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.

6a) $p(\theta|x,y) = \frac{p(\theta,x,y)}{p(x,y)} = p(y|x,\theta)\frac{p(\theta,x)}{p(x,y)} = p(y|x,\theta)\frac{p(\theta)\,p(x)}{p(y|x)\,p(x)}$

$= p(y|x,\theta)\frac{p(\theta)}{p(y|x)}$ $\quad p(y|x)$ indep. of θ.

$\Rightarrow \theta_{MAP} = \operatorname{argmax}_\theta p(y|x,\theta)\frac{p(\theta)}{p(y|x)} = \operatorname{argmax}_\theta p(y|x,\theta)p(\theta)$

6b) If we set our prior to be $p(\theta; 0, \tau^2 I) = \frac{\det(\tau^2 I)^{-1/2}}{\sqrt{2\pi}^k} \exp\left(\frac{1}{2}\theta^T \frac{1}{\tau^2 I}\right)$

then we get that $\theta_{MAP} = \operatorname{argmax}_\theta p(y|x,\theta) \dots \exp(\dots)$

$\Leftrightarrow \theta_{MAP} = \operatorname{argmax}_\theta \log p(y|x,\theta) - \log(\sqrt{2\pi}^{k/2}\tau) - \frac{\theta^T\theta}{2\tau^2}$

$= \operatorname{argmin}_\theta -\log p(y|x,\theta) + (2\tau^2)^{-1}\|\theta\|_2^2$, $\lambda = (2\tau^2)^{-1}$

6c) In other words, we have $p(\vec{y}|X,\theta) \sim \exp\frac{-1}{2\sigma^2}(\vec{y}-X\theta)^T(\vec{y}-X$

$\Rightarrow \theta_{MAP} = \operatorname{argmin}_\theta \frac{1}{2\sigma^2}(\vec{y}-X\vec{\theta})^T(\vec{y}-X\vec{\theta}) + \frac{\vec{\theta}^T\vec{\theta}}{2\tau^2} = \operatorname{argmin}_\theta \frac{1}{2\sigma^2}\left[\vec{y}^T\vec{y} + \vec{\theta}^T X^T X\vec{\theta} - \vec{\theta}^T X^T y - y^T X\vec{\theta}\right] + \frac{1}{2}\tau^2\theta^T\theta.$

($\vec{y}$: $n\times 1$, X: $n\times k$, $\vec{\theta}$: $k\times 1$)

Taking the derivative wrt θ of the loss gives:

$$\frac{\partial}{\partial \theta} Loss = \frac{1}{2}(X^T X \theta - 2X^T y) + \frac{\theta}{\tau^2} = 0$$

$$\Leftrightarrow \left(X^T X + \frac{\sigma^2}{\tau^2} I\right) \theta_{MAP} = X^T y$$

$$\Rightarrow \theta_{MAP} = \left[X^T X + \frac{\sigma^2}{\tau^2} I\right] X^T y$$

$$= \left(X^T X + 2\sigma^2 \lambda I\right) X^T y$$

6d) Now $\theta_{MAP} = \underset{\theta}{\operatorname{argmin}} - \log p(y|X,\theta) + \sum_{i=1}^{k} \frac{1}{b} |\theta_i|$

$$= \underset{\vec{\theta}}{\operatorname{argmin}} \frac{1}{2\sigma^2} |\vec{y} - X\vec{\theta}|_2^2 + \frac{1}{b} |\theta|_1$$

$$= \underset{\vec{\theta}}{\operatorname{argmin}} |\vec{y} - X\vec{\theta}|_2^2 + \frac{2\sigma^2}{b} |\theta|_1 \text{, with } \gamma = \frac{2\sigma^2}{b}$$