# Lecture January 20

Shrinkage methods:

## OLS

$$C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - x_{i*}\beta)^2$$

$$= \frac{1}{n} \|(y - X\beta)\|_2^2$$

$$\beta_{OLS}^{opt} = \hat{\beta} (= \hat{\theta}) = (X^T X)^{-1} X^T y$$

$$\tilde{y}^{OLS} = X\beta^{opt}$$

## Ridge

$$C(\beta) = \frac{1}{n} \|(y - X\beta)\|_2^2 + \lambda \|\beta\|_2^2$$

$$\|\beta\|_2^2 = \sum_{i=0}^{p-1} \beta_i^2$$

$$\frac{\partial C}{\partial \beta} = 0 \Rightarrow \beta_{Ridge}^{opt} = \underline{(X^T X + \lambda I)^{-1} X^T y}$$

$$\lambda \|\beta\|_2^2 \leq C < \infty$$

norm-2 of vector $x$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

## Lasso

$$C(\beta) = \frac{1}{n} \|(y - X\beta)\|_2^2 + \lambda \|\beta\|_1$$

## Example

- leave out intercept ($\beta_0$)

$n = p \qquad X \in \mathbb{R}^{n \times p}$

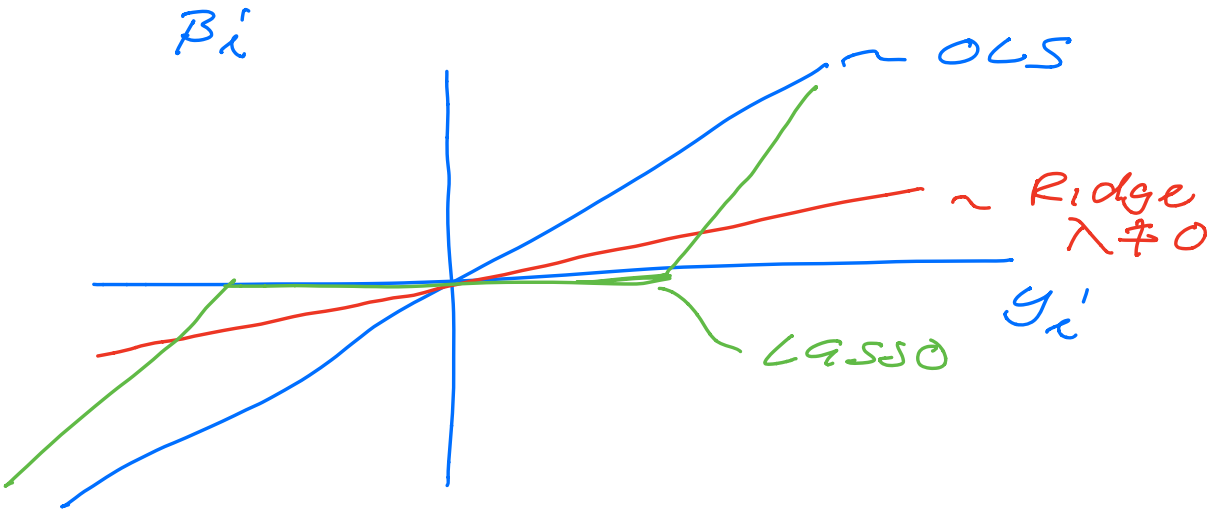
$$X = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}$$

## OLS (drop $\frac{1}{n}$)

$$C(\beta) = \sum_{i=0}^{n-1} (y_i - \beta_i)^2$$

$X_i * \beta$

$\beta_i^{opt} = y_i \qquad \hat{y} = X\beta^{opt}$




## Ridge

$$C(\beta) = \sum_{i=0}^{n-1} (y_i' - \beta_i)^2 + \lambda \sum_{j=0}^{n-1} \beta_j^2$$

