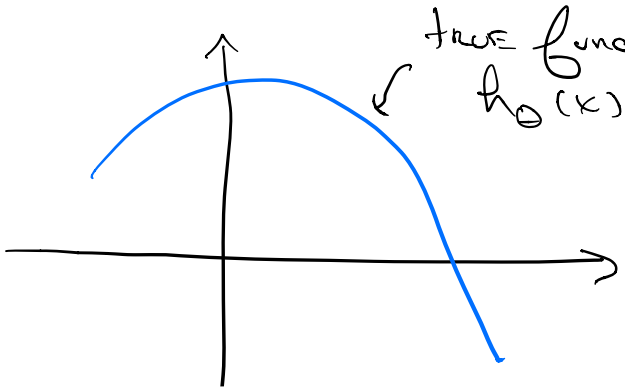
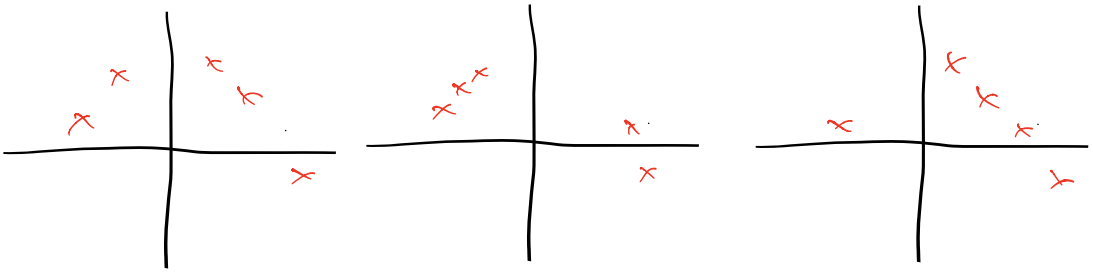
Bias & Variance: Classical & modern Elements

+ Classical theory
  - → Regularization
  - → Parameter Selection → Compute efficient: Successive Halving
    Data : k-fold
+ Modern theory (Bonus)

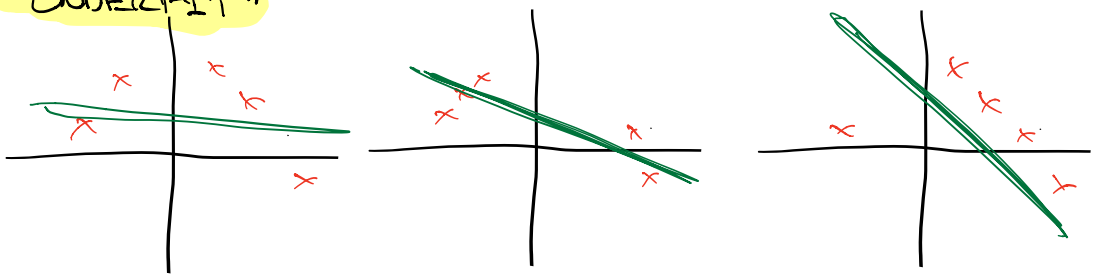


We don't get to see h_θ directly - only samples



What happens if we fit a line to these samples?

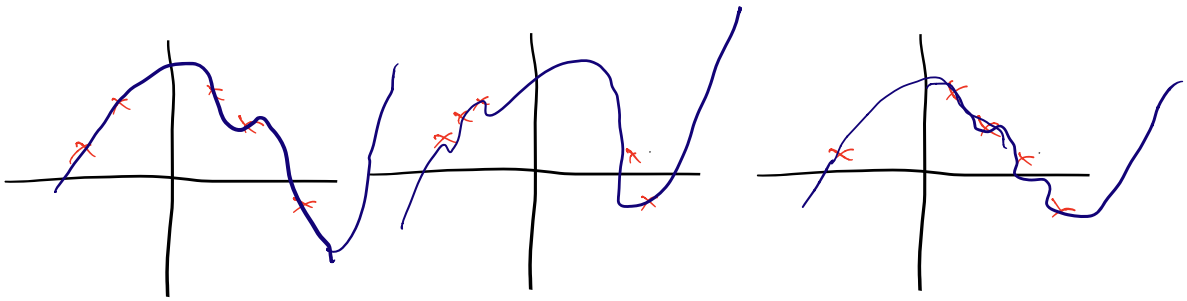
"UNDERFIT"



$S^{(1)}$ $S^{(2)}$ $S^{(3)}$

We informally call this underfit the error is pretty high (we use for def.)

