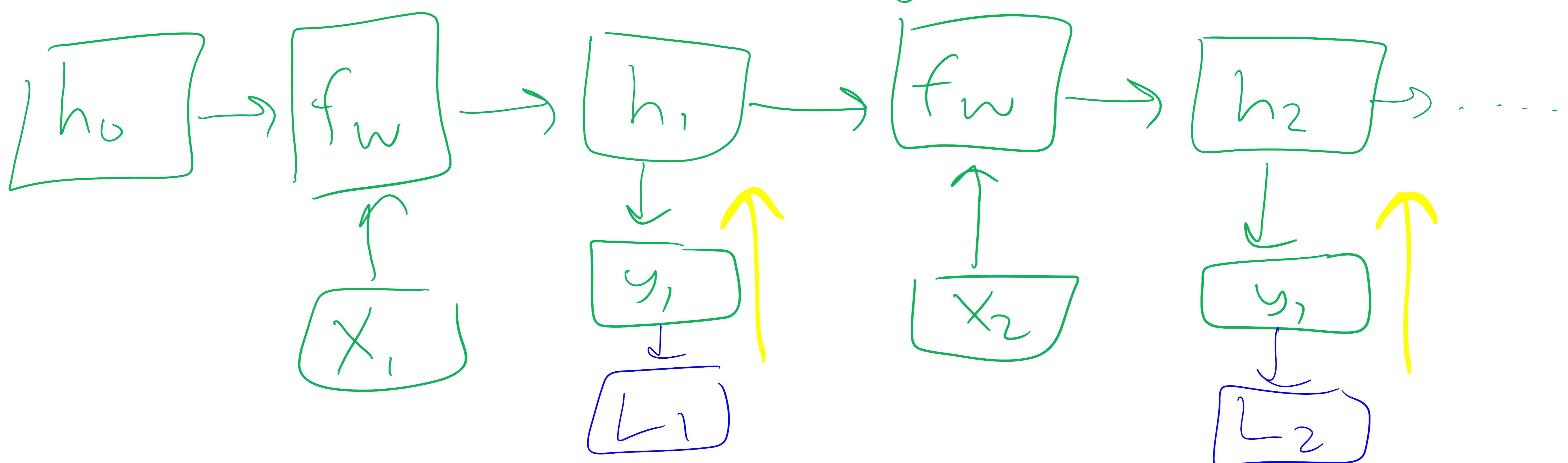


$$h_t = f_w(h_{t-1}, x_t)$$

↑
input

