Lecture 4:
Cross validation and Bias-Variance Tradeoff

- explaining test error using theoretical analysis

- HW0 due Wednesday October 8th midnight

Cross-validation

Leave-one-out Cross Validation and k-fold Cross Validation

LOO cross validation

LOO Cross validation

(n-fold CV)

$$\text{Error}_{LOOCV} \triangleq \frac{1}{n}\sum_i \text{Err}_i$$

$$\text{Error}_{\text{validate } i} \triangleq (Y_i - \hat{f}_{D/i}(X_i))^2$$

Pro: $\hat{f}_D \approx \hat{f}_{D/i}$

Con: n models train ⇒ Slow

k-fold Cross validation

$$\text{Error}_1 \triangleq \frac{1}{n/k}\sum_{(x,y)\in D_1} (Y - \hat{f}_{D/D_1}(x))^2 \Rightarrow \text{Err}_{\text{validate}} \triangleq \sum_{i=1}^{k} \text{Error}_i$$

Pro: only k-models trained ⇒ faster

Con: $\hat{f}_{D/D_1} \not\approx \hat{f}_D$ ($n - \frac{n}{k}$ vs. n samples)

LOO cross validation is (almost) unbiased estimate!

> When computing **LOOCV error, we only use $n-1$ data points**
  - So it's not estimate of true error of learning with n data points
  - learning with less data typically gives worse answer $\Rightarrow$ Usually validation error is **pessimistic**
  - but that bias is small, since $n-1$ is very close to n

> LOO is almost unbiased, and it is common to use LOO error for **model class** selection
  - E.g., picking degree is a model class selection since, for example, a set of all degree-5 polynomial functions is a **model class**

> But, LOOCV requires a lot of computational time
  - Suppose you have 100,000 data points
  - You implemented a great version of your learning algorithm that Learns in only 1 second
  - Computing LOO will take about 1 day.

Use *k*-fold cross validation

- Randomly **divide training data into *k* equal parts**
  - $D_1,\ldots,D_k$

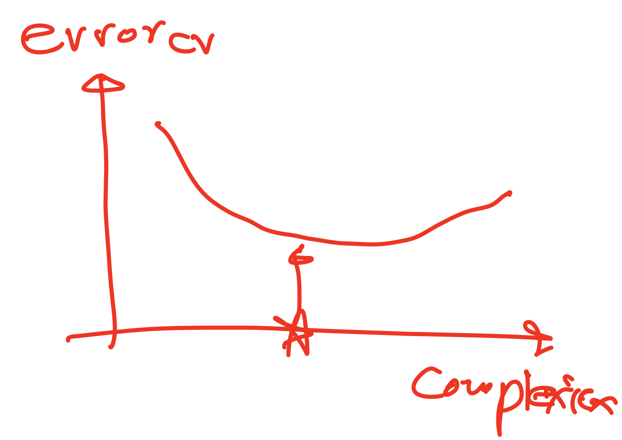
1	2	3	4	5
Train	Train	Validation	Train	Train

- For each *i*
  - Learn model $f_{D\backslash D_i}$ using data point not in D_i
  - Estimate error of $f_{D\backslash D_i}$ on validation set D_i:

$$\text{error}_{D_i}(f_{D\backslash D_i}) = \frac{1}{|D_i|} \sum_{(x_j,y_j)\in D_i} (y_j - f_{D\backslash D_i}(x_j))^2$$

- k-fold cross validation error is average over data splits:

$$\text{error}_{k-fold} = \frac{1}{k} \sum_{i=1}^{k} \text{error}_{D_i}(f_{D\backslash D_i})$$

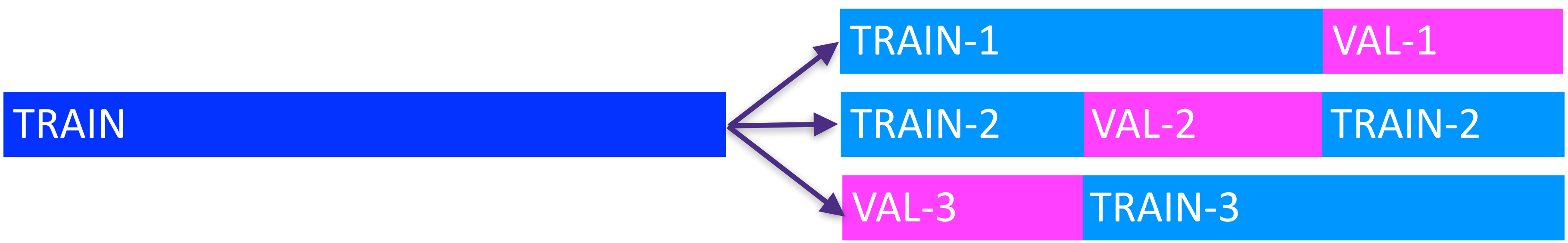


- k-fold cross validation properties:
  - Much faster to compute than LOO
  - More (pessimistically) biased – using much less data, only $\frac{k-1}{k}n$
  - Usually, **k = 10**

Recap

> Given a dataset, begin by splitting into

> Model selection: Use k-fold cross-validation on **TRAIN** to train predictor and choose hyper-parameters such as degree

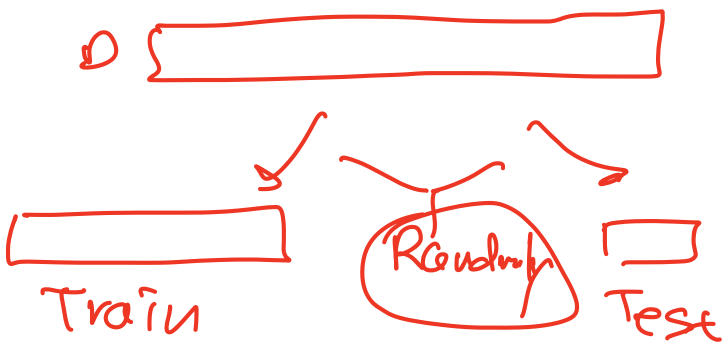


> Model assessment: Use **TEST** to assess the accuracy of the model you output

- **Never train or choose hyper-parameters based on the test data**

Example 1

- You wish to predict the stock price of zoom.us given historical stock price data
- You use all daily stock price up to Jan 1, 2020 as TRAIN and Jan 2, 2020 - April 13, 2020 as TEST
- What's wrong with this procedure?

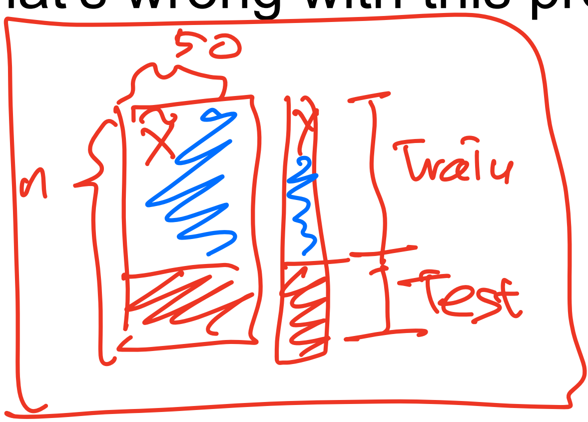


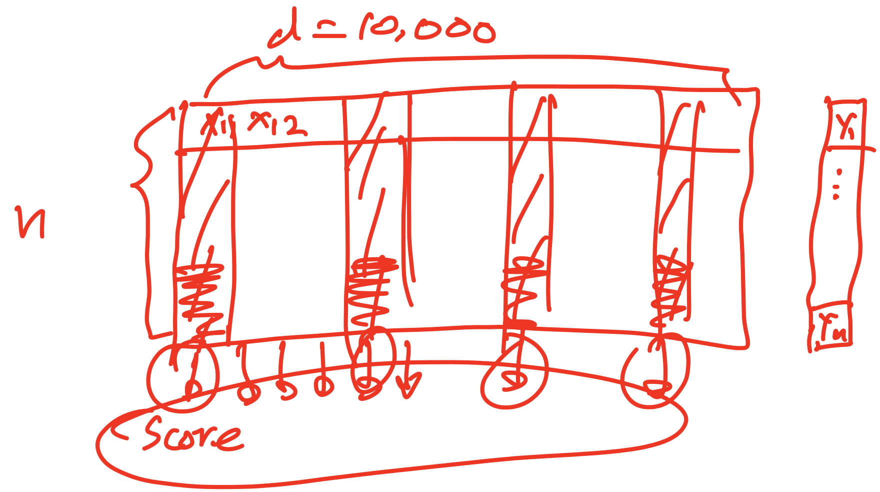
Example 2

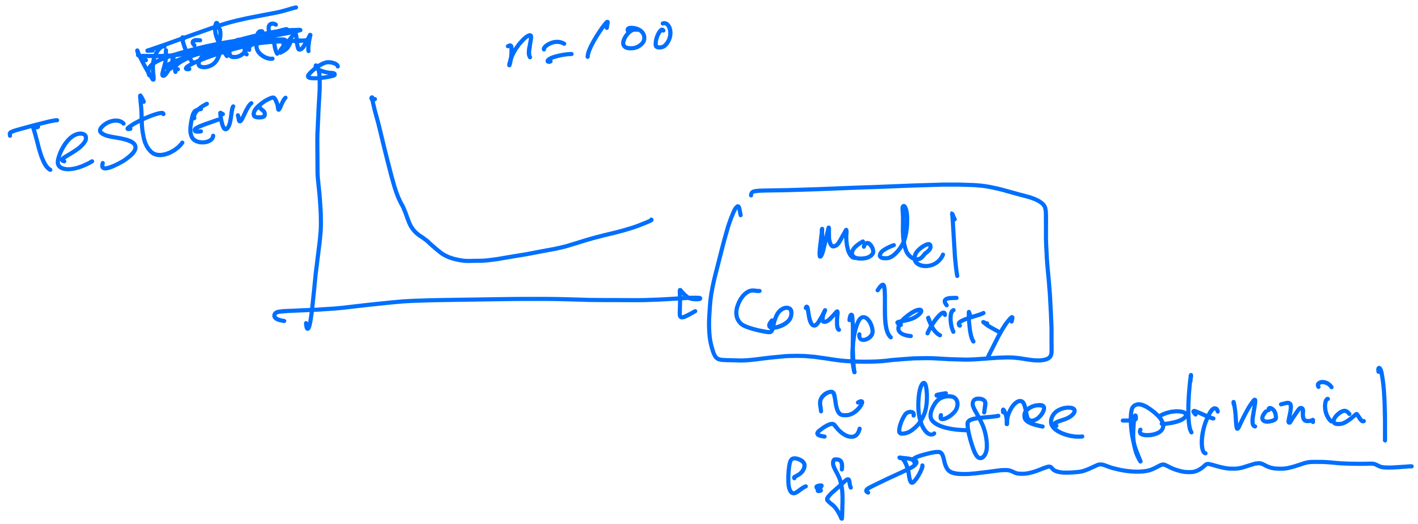
- Given 10,000-dimensional data and n examples, we pick a subset of 50 dimensions that have the highest correlation with labels in the entire dataset:

50 indices j that have largest $$\frac{\left|\sum_{i=1}^{n} x_{i,j} y_i\right|}{\sqrt{\sum_{i=1}^{n} x_{i,j}^2}}$$

- After picking our 50 features, we then break data into train and test dataset.
- We train linear regression on these selected features on the training set. We compute the test error and report it
- What's wrong with this procedure?