$$\frac{\partial C}{\partial \beta} = 0$$

$$\beta_i^{Ridge} = \frac{y_i'}{1+\lambda}$$

$$\tilde{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots$$

## Lasso

$$C(\beta) = \sum_i (y_i' - \beta_i)^2 + \lambda \sum_i |\beta_i|$$

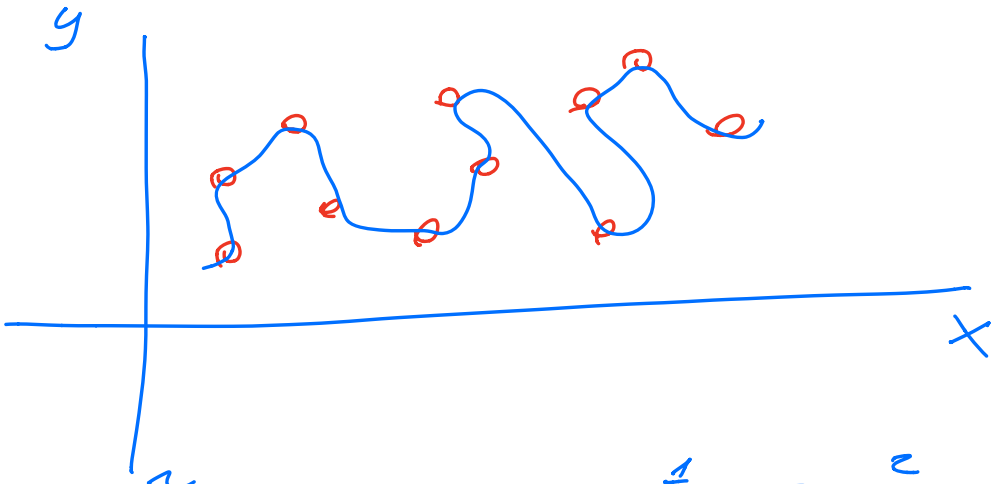
$$\lambda \sum_i \sqrt{\beta_i^2}$$

~ ~

$$\frac{\partial L}{\partial \beta} = 0 \quad \Rightarrow$$

$$-2 \sum_i (y_i' - \beta_i) + \lambda \Sigma \beta_i' / |\beta_i'|$$

$$\beta_i^{lasso} = \begin{cases} y_i' - \lambda/2 & \text{if } y_i' > \lambda/2 \\ y_i' + \lambda/2 & \text{if } y_1' < -\lambda/2 \\ 0 & \text{if } |y_i'| \le \lambda/2 \end{cases}$$




$$\tilde{y}_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1}$$

OLS $\quad y \cong \tilde{y} + \varepsilon$

$$\varepsilon \sim N(0, \sigma^2)$$

$$var(\beta^{opt}) = \sigma^2 (x^T x)^{-1}$$

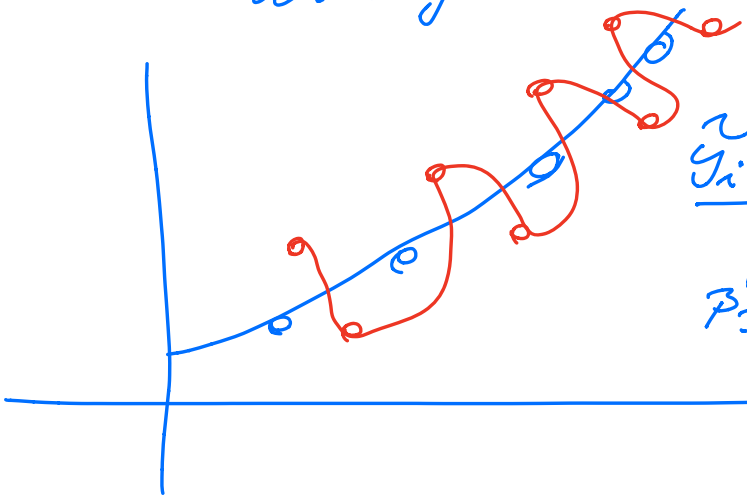
$\tilde{y}_i$ (5th-order)

$$\tilde{y}_i = (\beta_0 \pm STD\beta_0)_{5th} + (\beta_1 \pm STD\beta_1)_{5th} x_i^1$$

