UNIVERSITY OF TRENTO, ITALY

Training of Neural Networks with Few Labels

Machine Learning Assignment

Subhankar Roy

(Matricola no.: 181518)

Backpropagation

Training a neural network means finding the set of weights and biases that minimizes the cost function. The optimization method, viz.- gradient descent, tries to reach a global minima of the cost function by updating its parameters (W and b) depending on learning rate η and gradient of the cost function w.r.t. all the parameters. Calculation of the partial derivatives w.r.t all the parameters of the network is computationally demanding. Backpropagation algorithm finds the partial derivates in a very fast way. This is what makes learning very fast.

1) Binary classification

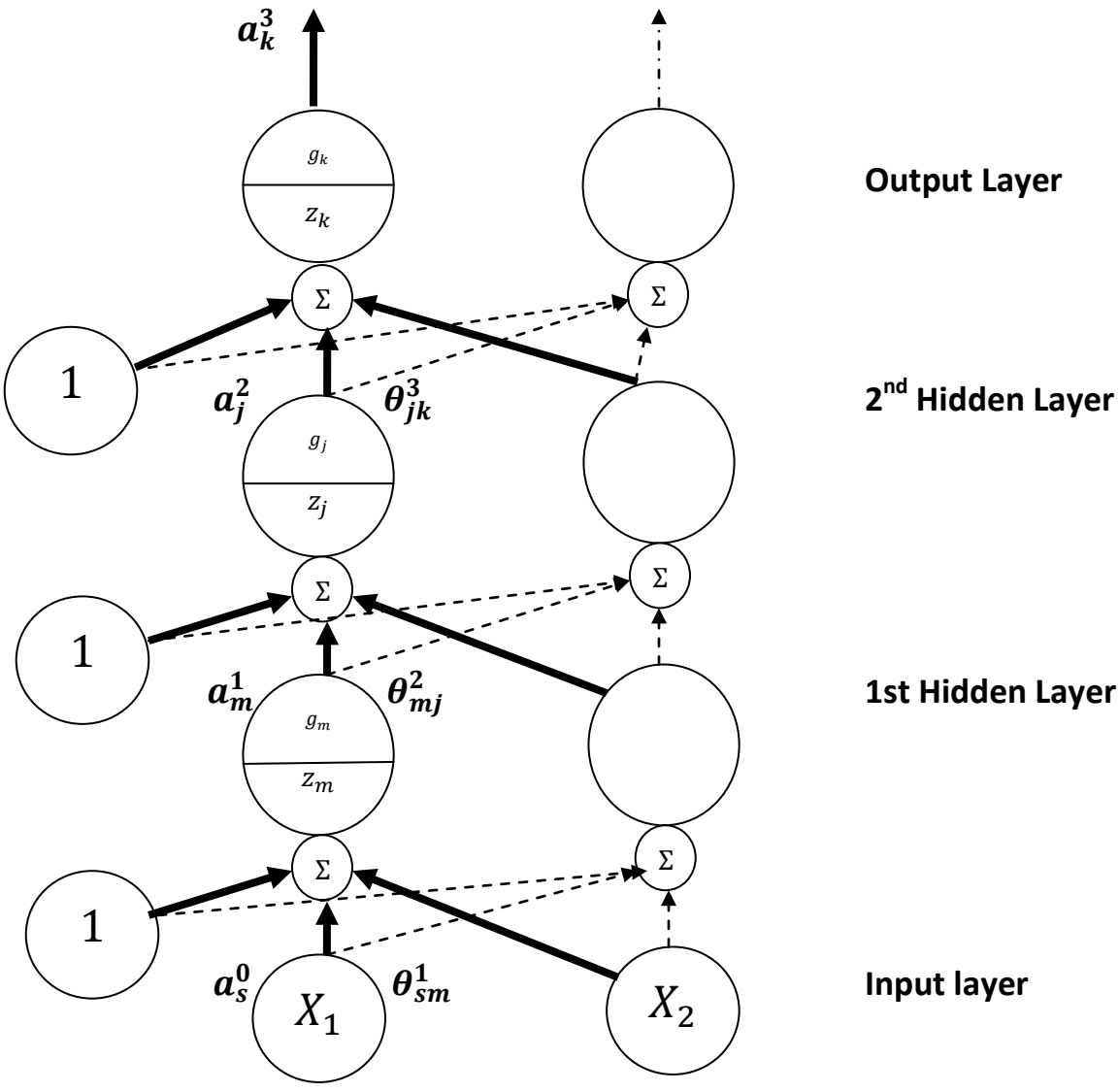


Fig 1. Feed forward Neural Network

Where,

a_k^3 = Output/activation received from the k^{th} node of the output layer.

