Bias/Variance and Regularization

Fundamentals of Data Science
12, 15 December 2022
Prof. Fabio Galasso

SAPIENZA
UNIVERSITÀ DI ROMA

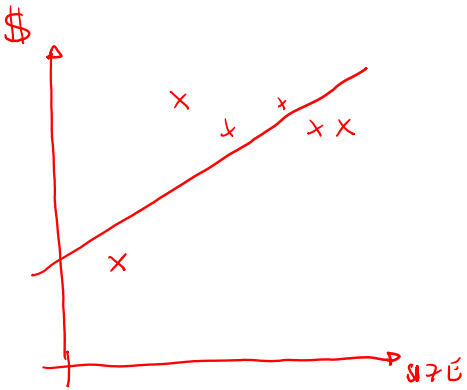
Outline

- Bias/Variance
- Regularization
- Hold-out Cross Validation

Bias/Variance

SAPIENZA
UNIVERSITÀ DI ROMA

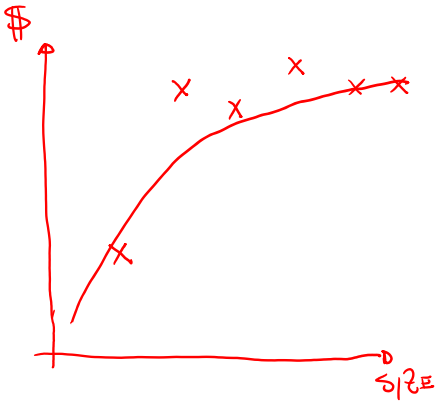
Bias/Variance
Regression



$\theta_0 + \theta_1 x$

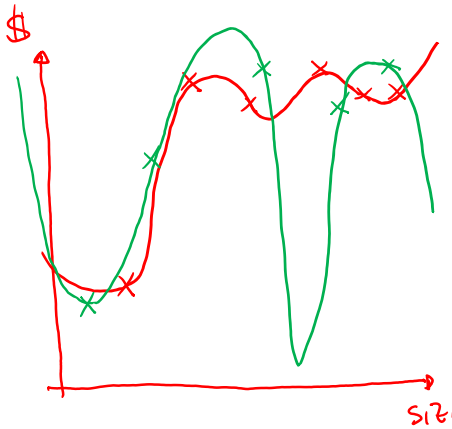
UNDERFIT

HIGH BIAS



$\theta_0 + \theta_1 x + \theta_2 x^2$

"JUST RIGHT"