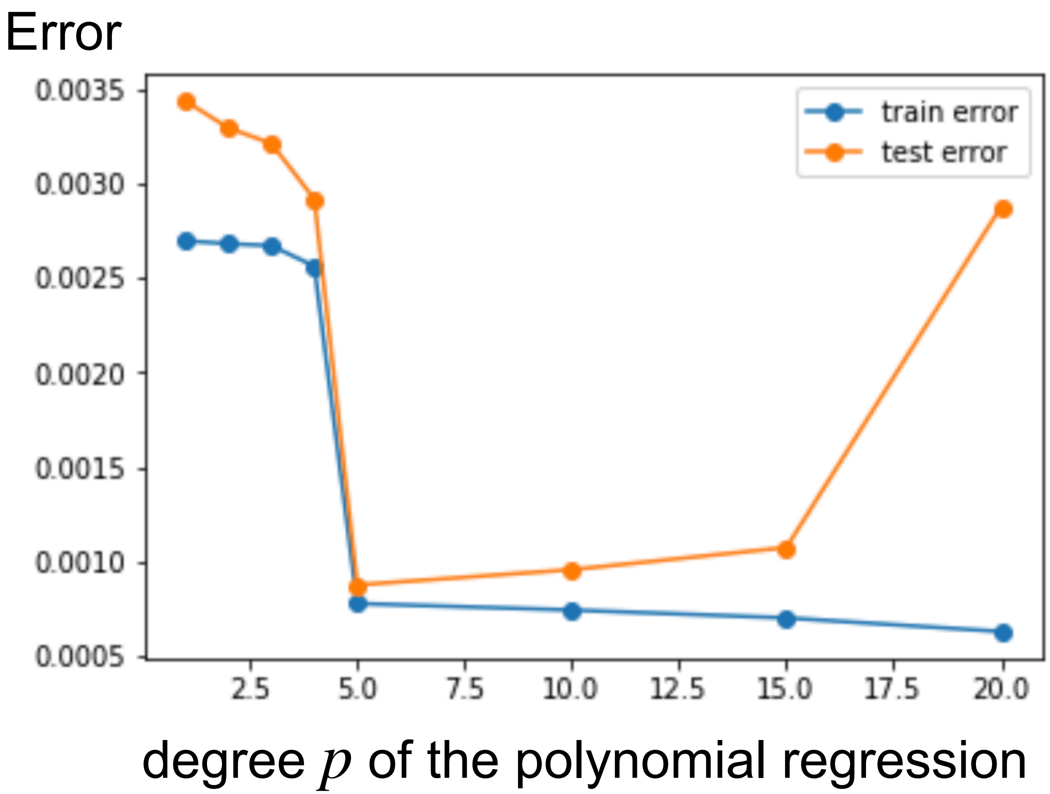


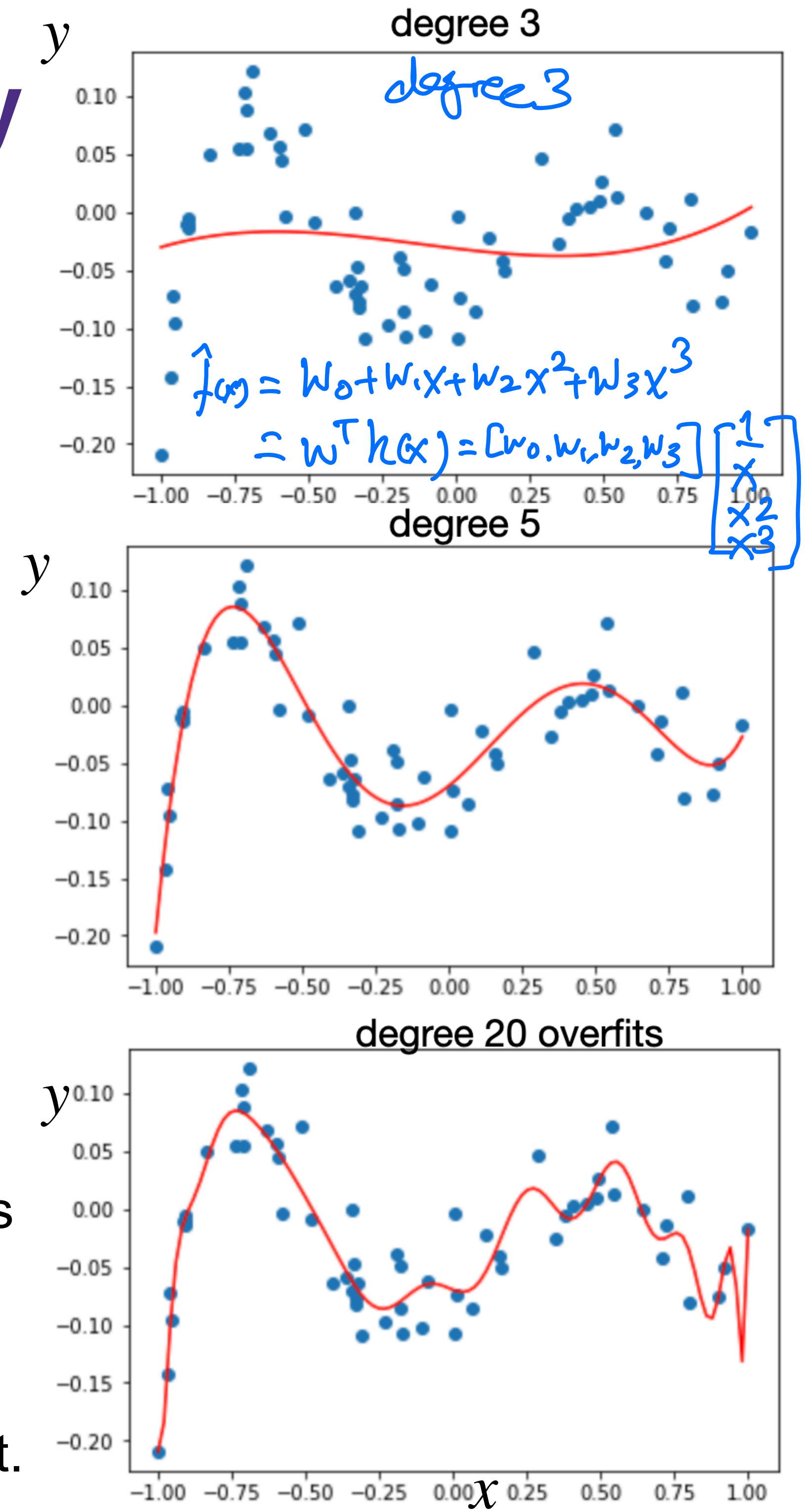


Bias-Variance Tradeoff

- explaining test error using theoretical analysis

Train/test error vs. complexity





- **Test error** has a U shape as we change the **model complexity**
- We want to theoretically explain and understand this important phenomenon in machine learning
- This is called **bias-variance tradeoff**
- Let's start with what an **optimal predictor** can achieve, and how practical predictor deviates from it.

Optimal prediction

Typical notation for this lecture:
X denotes a random variable
x denotes a deterministic instance

- Suppose data is generated from a statistical model $(X, Y) \sim P_{X,Y}$
  - and assume we know $P_{X,Y}$ (just for now to explain statistical learning)
- Then **learning** is to find a predictor $\eta : \mathbb{R}^d \to \mathbb{R}$ that minimizes
  $X \mapsto \eta(X)$
  - the expected error $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2]$ = Test Error (True Error)
    squared error
  - think of this random (X, Y) as a new sample you will encounter when you deployed your learned model, and we care about its average performance

η is not restricted to any class
there is no sample, we know $P_{X,Y}$

Optimal prediction

Typical notation:
X denotes a random variable
x denotes a deterministic instance

- Suppose data is generated from a statistical model $(X, Y) \sim P_{X,Y}$
  - and assume we know $P_{X,Y}$ (just for now to explain statistical learning)
- Then **learning** is to find a predictor $\eta : \mathbb{R}^d \to \mathbb{R}$ that minimizes
  - the expected error $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2]$
  - think of this random (X, Y) as a new sample you will encounter when you deployed your learned model, and we care about its average performance
- Since, we do not assume anything about the function $\eta(x)$, it can take any value for each $X = x$,
  - for example $\eta(1.0)$ has nothing to do with $\eta(1.1)$
- hence we can try to find the optimal prediction $\eta(x)$ for each value of $X = x$ separately

Typical notation:
X denotes a random variable
x denotes a deterministic instance

Optimal prediction

- Suppose data is generated from a statistical model $(X, Y) \sim P_{X,Y}$
  - and assume we know $P_{X,Y}$ (just for now to explain statistical learning)
- Then **learning** is to find a predictor $\eta : \mathbb{R}^d \to \mathbb{R}$ that minimizes
  - the expected error $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2]$
  - think of this random (X, Y) as a new sample you will encounter when you deployed your learned model, and we care about its average performance
- Since, we do not assume anything about the function $\eta(x)$, it can take any value for each $X = x$, hence we can try to find the optimal prediction $\eta(x)$ for each value of $X = x$ separately
  - $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2] = \mathbb{E}_{X\sim P_X}\left[\mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]\right]$

$$= \int \mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x] \, P_X(x) \, dx$$

