

1. Word-Level Neural Bigram Language Model

0.a Gradient of Softmax + CE

- Denote SCE - Softmax Cross Entropy function:

$$SCE(\theta, y) = CE(\hat{y}, y) = CE(\text{softmax}(\theta), y)$$

$$= -\sum_i y_i \log(\hat{y}_i)$$

if k - true label :

$$= -\log(\hat{y}_k) = -\log(\text{softmax}(\theta)_k) = -\log\left(\frac{\exp(\theta_k)}{\sum_j \exp(\theta_j)}\right)$$

$$= \log\left(\sum_j \exp(\theta_j)\right) - \theta_k$$

$$\implies SCE(\theta, y) = \log\left(\sum_j \exp(\theta_j)\right) - \theta_k$$

- Now let's calculate it's derivative with respect to some θ_i :

$$\frac{\delta SCE}{\delta \theta_i} = \frac{\log(\sum_j \exp(\theta_j))}{\delta \theta_i} - \frac{\delta \theta_k}{\delta \theta_i}$$

↓

$$\text{denote : } f = \sum_j \exp(\theta_j)$$

↓

$$\frac{\log(\sum_j \exp(\theta_j))}{\delta \theta_i} - \frac{\delta \theta_k}{\delta \theta_i}$$

$$= \frac{\log(f)}{\delta f} \frac{\delta f}{\delta \theta_i} - \frac{\delta \theta_k}{\delta \theta_i}$$

$$= \frac{1}{f} \sum_j \frac{\exp(\theta_j)}{\exp(\theta_i)} - \frac{\delta \theta_k}{\delta \theta_i}$$

$$= \frac{1}{f} \exp(\theta_i) - \frac{\delta \theta_k}{\delta \theta_i}$$

$$= \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)} - \frac{\delta \theta_k}{\delta \theta_i}$$

$$= \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)} - \mathbb{1}(i = k)$$

$$= \text{softmax}(\theta)_i - \mathbb{1}(i = k)$$

$$\implies \nabla_{\theta} SCE = \text{softmax}(\theta) - y = \hat{y} - y$$

- Basically the gradient of this function is prediction softmax vector minus true label one-hot vector.

0.b Gradients of NN

- In this section we use:

$$h = \sigma(xW_1 + b_1)$$

$$\theta = hW_2 + b_2$$

$$\hat{y} = \text{softmax}(\theta)$$

$$CE(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i)$$

$$\nabla_{\theta} SCE = \hat{y} - y$$

- Also notice that Jacobian matrix of a vector function sigmoid is:

$$da(\sigma(a)) = \text{diag}(\sigma'(a)) = \begin{bmatrix} \sigma'(a_1) & 0 & 0 & 0 \\ 0 & \sigma'(a_2) & 0 & 0 \\ 0 & 0 & \sigma'(a_i) & 0 \\ 0 & 0 & 0 & \sigma'(a_n) \end{bmatrix} ; \sigma'(a_i) = \sigma(a_i)(1 - \sigma(a_i))$$

- Let's write a formula for each gradient using differentials:

$$\begin{aligned}
\bullet \quad dx(SCE) &= dx(SCE(hW_2 + b_2)) \\
&= \nabla_{\theta} SCE \, dx(hW_2 + b_2) \\
&= \nabla_{\theta} SCE \, dx(hW_2) \\
&= \nabla_{\theta} SCE \, W_2^T \, dx(h) \\
&= \nabla_{\theta} SCE \, W_2^T \, dx(\sigma(xW_1 + b_1)) \\
&= \nabla_{\theta} SCE \, W_2^T \, \text{diag}(\sigma'(xW_1 + b_1)) \, dx(xW_1 + b_1) \\
&= \nabla_{\theta} SCE \, W_2^T \, \text{diag}(\sigma'(xW_1 + b_1)) \, W_1^T \, dx \\
&\implies \nabla_x SCE = \nabla_{\theta} SCE \, W_2^T \, \text{diag}(\sigma'(xW_1 + b_1)) \, W_1^T \\
&\implies \nabla_x SCE = (\hat{y} - y) \, W_2^T \, \text{diag}(\sigma'(xW_1 + b_1)) \, W_1^T
\end{aligned}$$

