

FYS-STK3155/4155,
Lecture September 15, 2025

FYS-STK3155/4155 September 15

Statistical interpretations
(Mainly OLS)

- Expectation values
- Central limit theorem
- Resampling methods and
Bias-variance tradeoff.

Basic assumptions

$$y = \underbrace{f(x)}_{\text{non-stochastic function}} + \underbrace{\Sigma}_{\text{noise}}$$

?

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

$$P_{\Sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma}$$

$$\exp\left[-x^2/2\sigma^2\right]$$

stochastic variable x

$$E[x^n] = \int_{x \in D} dx x^n P_X(x)$$

$$\left(\sum_{x_i' \in D} x_i'^n P_X(x_i') \right)$$

$$\mu_x = E[x] = \int_{x \in D} dx x P_X(x)$$

sample mean ($P_X(x)$ is unknown)

$$\bar{\mu}_x = \frac{1}{n} \sum_{i=1}^n x_i' \neq \mu_x$$

$$\text{var}[x] = E[x^2] - \mu_x^2$$

$$= \int_{x \in D} P_x(x) dx (x - \mu_x)^2$$

sample variance

$$\overline{\Delta_x}^2 = \frac{1}{n} \sum_{i=1}^n (x_i' - \bar{\mu}_x)^2$$

$$\neq \text{var}[x]$$

Design matrix

$$X = [\vec{x}_0 \vec{x}_1 \dots \vec{x}_{p-1}]$$

$$\text{cov}(\vec{x}_i, \vec{x}_j) =$$

$$\frac{1}{n} \sum_{i=0}^n \underbrace{(x_{ij} - \bar{\mu}_{x_j})(x_{ii} - \bar{\mu}_{x_i})}_{\text{number}}$$

$$\Rightarrow C[X] = \frac{1}{n} X' X$$

$$\underline{\text{SVD}} \quad X^T X v_i' = \sigma_i^2 v_i'$$

$$V = [\vec{v}_0 \vec{v}_1 \dots \vec{v}_{p-1}]$$

$$\tilde{y} = \underbrace{X \cdot \epsilon}_{\text{non-stochastic}} \simeq f(x)$$

$$y_i = \underbrace{\sum_{j=0}^{p-1} x_{ij}' \epsilon_j}_{X_i' * \epsilon} + \epsilon_i'$$

$$E[y_i] = ?$$

$$= E[x_i^* \epsilon_i] + E[\epsilon_i]$$

≈ 0

what does $P_Y(y)$ look like?

$$E[x_i^* \epsilon_i] = \int_{y \in D} P_Y(y) x_i^* \epsilon_i \times dy$$

$$\int_{y \in D} P_Y(y) dy = 1$$

$$\begin{aligned}
 E[y_i] &= X_i \theta \\
 &= \sum_{j=0}^{p-1} x_{ij} \theta_j
 \end{aligned}$$

$$E[y] = X\theta$$

Assumption

$$\begin{aligned}
 y_i &\sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(y_i - X_i \theta)^2}{2\sigma^2} \right] \\
 &\sim N(X_i \theta, \sigma^2)
 \end{aligned}$$

$$\text{var}[y_i] = \sigma^2$$

Can we derive OLS eq
using

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

$$= P_i$$

$$P(\vec{S} \times | \theta) = \prod_{i=0}^{n-1} P_i$$

$$y_i \sim i.i.d.$$

$$\hat{e} = \underset{e \in \mathbb{R}^D}{\operatorname{argmax}} P(\vec{S} \times | e)$$

$$\vec{\nabla}_e P(\vec{S} \times | e) = 0$$

$$C(\theta) = -\log P(\vec{y} | \theta)$$

$$= - \sum_{i=0}^{n-1} \log(y_i | \theta)$$

$$= + \frac{n}{2} \log 2\pi\sigma^2 + \sum_{i=0}^{n-1} \left(\frac{y_i - x_i^* \theta}{\sigma^2} \right)^2$$

$$\nabla_{\theta} C(\theta) = 0$$

$$C(\theta) = \frac{n}{2} \log 2\pi\sigma^2 + \frac{1}{2\sigma^2} \|g - X\theta\|_2^2$$