$+ \cdots$

$\tilde{y}_i$ (4th-order)

$(\beta_1 \pm STD\beta_1)_{5th}$ and compare with $(\beta_1 \pm STD\beta_1)_{4th}$ will fluctuate wildly.



$$\tilde{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_{p-1} x_i^{p-1}$$

$\beta_3^{(3rd)} \simeq \beta_3^{(4th)}$

## OLS and probability

$$D = \{(x_0, y_0), (x_0, y_1) \dots (x_{m-1}, y_{m-1})\}$$

$$y_i = x_i^* \beta + \varepsilon_i$$

$\varepsilon_i$ : $i.i.d \sim N(0, \sigma^2)$

$\downarrow$

$$N(x_i^* \beta, \sigma^2)$$

$$p(y_i | x_i \beta, \sigma^2) \sim N(x_i * \beta, \sigma^2)$$

Maximum Likelihood:

$$p(y | X\beta, \sigma^2) = \prod_{i=0}^{n-1} p(y_i | x_i * \beta \sigma^2)$$

$$\beta^{opt} = \underset{\beta \in \mathbb{R}^p}{\arg\max}\ p(y | \beta)$$

$$= \underset{\beta \in \mathbb{R}^p}{\arg\max} \sum_{i=0}^{n-1} \log p(y_i | x_i \beta)$$

$$C(\beta) = \sum_{i=0}^{n-1} \log p(y_i | x_i; \beta)$$

minimize

$$C(\beta) = -\sum_{i=0}^{n-1} \log \ldots$$

$-\log$ is a convex function

$$y_i \sim N(x_i * \beta, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(y_i - x_i * \beta)^2 / 2\sigma^2\right]$$

$$= p(y_i | x_{i*}\beta, \sigma^2)$$

$$\log p(y_i | x_{i*}\beta, \sigma^2) =$$

$$-(y_i - x_{i*}\beta)^2/2\sigma^2 \quad -\frac{1}{2}\log(2\pi\sigma^2)$$

$$+\sum_{i=0}^{n-1} (y_i - x_{i*}\beta)^2/2\sigma^2$$

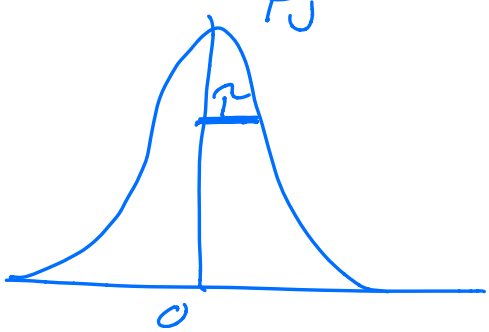
$$+ \frac{n}{2}\log(2\pi\sigma^2)$$

$$\frac{\partial C(\beta)}{\partial \beta} = 0 = \frac{\partial}{\partial \beta}\sum_{i=0}^{n-1} (y_i - x_{i*}\beta)^2$$

$$\Rightarrow \quad \beta^{opt} = (X^T X)^{-1} X^T y$$

Connection with Ridge?

assume $p(\beta) = \prod_{j=0}^{p-1} N(\beta_j | 0, \tau^2)$



Bayesian interpretation

$$p(y \mid x, \beta) = \prod_{i=0}^{n-1} N(y_i \mid x_i \beta, \tau^2) \times \prod_{j=0}^{p-1} N(\beta_j \mid 0, \tau^2)$$