This formula won't work for batched data, but we can write it in the more generic way using Hadamard product, i.e. elementwise multiplication - \circ :

$$\implies \nabla_x SCE = \{(\hat{y} - y) \, W_2^T \circ \sigma'(xW_1 + b_1)\} \, W_1^T$$

- Let's sanity check it's dimensions:

$x - 2 \times 10$ (batch of two vectors)

$W_1 - 10 \times 7$

$b_1 - 1 \times 7$

$\sigma(xW_1 + b_1) - 2 \times 7$ (bias is broadcasted)

$\sigma'(xW_1 + b_1) - 2 \times 7$ (bias is broadcasted)

$W_2 - 7 \times 5$

$b_2 - 1 \times 5$

$y - 2 \times 5$ (batch of two vectors)

$$\begin{aligned}
&\implies \nabla_x SCE = \{(\hat{y} - y) \, W_2^T \circ \sigma'(xW_1 + b_1)\} \, W_1^T \\
&\implies 2 \times 10 = \{(2 \times 5) \, (5 \times 7) \circ (2 \times 7)\} \, (7 \times 10)
\end{aligned}$$

All dimensions match.

$$\begin{aligned}
\bullet \quad dW_1(SCE) &= dW_1(SCE(hW_2 + b_2)) \\
&= \nabla_{\theta} SCE \, dW_1(hW_2 + b_2) \\
&= \nabla_{\theta} SCE \, dW_1(hW_2) \\
&= \nabla_{\theta} SCE \, W_2^T \, dW_1(h) \\
&= \nabla_{\theta} SCE \, W_2^T \, dW_1(\sigma(xW_1 + b_1)) \\
&= \nabla_{\theta} SCE \, W_2^T \, \text{diag}(\sigma'(xW_1 + b_1)) \, dW_1(xW_1 + b_1) \\
&= x^T \, \nabla_{\theta} SCE \, W_2^T \, \text{diag}(\sigma'(xW_1 + b_1)) \, dW_1 \\
&\implies \nabla_{W_1} SCE = x^T \, \nabla_{\theta} SCE \, W_2^T \, \text{diag}(\sigma'(xW_1 + b_1)) \\
&\implies \nabla_{W_1} SCE = x^T \, (\hat{y} - y) \, W_2^T \, \text{diag}(\sigma'(xW_1 + b_1))
\end{aligned}$$

Same trick with Hadamard product, i.e. elementwise multiplication - \circ :

$$\implies \nabla_{W_1} SCE = x^T \, \{(\hat{y} - y) \, W_2^T \circ \sigma'(xW_1 + b_1)\}$$

$$\begin{aligned}
\bullet \quad db_1(SCE) &= db_1(SCE(hW_2 + b_2)) \\
&= \nabla_{\theta} SCE \, db_1(hW_2 + b_2) \\
&= \nabla_{\theta} SCE \, db_1(hW_2) \\
&= \nabla_{\theta} SCE \, W_2^T \, db_1(h) \\
&= \nabla_{\theta} SCE \, W_2^T \, db_1(\sigma(xW_1 + b_1)) \\
&= \nabla_{\theta} SCE \, W_2^T \, diag(\sigma'(xW_1 + b_1)) \, db_1(xW_1 + b_1) \\
&= \nabla_{\theta} SCE \, W_2^T \, diag(\sigma'(xW_1 + b_1)) \, 1 \, db_1 \\
&\implies \nabla_{b_1} SCE = \nabla_{\theta} SCE \, W_2^T \, diag(\sigma'(xW_1 + b_1)) \\
&\implies \nabla_{b_1} SCE = (\hat{y} - y) \, W_2^T \, diag(\sigma'(xW_1 + b_1))
\end{aligned}$$