Or for discrete X,

$$= \sum_x \mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x] \, P_X(x)$$

Where we used the chain rule: $\mathbb{E}_{X,Y}[f(X, Y)] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[f(x, Y) \,|\, X = x]\right]$

Optimal prediction

- To find an optimal predictor, we can solve the optimization for each $X = x$ separately
  - $\eta(x) \; = \; \arg\min_{a_x \in \mathbb{R}} \mathbb{E}_{Y \sim P_{Y|X}}[(Y - a_x)^2 \,|\, X = x] = f(a)$
- The optimal solution is $\eta(x) = \mathbb{E}_{Y \sim P_{Y|X}}[Y \,|\, X = x]$, which is the best prediction in ℓ_2-loss/Mean Squared Error
- Claim: $\mathbb{E}_{Y \sim P_{Y|X}}[Y \,|\, X = x] \; = \; \arg\min_{a_x \in \mathbb{R}} \mathbb{E}_{Y \sim P_{Y|X}}[(Y - a_x)^2 \,|\, X = x]$
- Proof: $0 = \frac{\partial f(a)}{\partial a} = \mathbb{E}_{Y \sim P_{Y|X}}[-2(Y - a) \,|\, X = x] = 0$

  $\Rightarrow \mathbb{E}_{Y \sim P_{Y|X}}[Y|x] - \mathbb{E}_{Y \sim P_{Y|X}}[a|x] = 0$

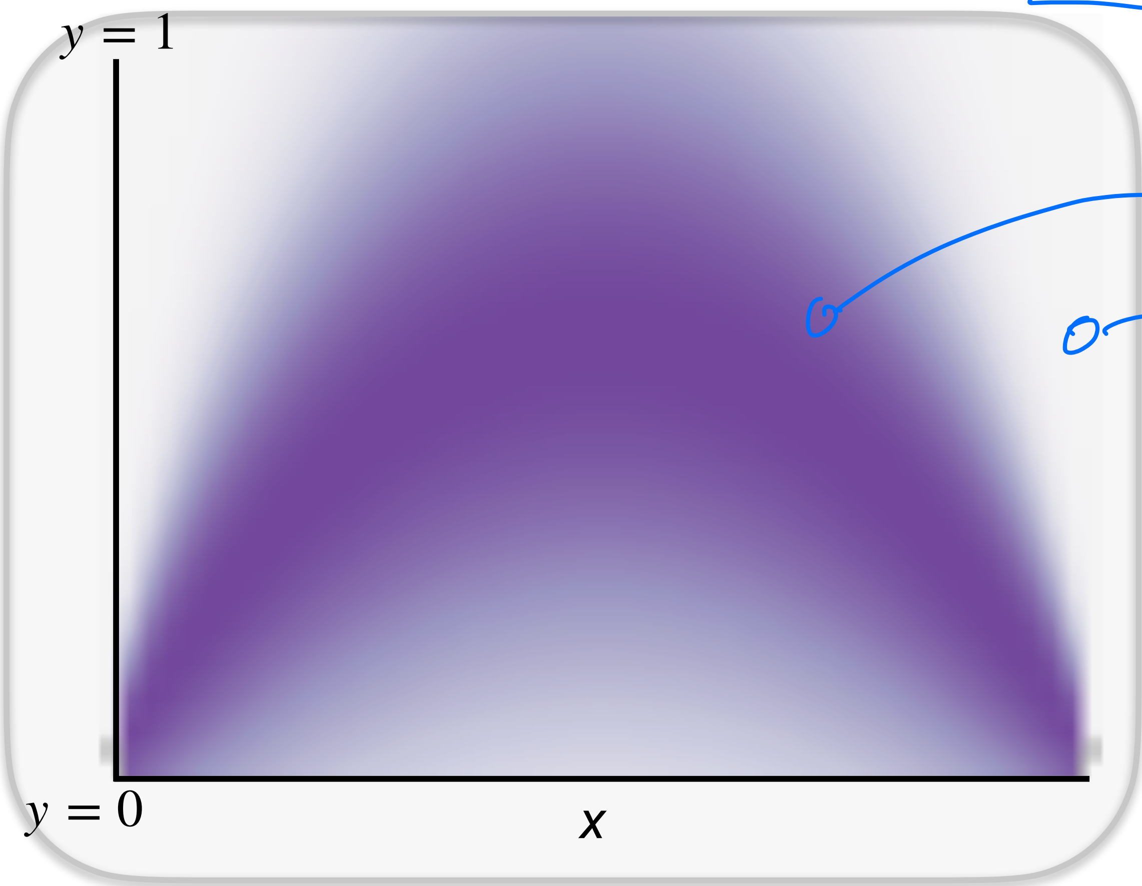
  $a = \mathbb{E}_{Y \sim P_{Y|X}}[Y \,|\, X = x]$
- Note that this optimal statistical estimator $\eta(x) = \mathbb{E}[Y \,|\, X = x]$ cannot be implemented as we do not know $P_{X,Y}$ in practice
- This is only for the purpose of conceptual understanding

Statistical Learning

- Consider a joint distribution $P_{XY}(x, y)$
- Our goal is to find the optimal predictor $\eta(x)$ to minimize $\mathbb{E}_{(X,Y)\sim P_{XY}}[(Y - \eta(X))^2]$

