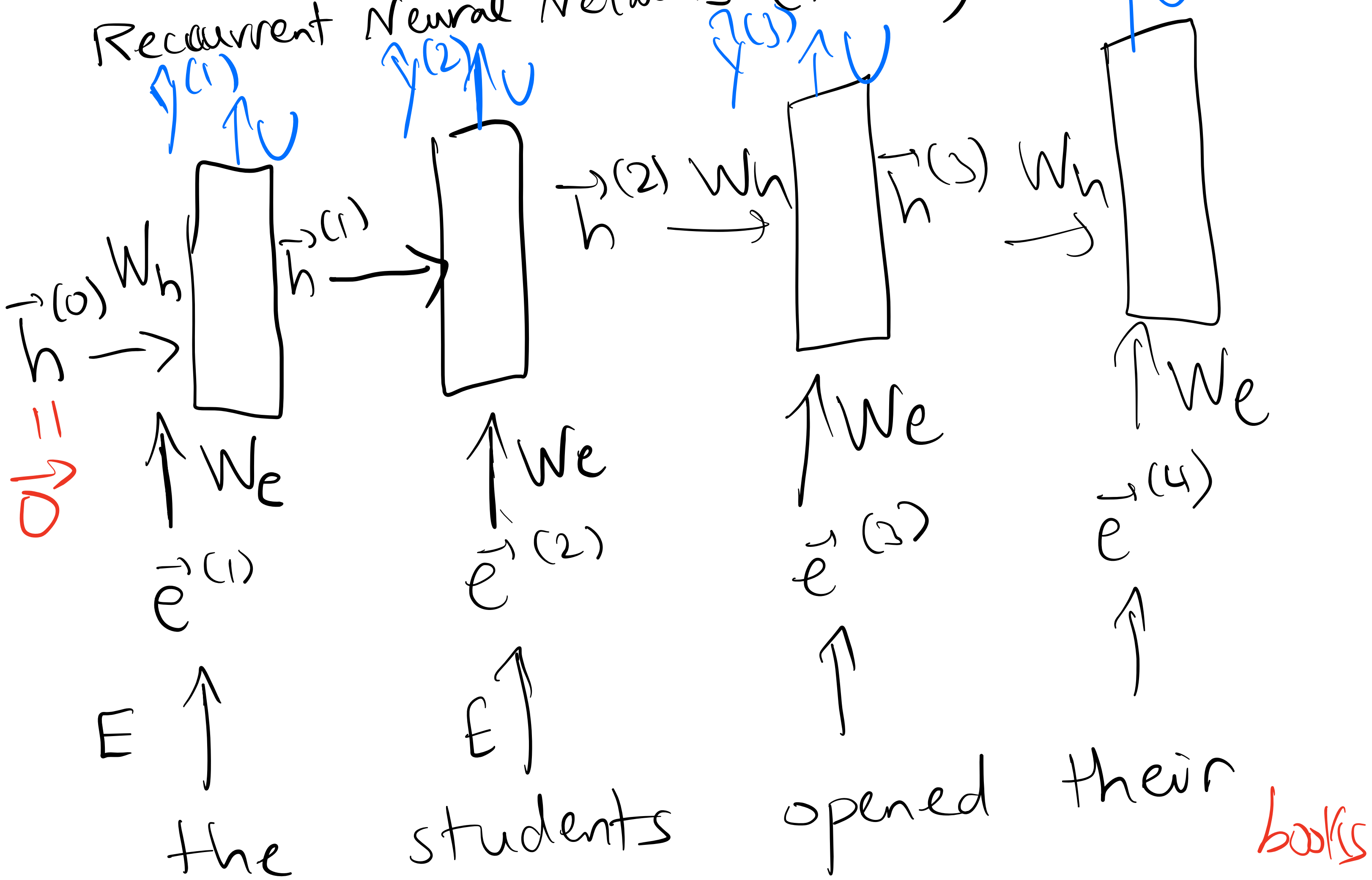
CS 6180 09/17

Recurrent Neural Networks (RNNs)



$$\vec{h}^{(t)} = \sigma\left(W_h h^{(t-1)} + W_e \vec{e}^{(t)} + \vec{b}_1\right)$$

$\hat{y}^{(1)}$ vs word students

We used cross entropy loss

$$\hat{y}^{(t)} = \mathrm{softmax}\left(P \vec{h}^{(t)} + \vec{b}_2\right)$$

$$\mathcal{L}^{(t)}(\theta) = -\sum_{w \in V} y_w^{(t)} \log \hat{y}_w$$

$\vec{y}$: one-hot of $\vec{x}^{(t+1)}$

Parameters :