Same trick with Hadamard product, i.e. elementwise multiplication - \circ :

$$\implies \nabla_{b_1} SCE \approx \{(\hat{y} - y) \, W_2^T\} \circ \sigma'(xW_1 + b_1)$$

Notice that dims don't match, because bias was broadcasted. To match the dims, we sum the gradient along batch dimension.

$$\implies \nabla_{b_1} SCE = \sum_{batch=0}^N \{(\hat{y} - y) \, W_2^T\} \circ \sigma'(xW_1 + b_1)_{batch}$$

$$\begin{aligned}
\bullet \quad dW_2(SCE) &= dW_2(SCE(hW_2 + b_2)) \\
&= \nabla_{\theta} SCE \, dW_2(hW_2 + b_2) \\
&= \nabla_{\theta} SCE \, dW_2(hW_2) \\
&= \nabla_{\theta} SCE \, h \, dW_2 \\
&= \sigma(xW_1 + b_1)^T \, \nabla_{\theta} SCE \, dW_2 \\
&\implies \nabla_{W_2} SCE = \sigma(xW_1 + b_1)^T \, \nabla_{\theta} SCE \\
&\implies \nabla_{W_2} SCE = \sigma(xW_1 + b_1)^T \, (\hat{y} - y)
\end{aligned}$$

$$\begin{aligned}
\bullet \quad db_2(SCE) &= db_2(SCE(hW_2 + b_2)) \\
&= \nabla_{\theta} SCE \, db_2(hW_2 + b_2) \\
&= \nabla_{\theta} SCE \, 1 \, dW_2 \\
&\implies \nabla_{b_2} SCE = \nabla_{\theta} SCE \\
&\implies \nabla_{b_2} SCE \approx (\hat{y} - y)
\end{aligned}$$

Adjust dims:

$$\implies \nabla_{b_2} SCE = \sum_{batch=0}^N (\hat{y} - y)_{batch}$$

0.c

0.d Perplexity

Dev perplexity : 112.889

Section 2 - Theoretical Inquiry of a Simple RNN Language Model

- (a) Some notation: Denote the elements of the matrices $\mathbf{H}, \mathbf{I}, \mathbf{U}, \mathbf{L}$ by $H_{ij}, I_{ij}, U_{ij}, L_{ij}$. Let $\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$ (Kronecker delta). Also note that $\mathbf{L}_{\mathbf{x}(t)} = \mathbf{e}^{(t)}$

We will also omit $|_{(t)}$ for notational simplicity but it is assumed when needed.

Let $\boldsymbol{\theta}_1^{(t)} = \mathbf{h}^{(t-1)}\mathbf{H} + \mathbf{e}^{(t)}\mathbf{I} + \mathbf{b}_1$ (logit vector), and denote its k -th element by $\theta_{1,k}^{(t)}$. Calculate its derivatives with respect to the various model parameters:

$$\begin{aligned}\frac{\partial \theta_{1,k}^{(t)}}{\partial H_{ij}} &= \frac{\partial}{\partial H_{ij}} \left(\sum_{\ell} h_{\ell}^{(t-1)} H_{\ell k} \right) = h_i^{(t-1)} \delta_{jk} \\ \frac{\partial \theta_{1,k}^{(t)}}{\partial I_{ij}} &= \frac{\partial}{\partial I_{ij}} \left(\sum_{\ell} e_{\ell}^{(t)} I_{\ell k} \right) = e_i^{(t)} \delta_{jk} \\ \frac{\partial \theta_{1,k}^{(t)}}{\partial b_{1,i}} &= \delta_{ik} \\ \frac{\partial \theta_{1,k}^{(t)}}{e_i^{(t)}} &= I_{ik} \\ \frac{\partial \theta_{1,k}^{(t)}}{\partial h_i^{(t-1)}} &= \frac{\partial}{\partial h_i^{(t-1)}} \left(\sum_{\ell} h_{\ell}^{(t-1)} H_{\ell k} \right) = H_{ik}\end{aligned}$$

