FYS-STK3155/4155 lecture
September 1, 2025

FYS-STK3155/4155 lecture September 1, 2025

Analyzing Ordinary least squares and Ridge Regression.

$$X^T X = X X^T = \mathbb{1} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & \ddots \; 1 \end{bmatrix}$$

$$\tilde{y} = X \underbrace{(X^T X)}_{\mathbb{1}}^{-1} X^T y$$

$$= y$$

SVD

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

$X = U \Sigma V^T$

$U \in \mathbb{R}^{n \times n}$

$UU^T = U^T U = \mathbb{1}$

$$U = \begin{bmatrix} \vec{u}_0 & \vec{u}_1 & \ldots & \vec{u}_{n-1} \\ & & & \end{bmatrix}$$

$\vec{u}_i^T \vec{u}_j = \delta_{ij}$

$$V \in \mathbb{R}^{p \times p}$$

$$V^T V = V V^T = \mathbb{1}$$

$$V = \begin{bmatrix} \vec{v}_0 & \vec{v}_1 & \cdots & \vec{v}_{p-1} \end{bmatrix}$$

$$\vec{v}_i^{\,T} \vec{v}_j = \delta_{i,j}$$

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

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

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

$\sigma_0 = 2 \quad 1 \quad \sigma_1 = \underline{1}$

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

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

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

$$= V \underbrace{\Sigma^T \Sigma}_{\begin{bmatrix} \sigma_0^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{p-1}^2 \end{bmatrix}} V^T$$

$$\left([V, \Sigma^T\Sigma] = 0 \right) \Rightarrow V\Sigma^T\Sigma$$

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

$$\boxed{\overline{X^T X} = \left(\overline{\Sigma^T \Sigma}\right) V \overline{V^T}} \cdot V$$

$$\overline{V^T} V = \mathbb{1}$$

$$\left(\overline{X^T X}\right) V = \overline{\Sigma^T \Sigma}\, V$$

$$V = \begin{bmatrix} \vec{v}_0 & \vec{v}_1 & \cdots & \vec{v}_{D-1} \\ & & & \end{bmatrix}$$

$$= \underbrace{\left(\overline{X^T X}\right)}_{A} \vec{v}_i = \sigma_i^2 \vec{v}_i$$

$$A \quad \vec{v}_i' = \lambda_i \vec{v}_i'$$

$$(X^T X)\vec{v}_i = \sigma_i^2 \vec{v}_i$$

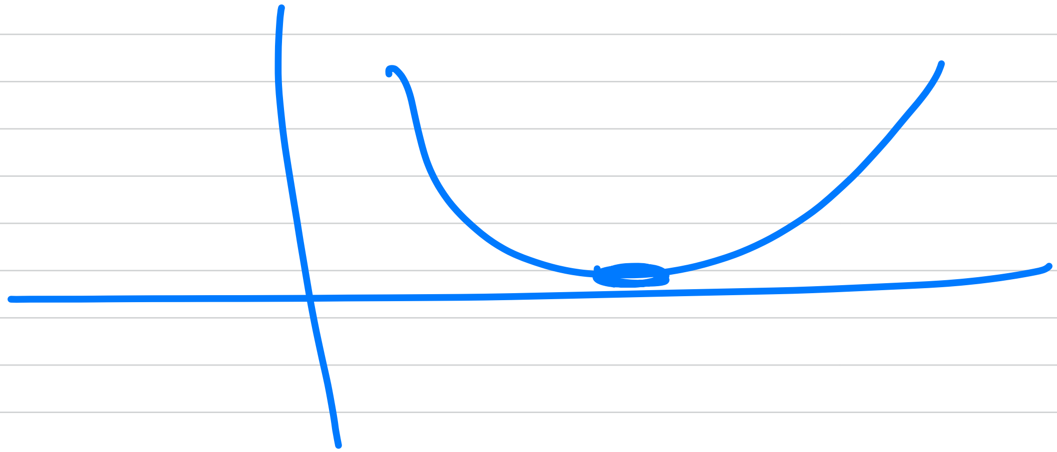
$$\sigma_i' \geq 0$$

MSE

$$C(\theta) = \frac{1}{n}(X\theta - y)^T(X\theta - y)$$

$$\frac{\partial C}{\partial \theta} \qquad \frac{\partial^2 C}{\partial \theta^2} = \frac{2}{n} X^T X$$

