2.2.2 Theory.

$$\beta^{ridge} = \arg\min_{\beta} \underbrace{\|y - X^T\beta\|^2 + \lambda\|\beta\|_2^2}_{f(\beta)}$$

where $f(\beta) \overset{def}{=} \|y - X^T\beta\|^2 + \lambda\|\beta\|_2^2$

$$= (y - X^T\beta)^T(y - X^T\beta) + \lambda\beta^T\beta$$

$$= y^Ty - y^TX^T\beta - \beta^TXy + \beta^TXX^T\beta$$
$$+ \lambda\beta^T\beta$$

$$= \beta^T(XX^T + \lambda I)\beta - 2y^TX^T\beta + y^Ty$$

differentiation rule

$$\frac{\partial f}{\partial \beta} = 2(XX^T + \lambda I)\beta - 2Xy$$

$$\frac{\partial^2 f}{\partial \beta^2} = 2(XX^T + \lambda I)$$

Notice for any vector $u \in \mathbb{R}^p$

$$u^T \frac{\partial^2 f}{\partial \beta^2} u = 2u^T(XX^T + \lambda I)u$$

$$= 2u^T XX^T u + 2\lambda u^T u$$

$$= 2\|X^T u\|^2 + 2\lambda \|u\|^2 \geq 0$$

$\frac{\partial^2 f}{\partial \beta^2}$ is pos. semi-def, so f is convex.

It suffices to find local minimizers.
To do this, we solve for the critical points from $\frac{\partial f}{\partial \beta} = 0$, i.e.

$$(XX^T + \lambda I)\beta^{ridge} \stackrel{!}{=} Xy$$

$$\beta^{ridge} = (XX^T + \lambda I)^{-1} Xy$$

(In fact,

$$f(\beta) - f(\beta^{ridge}) = \underbrace{\cancel{\nabla f(\beta^{ridge})}}_{=0} \cdot (\beta - \beta^{ridge})$$

$$+ \frac{1}{2!} (\beta - \beta^{ridge})^T \underbrace{\frac{\partial^2 f}{\partial \beta^2}}_{\text{pos semi-def}} (\beta - \beta^{ridge}) \geq 0)$$

$$\boxed{\text{In conclusion: } \beta^{ridge} = (XX^T + \lambda I)^{-1} X y \quad (\lambda > 0)}$$

Q: Why β^{ridge} reduces over-fitting compared to β^{ls}?

A: Without Loss of Generality, suppose

$X \in \mathbb{R}^{p \times n}$ has full row rank and $n > p$.

Let $X = U\Sigma V^T = U \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_p & \end{pmatrix}_{p\times n} V^T$

where U, V are orthogonal matrices,

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$.

For β^{ls}, it satisfies

$$XX^T\beta^{ls} = Xy = X(X^T\beta^* + \epsilon)$$

$$= XX^T\beta^* + X\epsilon$$

$$\overset{SVD}{\Longrightarrow} \quad \cancel{U}\Sigma\Sigma^T \underbrace{U^T\beta^{ls}}_{\overset{def}{=}\tilde{\beta}^{ls}} = \cancel{U}\Sigma\Sigma^T \underbrace{U^T\beta^*}_{\overset{def}{=}\tilde{\beta}^*} + \cancel{U}\Sigma \underbrace{V^T\epsilon}_{\overset{def}{=}\tilde{\epsilon}}$$

$$\Rightarrow \quad \Sigma\Sigma^T\tilde{\beta}^{ls} = \Sigma\Sigma^T\tilde{\beta}^* + \Sigma\tilde{\epsilon}$$

i.e. $\begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_p^2 \end{pmatrix} \tilde{\beta}^{ls} = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_p^2 \end{pmatrix} \tilde{\beta}^* +$

$$\begin{pmatrix} \sigma_1 & & & \vdots & \\ & \ddots & & \vdots & 0 \\ & & \sigma_p & \vdots & \end{pmatrix}_{p\times n} \tilde{\epsilon}$$

i.e. for each i, $\sigma_i^2 \tilde{\beta}_i^{ls} = \sigma_i^2 \tilde{\beta}^* + \sigma_i \tilde{\epsilon}_i$

$$\Rightarrow \quad \tilde{\beta}_i^{ls} = \tilde{\beta}_i^* + \frac{1}{\sigma_i}\tilde{\epsilon}_i, \quad i=1,\cdots,p$$