We have $\mathbf{h}^{(t)} = \sigma(\boldsymbol{\theta}_1^{(t)})$. Defining $\boldsymbol{\theta}_2^{(t)} = \mathbf{h}^{(t)}\mathbf{U} + \mathbf{b}_2$ (logit vector) and denoting its k -th element by $\theta_{2,k}^{(t)}$, by the chain rule we have

$$\begin{aligned}\frac{\partial \theta_{2,k}^{(t)}}{\partial H_{ij}} &= \frac{\partial}{\partial H_{ij}} \left(\sum_{\ell} h_{\ell}^{(t)} U_{\ell k} \right) \\ &= \sum_{\ell} \frac{\partial h_{\ell}^{(t)}}{\partial H_{ij}} U_{\ell k}\end{aligned}$$

$$\begin{aligned}
&= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) \frac{\partial \theta_{1,\ell}^{(t)}}{\partial H_{ij}} U_{\ell k} \\
&= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) h_i^{(t-1)} \delta_{j\ell} U_{\ell k} \\
&= \sigma'(\theta_{1,k}^{(t)}) h_i^{(t-1)} U_{jk}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta_{2,k}^{(t)}}{\partial I_{ij}} &= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) \frac{\partial \theta_{1,\ell}^{(t)}}{\partial I_{ij}} U_{\ell k} \\
&= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) e_i^{(t)} \delta_{j\ell} U_{\ell k} \\
&= \sigma'(\theta_{1,j}^{(t)}) e_i^{(t)} U_{jk}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta_{2,k}^{(t)}}{\partial b_{1,i}} &= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) \frac{\partial \theta_{1,\ell}^{(t)}}{\partial b_{1,i}} U_{\ell k} \\
&= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) \delta_{i\ell} U_{\ell k} \\
&= \sigma'(\theta_{1,i}^{(t)}) U_{ik}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta_{2,k}^{(t)}}{\partial e_i^{(t)}} &= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) \frac{\partial \theta_{1,\ell}^{(t)}}{\partial e_i^{(t)}} U_{\ell k} \\
&= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) I_{i\ell} U_{\ell k}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta_{2,k}^{(t)}}{b_{2,i}} &= \delta_{ik} \\
\frac{\partial \theta_{2,k}^{(t)}}{U_{ij}} &= \frac{\partial}{\partial U_{ij}} \left(\sum_{\ell} h_{\ell}^{(t)} U_{\ell k} \right) \\
&= h_i^{(t)} \delta_{jk}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta_{2,k}^{(t)}}{h_i^{(t-1)}} &= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) \frac{\partial \theta_{1,\ell}^{(t)}}{\partial h_i^{(t-1)}} U_{\ell k} \\
&= \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) H_{i\ell} U_{\ell k}
\end{aligned}$$

Recall that $\sigma'(t) = \sigma(t)(1 - \sigma(t))$.

Now since $J^{(t)} = \text{CE}(\mathbf{y}^{(t)}, \hat{\mathbf{y}}^{(t)})$ and $\hat{\mathbf{y}}^{(t)} = \text{softmax}(\boldsymbol{\theta}_2^{(t)})$, by the result from problem 1a $\nabla_{\boldsymbol{\theta}_2^{(t)}} J^{(t)} = \mathbf{y}^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})$, i.e. $\frac{\partial J^{(t)}}{\partial \theta_{2,k}^{(t)}} = y_k^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_k$. Therefore we can calculate the derivatives of J with respect to the various parameters by the chain rule:

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial H_{ij}} &= \sum_k \frac{\partial J^{(t)}}{\partial \theta_{2,k}^{(t)}} \frac{\partial \theta_{2,k}^{(t)}}{\partial H_{ij}} \\
&= \sum_k (y_k^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_k) \sigma'(\theta_{1,k}^{(t)}) h_i^{(t-1)} U_{jk} \\
\frac{\partial J^{(t)}}{\partial \mathbf{H}} &= \mathbf{h}^{(t-1)T} (\mathbf{y}^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})) \boldsymbol{\Sigma}(\boldsymbol{\theta}_1^{(t)}) \mathbf{U}^T
\end{aligned}$$

where we define $\boldsymbol{\Sigma}(\mathbf{v})$ as in problem 1a to be the diagonal matrix with i -th diagonal element $\sigma'(v_i) = \sigma(v_i)(1 - \sigma(v_i))$.

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial I_{ij}} &= \sum_k \frac{\partial J^{(t)}}{\partial \theta_{2,k}^{(t)}} \frac{\partial \theta_{2,k}^{(t)}}{\partial I_{ij}} \\
&= \sum_k (y_k^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_k) \sigma'(\theta_{1,j}^{(t)}) e_i^{(t)} U_{jk} \\
\frac{\partial J^{(t)}}{\partial \mathbf{I}} &= \mathbf{e}^{(t)T} (\mathbf{y}^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})) \mathbf{U}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_1^{(t)})
\end{aligned}$$

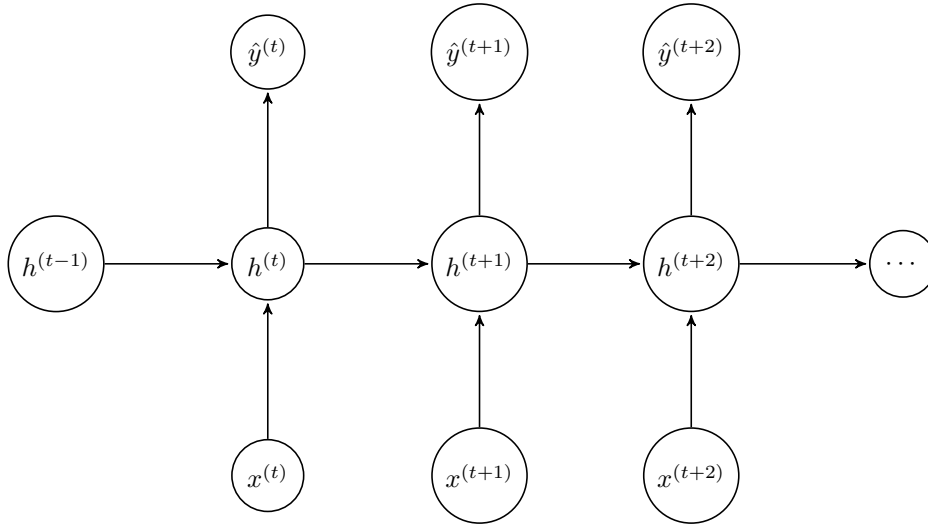
$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial b_{1,i}} &= \sum_k \frac{\partial J^{(t)}}{\partial \theta_{2,k}^{(t)}} \frac{\partial \theta_{2,k}^{(t)}}{\partial b_{1,i}} \\
&= \sum_k (y_k^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_k) \sigma'(\theta_{1,i}^{(t)}) U_{ik} \\
\frac{\partial J^{(t)}}{\partial \mathbf{b}_1} &= (\mathbf{y}^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})) \mathbf{U}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_1^{(t)})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial e_i^{(t)}} &= \sum_k \frac{\partial J^{(t)}}{\partial \theta_{2,k}^{(t)}} \frac{\partial \theta_{2,k}^{(t)}}{\partial e_i^{(t)}} \\
&= \sum_k (y_k^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_k) \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) I_{i\ell} U_{\ell k} \\
\frac{\partial J^{(t)}}{\partial \mathbf{L}_{\mathbf{x}^{(t)}}} &= \frac{\partial J^{(t)}}{\partial \mathbf{e}^{(t)}} = (\mathbf{y}^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})) \mathbf{U}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_1^{(t)}) \mathbf{I}^T
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial b_{2,i}} &= \sum_k \frac{\partial J^{(t)}}{\partial \theta_{2,k}^{(t)}} \frac{\partial \theta_{2,k}^{(t)}}{\partial b_{2,i}} \\
&= \sum_k (y_k^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_k) \delta_{ik} \\
&= (y_i^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_i) \\
\frac{\partial J^{(t)}}{\partial \mathbf{b}_2} &= \mathbf{y}^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial U_{ij}} &= \sum_k \frac{\partial J^{(t)}}{\partial \theta_{2,k}^{(t)}} \frac{\partial \theta_{2,k}^{(t)}}{\partial U_{ij}} \\
&= \sum_k (y_k^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_k) h_i^{(t)} \delta_{jk} \\
&= (y_j^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_j) h_i^{(t)} \\
\frac{\partial J^{(t)}}{\partial \mathbf{U}} &= \mathbf{h}^{(t)T} (\mathbf{y}^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)}))
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial h_i^{(t-1)}} &= \sum_k \frac{\partial J^{(t)}}{\partial \theta_{2,k}^{(t)}} \frac{\partial \theta_{2,k}^{(t)}}{\partial h_i^{(t-1)}} \\
&= \sum_k (y_k^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})_k) \sum_{\ell} \sigma'(\theta_{1,\ell}^{(t)}) H_{i\ell} U_{\ell k} \\
\frac{\partial J^{(t)}}{\partial \mathbf{h}^{(t-1)}} &= (\mathbf{y}^{(t)} - \text{softmax}(\boldsymbol{\theta}_2^{(t)})) \mathbf{U}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_1^{(t)}) \mathbf{H}^T
\end{aligned}$$