$\boxed{W_h}$, W_e, $\vec{b}_1$, $\cancel{U}$, $\vec{b}_2$

$$\frac{\partial \mathcal{L}^{(t)}}{\partial W_h} = ?$$

$f(x(t), y(t))$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x} \cdot \frac{\partial x}{\partial t} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial t}$$

$$\mathcal{L}^{(t)}(\theta) = -\sum_{w \in V} y_w^{(t)} \log \hat{y}_w^{(t)}$$

$$\mathcal{L}^{(t)} = \mathcal{L}^{(t)}\left(\vec{\hat{y}}^{(t)}\right)$$

$$= \mathcal{L}^{(t)}\left(\vec{\hat{y}}^{(t)}\left(\vec{h}^{(t)}\right)\right)$$

$$= \mathcal{L}^{(t)}\left(\vec{\hat{y}}^{(t)}\left(\vec{h}^{(t)}\left(W_h, \vec{h}^{(t-1)}(W_h)\right)\right)\right)$$

$$= \mathcal{L}^{(t)}\left(W_h, \vec{h}^{(t-1)}(W_h)\right)$$

used when computing $\vec{h}^{(t)}$

$$= \mathcal{L}^{(t)}\left(W_h\big|_{(t)}, \vec{h}^{(t-1)}(W_h)\right)$$

$$\frac{\delta f}{\delta t} = \frac{\delta f}{\delta x}\frac{\delta x}{\delta t} + \frac{\delta f}{\delta y}\frac{\delta y}{\delta t}$$

$$\frac{\partial \mathcal{L}^{(t)}}{\partial W_h} = \frac{\partial \mathcal{L}^{(t)}}{\partial W_h\big|_t} \cdot \frac{\partial W_h\big|_t}{\partial W_h} = 1$$

$$+ \frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \cdot \frac{\partial \vec{h}^{(t-1)}\left(W_h\big|_{t-1}, \vec{h}^{(t-2)}(W_h)\right)}{\partial W_h}$$

$$= \frac{\partial \mathcal{L}^{(t)}}{\partial W_h\big|_t} + \frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \frac{\partial \vec{h}^{(t-1)}}{\partial W_h}\left(W_h\big|_{t-1}, \vec{h}^{(t-2)}(W_h)\right)$$

$$= \frac{\partial \mathcal{L}^{(t)}}{\partial W_h\big|_t} + \frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}}\left[\frac{\partial \vec{h}^{(t-1)}}{\partial W_h\big|_{t-1}} \frac{\partial W_h\big|_{t-1}}{\partial W_h} + \frac{\partial \vec{h}^{(t-1)}}{\partial \vec{h}^{(t-2)}} \frac{\partial \vec{h}^{(t-2)}(\cdot,\cdot)}{\partial W_h}\right]$$

$$= \frac{\partial \mathcal{L}^{(t)}}{\partial W_h}\bigg|_t + \frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \frac{\partial \vec{h}^{(t-1)}}{\partial W_h}\bigg|_{t-1} \underbrace{\frac{\partial W_h|_{t-1}}{\partial W_h}}_{=1} + \frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \cdot \frac{\partial \vec{h}^{(t-1)}}{\partial \vec{h}^{(t-2)}} \frac{\partial \vec{h}^{(t-2)}}{\partial W_h}$$

$$= \frac{\partial \mathcal{L}^{(t)}}{\partial W_h}\bigg|_t + \underbrace{\frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \cdot \frac{\partial \vec{h}^{(t-1)}}{\partial W_h}\bigg|_{t-1}}_{\text{combining (reverse chain rule)}} + \frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \frac{\partial \vec{h}^{(t-1)}}{\partial \vec{h}^{(t-2)}} \frac{\partial \vec{h}^{(t-2)}}{\partial W_h}$$