For β^{ridge}, $(XX^T+\lambda I)\beta^{ridge} = Xy$

$$= \mathbf{X}(X^T\beta^* + \epsilon) = XX^T\beta^* + X\epsilon$$

$$\overset{SVD}{\Longrightarrow} \left(U\Sigma\Sigma^T U^T + \lambda I\right)\beta^{ridge} =$$

$$U\Sigma\Sigma^T U\beta^* + U\Sigma V^T\epsilon$$

i.e. $\cancel{U}\left(\Sigma\Sigma^T + \lambda I\right)\underbrace{U^T\beta^{ridge}}_{\tilde{\beta}^{ridge}}$

$$= \cancel{U}\Sigma\Sigma^T \underbrace{U^T\beta^*}_{\tilde{\beta}^*} + \cancel{U}\Sigma\underbrace{V^T\epsilon}_{\tilde{\epsilon}}$$

$$\Rightarrow \quad (\Sigma\Sigma^T + \lambda I)\,\tilde{\beta}^{ridge} = \Sigma\Sigma^T\tilde{\beta}^* + \Sigma\tilde{\epsilon}$$

i.e.
$$\begin{pmatrix} \sigma_1^2+\lambda & & \\ & \ddots & \\ & & \sigma_p^2+\lambda \end{pmatrix} \tilde{\beta}^{ridge}$$

$$= \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_p^2 \end{pmatrix} \tilde{\beta}^* + \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_p & \end{pmatrix} \tilde{\epsilon}$$

i.e. for each i, $\quad (\sigma_i^2+\lambda)\,\tilde{\beta}_i^{ridge} = \sigma_i^2\tilde{\beta}_i^* + \sigma_i\tilde{\epsilon}_i$

$$\Rightarrow \quad \tilde{\beta}_i^{ridge} = \frac{\sigma_i^2}{\sigma_i^2+\lambda}\tilde{\beta}_i^* + \frac{\sigma_i}{\sigma_i^2+\lambda}\tilde{\epsilon}_i$$

$$= \frac{\sigma_i^2}{\sigma_i^2+\lambda}\tilde{\beta}_i^* + \frac{1}{\sigma_i + \frac{\lambda}{\sigma_i}}\tilde{\epsilon}_i$$

$$= \frac{\sigma_i^2}{\sigma_i^2+\lambda}\tilde{\beta}_i^* + \underbrace{\frac{1}{\left(\sqrt{\sigma_i}-\sqrt{\frac{\lambda}{\sigma_i}}\right)^2+2\sqrt{\lambda}}}_{\le \frac{1}{2\sqrt{\lambda}}}\tilde{\epsilon}_i$$

Remark: (1) As $\lambda \to 0$, $\underbrace{\tilde{\beta}^{ridge}}_{= U^T\beta^{ridge}} \to \tilde{\beta}^{ls} = U^T\beta^{ls}$

thus $\beta^{ridge} \to \beta^{ls}$.

(2) If $\epsilon = 0$, $\tilde{\beta}^{ridge} = \frac{\sigma_i^2}{\sigma_i^2+\lambda}\tilde{\beta}^*$

thus $\beta^{ridge} = \frac{\sigma_i^2}{\sigma_i^2+\lambda}\beta^*$.

But $\tilde{\beta}^{ls} = \tilde{\beta}^*$, thus $\beta^{ls} = \beta^*$.