(b)

First we calculate the following auxiliary quantities, using all the notation and results from part (a) above:

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial \theta_2^{(t)}} &= \mathbf{y}^{(t)} - \hat{\mathbf{y}}^{(t)} \\
\frac{\partial J^{(t)}}{\partial h_i^{(t)}} &= \sum_j \frac{\partial J^{(t)}}{\partial \theta_{2,j}^{(t)}} \frac{\partial \theta_{2,j}^{(t)}}{\partial h_i^{(t)}} \\
&= \sum_j (y_j^{(t)} - \hat{y}_j^{(t)}) \frac{\partial \theta_{2,j}^{(t)}}{\partial h_i^{(t)}} \\
&= \sum_j (y_j^{(t)} - \hat{y}_j^{(t)}) \frac{\partial}{\partial h_i^{(t)}} \left(\sum_k h_k^{(t)} U_{kj} + b_{2,j} \right) \\
&= \sum_j (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial \mathbf{h}^{(t)}} &= (\mathbf{y}^{(t)} - \hat{\mathbf{y}}^{(t)}) \mathbf{U}^T \\
\frac{\partial J^{(t)}}{\partial \theta_{1,i}^{(t)}} &= \sum_j \frac{\partial J^{(t)}}{\partial h_j^{(t)}} \frac{\partial h_j^{(t)}}{\partial \theta_{1,i}^{(t)}} \\
&= \sum_{j,k} (y_k^{(t)} - \hat{y}_k^{(t)}) U_{jk} \frac{\partial h_j^{(t)}}{\partial \theta_{1,i}^{(t)}} \\
&= \sum_{j,k} (y_k^{(t)} - \hat{y}_k^{(t)}) U_{jk} \frac{\partial}{\partial \theta_{1,i}^{(t)}} \sigma(\theta_{1,j}^{(t)}) \\
&= \sum_{j,k} (y_k^{(t)} - \hat{y}_k^{(t)}) U_{jk} \delta_{ij} \sigma'(\theta_{1,i}^{(t)}) \\
&= \sum_k (y_k^{(t)} - \hat{y}_k^{(t)}) U_{ik} \sigma'(\theta_{1,i}^{(t)})
\end{aligned}$$

