Starting with the loss function:

$$\tilde{J}(\theta) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} -\log\left[\frac{e^{z^{(3)}_{y_i}}}{\sum e^{z^{(3)}_{i,j}}}\right]}_{J_{data}} + \underbrace{\lambda\left(\|W^{(1)}\|_2^2 + \|W^{(2)}\|_2^2\right)}_{J_{reg}}$$

Let $\psi(z^{(3)})$ be the probability matrix where each element $\psi(z^{(3)})_{ij}$ is the softmax probability for class j on sample i:

$$\psi(z^{(3)})_{ij} = \frac{\exp(z^{(3)}_{ij})}{\sum_k \exp(z^{(3)}_{ik})}$$

The loss for each sample i can be expressed as $-\log(\psi(z^{(3)}_{i,y_i}))$ where y_i is the correct class for sample i.

Finding the gradient: (for just

$$\frac{\partial(J(\theta))}{\partial z^{(3)}_{ij}} = \partial$$

predicted probability minus actual probability

$$\frac{\partial(-\log(\psi(z^{(3)})_{i,y_i})}{\partial(z^{(3)}_{ij})} = \psi(z^{(3)})_{ij} - 1 \quad \text{if } j = y_i$$

and

$$= \psi(z^{(3)})_{ij} - 0 \quad \text{if } j \neq y_i$$

$$\text{so } \frac{\partial(J_{data})}{\partial z^{(3)}} = \frac{1}{N}\left(\psi(z^{(3)}) - \Delta\right)$$ (putting results in matrix form)

since $\partial(J_{reg})/\partial(z^{(3)}) = 0$,

$$\frac{\partial J(\theta)}{\partial z^{(3)}} = \frac{1}{N}\left(\psi(z^{(3)}) - \Delta\right)$$

Starting with normal loss $J(\theta) = \frac{1}{N}\sum_{i=1}^{N} -\log\left[\frac{\exp(z_{y_i}^{(3)})}{\sum \exp(z_j^{(2)})}\right]$

$$\frac{\partial(J(\theta))}{\partial w^{(2)}} = \frac{\partial J}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial (w^{(2)})}$$

$$\frac{\partial z^{(3)}}{\partial (w^{(2)})} = \frac{\partial}{\partial (w^{(2)})}\left(w^{(2)} a^{(2)} + b^{(2)}\right)$$
$$= a^{(2)}$$

so $$\frac{\partial(J(\theta))}{\partial w^{(2)}} = \frac{1}{N}\left(\psi(z^{(3)}) - \Delta\right) \underbrace{a^{(2)'}}$$

↑ transpose of $a^{(2)}$ to make mult possible

Now for $\tilde{J}(\theta) = J(\theta) + J_{reg}$

$$\frac{\partial}{\partial w^{(2)}}(J_{reg}) = \frac{\partial}{\partial w^{(2)}}\left(\lambda\left(\|w^{(1)}\|_2^2 + \|w^{(2)}\|_2^2\right)\right)$$
$$= 2\lambda w^{(2)}$$

so $$\frac{\partial \tilde{J}(\theta)}{\partial w^{(2)}} = \frac{\partial J(\theta)}{\partial w^{(2)}} + \frac{\partial J_{reg}}{\partial w^{(2)}}$$
$$= \frac{1}{N}\left(\psi(z^{(3)}) - \Delta\right) a^{(2)'} + 2\lambda w^{(2)}$$

$$\frac{\partial z^{(3)}}{\partial b^{(2)}} = \frac{\partial}{\partial b^{(2)}}\left(w^{(2)} a^{(2)} + b^{(2)}\right) = 1$$

so $$\frac{\partial \tilde{J}(\theta)}{\partial \cancel{J(\theta)}\, b^{(2)}} = \cancel{\frac{\partial \tilde{J}(\theta)}{\partial z^{(3)}}} + \frac{\partial}{\partial b^{(2)}} J(\theta) + \frac{\partial}{\partial b^{(2)}} J_{reg}$$
$$= \frac{\partial J(\theta)}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial b^{(2)}} + \frac{\partial}{\partial b^{(2)}}\left(\lambda\left(\|w^{(1)}\|_2^2 + \|w^{(2)}\|_2^2\right)\right)$$
$$= \frac{1}{N}\left(\psi(z^{(3)}) - \Delta\right)(1) + 0$$
$$= \frac{1}{N}\left(\psi(z^{(3)}) - \Delta\right)$$