What happens if we use degree 5 polynomial?



$S^{(1)}$ $S^{(2)}$ $S^{(3)}$

"OVERFIT"

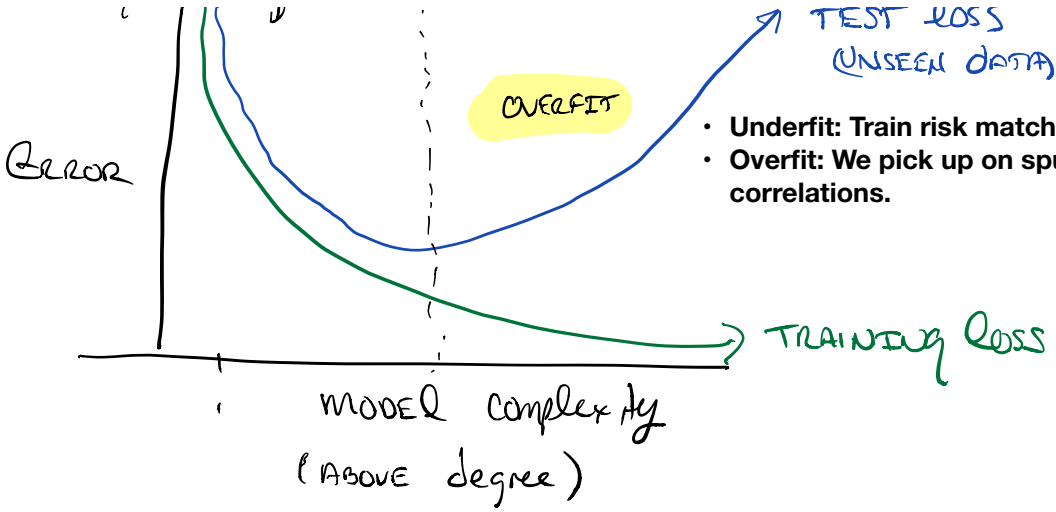
This fits EACH sample well. But the function totally CHANGES PER SAMPLE! (High variance)

WHAT IF WE USE QUADRATICS?

→ low error & low variance.

→ IT FITS

UNDERFIT → optimal complexity



- **Underfit: Train risk matches test risk.**
- **Overfit: We pick up on spurious correlations.**

NB: This is CLASSICAL BIAS VARIANCE.

→ helpful to understand many ML ideas.

→ Incomplete for modern models in important ways (more later)

More Formal Bias-Variance

Consider linear regression

$y \in \mathbb{R}$ — Output

$y = \theta \cdot x + \epsilon$

x: features (data/input) $\in \mathbb{R}^d$

θ: parameters in $\mathbb{R}^d$

$\epsilon \sim \mathcal{N}(0, \sigma^2)$

Procedure

Fix $x \in \mathbb{R}^d$, a test point (reason about error here)

1. Draw n points $(x^{(1)}, y^{(1)}) \dots (x^{(n)}, y^{(n)})$

CALL THIS S NOTE: $y^{(i)} = h_\theta(x^{(i)}) + \epsilon^{(i)} \sim N(0, \sigma^2)$

2. TRAIN A LINEAR REGRESSOR $h_S : \mathbb{R}^d \to \mathbb{R}$

3. DRAW TEST SAMPLE (x, y) IN WHICH

$$h_\theta(x) + \epsilon = y$$

$\epsilon \sim N(0, \sigma^2)$ NOISE

4. MEASURE $(h_S(x) - y)^2$

WE EXAMINE

$$\mathbb{E}_{\epsilon, S}[(h_S(x) - y)^2]$$

→ TWO SOURCES OF RANDOMNESS

GOAL: DECOMPOSE THIS ERROR

$$\mathbb{E}\left[(h_S(x) - (h_\theta(x) + \epsilon))^2\right] =$$

$$\underbrace{\mathbb{E}_\epsilon[\epsilon^2]}_{\sigma^2} + \underbrace{\mathbb{E}_S\left[(h_S(x) - h_\theta(x))^2\right]}_{\text{DEPENDS ON TRAIN SET } S} + 2\,\mathbb{E}\left[\epsilon\,(h_S(x) - h_\theta(x))\right]$$

