sum_{t=1}^{T} L_t$$

Let $Z_t = V \cdot h_t$

$\Rightarrow \hat{y}_t = \text{softmax}(Z_t)$

$$W = W - \alpha \cdot \frac{\partial L}{\partial W}$$

$$V = V - \alpha \cdot \frac{\partial L}{\partial V}$$

$$U = U - \alpha \cdot \frac{\partial L}{\partial U}$$

α = learning rate

$$\frac{\partial L}{\partial V} = \frac{\partial L_1}{\partial V} + \frac{\partial L_2}{\partial V} + \cdots + \frac{\partial L_T}{\partial V}$$

$$= \sum_{i=1}^{T} \frac{\partial L_i}{\partial V} = \sum_{i=1}^{T} \frac{\partial L_i}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial V}$$

$$\frac{\partial L_i}{\partial \hat{y}_i} = \frac{-y_i}{\hat{y}_i}$$

PTO

$$\underline{\underline{\hat{y}_i}} = \text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{K} e^{z_c}}$$

$K \rightarrow$ # of classes

MNIST = 10

Binary = 2

k - predicted label

Case 1 :- $i = K$

$$\frac{\partial \hat{y}_i}{\partial z_K} = \underbrace{\frac{e^{z_i}}{\sum_{c=1}^{K} e^{z_K}}}_{\hat{y}_i} - \underbrace{e^{z_i} \cdot \frac{e^{z_i}}{\left(\sum_{c=1}^{K} e^{z_c}\right)^2}}_{(\hat{y}_i)^2} = \hat{y}_i(1 - \hat{y}_i)$$

Case 2:- $i \neq K$

$$\frac{\partial \hat{y}_i}{\partial K} = -\frac{e^{z_i} e^{z_K}}{\left(\sum_{c=1}^{K} e^{s_c}\right)^2} = -\hat{y}_i \cdot \hat{y}_K$$

$$\Rightarrow \frac{\partial \hat{y}_i}{\partial z_K} = \begin{cases} \hat{y}_i(1-\hat{y}_i), & i = K \\ -\hat{y}_i \cdot \hat{y}_K, & i \neq K \end{cases}$$

$$\frac{\partial L_i}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_i} = \frac{-y_i}{\hat{y}_i} \begin{cases} \hat{y}_i(1-\hat{y}_i), & i=k \\ -\hat{y}_i \hat{y}_k, & i \neq k \end{cases}$$

$$= \begin{cases} -y_i(1-\hat{y}_i), & i=k \\ y_i \hat{y}_k, & i \neq k \end{cases}$$

$$= \underbrace{-y_i + y_i \hat{y}_i}_{i=k} + \sum_{i \neq k} y_i \hat{y}_k$$

$$= -y_k + y_k \hat{y}_k + \sum_{i \neq k} y_i \hat{y}_k$$

$$= -y_k + \hat{y}_k \left[y_k + \sum_{i \neq k} y_i \right]$$

$$= -y_k + \hat{y}_k \left[\sum_{i=1}^{K} y_i \right] \longrightarrow \text{one-hot labeled}$$

$$\begin{array}{cl} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & \begin{array}{l} \\ \\ \\ \longrightarrow \text{true class} \\ \\ \\ \underline{\Sigma = 1} \\ \\ \end{array} \end{array}$$

$$= -y_k + \hat{y}_k$$

$$\frac{\partial L}{\partial V} = \sum_{i=1}^{T} (\hat{y_i} - y_i) \cdot \frac{\partial z_i}{\partial V}$$

$$Z = V \cdot h_t$$

$$\Rightarrow \frac{\partial Z}{\partial V} = h_t$$

$$\underline{\underline{\frac{\partial L}{\partial V}}} = \underbrace{\sum_{i=1}^{T} (\hat{y_i} - y_i)}_{1^{st}\ \text{vector}} \otimes \underbrace{h_i}_{2^{nd}\ \text{vector}}$$

outer product

Eq 1

$\partial L / \partial W$

$$\frac{\partial L}{\partial W} = \frac{\partial L_1}{\partial W} + \frac{\partial L_2}{\partial W} + \cdots + \frac{\partial L_T}{\partial W}$$

