Lecture FYS-STK3155/4155, August 26, 2024

FYS-STK 3155/4155 AUG 26

$$C(\beta) = MSE = \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \sum_{j=0}^{p-1}\beta_j x_i^j\right)^2$$

$$\tilde{y}_i = \sum_{j=0}^{p-1}\beta_j x_i^j \qquad \text{(Model)}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$\uparrow$ optimal

X orthogonal $X^T X = X X^T = I$

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

$$\hat{y} = X\hat{\beta} = X\underbrace{(X^T X)^{-1} X^T}_{1} y$$

$$= y$$

$$A = X(X^T X)^{-1} X^T$$

$A^2 = A$ (projection matrix)

$$\tilde{y} = Ay$$

$\tilde{y}$ can be viewed as a projection of y in terms of the column vectors of X

$$X = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{p-1} \\ 1 & x_1 & & & \\ \vdots & \vdots & & & \\ 1 & x_{n-1} & \cdots & & x_{n-1}^{p-1} \end{bmatrix}$$

$$\frac{\partial C}{\partial \beta} = 0 = -\frac{2}{n} X^T (y - X\beta)$$

Ex 1

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = \frac{2}{n} \underbrace{X^T X}_{\text{Hessian}}$$

(MSE)

SVD of a matrix

Decompose Matrix X

$$X = U \Sigma V^T$$

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

$$U \in \mathbb{R}^{n \times n} \qquad U U^T = U^T U = \mathbf{1}$$

$$V \in \mathbb{R}^{p \times p} \qquad V V^T = V^T V = \mathbf{1}$$

$$U = [u_0\ u_1\ \cdots\ u_{m-1}]$$

$$u_i^T u_j = \delta_{ij}$$

$$V = [v_0\ v_1\ \cdots\ v_{p-1}]$$

$$v_i^T v_j = \delta_{ij}$$

$$\Sigma \in \mathbb{R}^{m \times p} = \begin{bmatrix} \sigma_0 & & 0 \\ & \sigma_1 & \\ 0 & & \ddots \\ & & & \sigma_{p-1} \\ & & 0 & \end{bmatrix}$$

$$\sigma_0 > \sigma_1 > \sigma_2 \cdots > \sigma_{p-1} > 0$$

Example 3×2

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\Sigma^T \Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$\Sigma \Sigma^T = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

2nd derivation

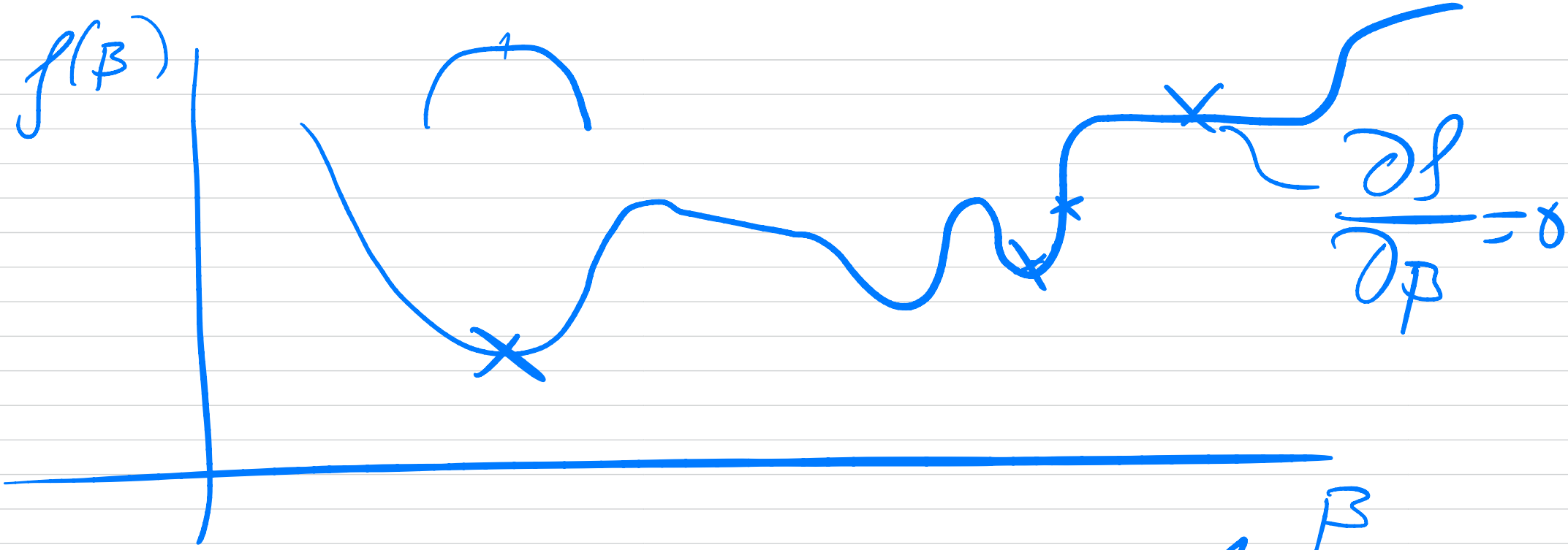
$$\underbrace{X^T X}_{p \times p} = V \Sigma^T \underbrace{U^T U}_{1_{m \times m}} \Sigma V^T$$