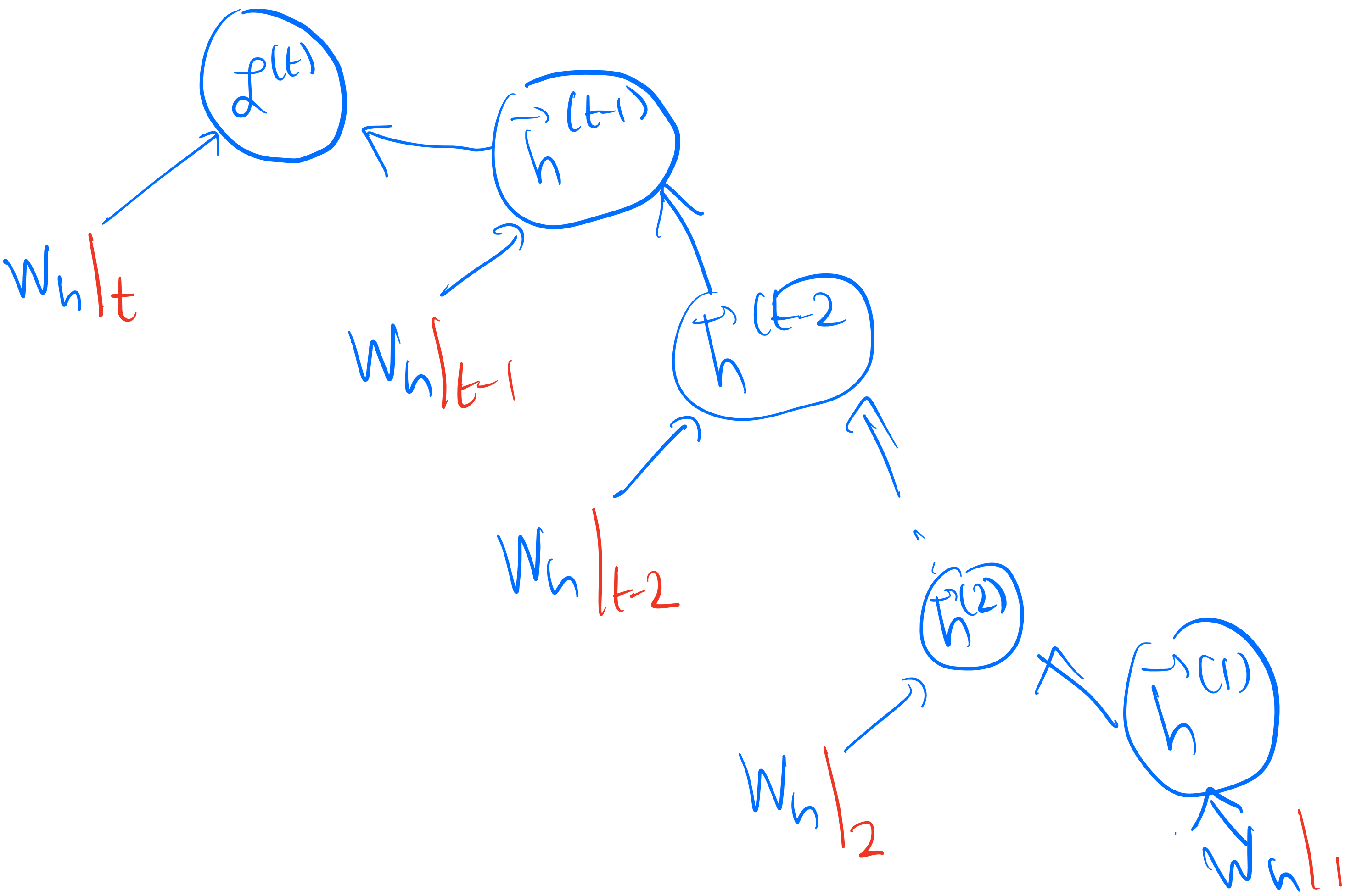
$$= \frac{\partial \mathcal{L}^{(t)}}{\partial W_h}\bigg|_t + \frac{\partial \mathcal{L}^{(t)}}{\partial W_h}\bigg|_{t-1} + \frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \cdot \frac{\partial \vec{h}^{(t-1)}}{\partial \vec{h}^{(t-2)}} \cdot \frac{\partial \vec{h}^{(t-2)}}{\partial W_h}$$

(repeat process)

