1. Generalization error of OLS

(a) Since $n \geq d$ and X is full rank. ($X \in \mathbb{R}^{n\times d}$)

$$d = \text{rank}(X) = \text{rank}(XX^T) = \text{rank}(X^TX)$$

We have $X^TX \in \mathbb{R}^{d\times d}$ is non-singular

Thus $\hat{\beta} = (X^TX)^{-1}X^Ty$

We also have $y = X\beta^* + e$, where $X = (x_1 \cdots x_n)^T$, $\beta^* \in \mathbb{R}^{d\times 1}$ the ground truth and $e = (\varepsilon_1 \cdots \varepsilon_n)^T$, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$

$$\mathbb{E}_{e,x}\hat{\beta} = \mathbb{E}(X^TX)^{-1}X^T(X\beta^* + e) = \beta^* + \mathbb{E}(X^TX)^{-1}X^Te$$

$$= \beta^* + \mathbb{E}_x(X^TX)^{-1}X^T\mathbb{E}e = \beta^* + \mathbb{E}_x 0 = \beta^* \quad \# \quad (\text{As } \mathbb{E}e = 0)$$

(b) $\mathbb{E}\|\hat{\beta} - \beta^*\|_2^2 = \mathbb{E}(\hat{\beta}^T - \beta^{*T})(\hat{\beta} - \beta^*) = \mathbb{E}\hat{\beta}^T\hat{\beta} - \beta^{*T}\beta^* - \beta^{*T}\beta^* + \beta^{*T}\beta^* = \mathbb{E}\left[y^TX(X^TX)^{-1}(X^TX)^{-1}X^Ty\right] - \beta^{*T}\beta^*$

$$= \mathbb{E}\left[(\beta^{*T} + e^TX(X^TX)^{-1})(\beta^* + (X^TX)^{-1}X^Te)\right] - \beta^{*T}\beta^*$$

$$= \mathbb{E}\left[e^TX(X^TX)^{-1}\beta^*\right] + \mathbb{E}\left[\beta^{*T}(X^TX)^{-1}X^Te\right] + \mathbb{E}\left[e^TX(X^TX)^{-2}X^Te\right]$$

$$= \mathbb{E}\left[e^TX(X^TX)^{-2}X^Te\right]$$

Denote by $Y = X(X^TX)^{-2}X^T$, then $Y \in \mathbb{R}^{n\times n}$ and $Y = Y^T$

$$\mathbb{E}\|\hat{\beta} - \beta^*\|_2^2 = \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[Y_{ij}\,\varepsilon_i\,\varepsilon_j\right] = \sum_{i=1}^n \mathbb{E}_x\left[Y_{ii}\,\mathbb{E}_{\varepsilon_i}\varepsilon_i^2\right]$$

$$= \sum_{i=1}^n \mathbb{E}_x[Y_{ii}]\,\sigma^2 = \sigma^2\,\mathbb{E}_x\sum_{i=1}^n Y_{ii} = \sigma^2\,\mathbb{E}_x\,\text{Tr}\,Y$$

$$= \sigma^2\,\mathbb{E}_x\,\text{Tr}\left[X(X^TX)^{-2}X^T\right] = \sigma^2\,\mathbb{E}_x\,\text{Tr}\left[X^TX(X^TX)^{-2}\right] = \sigma^2\,\text{Tr}\,\mathbb{E}_x\left[(X^TX)^{-1}\right] \quad \#$$

(c) We know that $X = (x_1 \cdots x_n)^T$, $x_i \overset{i.i.d}{\sim} N(0, I_d)$

then $X^TX = \sum_{i=1}^n x_ix_i^T \sim W(I_d, n)$, the Wishart distribution

Thus $(X^TX)^{-1} \sim W^{-1}(I_d, n)$, the inverse-Wishart distribution

We have $\mathbb{E}(X^TX)^{-1} = \frac{I_d}{n-d-1}$ (K Mardia 1979 Multivariate Analysis)

Thus $\mathbb{E}\|\hat{\beta} - \beta^*\|^2 = e(n, d, \sigma) = \sigma^2\,\text{tr}\left(\frac{I_d}{n-d-1}\right) = \frac{d\sigma^2}{n-d-1}.$

2. Equivalent forms of LASSO

$X \in \mathbb{R}^{n\times d}$ $y \in \mathbb{R}^d$ $\beta \in \mathbb{R}^d$

$S_1(\lambda) = \{\beta_1 \in \mathbb{R}^d : \beta_1 = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1\}$

