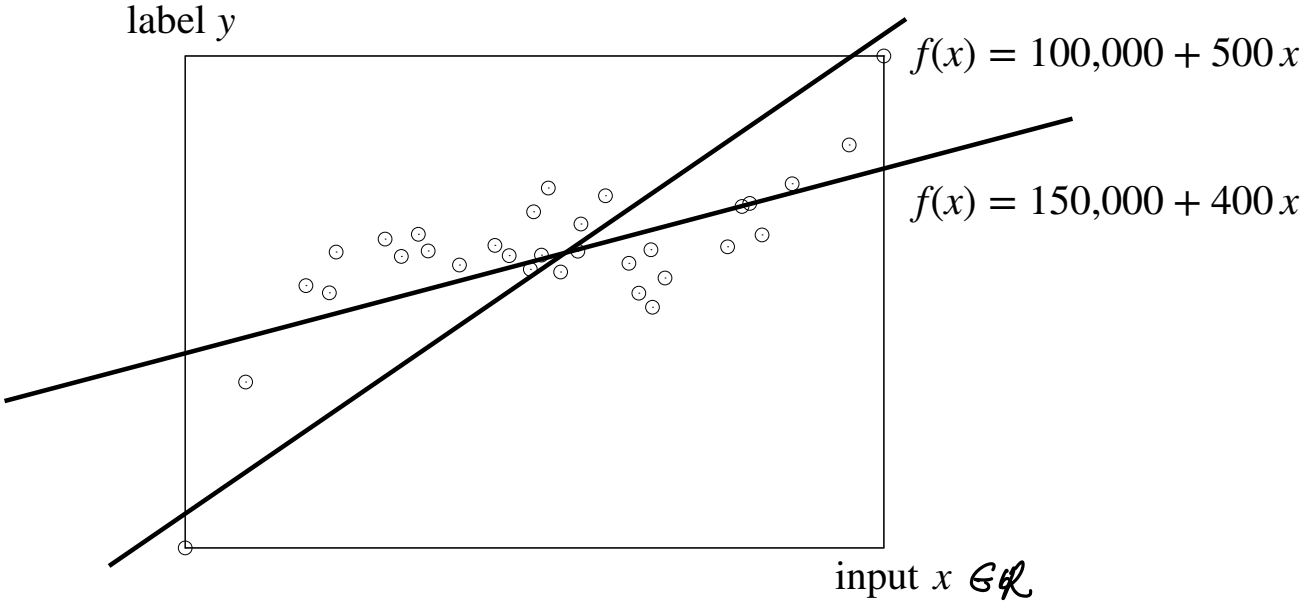
- HW0 due Tuesday midnight
- Extra office hours
  - **Monday:**
    - Tim Li, 10:30 - 11:30
    - **Sewoong Oh, 12:30 - 1:30**
    - Hugh Sun, 14:30 - 15:30
  - **Tuesday:**
    - Josh Gardner, 9:00 - 10:00
    - Hugh Sun, 14:30 - 15:30
    - Jakub Filipek, 16:00 - 17:00
    - **Pemi Nguyen, 17:00 -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

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

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

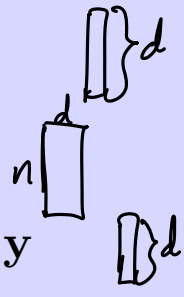
As we derived in class:

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

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

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

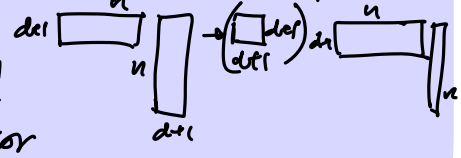
$$\widehat{b}_{\mathrm{LS}} = \frac{1}{n}\sum_{i=1}^{n} y_i - \mu^T\widehat{w}_{\mathrm{LS}}$$