a_j^2 = Activation from the j^{th} node of the 2^{nd} hidden layer.

a_m^1 = Activation from the m^{th} node of the 1^{st} hidden layer.

θ_{jk}^3 = Weight connecting the j^{th} node of the 2^{nd} hidden layer with k^{th} node of the output layer.

θ_{mj}^2 = Weight connecting the m^{th} node of the 1^{st} hidden layer with j^{th} node of the 2^{nd} hidden layer.

θ_{pm}^1 = Weight connecting the p^{th} node of the input layer with m^{th} node of the 1^{st} hidden layer.

g_k = Activation function of the output layer.

z_k = Sum of inputs of from all the neurons of the 2^{nd} hidden layer to the k^{th} node of the output layer.

The other notations have their usual significance.

The final formulation of the PU learning problem is given as:

$$\mathcal{J}(\theta) = \sum_{k=1}^{2}\left\{-\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k} f_k(x_i,\theta) + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k} \max\left\{f_k(x_i,\theta), \max\left\{0, \frac{1}{2} + \frac{f_k(x_i,\theta)}{2}\right\}\right\}\right\} + \lambda\sum_{\theta}\theta^2$$

$f_k(x_i,\theta) = a_{k,i}^3$

<u>Gradient calculation for the output layer</u>

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{jk}^3} = -\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k}\frac{\partial a_{k,i}^3}{\partial \theta_{jk}^3} + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k}\frac{\partial}{\partial \theta_{jk}^3}\max\left\{a_{k,i}^3, \max\left\{0, \frac{1}{2} + \frac{a_{k,i}^3}{2}\right\}\right\} + \lambda\sum_{\omega}\frac{\partial \theta^2}{\partial \theta_{jk}^3} \quad (1.1)$$

Lets calculate the term $\frac{\partial a_k^3}{\partial \theta_{jk}^3}$

$$\frac{\partial a_k^3}{\partial \theta_{jk}^3} = \frac{\partial g_k(z_k)}{\partial \theta_{jk}^3}$$

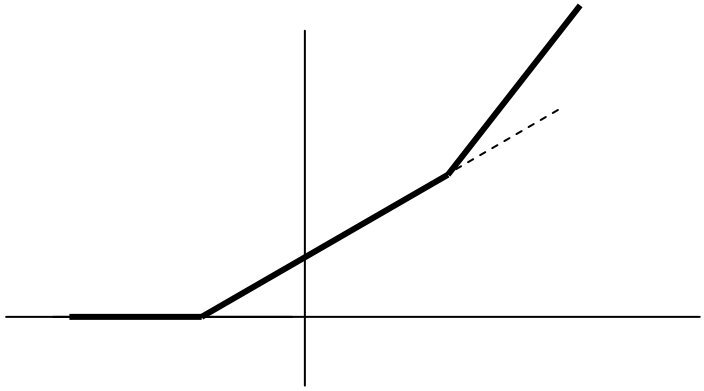
$$= g_k'(z_k).\frac{\partial z_k}{\partial \theta_{jk}^3}$$

$$= g'_k(z_k).\frac{\partial}{\partial \theta^3_{jk}}\left(\sum_j \theta^3_{jk}\, a^2_j\right)$$

$$\therefore \qquad \frac{\partial a^3_k}{\partial \theta^3_{jk}} = g'_k(z_k).a^2_j \tag{1.2}$$

Now, lets calculate the term $\frac{\partial}{\partial \theta^3_{jk}} \max\left\{a^3_k, \max\left\{0, \frac{1}{2} + \frac{a^3_k}{2}\right\}\right\}$

Before computing the term, let's have a look at the double hinge loss function.



