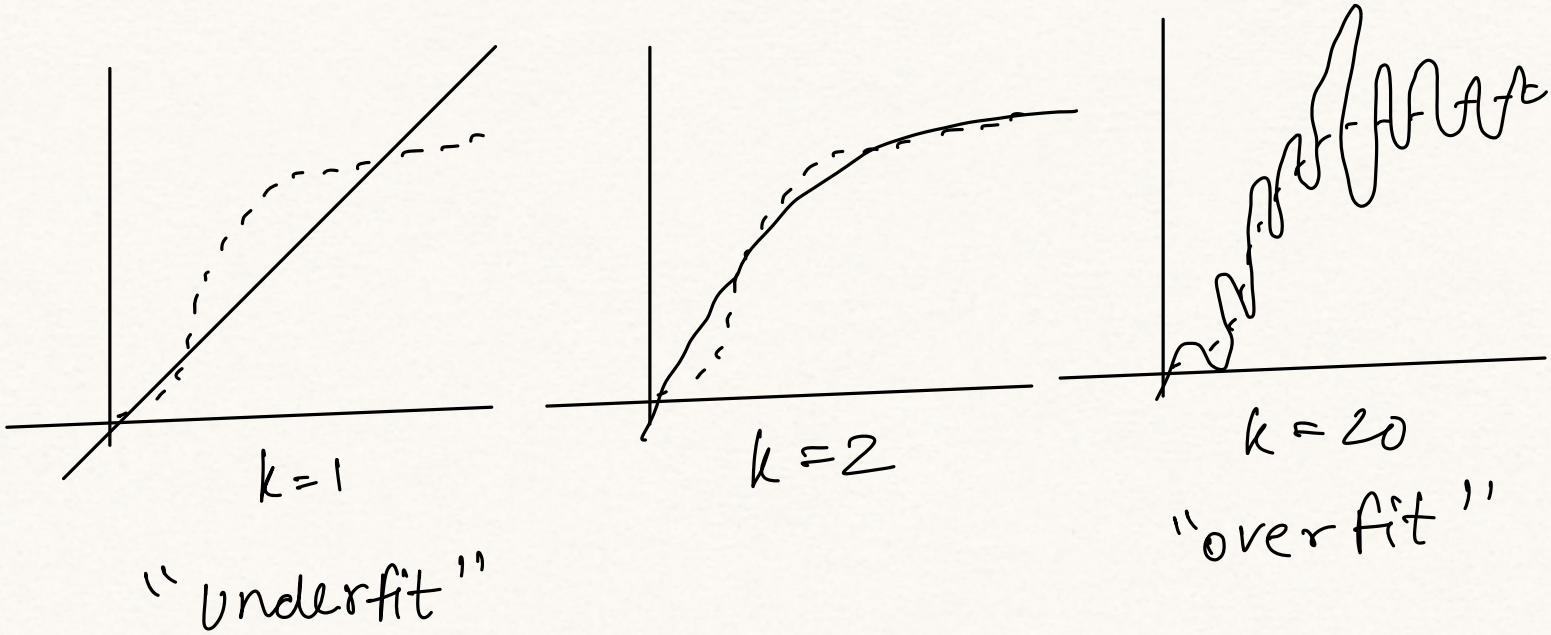
BIAS – VARIANCE

Generalization error (G.E)



"underfit" "overfit"

- Bias :- component of G.E due to "expressivity handicap"
- Variance : due to finite sample of training set.

For case of Squared Error

$$y^{(i)} = f(x^{(i)}) + \varepsilon^{(i)}$$

$$E[\varepsilon] = 0$$

$$\text{Var}[\varepsilon] = \sigma^2$$

$$\text{Test Error } [G.E] = \mathbb{E}\left[(y - \hat{f}_n(x))^2\right]$$

[expectation is over all ε in training set & test eg.]

