BIAS AND VARIANCE ANALYSIS

8/4/2023

SUPERVISED $\mathbb{R}$

$$S = \{(x^{(i)}, y^{(i)})\}^{n}, \quad x^{(i)} \in \mathbb{R}^{d}, \quad y \in \mathbb{R}$$

θ^{*} "TRUE PARAMETER" UNKNOWN

← feat →

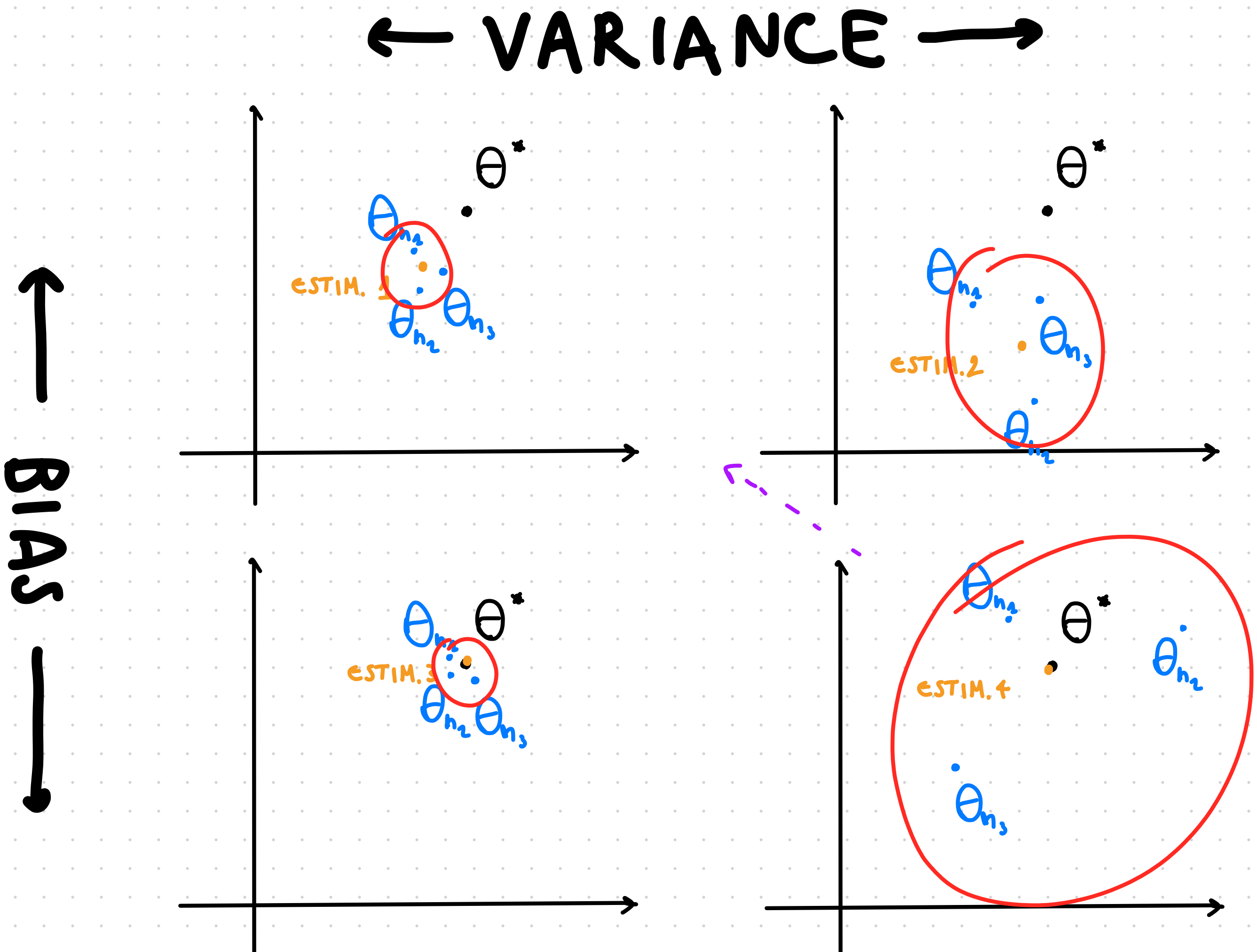
$X \in \mathbb{R}^{n \times d}$ "DESIGN MATRIX"