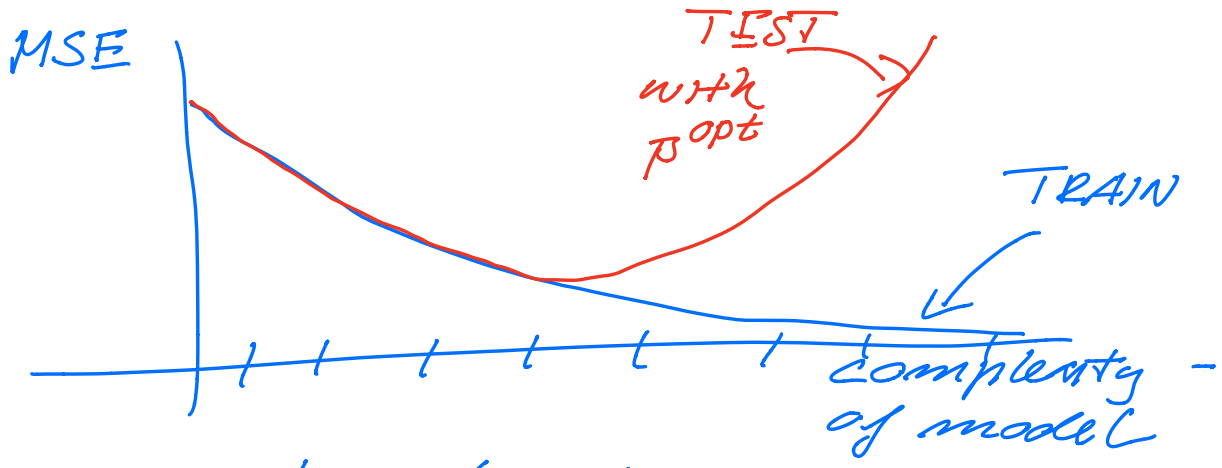
$$\beta^{opt} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; - \left( \sum_{i=0}^{n-1} \log\left(N(y_i \mid x_i * \beta, \tau^2)\right) + \sum_{j=0}^{p-1} \log N(\beta_j \mid 0, \tau^2) \right)$$

$$C(\beta) \propto \sum_{i=0}^{n-1} (y_i - x_i * \beta)^2 + \lambda \|\beta\|_2^2$$

$$\lambda \sim 1/\tau^2$$

Hastie et al chapter 3.4

Bias - Variance tradeoff




Hastie et al, figure 2.11

Bias - variance tradeoff:

Rewrite of MSE.

$$y = f(x) + \varepsilon$$

$$f(x) \approx \tilde{y}(x) = X\beta$$

$$C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \tilde{y}_i)^2$$

Expected value:

$$\mathbb{E}[x^i] = \int p(x) x^i dx$$

$$\mu = \int p(x) x \, dx$$

{ Discrete version

$$\sum_i p(x_i) x_i$$

In ML sample mean

$$\bar{\mu} = \frac{1}{n}\sum_i x_i \neq \mu$$

$$= \mathbb{E}[x] \quad (\langle x \rangle)$$

$$C(\beta) = \mathbb{E}\left[(y-\hat{y})^2\right]$$

(add and subtract $\mathbb{E}[y]$)

$$= \mathbb{E}\left[\underbrace{(y-\mathbb{E}[y])^2}_{\text{Bias}}\right]$$

$$+ \underbrace{\mathbb{E}\left[(\hat{y}-\mathbb{E}[\hat{y}])^2\right]}_{\text{variance (of the model)}} + \sigma^2$$

How to calculate the Bias and variance in a reliable?

- Resampling
  - * Bootstrapping
  - * Cross validation.

Bootstrap

1) pick randomly n-data from a sample of n-data

$$\underline{D} = \{x_1, x_2, x_3, \ldots x_n\}$$

$$D^* = \{\underbrace{x_3, x_1, x_1, x_{10}, \ldots x_{n-1}, x_{n-1}}_{n \text{ in total}}\}$$

2) compute sample expected value $\theta_1^*$

3) repeat as many as $\frac{1-2}{M\text{-times}}$ you want

4) compute final expected values

$$\theta^* = \frac{1}{M}\sum_{j=1}^{M} \theta_j^*$$

EFRON (1976)

used in small data sets.