5a) $CE^i = -\sum_{k=1}^{k} \underbrace{y_k^i}_{=\delta_{kl}^i} \log \hat{y}_k^i = -\log \hat{y}_l^i \overset{\text{ignoring superscript of } l}{=} -\log\left[\frac{e^{z_l^i}}{\sum_{j=1}^{k} e^{z_j^i}}\right]$

$$\frac{\partial CE^i}{\partial z_q^i} = -\frac{1}{\hat{y}_l^i}\,\frac{+e^{z_q^i}\delta_{ql}\Sigma - e^{z_l^i}e^{z_q^i}}{[\Sigma]^2} = \frac{+1}{\hat{y}_l^i}\left(\underbrace{\delta_{ql}}_{=\delta_{ql}\hat{y}_l^i}\hat{y}_q^i + \hat{y}_l^i\hat{y}_q^i\right)$$

$$= -\delta_{ql} + \hat{y}_q^i = \underline{\underline{-y_q^i + \hat{y}_q^i}} \Rightarrow \underline{\nabla_{z^i} CE^i = \underline{\underline{-y^i + \hat{y}^i}}} \in R^4$$

5b) Need to compute the gradients wrt W^1, W^2, b^1, b^2.

$$b_a^2: \underline{\frac{\partial CE^i}{\partial b_a^2}} = \frac{\partial CE^i}{\partial z_q^i}\frac{\partial z_q^i}{\partial b_a^2} = (\hat{y}_q^i - y_q^i)\cdot\delta_{aq} = \underline{\underline{\hat{y}_a^i - y_a^i}}$$

$$W^2: \frac{\partial CE^i}{\partial W_{\alpha\beta}^2} = \frac{\partial CE^i}{\partial z_q^i}\frac{\partial z_q^i}{\partial W_{\alpha\beta}^2} = (\hat{y}_q^i - y_q^i)\cdot(\delta_{aq}\,a_\beta) = \underline{\underline{(\hat{y}_a^i - y_a^i)\,a_\beta^i}}$$

$$b_a^1: \underline{\frac{\partial CE^i}{\partial b_a^1}} = \frac{\partial CE^i}{\partial z_q^i}\cdot\frac{\partial z_q^i}{\partial a_r^i}\frac{\partial a_r^i}{\partial b_a^i} = (\hat{y}_q^i - y_q^i)\cdot W_{qr}^2\,\sigma_r^i(1-\sigma_r^i)\,\delta_{ra}$$

$$= \underline{\underline{(\hat{y}_q^i - y_q^i)W_{qa}^2\,\sigma_a^i(1-\sigma_a^i)}} = (\hat{y}_q^i - y_q^i)W_{qa}^2\,a_a^i(1-a_a^i)$$

$$W^1: \underline{\frac{\partial CE^i}{\partial W_{\alpha\beta}^1}} = \frac{\partial CE^i}{\partial z_q^i}\frac{\partial z_q^i}{\partial a_r^i}\frac{\partial a_r^i}{\partial W_{\alpha\beta}^1} = (\hat{y}_q^i - y_q^i)W_{qr}^2\,\sigma_r^i(1-\sigma_r^i)(\delta_{ra}\,x_\beta^i)$$

$$= \underline{\underline{(\hat{y}_q^i - y_q^i)W_{qa}^2\,\sigma_a^i(1-\sigma_a^i)x_\beta^i}} = (\hat{y}_q^i - y_q^i)W_{qa}^2\,a_a^i(1-a_a^i)\,x_\beta^i$$

~~Thx Ashish~~ Nice done – my code works good!

5c) Both regularised & non-regularised converge well.
But the ~~two~~ remarkable thing here is that adding the regularisation had almost no effect on the training set accuracy & loss, but while giving a significant several percentage point increase in the dev set accuracy & loss. This is due to regularisation punishing overfitting (albeit at only ~~much lower~~ ~~cost~~ a modest increase in bias)

5d) 93% vs 97% for the test set, without and with regularisation.
This makes sense for the same reasons as above.