$S_2(t) = \{\beta_2 \in \mathbb{R}^d : \beta_2 = \arg\min_{\beta} \|y - X\beta\|_2^2, \text{ s.t. } \|\beta\|_1 \le t\}$

(a) Let $\beta_1, \beta_2 \in S_1(\lambda)$ and $C = \|y - X\beta_1\|_2^2 + \lambda\|\beta_1\|_1 = \|y - X\beta_2\|_2^2 + \lambda\|\beta_2\|_1$

Take any $0 < \alpha < 1$ if $X\beta_1 \ne X\beta_2$

$\|y - X(\alpha\beta_1 + (1-\alpha)\beta_2)\|_2^2 + \|\alpha\beta_1 + (1-\alpha)\beta_2\|_1 < \alpha\|\beta_1\|_1 + (1-\alpha)\|\beta_2\|_1$

$+ \alpha\|y - X\beta_1\|^2 + (1-\alpha)\|y - X\beta_2\|^2 = \alpha C + (1-\alpha)C = C$

The strict inequality is due to convexity of $\|x\|_1$ and strict convexity of $\|y - X\|_2^2 = X^TX - y^TX - X^Ty + y^Ty$ and $X\beta_1 \ne X\beta_2$.

Hence $\alpha\beta_1 + (1-\alpha)\beta_2$ attains a smaller value. Contradiction.

Thus we have $X\beta_1 = X\beta_2$. Since $C = \|y - X\beta_1\|_2^2 + \lambda\|\beta_1\|_1 = \|y - X\beta_2\|_2^2 + \lambda\|\beta_2\|_1$

$\|\beta_1\|_1 = \|\beta_2\|_1$ #

(b) First we prove $S_1(\lambda) \subseteq S_2(\varphi(\lambda))$

Set $\beta_3 \in S_1(\lambda)$. We have $\|\beta_3\|_1 = \varphi(\lambda)$

and $\forall \beta \in \mathbb{R}^d$, $\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \ge \|y - X\beta_3\|_2^2 + \lambda\|\beta_3\|_1$

Hence for all β such that $\|\beta\|_1 \le \varphi(\lambda)$,

$\|y - X\beta\|_2^2 \ge \|y - X\beta_3\|^2$ hence $\beta_3 \in S_2(\varphi(\lambda))$

Next we prove $S_1(\lambda) \supseteq S_2(\varphi(\lambda))$

Set $\beta_4 \in S_2(\varphi(\lambda))$ Take any $\beta_5 \in S_1(\lambda)$ (It's obvious that $S_1(\lambda) \ne \phi$ See next page for proof)

We have $|\beta_4| \le \varphi(\lambda) = |\beta_5|$

and $\|y - X\beta_4\|_2^2 \le \|y - X\beta_5\|_2^2$

hence for any $\beta \in \mathbb{R}^d$, $\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \ge \|y - X\beta_5\|_2^2 + \lambda\|\beta_5\|_1$

$\ge \|y - X\beta_4\|_2^2 + \lambda\|\beta_4\|_1 \Rightarrow \beta_4 \in S_1(\lambda)$

Hence $S_1(\lambda) = S_2(\varphi(\lambda))$ #

3. Norm Control of LASSO estimator

As in Prop. 1.5, we have

$$0 \le \frac{1}{2n} \|X(\hat\beta - \beta^*)\|_2 \le \frac{\|X^T\varepsilon\|_\infty}{n} \|\hat\beta - \beta^*\|_1 + \lambda_n (\|\beta^*\|_1 - \|\hat\beta\|_1)$$

$$\le \frac{\lambda_n}{2} \|\hat\beta - \beta^*\|_1 + \lambda_n (\|\beta^*\|_1 - \|\hat\beta\|_1) \quad \text{(given condition)}$$

Thus $0 \le \|\hat\beta - \beta^*\|_1 + 2\|\beta^*\|_1 - 2\|\hat\beta\|_1$

$\le \|\hat\beta\|_1 + \|\beta^*\|_1 + 2\|\beta^*\|_1 - 2\|\hat\beta\|_1$ (triangle inequality)

We have $\|\hat\beta\|_1 \le 3\|\beta^*\|_1$ #

$S_1(\lambda) \neq \phi$: $f(\beta) = \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ is continous function

$f(\beta) \geq 0$ Set $A = f(0) = \|y\|_2^2$ for $\|\beta\|_1 > \frac{A}{\lambda}$,

$f(\beta) \geq \lambda\|\beta\|_1 > A$ Let $f(\beta^*) = \inf_{\|\beta\|_1 \le \frac{A}{\lambda}} f(\beta)$ Then β^* is global minimum.

$\beta^* \in S_1(\lambda)$ #