$$\frac{\partial \tilde{J}(\theta)}{\partial a^{(2)}} = \frac{\partial J(\theta)}{\partial a^{(2)}} + \frac{\partial J_{reg}}{\partial a^{(2)}}$$

$$= \frac{\partial J(\theta)}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(2)}} + \frac{\partial}{\partial (a^{(2)})} \left(\lambda(\|w^{(1)}\|_2^2 + \|w^{(2)}\|_2^2)\right)$$

$$= \frac{1}{N} \left(\varphi(z^{(3)}) - \Delta\right) (w^{(2)})' + 0$$

$$= \frac{1}{N} \left(\varphi(z^{(3)}) - \Delta\right) \cdot w^{(2)'}$$

$$+ 0 \quad \left(\frac{\partial J_{reg}}{\partial z^{(2)}} = 0\right)$$

$$\frac{\partial \tilde{J}(\theta)}{\partial z^{(2)}} = \frac{\partial \tilde{J}(\theta)}{\partial a^{(2)}} \times \frac{\partial a^{(2)}}{\partial z^{(2)}}$$, since $a^{(2)} = \varphi(z^{(2)})$,

$$\frac{\partial a^{(2)}}{\partial z^{(2)}} = \begin{cases} 1 & \text{if } z^{(2)} > 0 \\ 0 & \text{if } z^{(2)} \leq 0 \end{cases}$$, so

$$\frac{\partial \tilde{J}(\theta)}{\partial z^{(2)}} = \begin{cases} \frac{1}{N} \left(\varphi(z^{(3)}) - \Delta\right) \cdot w^{(2)'} & \text{if } z^{(2)} > 0 \\ 0 & \text{if } z^{(2)} \leq 1 \end{cases}$$

$$z^{(2)} = w^{(1)} a^{(1)} + b^{(1)}$$

$$\frac{\partial z^{(2)}}{\partial w^{(1)}} = a^{(1)}$$, $$\frac{\partial z^{(2)}}{\partial b^{(1)}} = 1$$

$$\frac{\partial J_{reg}}{\partial w^{(1)}} = \frac{\partial}{\partial w^{(1)}} \left(\lambda(\|w^{(1)}\|_2^2 + \|w^{(2)}\|_2^2)\right) = 2\lambda w^{(1)}$$

$$\frac{\partial \tilde{J}(\theta)}{\partial w^{(1)}} = \frac{\partial J(\theta)}{\partial w^{(1)}} + \frac{\partial J_{reg}}{\partial w^{(1)}} \quad (\text{when } z^{(2)} > 0)$$

$$= \frac{\partial J(\theta)}{\partial z^{(2)}} \times \frac{\partial z^{(2)}}{\partial w^{(1)}} + 2\lambda w^{(1)}$$

$$= \frac{1}{N} \left(\varphi(z^{(3)}) - \Delta\right) \cdot a^{(1)'} + 2\lambda w^{(1)}$$

Similarly, (when $z^{(2)} > 0$)

$$\frac{\partial \tilde{J}(\theta)}{\partial b^{(1)}} = \frac{\partial J(\theta)}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial b^{(1)}} + \frac{\partial J_{reg}}{\partial b^{(1)}}$$

$$= \frac{1}{N}\left(\psi(z^{(2)}) - \Delta\right) \cdot (1) + 0$$

$$\frac{\partial \tilde{J}(\theta)}{\partial w^{(1)}} = \begin{cases} \frac{1}{N}\left(\psi(z^{(3)}) - \Delta\right) \cdot a^{(1)'} & \text{if } z^{(2)} > 0 \\ 0 & \text{if } z^{(2)} \leq 0 \end{cases}$$

$$\frac{\partial \tilde{J}(\theta)}{\partial b^{(1)}} = \begin{cases} \frac{1}{N}\left(\psi(z^{(3)}) - \Delta\right) & \text{if } z^{(2)} > 0 \\ 0 & \text{if } z^{(2)} < 0 \end{cases}$$