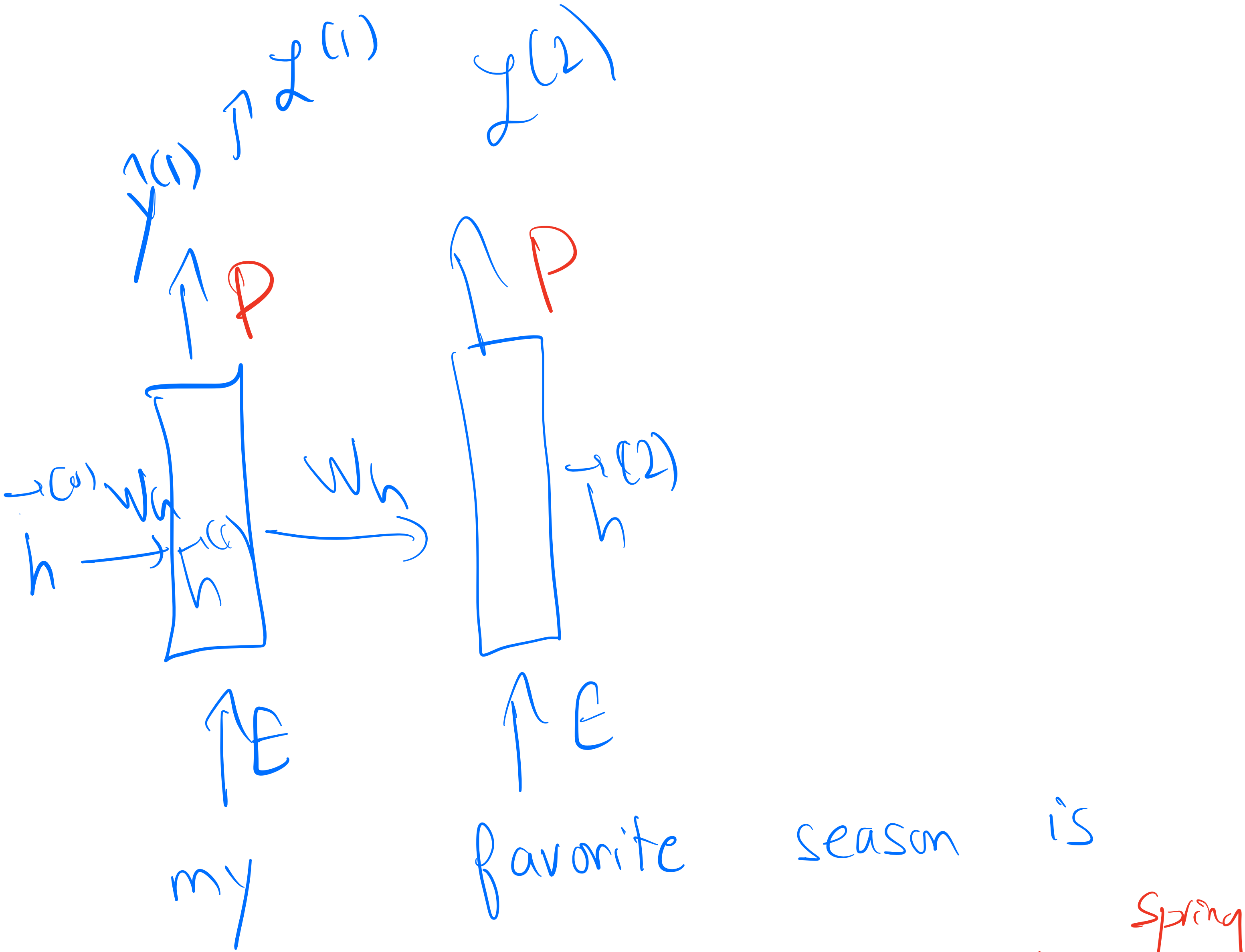
$$= \frac{\partial \mathcal{L}^{(t)}}{\partial W_h}\bigg|_t + \frac{\partial \mathcal{L}^{(t)}}{\partial W_h}\bigg|_{t-1} + \dots + \frac{\partial \mathcal{L}^{(t)}}{\partial W_h}\bigg|_2 + \frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \cdots \underbrace{\frac{\partial \vec{h}^{(1)}}{\partial W_h}}_{\substack{\text{depends on } W_h \text{ through only one channel} \\ \text{so same as } \frac{\partial \vec{h}^{(1)}}{\partial W_h}\big|_1}}$$

$$= \frac{\partial \mathcal{L}^{(t)}}{\partial W_h\big|_t} + \cdots + \frac{\partial \mathcal{L}^{(t)}}{\partial W_h\big|_2} + \underbrace{\frac{\partial \mathcal{L}^{(t)}}{\partial \vec{h}^{(t-1)}} \cdots \frac{\partial \vec{h}^{(1)}}{\partial W_h\big|_1}}_{\substack{\text{reverse chain rule} \\ \frac{\partial \mathcal{L}^{(t)}}{\partial W_h|_1}}}$$

$$= \sum_{i=1}^{t} \boxed{\frac{\partial \mathcal{L}^{(t)}}{\partial W_h\big|_i}}$$



how to compute $\dfrac{\partial \mathcal{L}^{(t)}}{\partial W_h|_i}$?



