Lecture FYS-STK, September 14, 2023

Resampling (Bootstrap, cross-validation)

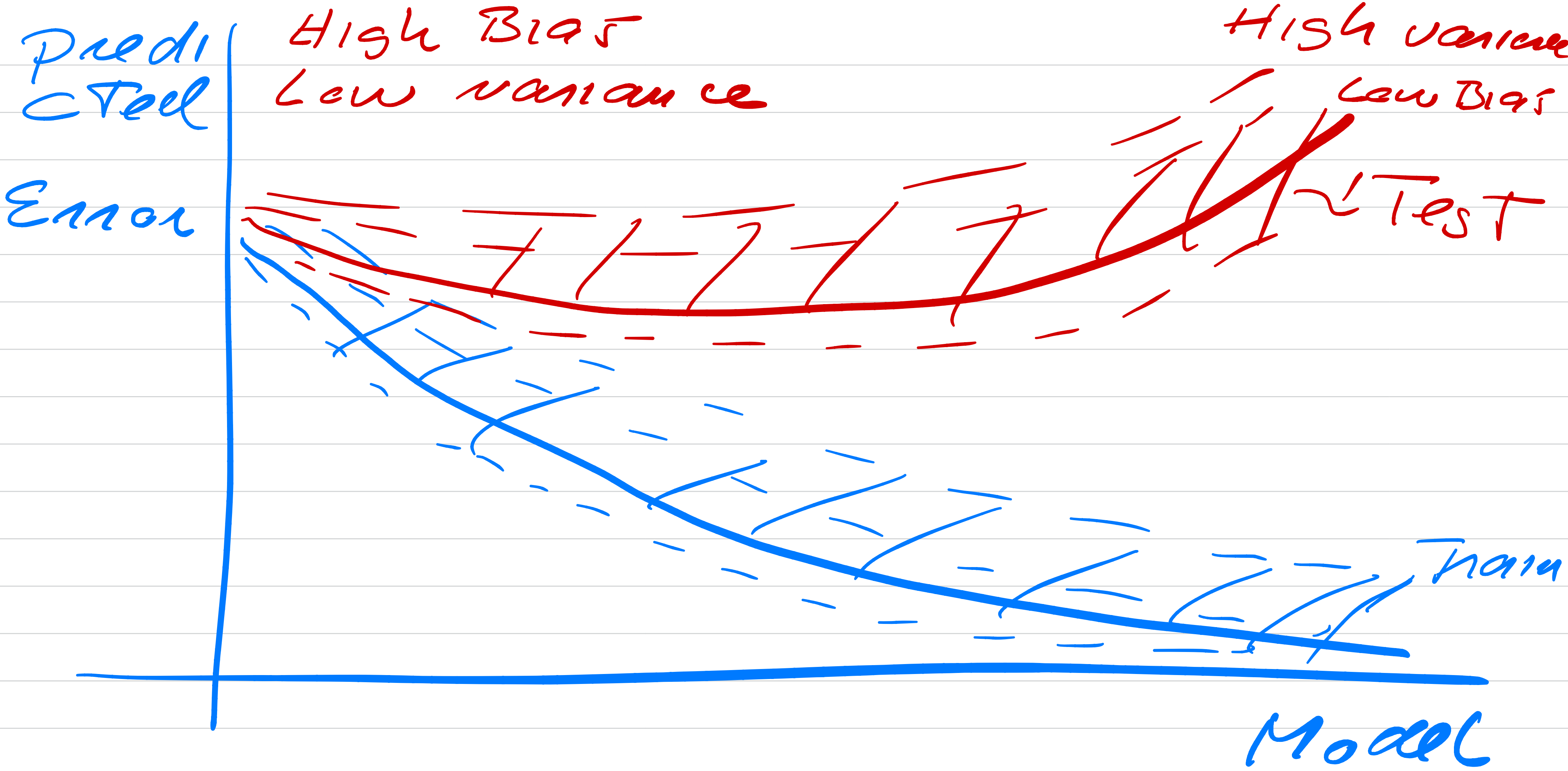
We have defined functions to assess model, MSE

$$\mathcal{L}(y_i, \tilde{y}_i)$$

model

with given training data $\mathcal{T}_i$ and test data $\tilde{\mathcal{T}}_i$

$$\mathrm{Err}_{Test} = \mathcal{L}(y_i, \tilde{y}_i; \mathcal{T}_i \tilde{\mathcal{T}}_i)$$



$$\mathrm{Err} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, \tilde{y}_i; T_i, \tilde{T}_i)$$

