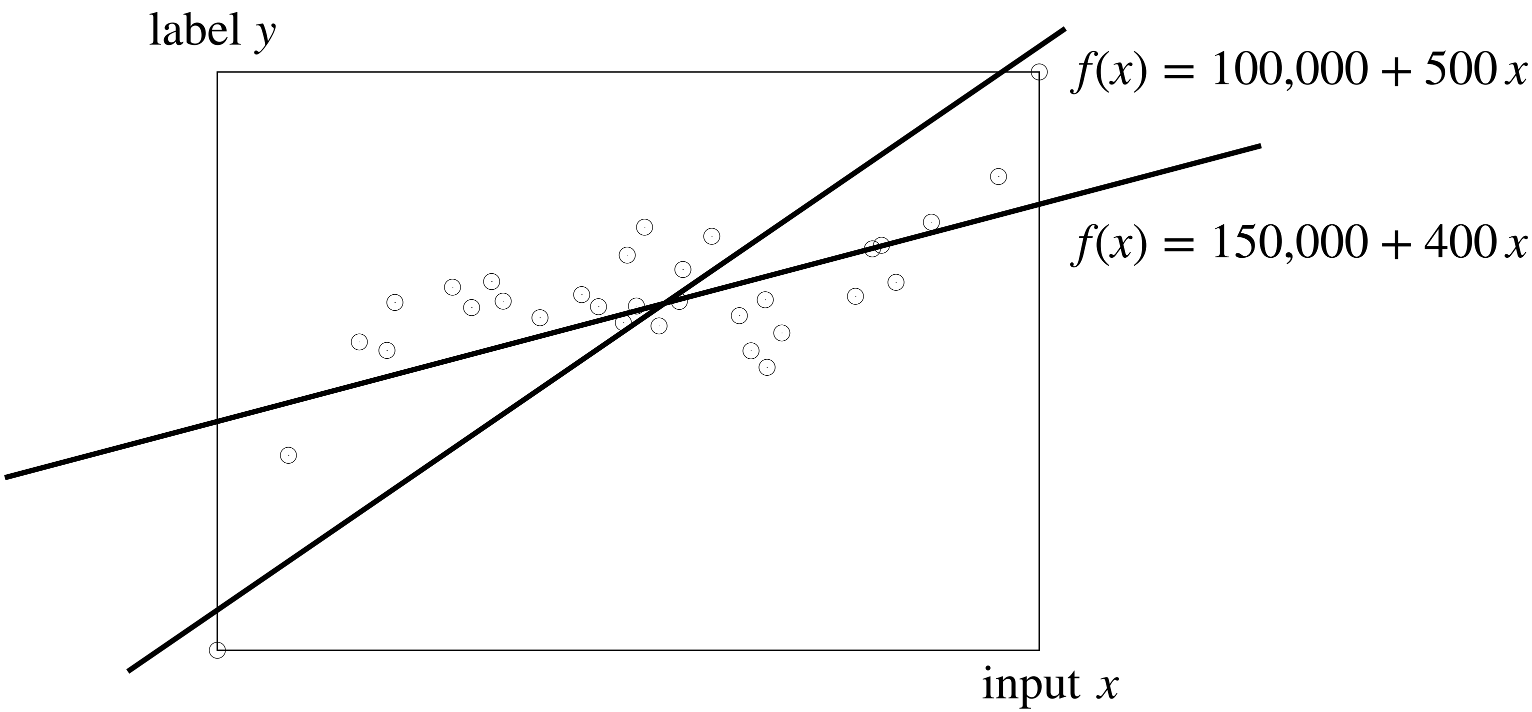
- HW0 due Tuesday midnight
- Extra office hours
  - **Monday:**
    - Tim Li, 10:30 - 11:30
    - **Sewoong Oh, 12:30 - 1:30**
    - Hugh Sun, 14:30 - 15:30
  - **Tuesday:**
    - Josh Gardner, 9:00 - 10:00
    - Hugh Sun, 14:30 - 15:30
    - Jakub Filipek, 16:00 - 17:00
    - **Pemi Nguyen, 17:00 - 20:00**

Lecture 4: Polynomial regression

- How to fit more complex data?

Recap: Linear Regression



- In general high-dimensions, we fit a linear model with intercept
  $y_i \simeq w^T x_i + b$, or equivalently $y_i = w^T x_i + b + \epsilon_i$
  with model parameters $(w \in \mathbb{R}^d, b \in \mathbb{R})$ that minimizes ℓ_2-loss

$$\mathcal{L}(w, b) \;=\; \sum_{i=1}^{n} \underbrace{(y_i - (w^T x_i + b))^2}_{\text{error } \epsilon_i}$$

Recap: Linear Regression

- The least squares solution, i.e. the minimizer of the ℓ_2-loss can be written in a **closed form** as a function of data $\mathbf{X}$ and $\mathbf{y}$ as

As we derived in class:

$$\mu = \frac{1}{n}\mathbf{X}^T\mathbf{1}$$

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\mu^T$$

$$\widehat{w}_{\text{LS}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{y}$$

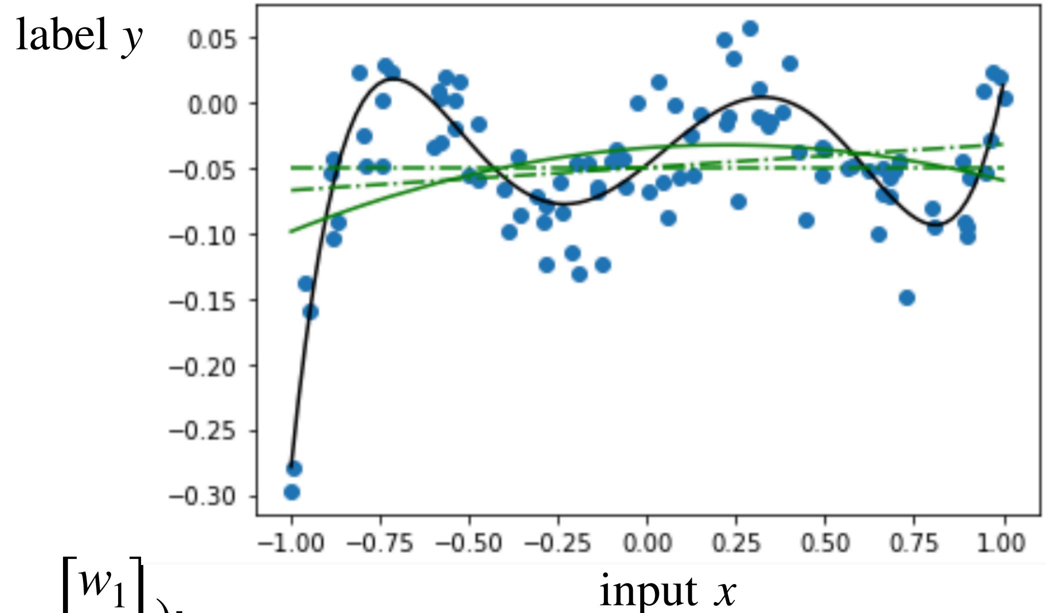
$$\widehat{b}_{\text{LS}} = \frac{1}{n}\sum_{i=1}^{n} y_i - \mu^T\widehat{w}_{\text{LS}}$$

or equivalently using straightforward linear algebra by setting the gradient to zero:

$$\begin{bmatrix} \widehat{w}_{\text{LS}} \\ \widehat{b}_{\text{LS}} \end{bmatrix} = \left(\begin{bmatrix} \mathbf{X}^T \\ \mathbf{1}^T \end{bmatrix} [\mathbf{X} \quad \mathbf{1}] \right)^{-1} \begin{bmatrix} \mathbf{X}^T \\ \mathbf{1}^T \end{bmatrix} \mathbf{y}$$

Quadratic regression in 1-dimension

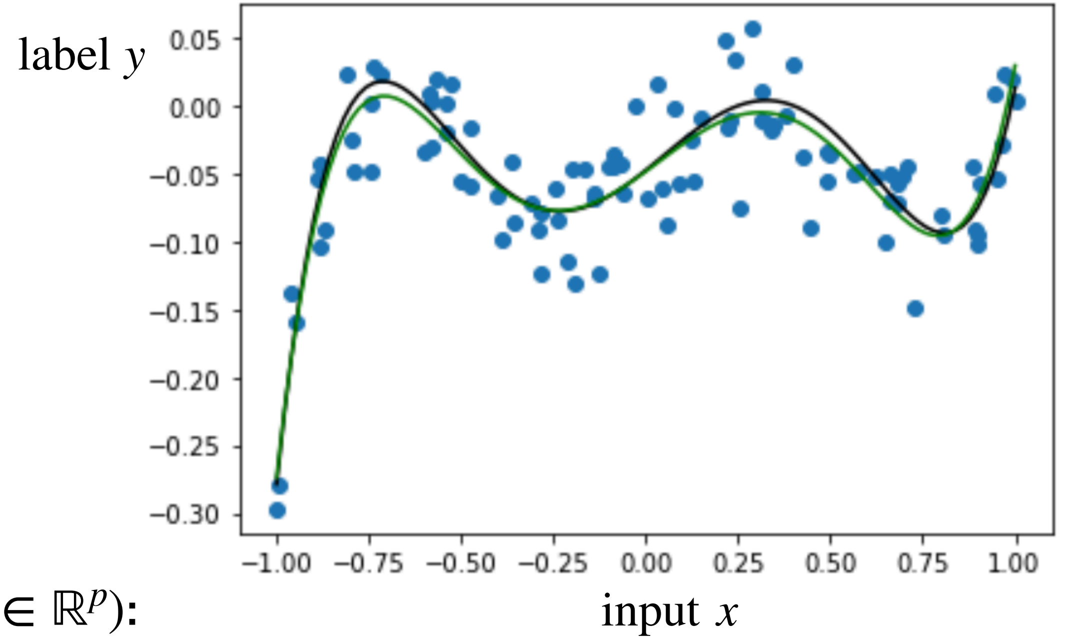
- **Data:** $\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$



- **Linear model with parameter (b, w_1):**
  - $y_i = b + w_1 x_i + \epsilon_i$
  - $\mathbf{y} = \mathbf{1}b + \mathbf{X}w_1 + \epsilon$
- **Quadratic model with parameter $(b, w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix})$:**
  - $y_i = b + w_1 x_i + w_2 x_i^2 + \epsilon_i$
  - Define $h : \mathbb{R} \to \mathbb{R}^2$ such that $x \mapsto h(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$
  - $y_i = b + h(x_i)^T w + \epsilon_i$
  - Treat $h(x)$ as new input features. Let $\mathbf{H} = \begin{bmatrix} h(x_1)^T \\ \vdots \\ h(x_n)^T \end{bmatrix}$. Replace x_i by $\begin{bmatrix} x_i \\ x_i^2 \end{bmatrix}$
  - $\mathbf{y} = \mathbf{1}b + \mathbf{H}w + \epsilon$