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

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

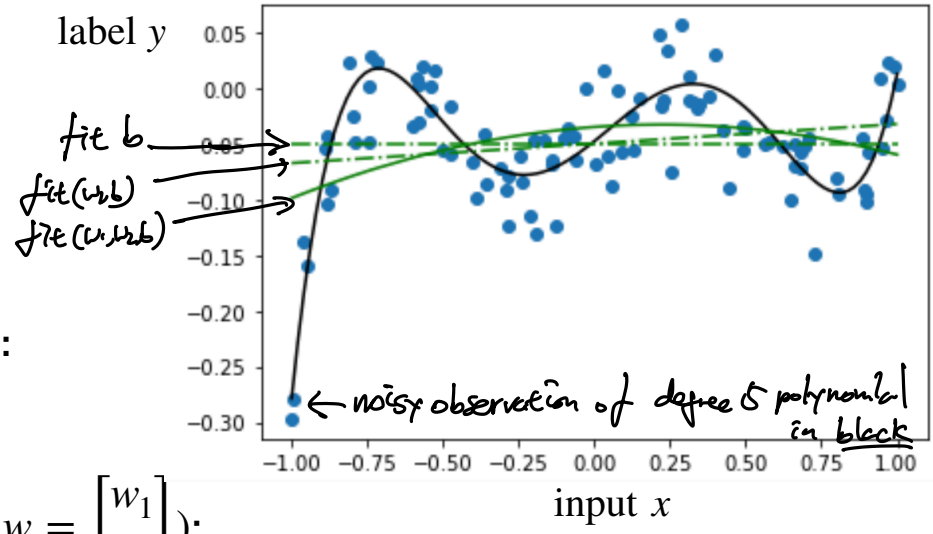
Concatenated as one vector



linear algebra

Quadratic regression in 1-dimension

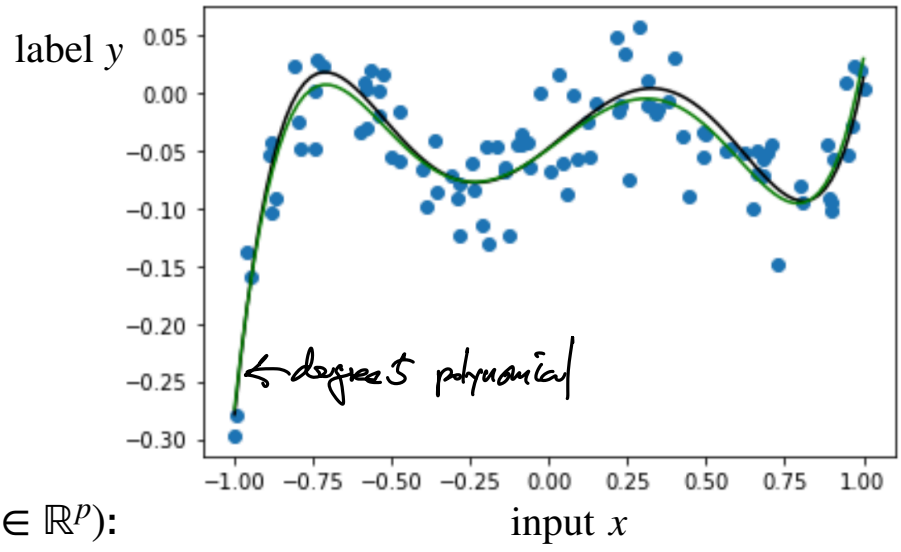
- **Data:** $\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$



- **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 $\mathbf{H} = \begin{bmatrix} h(x_1)^T \\ \vdots \\ h(x_n)^T \end{bmatrix}$
  - $\mathbf{y} = \mathbf{1}b + \mathbf{H}w + \epsilon$

* as p increases, we can represent more complex models.

Q. Which p should we use?
Q. What is downside of large p?
Q. What is downside of small p?

Degree-p polynomial regression in d-dimension

- **Data:** $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ & x_2^T & & \\ & \vdots & & \\ & x_n^T & & \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

let $d=2$, $p=3$

$h(x) = \begin{bmatrix} x_1 \\ x_2 \\ \hline x_1^2 \\ x_2^2 \\ \hline x_1^3 \\ x_2^3 \end{bmatrix}$

- **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 $\mathbf{H} = \begin{bmatrix} h(x_1)^T \\ \vdots \\ h(x_n)^T \end{bmatrix}$
  - $\mathbf{y} = \mathbf{1}b + \mathbf{H}w + \epsilon$

You could also use

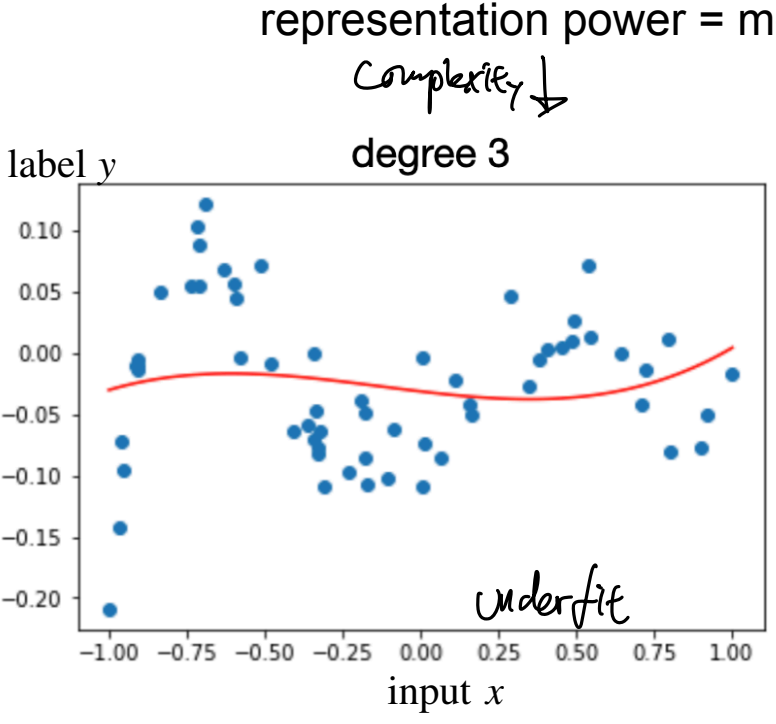
$h(x) = \begin{bmatrix} x_1 \\ x_2 \\ \hline x_1^2 \\ x_2^2 \\ x_1 x_2 \\ \hline x_1^3 \\ x_2^3 \\ x_1^2 x_2 \\ x_1 x_2^2 \\ \hline \sin(x_1 \cdot \frac{2\pi}{f}) \\ \log(x_1) \\ \vdots \end{bmatrix}$