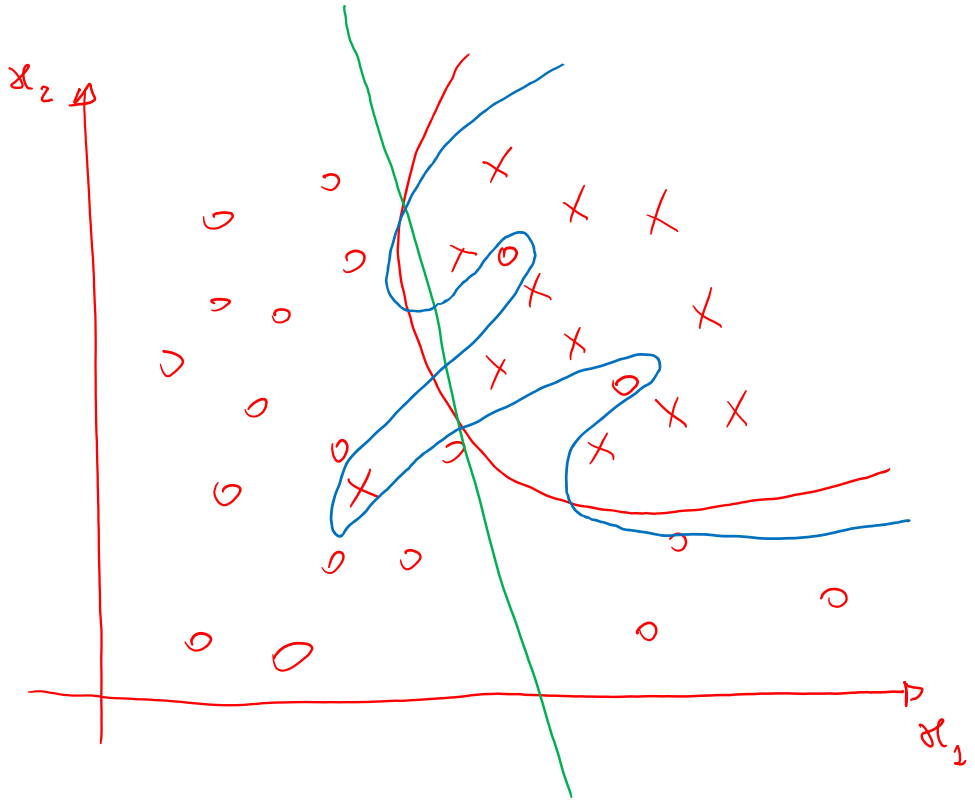


$\theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_5 x^5$

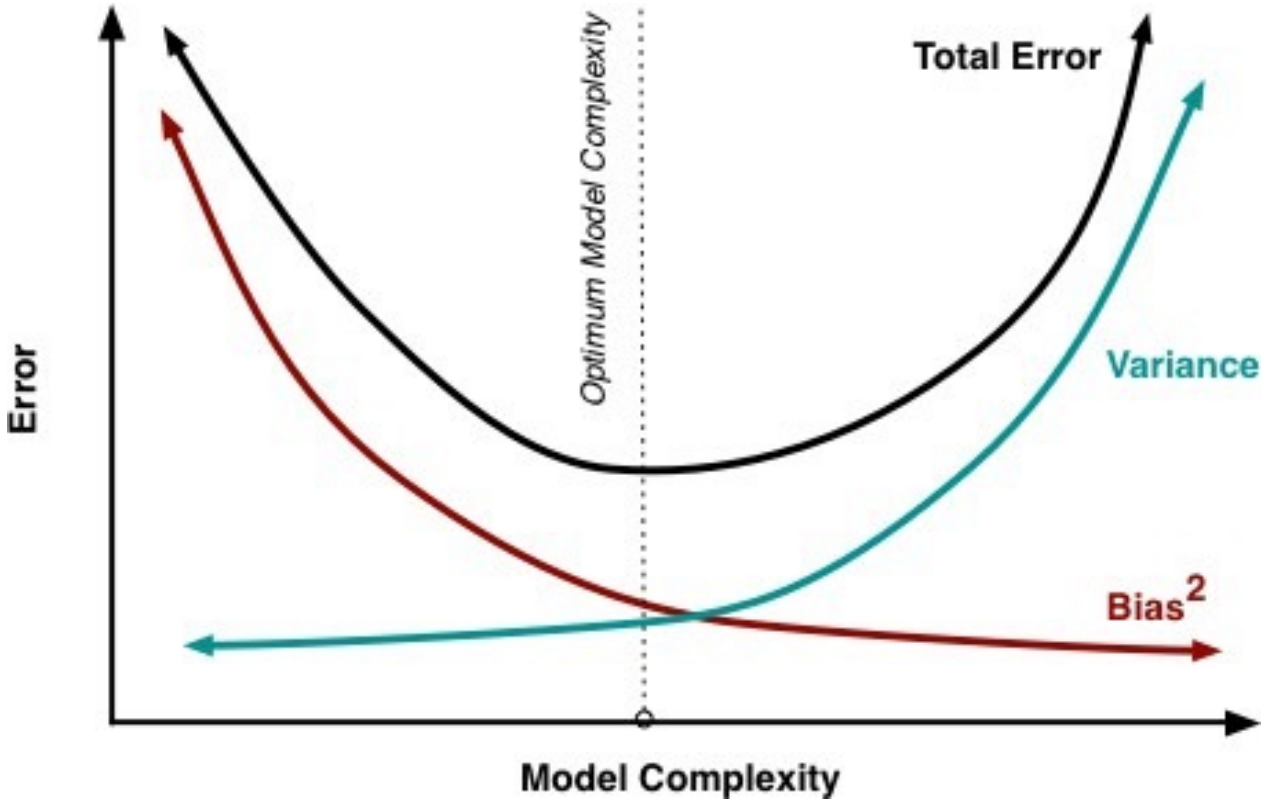
OVERFIT

HIGH VARIANCE

Bias/Variance
Classification



Bias-Variance Tradeoff



http://scott.fortmann-roe.com/docs/BiasVariance.html

Setup/Assumptions

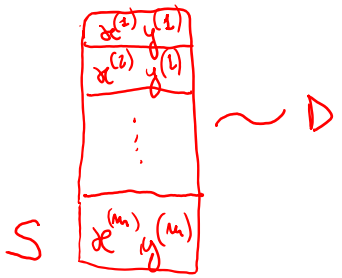
1) DATA DISTRIBUTION D

$(x, y) \sim D$ < TRAIN / TEST

2) INDEPENDENT SAMPLES

θ^* OR h^* "TRUE" PARAMETERS (NOT RANDOM)