$\mathbb{E}[\epsilon] = 0$ & ϵ, S INDEPENDENT

UNAVOIDABLE ERROR TERM

$h_{AVG}(x) \triangleq \mathbb{E}_S[h_S(x)]$ "long run average of trained x"

$$= \mathbb{E}_S\left[(h_\theta(x) - h_{AVG}(x) + h_{AVG}(x) - h_S(x))^2\right]$$

$$= \mathbb{E}_S\left[(h_\theta(x) - h_{AVG}(x))^2\right] + \mathbb{E}\left[(h_{AVG}(x) - h_S(x))^2\right] + \sigma$$ (cross term)

$$= (h_\theta(x) - h_{AVG}(x))^2 + VAR_S(h)$$

$(h_\theta(x) - h_{AVG}(x))^2$: does NOT depend on S; on CLASS of hypotheses → BIAS

$VAR_S(h)$: "VARIANCE OVER TRAINING SET" → VARIANCE

RECAP

$$\mathbb{E}_{\epsilon,S}[(y - h_S(x))^2] = \sigma^2 + \text{BIAS} + \text{VARIANCE}$$

(σ^2: NOISE TERM IN TEST DATA)

Examples

	LINEAR	degree 5
BIAS	LARGE	0 (fits every point!)
VARIANCE	lower	higher

just right fit combines both!

NOTE:
- Having different DEV / HOLDOUT SET allows us to ASSESS VARIANCE (and train STABILITY)
- IF we use model class that is expressive may need to "trust points less" (reduce VARIANCE) → Regularization

Regularization is at heart of both classical

and modern theory. Spend a little bit of time on this...

Regularization

Reduce variance to get robust model

→ can be Explicit (change model)

Implicit (procedure)

Most Classical Linear Regression

$$\underset{\theta \in \mathbb{R}^d}{\text{Argmin}} \; \frac{1}{2} \sum_{i=1}^{n} \left(x^{(i)} \cdot \theta - y^{(i)} \right)^2 + \frac{\lambda}{2} \|\theta\|_2^2$$

λ: Parameter $\in \mathbb{R}_+$

→ Penalty for really complex model (minimum norm solution)

$\lambda = 0$ → Ordinary least squares

$\lambda = 10^{100}$ → $\theta = 0$ probably looks pretty good!

Set λ to some value to balance loss & norm.

HW will show you!

Solution?

Fix $\lambda > 0$.

Take derivate wrt to θ

$$X^T X\theta - X^T y + \lambda\theta = 0$$

$$\boxed{(X^T X + \lambda I)\theta = X^T y}$$

Underdetermined case

If X^TX is not full rank $\nexists\ \lambda = 0$ may not have unique solution ($\forall X \in \mathbb{R}^{n \times d}$ $n < d$)

If X^TX is not full rank $\exists v$ s.t.

$v \neq 0$ but $X^TXv = 0$

if $X^TX\theta = X^Ty$ then $X^TX(\theta + v) = X^Ty$ as well $\Rightarrow$ no unique solution

$\Rightarrow$ if $\lambda > 0$ then it does have a unique solution, since $X^TX + \lambda I$ is full rank.

that is eigenvalues of X^TX $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2 \geq 0$

then $X^TX + \lambda I$ has eigenvalues

$\sigma_1^2 + \lambda,\ \sigma_2^2 + \lambda\ \dots,\ \sigma_n^2 + \lambda > 0$

In this case, $\theta_\lambda = (X^TX + \lambda I)^{-1}X^Ty$

Back to variance

$$\mathop{\mathbb{E}}_{S}\left[(h_S(x) - h_{Avg}(x))^2\right]$$

Variance in θ_λ the solution for fixed λ

$$VAR_S = \mathop{\mathbb{E}}_{S}\left[(\theta_\lambda \cdot x - \mathop{\mathbb{E}}_{S}[\theta_\lambda \cdot x])^2\right]$$

$$\approx \mathbb{E}\left[\|\theta_1 - \mathbb{E}[\theta_1]\|^2\right] \quad (*)$$