We have kinks at -1 and 1.

$$\frac{\partial}{\partial \theta^3_{jk}} \max\left\{a^3_k, \max\left\{0, \frac{1}{2} + \frac{a^3_k}{2}\right\}\right\} = \frac{\partial}{\partial a^3_k} \max\left\{a^3_k, \max\left\{0, \frac{1}{2} + \frac{a^3_k}{2}\right\}\right\}.\frac{\partial a^3_k}{\partial \theta^3_{jk}}$$

$$\delta^3_{k,i} = \begin{cases} g'_k(z_k)\ ;\ \text{for}\ \ a^3_{k,i}\ \geq 1 \\ \frac{1}{2}\, g'_k(z_k)\ ;\ \text{for}\ 1 > a^3_{k,i}\ \geq\ -1 \\ 0\ ;\ \text{otherwise} \end{cases} \tag{1.3}$$

The derivative of the regularization parameter:

$$\lambda \sum_\omega \frac{\partial \theta^2}{\partial \theta^3_{jk}} = 2\lambda \sum_j \theta^3_{jk} \tag{1.4}$$

Now plugging all the values obtained in equations (1.2), (1.3) and (1.4) into equation (1.1) we get,

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{jk}^3} = -\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k} \delta_{k,i}^3 \, a_{j,i}^2 + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k} \delta_{k,i}^3 a_{j,i}^2 + 2\lambda \sum_j \theta_{jk}^3 \tag{1.5}$$

Calculation of gradient for the 2nd hidden layer

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{mj}^2} = \sum_{k=1}^{2}\left\{-\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k} \frac{\partial a_{k,i}^3}{\partial \theta_{mj}^2} + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k} \frac{\partial}{\partial \theta_{mj}^2} \max\left\{a_{k,i}^3, \max\left\{0, \frac{1}{2} + \frac{a_{k,i}^3}{2}\right\}\right\}\right\} + \lambda \sum_{\omega} \frac{\partial}{\partial \theta_{mj}^2} \omega^2 \tag{1.6}$$

Let's calculate $\frac{\partial a_k^3}{\partial \theta_{mj}^2}$

$$\frac{\partial a_k^3}{\partial \theta_{mj}^2} = \frac{\partial g_k(z_k)}{\partial \theta_{mj}^2}$$

$$= g_k'(z_k).\frac{\partial}{\partial a_j^2}\left(\sum_j \theta_{jk}^3 \, a_j^2\right).\frac{\partial g_j(z_j)}{\partial \theta_{mj}^2}$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).\frac{\partial z_j}{\partial \theta_{mj}^2}$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).\frac{\partial}{\partial \theta_{mj}^2}\left(\sum_m \theta_{mj}^2 \, a_m^1\right)$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).a_m^1$$

$$= \delta_k^3.\theta_{jk}^3.g_j'(z_j).a_m^1 \qquad \text{From equation (1.3)}$$

$$= \delta_j^2.a_m^1 \tag{1.7}$$

Similarly, like the output layer, we can write the final gradient as:

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{mj}^2} = \sum_{k=1}^{2}\left\{-\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k} \delta_{j,i}^2.a_{m,i}^1 + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k} \delta_{j,i}^2.a_{m,i}^1\right\} + 2\lambda \sum_m \theta_{mj}^2 \tag{1.8}$$

Calculation of gradient for the 2nd hidden layer

$$\frac{\partial J(\theta)}{\partial \theta_{sm}^1} = \sum_{k=1}^{2}\left\{-\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k}\frac{\partial a_{k,i}^3}{\partial \theta_{sm}^1} + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k}\frac{\partial}{\partial \theta_{sm}^1}\max\left\{a_{k,i}^3, \max\left\{0, \frac{1}{2}+\frac{a_{k,i}^3}{2}\right\}\right\}\right\} + \lambda\sum_{\omega}\frac{\partial}{\partial \theta_{sm}^1}\omega^2$$

