1 Word-Level Neural Bigram Language Model

(a)

$$\left[\nabla_{\underline{\theta}} C E(\underline{Y}, \hat{\underline{Y}})\right]_i \overset{i=j}{=} \frac{d}{d\theta_j}\left\{-\log\left[\frac{\exp(\theta_j)}{\sum_j \exp(\theta_j)}\right]\right\} =$$

Assume: $(\underline{Y})_j = 1$

$$= -\left[\frac{\sum_j \exp(\theta_j)}{\exp(\theta_j)}\right] \cdot \left[\frac{\exp(\theta_j)\left[\left[\sum_j \exp(\theta_j)\right] - \exp(\theta_j)\right]}{\left[\sum_j \exp(\theta_j)\right]^2}\right] =$$

$$= -\left[\frac{\sum_j \exp(\theta_j) - \exp(\theta_j)}{\sum_j \exp(\theta_j)}\right] = \frac{\exp(\theta_j)}{\sum_j \exp(\theta_j)} - 1 =$$

$$= \text{softmax}(\theta_j) - 1.$$

... $\overset{i \neq j}{=}$ $\frac{d}{d\theta_i}\left[-\log\left[\frac{\exp(\theta_j)}{\sum_j \exp(\theta_j)}\right]\right] =$

$$= \frac{\sum_j \exp(\theta_j)}{\exp(\theta_j)} \cdot \left[\frac{\exp(\theta_j) \cdot \exp(\theta_i)}{\left(\sum_j \exp(\theta_j)\right)^2}\right] =$$

$$= \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)} = \text{softmax}(\theta_i).$$

$\Rightarrow$

$$\boxed{\nabla_{\underline{\theta}} CE(\underline{y}, \underline{\hat{y}}) = \underline{\hat{y}} - \underline{y}}$$

[a]

(b)

Denote $\underline{\theta} = \underline{h} W_2 + \underline{b}_2 \Rightarrow \underline{\hat{y}} = \text{Softmax}(\underline{\theta})$.

$\underline{r} = \underline{x} W_1 + \underline{b}_1 \Rightarrow \underline{h} = \sigma(\underline{r})$

$$\frac{\partial J}{\partial \underline{x}} = \left(\frac{\partial J}{\partial \underline{\theta}}\right) \cdot \left(\frac{\partial \underline{\theta}}{\partial \underline{h}}\right) \cdot \left(\frac{d\underline{h}}{\partial \underline{r}}\right) \cdot \left(\frac{\partial \underline{r}}{\partial \underline{x}}\right)$$

diagonal-Jacobian:

$$\begin{bmatrix} \sigma(\underline{r})_1 \cdot \sigma(\underline{r})_1 & & \\ & \ddots & \\ & & \sigma(\underline{r})_{dh} \cdot \sigma(\underline{r})_{dh} \end{bmatrix}$$

Chain rule:

$$= \underset{(1\times d_y)}{[\hat{y} - y]} \underset{(d_y \times d_h)}{[W_2^T]} \underset{d_h \times d_h}{[\quad M \quad]} \underset{(d_h \times d_x)}{[W_1^T]} =$$

[b]

$$= [\hat{y} - y][W_2^T][\text{diag}[h \circ (1-h)]] \cdot [W_1^T].$$

(d) perplexity: 112.81714028757714

[2] Theoretical Inquiry of simple RNN Language-model

- denote: $\underline{o}^t = \underline{h}^t V + \underline{b}_2$; $\left[\underline{\hat{y}}^t - \underline{y}^t\right] = \underline{\mu}^t$; $\underline{V}^t = \underline{h}^{t-1} H + \underline{e}^t I + \underline{b}_1$

(a) $$\frac{\partial J^{(t)}}{\partial U_{ij}} = \left(\frac{\partial J^t}{\partial \underline{o}^t}\right)\left(\frac{\partial \underline{o}^t}{\partial U_{ij}}\right) =$$

$$= \left[\underline{\hat{y}}^t - \underline{y}^t\right] \begin{bmatrix} 0 \\ \vdots \\ h_i^t \\ \vdots \\ 0 \end{bmatrix} \leftarrow j\text{'th row} \quad = \underbrace{\left[\underline{\hat{y}}^t - \underline{y}^t\right]_j}_{\text{Denote } \underline{\mu}^t} \cdot h_i^t$$

$$\Rightarrow \boxed{\frac{\partial J^{(t)}}{\partial U_{ij}}} = \begin{bmatrix} \mu_1^t \cdot h_1^t & \cdots & \mu_{in}^t \cdot h_1^t \\ \mu_1^t \cdot h_2^t & & \\ \vdots & & \\ \mu_1^t \cdot h_{dh}^t & \cdots & \mu_{|V|}^t \cdot h_{dh}^t \end{bmatrix} = \left[\underline{h}^t\right]^T \cdot \left[\underline{\hat{y}}^t - \underline{y}^t\right] =$$

$$\boxed{\underline{h}^t \otimes \left[\underline{\hat{y}}^t - \underline{y}^t\right] = \left(\underline{h}^t\right)^T \cdot \left[\underline{\hat{y}}^t - \underline{y}^t\right]}$$