S

TO Simplify Analysis,

AND only RANDOMNESS IN DRAW IS THE TRAIN Points Eaqual

$$y = X\cdot\theta + r \qquad r \in N(0, c^2 I_n)$$

RANDOM NOISE PER Point fixed X

$$\textcircled{*} = \mathbb{E}\left[\| \underbrace{(X^T X + \lambda)^{-1} X^T}_{A} r \|^2\right]$$

then $Ar \sim N(0, c^2 AA^T)$

Hence, $\leq c^2 \frac{\sigma_{max}^2}{(\sigma_{max}^2 + \lambda)^2}$

SO AS λ INCREASES, VARS DECREASES

Bonus Observation CAN sometimes implicitly Regularize as well ((Suprisingly Important! In modern theory)

thought Experiment, WE RUN gradient descent with $\lambda = 0$ IN underdetermined CASE

Claim if WE initialize TO $\theta^{(0)}$ then our solution IS θ_n

$$\theta_{GD} = P_{\text{NULL}(X)}(\theta_{GD}) + P_{\text{SPAN}(X)}(\theta_{GD})$$

$$= \theta_0 !$$

why? $\theta^{(t+1)} = \theta^{(t)} - \alpha X^T(X^T\theta - y)$

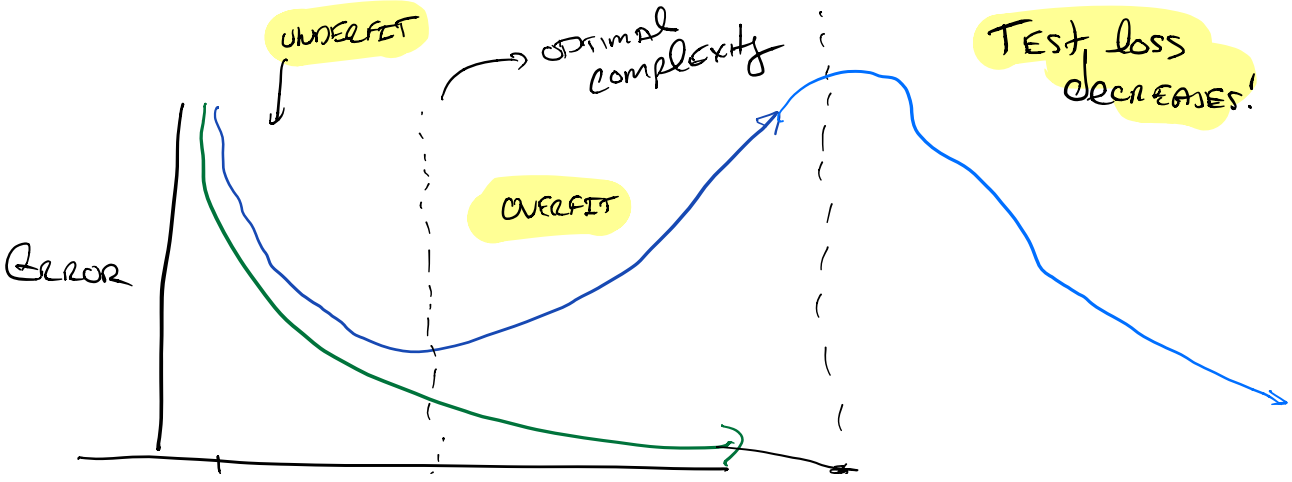
only changes in SPAN(X)!

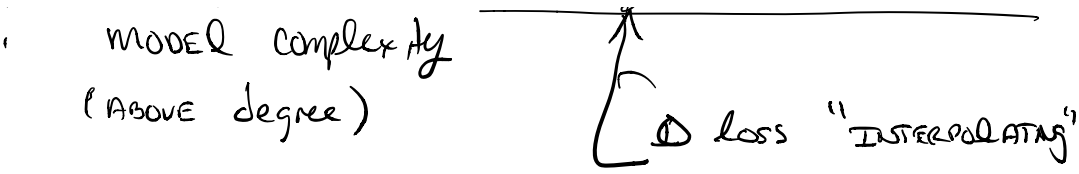
Observation: We can Regularize by Initializing!
set $\theta^{(0)} = 0$ has good properties.

$\Rightarrow$ deep learning is undetermined (often)
and so initialization plays major role!