We need to calculate $\frac{\partial a_k^3}{\partial \theta_{sm}^1}$

$$\frac{\partial a_k^3}{\partial \theta_{sm}^1} = g_k'(z_k)\frac{\partial z_k}{\partial \theta_{sm}^1}$$

$$= g_k'(z_k).\frac{\partial z_k}{\partial a_j^2}.\frac{\partial a_j^2}{\partial \theta_{sm}^1}$$

$$= g_k'(z_k).\frac{\partial}{\partial a_j^2}\left(\sum_j \theta_{jk}^3\, a_j^2\right).\frac{\partial g_j(z_j)}{\partial \theta_{sm}^1}$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).\frac{\partial z_j}{\partial \theta_{sm}^1}$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).\frac{\partial z_j}{\partial a_m^1}.\frac{\partial a_m^1}{\partial \theta_{sm}^1}$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).\frac{\partial}{\partial a_m^1}\left(\sum_m \theta_{mj}^2\, a_m^1\right)\frac{\partial g_m(z_m)}{\partial \theta_{sm}^1}$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).\theta_{mj}^2.g_m'(z_m).\frac{\partial z_m}{\partial \theta_{sm}^1}$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).\theta_{mj}^2.g_m'(z_m).\frac{\partial}{\partial \theta_{sm}^1}\left(\sum_s \theta_{sm}^1\, x_s\right)$$

$$= g_k'(z_k).\theta_{jk}^3.g_j'(z_j).\theta_{mj}^2.g_m'(z_m).x_s$$

$= \delta_k^3.\theta_{jk}^3.g_j'(z_j).\theta_{mj}^2.g_m'(z_m).x_s$ From equation (1.3)

$= \delta_j^2\theta_{mj}^2.g_m'(z_m).x_s$ From equation (1.7)

$\therefore \frac{\partial a_k^3}{\partial \theta_{sm}^1} = \delta_m^1.x_s$ (1.9)

Hence, we can write the derivative of the cost function as:

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{mj}^{2}} = \sum_{k=1}^{2} \left\{ -\frac{\pi_k}{|D_p^k|} \sum_{x_i \epsilon D_p^k} \delta_{m,i}^{1} . x_{s,i} + \frac{1}{|D_n^k|} \sum_{x_i \epsilon D_n^k} \delta_{m,i}^{1} . x_{s,i} \right\} + 2\lambda \sum_{s} \theta_{sm}^{1} \tag{1.10}$$

2) Multiclass classification

In this problem I have considered that this feed-forward network has 1.) **L** layers (inclusive of input and output layer) 2.) **K** output neurons 3.) $\boldsymbol{\ell}$ hidden layers.