Bias-variance tradeoff for MSE

$$MSE = \frac{1}{N}\sum_{i=0}^{N-1}(y_i - \tilde{y}_i)^2$$

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

$$= \mathbb{E}\left[y^2\right] - 2\mathbb{E}\left[y\tilde{y}\right] + \mathbb{E}\left[\tilde{y}^2\right]$$

$$\mathbb{E}\left[\tilde{y}^2\right] = \mathrm{var}\left[\tilde{y}\right] + \left(\mathbb{E}\left[\tilde{y}\right]\right)^2$$

$$\mathbb{E}\left[y^2\right] = \mathbb{E}\left[(f+\varepsilon)^2\right] =$$

$$\mathbb{E}\left[f^2\right] + 2\underbrace{\mathbb{E}\left[f\varepsilon\right]}_{=0} + \underbrace{\mathbb{E}\left[\varepsilon^2\right]}_{=\sigma^2}$$

$$= f^2 + 2f \cdot 0 + \sigma^2$$

$$\mathbb{E}[y\tilde{y}] = \mathbb{E}[(f+\varepsilon)\tilde{y}]$$

$$= \mathbb{E}[f\tilde{y}] + \mathbb{E}[\varepsilon \cdot \tilde{y}]$$

$$= f\,\mathbb{E}[\tilde{y}] + \mathbb{E}[\tilde{y}]\underbrace{\mathbb{E}[\varepsilon]}_{=0}$$

collecting

$$f^2 - 2f\,\mathbb{E}[\tilde{y}] + \sigma^2 + \mathrm{Var}[\tilde{y}] + (\mathbb{E}[\tilde{y}])^2$$