PROCESS OF LEARNING



RANDOM VARIABLE

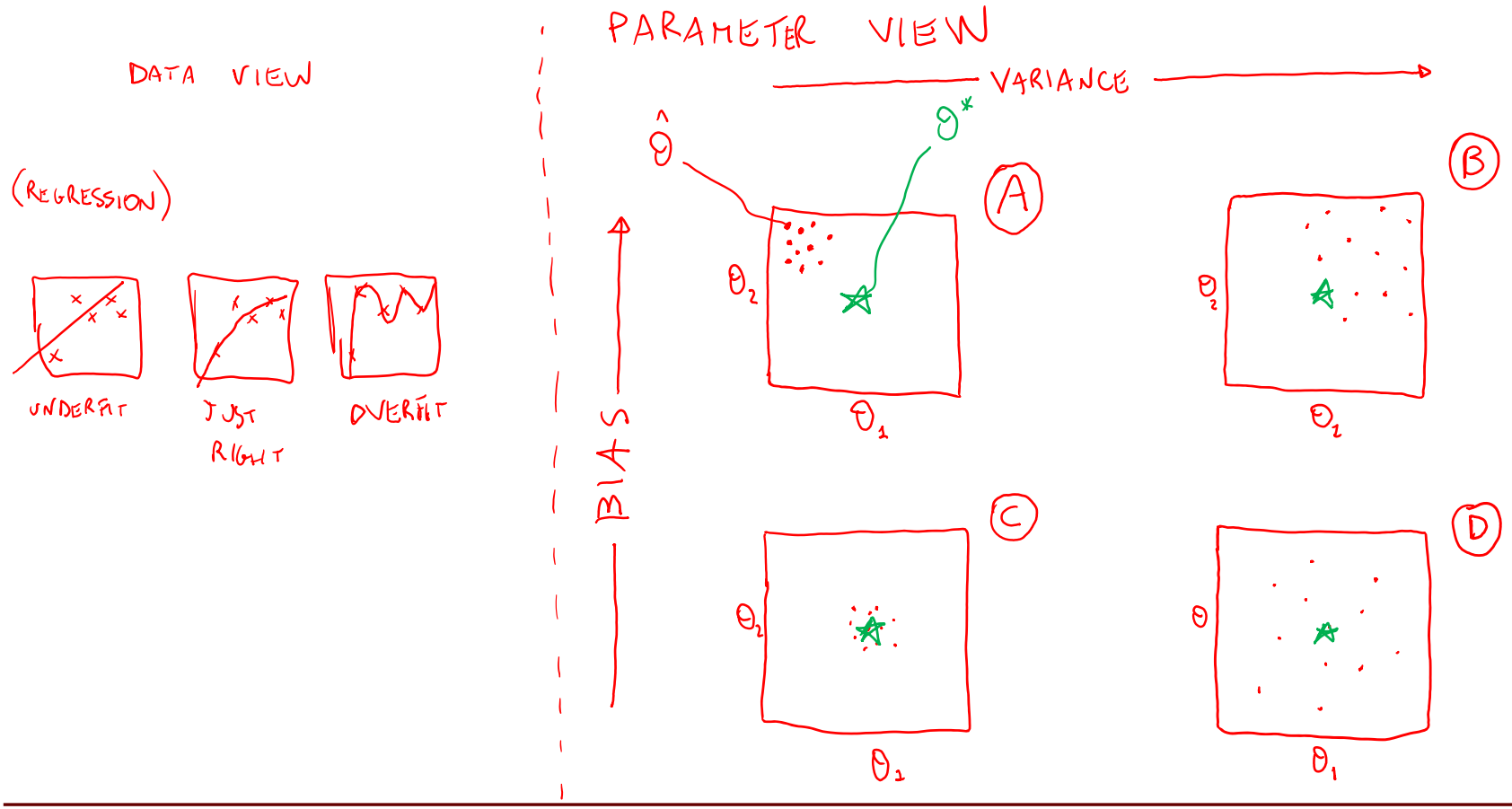
LEARNING ALGORITHM "ESTIMATOR"

DETERMINISTIC FUNCTION

$\hat{h}$ OR $\hat{\theta}$ $\sim$ SAMPLING DISTN.

RANDOM VARIABLE

Bias/Variance
Data and Parameter View



Consistency, Statistical Efficiency, Bias

$m \to \infty$

$\mathrm{VAR}[\hat{\theta}] \to 0$

E.g. $\Theta\left(\frac{1}{m^2}\right)$ $\Theta(e^{-m})$

"STATISTICAL EFFICIENCY": RATE $\mathrm{VAR}[\hat{\theta}] \to 0$ AS $m \to \infty$

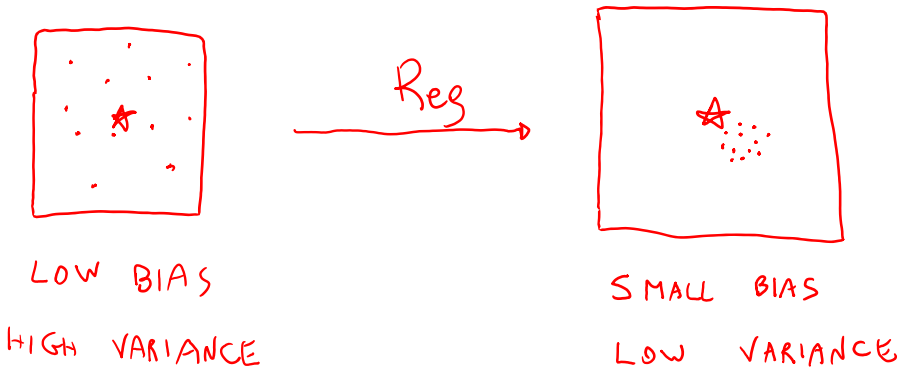
$\hat{\theta} \to \theta^*$ $m \to \infty$: CONSISTENT