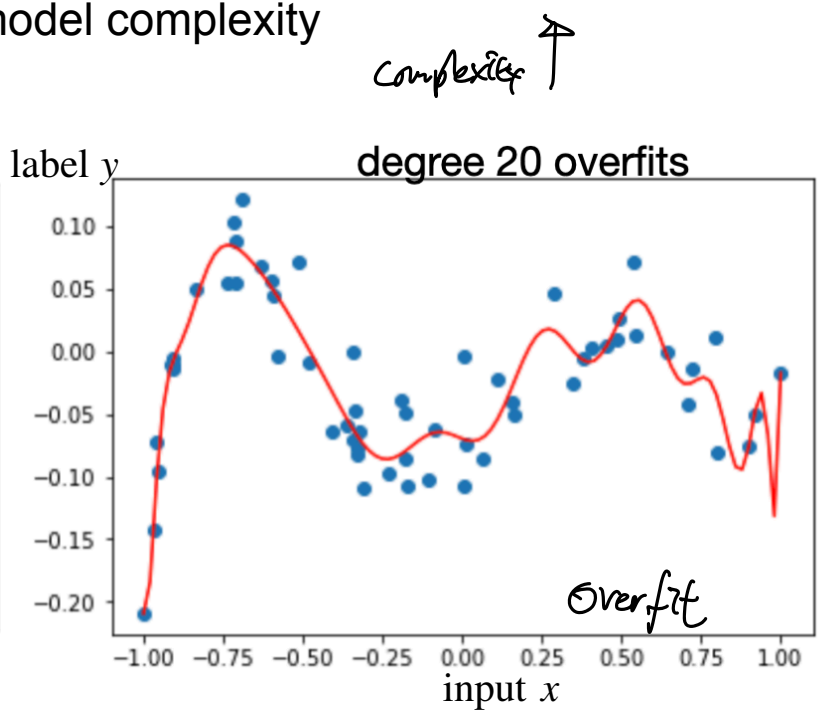
* Requires Domain knowledge.

- 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

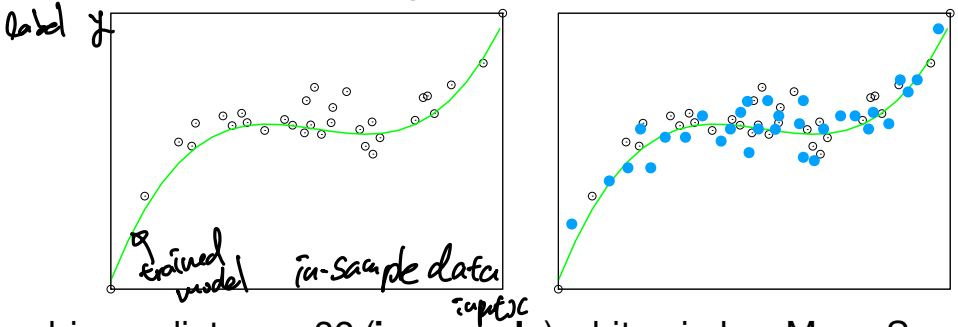




- How do we determine which is better model?

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 lecture)
- 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
  - $$\text{minimize} \quad \mathscr{L}_{\text{train}}(w) = \frac{1}{|S_{\text{train}}|} \sum_{i \in S_{\text{train}}} (y_i - x_i^T w)^2$$
- **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 ← random split is important
- **test set should never be used in training**

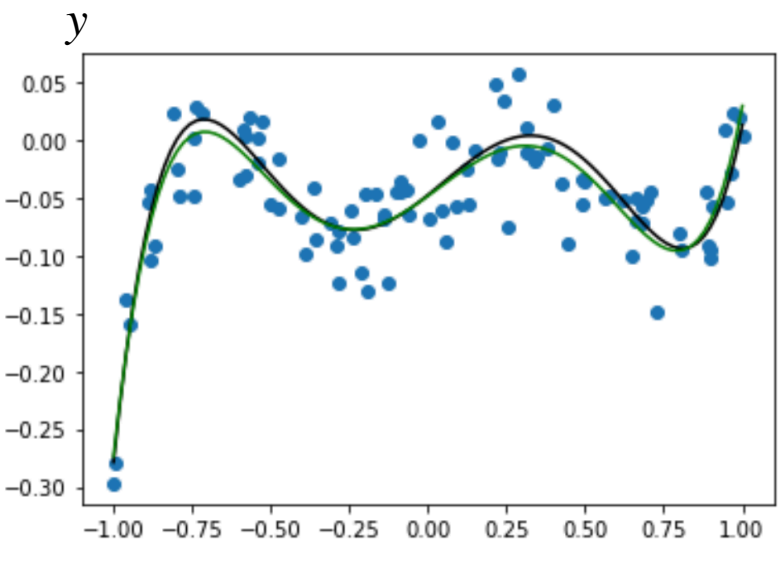
We say a model w or predictor is **overfit** if $\mathscr{L}_{\text{train}}(w) \ll \mathscr{L}_{\text{test}}(w)$