↑ eq ↓

$\hat{\theta}_n$ - ESTIMATOR OF θ USING n TRAINING EXAMPLES

$$\boxed{\begin{array}{c} x^{(1)}, y^{(1)} \\ x^{(2)}, y^{(2)} \\ \vdots \\ x^{(n)}, y^{(n)} \end{array}} \overset{\text{IID}}{\sim} P \Rightarrow \boxed{\text{ESTIMATOR}} \Rightarrow \hat{\theta}_n$$

ex


VARIANCE
BIAS
θ^*
θ_{n_1}
θ_{n_2}
θ_{n_3}
ESTIM. 1
ESTIM. 2
ESTIM. 3
ESTIM. 4

ESTIMATOR IS A SAMPLING DISTRIBUTION AND THE θ ARE SAMPLES

$$\text{BIAS}(\hat{\theta}_n) \equiv \mathbb{E}[\hat{\theta}_n - \theta^*]$$, IF UNBIASED $\mathbb{E}[\hat{\theta}_n] = \theta^*$

IS A THEORETICAL PROPRIETY THAT CAN ONLY BE OBTAINED ANALITICAL=LY

$$\text{VARIANCE}(\hat{\theta}_n) \equiv \text{Cov}[\hat{\theta}_n]$$

IF WE USE BOOTSTRAP (ESTIMATION W/ DIFFERENT DATA SETS) WE CAN GET GOOD ESTIMATES

AND THE MEAN SQUARED ERROR (MSE) WILL BE

$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2]$$

$$\text{MSE} = tr(\text{VAR}(\hat{\theta}_n)) + \|\text{BIAS}(\hat{\theta}_n)\|^2$$

BY LOWERING THE VAR WE CAN INCREASE THE BIAS AND VICEVERSA. WE NEED TO FIND A SWEET SPOT. LET'S TAKE A LOOK AT LINEAR REGRESSION W/ L_2 REGULARIZATION. WE DEFINED THE LOSS $J(\theta)$

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

$$\hat{\theta}_n = \arg \min_{\theta \in \mathbb{R}^d} J(\theta)$$

$$= \left(X^T X + \lambda I \right)^{-1} X^T \vec{y}$$

$$X^T X = U \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix} U^T \qquad (X^T X \text{ IS PSD BY } \underline{\text{DEF}})$$

$$X^T X + \lambda I = U \begin{bmatrix} \sigma_1^2 + \lambda & & \\ & \ddots & \\ & & \sigma_n^2 + \lambda \end{bmatrix} U^T$$

$$(X^TX+\lambda I)^{-1}X^T\vec{y} = (X^TX+\lambda I)^{T}\left(X^T(X\theta^*+\vec{\varepsilon})\right)$$

$$\hat{\theta}_n = (X^TX+\overset{(*)}{\lambda I})^{-1}X^TX\theta^* + \left[(X^TX+\lambda I)^{-1}X^T\right]\vec{\varepsilon}$$

$$\mathbb{E}[\hat{\theta}_n] = \underbrace{\mathbb{E}\Big[\underbrace{(X^TX+\lambda I)^{-1}X^TX\theta^*}_{\text{CONST}}\Big]}_{(*)} + \underbrace{\mathbb{E}\Big[\left[(X^TX+\lambda I)^{-1}X^T\right]\overset{0}{\vec{\varepsilon}}\Big]}_{0}$$

$$\mathbb{E}[\hat{\theta}_n] = (X^TX+\lambda I)^{-1}X^TX\theta^*$$

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

THIS WAS BIAS, THE MORE λ IS USED, THE MORE THE BIAS.

WHAT ABOUT VAR?

$$\mathrm{Var}(\hat{\theta}_n) = \mathrm{Cov}[\hat{\theta}_n]$$

$$= u \begin{bmatrix} \frac{\gamma^2 \sigma_1^2}{(\sigma_1^2+\lambda)^2} & & \\ & \ddots & \\ & & \frac{\gamma^2 \sigma_n^2}{(\sigma_n^2+\lambda)^2} \end{bmatrix} u^T$$

WE ARE INTRODUCING REGULARIZATION, REDUCING VAR BUT INDUCING BIAS

HOW DO WE FIND THE SWEET SPOT?