$$\nabla_{\theta} C(\theta) = 0 \Rightarrow$$

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

Resampling with Bootstrap

Dæmain

$$D = [z_0 z_1 \dots z_{n-1}]$$

$$\bar{\mu}_z = \frac{1}{n} \sum_{i=0}^{n-1} z_i \quad (\neq \mu_z)$$

$$\text{var}[z] \neq \frac{1}{n} \sum_{i=0}^{n-1} (z_i - \bar{\mu}_z)^2$$

with replacement

Reshuffle data randomly

$$D' = [z'_0 z'_1 \dots z'_i \dots z'_{n-1}]$$

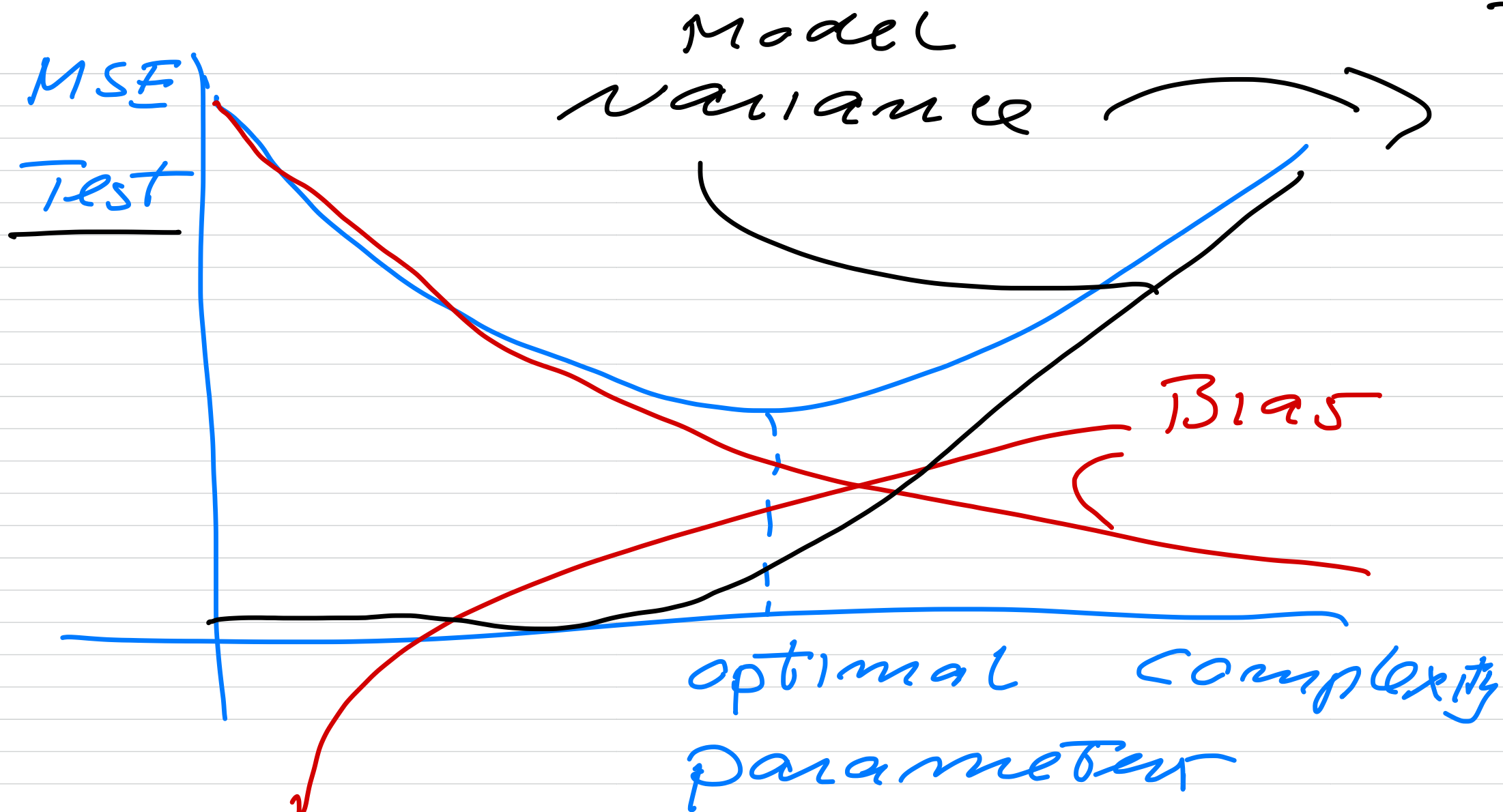
$z'_0 = z_0$ $z'_i = \bar{\mu}_z^2$ $z'_0 = z_0$

