1.) Ridge regression solves $\hat{\beta}_\tau = \text{argmin}_\beta (y - X\beta)^T (y - X\beta) + \tau \beta^T \beta$ with $\tau \geq 0$

Prove: $\mathbb{E}[\hat{\beta}_\tau] = S_\tau^{-1} S \beta^*$ $\wedge$ $Cov[\hat{\beta}_\tau] = S_\tau^{-1} S S_\tau^{-1} \sigma^2$

$cov(a,b) = \mathbb{E}[(a-\bar{a})(b-\bar{b})]$

where $S = X^T X$ $\wedge$ $S_\tau = X^T X + \tau \mathbb{1}_D$

SVD: $X = U \Lambda V^T = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_m \\ | & & | \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_r} \end{pmatrix} \begin{pmatrix} | & & | \\ v_1 & \cdots & v_m \\ | & & | \end{pmatrix}$

↳ $S = X^T X = V \Lambda^2 V^T$

↳ $S_\tau = X^T X + \tau \mathbb{1}_D = V \Lambda^2 V^T + \tau \mathbb{1}_D$

Solution to Ridge regression: $\beta_\tau = (X^T X + \tau \mathbb{1}_D)^{-1} X^T y$ (from lecture)

True model: $y = X\beta^* + \varepsilon$

↳ $\hat{\beta}_\tau = (X^T X + \tau \mathbb{1}_D)^{-1} X^T (X\beta^* + \varepsilon)$

$= (X^T X + \tau \mathbb{1}_D)^{-1} X^T X \beta^* + (X^T X + \tau \mathbb{1}_D)^{-1} X^T \varepsilon$

$= S_\tau^{-1} S \beta^* + S_\tau^{-1} X^T \varepsilon$

↳ $\mathbb{E}[\hat{\beta}_\tau] = \mathbb{E}[S_\tau^{-1} S \beta^* + S_\tau^{-1} X^T \varepsilon] = S_\tau^{-1} S \beta^*$ q.e.d.

$\mathbb{E}[S_\tau^{-1} X^T \varepsilon] = 0$ for $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ centred around 0

$Cov[\hat{\beta}_\tau] = Cov[S_\tau^{-1} S \beta^* + S_\tau^{-1} X^T \varepsilon]$

2.) $$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 = \sum_{i=1}^{N} -2 X_i (y_i^* - X_i \beta) \overset{!}{=} 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{N} X_i y_i^* = \sum_{i=1}^{N} X_i X_i \beta$$

$$\Leftrightarrow \quad X^T y^* = X^T X \beta$$

$$\mu_{-1} = \frac{1}{N_{-1}} \sum_{i: y_i^* = -1} X_i = \frac{2}{N} \sum_{i: y_i^* = -1} X_i \qquad \wedge \qquad \mu_1 = \frac{2}{N} \sum_{i: y_i^* = 1} X_i$$

$$X^T y_i^* = \sum_{i: y_i^* = -1} X_i - \sum_{i: y_i^* = 1} X_i = \frac{N}{2} \mu_1 - \frac{N}{2} \mu_{-1} = \frac{N}{2} (\mu_1 - \mu_{-1})$$

$$X^T X = \sum X_i^T X_i = \sum_{i: y_i^* = -1} X_i^T X_i + \sum_{i: y_i^* = 1} X_i^T X_i \overset{\text{since data is centred}}{=} \sum_{i: y_i^* = -1} (X_i - \mu_{-1})^T (X_i - \mu_{-1}) + \sum_{i: y_i^* = 1} (X_i - \mu_1)^T (X_i - \mu_1)$$

$$= N \cdot \Sigma$$

$$\text{L.S.} \quad \Sigma \cdot \beta + \frac{1}{4} (\mu_1 - \mu_{-1})^T (\mu_1 - \mu_{-1}) \cdot \beta = \frac{1}{2} (\mu_1 - \mu_{-1})^T$$