$= \mathbb{E}[Y \mid X = x]$

$$P_{XY}(X = x, Y = y)$$

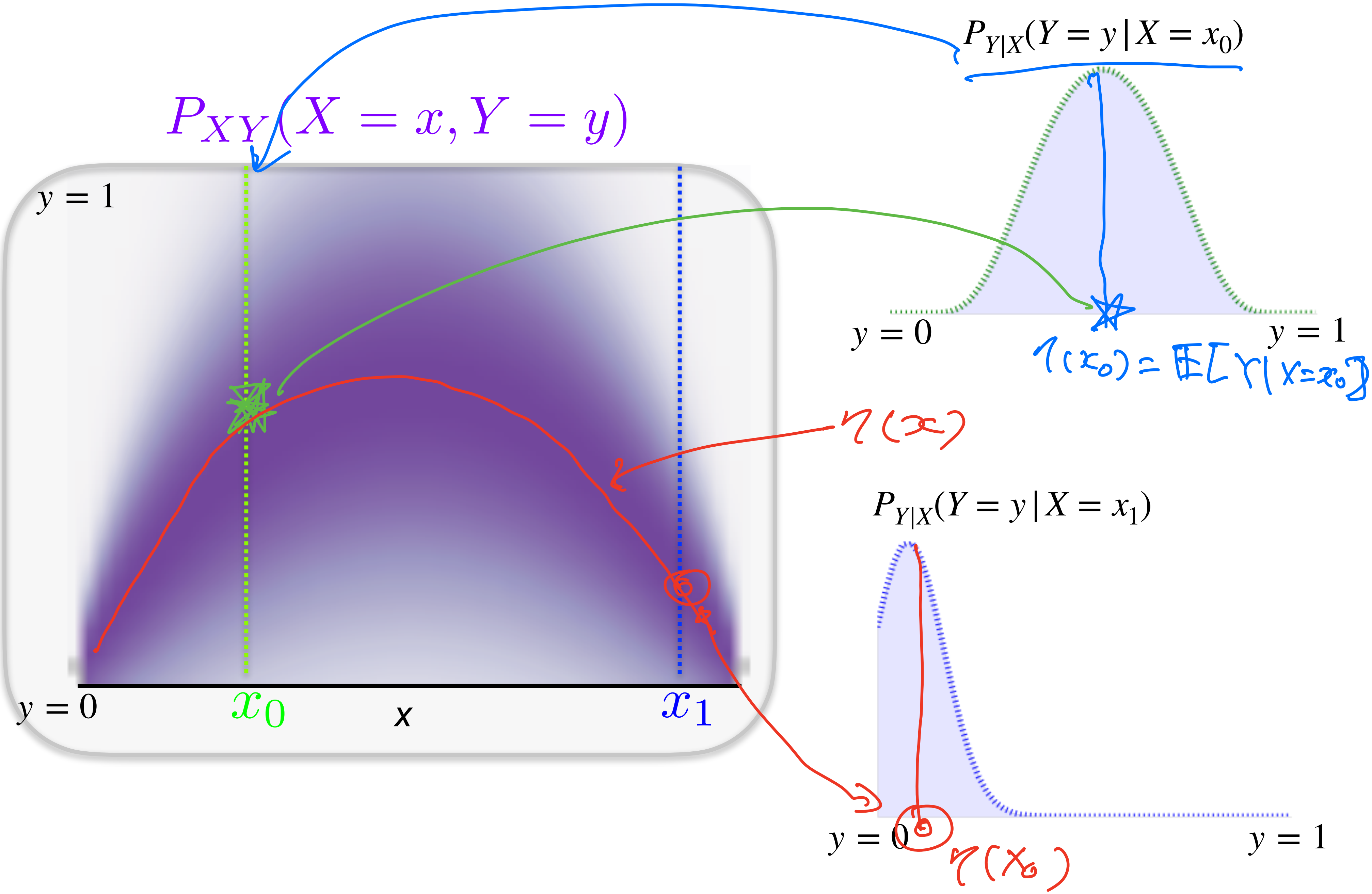


Darker = larger $P_{XY}(x,y)$

lighter = $P_{XY}(x,y) \downarrow$

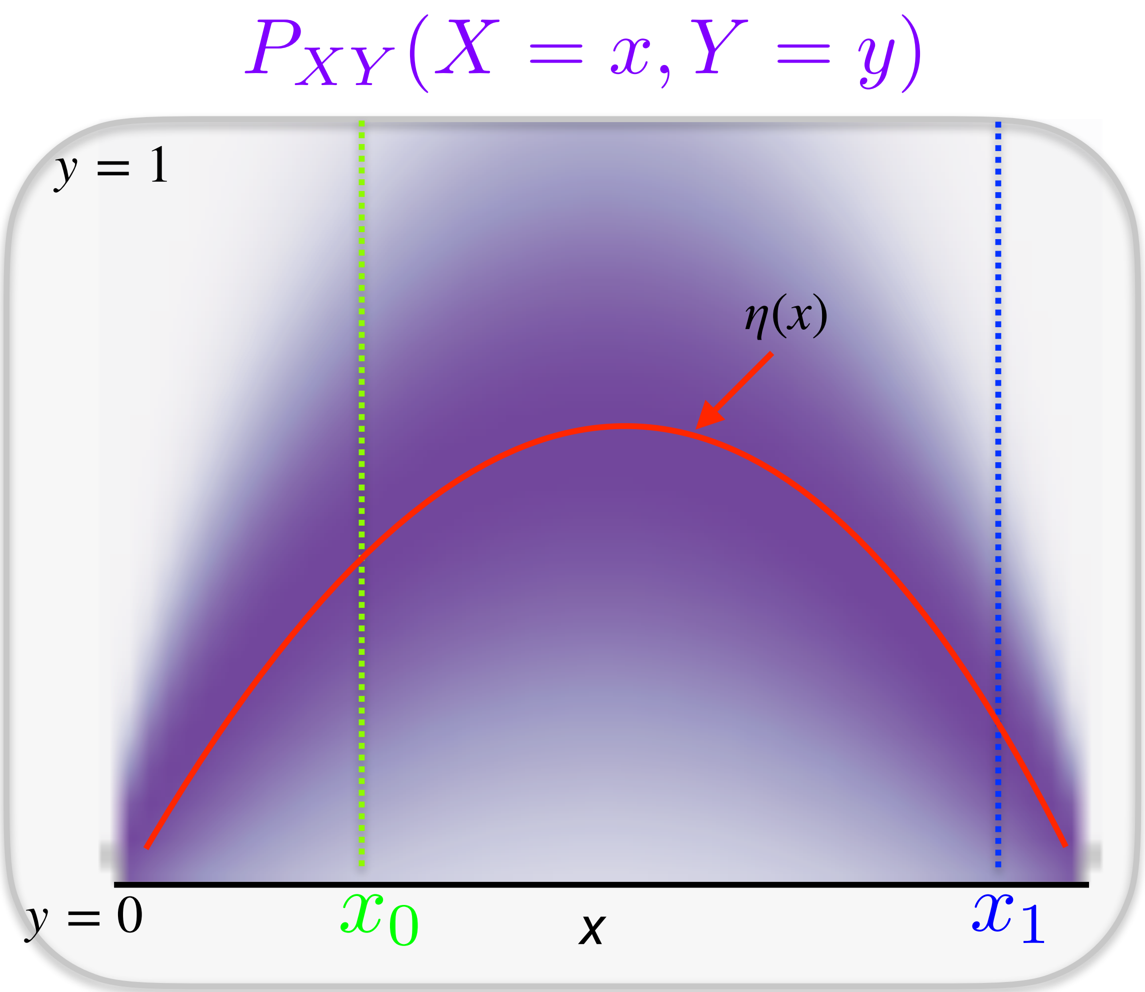
Statistical Learning

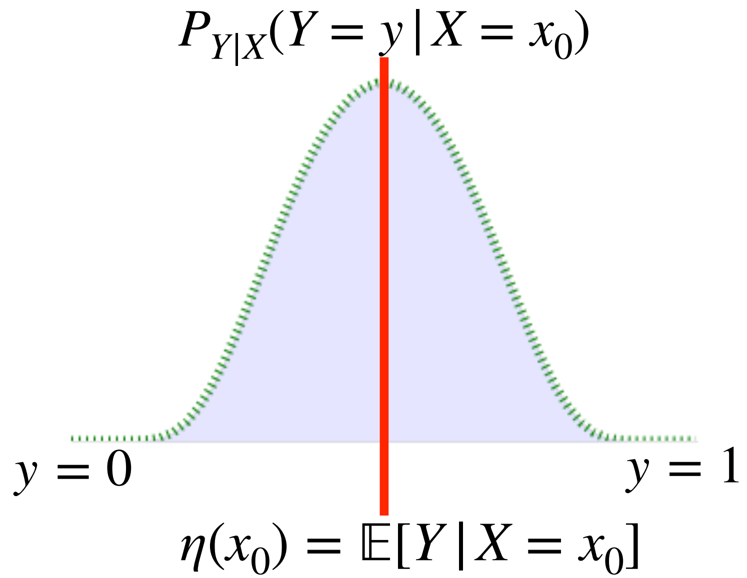
- The optimal predictor is given by $\eta(x) = \mathbb{E}_{P_{Y|X}}[Y \mid X = x]$

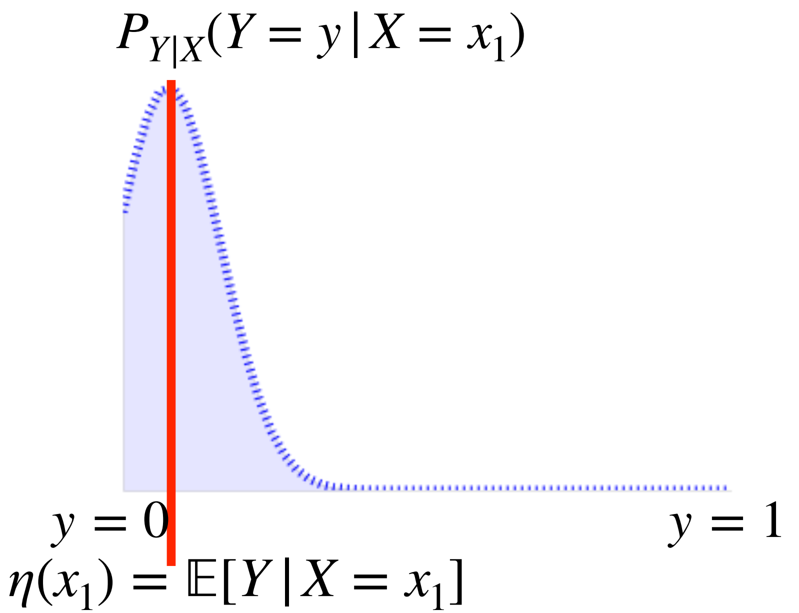


Statistical Learning

- The optimal predictor is given by
  $\eta(x) = \mathbb{E}_{P_{Y|X}}[Y \mid X = x]$

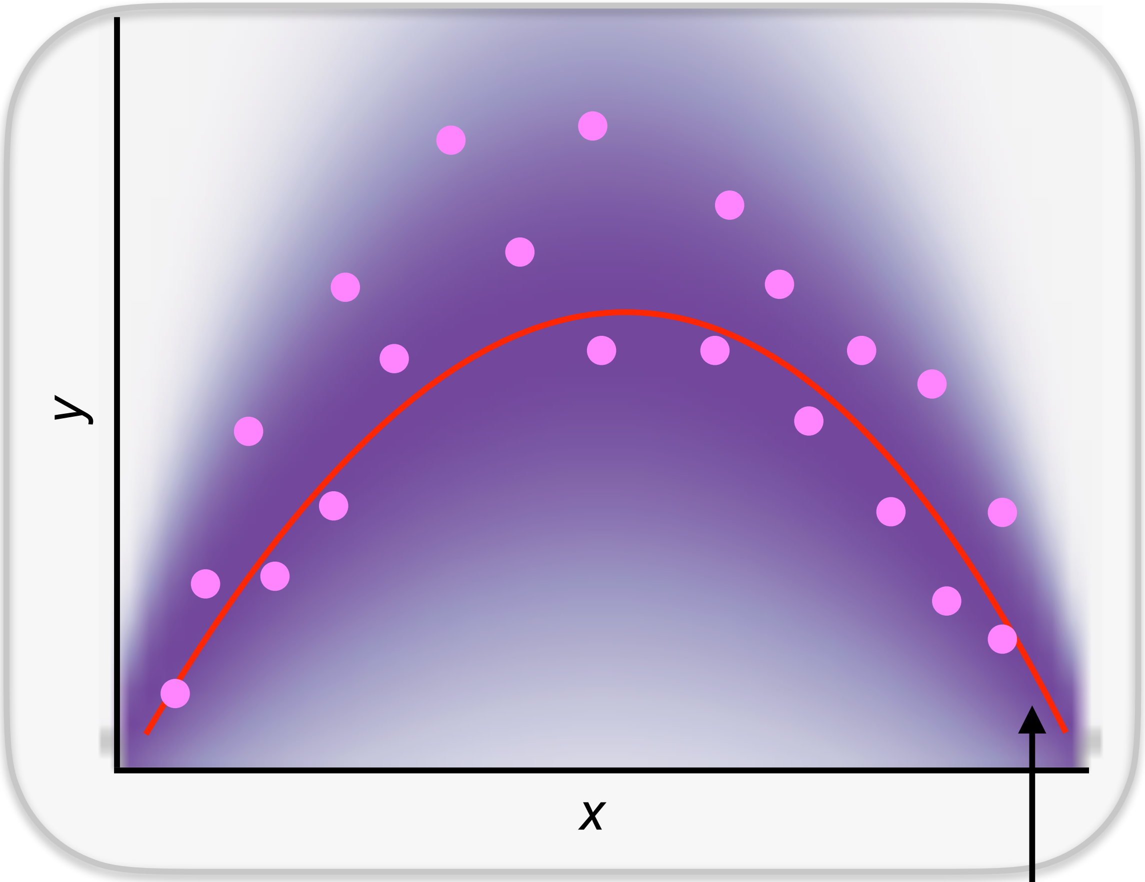






Statistical Learning

$P_{XY}(X = x, Y = y)$



But we do not know $P_{X,Y}$,

and so we cannot compute $\eta(x)$.