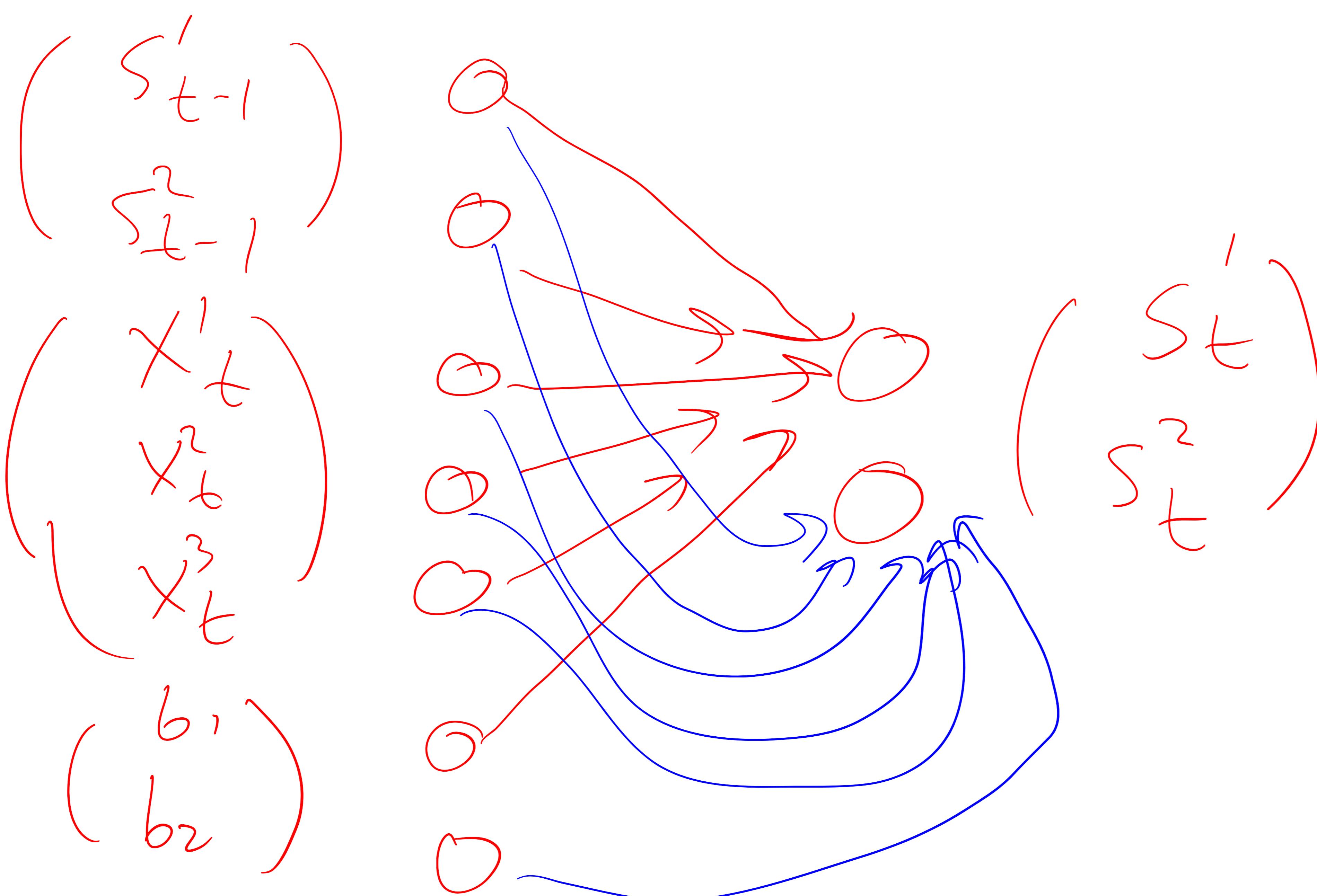
$$h_t = \tanh(W_h h_{t-1} + W_x x_t)$$

$$y_t = W_y h_t$$