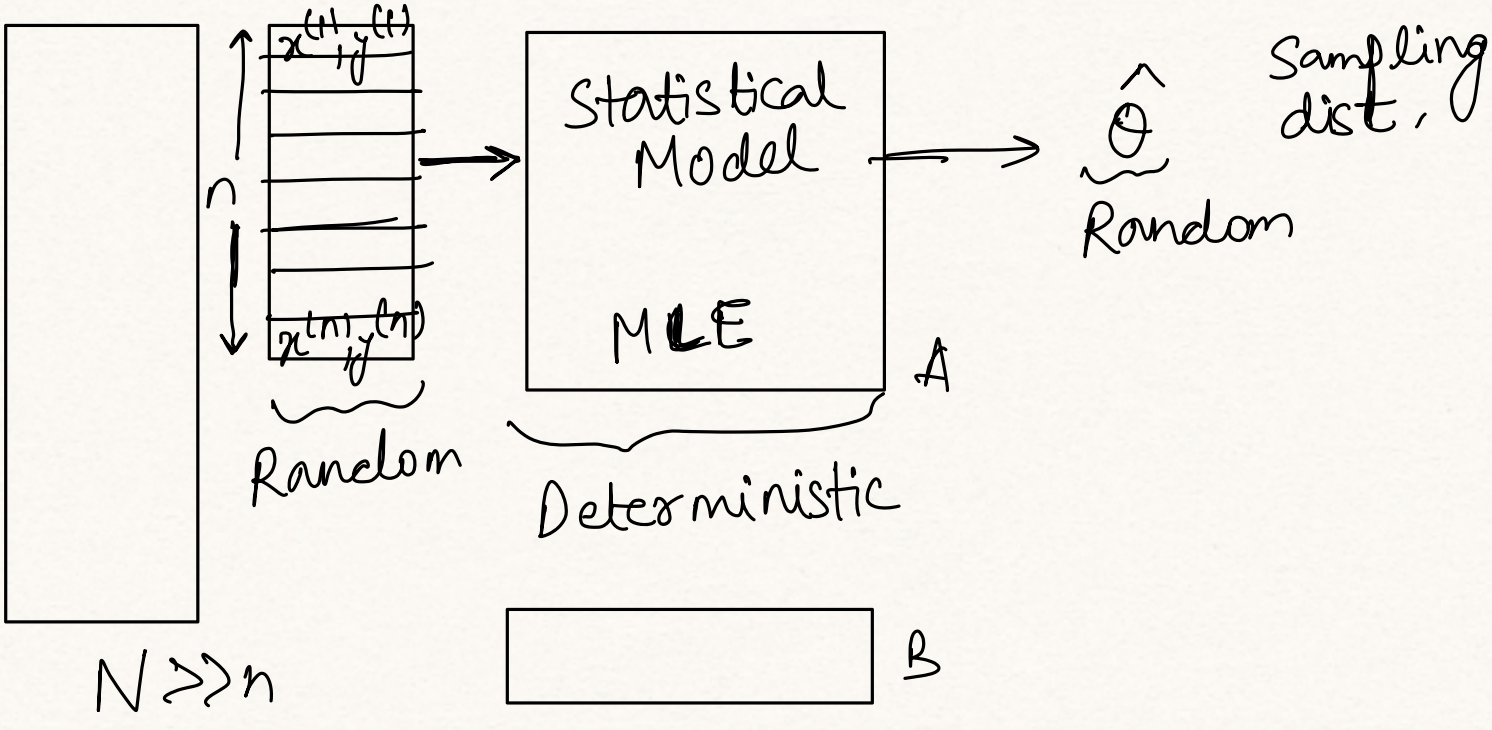
$$= \sigma^2 + \mathbb{E}\left[f(x) - \hat{f}_n(x)\right]^2 + \mathrm{Var}\left[\hat{f}_n(x)\right]$$