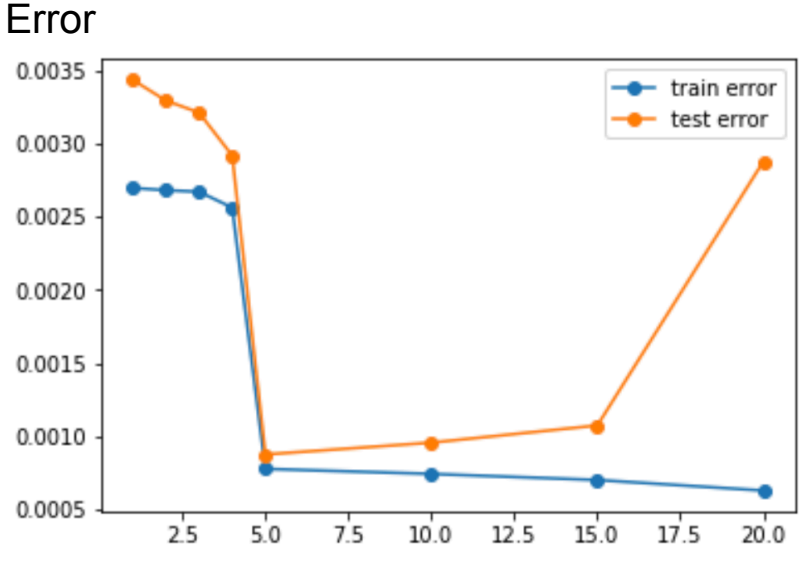
	small training error	large training error
small test error	generalizes well performs well	possible, but unlikely
large test error	fails to generalize **Overfitting**	generalizes well performs poorly

decrease model complexity by e.g. adding regularization.

increase model complexity by e.g. adding more features.

How do we choose which model to use?





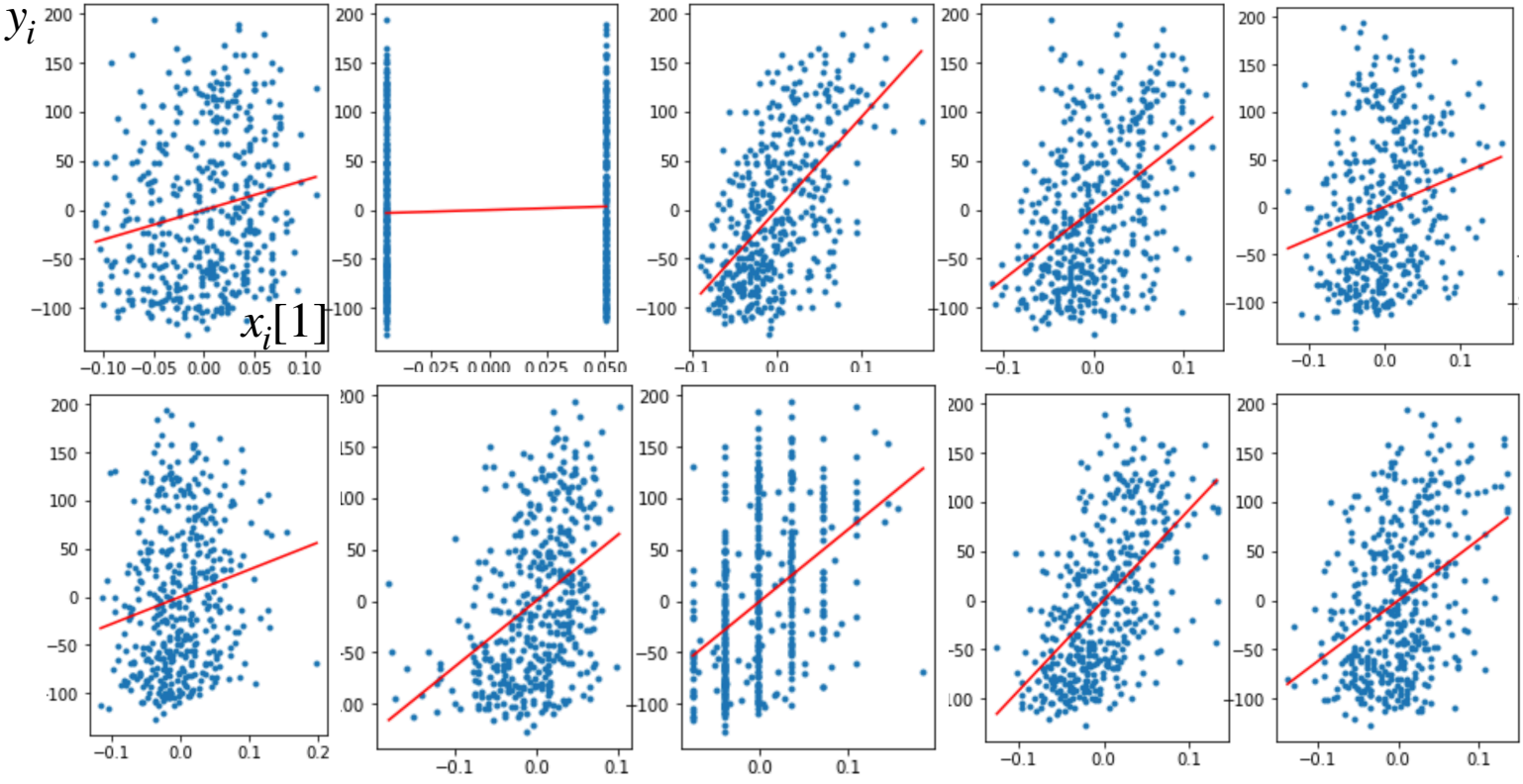
* One-way to do it, we learn more proper way later: (cross-validation)