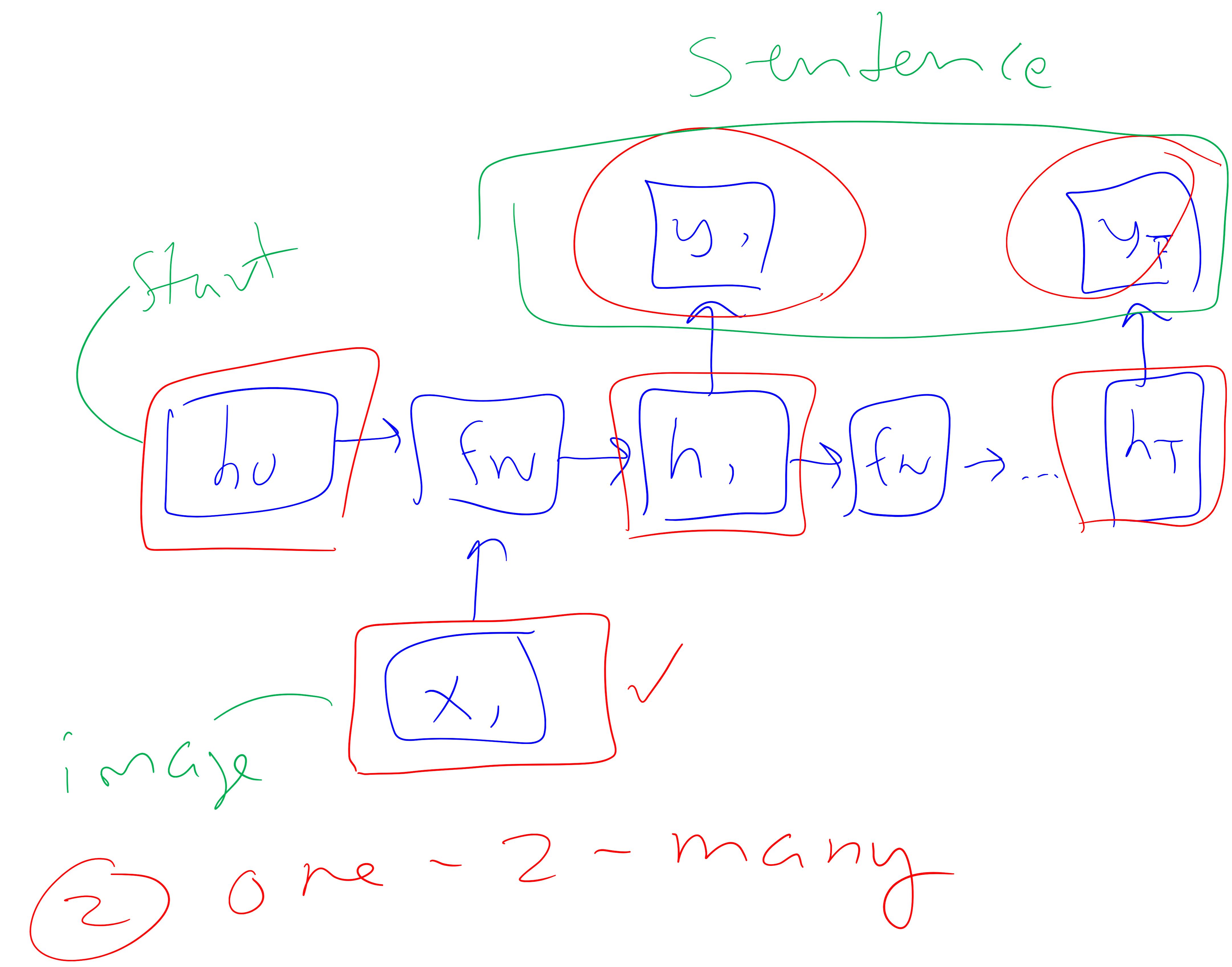
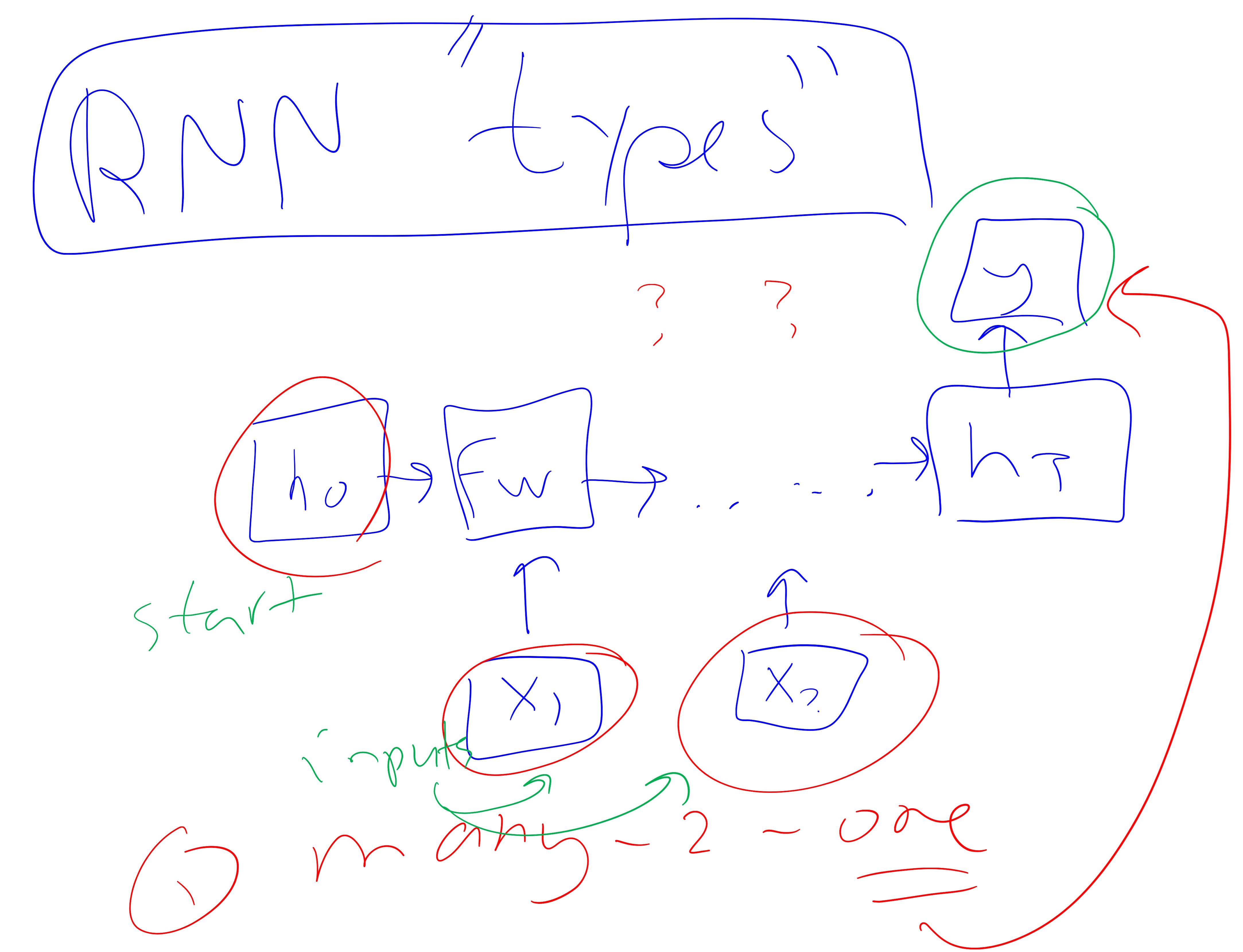