$E[\hat{\theta}] = \theta^*$ for all m : UNBIASED

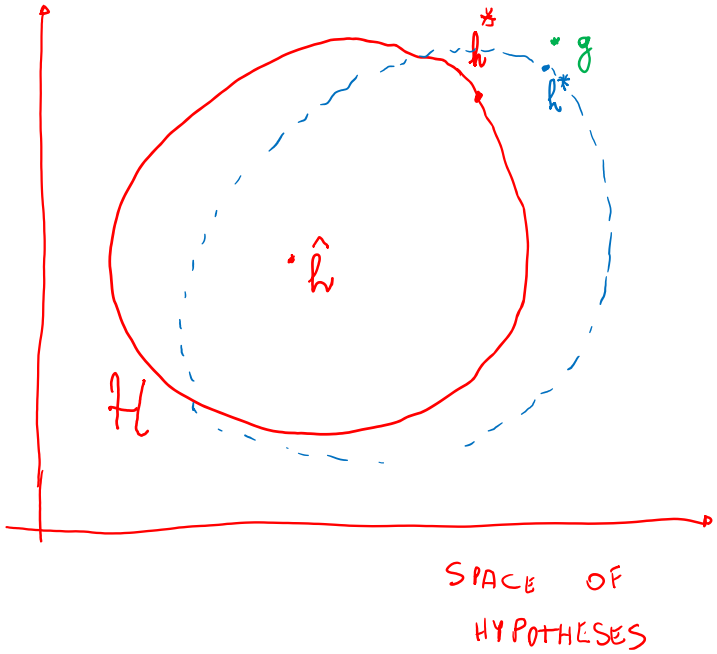
Fighting High Variance

1) $m \to \infty$

2) REGULARIZATION



Generalization and Empirical Risk



g BEST POSSIBLE HYPOTHESIS

h^* BEST IN CLASS $\mathcal{H}$

$\hat{h}$ LEARNT FROM FINITE DATA

$$\mathcal{E}(h) = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}\{h(x) \neq y\}\right]$$ RISK GENERALIZATION ERROR

$$\hat{\mathcal{E}}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}$$ EMPIRICAL RISK

Approximation and Estimation Error

$\mathcal{E}(g)$: BAYES ERROR / IRREDUCIBLE ERROR

$\mathcal{E}(h^*) - \mathcal{E}(g)$: APPROXIMATION ERROR } CLASS

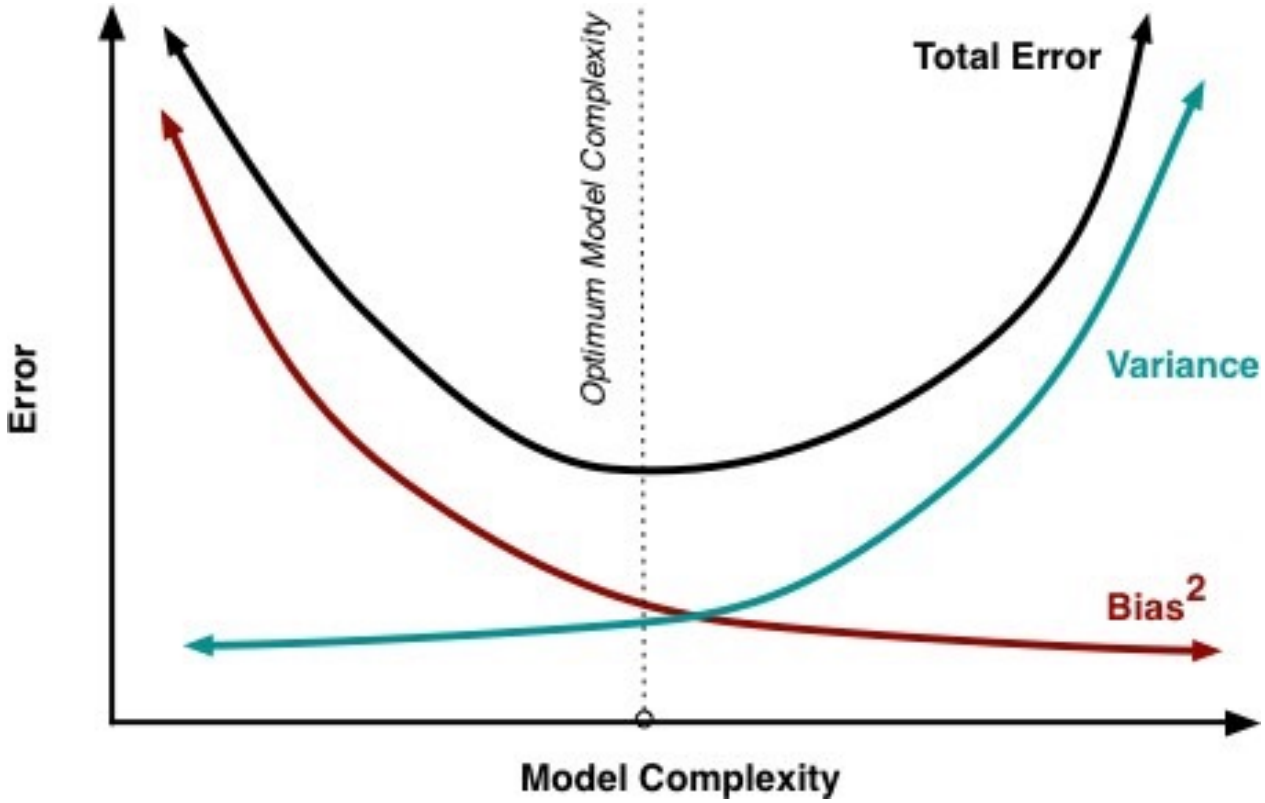
$\mathcal{E}(\hat{h}) - \mathcal{E}(h^*)$: ESTIMATION ERROR } DATA