$$\frac{\partial J^{(t)}}{\partial \theta_1^{(t)}} = (\mathbf{y}^{(t)} - \hat{\mathbf{y}}^{(t)}) \mathbf{U}^T \boldsymbol{\Sigma}(\theta_1^{(t)})$$

$$\begin{aligned}
\frac{\partial h_i^{(t)}}{\partial e_j^{(t)}} &= \frac{\partial}{\partial e_j^{(t)}} \sigma \left(\sum_k h_k^{(t-1)} H_{ki} + \sum_k e_k^{(t)} I_{ki} + b_{1,i} \right) \\
&= I_{ji} \sigma' \left(\sum_k h_k^{(t-1)} H_{ki} + \sum_k e_k^{(t)} I_{ki} + b_{1,i} \right) \\
&= I_{ji} \sigma'(\theta_{1,i}^{(t)})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial h_i^{(t)}}{\partial H_{jk}} &= \frac{\partial}{\partial H_{jk}} \sigma \left(\sum_\ell h_\ell^{(t-1)} H_{\ell i} + \sum_\ell e_\ell^{(t)} I_{\ell i} + b_{1,i} \right) \\
&= h_j^{(t-1)} \delta_{ik} \sigma'(\theta_{1,i}^{(t)})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial h_i^{(t)}}{\partial I_{jk}} &= \frac{\partial}{\partial I_{jk}} \sigma \left(\sum_\ell h_\ell^{(t-1)} H_{\ell i} + \sum_\ell e_\ell^{(t)} I_{\ell i} + b_{1,i} \right) \\
&= e_j^{(t)} \delta_{ik} \sigma'(\theta_{1,i}^{(t)})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial h_i^{(t)}}{\partial b_{1,j}} &= \frac{\partial}{\partial b_{1,j}} \sigma \left(\sum_\ell h_\ell^{(t-1)} H_{\ell i} + \sum_\ell e_\ell^{(t)} I_{\ell i} + b_{1,i} \right) \\
&= \delta_{ij} \sigma'(\theta_{1,i}^{(t)})
\end{aligned}$$

Now we can calculate the desired quantities:

