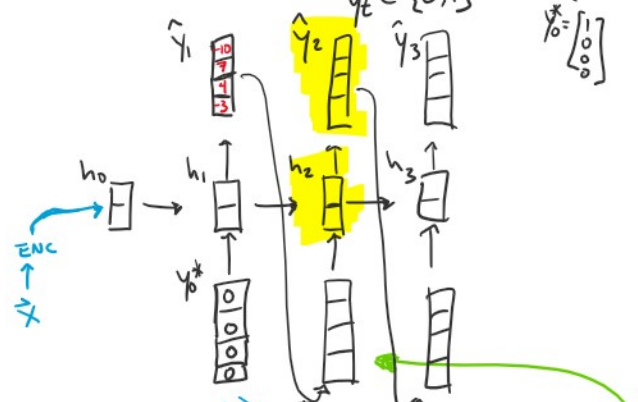
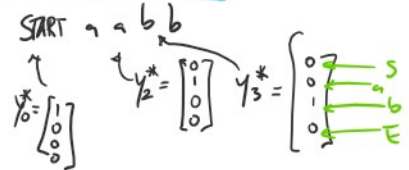


next section

**RNN-LMs**

$\vec{x} = [x_1, \dots, x_T]$      $x_t \in \{A, B\}$     **AA BB**

gold:  $\vec{y} = [y_1, \dots, y_T]$      $y_t \in \{a, b\}$



$P(y_1, \dots, y_T | \vec{x}) = \prod_{t=1}^T P(y_t | y_{t-1}, \dots, y_1, \vec{x})$

Q: What do we feed in here at train time?  $y_i^*$  (MLE)  
 " " " " " " test time?

$\text{onehot}(\arg\max_k \hat{y}_{ik}) = 1 \equiv "a"$   
 $= [0, 1, 0, 0]^T$

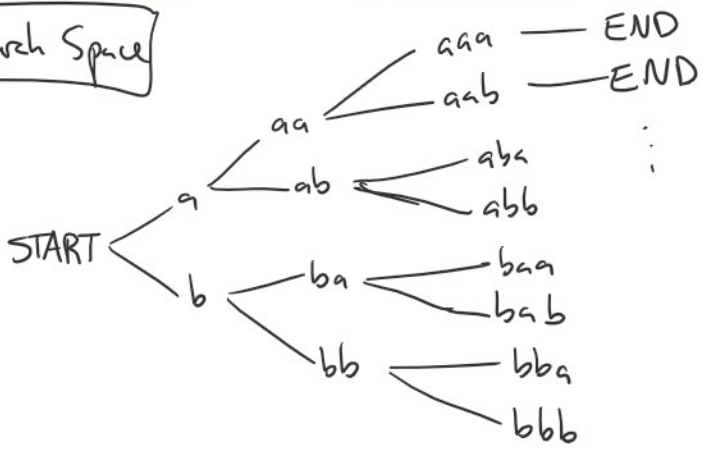
Q: What are the parameters? (Assume an Elman network)

- hid  $\rightarrow$  hid  $W^{(h \rightarrow h)} \in \mathbb{R}^{2 \times 2}$
- inp  $\rightarrow$  hid  $W^{(y \rightarrow h)} \in \mathbb{R}^{4 \times 2}$
- hid  $\rightarrow$  out  $W^{(h \rightarrow y)} \in \mathbb{R}^{2 \times 4}$

$y_2 = W^{(h \rightarrow y)} h_2$   
 $h_2 = \tanh(W^{(h \rightarrow h)} h_1 + W^{(y \rightarrow h)})$