$$\frac{\partial \mathcal{L}^{(1)}}{\partial W_h} = \frac{\partial \mathcal{L}^{(1)}}{\partial W_h|_1}\left(\hat{y}^{(1)}\left(\vec{h}^{(1)}\left(W_h|_1\right)\right)\right)$$

$$= \frac{\partial \mathcal{L}^{(1)}}{\partial \hat{y}^{(1)}} \cdot \frac{\partial \hat{y}^{(1)}}{\partial \vec{h}^{(1)}} \frac{\partial \vec{h}^{(1)}}{\partial W_h \big|_1}$$

derivative of cross entropy loss

derivative of softmax

derivative of sigmoid

* $\mathcal{L}^{(2)}$:

$$\frac{\partial \mathcal{L}^{(2)}}{\partial W_h} = \underbrace{\frac{\partial \mathcal{L}^{(2)}}{\partial W_h \big|_2}} + \boxed{\frac{\partial \mathcal{L}^{(2)}}{\partial W_h \big|_1}}$$

exactly same as above but shifted to

the left.

$$\frac{\partial \mathcal{L}^{(2)}}{\partial W_h|_1} = \frac{\partial \mathcal{L}^{(2)}}{\partial \vec{h}^{(2)}} \cdot \frac{\partial \vec{h}^{(2)}}{\partial \vec{h}^{(1)}} \cdot \frac{\partial \vec{h}^{(1)}}{\partial W_h|_1}$$

$$= \frac{\partial \mathcal{L}^{(2)}}{\partial \hat{y}^{(2)}} \cdot \frac{\partial \hat{y}^{(2)}}{\partial \vec{h}^{(2)}} \; \frac{\partial \vec{h}^{(2)}}{\partial \vec{h}^{(1)}} \; \frac{\partial \vec{h}^{(1)}}{\partial W_h^{(1)}}$$

derivative of cross entropy

derivative of softmax

derivative of sigmoid

derivative of sigmoid

$$\mathcal{L} = -\log \hat{y}_{x_{t+1}}$$ (see part 1 of P1 of HW1)

$$\frac{\partial \mathcal{L}}{\partial \hat{y}} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial \hat{y}_1} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \hat{y}_{x_{t+1}}} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \hat{y}_{|V|}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -\frac{1}{\hat{y}_{x_{t+1}}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

vector
$|V| \times 1$

Now we can train our model, let's evaluate it (through testing)

* Perplexity

$$\text{perplexity of a language model} = \prod_{t=1}^{T} \left(\frac{1}{P\left(\vec{x}^{(t+1)} \mid \vec{x}^{(1)}, \ldots, \vec{x}^{(t)}\right)} \right)^{\frac{1}{T}}$$

$$= \prod_{t=1}^{T} \left(\frac{1}{\hat{y}_{x_{t+1}}} \right)^{1/T}$$

$$= \exp\left(\log \left(\prod_{t=1}^{T} \left(\frac{1}{\hat{y}_{x_{t+1}}} \right)^{1/T} \right) \right)$$

$$= \exp\left(\sum_{t=1}^{T} \log \left(\frac{1}{\hat{y}_{x_{t+1}}} \right)^{1/T} \right)$$

$$= \exp\left(\sum_{t=1}^{T} \frac{1}{T} \log \left(\frac{1}{\hat{y}_{x_{t+1}}} \right) \right)$$

$$= \exp\Big(\underbrace{-\frac{1}{T} \sum_{t=1}^{T} \log (\hat{y}_{x_{t+1}})}_{\mathcal{L}(\theta)} \Big)$$

$$= \underline{\exp}\left(\mathcal{L}(\theta) \right) \rightarrow \text{want small for a good model}$$

why exponential ?

- model predicts perfectly
  $\Rightarrow$ loss function $= 0$
  $\Rightarrow$ perplexity $= e^0 = 1$

- model thinks all words are equally likely.

$$-\frac{1}{T}\sum_{t=1}^{T} \log\left(\hat{y}_{x_{t+1}}\right)$$

$$= -\frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{1}{|V|}\right)$$

$$= -\frac{1}{T}\sum_{t=1}^{T} -\log(|V|)$$

$$= \frac{1}{T} \cdot T \cdot \log(|V|)$$

$$= \log(|V|).$$

$$\text{perplexity} = \exp(\log|V|)$$

$$= |V|.$$

the # of choices that the model will be perplexed (confused) with ,