Repeat B -times

$$\bar{\mu}_B = \frac{1}{B} \sum_{b=1}^B \bar{\mu}_b$$

hope is that this gets
closer to true mean

$$\bar{\sigma}_B^2 = \frac{1}{B} \sum_{b=1}^B (\bar{\mu}_b - \bar{\mu}_B)^2$$



$$\frac{1}{n} \sum_{i=0}^{n-1} (y_i - E[\hat{y}])^2$$

(f_i)

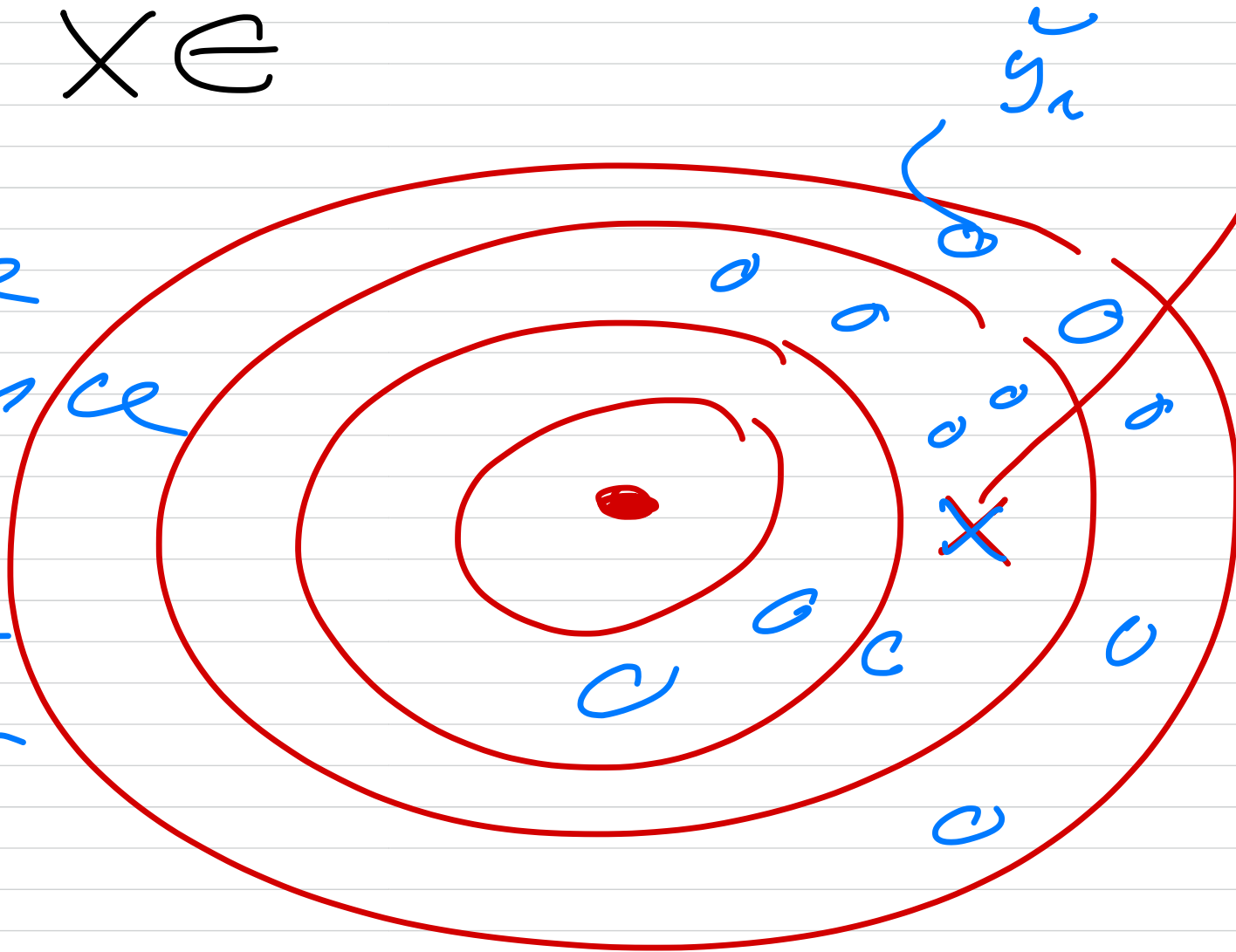
$$\text{var} [\tilde{y}] = \frac{1}{n} \sum_{i=0}^{n-1} (\tilde{y}_i - E[\tilde{y}])^2$$