FINDING THE SWEET SPOT BY LOOKING AT THE MEAN OF VAR AND BIAS IN TERMS OF PREDICTION

$$S = \{x^{(i)}, y^{(i)}\}_{i=1}^{n}$$

$$y = f(x) + \varepsilon \qquad \mathbb{E}[\varepsilon] = 0, \quad V[\varepsilon] = \sigma^2$$

"TRUE" f

$$f'(x) = \mathbb{E}[y \mid x = x']$$

$$\hat{f}_n(x) = \vec{y}$$

THE $MSE_{f_n(x)}$

$$MSE_{f_n(x)} = \mathbb{E}\left[\left(y_* - \vec{f}_n(x_*)\right)^2\right]$$

x_*, y_* IS "TEST EXAMPLE"

$$= \mathbb{E}\left[\left(\varepsilon + f(x_*) - \hat{f}_n(x_*)\right)^2\right]$$

$$= \ldots = \underbrace{\sigma^2}_{\substack{\text{IRREDUCIABLE} \\ \text{ERROR} \\ \text{COMPONENT}}} + \underbrace{\mathbb{E}\left[\hat{f}_n(x_*) - f(x_*)\right]^2}_{\text{BIAS}} + \underbrace{\mathbb{V}\left[\hat{f}_n(x_*)\right]}_{\text{VAR}}$$

WILL ALWAYS EXIST, EVEN W/ TRUE f GIVEN FROM RANDOMNESS ← IRREDUCIABLE ERROR COMPONENT

VAR: HERE LIES RANDOMNESS (NOISE)

NOTE

$$n \longrightarrow \infty$$

$$\text{VAR}[\hat{\theta}_n] \longrightarrow 0$$

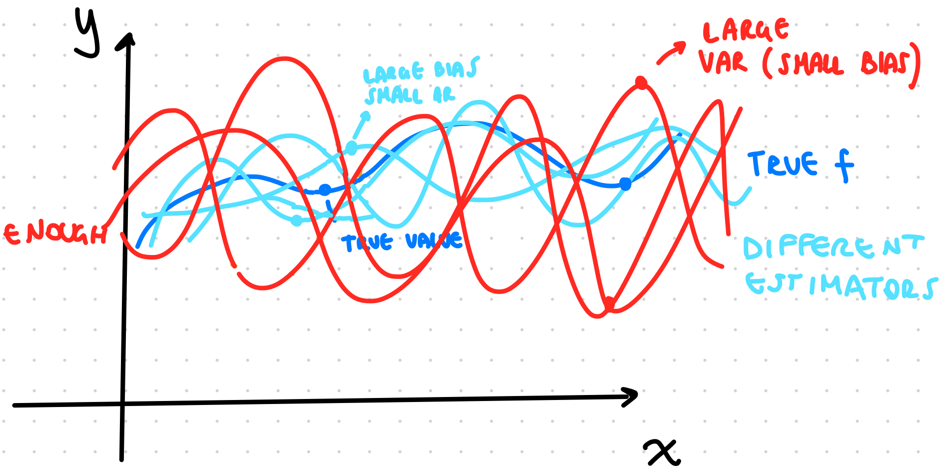
GOING BACK TO L.R. W/ L_2 REG.

$$\text{BIAS}(\hat{f}_n) = \text{BIAS}(\hat{\theta}_n)^T x_*$$

$$\text{VAR}(\hat{f}_n) = x_*^T \, \text{VAR}[\hat{\theta}_n] \, x_*$$

THE CURVES ARE ESTIMATORS

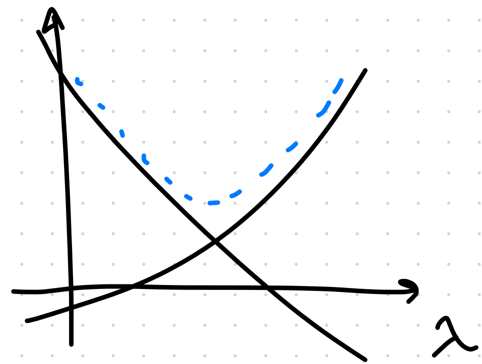
- UNDERFITTING NOT ENOUGH REG.
- OVERFITTING



AS WE INCREASE λ INCREAS BIAS → REDUCE VAR

AND THERE IS SOME σ^2

THE MSE WILL BE



TO KNOW IF WE ARE OVER/UNDER FITTING THERE ARE HEURISTICS

	EX 1	EX 2	EX 3
	TEST	TEST TRAIN	TRAIN
ERROR			TEST
	TRAIN		
	OVERFITTING	UNDERFITTING	NOISE / BUG

THROWING MORE DATA LOWERS TRAIN ERROR

THROW COMPUTING LOWER TEST ERR.