Question 3.1

Given : $\hat{\omega}$ for L2 reg. cost is :

$$\hat{\omega} = Q(\Lambda + \alpha I)^{-1} \Lambda Q^T \omega^*$$

Second order Taylor expansion to approximate reg. cost $\hat{J}(\theta)$.

$$\hat{J}(\omega) = J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H (\omega - \omega^*)$$

where ω^* is optimal sol. on J.

$\downarrow$ find gradient $\nabla_\omega \hat{J}(\omega)$

$$\nabla_\omega J(\omega^*) = 0$$

ω^* is a minimum on J, so gradient is zero.

$$\nabla_\omega \hat{J}(\theta) = H(\omega - \omega^*)$$

$\downarrow$ find weight update at iteration t

$$\omega^t = \omega^{t-1} - \varepsilon \underbrace{H(\omega^{t-1} - \omega^*)}_{\text{gradient evaluated at } \omega^{t-1}}$$

$$\omega^t - \omega^* = (I - \varepsilon H)(\omega^{t-1} - \omega^*)$$

$\downarrow$ decompose $H \rightarrow Q \Lambda Q^T$

$$\omega^t - \omega^* = (I - \varepsilon Q \Lambda Q^T)(\omega^{t-1} - \omega^*)$$

$$= Q(I - \varepsilon \Lambda) Q^T (\omega^{t-1} - \omega^*)$$

- $\omega^{(0)} = [0 .. 0]$, initialization from origin.
- learning rate is small

$$\omega^t = Q[I - (I - \varepsilon \Lambda)^t] Q^T \omega^*$$

$$\hat{\omega} = Q(\Lambda + \alpha I)^{-1} \Lambda Q^T \omega^* \quad \text{(L2 reg.)}$$

$$= Q[I - (\Lambda + \alpha I)^{-1} \alpha] Q^T \omega^*$$

Condition for $\omega^t = \hat{\omega}$:

$$(I - \varepsilon \Lambda)^t = (\Lambda + \alpha I)^{-1} \alpha$$

$\downarrow$ $\varepsilon \lambda_i$ is small

$$t \approx \frac{1}{\varepsilon \alpha}$$

$\therefore$ t and α are inversely related.

early stopping is equivalent to using a large regularization constant.

Question 3.2

1) Verify $\underbrace{\nabla_{\omega^{(k)}} \mathcal{J}}_{m\times n} = \underbrace{g}_{m\times 1} \underbrace{h^{(k-1)T}}_{1\times n} + \lambda \underbrace{\nabla_{\omega^{(k)}} \Omega(\theta)}_{m\times n}$

$\boxed{\text{non-regularized}}$

$$\underbrace{\nabla_{\omega^{(k)}} \mathcal{J}}_{m\times n} = \underbrace{\nabla_{a^{(k)}} \mathcal{J}}_{m\times 1} \cdot \underbrace{\frac{\partial a^{(k)}}{\partial \omega^{(k)}}}_{1\times n} \quad \text{(chain rule)}$$

this is g

a : $m\times 1$ ⇒ ω : $m\times n$

this is vector to matrix diff, but rows in ω are independent. Let's break it down to scalar to vector operation as follows :

Take $a_1^{(k)}$, the first element (scalar) in a ; and $\omega^{(k)}_{1,:}$, the first role of matrix $\omega^{(k)}$ $(1\times n)$

$$\frac{\partial a_1^{(k)}}{\partial \omega^{(k)}_{1,:}} = \begin{bmatrix} \frac{\partial f}{\partial \omega_{1,1}} \\ \vdots \\ \frac{\partial f}{\partial \omega_{1,n}} \end{bmatrix}, \quad \text{where } f = \sum_{i=1}^{n} \omega^{(k)}_{1,i} \cdot h^{(k-1)}_i$$

$$\frac{\partial a_1^{(k)}}{\partial \omega^{(k)}_{1,:}} = \begin{bmatrix} h_1^{(k-1)} \\ h_2^{(k-1)} \\ \vdots \\ h_n^{(k-1)} \end{bmatrix}$$

$\frac{\partial a_i^{(k)}}{\partial \omega^{(k)}_{i,:}}$ for all $i \in 1 \dots m$ generate the same result .

Hence , $\frac{\partial a_i^{(k)}}{\partial \omega_i^{(k)}} = \begin{bmatrix} h_1^{(k-1)} \\ h_2^{(k-1)} \\ \vdots \\ h_n^{(k-1)} \end{bmatrix}$

∵ this does not depend on i

∴ it can be applied to each and every element in a (outter product)

$$\therefore \nabla_{\omega^{(k)}} \mathcal{J} = \nabla_{a^{(k)}} \mathcal{J} \cdot h^{(k-1)T}$$

$\boxed{\text{regularized}}$

$$\mathcal{J}\_reg = \lambda \Omega(\underbrace{\omega_1, \omega_2 \cdots \omega_l}_{\text{independent}}) \quad \therefore \text{no chain rule}$$

$$\frac{\partial \mathcal{J}\_reg}{\omega_k} = \lambda \nabla_{\omega^{(k)}}(\theta)$$

$\boxed{\text{overall}}$ $\nabla_{\omega^{(k)}} \mathcal{J} = g h^{(k-1)T} + \lambda \nabla_{\omega^{(k)}} \Omega(\theta)$

2) Verify $\underbrace{\nabla_{h^{(k-1)}} \mathcal{J}}_{n\times 1} = \underbrace{\omega^{(k)T}}_{n\times m} \cdot \underbrace{g}_{m\times 1}$

$$\nabla_{h^{(k-1)}} \mathcal{J} = \left(\frac{\partial a^{(k)}}{\partial h^{(k-1)}}\right)^T \cdot \nabla_{a^{(k)}} \mathcal{J}$$

this is g

$$\frac{\partial a^{(k)}}{\partial h^{(k-1)}} = \underbrace{\begin{bmatrix} \frac{\partial a_1}{\partial h_1} & \frac{\partial a_1}{\partial h_2} & \cdots & \frac{\partial a_1}{\partial h_n} \\ \frac{\partial a_2}{\partial h_1} & \ddots & & \\ \vdots & & \ddots & \\ \frac{\partial a_m}{\partial h_1} & & & \ddots \end{bmatrix}}_{n} \Bigg\} m$$

During forward propagation, $a^{(k)} = \omega^{(k)} h^{(k-1)}$, so $\frac{\partial a^{(k)}}{\partial h^{(k-1)}}$ is exactly $\omega^{(k)}$.

$$\therefore \nabla_{h^{(k-1)}} \mathcal{J} = \omega^{(k)T} \cdot g$$