② Softmax Classifier Gradient

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)}); \quad x^{(j)} \in \mathbb{R}^n, \quad y^{(j)} \in \{1, \ldots, c\}, \quad j=1, \ldots, m$$

$$\theta = \{w_i, b_i\}_{i=1,\ldots,c}$$

$$\Pr\left(y^{(j)} = i \mid x^{(j)}, \theta\right) = \text{softmax}_i\left(x^{(j)}\right)$$

$$\text{softmax}_i(x) = \frac{e^{w_i^T x + b_i}}{\sum_{k=1}^{c} e^{w_k^T x + b_k}}$$

$$\boxed{\tilde{x} = \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \tilde{w}_i = \begin{bmatrix} w_i \\ b_i \end{bmatrix} \quad \Rightarrow a_i(x) = \tilde{w}_i^T \tilde{x}}$$

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

$$= \prod_{i=1}^{m} p\left(y^{(i)} \mid x^{(i)}, \theta\right) p\left(x^{(i)} \mid \theta\right)$$

$$\underset{\theta}{\text{argmax}} \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}, \theta)\, p(x^{(i)} \mid \theta) = \underset{\theta}{\text{argmax}} \prod_{i=1}^{m} p\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

$$= \underset{\theta}{\text{argmax}} \sum_{i=1}^{m} \log \text{softmax}_{y^{(i)}}\left(x^{(i)}\right) = \underset{\theta}{\text{argmax}} \sum_{i=1}^{m} \log \left[\frac{e^{a_{y^{(i)}}(x^{(i)})}}{\sum_{j=1}^{c} e^{a_j(x^{(i)})}} \right]$$

$$= \underset{\theta}{\text{argmax}} \frac{1}{m} \sum_{i=1}^{m} \left[a_{y^{(i)}}\left(x^{(i)}\right) - \log \sum_{j} e^{a_j(x^{(i)})} \right]$$

$$= \underset{\theta}{\text{argmin}} \frac{1}{m} \sum_{i=1}^{m} \left[\log \sum_{j=1}^{c} e^{a_j(x^{(i)})} - a_{y^{(i)}}\left(x^{(i)}\right) \right]$$

$$\boxed{\mathcal{L}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[\log \sum_{j=1}^{c} e^{a_j(x^{(i)})} - a_{y^{(i)}}\left(x^{(i)}\right) \right]}$$

$$\nabla_{\tilde{w}_i}\left(\log\sum_{j=1}^{c} e^{a_j(x)}\right) = \nabla_{\tilde{w}_i}\left(\log\left[e^{\tilde{w}_1^T\tilde{x}} + e^{\tilde{w}_2^T\tilde{x}} + \cdots + e^{\tilde{w}_i^T\tilde{x}} + \cdots + e^{\tilde{w}_c^T\tilde{x}}\right]\right)$$

$$\frac{\partial}{\partial \tilde{w}_{i1}}\left(\log\left[e^{\tilde{w}_1^T\tilde{x}} + \cdots + e^{\tilde{w}_{i1}x_1 + \tilde{w}_{i2}x_2 + \cdots \tilde{w}_{in}x_n + \tilde{w}_{i0}} + \cdots + e^{\tilde{w}_c^T\tilde{x}}\right]\right)$$

$$= \frac{1}{\sum_j e^{a_j(x)}}\frac{\partial}{\partial \tilde{w}_{i1}}\left[\underbrace{e^{\tilde{w}_1^T\tilde{x}} + \cdots}_{0} + e^{\tilde{w}_{i1}x_1 + \tilde{w}_{i2}x_2 + \cdots \tilde{w}_{in}x_n + \tilde{w}_{i0}} \underbrace{+ \cdots + e^{\tilde{w}_c^T\tilde{x}}}_{0}\right]$$

$$= \frac{1}{\sum_j e^{a_j(x)}} x_1 e^{\tilde{w}_i^T\tilde{x}} = \frac{e^{\tilde{w}_i^T\tilde{x}}}{\sum_j e^{a_j(x)}} x_1$$

$$\frac{\partial}{\partial \tilde{w}_{i2}}\left(\log\sum_{j=1}^{c} e^{a_j(x)}\right) = \frac{e^{\tilde{w}_i^T\tilde{x}}}{\sum_j e^{a_j(x)}} x_2$$

$$\boxed{\nabla_{\tilde{w}_i}\left(\log\sum_{j=1}^{c} e^{a_j(x)}\right) = \frac{e^{a_i(x)}}{\sum_{j=1}^{c} e^{a_j(x)}}\tilde{x}}$$

$$\nabla_{\tilde{w}_i} a_{y(k)}(x) = \begin{cases} 0 & \text{if } y(k) \neq i \\ \tilde{x} & \text{if } y(k) = i \end{cases}$$

$\hookrightarrow y(k) \neq i \Rightarrow 0$

$y(k) = i \Rightarrow \nabla_{\tilde{w}_i}\tilde{w}_i^T\tilde{x} = \tilde{x}$

$$\Rightarrow \nabla_{\tilde{w}_i}\mathcal{L} = \frac{1}{m}\sum_{k=1}^{m}\left[\frac{e^{a_i(x^{(k)})}}{\sum_j e^{a_j(x^{(k)})}}\tilde{x}^{(k)} - \delta_{y^{(k)},i}\tilde{x}^{(k)}\right]$$

$$\boxed{\nabla_{w_i}\mathcal{L} = \frac{1}{m}\sum_{k=1}^{m}\left[\frac{e^{a_i(x^{(k)})}}{\sum_j e^{a_j(x^{(k)})}} - \delta_{y^{(k)},i}\right]x^{(k)}}$$

$$\boxed{\nabla_{b_i}\mathcal{L} = \frac{1}{m}\sum_{k=1}^{m}\left[\frac{e^{a_i(x^{(k)})}}{\sum_j e^{a_j(x^{(k)})}} - \delta_{y^{(k)},i}\right]}$$