1) $\tilde{\beta} = (H+D)y = Hy + Dy = \beta^* + Dy$

$E(\tilde{\beta}) = E(\beta^*) + E(Dy) = \beta + E(Dy)$

$Var(\tilde{\beta}) = Var(\beta^* + Dy) = Var(\beta^*) + D\,Var(y)\,D^T$

$= Var(\beta^*) + D(y\,y^T)D^T \geqslant Var(\beta^*)$

$Var(\tilde{\beta}) > Var(\beta^*) \quad \forall\, D \neq 0$

She assumption made is that $Y = X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$

2.1) $\beta^*_{ridge} = (X^T X + \lambda I)^{-1} X^T Y \Rightarrow$

$E(\beta^*_{ridge}) = E(\underline{(X^T X + \lambda I)^{-1} X^T} Y) = (X^T X + \lambda I)^{-1} X^T E(Y)$

$Y = X\beta + \varepsilon \Rightarrow E(\beta^*_{ridge}) = (X^T X + \lambda I)^{-1} X^T (E(X\beta) + \cancel{E(\varepsilon)}^{0})$

$E(\beta^*_{ridge}) = (X^T X + \lambda I)^{-1} X^T X \beta \neq \beta$ (unless $\lambda = 0$)

2.2) $X = UDV^T$ and $X^T = VDU^T$

$\beta^*_{ridge} = (VD\cancel{U^T U}DV^T + \lambda I)^{-1} VDU^T Y = (VDDV^T + \lambda I)^{-1} VDU^T Y$

$= (VDDV^T + V\lambda I V^T)^{-1} VDU^T Y = \cancel{V}(DD + \lambda I)^{-1} \cancel{V^T V} DU^T Y$

$= V(DD + \lambda I)^{-1} DU^T Y$

It is usually much easier to find $(DD + \lambda I)^{-1}$ than $(X^T X + \lambda I)^{-1}$. Sherefore, it is worth to do this decomposition when the computational cost of this operation is lower than the one of finding $(X^T X + \lambda I)^{-1}$. An example of this is when we are testing multiplble λ.

2.3) $Var((X^T X+\lambda I)^{-1} X^T y))$

$= (X^T X+\lambda I)^{-1} X^T Var(Y) X (X^T X+\lambda I)^{-1}$ but $Y = X\beta + \varepsilon \Rightarrow Var(Y) = Var(\varepsilon)$ $= \sigma^2$

$= \sigma^2 (X^T X+\lambda I)^{-1} X^T X (X^T X+\lambda I)^{-1}$

$Var(\beta^*_{OLS}) = \sigma^2 (X^T X)^{-1} \cdot X^T X \cdot (X^T X)^{-1}$

$\lambda \geq 0 \Rightarrow (X^T X+\lambda I)^{-1} \leq (X^T X)^{-1}$

$\Downarrow$

$\sigma^2 (X^T X+\lambda I)^{-1} X^T X (X^T X+\lambda I)^{-1} \leq \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$

$Var(\beta^*_{ridge}) \leq Var(\beta^*_{OLS})$

2.4) $(X^T X+\lambda_1 I)^{-1} X^T X (X^T X+\lambda_1 I)^{-1} < (X^T X+\lambda_2 I)^{-1} X^T X (X^T X+\lambda_2 I)^{-1}$

for $\lambda_1 > \lambda_2$ $\Rightarrow$ when λ increases the varianc decreases.

$b(\beta^*_{ridge}) = E(\beta^*_{ridge}) - \beta = (X^T X+\lambda I)^{-1} X^T X \beta - I\beta$

$= ((X^T X+\lambda I)^{-1} X^T X - I)\beta$ $\quad \lambda = 0 \Rightarrow b(\beta^*_{ridge}) = 0$

as λ increases, the bias increase in absolute value.

2.5) $\beta^*_{ridge} = (X^T X+\lambda I)^{-1} X^T Y = (1+\lambda)^{-1} I X^T Y = \dfrac{X^T Y}{(1+\lambda)}$

$\beta^*_{OLS} = (X^T X)^{-1} X^T Y = X^T Y$ $<$

$\beta^*_{ridge} = \dfrac{\beta^*_{OLS}}{(1+\lambda)}$

3) $f(\beta) = (y - X\beta)^T (y - X\beta) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$

$$= y^T y - 2y^T X\beta + \beta^T \overset{I}{\cancel{X^T X}} \beta + \lambda_2 \beta^T \beta + \lambda_1 \|\beta\|_1$$

$$= y^T y - 2y^T X\beta + (1+\lambda_2)\beta^T\beta + \lambda_1 \|\beta\|_1$$

$$\partial f(\beta) = -2(y^T X)^T + 2(1+\lambda_2)\beta + \lambda_1 \partial\|\beta\|_1$$

$$\arg\min_{\beta} f(\beta) \Rightarrow 0 = -2X^T y + 2(1+\lambda_2)\beta^* + \lambda_1 \partial\|\beta^*\|_1$$

$$\partial\|\beta_j\|_1 \begin{cases} [-1, 1], & \beta_j = 0 \\ 1, & \beta_j > 0 \\ -1, & \beta_j < 0 \end{cases} \qquad \beta^* \in \frac{2X^T y - \lambda \partial\|\beta\|_1}{2(1+\lambda_2)}$$

for each β_j in $\vec{\beta}$

$$\beta_j^* \in \begin{cases} \left\{ \dfrac{X^T y - \frac{\lambda_1}{2}}{1+\lambda_2} \right\}, & \beta_j^* > 0 \\ \left\{ \dfrac{X^T y + \frac{\lambda_1}{2}}{1+\lambda_2} \right\}, & \beta_j^* < 0 \\ \left[\dfrac{X^T y - \frac{\lambda_1}{2}}{1+\lambda_2}, \dfrac{X^T y + \frac{\lambda_1}{2}}{1+\lambda_2} \right], & \beta_j^* = 0 \end{cases}$$

$$\beta^*_{OLS} = (\overset{I}{\cancel{X^T X}})^{-1} X^T y = X^T y$$

$$\Rightarrow \beta_j^* = \begin{cases} \dfrac{\beta^*_{OLS_j} \pm \frac{\lambda_1}{2}}{1+\lambda_2}, & \beta_{OLS_j} \neq \frac{\lambda_1}{2} \\ \left[0, \dfrac{\lambda_1}{1+\lambda_2} \right], & \beta_{OLS_j} = \frac{\lambda_1}{2} \end{cases}$$