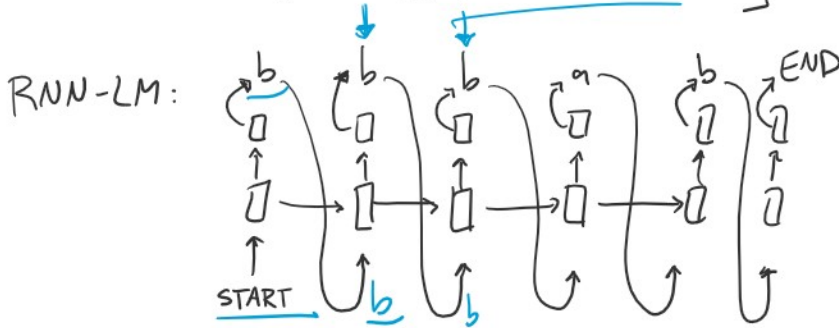
Q: How to compute the log-probability of "START a b END"  
 LEFT AS EXERCISE

**Search Space**

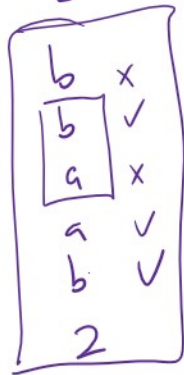
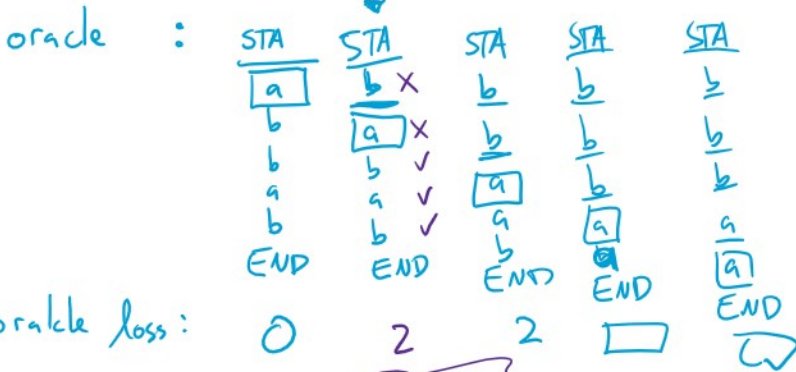


L2S Assume  $|y^*| = |\hat{y}| = T$  [DAGGER]

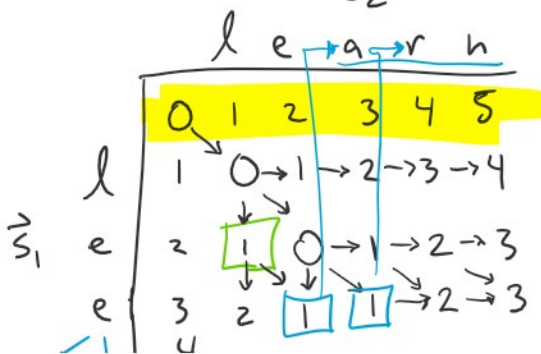
$$\text{loss}(y^*, \hat{y}) = \left[ \sum_{t=1}^T \mathbb{1}(y_t^* \neq \hat{y}_t) \right] + \left[ (2T) \mathbb{1}(\text{>3 b's in } \hat{y}) \right]$$



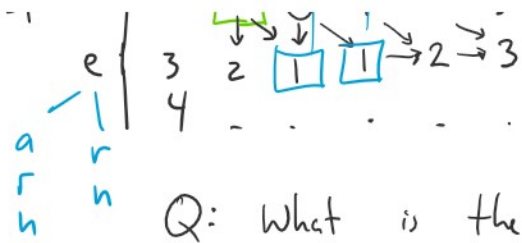
$y^*$  : a b b a b



Edit Distance



$$e[i][j] = \min \begin{pmatrix} e[i-1][j] + 1 \\ e[i][j-1] + 1 \\ e[i-1][j-1] + \mathbb{1}(s_1[i] \neq s_2[j]) \end{pmatrix}$$



Q: What is the optimal target (OCD) for

- ①  $y^* = \text{learn}$        $\hat{y} = \text{lee}$        $[a] r n$
- ②  $y^* = \text{learn}$        $\hat{y} = \text{lee}$       set of size two  
 $[a] r n$  or  
 $[r] n$