$p \times p$ $\quad p \times m$ $\quad n \times p$ $\quad p \times p$

$$= V \Sigma^T \Sigma V^T$$

$$X^T X V = V \Sigma^T \Sigma \underbrace{V^T V}_{I_{p \times p}}$$

$$X^T X v_i = v_i \sigma_i^2 \Rightarrow \text{convex optimization}$$



$$\hat{y} = X\beta = X(X^T X)^{-1} X^T y$$

with SVD

$$X = U \Sigma V^T$$

$$\tilde{y}_{OLS} = \underbrace{U \Sigma V^T}_{n \times p} \left(\underbrace{V \Sigma^T \Sigma V^T}_{p \times p} \right)^{-1} \underbrace{V \Sigma^T U^T y}_{p \times n}$$

$\uparrow$ n $\qquad$ n

$$\frac{1}{V \Sigma^T \Sigma V^T}$$

A, B are square invertible matrices

$$\frac{1}{AB} = A^{-1} B^{-1}$$

$$\tilde{y} = \left(\sum_{j=0}^{p-1} u_j u_j^T \right) y$$

if $\forall \sigma_i > 0$

$$\tilde{y} = \sum_{j=0}^{n-1} u_j u_j^T y$$

$$= \mathbb{1}$$

$$\tilde{y} = y$$

Why interesting?

our Next Method:

Ridge Regression.

Trick in case you cannot invert $X^T X$

$$X^T X \Rightarrow X^T X + \lambda \mathbb{I}_{p \times p}$$

$$\hat{\beta}_{Ridge} = \left(X^T X + \lambda \mathbb{I}\right)^{-1} X^T y$$

$$\lambda > 0$$

Another cost function

$$C_{Ridge} = \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \sum_{j=0}^{p-1}\beta_j x_i^j\right)^2 + \lambda \sum_{j=0}^{p-1}\beta_j^2$$

↑ Hyperparameter

constraint

$$\sum_{j=0}^{p-1}\beta_j^2 < t < \infty$$

$$C_{LASSO} = \frac{1}{n} \sum_{i=0}^{n-1} \left(y_i - \sum_{j=0}^{p-1} \beta_j x_i^j \right)^2$$

$$+ \underbrace{\lambda \sum_{j=0}^{p-1} |\beta_j|}_{\lambda \|\beta\|_1}$$

$$\frac{\partial C_{Ridge}}{\partial \beta} = 0 = -\frac{2}{n} X^T (y - X\beta) - 2\lambda\beta$$

$$-X^T(y - X\beta) - \underbrace{n\lambda\beta}_{\lambda \Rightarrow \lambda}$$

$$\beta_{Ridge} = \left(X^T X + \lambda I\right)^{-1}_{p \times p} X^T y$$

Lasso

$$\frac{d|\beta|}{d\beta} = \begin{cases} +1 & \beta > 0 \\ -1 & \beta < 0 \end{cases}$$

$$\tilde{y}_{Ridge} = X\underbrace{\left(X^T X + \lambda I\right)^{-1} X^T}_{U \Sigma V^T} y$$

SVD

$$= \left[\sum_{j=0}^{p-1} \left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right) u_j u_j^T\right] y$$

$$\sigma_0^2 > \sigma_1^2 - - - > \sigma_{p-1}^2 > 0$$

with $\downarrow$ small and λ, we can suppress (reduce) the role of a specific component u_j $\Rightarrow$ reduction of degrees of freedom.