$x^2 + 1 = f(x)$

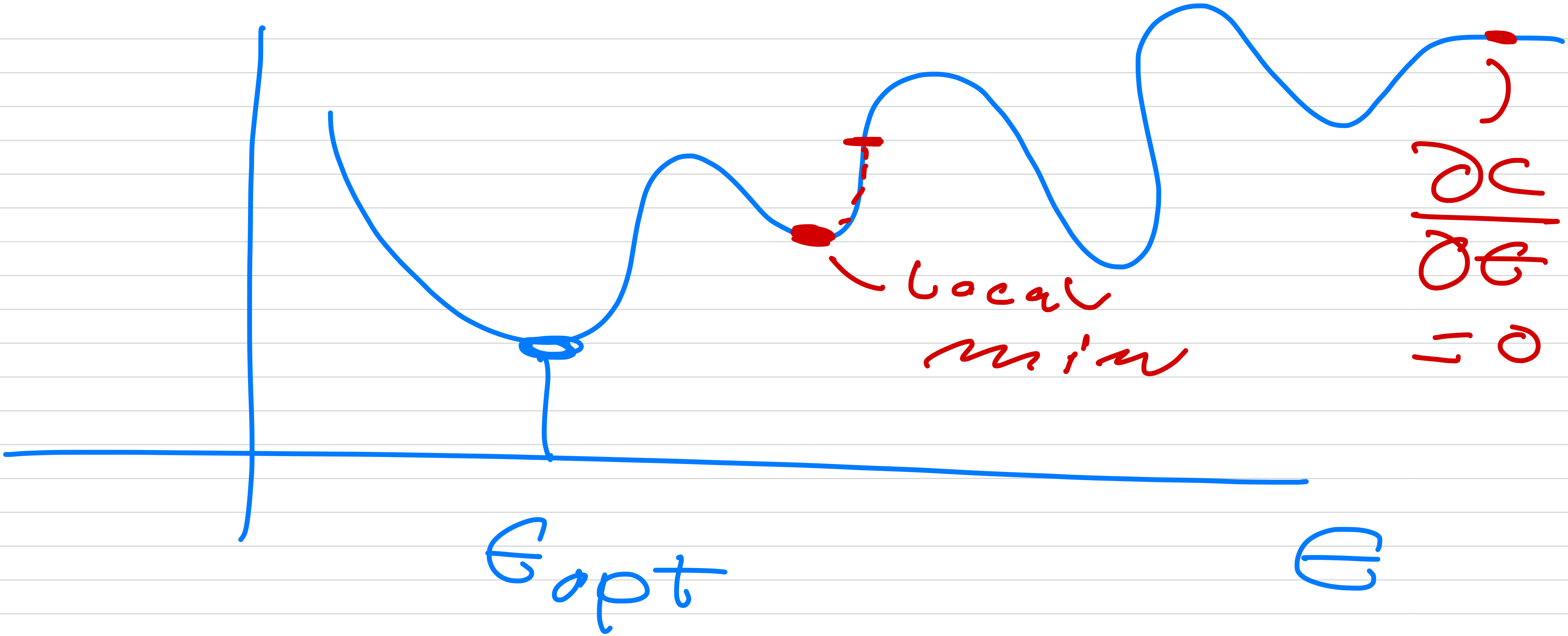


$$\frac{\partial^2 f}{\partial x^2} = 2 > 0$$

Ridge :

$$\frac{\partial^2 C_{Ridge}}{\partial \theta^2} = 2\left(\frac{x^T x}{n} + \lambda\right)$$

