Q1.

Given $\hat{Y} = Xw$

L_2 penalty term

MSE loss, $J(w) = (\hat{Y} - Y)(\hat{Y} - Y)^T + \lambda \|w\|_2^2$

$= (Xw - Y)(Xw - Y)^T + \lambda w w^T$

$$\nabla_w J = \frac{\partial J}{\partial w} = \frac{\partial}{\partial w}\left\{(Xw - Y)(Xw - Y)^T + \lambda w w^T\right\}$$

$$\because \frac{\partial (u u^T)}{\partial x} = u^T \frac{\partial u}{\partial x} + u^T \frac{\partial u}{\partial x} = 2u^T \frac{\partial u}{\partial x}$$

$$\Rightarrow 2X(Xw - Y)^T + 2\lambda w = 0$$

For optimal weights $\nabla_w J = 0$

$$\Rightarrow X(Xw - Y)^T = -\lambda w$$

$$\Rightarrow X^T(Xw - Y) = \lambda w$$

$$\Rightarrow X^T X w - X^T Y = \lambda w$$

$$\Rightarrow w(X^T X - \lambda I) = X^T Y$$

$$\Rightarrow \boxed{w = X^T Y (X^T X - \lambda I)^{-1}}$$

Q2. ▷ Weights and Inputs (activations) have zero mean

▷ Weights and Inputs are independent & identically distributed

$$z_i^{(l)} = \sum_{j=1}^{n_{inputs}} a_j^{(l-1)} w_{ji}^{(l)}$$, here $W_{ji} \sim \mathcal{U}(-\alpha, \alpha)$ i.e., weights are uniformly distributed

$$Var(z^{(l)}) = \sum_{j=1}^{n_{l-1}} Var(a_j^{(l-1)} w_{ji}^{(l)})$$

$$\Rightarrow \sum_{j=1}^{n_{l-1}} \left[E(a_j^{(l-1)})^2 \, Var(w_{ji}^{(l)}) + E(W_{ji}^{(l)})^2 \, Var(a_j^{(l-1)}) + Var(a_j^{(l-1)}) \, Var(W_{ji}^{(l)}) \right]$$

$$\boxed{\begin{aligned} Var(XY) = {} & E[X]^2 \, Var(Y) + \\ & E[Y]^2 \, Var(X) + \\ & Var(X) \, Var(Y) \end{aligned}}$$

$$\Rightarrow \quad Var(z^{(l)}) = n_{l-1} \, Var(a_j^{(l-1)}) \, Var(W_{ji}^{(l)})$$

Since $Var(z^{(l)}) = Var(a_j^{(l-1)})$

$$\Rightarrow \quad Var(W_{ji}^{(l)}) = \frac{1}{n_{l-1}}$$

$n_{l-1} \longrightarrow$ number of inputs in $l-1$ layer.