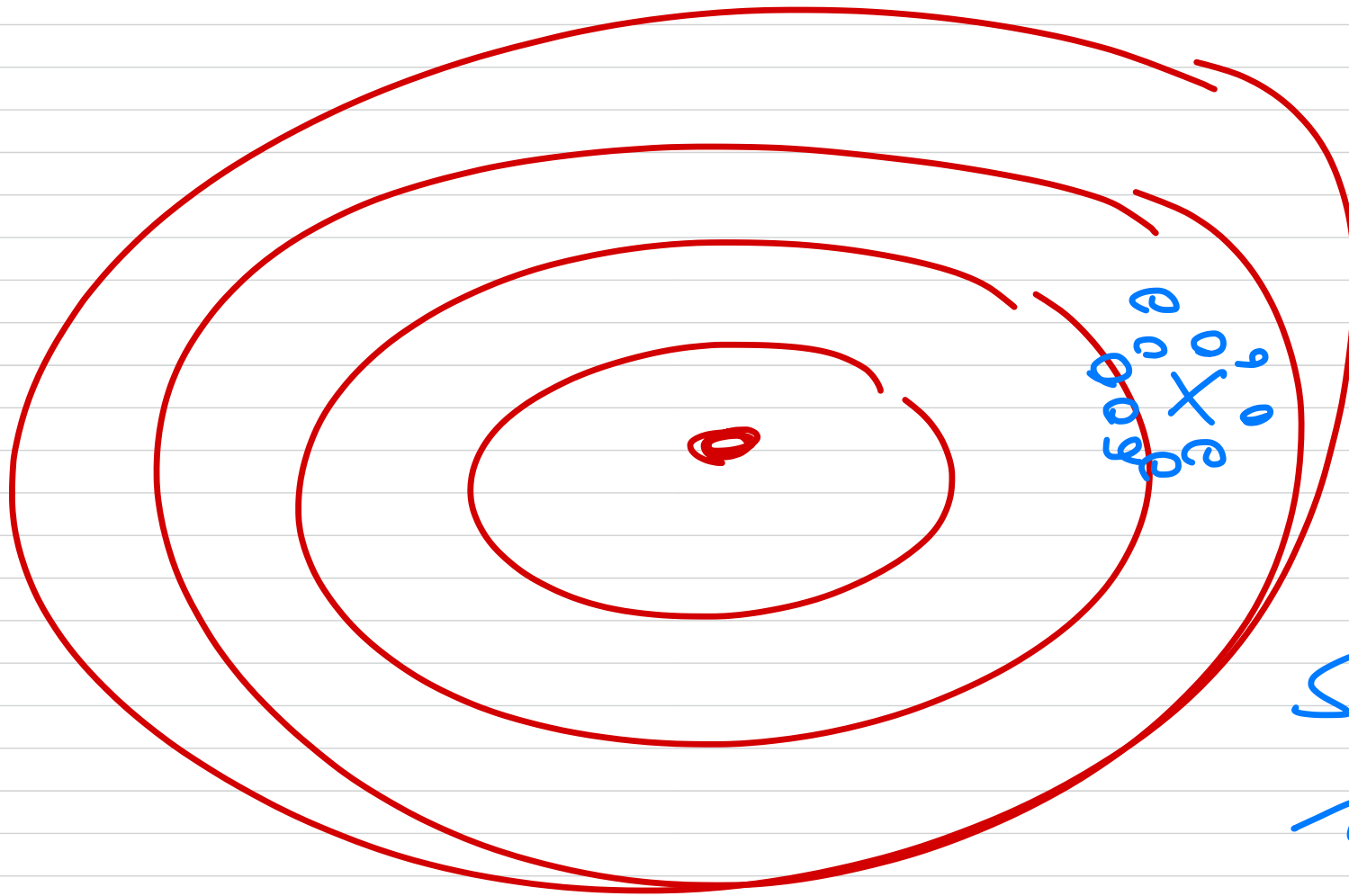
$$\tilde{y} = X \hat{\beta}$$

Large
variance

Large

Bias