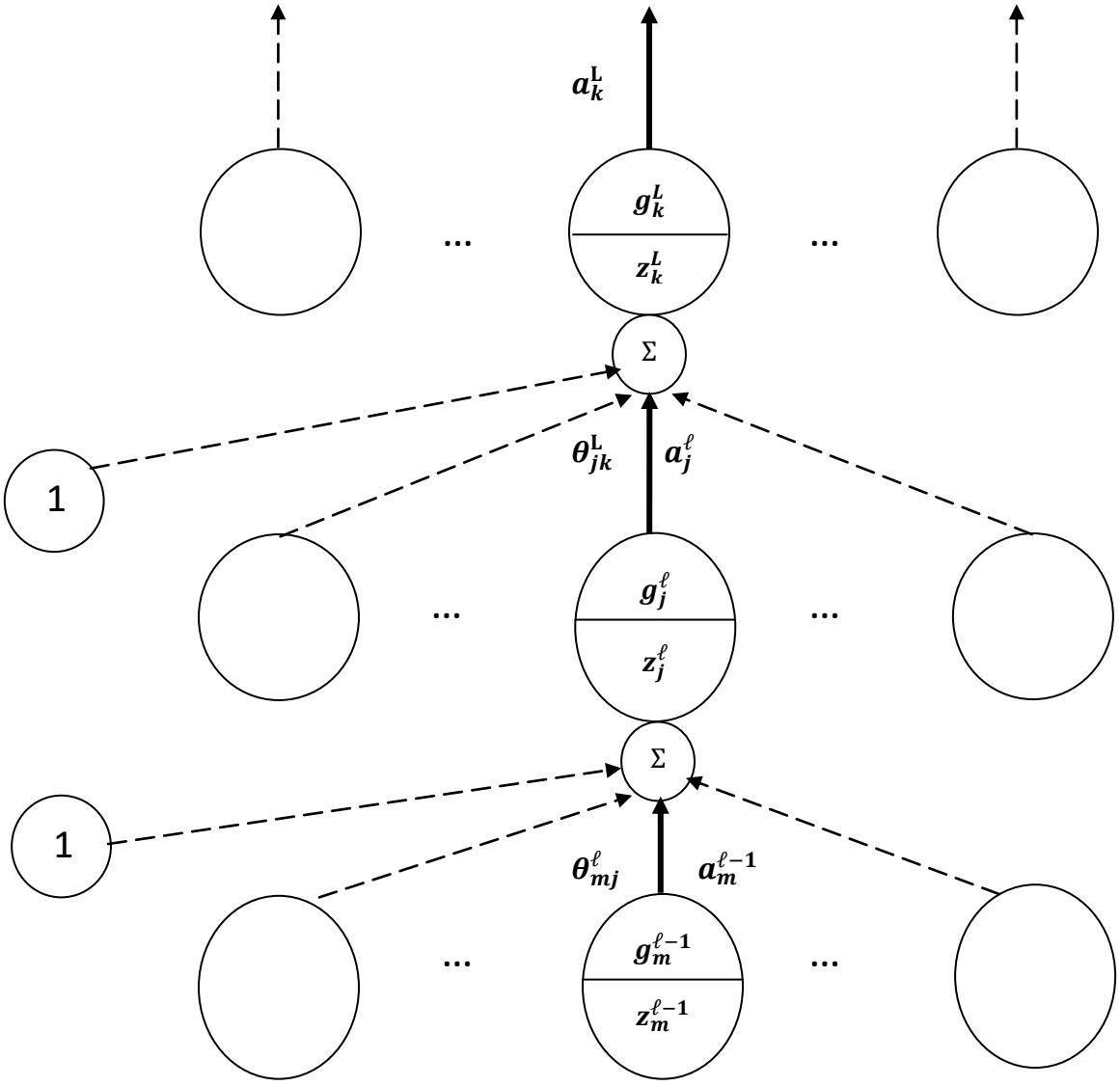


Fig 2. Feed-forward neural network with L layers and K outputs

Where,

a_k^{L} = Output/activation received from the k^{th} node of the output layer.

a_j^{ℓ} = Activation from the j^{th} node of the ℓ^{th} hidden layer. Here ℓ is the ultimate hidden layer.

$a_m^{\ell-1}$ = Activation from the m^{th} node of the penultimate hidden layer.

θ_{jk}^{L} = Weight connecting the j^{th} node of the ℓ^{th} hidden layer with k^{th} node of the output layer.

θ_{mj}^{ℓ} = Weight connecting the m^{th} node of the $(\ell-1)^{th}$ hidden layer with j^{th} node of the ℓ^{th} hidden layer.

g_k^{L} = Activation function of the k^{th} node of the output layer.

z_k^{L} = Sum of inputs of from all the neurons of the ℓ^{th} hidden layer to the k^{th} node of the output layer.

Likewise, the other notations have their usual significance.

The cost function to be minimized:

$$\mathcal{J}(\theta) = \sum_{k=1}^{\mathrm{K}} \left\{ -\frac{\pi_k}{|D_p^k|} \sum_{x_i \epsilon D_p^k} f_k(x_i, \theta) + \frac{1}{|D_n^k|} \sum_{x_i \epsilon D_n^k} \max\left\{ f_k(x_i, \theta), \max\left\{ 0, \frac{1}{2} + \frac{f_k(x_i, \theta)}{2} \right\} \right\} \right\} + \lambda \sum_{\theta} \theta^2 \qquad (2.1)$$

$f_k(x_i, \theta) = a_{k,i}^{\mathrm{L}}$

<u>Gradient calculation for the output layer</u>

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{jk}^{\mathrm{L}}} = -\frac{\pi_k}{|D_p^k|} \sum_{x_i \epsilon D_p^k} \frac{\partial a_{k,i}^{\mathrm{L}}}{\partial \theta_{jk}^{\mathrm{L}}} + \frac{1}{|D_n^k|} \sum_{x_i \epsilon D_n^k} \frac{\partial}{\partial \theta_{jk}^{\mathrm{L}}} \max\left\{ a_{k,i}^{\mathrm{L}}, \max\left\{ 0, \frac{1}{2} + \frac{a_{k,i}^{\mathrm{L}}}{2} \right\} \right\} + \lambda \sum_{\omega} \frac{\partial \theta^2}{\partial \theta_{jk}^{\mathrm{L}}} \qquad (2.2)$$