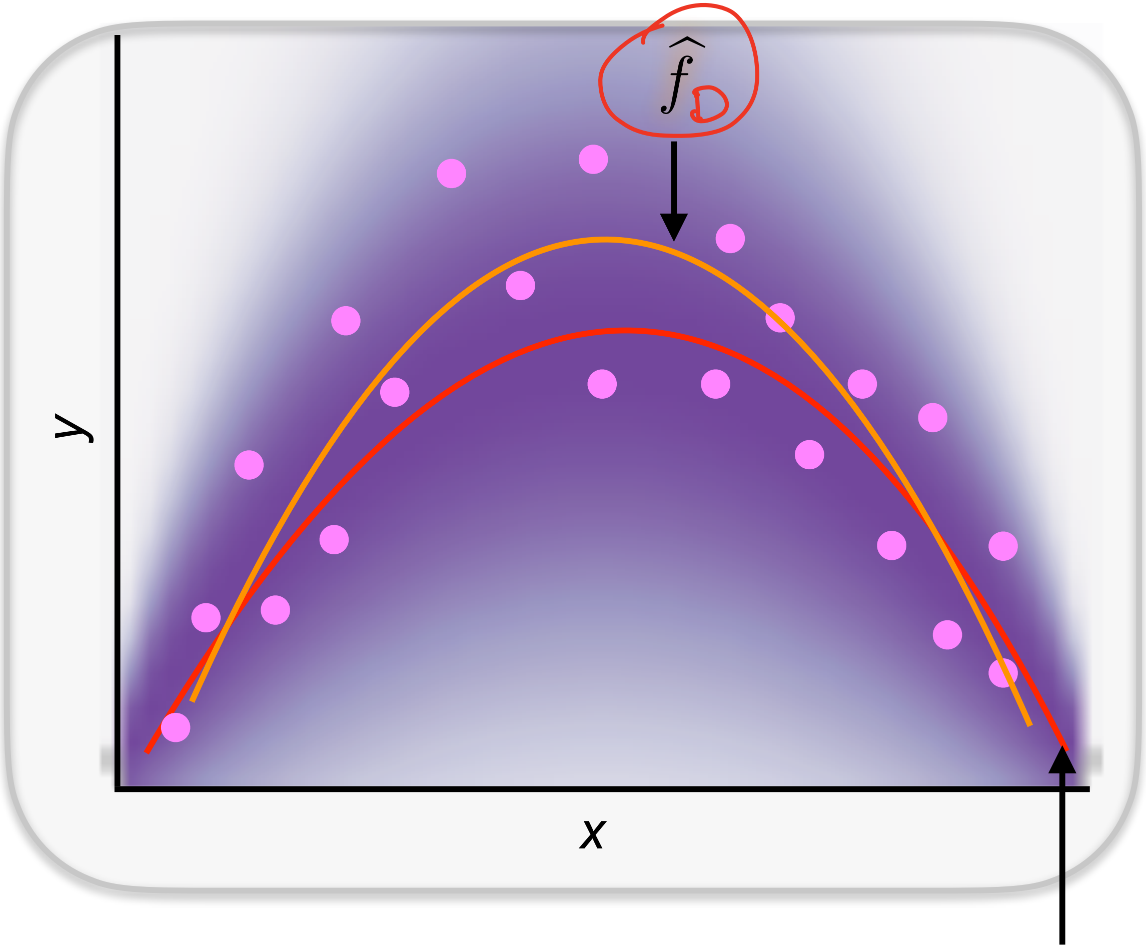
We only have samples.

What can we do?

$\eta(x) = \mathbb{E}_{Y|X}[Y \,|\, X = x]$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



$$\mathbb{E}_{Y|X}[Y|X = x]$$

Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \ldots, n$$

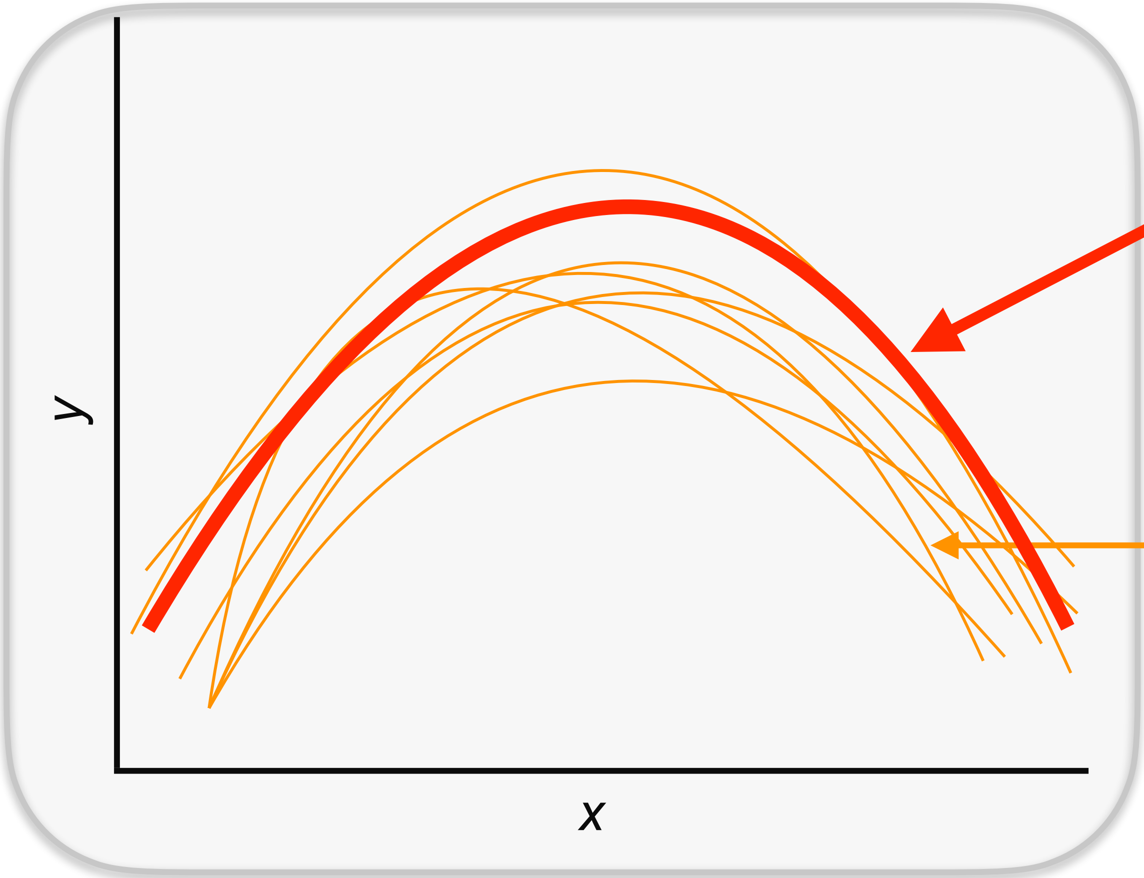
So we need to restrict our predictor to **a function class** (e.g., linear, degree-p polynomial) to avoid overfitting and **minimize empirical error**:

$$\hat{f}_D = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

We care about how our predictor performs on future unseen data

True Error of $\hat{f}$: $\mathbb{E}_{X,Y}[(Y - \hat{f}_D(X))^2]$

True error $\mathbb{E}_{X,Y}[(Y - \hat{f}_D(X))^2]$ is random because $\hat{f}_D$ is random (whose randomness comes from training data D)



Optimal predictor

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

Learned predictor

$$\hat{f}_{\mathcal{D}} = \arg\min_{f \in \mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

- We are interested in the **True Error** of a (random) learned predictor:

$$\mathbb{E}_{X,Y}[(Y - \hat{f}_{\mathcal{D}}(X))^2]$$

Each draw of $D = \{(x_i, y_i)\}_{i=1}^n$ results in different $\hat{f}_D$

Demo bias-variance trade-off

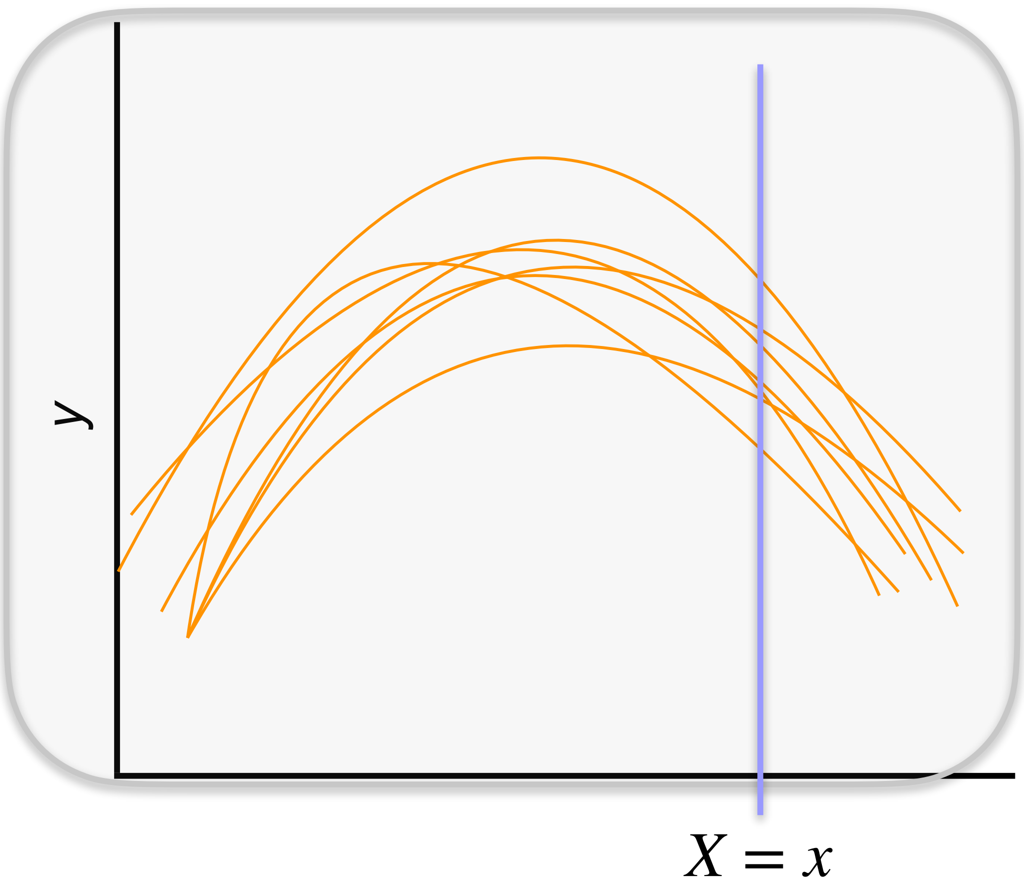
Exercise: Given joint distribution $(X, Y) \sim \mathcal{N}([\mu_X, \mu_Y], \begin{bmatrix} \Sigma_X, \Sigma_{XY} \\ \Sigma_{XY}, \Sigma_Y \end{bmatrix})$,

- is $E_{X,Y}[X + Y]$ a random variable or not?
- is $E_{X|Y}[X + Y \mid Y]$ a random variable or not?
- is $E_{X|Y}[X + Y \mid Y = y]$ a random variable or not?