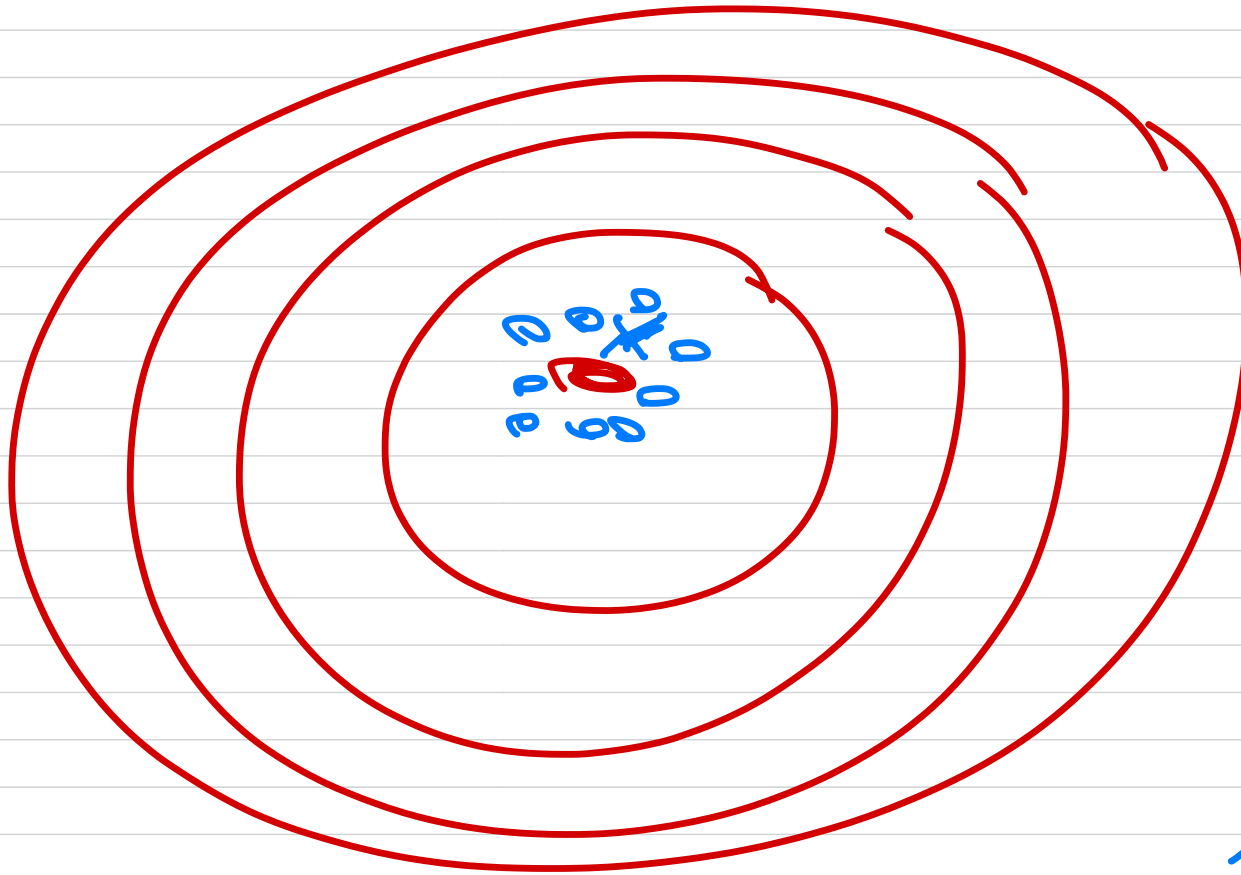


90%
Confidence
Interval

Small
variance

Large
Bias

ideal



Low
variance

Low
Bias