1. first use 60 data points to train and 60 data points to test and train several models to et 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

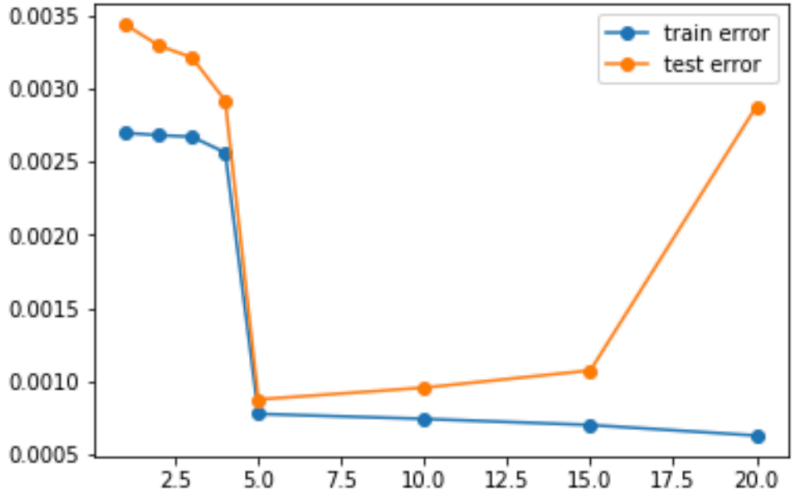
Questions?

Bias-Variance Tradeoff

- explaining test error using theoretical analysis

Train/test error vs. model complexity

Error



Model complexity

e.g., degree p of the polynomial model, number of features used in diabetes,

Usually related to the dimension of the model parameter

Statistical learning

- Suppose data is generated from a statistical model $(X, Y) \sim P_{X,Y}$
  - and we know $P_{X,Y}$
- Then learning is to find a function $\eta : \mathbb{R}^d \to \mathbb{R}$ that minimizes
  - the expected error $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2]$
- Since, we do not assume anything about the function $\eta(X)$, it can take any value for each X, hence the optimization can be broken into sum (or more precisely integral) of multiple objective functions, each involving a specific value $X = x$
  - $$\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2] = \mathbb{E}_{X\sim P_X}\left[\mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]\right]$$
  $$= \int \mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]\, P_X(x)\, dx$$

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

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

Statistical learning

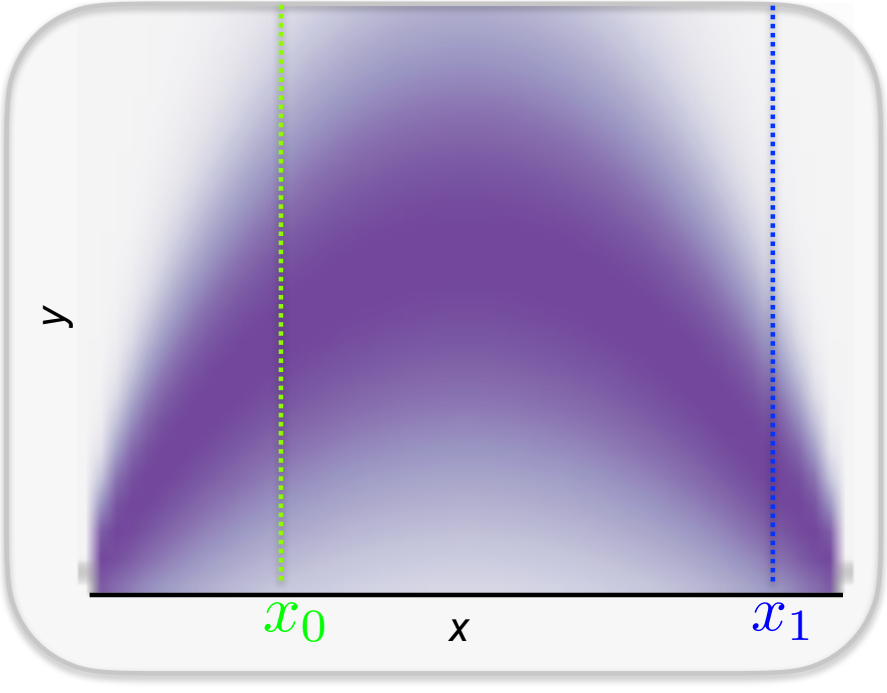
- We can solve the optimization for each $X = x$ separately
  - $\eta(x) = \arg\min_{a \in \mathbb{R}} \mathbb{E}_{Y \sim P_{Y|X}}[(Y - a)^2 \,|\, X = x]$
- The optimal solution is $\eta(x) = \mathbb{E}_{Y \sim P_{Y|X}}[Y|X=x]$,
  which is the best prediction in ℓ_2-loss/Mean Squared Error