True error $\mathbb{E}_{X,Y}[(Y - \hat{f}_D(X))^2]$ is random because $\hat{f}_D$ is random (whose randomness comes from training data D)

- We are interested in the **True Error** of a (random) learned predictor:

$$\mathbb{E}_{X,Y}[(Y - \hat{f}_{\mathcal{D}}(X))^2]$$



- But the analysis can be done for each $X = x$ separately, so we analyze the conditional **true error**:

$$\mathbb{E}_{Y|X}[(Y - \hat{f}_{\mathcal{D}}(x))^2 \,|\, X = x]$$

Each draw of $D = \{(x_i, y_i)\}_{i=1}^n$ results in different $\hat{f}_D$

True error $\mathbb{E}_{X,Y}[(Y - \hat{f}_D(X))^2]$ is random because $\hat{f}_D$ is random (whose randomness comes from training data D)

- We are interested in the **True Error** of a (random) learned predictor:

$$\mathbb{E}_{X,Y}[(Y - \hat{f}_{\mathcal{D}}(X))^2]$$

- But the analysis can be done for each $X = x$ separately, so we analyze the conditional **true error**:

$$\mathbb{E}_{Y|X}[(Y - \hat{f}_{\mathcal{D}}(x))^2 \,|\, X = x]$$

- And we care about the **average conditional true error**, averaged over training data:

$$\mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_{Y|X}[(Y - \hat{f}_{\mathcal{D}}(x))^2 \,|\, X = x]\Big]$$

written compactly as

$$= \mathbb{E}_{\mathcal{D},Y|x}[(Y - \hat{f}_{\mathcal{D}}(x))^2]$$

$$= \mathbb{E}[(Y - \hat{f}_{\mathcal{D}}(x))^2]$$

Each draw of $D = \{(x_i, y_i)\}_{i=1}^n$ results in different $\hat{f}_D$

Bias-variance tradeoff

Ideal predictor

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

Learned predictor

$$\hat{f}_{\mathcal{D}} = \arg\min_{f \in \mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

- **Average conditional true error**:

$$\mathbb{E}_{\mathcal{D}, Y|x}[(Y - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}, Y|x}[(\underbrace{Y - \eta(x)}_{A} + \underbrace{\eta(x) - \hat{f}_{\mathcal{D}}(x)}_{B})^2]$$

$\mathbb{E}A = 0.$

$\mathbb{E}[Y - \eta(x)] = \mathbb{E}[Y] - \eta(x)$

by (*)

$$= \underbrace{\mathbb{E}[(Y - \eta(x))^2]}_{\text{Irreducible Error}} + \underbrace{\mathbb{E}[(\eta(x) - \hat{f}_{D}(x))^2]}_{\text{Estimation Error}}$$

$\mathbb{E}[B] = 0$, $A \perp\!\!\!\perp B$ independent.

$$\mathbb{E}[(A+B)^2] = \mathbb{E}[A^2] + 2\mathbb{E}[AB] + \mathbb{E}[B^2] \quad (*)$$

$2\mathbb{E}[AB] = 2\mathbb{E}[A] \cdot \mathbb{E}[B]$ ← indep.

$= 0$ ← $\mathbb{E}[B] = 0$

Bias-variance tradeoff

Bias-variance tradeoff

Ideal predictor

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

Learned predictor

$$\hat{f}_{\mathcal{D}} = \arg\min_{f \in \mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

- **Average conditional true error**:

$$\mathbb{E}_{\mathcal{D},Y|x}[(Y - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D},Y|x}[(Y - \eta(x) + \eta(x) - \hat{f}_{\mathcal{D}}(x))^2]$$

$$= \mathbb{E}_{\mathcal{D},Y|x}\left[(Y - \eta(x))^2 + 2(Y - \eta(x))(\eta(x) - \hat{f}_{\mathcal{D}}(x)) + (\eta(x) - \hat{f}_{\mathcal{D}}(x))^2 \right]$$

$$= \mathbb{E}_{Y|x}[(Y - \eta(x))^2] + 2\mathbb{E}_{\mathcal{D},Y|x}[\underbrace{(Y - \eta(x))}_{\text{expectation } 0}(\eta(x) - \hat{f}_{\mathcal{D}}(x))] + \mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]$$

(this follows from independence of $\mathcal{D}$ and (X, Y) and $\mathbb{E}_{Y|x}[Y - \eta(x)] = \mathbb{E}[Y|X = x] - \eta(x) = 0$)

$$= \mathbb{E}_{Y|x}[(Y - \eta(x))^2] + \mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]$$

Irreducible error

Caused by

(a) stochastic label noise in $P_{Y|X=x}$

(b) cannot be avoided

Average learning error

Caused by

(a) either using too "simple" of a model or

(b) not enough data to learn the model accurately

Bias-variance tradeoff

Ideal predictor

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

Learned predictor

$$\hat{f}_{\mathcal{D}} = \arg\min_{f \in \mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

- **Average learning error**:

$$\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}\Big[\Big(\underbrace{\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]}_{A} + \underbrace{\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)}_{B}\Big)^2\Big]$$

by (*)

$$= \underbrace{\mathbb{E}[(\eta(x) - \mathbb{E}[\hat{f}_D(x)])^2]}_{A^2} + \underbrace{\mathbb{E}[(\mathbb{E}[\hat{f}_D(x)] - \hat{f}_D(x))^2]}_{B^2}$$

Bias Squared

Variance of $\hat{f}_D(x)$

Bias-variance tradeoff

Ideal predictor

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

Learned predictor

$$\hat{f}_{\mathcal{D}} = \arg\min_{f \in \mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

- **Average learning error**:

$$\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}\Big[\big(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\big)^2\Big]$$

$$= \mathbb{E}_{\mathcal{D}}\Big[\big(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\big)^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)) + (\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2\Big]$$

$$= \underbrace{\big(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\big)^2}_{\text{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}\Big[\big(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\big)^2\Big]}_{\text{variance}}$$

biased squared **variance**

[Homework 1 Problem B2 derives similar trade-off for a specific estimator]

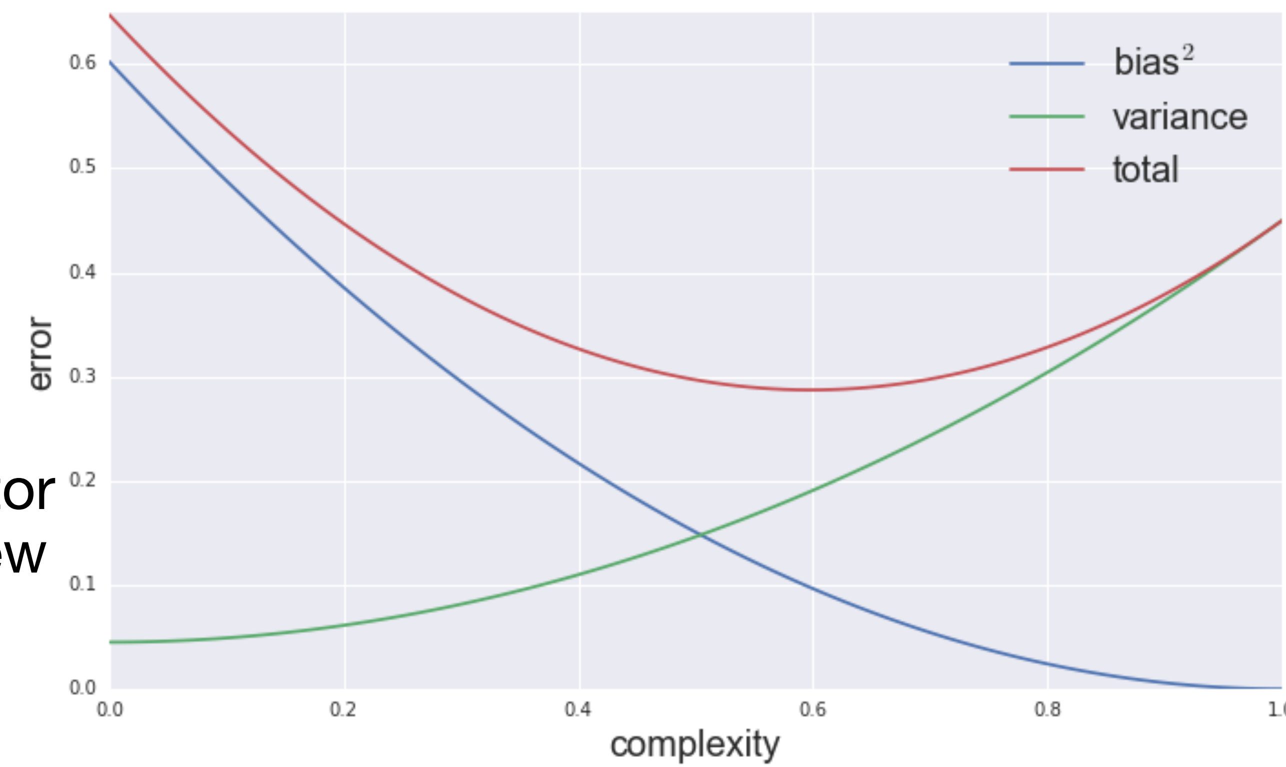
Bias-variance tradeoff

- **Average conditional true error**:

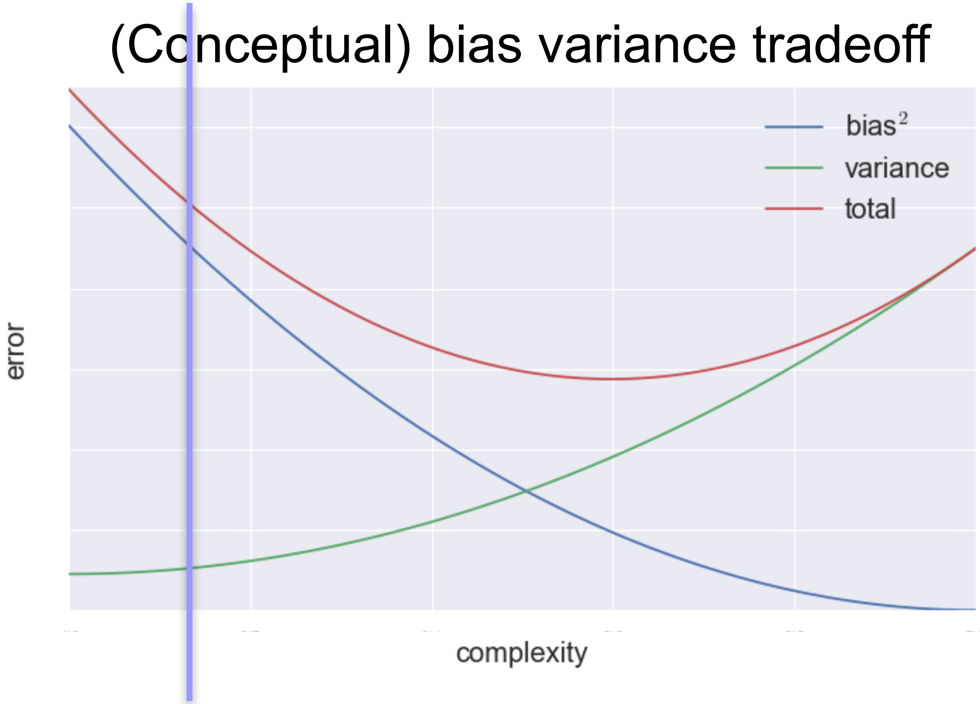
$$\mathbb{E}_{\mathcal{D},Y|x}[(Y - \hat{f}_{\mathcal{D}}(x))^2] = \underbrace{\mathbb{E}_{Y|x}\left[(Y - \eta(x))^2\right]}_{\textbf{irreducible error}}$$