$$= \mathbb{E}\left[\left(f - \mathbb{E}[\hat{y}]\right)^2\right] + \sigma^2$$

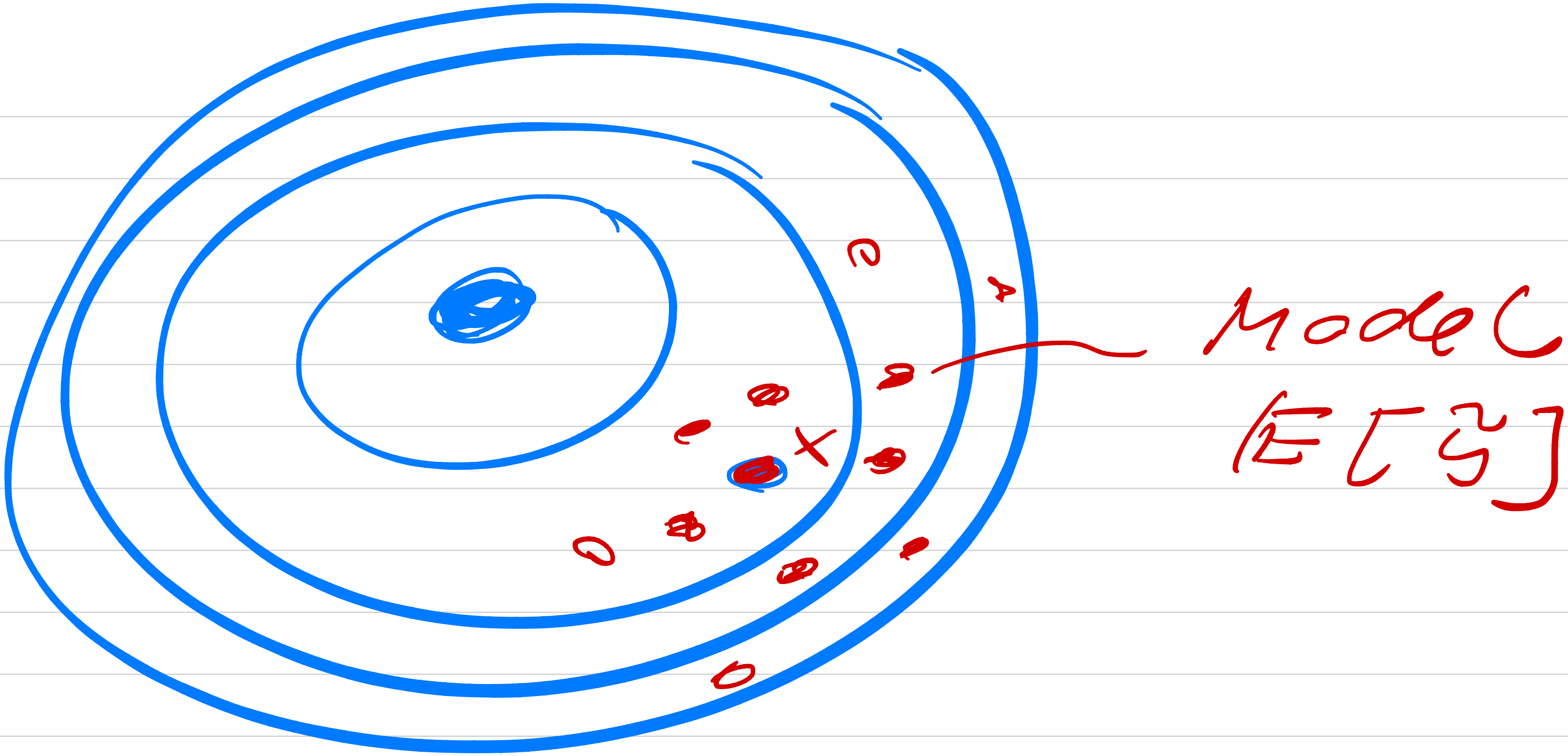
$$+ \operatorname{var}[\hat{y}]$$

Bias

variance

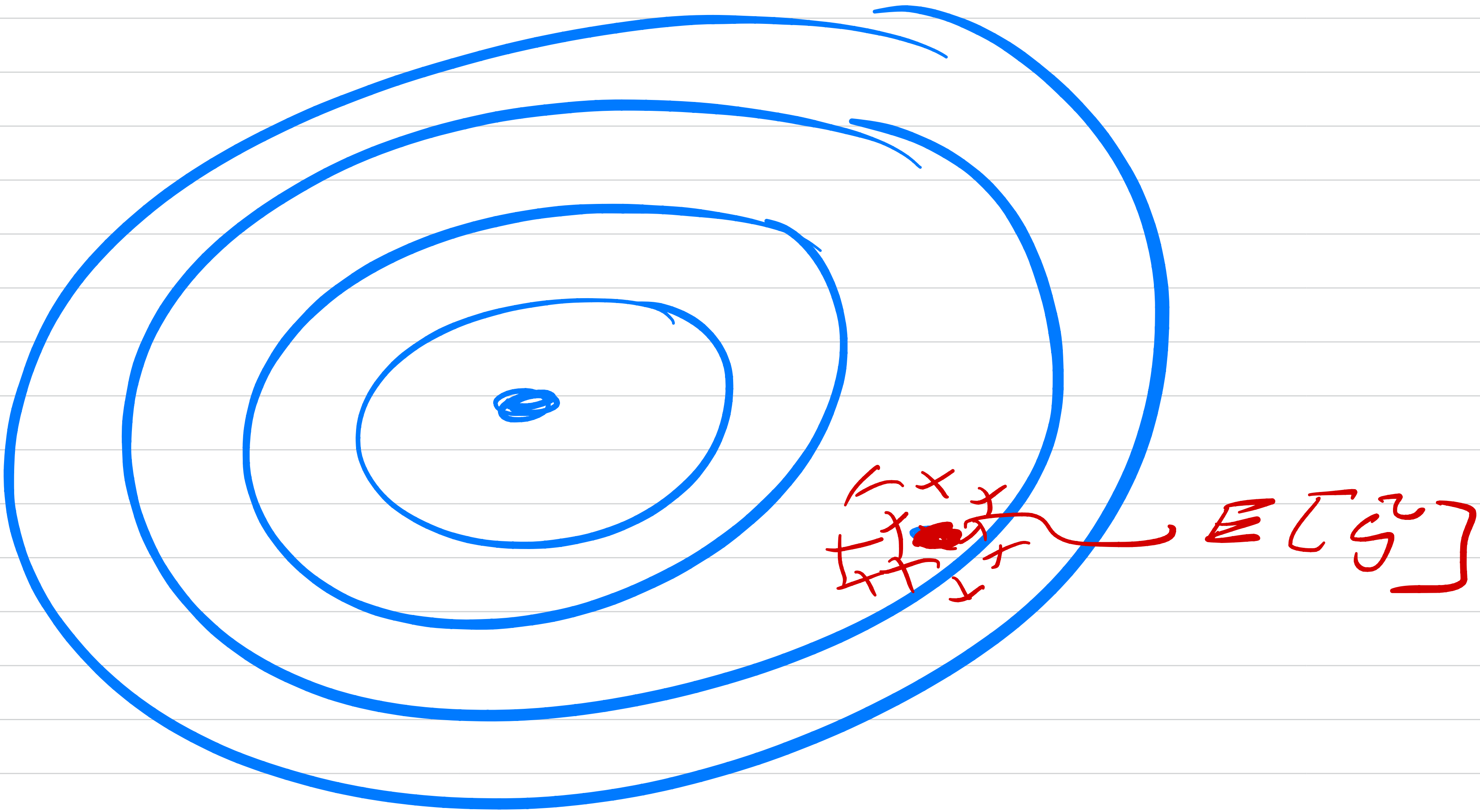
$$\text{Bias} = \frac{1}{n} \sum_{i=0}^{n-1} \left(y_i - \mathbb{E}[\hat{y}]\right)^2$$