Statistical Learning

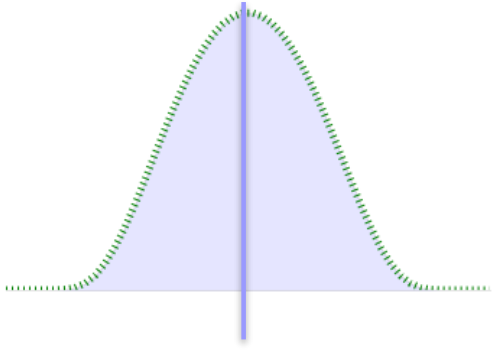
Ideally, we want to find:

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

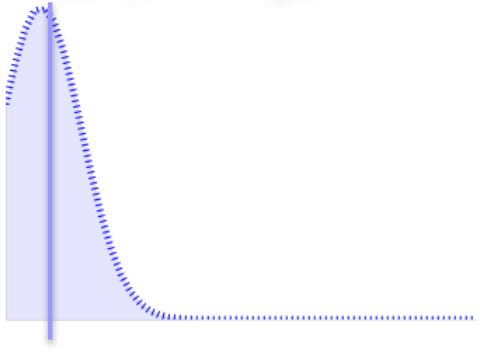
$P_{XY}(X = x, Y = y)$



$P_{XY}(Y = y|X = x_0)$

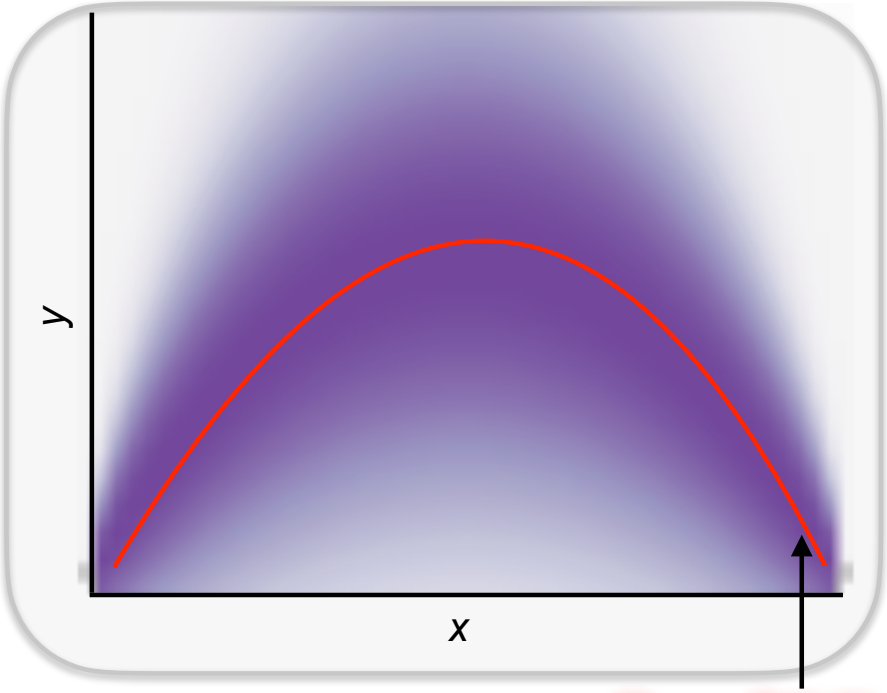


$P_{XY}(Y = y|X = x_1)$



Statistical Learning

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

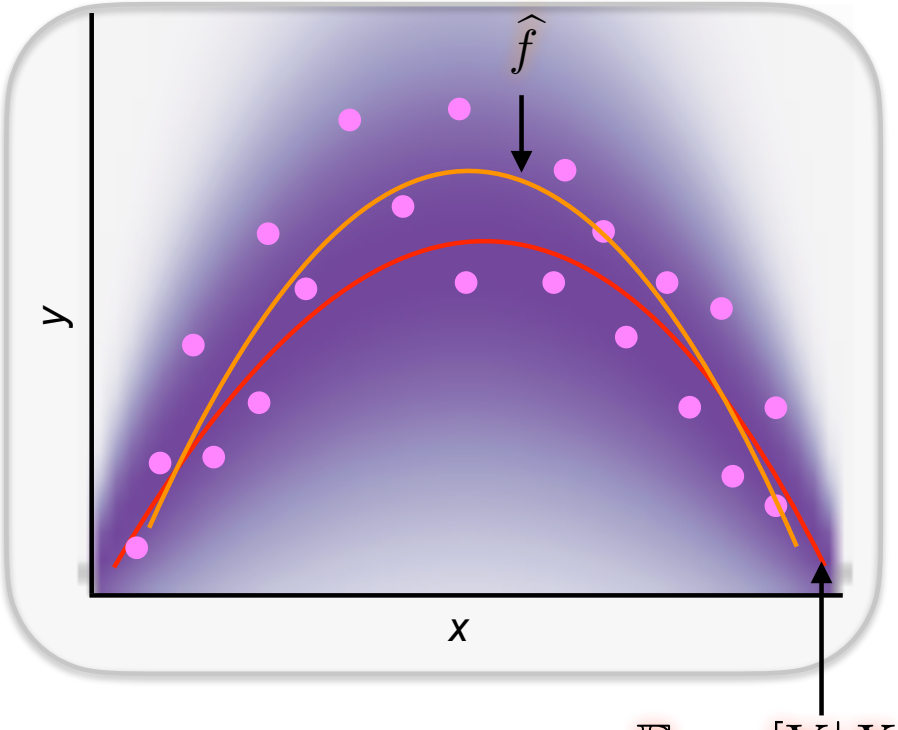


Ideally, we want to find:

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

Statistical Learning

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



Ideally, we want to find:

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

But we only have samples:

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

and are restricted to a function class (e.g., linear) so we compute:

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

We care about future predictions: $\mathbb{E}_{XY}[(Y - \widehat{f}(X))^2]$

Future prediction error $\mathbb{E}_{X,Y\sim P_{X,Y}}[(Y-\hat{f}(X))^2]$ is random because $\hat{f}$ is random

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

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