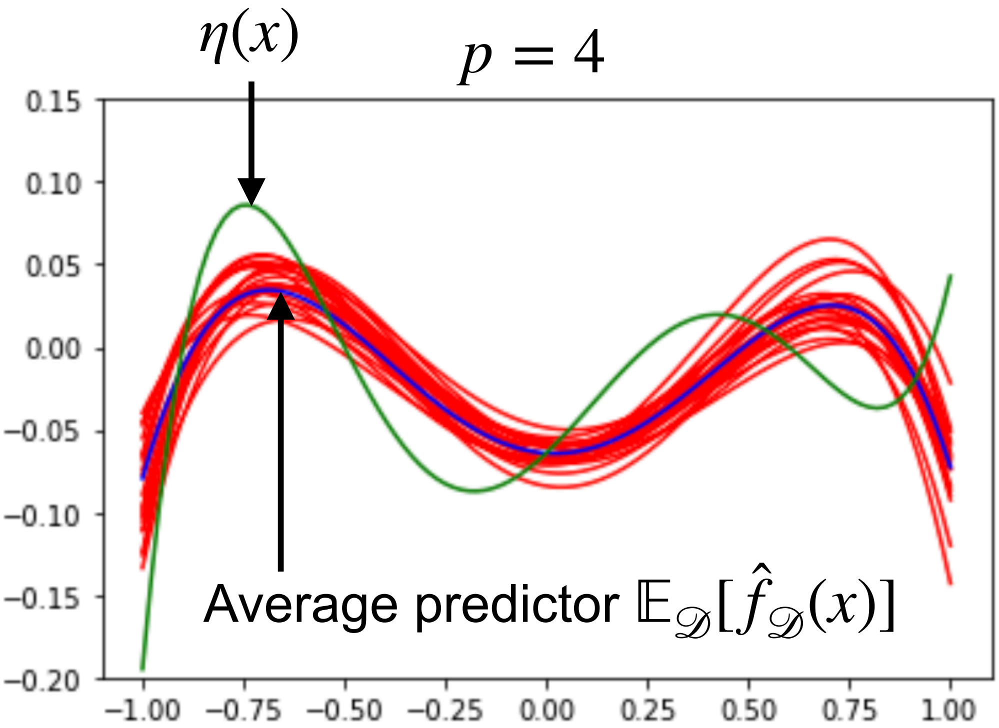
$$+ \underbrace{\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2}_{\textbf{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]}_{\textbf{variance}}$$

- **Bias squared:** measures how the predictor is mismatched with the best predictor in expectation
- **variance:** measures how the predictor varies each time with a new training datasets



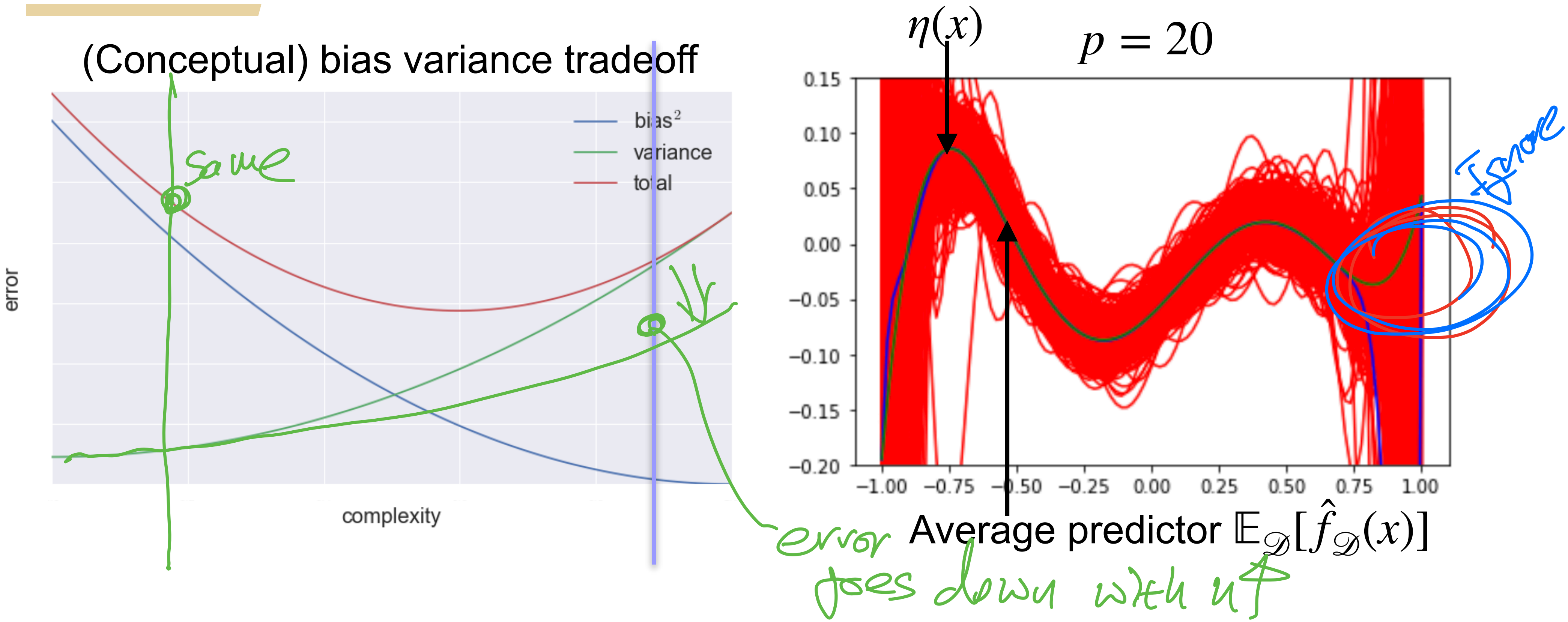
Recap: Bias-variance tradeoff with simple model





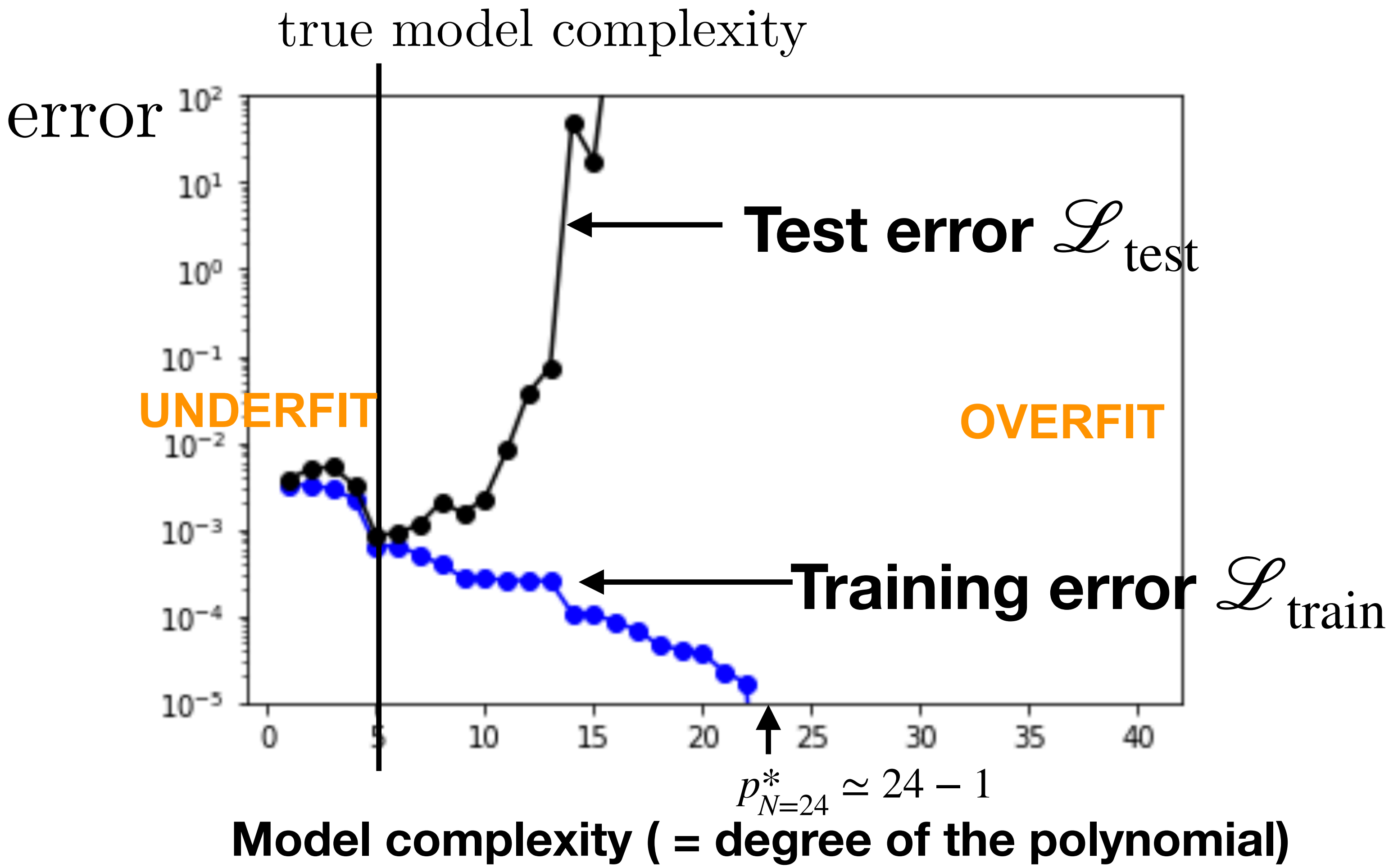
- When model **complexity is low** (lower than the optimal predictor $\eta(x)$)
  - Bias^2 of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2$, is large
  - Variance of our predictor, $\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]$, is small
  - If we have more samples, then
    - Bias
    - Variance
    - Because Variance is already small, overall test error