$$\operatorname{var}[\hat{y}] = \frac{1}{n} \sum_{i=0}^{n-1} \left(\hat{y}_i - \mathbb{E}[\hat{y}]\right)^2$$



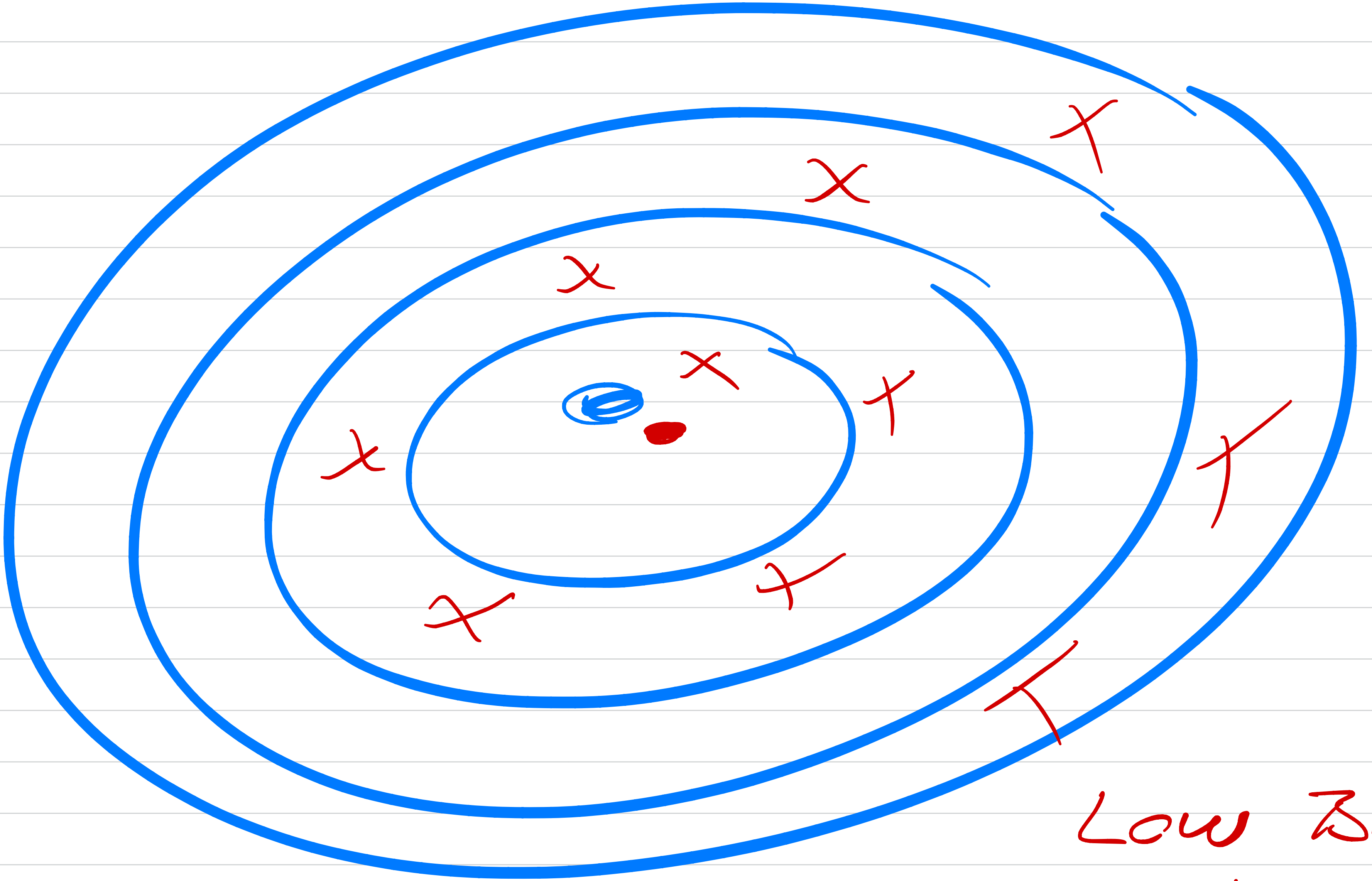
High Bias

High variance