Degree-p polynomial regression in 1-dimension

- **Data:** $\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$



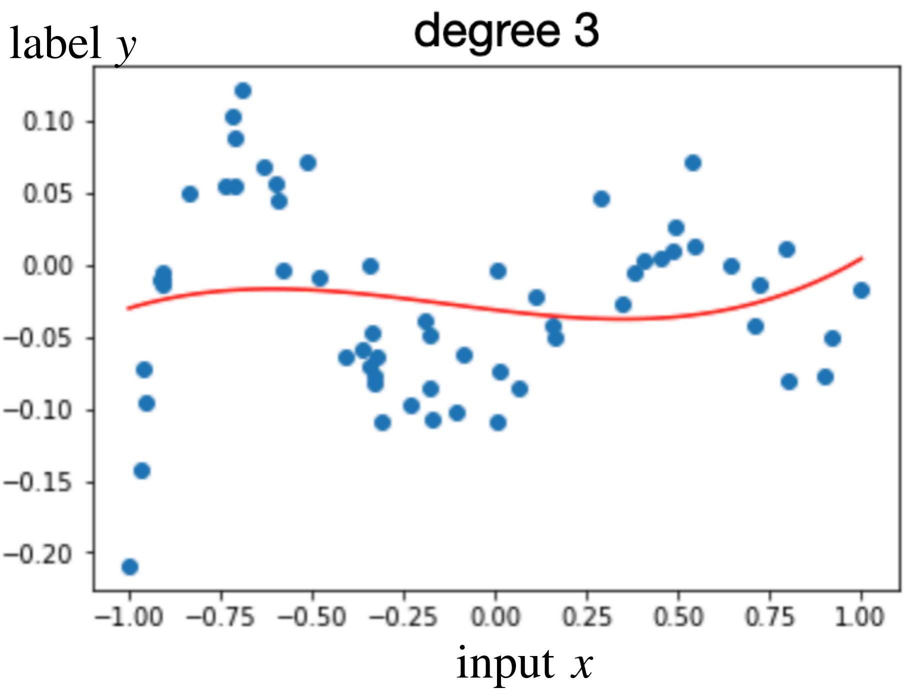
- **Linear model with parameter** (b, w_1):
    - $y_i = b + w_1 x_i + \epsilon_i$
    - $\mathbf{y} = \mathbf{1}b + \mathbf{X}w_1 + \epsilon$
- **Degree-p model with parameter** $(b, w \in \mathbb{R}^p)$:
    - $y_i = b + w_1 x_i + \cdots + w_p x_i^p + \epsilon_i$
    - Define $h : \mathbb{R} \to \mathbb{R}^p$ such that $x \mapsto h(x) = \begin{bmatrix} x \\ \vdots \\ x^p \end{bmatrix}$
    - $y_i = b + h(x_i)^T w + \epsilon_i$
    - Treat $h(x)$ as new input features and let $\mathbf{H} = \begin{bmatrix} h(x_1)^T \\ \vdots \\ h(x_n)^T \end{bmatrix}$
    - $\mathbf{y} = \mathbf{1}b + \mathbf{H}w + \epsilon$

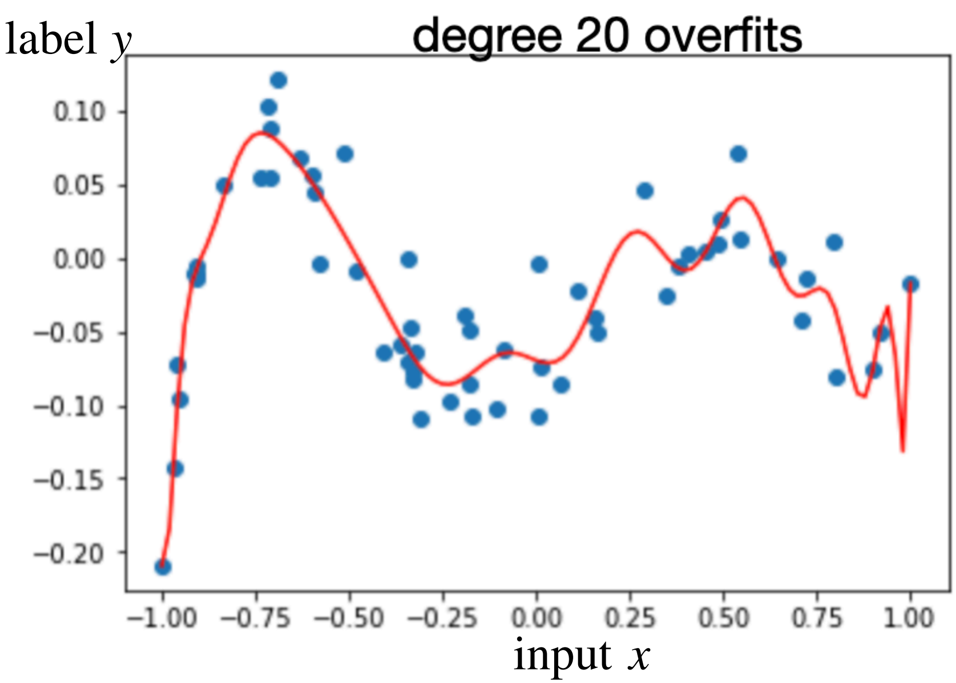
Degree-p polynomial regression in d-dimension

- **Data:** $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ & x_2^T & & \\ & \vdots & & \\ & x_n^T & & \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$
- **Degree-p model with parameter $(b, w \in \mathbb{R}^{dp})$:**
  - $y_i = b + x_i^T w_1 + \cdots + (x_i^p)^T w_p + \epsilon_i$, where $x_i^P = \begin{bmatrix} x_{i1}^p \\ \vdots \\ x_{id}^p \end{bmatrix}$
  - Define $h : \mathbb{R}^d \to \mathbb{R}^{dp}$ such that $x \mapsto h(x) = \begin{bmatrix} x \\ \vdots \\ x^p \end{bmatrix} \in \mathbb{R}^{dp}$
  - $y_i = b + h(x_i)^T w + \epsilon_i$
  - Treat $h(x)$ as new input features and let $\mathbf{H} = \begin{bmatrix} h(x_1)^T \\ \vdots \\ h(x_n)^T \end{bmatrix}$
  - $\mathbf{y} = \mathbf{1}b + \mathbf{H}w + \epsilon$
- In general, any feature $h(x)$ can be used, e.g., $\sin(ax + b)$, $e^{-b(x-a)^2}$, $\log x$, etc.

Which p should we choose?

- First instance of class of models with different representation power = model complexity

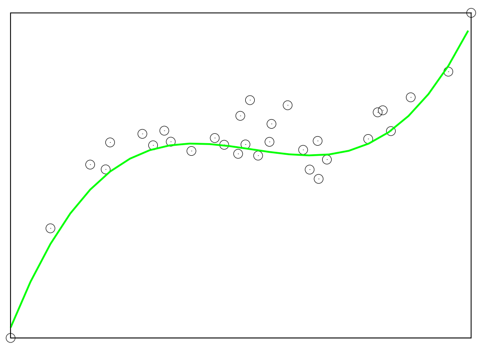


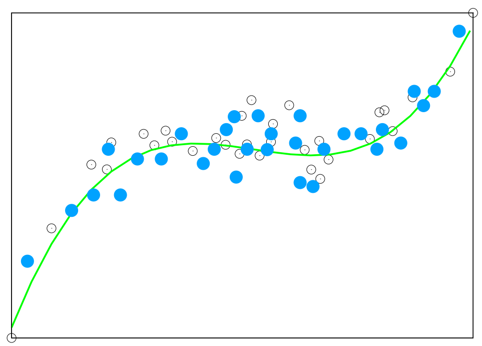


- How do we determine which is better model?

Generalization

- we say a predictor **generalizes** if it performs as well on unseen data as on training data
- formal mathematical definition involves probabilistic assumptions (coming later in this week)
- the data used to train a predictor is **training data** or **in-sample data**
- we want the predictor to work on **out-of-sample data**
- we say a predictor **fails to generalize** if it performs well on in-sample data but does not perform well on out-of-sample data





- **train** a cubic predictor on 32 (**in-sample**) white circles: Mean Squared Error (MSE) 174
- **predict** label y for 30 (**out-of-sample**) blue circles: MSE 192
- conclude this predictor/model generalizes, as in-sample MSE $\simeq$ out-of-sample MSE

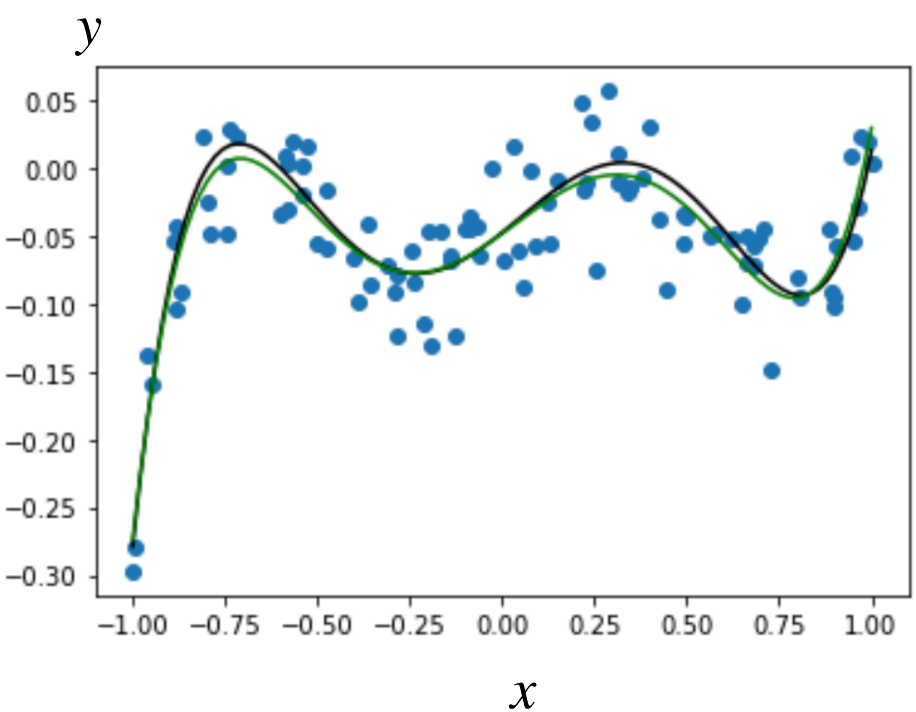
Split the data into training and testing