Recap: Bias-variance tradeoff with simple model



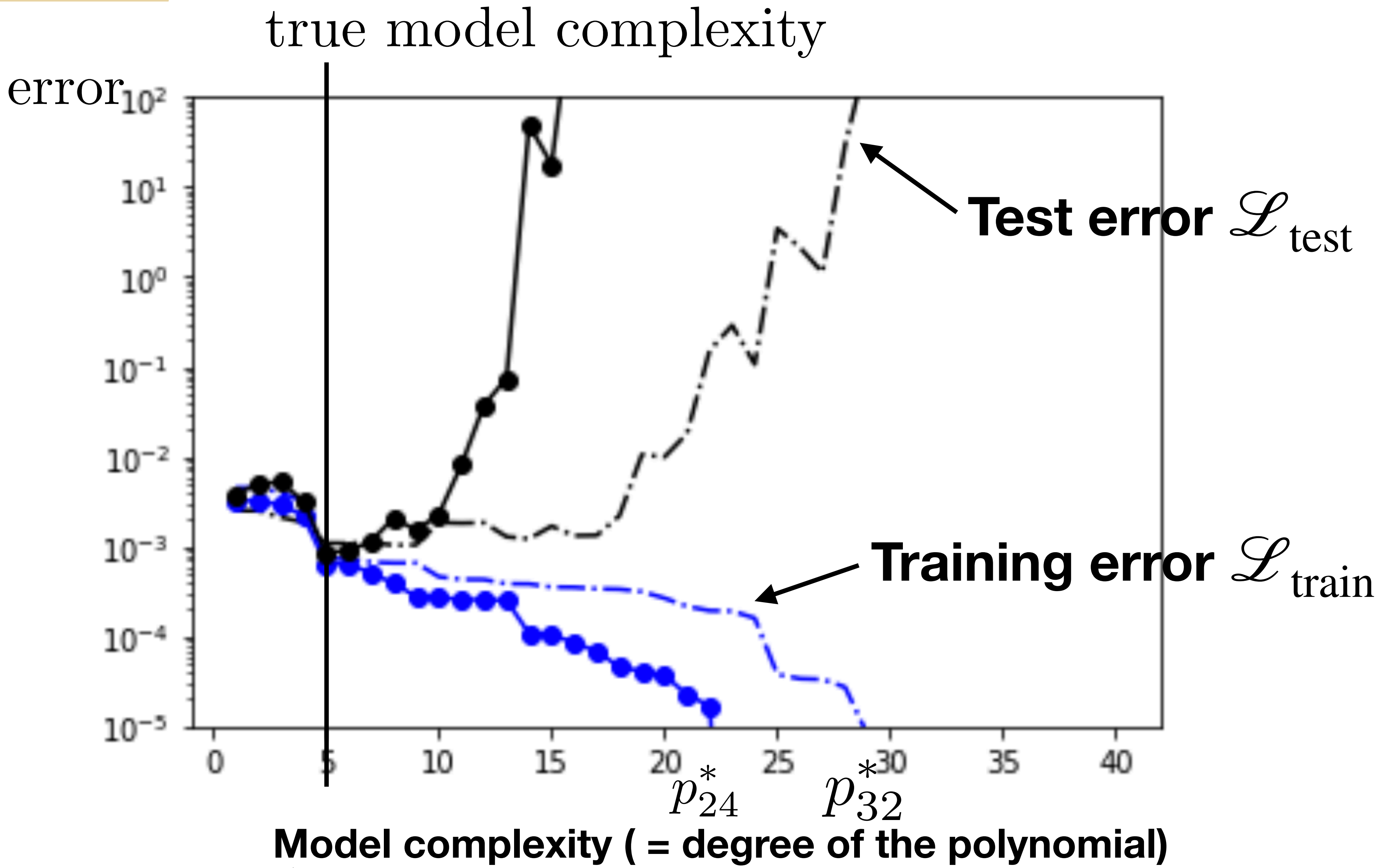
- When model complexity is high (higher than the optimal predictor $\eta(x)$)
  - Bias^2 of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2$, is small
  - Variance of our predictor, $\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]$, is large
  - If we have more samples, then
    - Bias does not change
    - Variance goes down
    - Because Variance is dominating, overall test error go down

- let us first fix sample size ***N=30***, collect one dataset of size N i.i.d. from a distribution, and fix one training set S_{train} and test set S_{test} via 80/20 split
- then we run multiple validations and plot the computed MSEs for all values of ***p*** that we are interested in



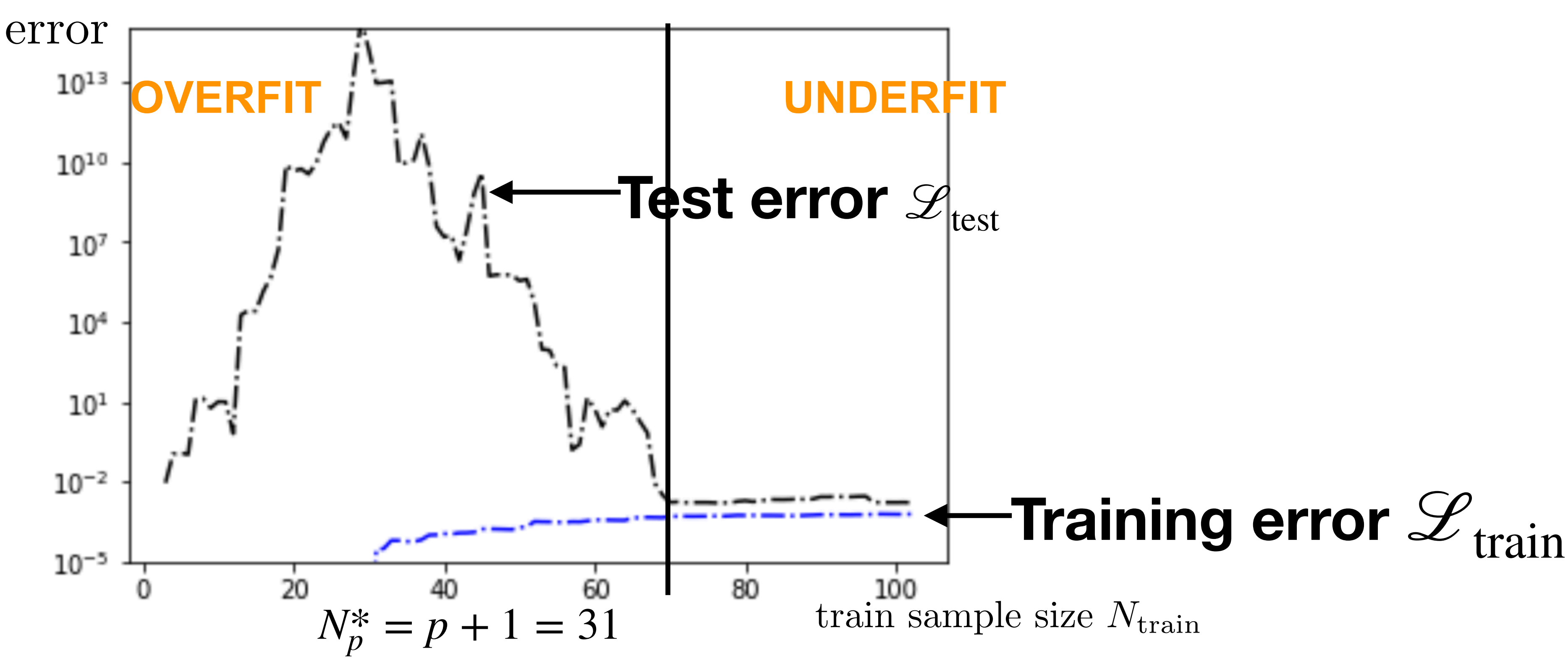
- Given sample size N there is a threshold, p^*_N, where training error is zero
- Training error is **always** monotonically non-increasing
- Test error has a trend of going down and then up, but fluctuates

- let us now repeat the process changing the sample size to **N=40** , and see how the curves change



- The threshold, p^*_N, moves right
- Training error tends to increase, because more points need to fit
- Test error tends to decrease, because Variance decreases

- let us now fix predictor model complexity ***p=30***, collect multiple datasets by starting with 3 samples and adding one sample at a time to the training set, but keeping a large enough test set fixed
- then we plot the computed MSEs for all values of train sample size ***Ntrain*** that we are interested in



- There is a threshold, N_p^*, below which training error is zero (extreme overfit)
- Below this threshold, test error is meaningless, as we are overfitting and there are multiple predictors with zero training error some of which have very large test error
- Test error tends to decrease
- Training error tends to increase

Questions?

- Good questions on Ed Discussion
  - Will we be tested in “bias”?
    - Bias shows up in many places, and you will have to know the concept. Anything that is taught in lectures can show up in the exams.
  - Why do we use x to denote a column vector and not a row vector?
  - $\theta*$ is the same as θ_*, it is just my writing that is not always consistent
  - What is $\theta*$?
    - The reason it is unnatural to think about $\theta*$ is that it is something that does not exist in reality. Only time it exists is when you generate simulated data yourself (like in lecture notes and homework).
    - The right interpretation is that we hypothesize that nature has chosen to generate the data from a distribution, which can be written as $P(\,\cdot\,;\theta*)$.
    - Whether this assumption is correct or not, we are deciding to go ahead with our MLE process.
    - That gives us some MLE estimate and corresponding distribution $P(\,\cdot\,;\theta^*_{\text{MLE}})$. What we do with it, and what we believe about it is up to us. (Hence you need to check your accuracy on a holdout set, which we will learn later)

[Homework 1 Problem A4 analyzes similar bias-variance tradeoffs]