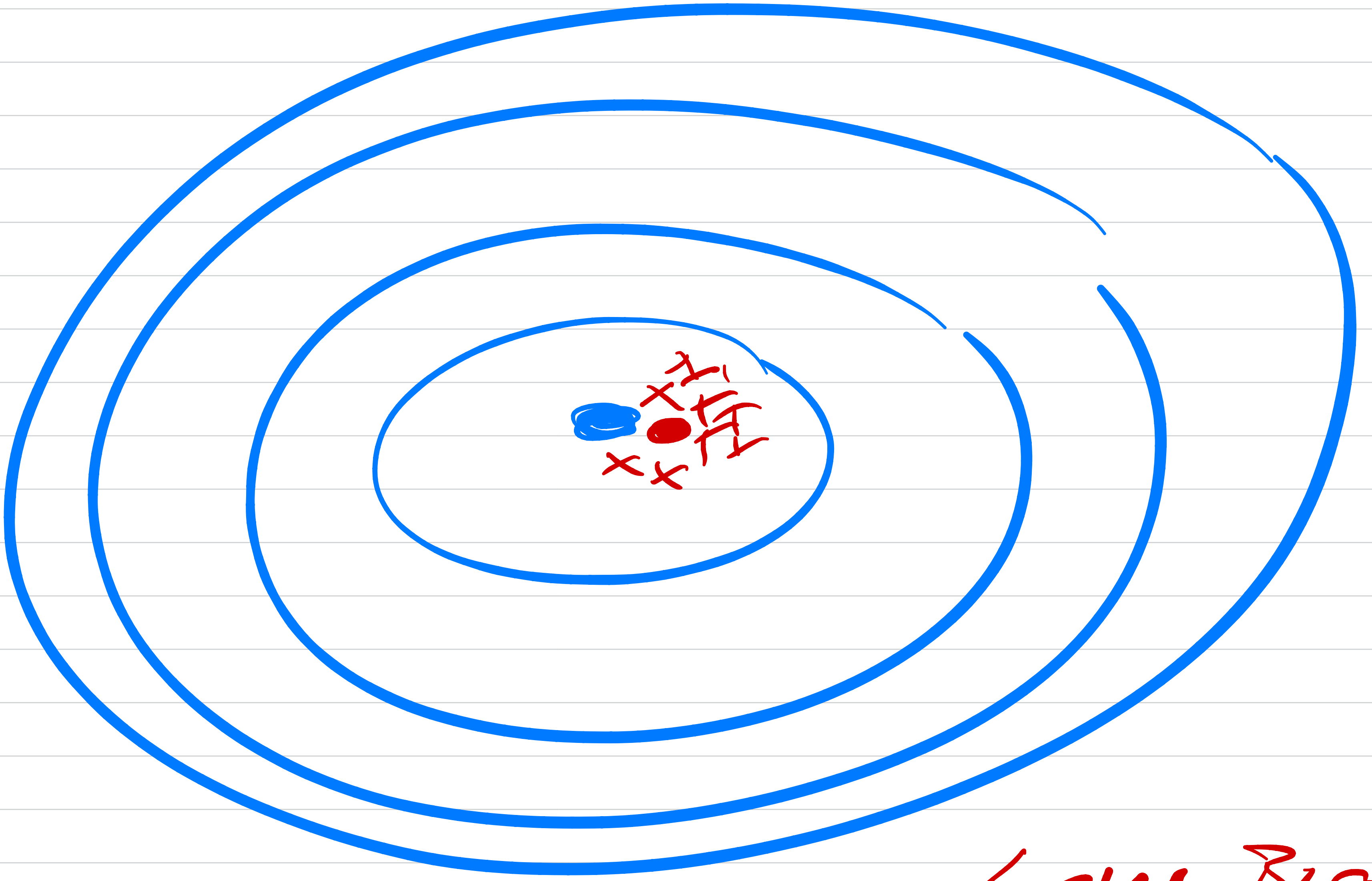


High Bias

Low variance



Low Bias
High variance



Low Bias
Low variance

Bootstrap

Original data set

$D: \{ z_1, z_2, \ldots z_N \}$

i.i.d.