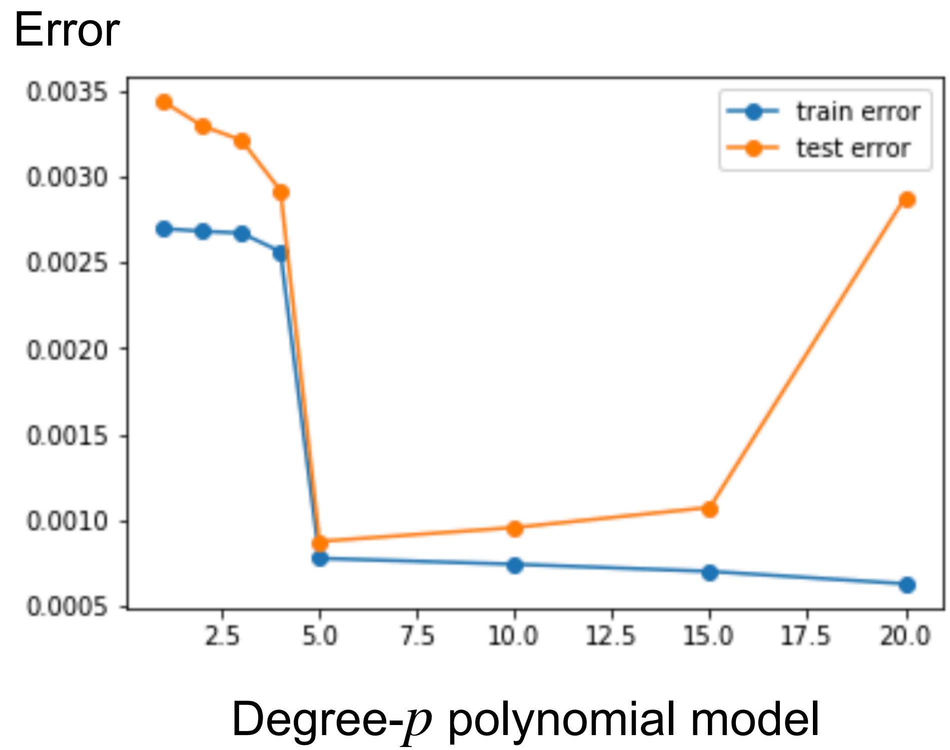
- a way to mimic how the predictor performs on unseen data
- given a single dataset $S = \{(x_i, y_i)\}_{i=1}^n$
- we split the dataset into two: training set and test set
- selection of data train/test should be done randomly (80/20 or 90/10 are common)
- **training set** used to train the model
  - $\text{minimize} \;\; \mathscr{L}_{\text{train}}(w) = \frac{1}{|S_{\text{train}}|} \sum_{i \in S_{\text{train}}} (y_i - x_i^T w)^2$
- **test set** used to evaluate the model
  - $\mathscr{L}_{\text{test}}(w) = \frac{1}{|S_{\text{test}}|} \sum_{i \in S_{\text{test}}} (y_i - x_i^T w)^2$
- this assumes that test set is similar to unseen data
- **test set should never be used in training**

We say a model w or predictor **overfits** if $\mathscr{L}_{\text{train}}(w) \ll \mathscr{L}_{\text{test}}(w)$

	small training error	large training error
small test error	generalizes well performs well	possible, but unlikely
large test error	fails to generalize **Overfitting**	generalizes well performs poorly

How do we choose which model to use?



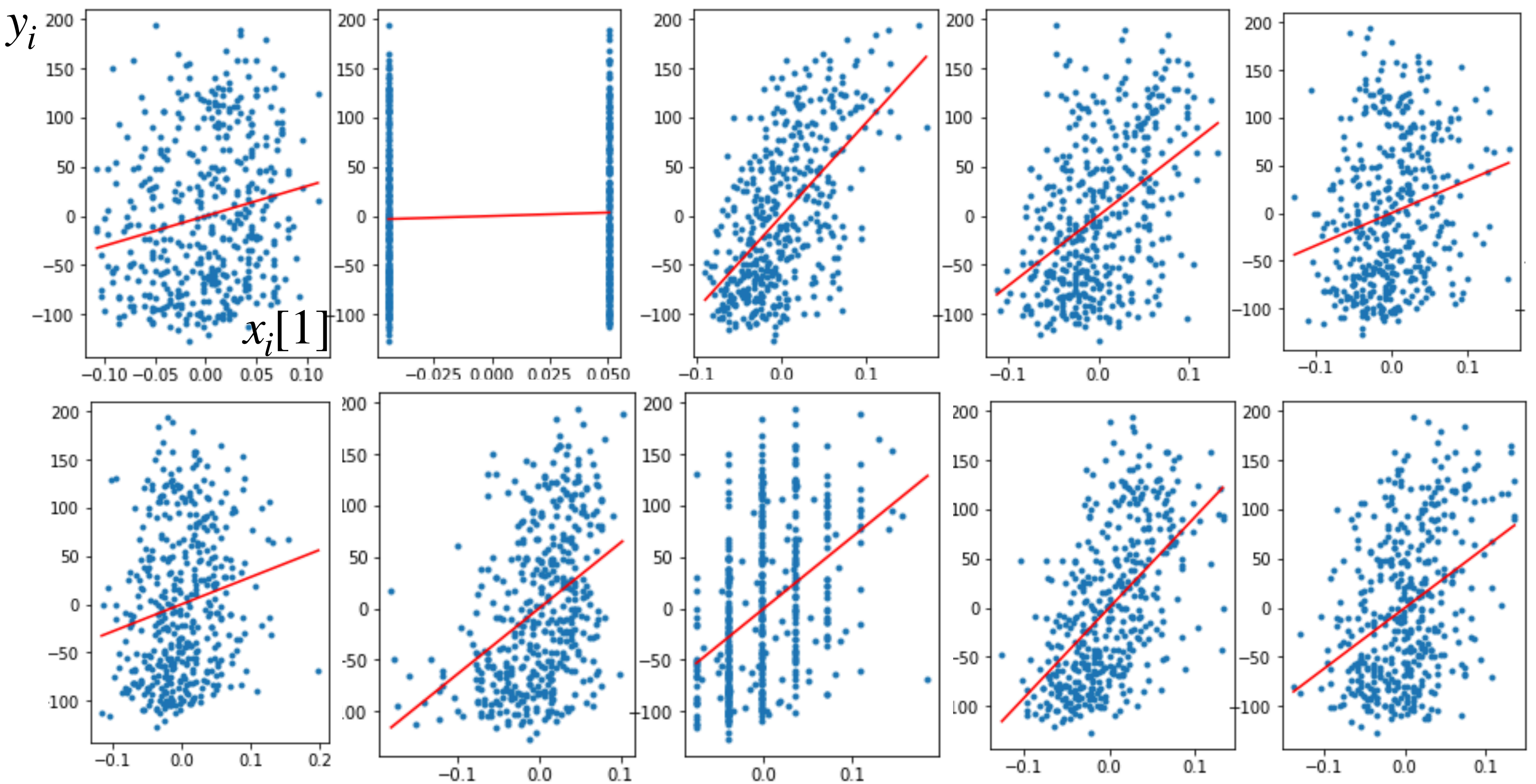


1. first use 60 data points to train and 60 data points to test and train several models to get the above graph on the right
2. then choose degree $p = 5$, since it achieves **minimum test error**
3. now re-train on all 120 data points with degree 5 polynomial model

demo2_lin.ipynb

Another example: Diabetes

- Example: Diabetes
  - 10 explanatory variables
  - from 442 patients
  - we use half for train and half for validation



Features	Train MSE	Test MSE
All	2640	3224
S5 and BMI	3004	3453
S5	3869	4227
BMI	3540	4277
S4 and S3	4251	5302
S4	4278	5409
S3	4607	5419
None	5524	6352

- **test MSE is the primary criteria for model selection**
- Using only 2 features (S5 and BMI), one can get very close to the prediction performance of using all features
- Combining S3 and S4 does not give any performance gain

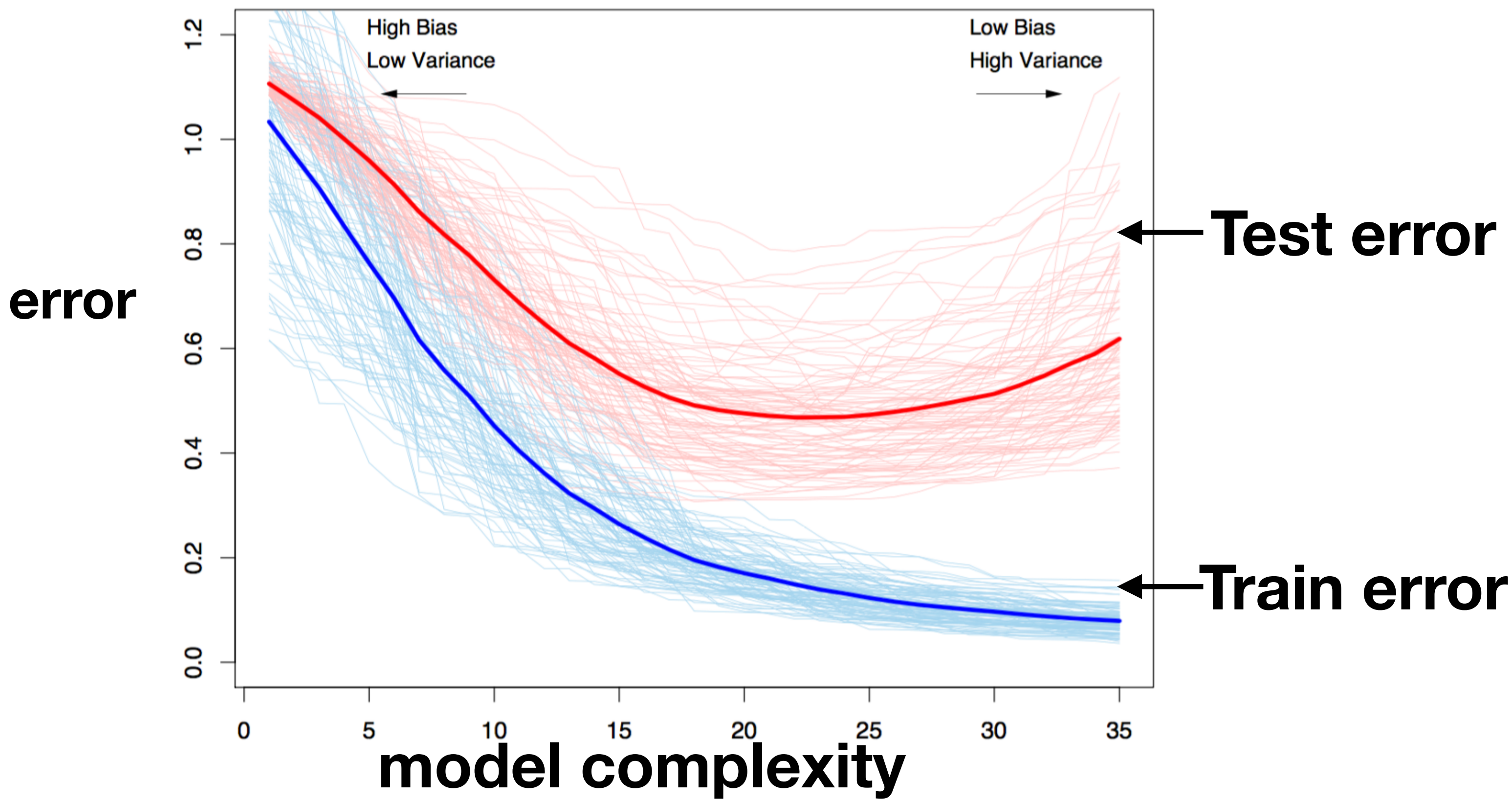
demo3_diabetes.ipynb

What does the bias-variance theory tell us?