$\mathcal{E}(\hat{h})$: EST. ERROR + APPROX. ERROR + IRREDUCIBLE ERROR

= EST. VAR + EST. BIAS + APPROX. ERROR + IRR. ERROR

$\mathcal{E}(\hat{h})$: VARIANCE + BIAS + IRREDUCIBLE

Bias-Variance Tradeoff



http://scott.fortmann-roe.com/docs/BiasVariance.html

Outline

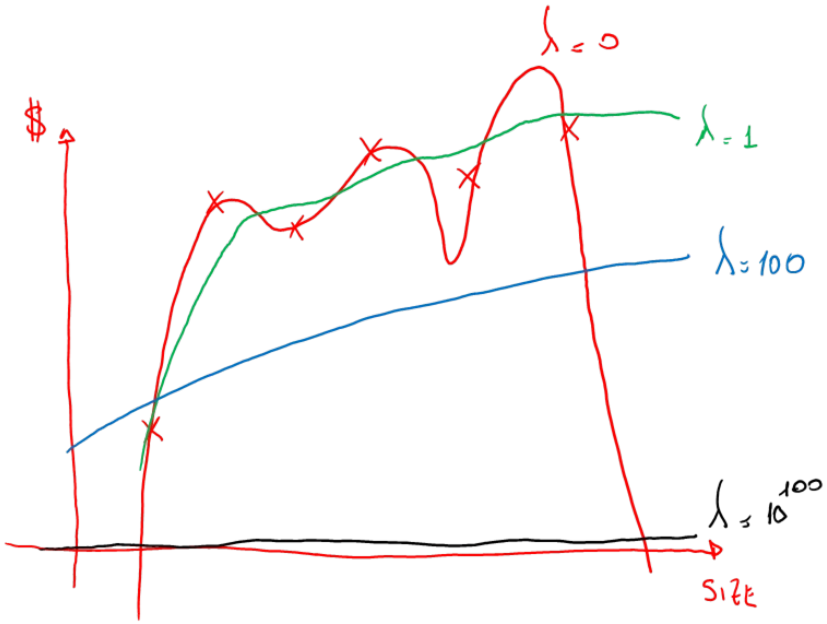
- Bias/Variance
- Regularization
- Hold-out Cross Validation

Regularization

SAPIENZA
UNIVERSITÀ DI ROMA

Regularization

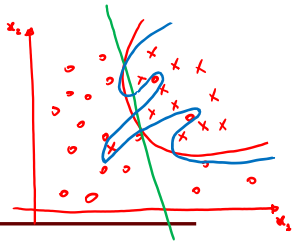
$$\min_{\theta} \frac{1}{2} \sum_{i=1}^{m} \| y^{(i)} - \theta^T x^{(i)} \|^2 + \underbrace{\frac{\lambda}{2} \| \theta \|^2}_{\text{REGULARIZATION}}$$



$$\text{ARGMAX}_{\theta} \sum_{i=1}^{m} \log P(y^{(i)} | x^{(i)}; \theta) - \lambda \| \theta \|^2$$

$h_\theta(x) \approx \emptyset$

Regularization and Gradient Descent



$$\underset{\theta}{\text{ARGMAX}} \sum_{i=1}^{m} \log p\left(y^{(i)} \mid x^{(i)}; \theta\right) - \frac{\lambda}{2}\|\theta\|^2 \longrightarrow -\frac{\lambda}{2}\sum_{j=1}^{n}\theta_j^2$$

NO REGULARIZATION FOR $j=0$

$$\ell(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right) - \frac{\lambda}{2}\sum_{j=1}^{n}\theta_j^2$$

$$\theta_j = \theta_j + \alpha\left[\sum_{i=1}^{m}\left(y^{(i)} - h_\theta\left(x^{(i)}\right)\right)x_j^{(i)} - \lambda\theta_j\right] \qquad j \in [1, 2, \ldots n]$$

$$\frac{\partial}{\partial\theta_j}\ell(\theta)$$

FOR LOGISTIC REGRESSION: $h_\theta\left(x^{(i)}\right) = g\left(\theta^T x^{(i)}\right)$

$$\theta_0 = \theta_0 + \alpha\sum_{i=1}^{m}\left(y^{(i)} - h_\theta\left(x^{(i)}\right)\right)$$