$$[S_t] = \mathcal{P} (w_{S_{t-1}} + Ux_t + b)$$

$$\begin{aligned}
 \begin{pmatrix} S_t^1 \\ S_t^2 \end{pmatrix} &= \mathcal{P} \left\{ \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \begin{pmatrix} S_{t-1}^1 \\ S_{t-1}^2 \end{pmatrix} + \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \end{pmatrix} \begin{pmatrix} x_t^1 \\ x_t^2 \\ x_t^3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\} \\
 &= \mathcal{P} \left\{ \begin{pmatrix} w_{11} S_{t-1}^1 + w_{12} S_{t-1}^2 \\ w_{21} S_{t-1}^1 + w_{22} S_{t-1}^2 \end{pmatrix} + \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \end{pmatrix} \begin{pmatrix} x_t^1 \\ x_t^2 \\ x_t^3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\} \\
 &= \mathcal{P} \left\{ \begin{pmatrix} \underline{w_{11}} & \underline{w_{12}} \\ \underline{w_{21}} & \underline{w_{22}} \end{pmatrix} \begin{pmatrix} \underline{u_{11}} & \underline{u_{12}} & \underline{u_{13}} \\ \underline{u_{21}} & \underline{u_{22}} & \underline{u_{23}} \end{pmatrix} \begin{pmatrix} S_{t-1}^1 \\ S_{t-1}^2 \\ x_t^1 \\ x_t^2 \\ x_t^3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\}
 \end{aligned}$$