Bias-variance tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

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

Bias-variance tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

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

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

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

irreducible error
Caused by stochastic label noise

learning error
Caused by either using too "simple" of a model or not enough data to learn the model accurately

Bias-variance tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

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

Bias-variance tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

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

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

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

biased squared **variance**

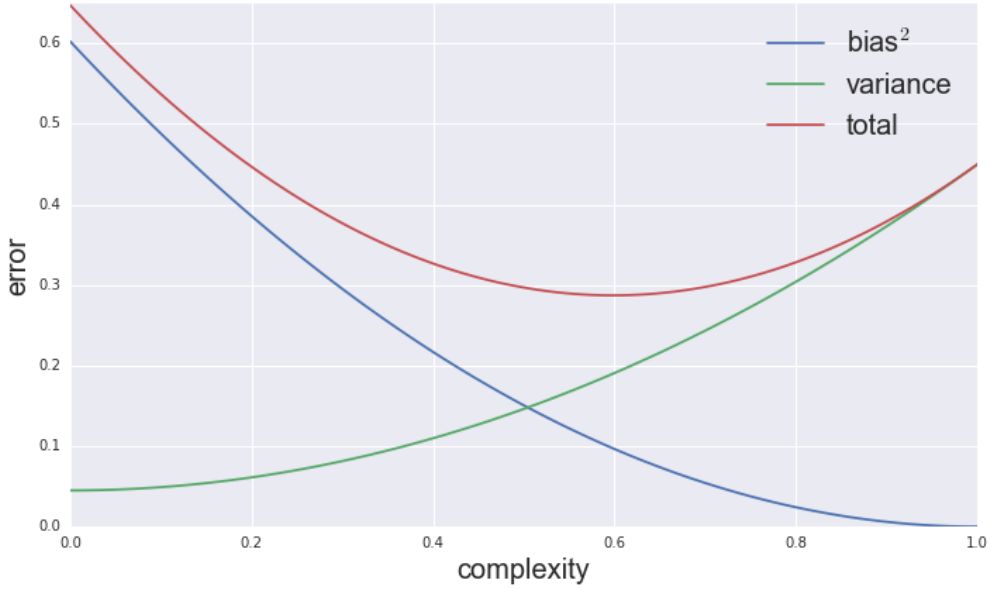
Bias-variance tradeoff

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

irreducible error

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

biased squared **variance**



Lecture 6: Bias-Variance Tradeoff (continued)

