1. Ridge regression solves the regularized least squares problem

$$\hat{\beta}_T = \text{argmin}_\beta (y - X\beta)^T (y - X\beta) + T\beta^T\beta$$

Here we can use the singular value decomposition of the design matrix X.

$$X = UDV^T$$

- U is an orthogonal matrix of size $N \times N$
- D is an diagonal matrix of size $N \times P$ with non-negative diagonal elements sorted in descending order.
- V is an orthogonal matrix of size $P \times P$.

By substituting the SVD of X into the ridge regression problem,

$$\beta_T = \text{arg min}_\beta (y - UDV^T\beta)^T (y - UDV^T\beta) + T\beta^T\beta$$

$$= \text{arg min}_\beta (y - UD\hat{\beta})^T (y - UD\hat{\beta}) + T(\hat{\beta}^T\hat{\beta})$$

where $\hat{\beta} = V^T\beta$

Here, we expanding the above expression:

$$(y - UD\hat{\beta})^T (y - UD\hat{\beta}) + T\hat{\beta}^T\hat{\beta} = y^Ty - 2\hat{\beta}^T U^T y D + \hat{\beta}^T D^T$$

$$= y^Ty - 2\hat{\beta}^T U^T y D + \hat{\beta}^T (D$$

$$(y - UD\hat{\beta})^T (y - UD\hat{\beta}) + T\hat{\beta}^T\hat{\beta} = y^Ty - 2\hat{\beta}^T U^T y D + \hat{\beta}^T D^T \hat{\beta} + T\hat{\beta}^T\hat{\beta}$$

$$= y^Ty - 2\hat{\beta}^T U^T y D + \hat{\beta}^T (D^T D + TI)\hat{\beta}$$

Those upper expression is minimized when

$$\hat{\beta} = (D^T D + TI)^{-1} U^T y D$$

$$\beta_\tau = V(D^TD + \tau I)^{-1} U^T yD$$
$$= V(D^TD + \tau I)^{-1} V^T V U^T yD$$
$$= V(D^TD + \tau I)^{-1} V^T X^T y$$
$$= X^T y \,(VD(D^TD + \tau I)^{-1} V^T)^T$$
$$= X^T y\,(VD^T D(D^TD + \tau I)^{-1} V^T)$$
$$= X^T y\,(VD^T(D^TD + \tau I)^{-1} V^T V)$$
$$= X^T y\,(VD^T(D^TD + \tau I)^{-1})$$
$$= X^T y\,(VD^{-1}(D^TD + \tau I)^{-1})$$
$$= X^T y\,(D^TD + \tau I)^{-1} D^{-T} V^T$$
$$= X^T y\; S_\tau^{-1} S^{-1}$$

And here, $S = X^T X$ is the ordinary scatter matrix and $S_\tau = X^T X + \tau I$ is the ~~regularized~~ regularized scatter matrix

Expectation of β_τ is given by:

$$E[\beta_\tau] = X^T y\, S_\tau^{-1} S^{-1} = (S_\tau)^{-1} S \beta^*$$

$\therefore$ β^* is the true model coefficient

We substitute the expression for $\hat{\beta}$ back to find the covariance of β_τ.

$$(y - UD\hat{\beta})^T (y - UD\hat{\beta}) + \tau\hat{\beta}^T\hat{\beta} = y^Ty - 2\hat{\beta}^T U^T yD + \hat{\beta}^T D^T D\hat{\beta} + \tau\hat{\beta}^T\hat{\beta}$$
$$= y^Ty - 2y^T UDV^T\hat{\beta} + \hat{\beta}^T VDU^T UDV^T\hat{\beta} + \tau\hat{\beta}^T\hat{\beta}$$
$$= y^Ty - 2y^T X\beta + \beta^T X^T X\beta + \tau\beta^T\beta$$
$$= y^Ty - 2y^T X\beta + \beta^T S\beta + \tau\beta^T\beta$$
$$= y^Ty - 2y^T X\beta + \beta^T(S + \tau I)\beta$$

$$S + \tau I = S_\tau$$

The covariance of β_T :

$$Cov[\beta_T] = (S_T)^{-1}\sigma^2$$
$$= (S+\tau I)^{-1}\sigma^2$$
$$= (S_T^{-1})^T\sigma^2$$
$$= (S_T^{-1})\sigma^2$$
$$= (S_T)^{-1} S (S_T)^{-1}\sigma^2$$

- The computation and covariance matrix of the regularized solution, averaged over all possible training sets of size N, are :

$$E[\beta_T] = (S_T)^{-1} S\beta^*$$

&

$$Cov[\beta_T] = (S_T)^{-1} S (S_T)^{-1}\sigma^2$$

where $S = X^T X$ & $S_T = X^T X + \tau I$.

2) Expanding the square loss term

$$\frac{\partial}{\partial \beta}\left(-\sum_{i=1}^{N}(y_i^* - X_i \cdot \beta)^2\right)$$

$$= \sum_{i=1}^{N} 2(y_i^* - X_i \cdot \beta)(-X_i)$$

$$\rightarrow -2\sum_{i=1}^{N} X_i^T (y_i^* - X_i \cdot \beta)$$

$\rightarrow y_i^* \rightarrow y_i$ (true class label) & rewrite eq^n:

$$-2\sum_{i=1}^{N} X_i^T (y_i - X_i \cdot \beta) = -2\sum_{i=1}^{N}(X_i^T y_i - X_i^T \cdot X_i \cdot \beta)$$

Distribute the transpose:

$$-2\sum_{i=1}^{N}(y_i^T X_i - \beta^T X_i^T X_i)$$

$\sum_{i=1}^{N} y_i^T X_i$ gets cancels out since classes are balanced

& $\sum_{i=1}^{N} y_i = 0$

→ Simplify the equation:

$$-2\sum_{i=1}^{N}(-\beta^T X_i^T X_i) = 2\sum_{i=1}^{N}\beta^T X_i^T X_i$$

Re-arrange the terms:

$$2\sum_{i=1}^{N} \beta^T X_i^T X_i = 2N\beta^T\hat{\Sigma}$$ where $\hat{\Sigma}$ is covariance matrix

& $$\sum_{i=1}^{N} X_i^T X_i = N\hat{\Sigma}$$

→ set the derivative to 0:

$$2N\beta^T\hat{\Sigma} + \frac{1}{4}(\mu_1-\mu_{-1})^T(\mu_1-\mu_{-1})\beta = 0$$

→ Rearrange the eqⁿ:

$$\hat{\Sigma}\beta + \frac{1}{4N}(\mu_1-\mu_{-1})^T(\mu_1-\mu_{-1})\beta = 0$$

→ Rewrite $(\mu_1-\mu_{-1})\beta$ as τ' for scalar τ'
& move second term to right hand side:

$$\hat{\Sigma}\beta = -\frac{1}{4N}(\mu_1-\mu_{-1})^T(\mu_1-\mu_{-1})\beta$$

$$\Rightarrow \hat{\Sigma}\beta = -\frac{1}{2N}(\mu_1-\mu_{-1})^T(\mu_1-\mu_{-1})\beta$$

→ Rearrange the eqⁿ with $\tau = \frac{1}{2} - \frac{\tau'}{4}$:

$$\hat{\Sigma}\beta = \tau(\mu_1-\mu_{-1})^T$$

Divide the eqⁿ by $\hat{\Sigma}$: we get

$$\boxed{\hat{\beta}_{OLS} = \tau\Sigma^{-1}(\mu_1-\mu_{-1})^T \text{ with } \tau = \frac{1}{2} - \frac{\tau'}{4}}$$