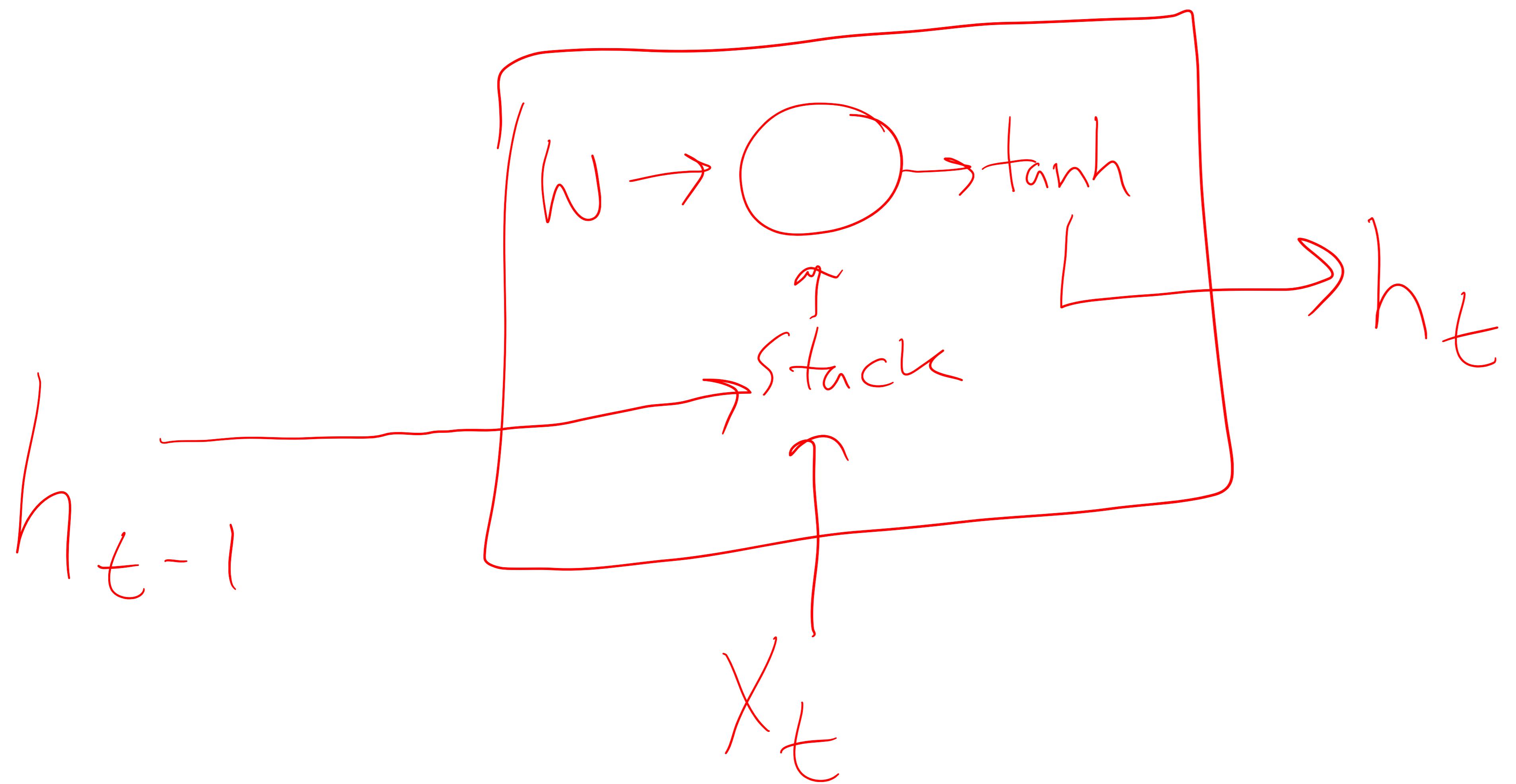
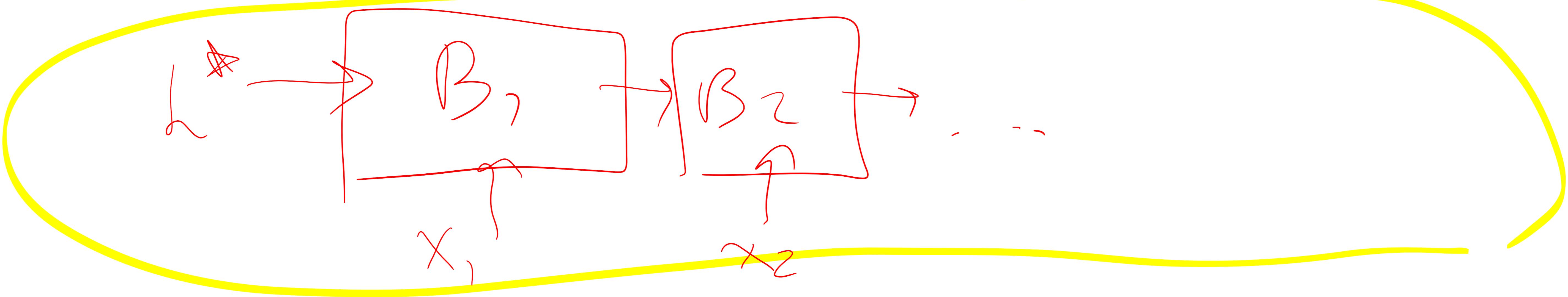
signal - 2 - text