Let's calculate $\frac{\partial a_k^{\mathrm{L}}}{\partial \theta_{jk}^{\mathrm{L}}}$

$$\frac{\partial a_k^{\mathrm{L}}}{\partial \theta_{jk}^{\mathrm{L}}} = \frac{\partial g_k^{\mathrm{L}}(z_k^{\mathrm{L}})}{\partial \theta_{jk}^{\mathrm{L}}}$$

$$= g_k^{L'}(z_k^L).\frac{\partial z_k^L}{\partial \theta_{jk}^L}$$

$$= g_k^{L'}(z_k^L).\frac{\partial}{\partial \theta_{jk}^L}\left(\sum_j \theta_{jk}^L a_j^{\ell}\right)$$

$$= g_k^{L'}(z_k^L).a_j^{\ell}$$

$$\therefore \quad \frac{\partial a_k^L}{\partial \theta_{jk}^L} = g_k^{L'}(z_k^L).a_j^{\ell} \tag{2.3}$$

Now let's calculate : $\frac{\partial}{\partial \theta_{jk}^L}\max\left\{a_k^L, \max\left\{0, \frac{1}{2}+\frac{a_k^L}{2}\right\}\right\}$

$$\frac{\partial}{\partial \theta_{jk}^L}\max\left\{a_k^L, \max\left\{0, \frac{1}{2}+\frac{a_k^L}{2}\right\}\right\} = \frac{\partial}{\partial a_k^L}\max\left\{a_k^L, \max\left\{0, \frac{1}{2}+\frac{a_k^L}{2}\right\}\right\}.\frac{\partial a_k^L}{\partial \theta_{jk}^L}$$

Following the graph of loss function from the section 1, we can write:

$$\delta_{k,i}^L = \begin{cases} g_k^{L'}(z_k^L)\ ; \text{ for } a_{k,i}^L \geq 1 \\ \frac{1}{2}g_k^{L'}(z_k^L)\ ; \text{ for } 1 > a_{k,i}^L \geq -1 \\ 0\ ; \text{otherwise} \end{cases} \tag{2.4}$$

The derivative of the regularization parameter:

$$\lambda \sum_{\omega} \frac{\partial \theta^2}{\partial \theta_{jk}^L} = 2\lambda \sum_j \theta_{jk}^L \tag{2.5}$$

Now plugging all the values obtained in equations (2.3), (2.4) and (2.5) into equation (2.2) we get,

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{jk}^L} = -\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k} \delta_{k,i}^L a_{j,i}^{\ell} + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k} \delta_{k,i}^L a_{j,i}^{\ell} + 2\lambda \sum_j \theta_{jk}^L \tag{2.6}$$

Gradient calculation for the last hidden layer

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{mj}^{\ell}} = \frac{\partial}{\partial \theta_{mj}^{\ell}}\left[\sum_{k=1}^{\mathbf{K}}\left\{-\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k} a_{k,i}^{\mathrm{L}} + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k} \max\left\{a_{k,i}^{\mathrm{L}}, \max\left\{0, \frac{1}{2} + \frac{a_{k,i}^{\mathrm{L}}}{2}\right\}\right\}\right\} + \lambda \sum_{\theta} \theta^2\right]$$

$$= \sum_{k=1}^{\mathbf{K}}\left\{-\frac{\pi_k}{|D_p^k|}\sum_{x_i \epsilon D_p^k} \frac{\partial a_{k,i}^{\mathrm{L}}}{\partial \theta_{mj}^{\ell}} + \frac{1}{|D_n^k|}\sum_{x_i \epsilon D_n^k} \frac{\partial}{\partial \theta_{mj}^{\ell}} \max\left\{a_{k,i}^{\mathrm{L}}, \max\left\{0, \frac{1}{2} + \frac{a_{k,i}^{\mathrm{L}}}{2}\right\}\right\}\right\} + \lambda \sum_{\theta} \frac{\partial \theta^2}{\partial \theta_{mj}^{\ell}} \qquad (2.7)$$