- **Train error** (random variable, randomness from $\mathcal{D}$)
  - Use $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \sim P_{X,Y}$ to find $\widehat{w}$
  - Train error: $\mathcal{L}_{\text{train}}(\widehat{w}_{\text{LS}}) = \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - \widehat{w}^T x_i)^2$
- recall the **test error** is an unbiased estimator of the **true error**
- **True error** (random variable, randomness from $\mathcal{D}$)
  - True error: $\mathcal{L}_{\text{true}}(\widehat{w}) = \mathbb{E}_{(x,y)\sim P_{X,Y}}[(y - \widehat{w}^T x)^2]$
- **Test error** (random variable, randomness from $\mathcal{D}$ and $\mathcal{T}$)
  - Use $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^m \sim P_{X,Y}$
  - Test error: $\mathcal{L}_{\text{test}}(\widehat{w}) = \frac{1}{|\mathcal{T}|} \sum_{(x_i,y_i)\in\mathcal{T}} (y_i - \widehat{w}^T x_i)^2$
- theory explains **true error**, and hence expected behavior of the (random) **test error**

What does the bias-variance theory tell us?

- Train error is optimistically biased (i.e. smaller) because the trained model is minimizing the train error
- Test error is unbiased estimate of the true error, if test data is never used in training a model or selecting the model complexity
- Each line is an i.i.d. instance of $\mathscr{D}$ and $\mathscr{T}$

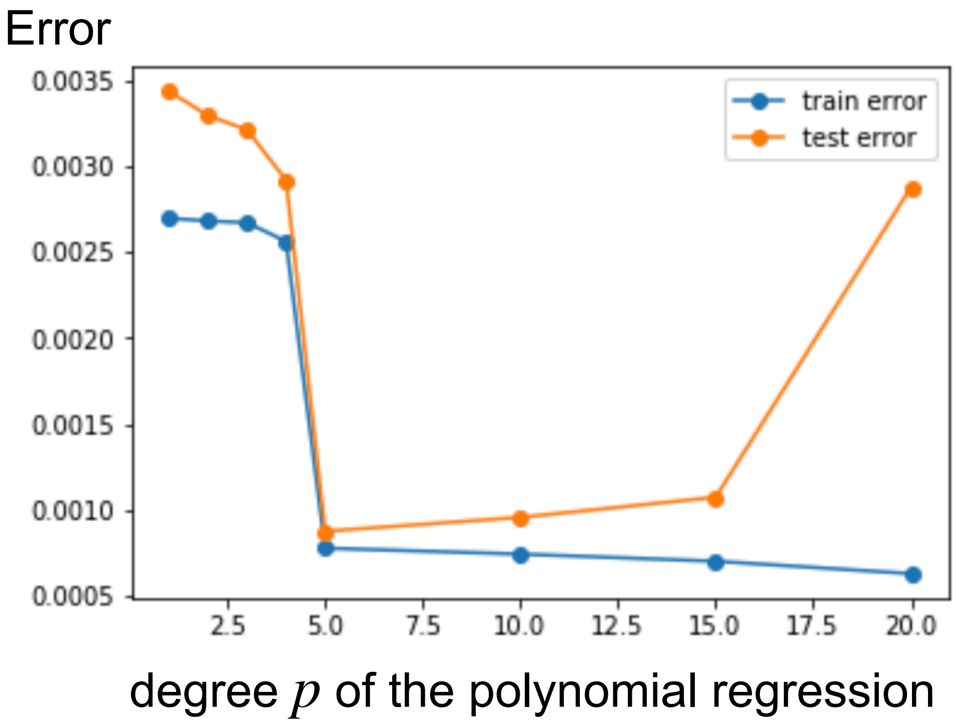


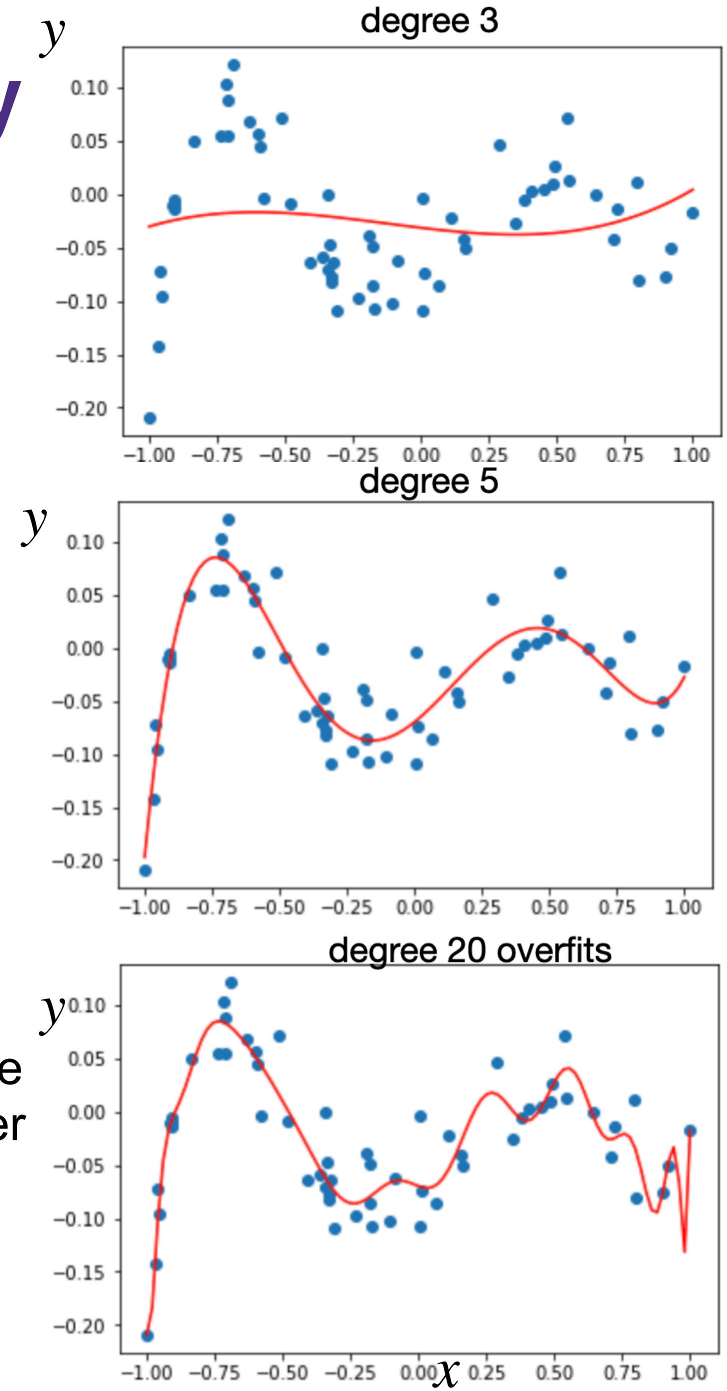
Questions?

Lecture 5:
Bias-Variance Tradeoff

- explaining test error using theoretical analysis

Train/test error vs. complexity





- **Model complexity** e.g., degree p of the polynomial model, number of features used in diabetes example
  - Related to the dimension of the model parameter
- **Train error** monotonically decreases with model complexity
- **Test error** has a U shape

Statistical learning

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 $X = x$, hence the optimization can be broken into sum (or more precisely integral) of multiple objective functions, each involving a specific value $X = x$
  - $$\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2] = \mathbb{E}_{X\sim P_X}\Big[\mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]\Big]$$
    $$= \int \mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]\, P_X(x)\, dx$$

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

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

Statistical learning

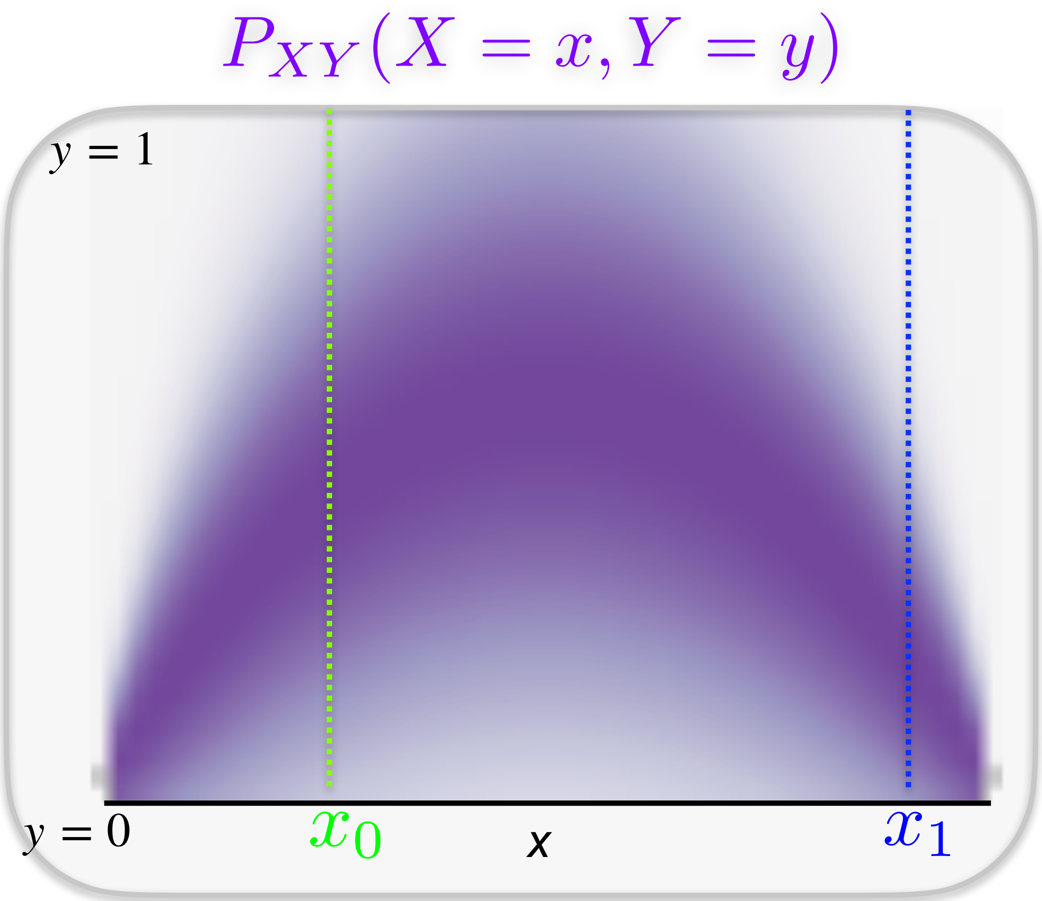
- We can solve the optimization for each $X = x$ separately
  - $\eta(x) \;=\; \arg\min_{a \in \mathbb{R}} \mathbb{E}_{Y \sim P_{Y|X}}[(Y - a)^2 \,|\, X = x]$
- The optimal solution is $\eta(x) = \mathbb{E}_{Y \sim P_{Y|X}}[Y \,|\, X = x]$,
  which is the best prediction in ℓ_2-loss/Mean Squared Error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