Outer-product.

Outer-product [when viewed as row vectors!].

$$\left(\frac{\partial J^t}{\partial o^t}\right) = \left[\underline{\hat{y}}^t - \underline{y}^t\right]; \quad \frac{\partial J^t}{\partial V^t} = \left(\frac{\partial J^t}{\partial o^t}\right)\left(\frac{\partial o^t}{\partial h^t}\right)\left(\frac{\partial h^t}{\partial V^t}\right) =$$

$$= \left[\underline{\hat{y}}^t - \underline{y}^t\right] U^T \cdot \text{diag}\left[h^t \circ (1 - h^t)\right]$$

$\circ$: Hadamard-product.

(1.)

$$\frac{\partial J^t}{\partial \underline{b}_{2i}} = \left(\frac{\partial J^t}{\partial \underline{\theta}_2^t}\right)\left(\frac{\partial \underline{\theta}_2^t}{\partial \underline{b}_{2i}}\right) =$$

$$= [\underline{y}^t - \underline{\hat{y}}^t] \cdot \begin{bmatrix} \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \leftarrow i\text{'th row} = [\underline{\mu}^t]_i$$

$$\Rightarrow \boxed{\frac{\partial J^t}{\partial \underline{b}_2} = [[\underline{\mu}^t]_1 \cdots [\underline{\mu}^t]_{M_1}] = \underline{\mu}^t = [\underline{\hat{y}}^t - \underline{y}^t]}.$$

(2)

$$\frac{\partial J^t}{\partial b_1} = \left(\frac{\partial J^t}{\partial V^t}\right)\left(\frac{\partial V^t}{\partial b_1}\right) = \left(\frac{\partial J^t}{\partial V^t}\right) \cdot I_{D_h \times D_h}$$

$$= \left[\hat{y}^t - y^t\right] U^T \cdot \text{diag}\left[h^t \circ (1 - h^t)\right].$$

$$\frac{\partial J^t}{\partial I} = \left(\frac{\partial J^t}{\partial V^t}\right)\left(\frac{\partial V^t}{\partial I}\right) = \left[(e^t)^T \cdot \left[\hat{y}^t - y^t\right] U^T \cdot \text{diag}\left[h^t \circ (1 - h^t)\right]\right].$$

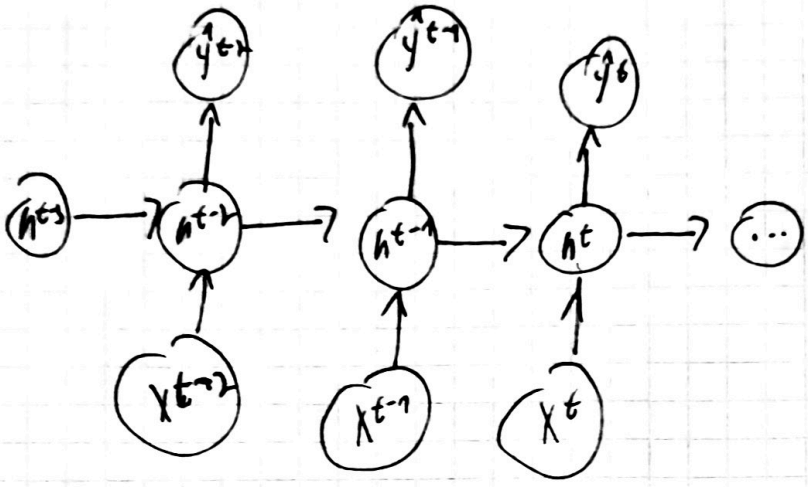
$$\frac{\partial J^t}{\partial H} = \left(\frac{\partial J^t}{\partial V^t}\right)\left(\frac{\partial V^t}{\partial H}\right) = \left[(h^{t-1})^T \cdot \left[\hat{y}^t - y^t\right] U^T \cdot \text{diag}\left[h^t \circ (1 - h^t)\right]\right].$$