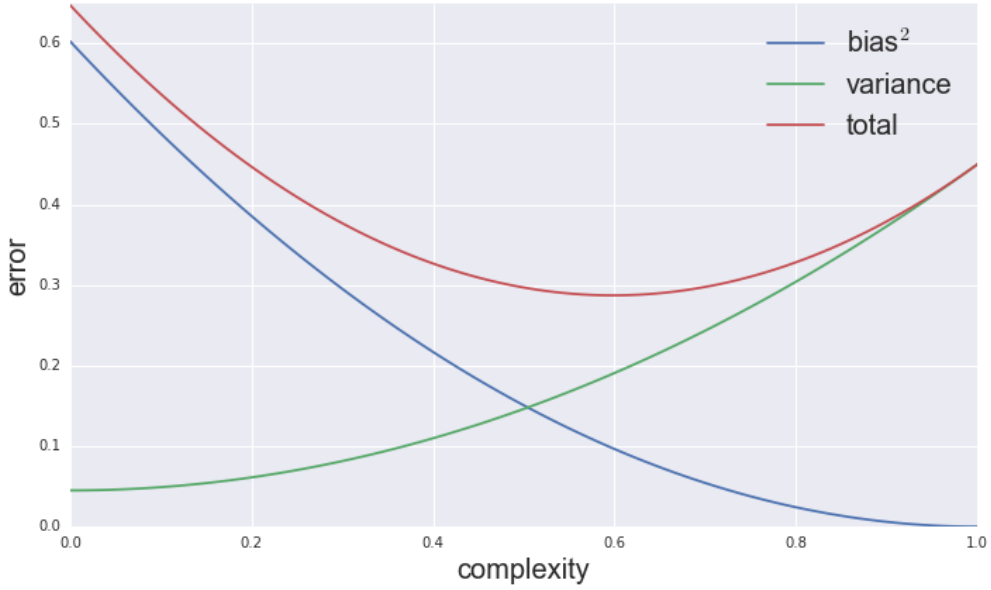
Recap: Bias-variance tradeoff

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

irreducible error

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

biased squared **variance**



Bias-variance tradeoff for linear models

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

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

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

$$\widehat{f}_{\mathcal{D}}(x) =$$

Bias-variance tradeoff for linear models

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

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

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

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

$\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x] = \sigma^2$

irreducible error

$(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2$ ·

biased squared

Bias-variance tradeoff for linear models

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

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

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

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

variance

Bias-variance tradeoff for linear models

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

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

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

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}} = \mathbb{E}_{\mathcal{D}}[x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x]$$

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

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

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} x_i x_i^T \overset{n \text{ large}}{\to} n\Sigma \qquad \Sigma = \mathbb{E}[XX^T], \quad X \sim P_X$$

$$\mathbb{E}_{X=x}\left[\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]\right] = \frac{\sigma^2}{n}\mathbb{E}_X[\text{Trace}(\Sigma^{-1}XX^T)] = \frac{d\sigma^2}{n}$$

Bias-variance tradeoff for linear models

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

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

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

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

$$\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x] = \sigma^2$$

irreducible error

$$(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2 = 0$$

biased squared

$$\mathbb{E}_{X=x}\left[\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]\right] = \frac{d\sigma^2}{n}$$

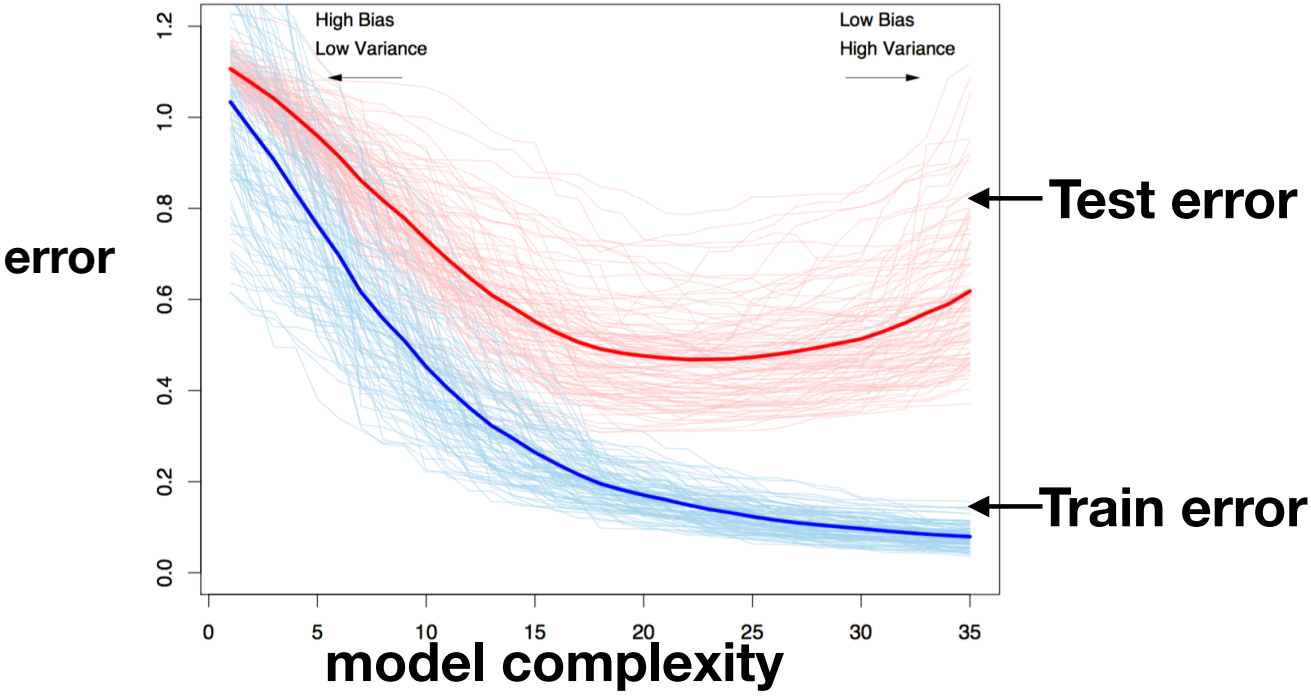
variance

What does the bias-variance theory tell us?

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