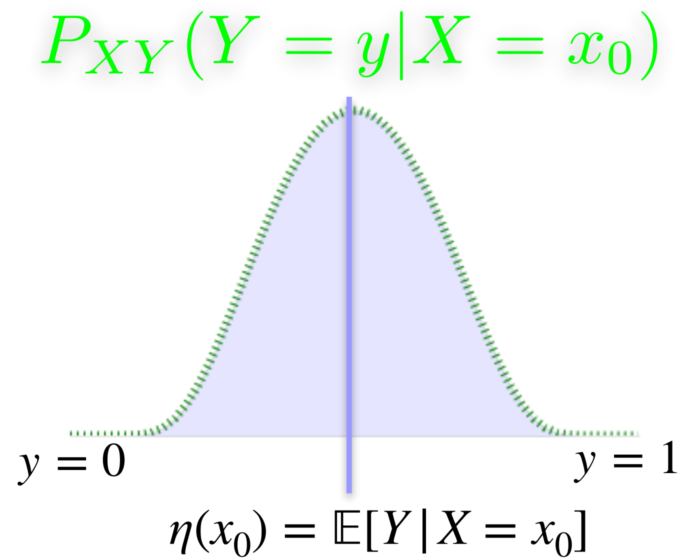
- Note that this optimal statistical estimator $\eta(x) = \mathbb{E}[Y \,|\, X = x]$ cannot be implemented as we do not know $P_{X,Y}$ in practice
- This is only for the purpose of conceptual understanding

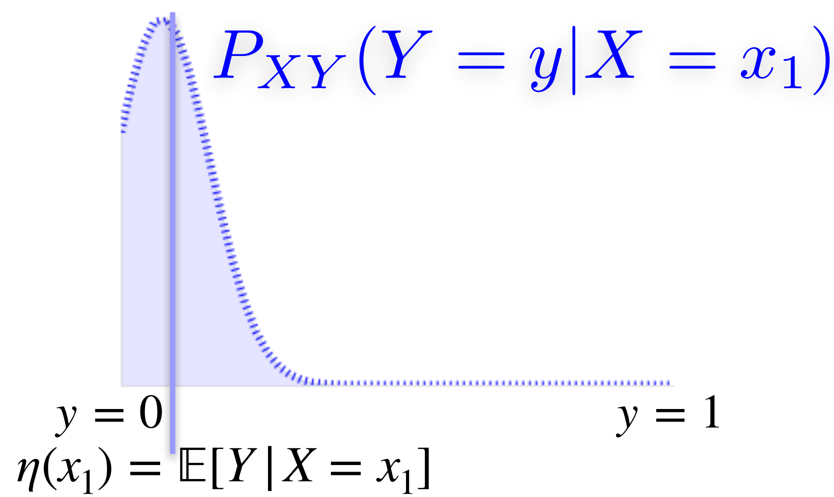
Statistical Learning

Ideally, we want to find:

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

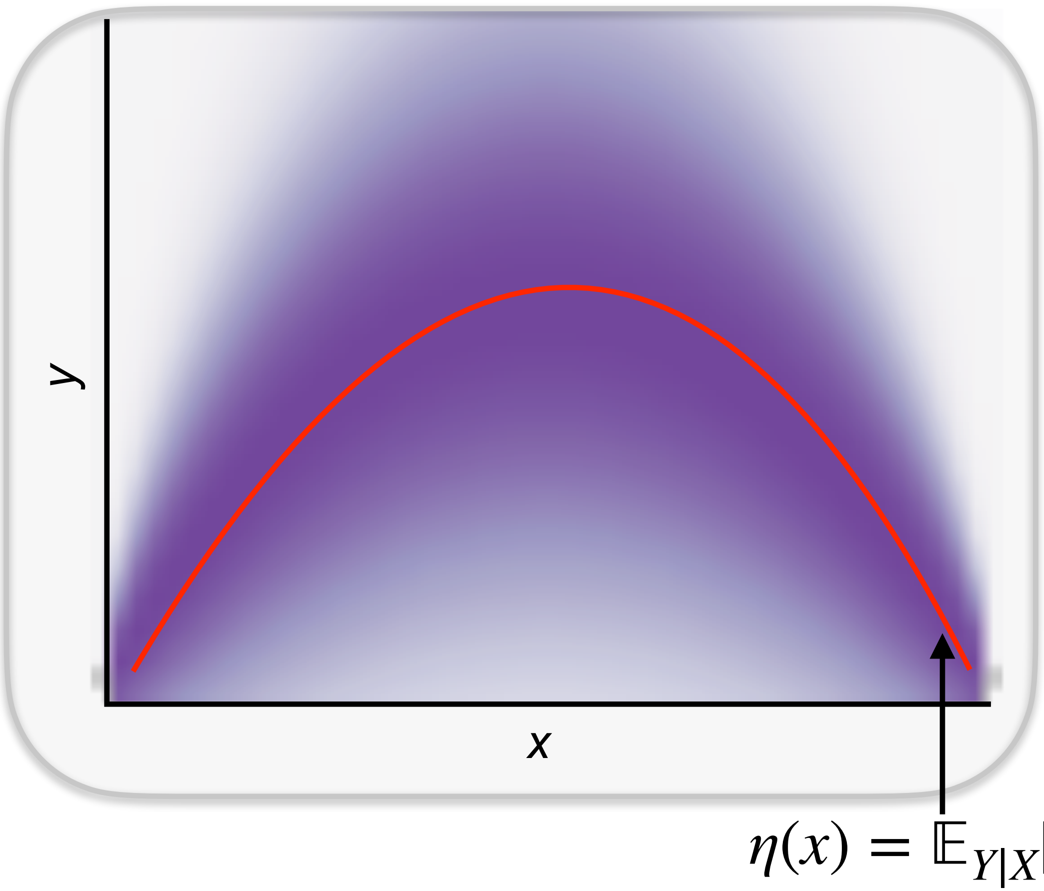






Statistical Learning

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



Ideally, we want to find:

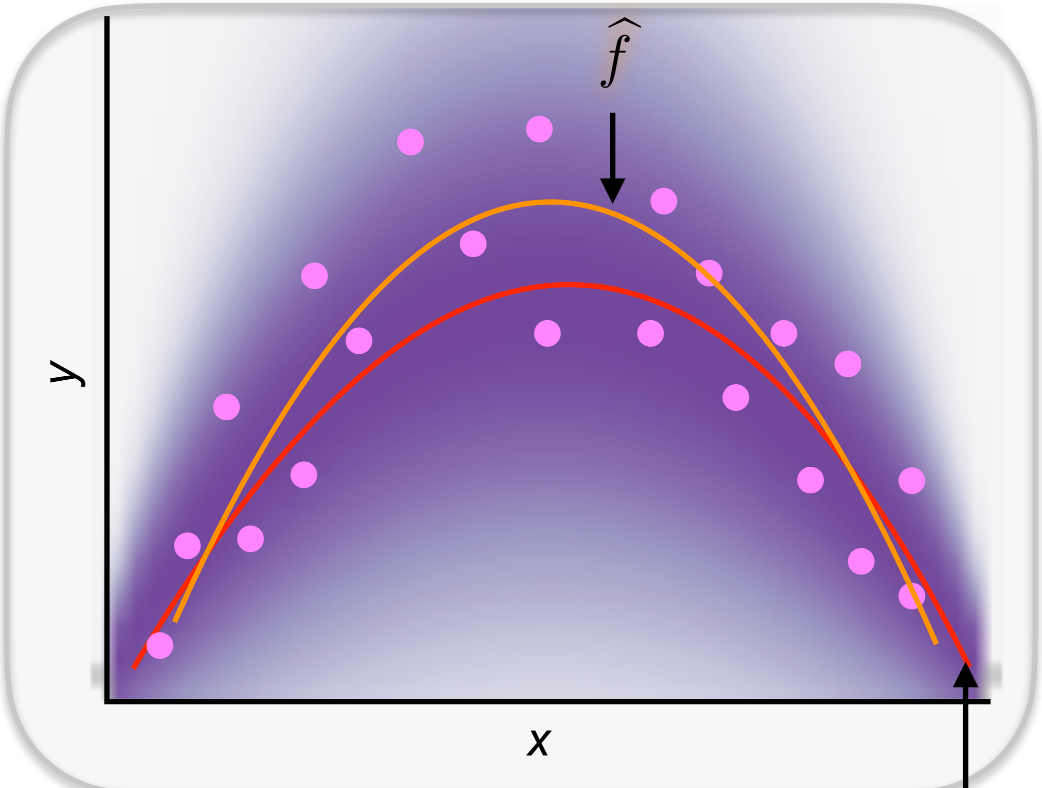
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

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

We only have samples.