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial L_{\mathbf{x}^{(t-1)},k}} &= \frac{\partial J^{(t)}}{\partial e_k^{(t-1)}} = \sum_i \frac{\partial J^{(t)}}{\partial \theta_{1,i}^{(t)}} \frac{\partial \theta_{1,i}^{(t)}}{\partial e_k^{(t-1)}} \\
&= \sum_{i,j} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) \frac{\partial \theta_{1,i}^{(t)}}{\partial e_k^{(t-1)}} \\
&= \sum_{i,j} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) \frac{\partial}{\partial e_k^{(t-1)}} \left(\sum_{\ell} h_{\ell}^{(t-1)} H_{\ell i} \right) \\
&= \sum_{i,j,\ell} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{\ell i} \frac{\partial h_{\ell}^{(t-1)}}{\partial e_k^{(t-1)}} \\
&= \sum_{i,j,\ell} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{\ell i} I_{k\ell} \sigma'(\theta_{1,\ell}^{(t-1)}) \\
\frac{\partial J^{(t)}}{\partial \mathbf{L}_{\mathbf{x}^{(t-1)}}} &= (\mathbf{y}^{(t)} - \hat{\mathbf{y}}_k^{(t)}) \mathbf{U}^T \Sigma(\theta_1^{(t)}) \mathbf{H}^T \Sigma(\theta_1^{(t-1)}) \mathbf{I}^T \\
\left. \frac{\partial J^{(t)}}{\partial H_{mn}} \right|_{(t-1)} &= \sum_{i,j,\ell} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{\ell i} \left. \frac{\partial h_{\ell}^{(t-1)}}{\partial H_{mn}} \right|_{(t-1)} \quad (\text{by the same logic as above}) \\
&= \sum_{i,j,\ell} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{\ell i} h_m^{(t-2)} \delta_{\ell n} \sigma'(\theta_{1,\ell}^{(t-1)}) \\
&= \sum_{i,j} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{ni} h_m^{(t-2)} \sigma'(\theta_{1,n}^{(t-1)}) \\
\left. \frac{\partial J^{(t)}}{\partial \mathbf{H}} \right|_{(t-1)} &= \mathbf{h}^{(t-2)T} (\mathbf{y}^{(t)} - \hat{\mathbf{y}}_k^{(t)}) \mathbf{U}^T \Sigma(\theta_1^{(t)}) \mathbf{H}^T \Sigma(\theta_1^{(t-1)}) \\
\left. \frac{\partial J^{(t)}}{\partial I_{mn}} \right|_{(t-1)} &= \sum_{i,j,\ell} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{\ell i} \left. \frac{\partial h_{\ell}^{(t-1)}}{\partial I_{mn}} \right|_{(t-1)} \quad (\text{by the same logic as above}) \\
&= \sum_{i,j,\ell} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{\ell i} e_m^{(t-1)} \delta_{\ell n} \sigma'(\theta_{1,\ell}^{(t-1)}) \\
&= \sum_{i,j} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{ni} e_m^{(t-1)} \sigma'(\theta_{1,n}^{(t-1)}) \\
\left. \frac{\partial J^{(t)}}{\partial \mathbf{I}} \right|_{(t-1)} &= \mathbf{e}^{(t-1)T} (\mathbf{y}^{(t)} - \hat{\mathbf{y}}_k^{(t)}) \mathbf{U}^T \Sigma(\theta_1^{(t)}) \mathbf{H}^T \Sigma(\theta_1^{(t-1)}) \\
\left. \frac{\partial J^{(t)}}{\partial b_{1,k}} \right|_{(t-1)} &= \sum_{i,j,\ell} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{\ell i} \left. \frac{\partial h_{\ell}^{(t-1)}}{\partial b_{1,k}} \right|_{(t-1)} \\
&= \sum_{i,j,\ell} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{\ell i} \delta_{\ell k} \sigma'(\theta_{1,\ell}^{(t-1)}) \\
&= \sum_{i,j} (y_j^{(t)} - \hat{y}_j^{(t)}) U_{ij} \sigma'(\theta_{1,i}^{(t)}) H_{ki} \sigma'(\theta_{1,k}^{(t-1)}) \\
\left. \frac{\partial J^{(t)}}{\partial \mathbf{b}_1} \right|_{(t-1)} &= (\mathbf{y}^{(t)} - \hat{\mathbf{y}}_k^{(t)}) \mathbf{U}^T \Sigma(\theta_1^{(t)}) \mathbf{H}^T \Sigma(\theta_1^{(t-1)})
\end{aligned}$$

3. GRU question

Advantage Smaller discrete space - There are about 100 English-language characters in common usage if we include all punctuation marks. By contrast, a vocabulary is many thousands of words. For character-based model we need about 100 embeddings to represent all possible tokens.

Disadvantage Char-level models can generate unusual words. Word-level models can't generate mistyped words as these are not in their vocabulary.

4. Perplexity

Write $p_i = p(s_i | s_1, \dots, s_{i-1})$. Then for any $b > 0$,

$$b^{-\frac{1}{M} \sum_{i=1}^M \log_b p_i} = (b^{\sum_{i=1}^M \log_b p_i})^{-1/M} = \left(\prod_{i=1}^M b^{\log_b p_i} \right)^{-1/M} = \left(\prod_{i=1}^M p_i \right)^{-1/M}$$

Therefore this expression has the same value for any such b , and in particular for $b = 2$ and $b = e$ which are the two given expressions.