$\lambda \geq 0$


local min
$\frac{\partial C}{\partial \theta} = 0$
θ_{opt}
θ

SVD, OLS & Ridge

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

$$\hat{y} = X\hat{\theta} =$$

$$\cancel{X}(X^T X)^{-1} X^T y$$

$$= \left\{ U\Sigma V^T \left(V\Sigma^T U^T U \Sigma V^T \right)^{-1} \right.$$

$$\left. \times V\Sigma^T U^T \right\} y$$

$$\left(V \Sigma^T \Sigma V^T\right)^{-1}$$

A, B are square matrices and invertible

$$(AB)^{-1} = B^{-1} A^{-1}$$

$$\tilde{y} = \left(U \Sigma \left(\Sigma^T \Sigma\right)^{-1} \Sigma^T U^T\right) y$$

$$= \left(\sum_{i=0}^{p-1} u_i u_i^T\right) y$$

$$\tilde{y}_{Ridge} = \left(\sum_{i=0}^{p-1} u_i u_i^T \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \right) y$$

$$\theta^T = [\theta_0, \theta_1, \ldots, \theta_{p-1}]$$

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

$$\sigma_i \geq 0$$

$\lambda \to \infty$ (or large)

Degrees of freedom through u_i will be suppressed.

Taylor - expansion of

$c(\theta)$ around $\hat{\theta}$

$\hat{\theta} - \theta^{(n)}$

$$\frac{dc}{d\theta} = g \quad \text{at } \theta = \theta^{(n)}$$

$$c(\hat{\theta}) = c(\theta^{(n)}) + g(\theta^{(n)})(\hat{\theta} - \theta^{(n)}) + \frac{1}{2}\frac{d^2c}{d\theta^2}(\hat{\theta} - \hat{\theta}^{(n)})^2 + \dots$$

Truncate at $\dfrac{d^2c}{d\theta^2}\Big|_{\theta=\theta^{(n)}}$

$$c(\epsilon^1) \simeq c(\epsilon^{(n)}) + \underbrace{\frac{dc}{d\epsilon}\Big|_{\epsilon=\epsilon^{(n)}}}_{g(\epsilon^{(n)})} \underbrace{(\epsilon^1 - \epsilon^{(n)})}_{\delta^{(n)}} + \frac{1}{2} \underbrace{\frac{d^2c}{d\epsilon^2}}_{H(\epsilon^{(n)})}\Big|_{\epsilon=\epsilon^{(n)}} (\epsilon^1 - \epsilon^{(n)})^2$$

$$= c\left(\epsilon^{(n)}\right) + g^{(n)} b^{(n)}$$

$$+ \frac{1}{2}\left(b^{(n)}\right)^2 H^{(n)}$$

$$\frac{dc}{db} = g^{(n)} + b^{(n)} H^{(n)}$$

$$= 0 \implies$$

$$b^{(n)} = \hat{\epsilon} - \epsilon^{(n)} = -\left(H^{(n)}\right)^{-1} g^{(n)}$$

