$$loss = \frac{1}{N}\sum_{i=0}^{N-1} loss_i$$

$$loss_i = -\log(prob_{iy_i})$$

其中 $prob_{iy_i} = \dfrac{e^{S_{iy_i}}}{\sum_{j=0}^{C-1} e^{S_{ij}}}$

考虑到指数可能很大，而

$$prob_{iy_i} = \frac{e^{S_{iy_i}}}{\sum_{j=0}^{C-1} e^{S_{ij}}} = \frac{k e^{S_{iy_i}}}{k\sum_{j=0}^{C-1} e^{S_{ij}}} = \frac{e^{S_{iy_i}+\log k}}{\sum_{j=0}^{C-1} e^{S_{ij}+\log k}}$$

这里取 $\log k = -\max(S_{ik}) \quad (k=0,1,\cdots C-1)$

$$S_{ij} = X_{i,:} W_{:,j}$$

$$S_{iy_i} = X_{i,:} W_{:,y_i}$$

$$\frac{\partial loss}{\partial W_{:,j}} = \frac{\partial loss}{\partial loss_i} \cdot \frac{\partial loss_i}{\partial prob_{iy_i}} \cdot \frac{\partial prob_{iy_i}}{\partial S_{ij}} \cdot \frac{\partial S_{ij}}{\partial W_{:,j}}$$

$$= 1 \cdot \left(-\frac{1}{prob}\right) \cdot \frac{\partial prob_{iy_i}}{\partial S_{ij}} \cdot \left(X_i^T\right)$$

$$\frac{\partial\left(\frac{e^{S_{iy_i}}}{\sum_{j=0}^{C-1} e^{S_{ij}}}\right)}{\partial S_{ij}}$$

当 $j \neq y_i$，有

$$原式 = \frac{\partial prob_{iy_i}}{\partial\left(\sum_{j=0}^{C-1} e^{S_{ij}}\right)} \cdot \frac{\partial\left(\sum_{j=0}^{C-1} e^{S_{ij}}\right)}{\partial S_{ij}}$$

$$= -\frac{e^{S_{iy_i}}}{\left(\sum_{j=0}^{C-1} e^{S_{ij}}\right)^2} \cdot \left(1 \cdot e^{S_{ij}}\right)$$

当 $j = y_i$，有

$$原式 = \frac{e^{S_{iy_i}} \cdot \sum_{j=0}^{C-1} e^{S_{ij}} - e^{S_{ij}} \cdot e^{S_{iy_i}}}{\left(\sum_{j=0}^{C-1} e^{S_{ij}}\right)^2}$$

代入 $j \neq y_i$

$$\frac{\partial loss_i}{\partial W_{:,j}} = 1 \cdot \left(-\frac{\sum_{j=0}^{C-1} e^{S_{ij}}}{e^{S_{iy_i}}}\right) \cdot \left(-\frac{e^{S_{iy_i}} e^{S_{ij}}}{\left(\sum_{j=0}^{C-1} e^{S_{ij}}\right)^2}\right) \cdot X_i^T$$

$$= prob_{ij} \cdot X_i^T$$

代入 $j = y_i$

$$\frac{\partial loss_i}{\partial W_{:,j}} = 1 \cdot \left(-\frac{\sum_{j=0}^{C-1} e^{S_{ij}}}{e^{S_{iy_i}}}\right)\left(\frac{e^{S_{iy_i}}}{\sum_{j=0}^{C-1} e^{S_{ij}}} - \frac{e^{S_{iy_i}} e^{S_{ij}}}{\left(\sum_{j=0}^{C-1} e^{S_{ij}}\right)^2}\right) \cdot X_i^T$$

$$= (prob_{ij} - 1) X_i^T$$

$$\therefore \frac{\partial loss}{\partial W} = prob \cdot X^T$$

成

$$\frac{\partial\, loss_i}{\partial W_{:,j}} = \frac{\partial\left(-\log \frac{e^{S_{iy_i}}}{\sum_{j=0}^{C-1} e^{S_{ij}}}\right)}{\partial W_{:,j}}$$

$$= \frac{\partial\left(\log \sum_{j=0}^{C-1} e^{S_{ij}} - \log e^{S_{iy_i}}\right)}{\partial W_{:,j}}$$

对 $W_{:,j}$ 来说，只有 $S_{ij} = X_i W_{:,j}$ 出现

$$\therefore \frac{\partial\, loss_i}{\partial W_{:,j}} = \frac{\partial\, loss_i}{\partial S_{ij}} \cdot \frac{\partial S_{ij}}{\partial W_{:,j}}$$

$$= \left(\frac{1}{\sum_{j=0}^{C-1} e^{S_{ij}}} \cdot e^{S_{ij}} - \frac{\partial \log e^{S_{iy_i}}}{\partial S_{ij}}\right)\left(X_i^T\right)$$

$j = y_i$，有 $\frac{\partial \log e^{S_{iy_i}}}{\partial S_{ij}} = 1$

$j \neq y_i$，有 $\frac{\partial \log e^{S_{iy_i}}}{\partial S_{ij}} = 0$

令 $\text{probility} = \frac{e^{s_{ij}}}{\sum_{j=0}^{c-1} e^{s_{ij}}} = \frac{e^{s_{ij}+\log K}}{\sum_{j=0}^{c-1} e^{s_{ij}+\log K}}$

∴ 综上

$j \neq y_i$, $\frac{\partial \text{loss}_i}{\partial w_{:,j}} = \text{prob}_{ij} \cdot X^T$

$j = y_i$, $\frac{\partial \text{loss}_i}{\partial w_{:,j}} = (\text{prob}_{ij} - 1) \cdot X^T$