$$= \sum_{i=1}^{T} \frac{\partial L_i}{\partial W}$$

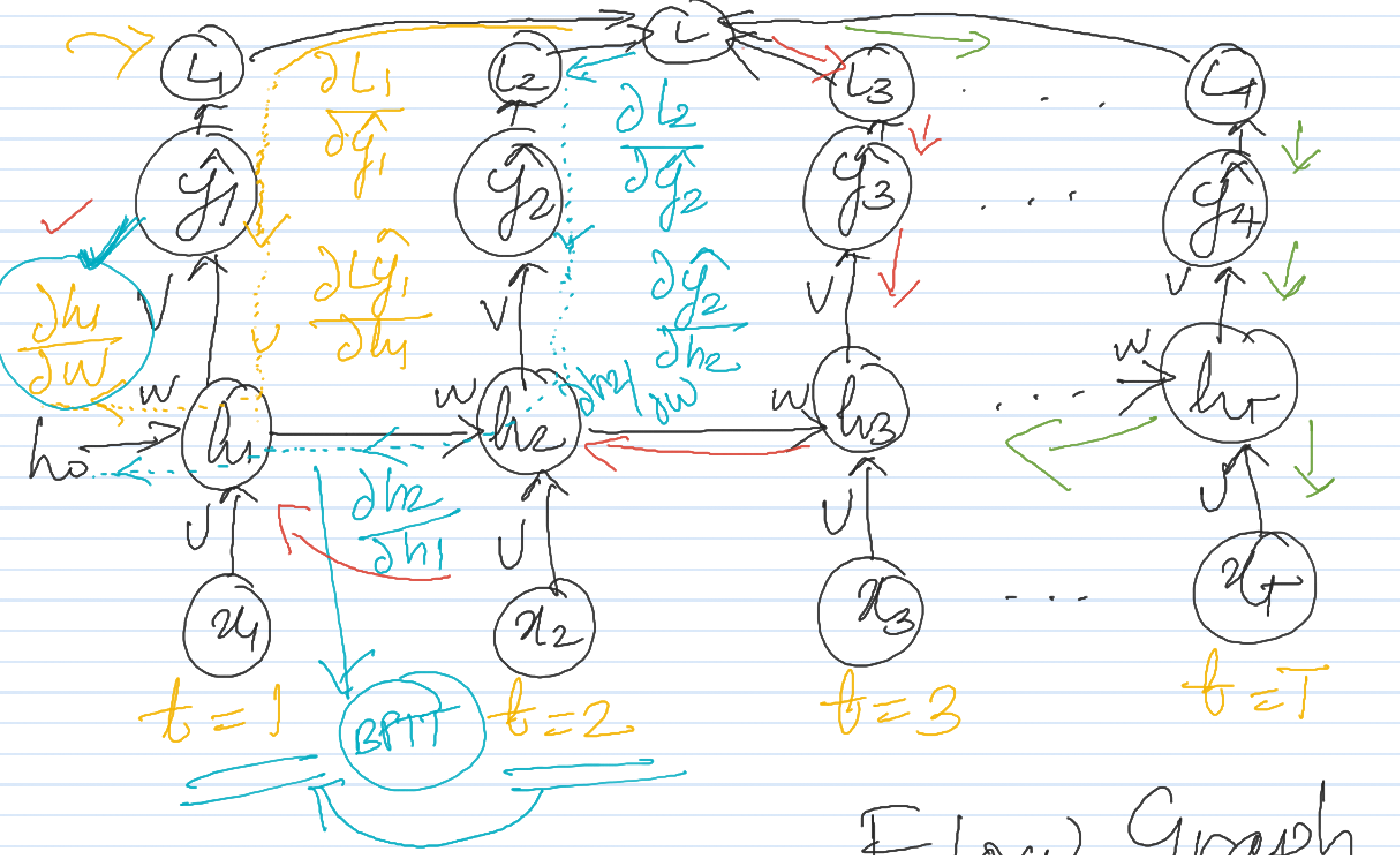
$$h_1 = \tanh(U \cdot x_1 + W \cdot h_0)$$

$$\frac{\partial L_1}{\partial W} = \underset{\Uparrow \atop -y_1/\hat{y}_1}{\frac{\partial L_1}{\partial \hat{y}_1}} \cdot \underset{\Uparrow \atop \hat{y}_1 - y_1}{\frac{\partial \hat{y}_1}{\partial h_1}} \cdot \frac{\partial h_1}{\partial W} \leftarrow ? \quad \text{——— } 2$$

$$\frac{\partial h_1}{\partial W} = \frac{\partial \tanh(U \cdot x_1 + W \cdot h_0)}{\partial W}$$