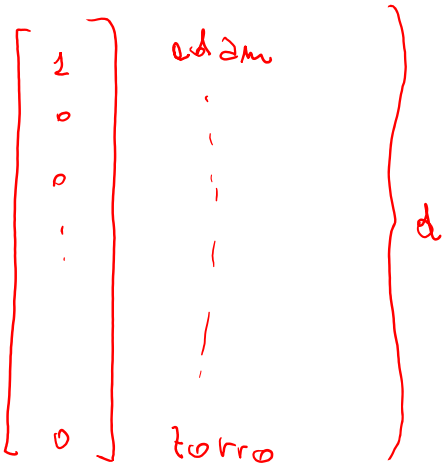
SINCE $x_0^{(i)} = 1$ AND NO REG FOR θ_0

ALSO APPLICABLE TO LINEAR REG $h_\theta\left(x^{(i)}\right) = \theta^T x^{(i)}$

Application to Text Classification

$m = 100$ EXAMPLES

$n = d = 10'000$

$x =$ 

Regularization: Probabilistic Interpretation

$S = \{x^{(i)}, y^{(i)}\}_{i=1}^{m}$ $\quad P(\theta \mid S) = \frac{P(S \mid \theta)\, P(\theta)}{P(S)}$

$$\underset{\theta}{\text{ARGMAX}}\ P(\theta \mid S) = \underset{\theta}{\text{ARGMAX}}\ P(S \mid \theta)\, P(\theta) = \underset{\theta}{\text{ARGMAX}} \left(\prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta) \right) P(\theta)$$

LOGISTIC REGRESSION

$P(\theta)$: $\theta \sim \mathcal{N}(0, \tau^2 I)$ $\quad P(\theta) = \frac{1}{\sqrt{2\pi}\, |\tau^2 I|^{1/2}} \exp\left(-\frac{1}{2} \theta^T (\tau^2 I)^{-1} \theta\right)$

FREQUENTIST : ARGMAX $P(S \mid \theta)$ MLE

BAEYSIAN : PRIOR DISTN. $P(\theta)$ $\underset{\theta}{\text{ARGMAX}}\ P(\theta \mid S)$ MAP MAXIMUM A POSTERIORI

Generative vs. Discriminative; Bayesian vs. Frequentist

$y \rightarrow$ CLASS

$x \rightarrow$ INPUT FEATURES

$\theta \rightarrow$ PARAMETERS

	FREQUENTIST	BAYESIAN
DISCRIMINATIVE	$p(y; x, \theta)$	$p(y, \theta; x) = p(y \mid \theta; x)\, p(\theta)$
GENERATIVE	$p(y, x; \theta)$	$p(y, x, \theta) = p(y, x \mid \theta)\, p(\theta)$