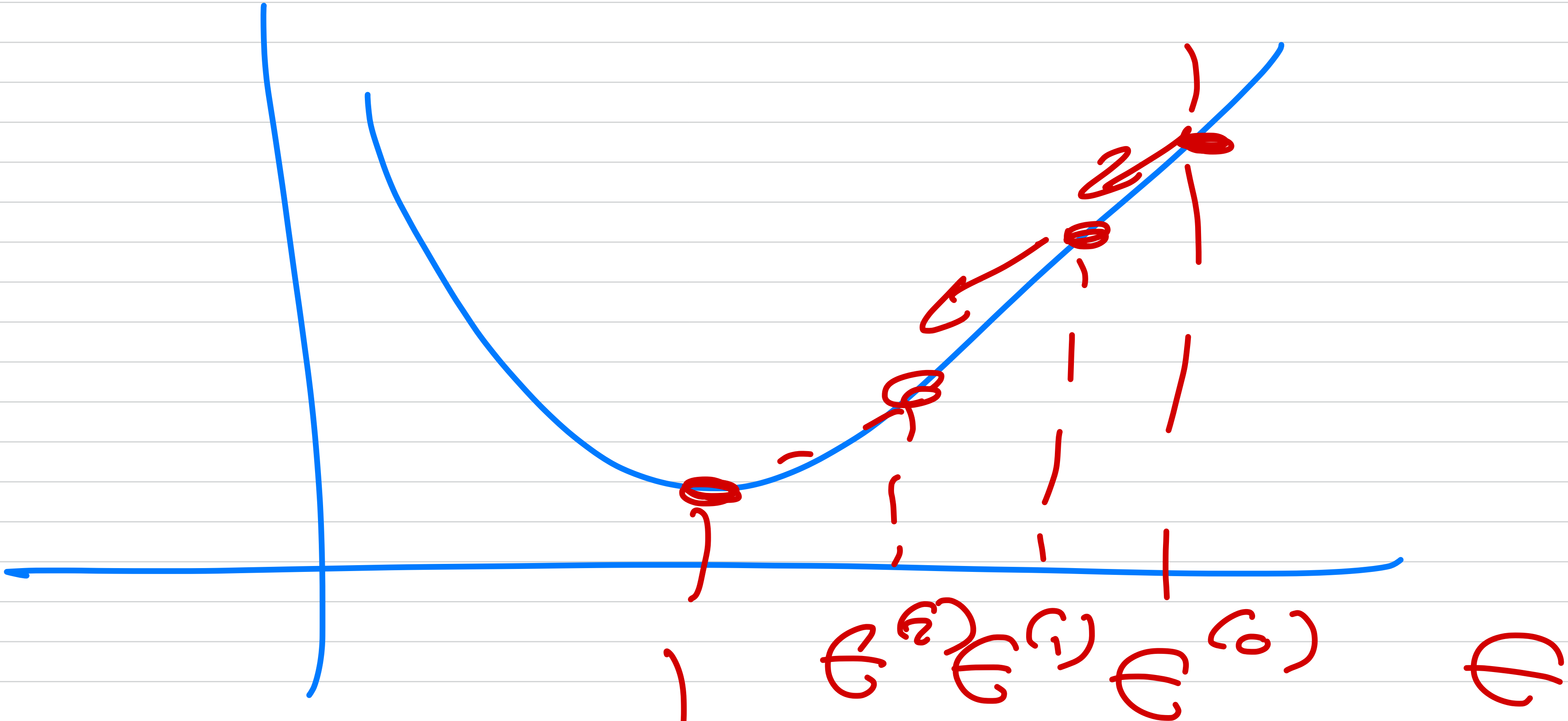
$$\hat{\epsilon} = \epsilon^{(n)} - \left(H^{(n)}\right)^{-1} g^{(n)}$$

$$\hat{\epsilon} \to \epsilon^{(n+1)} = \epsilon^{(n)} - \left(H^{(n)}\right)^{-1} g^{(n)}$$

recipe: start iteration with a guess $\epsilon^{(0)}$

keep iterating till

$$\left| \epsilon^{(n+1)} - \epsilon^{(n)} \right| \leq \varepsilon$$



$$\hat{\theta} = \theta^{(n+1)}$$

$$\boxed{\theta^{(n+1)} = \theta^{(n)} - \left(H^{(n)}\right)^{-1} g^{(n)}}$$

OLS

$$\nabla_\theta C(\theta) = \frac{2}{n}\left(X^T X\theta - X^T y\right)$$

$$\nabla^2_\theta C(\theta) = \frac{2}{n} X^T X = H$$

independent of θ

$$\theta^{(n+1)} = \theta^{(n)} - \left(H(\theta^{(n)})\right)^{-1} \nabla_\theta C(\theta^{(n)})$$

vector

H = Hessian matrix

$$H = \begin{bmatrix} \frac{\partial^2 C}{\partial \theta_0^2} & \frac{\partial^2 C}{\partial \theta_0 \partial \theta_1} & \cdots & \frac{\partial^2 C}{\partial \theta_0 \partial \theta_{p-1}} \\ \vdots & & & \\ \frac{\partial^2 C}{\partial \theta_{p-1} \theta_0} & \cdots & & \frac{\partial^2 C}{\partial \theta_{p-1}^2} \end{bmatrix}$$

$$\theta^{(n+1)} = \theta^{(n)} - \left(H(\theta^{(n)})\right)^{-1} \nabla_\theta C(\theta^{(n)})$$

$$\theta^{(n+1)} = \theta^{(n)} - \gamma \nabla_\theta \left(\ell\left(\theta^{(n)}\right) \right)$$

learning rate

constant γ : GD

$\gamma^{(n)}$: $\begin{cases} \text{momentum} \\ \text{ADAM} \\ \text{RMSprop} \\ \text{Adagrad} \\ \vdots \end{cases}$