$$\frac{\partial J^t}{\partial L_{x^t}} = \left(\frac{\partial J^t}{\partial V^t}\right)\left(\frac{\partial V^t}{\partial e^t}\right)\left(\frac{\partial e^t}{\partial L_{x^t}}\right) = \left[\left[\hat{y}^t - y^t\right] U^T \cdot \text{diag}\left[h^t \circ (1 - h^t)\right] \cdot I^T\right]$$

$$\frac{\partial J^t}{\partial h^{t-1}} = \left(\frac{\partial J^t}{\partial V^t}\right)\left(\frac{\partial V^t}{\partial h^{t-1}}\right) = \left[\hat{y}^t - y^t\right] U^T \cdot \text{diag}\left[h^t \circ (1 - h^t)\right] \cdot H^T = \left[\delta^{t-1}\right]$$

(2)

(b)



V^{t-1}

$$h^{t-1} = \sigma(h^{t-2}H + e^{t-1}I + b_1)$$

$$h^{t} = \sigma(h^{t-1}H + e^{t}I + b_1)$$

$$\left(\frac{\partial J^t}{\partial I}\right)\Big|_{t-1} = \left(\frac{\partial J^t}{\partial h^{t-1}}\right)\left[\left(\frac{\partial h^{t-1}}{\partial V^{t-1}}\right)\left(\frac{\partial V^{t-1}}{\partial I}\right)\right] =$$

$$= [e^{t-1}]^T [\delta^{t-1}][\text{diag}[h^{t-1} \circ (1 - h^{t-1})]].$$

$$\left(\frac{\partial J^t}{\partial H}\right)\Big|_{t-1} = \left(\frac{\partial J^t}{\partial h^{t-1}}\right)\left(\frac{\partial h^{t-1}}{\partial V^{t-1}}\right)\left(\frac{\partial V^{t-1}}{\partial H}\right) =$$

$$= [h^{t-2}]^T [\delta^{t-1}][\text{diag}[h^{t-1} \circ (1 - h^{t-1})]]$$

$$\left(\frac{\partial J^t}{\partial L_x^{t-1}}\right)\Big|_{t-1} = \left(\frac{\partial J^t}{\partial h^{t-1}}\right)\cdot\left(\frac{\partial h^{t-1}}{\partial v^{t-1}}\right)\cdot\left(\frac{\partial v^{t-1}}{\partial e^{t-1}}\right)\left(\frac{\partial e^{t-1}}{\partial L_x^{t-1}}\right) =$$

$$= \left(\delta^{t-1}\right)\cdot\left(\text{diag}\left[h^{t-1}\circ(1-h^{t-1})\right]\right)\cdot\left(I\right)^T$$

$$\left(\frac{\partial J^t}{\partial b_1}\right)\Big|_{t-1} = \left(\frac{\partial J^t}{\partial h^{t-1}}\right)\left(\frac{\partial h^{t-1}}{\partial v^{t-1}}\right)\left(\frac{\partial v^{t-1}}{\partial b_1}\right) =$$