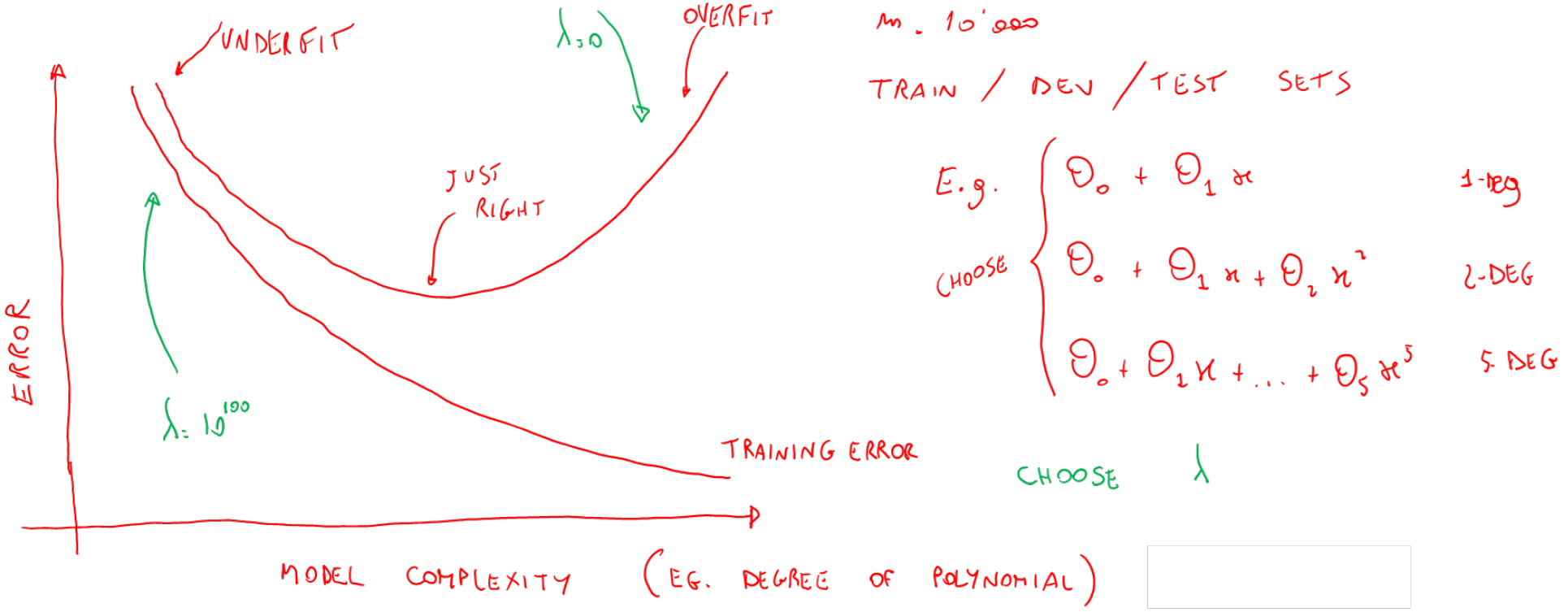
Outline

- Bias/Variance
- Regularization
- Hold-out Cross Validation

Hold-out Cross Validation

SAPIENZA
UNIVERSITÀ DI ROMA

Model Complexity Vs Error



Train/Dev/Test

1. Split S into S_{train}, S_{dev}, S_{test}

DEV = DEVELOPMENT / CROSS VALIDATION

2. Train each model S_{train}

GET HYPOTHESES h_θ'S E.G. VARIOUS DEGREE OF POLYNOMIAL

3. Choose the model with lowest error on S_{dev}

↳ NOT ON S_{TRAIN}

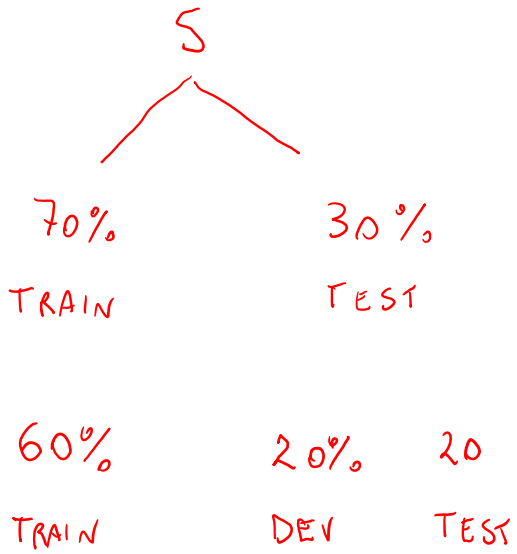
- Optional: evaluate on S_{test} and report performance

↳ NOT ON S_{DEV}

DEGREE	S_{DEV} ERROR	S_{TEST} ERROR
1	10	10
2	5.1	5.0
3	5.0	5.0
4	4.9	5.0
5	7	7
6	10	10

(SIMPLE) HOLD-OUT CROSS VALIDATION

Hold-out Cross Validation

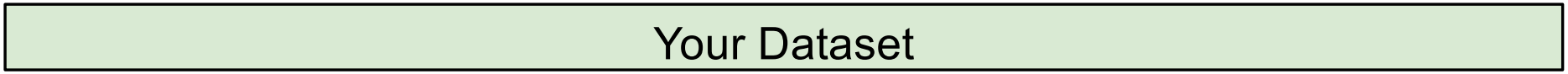


10'000'000

TRAIN	6,000,000
DEV	2,000,000
TEST	2,000,000

Setting Hyperparameters

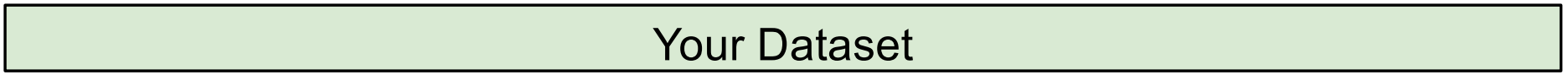
Idea #1: Choose hyperparameters that work best on the data



Setting Hyperparameters

Idea #1: Choose hyperparameters that work best on the data

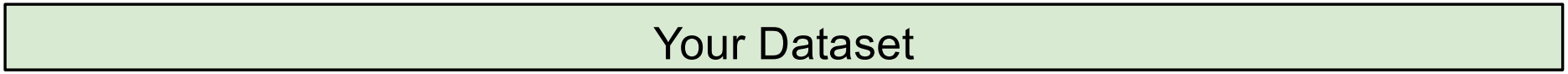
BAD: would always choose most complex model



Setting Hyperparameters

Idea #1: Choose hyperparameters that work best on the data

BAD: would always choose most complex model

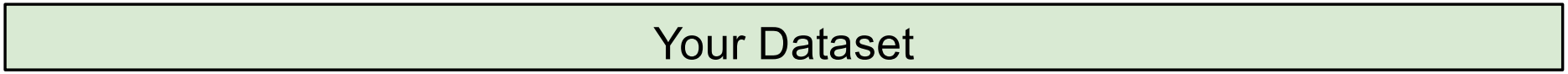


Idea #2: Split data into **train** and **test**, choose hyperparameters that work best on test data

