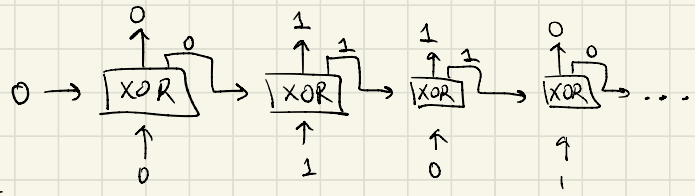
1. input $x = [0, 1, 0, 1, 1, 0]$
output $y = [0, 1, 1, 0, 1, 1]$

$h_0 = 0$

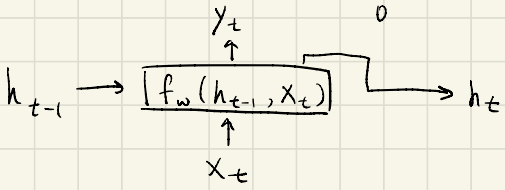
$h_t = h_{t-1}$ XOR x_t *can't use XOR, but XOR is the relationship

$y_t = h_t$

we want basically



$\Rightarrow$



let $f_w(h_{t-1}, x_t) = h_{t-1} \cdot \bar{x}_t + \bar{h}_{t-1} \cdot x_t$

We see that this acts as an XOR gate for binary inputs

h_{t-1}	x_t	$f_w(h_{t-1}, x_t)$
0	0	0
0	1	1
1	0	1
1	1	1

2. $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$

$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$

$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$

$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

$o^t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$

$h_t = o^t * \tanh(C_t)$

Assume $h_0 = 0 \quad C_0 = 0$

let $W_f = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad b_f = 0$

$W_i = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad b_i = 0$

$W_c = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad b_c = 0$

$W_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad b_o = 1$

h_{t-1}	x_t	f_t	i_t	$\tilde{C}_t$	o^t	C_t	h_t
0	0	0	0	0	1	0	0
0	1	0	1	1	1	1	1
1	0	1	0	0	1	1	1
1	1	0	0	1	1	0	0

← note $C_{t-1} = h_{t-1}$

can see that we get XOR from h_{t-1} and x_t

3. $\text{best}_{\leq t}$ is defined at t

highest scoring beam in B_t is y_t^1 with score s_t^1

$s_t^1 \leq \text{best}_{\leq t}$

probability $\in [0,1]$ due to $\sum_x p(y_{t+1} | x, y_{t+1}) = 1$

$\text{best}_{\leq t+1} = s_t^1 + \log p(y_x | x, y_{<t})$

$s_t^i \leq s_t^1 \leq \text{best}_{\leq t}$ for all other beams in B_t

$s_{t+1}^1 = s_t^1 + \log p(y_x | x, y_{<t})$ and since probability $\in [0,1]$

all beams in $t' > t$ will be no better than $\text{best}_{\leq t}$

4. $h_t = W^T h_{t-1}$

let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of weight matrix $W \in \mathbb{C}^{n \times n}$

$$\rho(W) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$$

Assume initial hidden state h_0

$$h_t = W^T h_{t-1}$$
$$h_{t-1} = W^T h_{t-2}$$
$$\vdots$$
$$h_1 = W^T h_0 \Rightarrow h_t = (W^T)^t h_0$$

take derivative $\frac{dh_t}{dh_0} = (W^T)^t$

decompose $W = PDP^{-1}$ where P is eigenvectors of W
D is diagonal matrix of W's eigenvalues

$$= ([PDP^{-1}]^T)^t$$
$$= ([P^{-1}]^T D^T P^T)^t \quad \text{note } D = D^T$$
$$= ([P^{-1}]^T D P^T)^t$$
$$= [P^{-1}]^T D^t P^T$$

we can see as $t >> 0$:

if $\rho(W) > 1$ then at least one value of D^t will exponentially grow toward $+\infty$ ← exploding gradient

if $\rho(W) < 1$ then all eigenvalues will grow towards 0 ← vanishing gradient

∴ eigenvalues determine whether exploding/vanishing gradient

5. a) $h_i^{t+1} = q(h_i^t, \text{Agg}(H_{i,t}'))$

$= q(h_i^t, \sum_{j \in N(v_i)} f_{j,i}(h_j^t))$

$= q(h_i^t, \sum_{j \in N(v_i)} (v^t h_j^t))$ where $v_j \in N(v_i)$

5.b) $h_1^t = [1, -1]$ $h_2^t = [-1, 1]$ $h_3^t = [0, -1]$ $h_4^t = [1, 0]$

$Agg(H_{1t}^l) = [0.6, 0.2, 0.2] \begin{bmatrix} f(h_2^t) \\ f(h_3^t) \\ f(h_4^t) \end{bmatrix}$

$f(x) = 2x$

$h_1^{t+1} = q(h_1^t, Agg(H_{1t}^l)) = W(h_1^t)^T + \max\{Agg(H_{1t}^l), 0\}$

where $W = [1, 1]$

$f(h_2^t) = [-2, 2]$

$f(h_3^t) = [0, -2]$

$f(h_4^t) = [2, 0]$

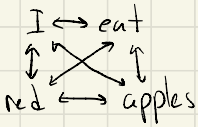
$Agg(H_{1t}^l) = [0.6, 0.2, 0.2] \begin{bmatrix} [-2, 2] \\ [0, -2] \\ [2, 0] \end{bmatrix} = [-0.8, 0.8]$

$h_1^{t+1} = W(h_1^t) + \max\{Agg(H_{1t}^l), 0\}$

$= [1, 1] \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \max\{[-0.8, 0.8], 0\}$

$= 0 + [0, 0.8] = [0, 0.8]$

5.c)



$2 \times 6 = \boxed{12}$

5. d) $h_i^{l+1} = \sum_{j \in S} \left(\text{softmax}_j \left(Q^l h_i^l \cdot K^l h_i^l \right) V^l h_j^l \right)$ (17)

$h_i^{l+1} = q\left(h_i^t, \sum_{j \in N(v_i)} \left(V^l h_j^l \right) \right)$ (15) from part a

$W_{ij} = \text{softmax}_j \left(Q^l h_i^l \cdot K^l h_j^l \right)$

$\Rightarrow$ (17) is $h_i^{l+1} = \sum_{j \in N(v_i)} W_{ij} \left(V^l h_j^l \right) = W(h_i^t)^T \sum_{j \in N(v_i)} \left(V^l h_j^l \right)$

5. e) For fully connected graphs, the number of connections increase exponentially, thus it becomes exponentially more expensive to learn longer term dependencies.