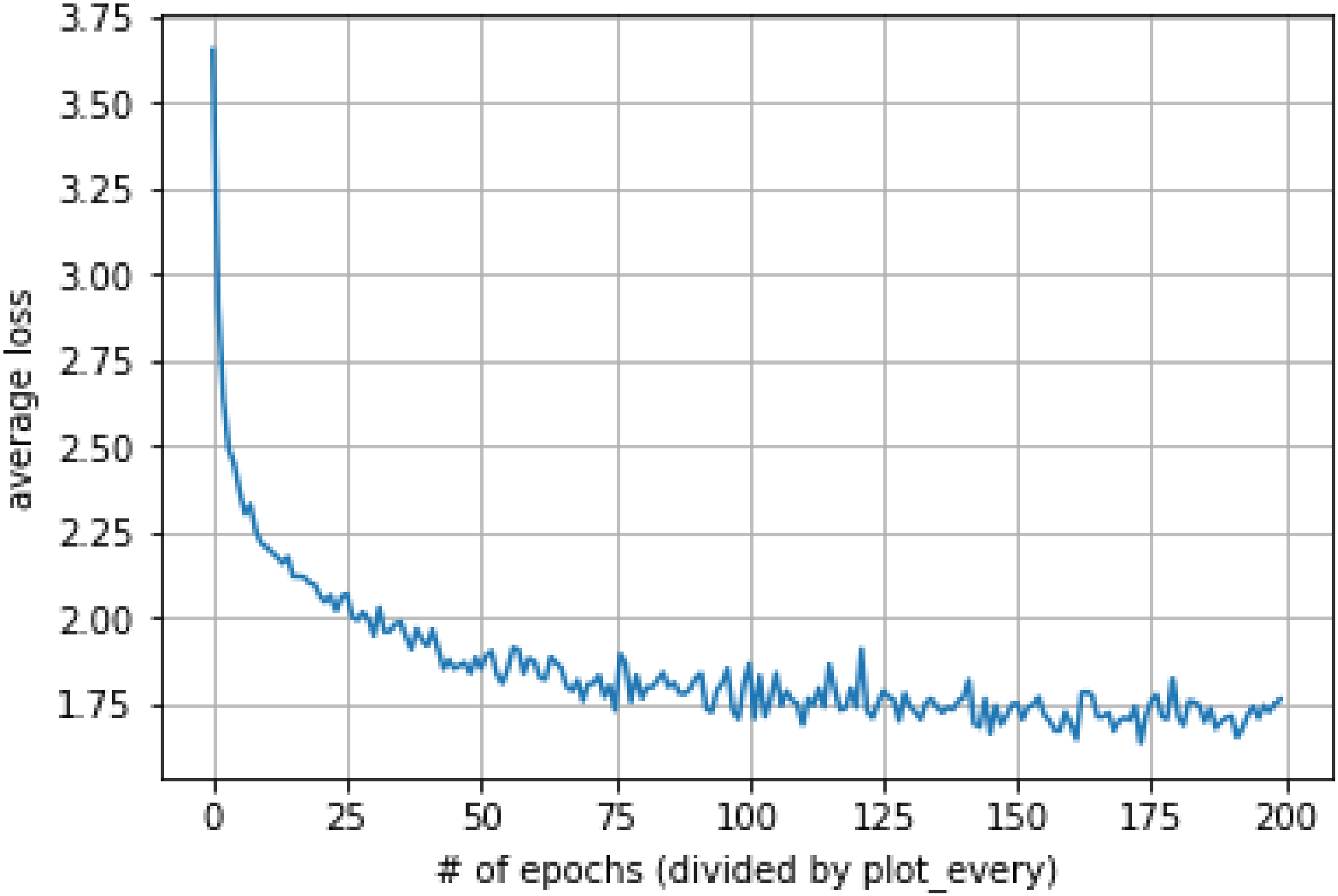
$$= \left(\delta^{t-1}\right)\left(\text{diag}\left[\sigma(v^{t-1})\circ\sigma(-v^{t-1})\right]\right)\left(I_{DH\times D_h}\right) =$$

$$= \left(\delta^{t-1}\right)\left[\text{diag}\left[h^{t-1}\circ(1-h^{t-1})\right]\right].$$

b

[3] Generating Shakespeare Using a Character-level Language Model.

- One very noticable advantage a char-based model would have over a word-based model is in terms of computational resources. The dimension $|V|$ is absolutely huge in comparison to the number of characters in the language [for most languages]. Working with matrices of a much smaller dimension leads to a very substantial improvment in terms of runtime & space required.
- Also, a char-based model would also be more flexible in it's ability to treat language-anomalities [since the number of different characters is limited whilst that of words is possibly infinite].
- ↳ That same trait might also come as a disadvantage in some tasks where we wish to avoid gibberish-meaningless words from being generated.


3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
average loss
0
25
50
75
100
125
150
175
200
# of epochs (divided by plot_every)

4 perplexity:

- Denote: $q(s_i) = p(s_i \mid s_1, \dots s_{i-1})$

$$b^{\left[-\frac{1}{m}\sum_{i=1}^{m}\log_b(q(s_i))\right]} = b^{\left[-\frac{1}{m}\log_b\left[\prod_{i=1}^{m} q(s_i)\right]\right]} =$$

$$= b^{\log_b\left[\left[\prod_{i=1}^{m} q(s_i)\right]^{-\frac{1}{m}}\right]} = \left[\prod_{i=1}^{m} q(s_i)\right]^{-\frac{1}{m}}.$$

לכן נוכל לקחת בסיס b, ובפרט עבור b=2,e

$\Leftarrow$

$$2^{-\frac{1}{m}\sum_{i=1}^{m}\log_2(q(s_i))} = e^{-\frac{1}{m}\sum_{i=1}^{m}\ln(q(s_i))}$$

כנדרש.

(7.)