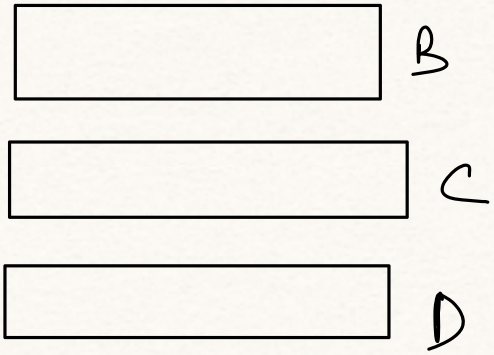
σ^2: irreducible error

$\mathbb{E}\left[f(x) - \hat{f}_n(x)\right]^2$: $(\mathrm{Bias})^2$

$\mathrm{Var}\left[\hat{f}_n(x)\right]$: variance

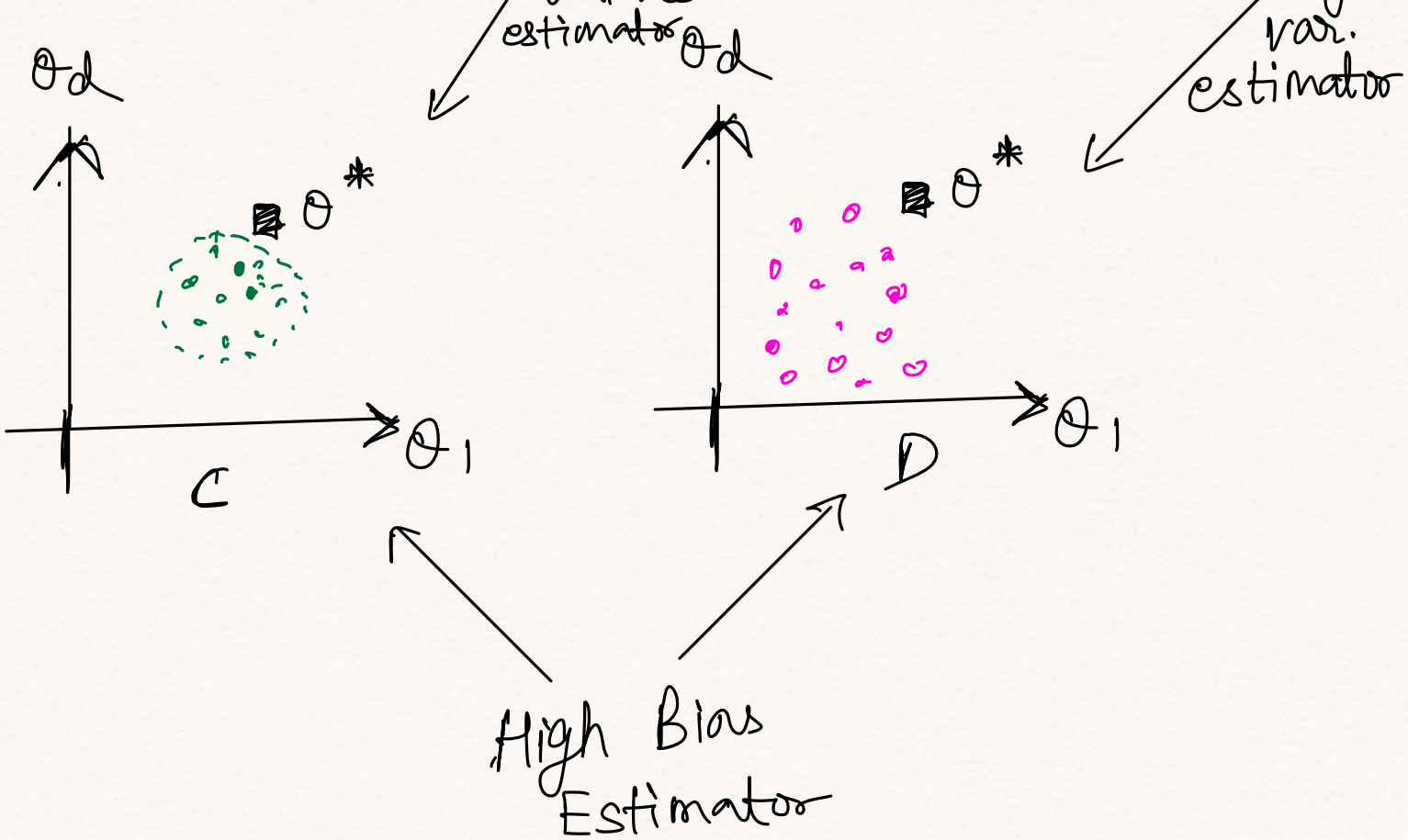
$(x, y) \sim \mathrm{Dist.}(\theta)$





For n

θ_d



In case of N,
in each case variance will come down but bias remains unchanged.
(comes down, but unbiased or biased thing remains unchanged)