Let's calculate $\frac{\partial a_k^{\mathrm{L}}}{\partial \theta_{mj}^{\ell}}$

$$\frac{\partial a_k^{\mathrm{L}}}{\partial \theta_{mj}^{\ell}} = \frac{\partial g_k^{\mathrm{L}}(z_k^{\mathrm{L}})}{\partial \theta_{mj}^{\ell}}$$

$$= g_k^{\mathrm{L}'}(z_k^{\mathrm{L}}).\frac{\partial z_k^{\mathrm{L}}}{\partial \theta_{mj}^{\ell}}$$

$$= g_k^{\mathrm{L}'}(z_k^{\mathrm{L}}).\frac{\partial}{\partial \theta_{mj}^{\ell}}\left(\sum_j \theta_{jk}^{\mathrm{L}} a_j^{\ell}\right)$$

$$= g_k^{\mathrm{L}'}(z_k^{\mathrm{L}}).\frac{\partial}{\partial a_j^{\ell}}\left(\sum_j \theta_{jk}^{\mathrm{L}} a_j^{\ell}\right).\frac{\partial a_j^{\ell}}{\partial \theta_{mj}^{\ell}}$$

$$= g_k^{\mathrm{L}'}(z_k^{\mathrm{L}}).\theta_{jk}^{\mathrm{L}}.\frac{\partial g_j^{\ell}(z_j^{\ell})}{\partial \theta_{mj}^{\ell}}$$

$$= g_k^{\mathrm{L}'}(z_k^{\mathrm{L}}).\theta_{jk}^{\mathrm{L}}.g_j^{\ell'}(z_j^{\ell}).\frac{\partial z_j^{\ell}}{\partial \theta_{mj}^{\ell}}$$

$$= g_k^{\mathrm{L}'}(z_k^{\mathrm{L}}).\theta_{jk}^{\mathrm{L}}.g_j^{\ell'}(z_j^{\ell}).\frac{\partial}{\partial \theta_{mj}^{\ell}}\left(\sum_m \theta_{mj}^{\ell} a_m^{\ell-1}\right)$$

$$= g_k^{\mathrm{L}'}(z_k^{\mathrm{L}}).\theta_{jk}^{\mathrm{L}}.g_j^{\ell'}(z_j^{\ell}).a_m^{\ell-1}$$

$$= \delta_k^{\mathrm{L}}.\theta_{jk}^{\mathrm{L}}.g_j^{\ell'}(z_j^{\ell}).a_m^{\ell-1}$$

$$\therefore \qquad \frac{\partial a_k^{\mathrm{L}}}{\partial \theta_{mj}^{\ell}} = \delta_j^{\ell} . a_m^{\ell-1} \tag{2.8}$$

The value of $\frac{\partial}{\partial \theta_{mj}^{\ell}} \max\left\{\mathrm{a}_{\mathrm{k}}^{\mathrm{L}}, \max\left\{0, \frac{1}{2} + \frac{\mathrm{a}_{\mathrm{k}}^{\mathrm{L}}}{2}\right\}\right\}$ can be calculated by using the value of the above formulation.

Thus, substituting all the obtained values in equation (2.7), we get

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta_{mj}^{\ell}} = \sum_{k=1}^{\mathbf{K}} \left\{ -\frac{\pi_k}{\left|D_p^k\right|} \sum_{x_i \epsilon D_p^k} \delta_{j,i}^{\ell} . a_{m,i}^{\ell-1} + \frac{1}{\left|D_n^k\right|} \sum_{x_i \epsilon D_n^k} \delta_{j,i}^{\ell} . a_{m,i}^{\ell-1} \right\} + 2\lambda \sum_{m} \theta_{mj}^{\ell} \tag{2.9}$$

To generalize, we can say that in order to calculate gradients with respect to the weights at any layer ℓ in a network with arbitrary hidden layers we just need to calculate the backpropagated error signal that trickles to that particular layer δ^{ℓ} and multiply it with the feed-forward activation $a^{\ell-1}$ feeding into that layer.