2) $P_\theta(y^{(j)} = i \mid x^{(j)}, \theta) = \text{softmax}_i(x^{(j)})$

$$\text{softmax}_i(x) = \frac{e^{\tilde{w}_i^T x}}{\sum_{k=1}^{C} e^{\tilde{w}_k^T x}}$$

ASSUME: $\tilde{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}$ $\tilde{w} = \begin{bmatrix} w \\ b \end{bmatrix}$

LIKELIHOOD:

$$L_H = p(\tilde{x}^{(1)}, \dots, \tilde{x}^{(m)}, y^{(1)} \dots y^{(m)} \mid \theta) = \prod_{j=1}^{m} p(x^{(j)}, y^{(j)} \mid \theta)$$

$$= \prod_{j=1}^{m} p(\tilde{x}^{(j)} \mid \theta)\, p(y^{(j)} \mid \tilde{x}^{(j)}, \theta)$$

$$= \prod_{j=1}^{m} p(y^{(j)} \mid \tilde{x}^{(j)}, \theta)$$

$$= \prod_{j=1}^{m} \text{softmax}_{y^{(j)}}(\tilde{x}^{(j)})$$

$$= \prod_{j=1}^{m} \left[\frac{e^{a_{y^{(j)}}(\tilde{x}^{(j)})}}{\sum_{k=1}^{C} e^{a_k(\tilde{x}^{(j)})}} \right]$$

where $a_k(\tilde{x}^{(j)}) = \tilde{w}_k^T \tilde{x}^{(j)}$

$$\log L_H = L = \sum_{j=1}^{m} \left[\left(a_{y^{(j)}}(\tilde{x}^{(j)}) \right) - \log \sum_{k=1}^{C} e^{a_k(\tilde{x}^{(j)})} \right]$$

WE WANT TO MAXIMIZE THE LOG LIKELIHOOD:

$$\frac{\partial L}{\partial \theta} = 0$$

$$\nabla_{\tilde{\omega}_i} L = \frac{\partial L}{\partial \tilde{\omega}_i} = \frac{\partial L}{\partial a_i} \times \frac{\partial a_i}{\partial \tilde{\omega}_i}$$

$$= \sum_{j=1}^{m} \frac{-1}{\sum_{k=1}^{c} e^{a_k(\tilde{x}(j))}} \times e^{a_k(\tilde{x}(j))} \times \tilde{x}(j)$$

$$+ \sum_{\substack{j=1 \\ \& \\ y(j)=i}}^{m} \tilde{x}(j)$$

//

TAKE $I_{ji} = \begin{cases} 1, & y(j) = i \\ 0, & y(j) \neq i \end{cases}$

$$\nabla_{\tilde{\omega}_i} L = \sum_{j=1}^{m} \left[I_{ji} - \frac{e^{a_k(\tilde{x}(j))}}{\sum_{k=1}^{c} e^{a_k(\tilde{x}(j))}} \right] \cdot \tilde{x}(j)$$

$$\nabla_{w_i} L = \sum_{j=1}^{m} \left[I_{ji} - \frac{e^{a_k(x^{(j)})}}{\sum_{k=1}^{c} e^{a_k(x^{(j)})}} \right] \cdot x^{(j)}$$

$$\nabla_{b_i} L = \sum_{j=1}^{m} \left[I_{ji} - \frac{e^{a_k(x^{(j)})}}{\sum_{k=1}^{c} e^{a_k(x^{(j)})}} \right]$$

WE CAN ALSO DERIVE FOR A SINGLE SAMPLE:

$$L_j(\theta) = a_{y^{(j)}}(\tilde{x}^{(j)}) - \log \sum_{k=1}^{c} e^{a_k(\tilde{x}^{(j)})}$$

$$= \log \left(\frac{e^{a_{y^{(j)}}(\tilde{x}^{(j)})}}{\sum_{k=1}^{c} e^{a_k(\tilde{x}^{(j)})}} \right)$$

$$a_k(\tilde{x}^{(j)}) = \tilde{w}_k^T \tilde{x}^{(j)}$$

LET $$\sigma_{y^{(j)}}(\tilde{x}^{(j)}) = \frac{e^{a_{y^{(j)}}(\tilde{x}^{(j)})}}{\sum_{k=1}^{c} e^{a_k(\tilde{x}^{(j)})}}$$