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:

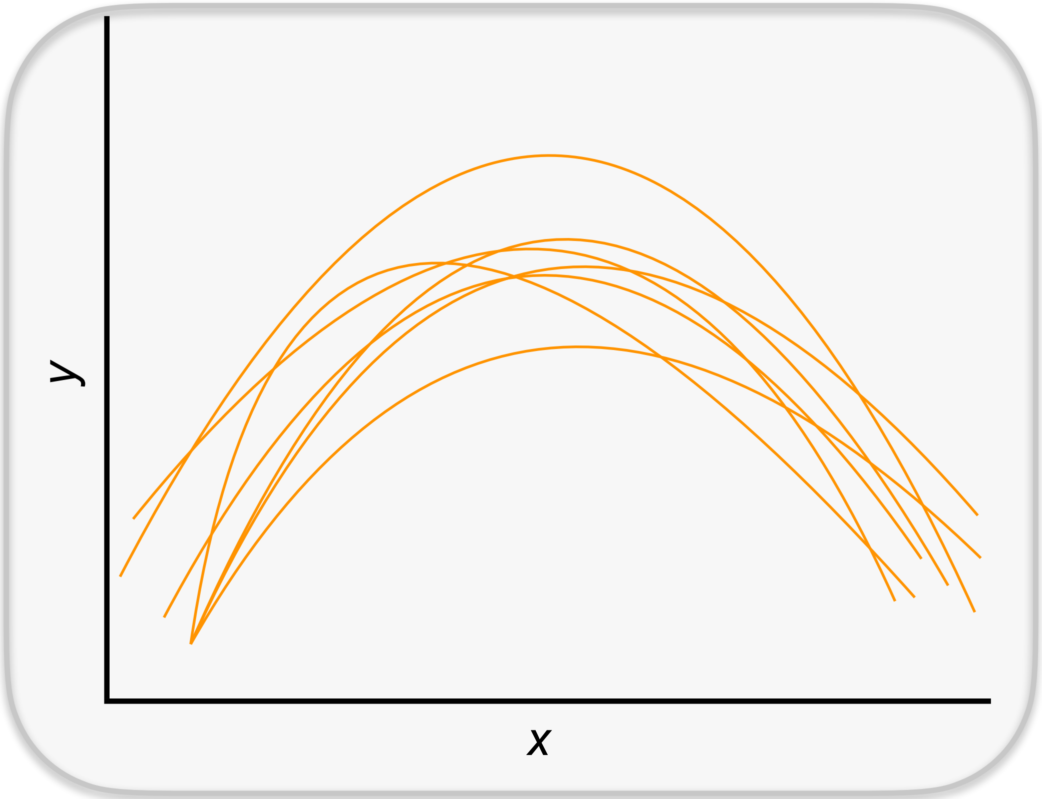
$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

We care about how our predictor performs on future unseen data

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

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

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



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

Bias-variance tradeoff

Notation:
I use predictor/model/estimate, interchangeably

Ideal predictor

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

Learned predictor

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

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

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**:

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

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**:

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

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

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

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

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

Irreducible error
(a) Caused by stochastic label noise in $P_{Y|X=x}$
(b) cannot be reduced

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**:

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

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**:

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

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

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

Bias-variance tradeoff

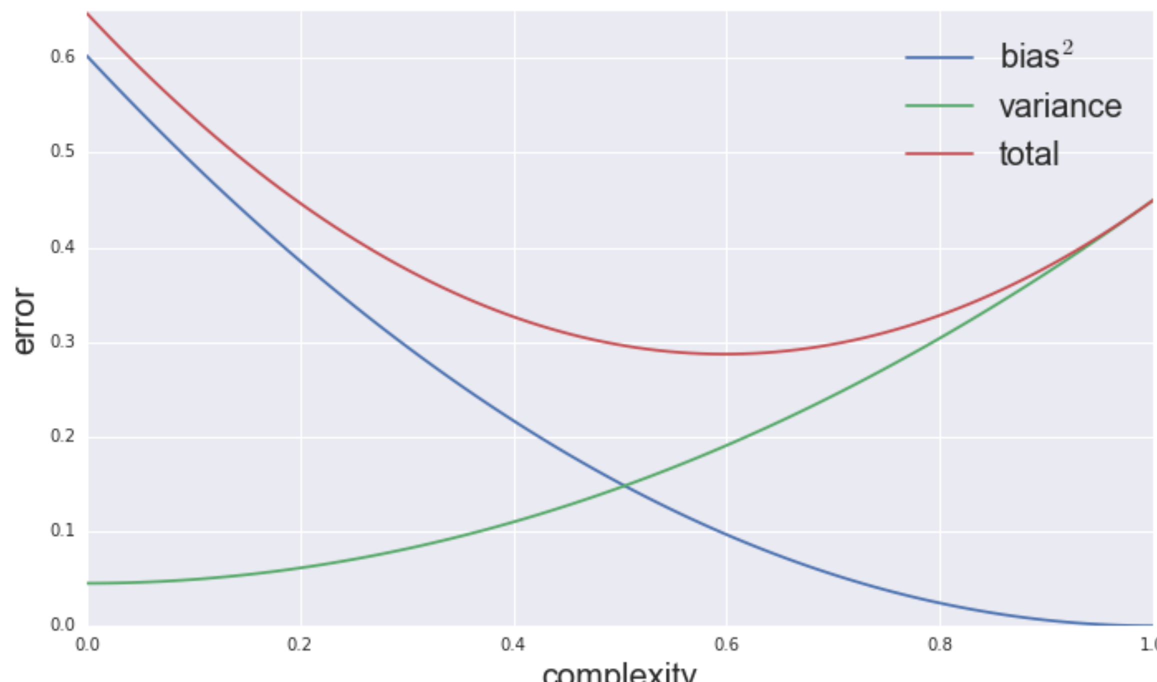
- **Average conditional true error**:

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

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

Bias squared:
measures how the predictor is mismatched with the best predictor in expectation

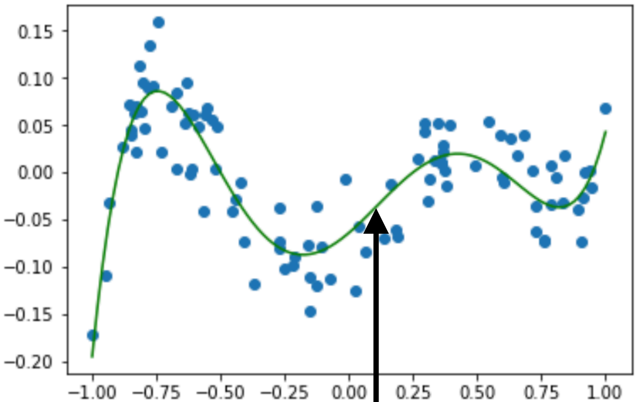
variance:
measures how the predictor varies each time with a new training datasets



Questions?

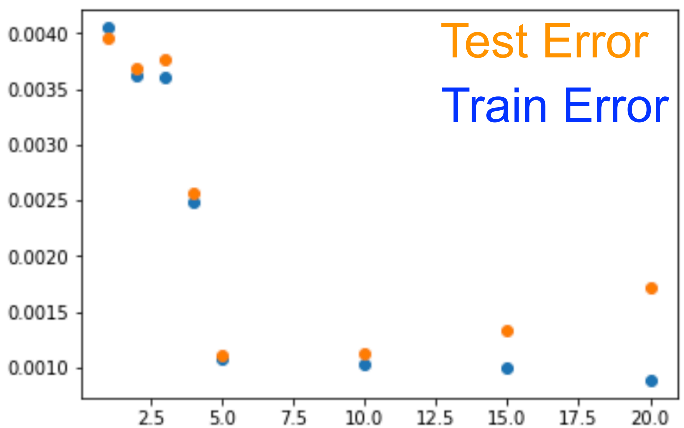
Lecture 6: Bias-Variance Tradeoff (continued)

Test error vs. model complexity



Optimal predictor $\eta(x)$ is degree-5 polynomial

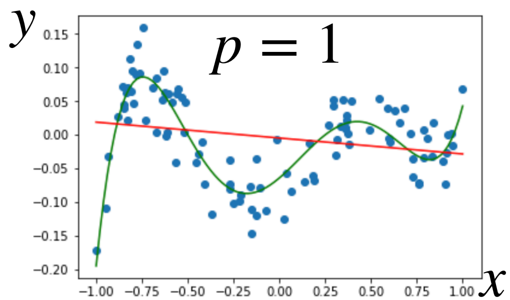
Error

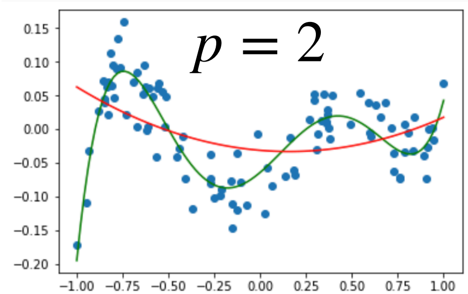


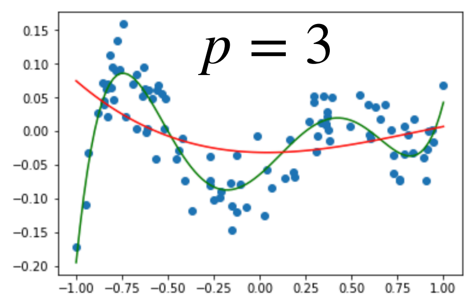
degree p of the polynomial regression

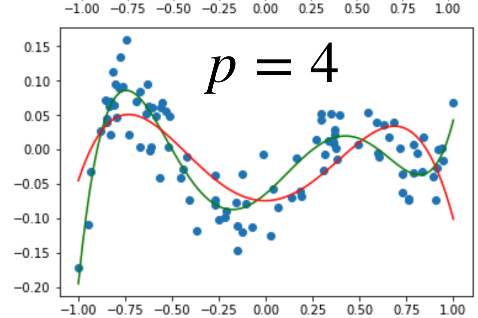
demo4_tradeoff.ipynb

Simple model:
Model complexity is below the complexity of $\eta(x)$

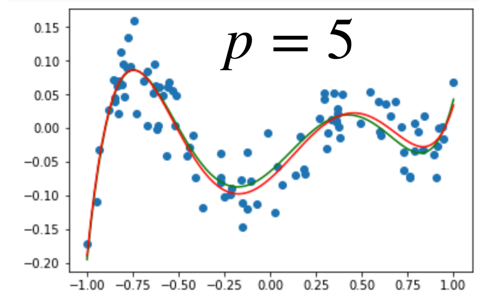


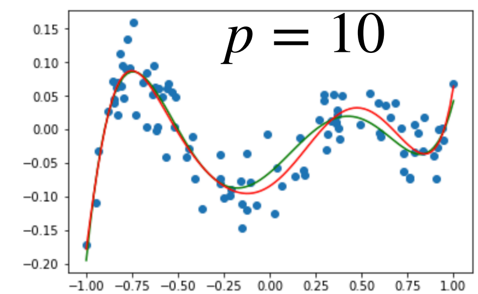


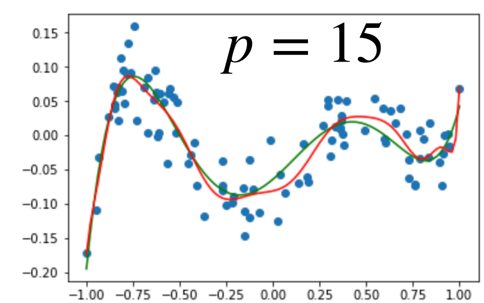


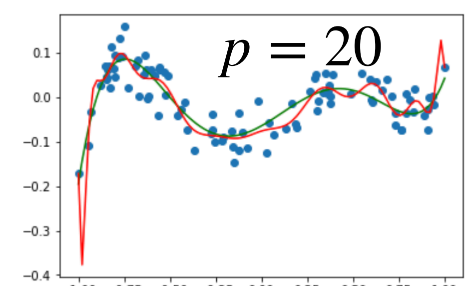


Complex model:

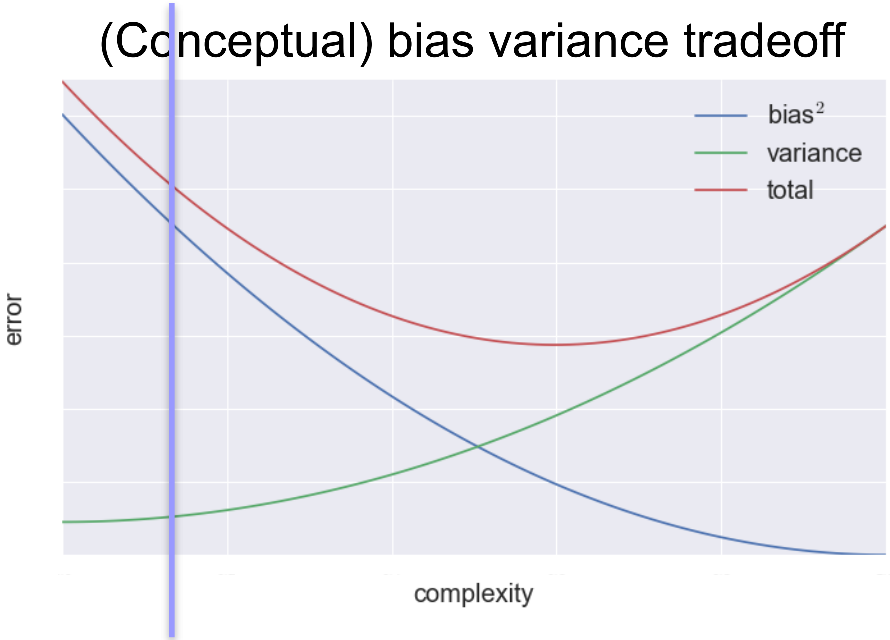


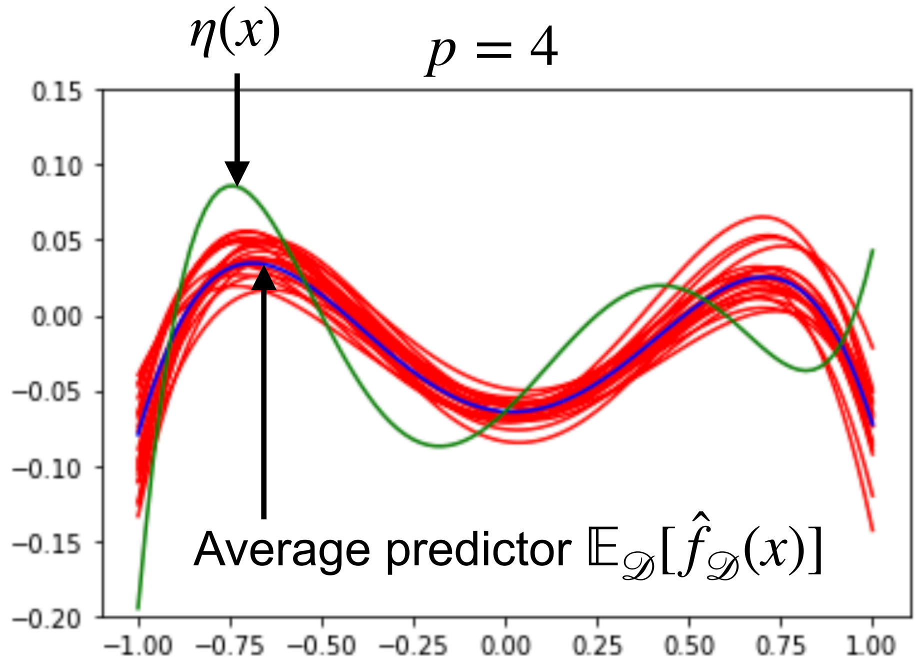






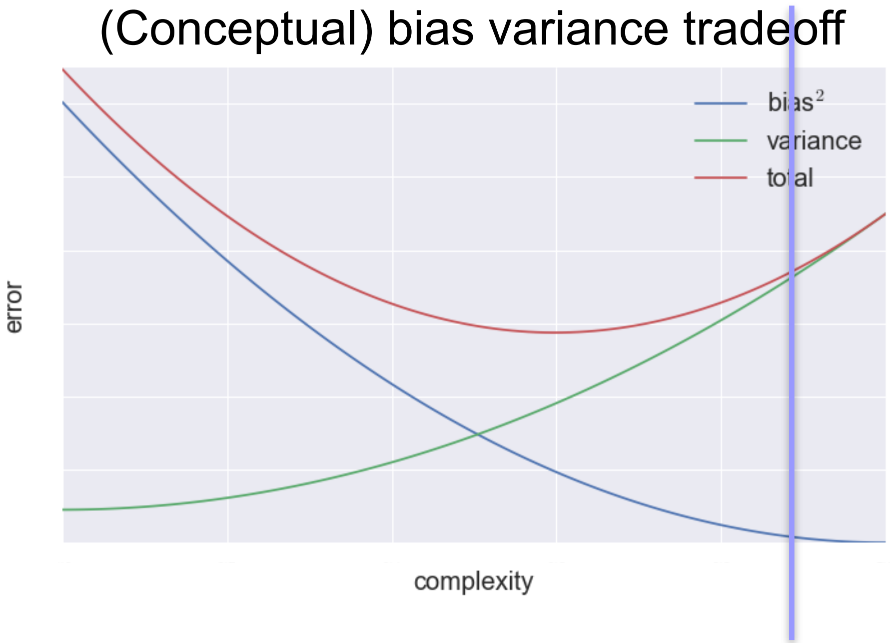
Recap: Bias-variance tradeoff with simple model

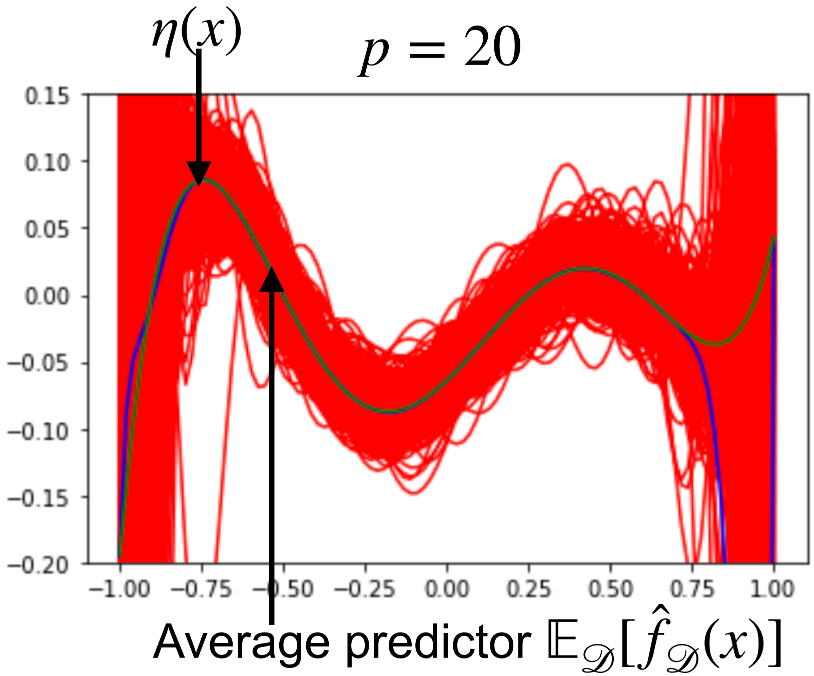




- When model **complexity is low** (lower than the optimal predictor $\eta(x)$)
  - Bias^2 of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2$, is large
  - Variance of our predictor, $\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]$, 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

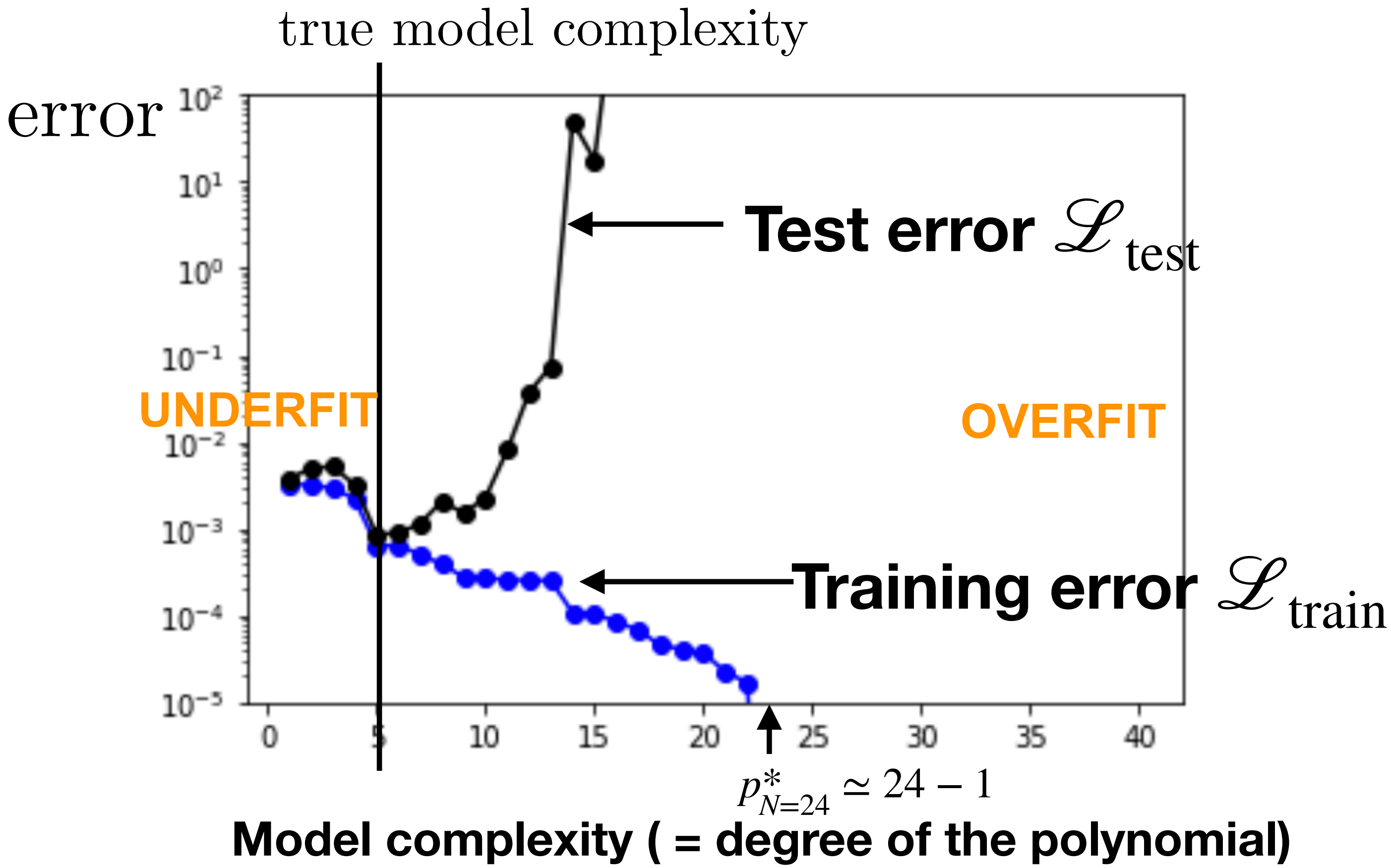




- When model complexity is high (higher than the optimal predictor $\eta(x)$)
  - Bias of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2$, is small
  - Variance of our predictor, $\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]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

