# Part One: Back-propogation

### Minghao Zhang
[mehooz.weebly.com](mehooz.weebly.com)

## Block One: Gradients of some basic layers

**(i)**

According to the algorithm, we have
$$\frac{\partial y_i}{\partial \beta} = 1, \; \frac{\partial y_i}{\partial \gamma} = \hat{x}_i$$

**(ii)**

As when $i \neq j$, there is no relation originally, so $\frac{\partial y_i}{\partial x_i} = 0$, $\forall i$. Totally, we can get

$$\frac{\partial y_j}{\partial x_i} = \begin{cases} 0 & i \neq j \; or \; r_j < p \\ \frac{1}{1-p} & i = j \; and \; r_j \geq p \end{cases}$$

**(iii)**

According to the definition, let $\sigma(z)$ denotes the $softmax$ function, then

$$\frac{\partial \sigma(z)_j}{\partial z_i} = \begin{cases} -\sigma(z)_i \sigma(z)_j & i \neq j \\ \sigma(z)_i(1 - \sigma(z)_i) & i = j \end{cases}$$

Specifically, we have

$$\frac{\partial \sigma(z)_j}{\partial z_i} = \begin{cases} -\dfrac{e^{z_i + z_j}}{(\sum\limits_1^k e^{z_j})^2} & i \neq j \\[4mm] \dfrac{e^{z_i}(\sum\limits_{j \neq i} e^{z_j})}{(\sum\limits_1^k e^{z_j})^2} & i = j \end{cases}$$

## Block Two: Feed-forward and back-propagation of the multi-task network

All notations are consistent with the Figure 3 in homework instruction.

One page for one subsection in the following.

**(i)**

$$z_{FC_{1a}} = \theta_{1a}x + b_{1a}$$

$$a_{FC_{1a}} = ReLU(z_{FC_{1a}})$$

$$a_{DP_{1a}} = a_{FC_{1a}} \odot M$$

$$\hat{y}_a = \theta_{2a}a_{DP_{1a}} + b_{2a}$$

$$z_{FC_{1b}} = \theta_{1b}x + b_{1b}$$

$$a_{FC_{1b}} = ReLU(z_{FC_{1b}})$$

$$a_{BN_{1b}} = BN_{\gamma,\beta}(z_{BN_{1b}})$$

$$z_{FC_{2b}} = \theta_{2b}(a_{FC_{2a}} \oplus (a_{BN_{1b}})) + b_{2b}$$

$$\hat{y}_b = softmax(z_{FC_{2b}})$$

$$L(x, y_a, y_b; \theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\|y_{ai} - \hat{y}_{ai}\|_2^2 - \sum_{k=1}^{n_{yb}}y_{bi}^k \log(\hat{y}_{bi}^k)\right)$$

**(ii)**

$$\frac{\partial L}{\partial z_{FC_{2b}}} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_b^{(i)} - y_b^{(i)}), \ got \ residual \ \delta^{(FC_{2b})}$$

$$\frac{\partial L}{\partial \theta_{2b}} = \delta^{(EC_{2b})}(\hat{y}_a \oplus a_{BN_{1b}})^T$$

$$\frac{\partial L}{a_{BN_{1b}}} = \delta^{(FC_{2b})}\theta_{2b}^T, \ got \ residual \ \delta^{(BN)_{1b}}$$

$$\frac{\partial L}{\partial \gamma} = \delta^{(BN_{1b})}\hat{a}_{FC_{1b}}^T, \ \ \frac{\partial L}{\partial \beta} = \sum_{i=1}^{n_{ya}}\delta^{(BN_{1b})}$$

$$\frac{\partial a_{BN_{1b}j}}{\partial a_{FC_{1b}i}} = \begin{cases} \gamma(\sigma_b^2 + \varepsilon)^{-\frac{3}{2}}(\frac{m-1}{m}(\sigma_b^2 + \varepsilon) - \frac{1}{m}(a_{FC_{1b}j} - \sigma_B)(a_{FC_{1b}i} - \sigma_B)) \ i = j \\ \gamma(\sigma_b^2 + \varepsilon)^{-\frac{3}{2}}(-\frac{1}{m}(\sigma_b^2 + \varepsilon) - \frac{1}{2}(a_{FC_{1b}j} - \sigma_B)(a_{FC_{1b}i} - \sigma_B)) \ i \neq j \end{cases}$$

$$\frac{\partial L}{\partial a_{FC_{1b}i}} = \sum_{j=1}^{m}\delta^{(BN_{1b})}\frac{\partial a_{BN_{1b}j}}{\partial a_{FC_{1b}i}}$$

$$\frac{\partial a_{FC_{1b}i}}{\partial z_{FC_{1b}i}} = sgn(z_{FC_{1b}i})$$

$$\frac{\partial L}{\partial z_{FC_{1b}i}} = \frac{\partial L}{\partial a_{FC_{1b}i}} \odot sgn(z_{FC_{1b}i}), \ got \ residual \ \delta^{(FC_{1b})}$$

$$\frac{\partial L}{\theta_{1b}} = \delta^{(FC_{1b})}x^T$$

$$\frac{\partial L}{\partial \hat{y}_a} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_a^{(i)} - y_a^{(i)}) + \delta^{(FC_{2b})}\theta_{2b}^T, \ got \ residual \ \delta^{(ya)}$$

$$\frac{\partial L}{\partial \theta_{2a}} = \delta^{(ya)}a_{DP_{1a}}^T$$

$$\frac{\partial L}{\partial a_{DP_{1a}}} = \delta^{(ya)}, \ got \ residual \ \delta^{(DP)}$$

$$\frac{\partial L}{\partial a_{FC_{1a}i}} = \begin{cases} \frac{1}{1-p}\delta_i^{(DP)} \ r_i \geq p \\ 0 \ r_i < p \end{cases}$$

$$\frac{\partial L}{\partial z_{FC_{1a}i}} = \begin{cases} \frac{1}{1-p}\delta_i^{(DP)}sgn(z_{FC_{1a}i}) \ r_i \geq p \\ 0 \ r_i < p \end{cases}$$

$$got \ residual \ \delta^{(FC_{1a})}$$

$$\frac{\partial L}{\partial \theta_{1a}} = \delta^{(FC_{1a})}x^T$$