(3) many -2- many

(4) one - 2 - one

Block



$$\begin{aligned} h_t &= \tanh(W_{hh} h_{t-1} + W_{xh} x_t) \\ &= \tanh((W_{hh} \quad W_{xh})(\begin{matrix} h_{t-1} \\ x_t \end{matrix})) \\ &= \tanh(W(\begin{matrix} h_{t-1} \\ x_t \end{matrix})) \end{aligned}$$

Issue) vanishing backprop

mul by 0's or mul $> 1, \dots$ vanishing
and exploding gradients

BPTT back prop through time

~~simple, shared weights and unrolled~~

$$h_t = \tanh(Ux_t + Wh_{t-1})$$
$$y_t = \text{softmax}(Vh_t)$$

softmax \Rightarrow $\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$

Diagram showing the forward pass of a RNN cell. The hidden state h_t is calculated as the tanh of the input Ux_t plus the previous hidden state Wh_{t-1} . The output y_t is the softmax of Vh_t . A red arrow labeled t points from the output y_t back to the hidden state h_t , indicating the unrolled nature of the computation.

Error fx

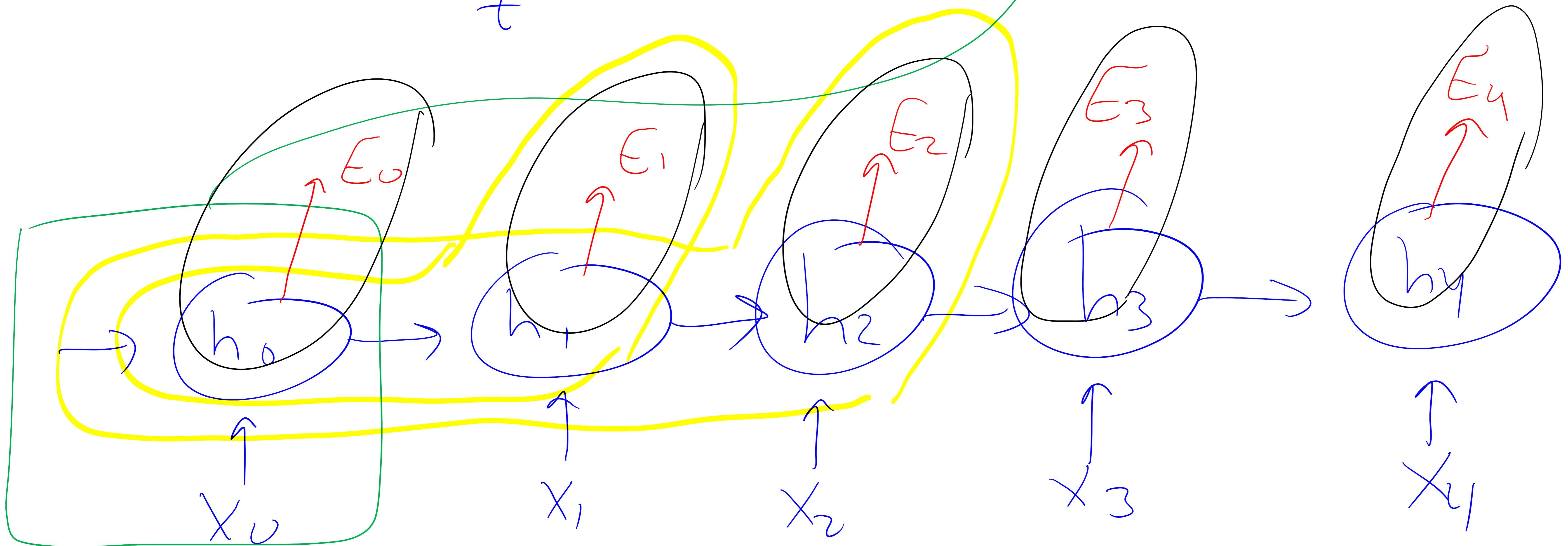
$$E_t(y_t, \hat{y}_t) = -y_t \ln \hat{y}_t$$