Setting Hyperparameters

Idea #1: Choose hyperparameters that work best on the data

BAD: would always choose most complex model



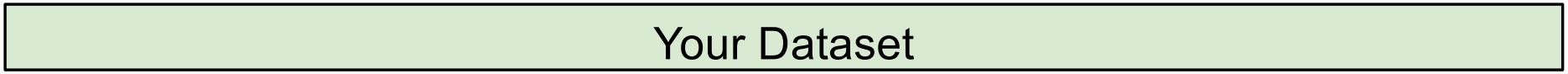
Idea #2: Split data into **train** and **test**, choose hyperparameters that work best on test data

BAD: No idea how algorithm will perform on new data

Setting Hyperparameters

Idea #1: Choose hyperparameters that work best on the data

BAD: would always choose most complex model

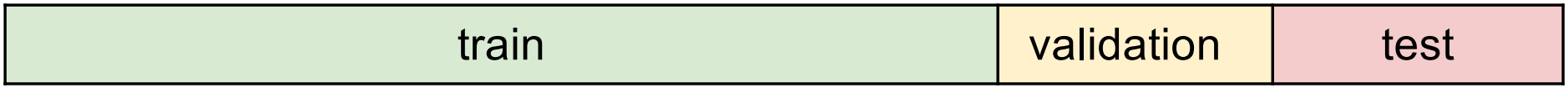


Idea #2: Split data into **train** and **test**, choose hyperparameters that work best on test data

BAD: No idea how algorithm will perform on new data

Idea #3 (**hold-out cross validation**): Split data into **train**, **val**, and **test**; choose hyperparameters on val and evaluate on test

Better!



Setting Hyperparameters

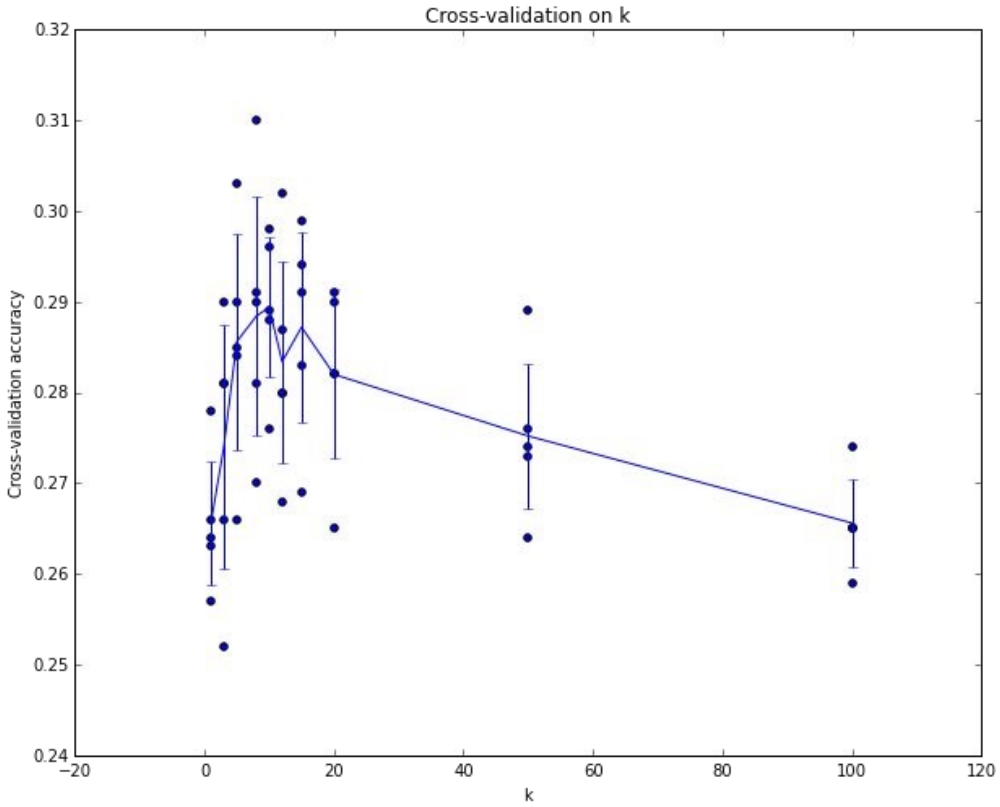
Your Dataset

Idea #4: **k-fold cross-validation**: Split data into **folds**,
try each fold as validation and average the results

fold 1	fold 2	fold 3	fold 4	fold 5	test
fold 1	fold 2	fold 3	fold 4	fold 5	test
fold 1	fold 2	fold 3	fold 4	fold 5	test

Useful for small datasets, but not used too frequently in deep learning

Setting Hyperparameters



Example of
5-fold cross-validation
for the value of **k.**

Each point: single
outcome.

The line goes
through the mean, bars
indicated standard
deviation

(Seems that k ~= 7 works best for this data)

References

- Sections 1.2.3, 1.3 and 3.2 in [Bishop, 2006. Pattern Recognition and Machine Learning]
- Sections 5-12, 20-32 in [Andrew Ng, 2019. Machine Learning Yearning]

Thank you

Acknowledges: slides and material from Andrew Ng, Joe Blitzstein, Hanspeter Pfister, Verena Kaynig-Fittkau

SAPIENZA
UNIVERSITÀ DI ROMA