$$\frac{\partial \sigma_{y(j)}(\tilde{x}(j))}{\partial a_k(\tilde{x}(j))}$$

$\rightarrow$ if $y(j) = k$

$$\frac{\partial \sigma_k(\tilde{x}(j))}{\partial a_k(\tilde{x}(j))} = \frac{\partial}{\partial a_k(\tilde{x}(j))}\left(\frac{e^{a_k(\tilde{x}j)}}{\sum_{k=1}^{c} e^{a_k(\tilde{x}(j))}}\right)$$

$$= \frac{\sum_{k=1}^{c} e^{a_k(\tilde{x}(j))} \cdot e^{a_k(\tilde{x}(j))} - e^{a_k(\tilde{x}(j))} \cdot e^{a_k(\tilde{x}(j))}}{\left[\sum_{k=1}^{c} e^{a_k(\tilde{x}(j))}\right]^2}$$

$$= \sigma_{y(j)}(\tilde{x}(j))\left[1 - \sigma_{y(j)}(\tilde{x}(j))\right]$$

$\rightarrow$ if $y(j) \neq k$

$$\frac{\partial \sigma_{y(j)}(\tilde{x}(j))}{\partial a_k(\tilde{x}(j))} = \frac{\partial}{\partial a_k(\tilde{x}(j))}\left(\frac{e^{a_{y(j)}(\tilde{x}j)}}{\sum_{k=1}^{c} e^{a_k(\tilde{x}(j))}}\right)$$

$$= \frac{\displaystyle\sum_{k=1}^{C} e^{a_k(\tilde{x}(j))} \cdot 0 - e^{a_{y(j)}(\tilde{x}(j))} \cdot e^{a_k(\tilde{x}(j))}}{\left[\displaystyle\sum_{k=1}^{C} e^{a_k(\tilde{x}(j))}\right]^2}$$

$$= -\sigma_{y(j)}(\tilde{x}(j))\,\sigma_k(\tilde{x}(j))$$

$$\frac{\partial L_j(\theta)}{\partial a_k(\tilde{x}(j))} = \frac{\partial L_j(\theta)}{\partial \sigma_{y(j)}(\tilde{x}(j))} \cdot \frac{\partial \sigma_{y(j)}(\tilde{x}(j))}{\partial a_k(\tilde{x}(j))}$$

$$L_j(\theta) = \log\left(\frac{e^{a_{y(j)}(\tilde{x}(j))}}{\displaystyle\sum_{k=1}^{C} e^{a_k(\tilde{x}(j))}}\right)$$

$$= \log\left(\sigma_{y(j)}(\tilde{x}(j))\right)$$

$$\frac{\partial L_j(\theta)}{\partial a_k(\tilde{x}(j))} = \begin{cases} (1 - \sigma_{y(j)}(\tilde{x}(j)), & y(j) = k \\ -\sigma_k(\tilde{x}(j)), & y(j) = k \end{cases}$$

$$
\frac{\partial L_j(\theta)}{\partial \tilde{w}_k} = \begin{cases} \left(1 - \sigma_{y(j)}\left(\tilde{x}^{(j)}\right)\right)\tilde{x}^{(j)}, & y^{(j)} = k \\ \quad \hookrightarrow \text{here } \tilde{x}^{(j)} = \begin{bmatrix} x^{(j)} \\ 1 \end{bmatrix} & \\ -\sigma_k\left(\tilde{x}^{(j)}\right)\tilde{x}^{(j)}, & y^{(j)} = k \end{cases}
$$

$$
\frac{\partial L_j(\theta)}{\partial w_k} = \begin{cases} \left(1 - \sigma_{y(j)}\left(x^{(j)}\right)\right)x^{(j)}, & y^{(j)} = k \\ \quad \hookrightarrow \text{here } \tilde{x}^{(j)} = \begin{bmatrix} x^{(j)} \\ 1 \end{bmatrix} & \\ -\sigma_k\left(x^{(j)}\right)x^{(j)}, & y^{(j)} = k \end{cases}
$$

$$
\frac{\partial L_j(\theta)}{\partial b_k} = \begin{cases} \left(1 - \sigma_{y(j)}\left(x^{(j)}\right)\right), & y^{(j)} = k \\ \quad \hookrightarrow \text{here } \tilde{x}^{(j)} = \begin{bmatrix} x^{(j)} \\ 1 \end{bmatrix} & \\ -\sigma_k\left(x^{(j)}\right), & y^{(j)} = k \end{cases}
$$