$$= h_0\left(1 - \tanh^2(U \cdot x_1 + W \cdot h_0)\right)$$

$$\frac{\partial L_2}{\partial W} = \frac{2}{0}$$


L
L_1
L_2
L_3
L_T
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_3$
$\hat{y}_T$
h_0
h_1
h_2
h_3
h_T
x_1
x_2
x_3
x_T
V
W
U
$\frac{\partial L_1}{\partial \hat{y}_1}$
$\frac{\partial \hat{y}_1}{\partial h_1}$
$\frac{\partial h_1}{\partial W}$
$\frac{\partial L_2}{\partial \hat{y}_2}$
$\frac{\partial \hat{y}_2}{\partial h_2}$
$\frac{\partial h_2}{\partial W}$
$\frac{\partial h_2}{\partial h_1}$
BPTT
$t=1$
$t=2$
$t=3$
$t=T$

Flow Graph

$$\frac{\partial L_2}{\partial W} = \frac{\partial L_2}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial W}$$

$$h_2 = \tanh(U \cdot x_2 + \underline{\underline{W \cdot h_1}})$$

$$h_1 = \tanh(U \cdot x_1 + \underline{\underline{W}} \cdot h_0)$$

h_1 is also dependent upon W.

$$\Rightarrow \frac{\partial h_2}{\partial W} = \underbrace{\frac{\partial h_2}{\partial W}}_{\text{explicit}} + \frac{\partial h_2}{\partial h_1} \cdot \underbrace{\frac{\partial h_1}{\partial W}}_{\text{implicit}}$$

$$\frac{\partial L_2}{\partial W} = \frac{\partial L_2}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial h_2} \left[\frac{\partial h_2}{\partial W} + \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W} \right]$$

$$= \frac{\partial L_2}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial W} + \frac{\partial L_2}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W}$$

$$\frac{\partial L_3}{\partial W} = \frac{\partial L_3}{\partial \hat{y}_3} \cdot \frac{\partial \hat{y}_3}{\partial h_3} \cdot \left[\frac{\partial h_3}{\partial W} + \frac{\partial h_3}{\partial h_2} \left[\frac{\partial h_2}{\partial W} + \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W} \right] \right]$$

$$= \frac{\partial L_3}{\partial \hat{y}_3} \cdot \frac{\partial \hat{y}_3}{\partial h_3} \cdot \frac{\partial h_3}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3} \cdot \frac{\partial \hat{y}_3}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial W}$$

$$+ \frac{\partial L_3}{\partial \hat{y}_3} \cdot \frac{\partial \hat{y}_3}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W},$$

Revisit forward prop:-

$$h\_unactivated = U \cdot x_t + W \cdot h_{t-1}$$

$$h\_activated = \tanh(h\_unactivated)$$

$$\hat{y}\_unactivated = V \cdot h\_activated$$

$$\hat{y}\_activated = \mathrm{softmax}(\hat{y}\_unactivated)$$

$$U\_frd = U \cdot x_t$$

$$W\_frd = W \cdot h_{t-1}$$

$$\frac{\partial L}{\partial V} = \frac{\partial L}{\partial y\_activated} \times \frac{\partial y\_activated}{\partial y\_unactivated} \times \frac{\partial y\_unactivated}{\partial V}$$

$$= (\hat{y}_t - y_t) \times h_t \longrightarrow A$$

$$\frac{\partial L}{\partial U} = \frac{\partial L}{\partial y\_act} \times \frac{\partial y\_act}{\partial y\_unact} \times \frac{\partial y\_unact}{\partial h\_act} \times \frac{\partial h\_act}{\partial h\_unact} \times \frac{\partial h\_unact}{\partial U_{frd}} \times \frac{\partial U_{frd}}{\partial U}$$

$$= (\hat{y}_t - y_t) \times V^T \times \left[1 - \tanh^2(h\_unact)\right] \times [1 \cdot\cdot 1] \times X^T$$

$$\longrightarrow C$$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y\_activated} \times \frac{\partial y\_activated}{\partial y\_unactivated} \times \frac{\partial y_{unac}}{\partial h_{activ}} \times \frac{\partial h_{activ}}{\partial h_{unactiv}}$$

$$\times \frac{\partial h_{unactiv}}{\partial W_{frd}} \times \frac{\partial W_{frd}}{\partial W}$$

$$= \underline{(\hat{y}_t - y_t)} \times V^T \times \left[1 - \tanh^2(h\_unactiv)\right] \times \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \times \underline{\underline{h_{t-1}}}$$

$\hookrightarrow$ B