$$E[\hat{\theta} - \theta^*] = \text{Bias}$$

$$\text{Var}[\hat{\theta}] = \text{Variance}$$

If Bias $\longrightarrow 0$ as $n \to \infty$

$\Rightarrow$ Consistent Estimator

$\lim_{n \to \infty} \text{Var}[\hat{\theta}] \xrightarrow[\text{rate}]{} 0$

↑ statistical efficiency

$\frac{1}{n} / \frac{1}{\sqrt{n}}$

Underfitting $\approx$ High Bias

Overfitting $\approx$ High Variance

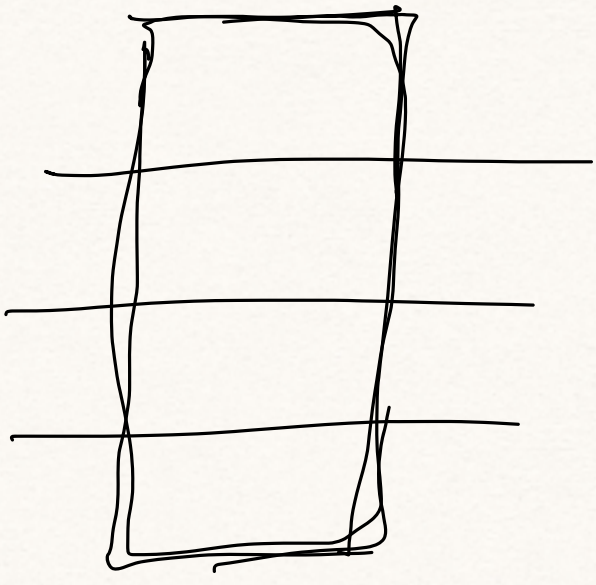
Goal : Do well on Generalization error.

- Cross validation

split it into
- Train → Do well on training set
- Valid/dev → Do well in G.E
- Test → Get an estimate on G.E.

K-fold Cross validation

Small Data sets



Model-1

fold-1 : valid

fold 2-n : training

Model-2

fold-2 : valid

fold 1,3-n : training.

leave-one-out C.V (k=n)

Regularization

Encouraging small $\|\theta\|$

$$J(\theta) = \sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)}))^2 + \lambda \|\theta\|_2^2$$

$$h_\theta(x) = \theta^T x$$

$$\sum_{i=1}^{d} \theta_i^2$$

$$\|\theta\|_1 = \sum_{i=1}^{d} |\theta_i|$$

L_2 - Regularized Linear Regression

$$J(\theta) = \left(\sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2 \right) + \underbrace{\lambda}_{>0} \|\theta\|_2^2$$

$$\hat{\theta}_n = (X^T X + \lambda I)^{-1} X^T \vec{y}$$

This term is extra

$$(X^T X + \lambda I)^{-1} = U \begin{bmatrix} (\sigma_1^2 + \lambda)^{-1} & & 0 \\ & \ddots & \\ 0 & & (\sigma_d^2 + \lambda)^{-1} \end{bmatrix} U^T$$

$$E[\hat{\theta}_n] = \left[U \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2 + \lambda} & & 0 \\ & \ddots & \\ 0 & & \frac{\sigma_d^2}{\sigma_d^2 + \lambda} \end{bmatrix} U^T \right] \theta^*$$

$$\mathrm{Cov}(\hat{\theta}_n) = V \begin{bmatrix} \frac{\tau^2 \sigma_1^2}{(\sigma_1^2+\lambda)^2} & & \\ & \ddots & \\ & & \frac{\tau^2 \sigma_d^2}{(\sigma_d^2+\lambda)^2} \end{bmatrix} V^T$$

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

$$y = \theta^T x + \varepsilon$$

$\lambda \uparrow \Rightarrow$ var $\downarrow$, Bias $\uparrow$

$\lambda \downarrow \Rightarrow$ var $\uparrow$, Bias $\downarrow$

$$MSE[\hat{f}_n] = \tau^2 + \underbrace{E[\hat{f}_n(x^*) - f(x^*)]^2}_{(\text{Bias})^2} + V[\hat{f}_n(x^*)]$$

τ^2: irreducible error.

$V[\hat{f}_n(x^*)]$: variance

$$f(x) = \theta^{*T} x$$

$$\text{Bias}(\hat{f}_n(x^*)) = E[\hat{f}_n(x^*) - f(x^*)]$$

$$= E[\hat{\theta}_n^T x^* - \theta^{*T} x^*]$$

$$= E[\hat{\theta}_n - \theta^*]^T x^*$$

$$= E[\hat{\theta}_n - \theta^*]^T x^*$$

$$= \text{Bias}(\hat{\theta}_n)^T x^*$$

$$\text{Var}(\hat{f}_n) = x^T \text{Cov}(\hat{\theta}_n) x$$

Heuristics for Bias & Variance

Training Error $\approx$ Bias

Cross Validation Error $-$ Training Error $\approx$ Variance

To fight Bias

* Make model larger
* Add more features
* Reduce regularization
* More complex model

To fight Variance

* collect more data
* Increase regularization
* simpler model.