Partial derivative of the softmax activation

$$a_{ij} = \frac{e^{z_{ij}}}{\sum_{s=1}^{C} e^{z_{is}}} = \frac{e^{z_{ij}}}{\Sigma}$$

$$\frac{\partial a_{ij}}{\partial z_{ik}} = \frac{\Sigma \frac{\partial e^{z_{ij}}}{\partial z_{ik}} - e^{z_{ij}} \frac{\partial \Sigma}{\partial z_{ik}}}{\Sigma^2} = \frac{\frac{\partial e^{z_{ij}}}{\partial z_{ik}}}{\Sigma} - \frac{e^{z_{ij}} \frac{\partial \Sigma}{\partial z_{ik}}}{\Sigma^2} = \frac{e^{z_{ij}} \mathbb{I}_{\{j=k\}}}{\Sigma} - \frac{e^{z_{ij}} \frac{\partial \sum_{s=1}^{C} e^{z_{is}}}{\partial z_{ik}}}{\Sigma^2} = \frac{e^{z_{ij}} \mathbb{I}_{\{j=k\}}}{\Sigma} - \frac{e^{z_{ij}} e^{z_{ik}}}{\Sigma^2}$$

$$= \frac{e^{z_{ij}}}{\Sigma}\left(\mathbb{I}_{\{j=k\}} - \frac{e^{z_{ik}}}{\Sigma}\right) = a_{ij}\left(\mathbb{I}_{\{j=k\}} - a_{ik}\right)$$

Loss for one observation

$$L_{ij} = -\sum_{j=1}^{C} \left\{\mathbb{I}_{\{j=y_i\}} \log a_{ij}\right\} = -\mathbb{I}_{\{j=y_i\}} \log a_{ij}$$

$$\frac{\partial L_i}{\partial z_{ik}} = -\sum_{j=1}^{C} \left\{\mathbb{I}_{\{j=y_i\}} \frac{\partial \log a_{ij}}{\partial z_{ik}}\right\} = -\sum_{j=1}^{C} \left\{\mathbb{I}_{\{j=y_i\}} \frac{1}{a_{ij}} \frac{\partial a_{ij}}{\partial z_{ik}}\right\} = -\sum_{j=1}^{C} \left\{\mathbb{I}_{\{j=y_i\}} \left(\mathbb{I}_{\{j=k\}} - a_{ik}\right)\right\}$$

$$= -\mathbb{I}_{\{k=y_i\}}(1 - a_{ik}) + \sum_{j=1, j\neq k}^{C} \mathbb{I}_{\{j=y_i\}} a_{ik} = -\mathbb{I}_{\{k=y_i\}} + a_{ik}\mathbb{I}_{\{k=y_i\}} + \sum_{j=1, j\neq k}^{C} \mathbb{I}_{\{j=y_i\}} a_{ik}$$

$$= -\mathbb{I}_{\{k=y_i\}} + a_{ik} \sum_{j=1}^{C} \mathbb{I}_{\{j=y_i\}} = -\mathbb{I}_{\{k=y_i\}} + a_{ik}$$

Loss across all observations

$$L = \frac{1}{N} \sum_{i=1}^{N} L_i$$

$$\frac{\partial L}{\partial L_i} = \frac{1}{N}$$

$$\frac{\partial L}{\partial z_{ik}} = \frac{\partial L}{\partial L_i} \frac{\partial L_i}{\partial z_{ik}} = \frac{1}{N}\left(a_{ik} - \mathbb{I}_{\{k=y_i\}}\right)$$

$$\frac{\partial L}{\partial \boldsymbol{z}} = \frac{\partial L}{\partial L_i} \frac{\partial L_i}{\partial \boldsymbol{z}} = \frac{1}{N}\left(\boldsymbol{a} - \mathbb{I}_{\{\boldsymbol{z}[range(N), \boldsymbol{y}]\}}\right)$$

Python

$$\frac{\partial L}{\partial \boldsymbol{z}^{(2)}} = \frac{\partial L}{\partial L_i} \frac{\partial L_i}{\partial \boldsymbol{z}^{(2)}} = \frac{1}{N}\left(\boldsymbol{a}^{(2)} - \mathbb{I}_{\{\boldsymbol{z}^{(2)}[range(N), \boldsymbol{y}]\}}\right)$$

$$\boldsymbol{z}^{(1)} = X \cdot W^{(1)} + \boldsymbol{b}^{(1)} \quad (N, H) = (N, D) \cdot (D, H) + (H,)$$

$$\frac{\partial L}{\partial W^{(1)}} = X^{\top} \cdot \frac{\partial L}{\partial \boldsymbol{z}^{(1)}} \quad (D, H) = (D, N) \cdot (N, H)$$

$$\frac{\partial L}{\partial \boldsymbol{b}^{(1)}} = sum\left(\frac{\partial L}{\partial \boldsymbol{z}^{(1)}}, axis = 0\right) \quad (H,) = (H,)$$

$$\boldsymbol{z}^{(2)} = \boldsymbol{a}^{(1)} \cdot W^{(2)} + \boldsymbol{b}^{(2)} \quad (N, C) = (N, H) \cdot (H, C) + (C,)$$

$$\frac{\partial L}{\partial W^{(2)}} = \boldsymbol{a}^{(1)\top} \cdot \frac{\partial L}{\partial \boldsymbol{z}^{(2)}} \quad (H, C) = (H, N) \cdot (N, C)$$

$$\frac{\partial L}{\partial \boldsymbol{a}^{(1)}} = \frac{\partial L}{\partial \boldsymbol{z}^{(2)}} \cdot W^{(2)\top} \quad (N, H) = (N, C) \cdot (C, H)$$