IN FACT, SGD plays a starring role in
modern theory of BIAS VARIANCE

Belkin et. al 2018 "DOUBLE DESCENT"





first OBSERVED (widely) for DEEP NETS,
But ALSO TRUE for kernels.

SGD Regularizes by picking MIN NORM Solution

MEMORIZATION AND GENERALIZATION!

OTHER METHODS of BIAS & VARIANCE

+ DATA Augmentation (SEE SHARON Li BLOG on DATA)
+ Drop out "DATA ADAPTIVE"
+ OPTIMIZATION Algorithms (PROXIMAL POINT METHODS)
:

LOTS MORE TO DO!

Picking Hyper Parameters

THREE SETS OF LABELED DATA

TRAIN - FIT PARAMETERS

DEV - "FIT" "HYPERPARAMETERS" e.g. λ

TEST (BLIND)

Our first example

for degree $d \in \{0, 1, \ldots, K\}$

TRAIN model(d) ON TRAIN SET $\mapsto h_d$

SCORE h_d ON DEV SET

Pick BEST SCORE, hope for best on Test

If we have INFINITELY many models we can grid search e.g.

for each $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, \ldots\}$

→ SAME PROCESS

WHY DO WE SCORE ON DEV, NOT TRAIN?

IMPROVEMENTS

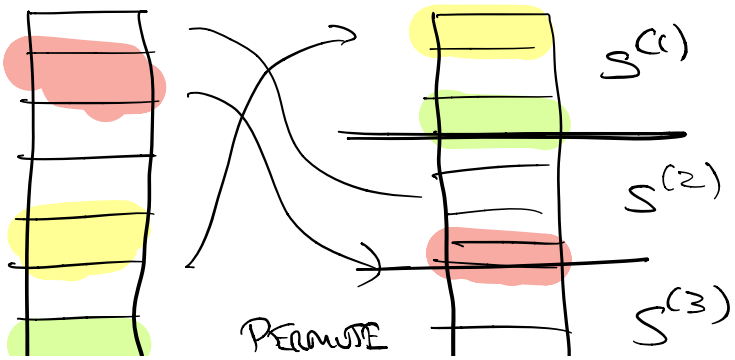
→ DATA Efficiency: MAKE BEST USE of DATA IN TRAIN/DEV
"CLASSICAL STATS"

→ Compute Efficiency: MANY RELATED hyper PARAMETERS → MANY MODELS

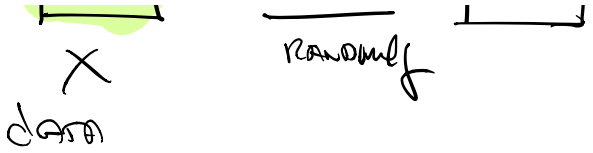
"MODERN ML STARTUP"
COMBINATORIAL Explosion.

DATA k-fold CROSS VALIDATION

k=3 but 5, 10, ... typical

1. 

MAKE k=3 SETS



2. TRAIN
$S^{(1)}, S^{(2)}$
$S^{(1)}, S^{(3)}$
$S^{(2)}, S^{(3)}$

SCORE
$S^{(3)}$
$S^{(2)}$
$S^{(1)}$

3. COMBINE SCORES (AVERAGE)

→ USE THIS TO PICK BEST.

Computational

MOTIVATION: REGULARIZER, DROPOUT RATE, STEPSIZE, DIMENSIONS -- MANY LAYERS!

Practical trick:

1. TUNE 1 PARAMETER AT A TIME
2. SWEEP OVER ALL PARAMETERS

CHOICES

$2(5 + 6 + 7) < 5 \cdot 6 \cdot 7$

MORE ADVANCED HYPERBAND (JAMIESON '15)

RUN all $5 \cdot 6 \cdot 7$ models

→ But for just a few steps

Pick top half, run for a 2^k steps

$\Rightarrow$ Repeat ---

Each round, we use same number of resources $\rightarrow$ but with fewer models

Run $\log_2(5 \cdot 6 \cdot 7)$ rounds

Lots more to do here (clear across runs?)

Recap:

Bias & variance $\rightarrow$ classical / modern

Regularization Explicit and Implicit

Tuning Cross Validation & Hyperparameter search