Cross entropy

Diagram showing the error function $E_t(y_t, \hat{y}_t) = -y_t \ln \hat{y}_t$. A red bracket groups the term $y_t \ln \hat{y}_t$. A blue arrow points from the softmax formula above down to this term, indicating they are equivalent. Below the error function, the text "Cross entropy" is written.

$$E(y, \hat{y}) = \sum_t E_t(y_t, \hat{y}_t)$$

$$= - \sum_t y_t \ln(\hat{y}_t)$$



Back prop

Question

$$\frac{\partial E}{\partial w} = \sum_t \frac{\partial E_t}{\partial w} \text{ longer...}$$

try

$$\frac{\partial E}{\partial v} = \sum_t \frac{\partial E_t}{\partial v} \checkmark$$

$$\frac{\partial E_3}{\partial V} = \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial V} \rightarrow z_3 = V h_3$$

just depends on v, current "current values" y_3 , h_3 , \hat{y}_3

$$= \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial z_3} \frac{\partial z_3}{\partial V}$$

post-thur

since $\frac{\partial z_3}{\partial V} = h_3$

$$= (G_3 - y_3) \otimes h_3$$

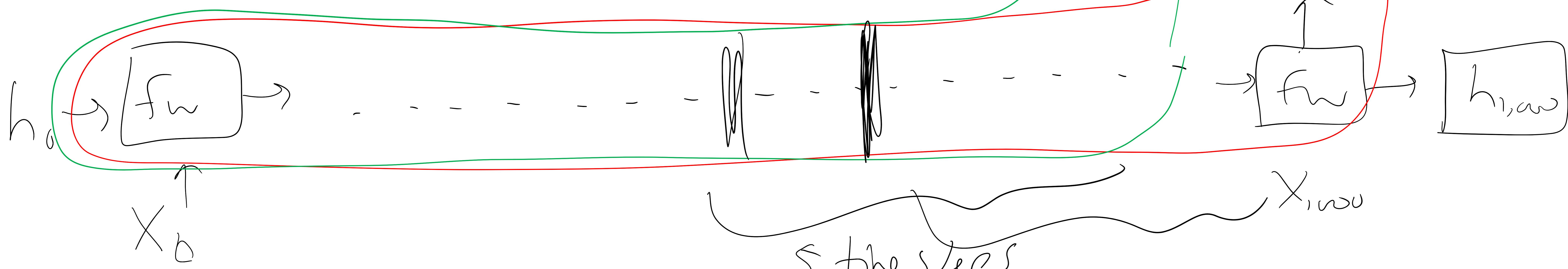
inter product

$$\frac{\partial E_3}{\partial W} = \frac{\partial E_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial h_3} \frac{\partial h_3}{\partial W}$$

note: $h_3 = \tanh(Ux_3 + Wh_2)$

π
depends on h_2 , which depends on h ,

$$\frac{\partial E_3}{\partial w} = \sum_{k=0}^3 \left(\frac{\partial E_3}{\partial y_3} \frac{\partial y_3}{\partial h_3} \right) \left(\frac{\partial h_3}{\partial h_k} \frac{\partial h_k}{\partial w} \right)$$



TBTT \rightarrow truncated BTT
 (fixed horizon)

approx...