(i) Draw a Bootstrap sample

$D_1^* = \{ z_1^* z_2^* \ldots z_N^* \}$

with replacement

Compute $\hat{e}_1^*$ (mean value)

(ii) Repeat B times getting estimates

$\hat{\theta}_1^* \quad \hat{\theta}_2^* \quad - - \quad \hat{\theta}_B^*$

$$\bar{\theta} = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j^*$$

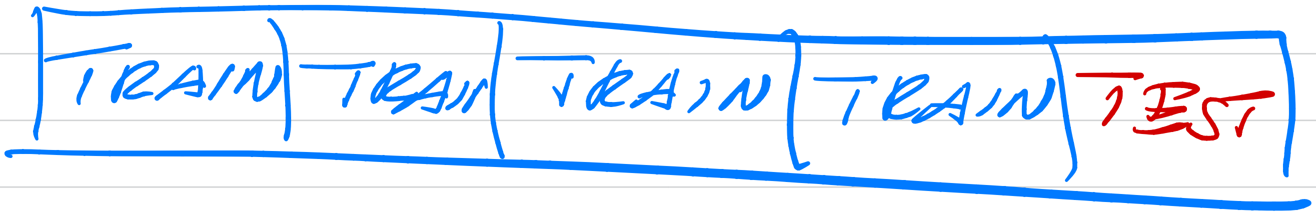
(iii) can compute variance

$$\sigma_\theta^2 = \frac{1}{B} \sum_{j=1}^{B} \left(\hat{\theta}_j^* - \bar{\theta} \right)^2$$

Resampling with cross-validation

k-Folds : splitting data in k-subsets

k = 5

1) 

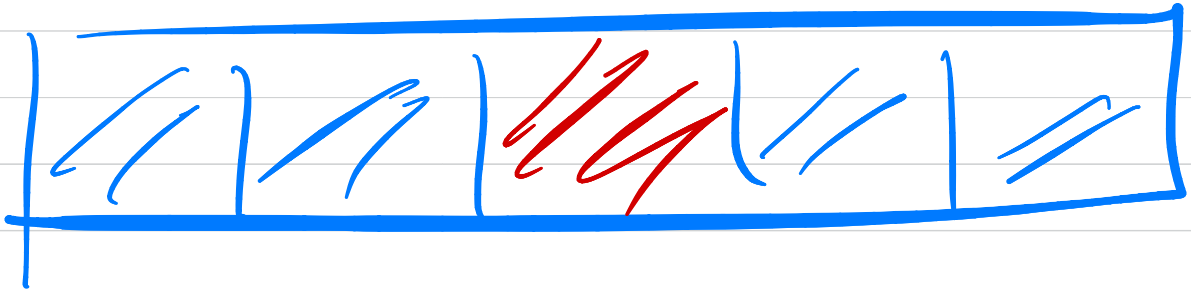
Data

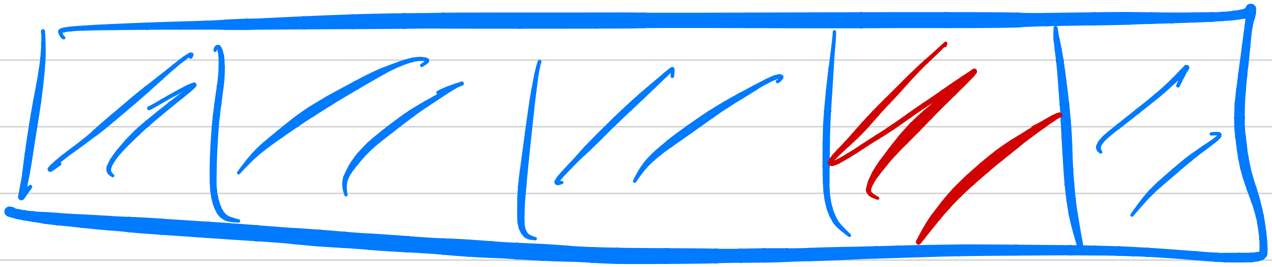
$E_{in} Test = E_{in}^{(1)}$

2) 

$E_{in}^{(2)}$

3)  $\varepsilon_{11}^{(3)}$

4)  $\varepsilon_{11}^{(4)}$

5)  $\varepsilon_{11}^{(5)}$

$$\varepsilon_{11} = \frac{1}{5}\sum_{i=1}^{5}\varepsilon_{11}^{(i)}$$

Typical values of $K \sim 5-10$

But need to test that MSE stabilizes