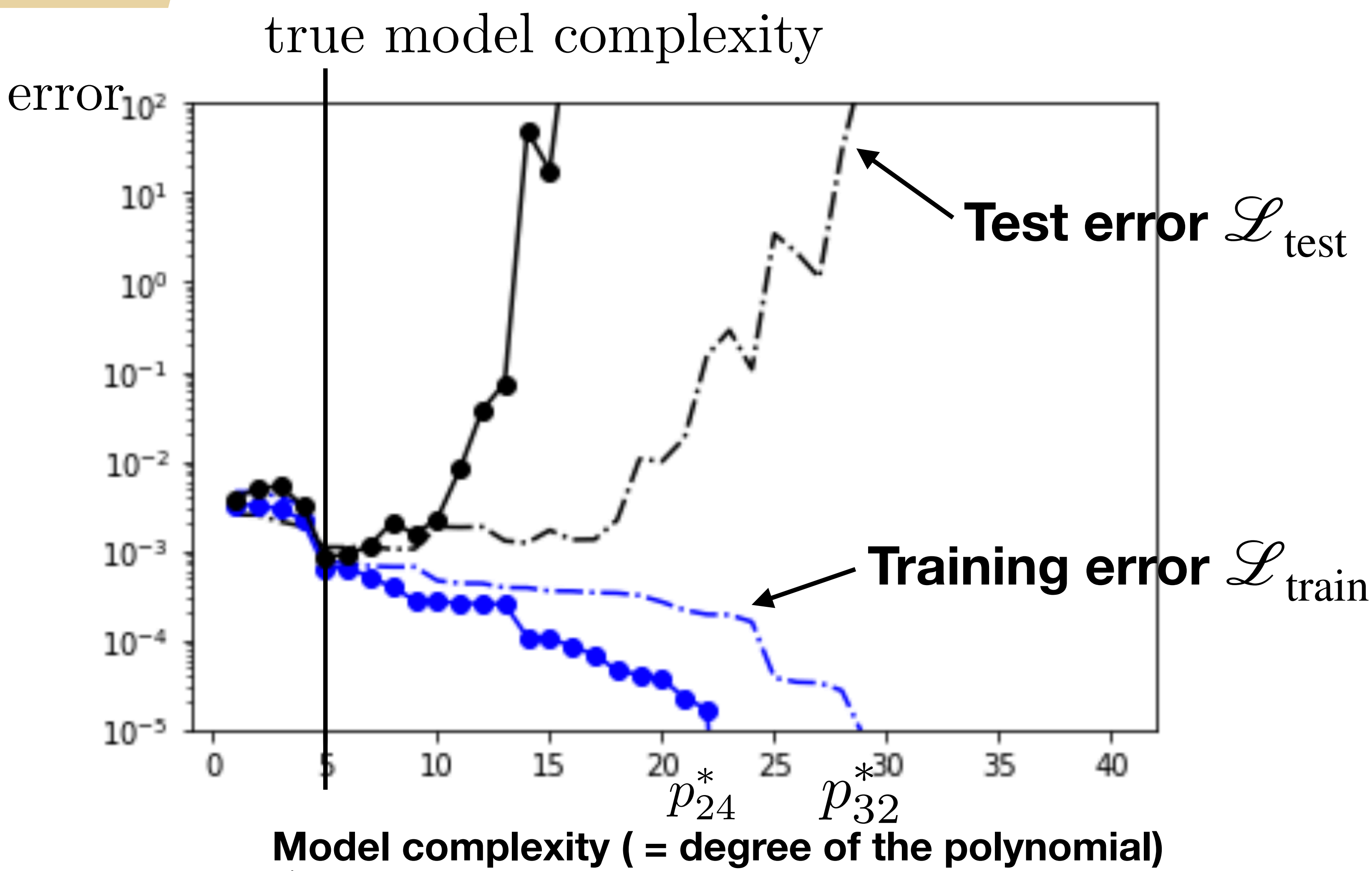
- let us first fix sample size ***N=30***, collect one dataset of size N i.i.d. from a distribution, and fix one training set S_{train} and test set S_{test} via 80/20 split

- then we run multiple validations and plot the computed MSEs for all values of ***p*** that we are interested in



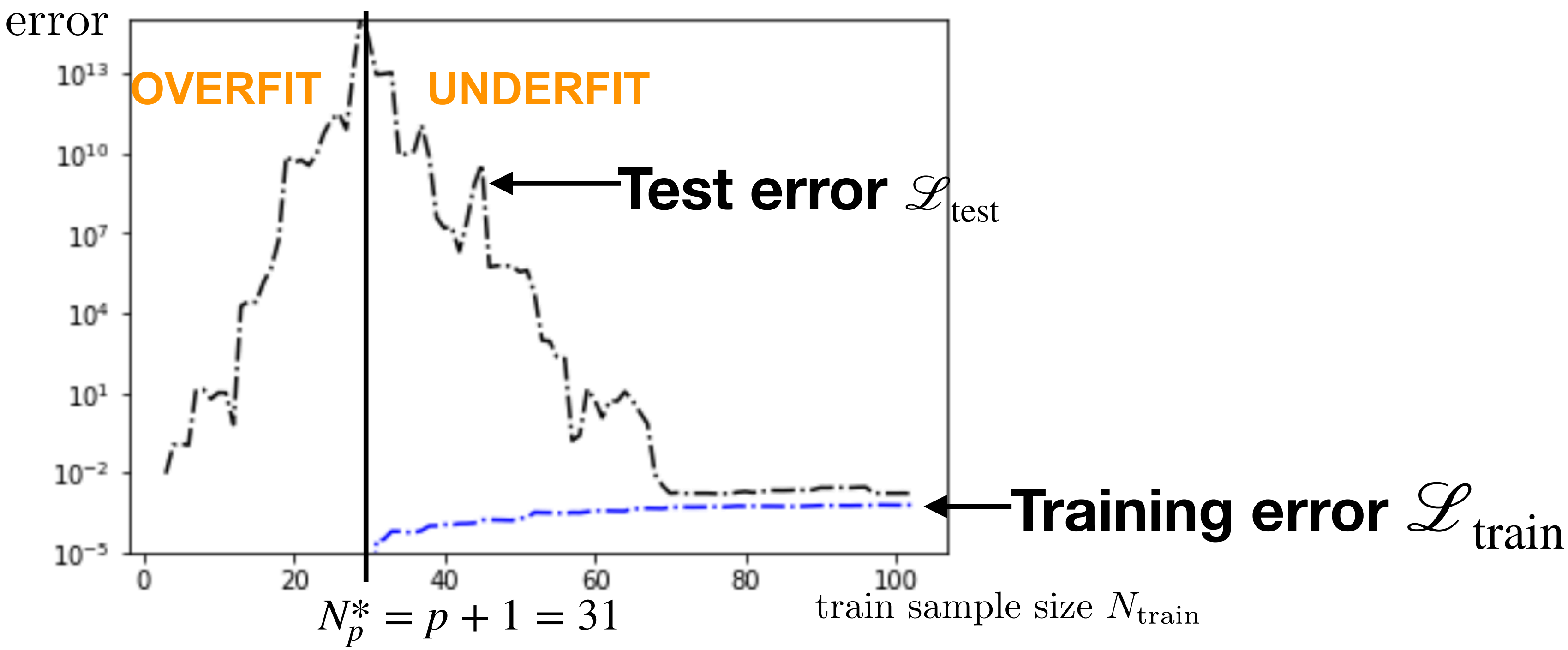
- Given sample size N there is a threshold, p^*_N, where training error is zero
- Training error is **always** monotonically non-increasing
- Test error has a trend of going down and then up, but fluctuates

- 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

lecture2_polynomialfit.ipynb

Bias-variance tradeoff for linear models

If $Y_i = X_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$\mathbf{y} = \mathbf{X} w^* + \epsilon$$

$$\widehat{w}_{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} =$$

$$=$$

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

$$\hat{f}_{\mathcal{D}}(x) = x^T \widehat{w}_{\text{MLE}} =$$

Bias-variance tradeoff for linear models

If $Y_i = X_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$\mathbf{y} = \mathbf{X}w^* + \epsilon$$

$$\widehat{w}_{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w^* + \epsilon)$$

$$= w^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

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

$$\hat{f}_{\mathcal{D}}(x) = x^T \widehat{w}_{\text{MLE}} = x^T w^* + x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

- Irreducible error: $\mathbb{E}_{X,Y}[(Y - \eta(x))^2 \,|\, X = x] =$
- Bias squared: $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] \right)^2 =$
  (is independent of the sample size!)

Bias-variance tradeoff for linear models

If $Y_i = X_i^T w^* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$\widehat{w}_{\text{MLE}} = w^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\eta(x) = x^T w^*$$

$$\hat{f}_{\mathcal{D}}(x) = x^T w^* + x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

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

Bias-variance tradeoff for linear models

If $Y_i = X_i^T w* + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$\widehat{w}_{\text{MLE}} = w* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\eta(x) = x^T w*$$

$$\hat{f}_{\mathcal{D}}(x) = x^T w* + x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

- Variance: $\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2\right] = \mathbb{E}_{\mathcal{D}}[x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x]$

$$= \sigma^2\, \mathbb{E}_{\mathcal{D}}[x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x]$$

$$= \sigma^2\, x^T\mathbb{E}_{\mathcal{D}}[(\mathbf{X}^T\mathbf{X})^{-1}]x$$

- To analyze this, let's assume that $X_i \sim \mathcal{N}(0,\mathbf{I})$ and number of samples, n, is large enough such that $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$ with high probability and $\mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}] \simeq \frac{1}{n}\mathbf{I}$, then
    - Variance is $\dfrac{\sigma^2 x^T x}{n}$, and decreases with increasing sample size n

Questions?