

Cross-Validation

How... How... How???????

- > How do we pick the number of basis functions...
- > We could use the test data, but...

(LOO) Leave-one-out cross validation

- > Consider a validation set with 1 example:
 - D – training data
 - $D \setminus j$ – training data with j th data point (x_j, y_j) moved to validation set
- > Learn classifier $f_{D \setminus j}$ with $D \setminus j$ dataset
- > Estimate true error as squared error on predicting y_j :
 - Unbiased estimate of error $\text{error}_{\text{true}}(f_{D \setminus j})!$

(LOO) Leave-one-out cross validation

- > Consider a validation set with 1 example:
 - D – training data
 - $D \setminus j$ – training data with j th data point (x_j, y_j) moved to validation set
- > Learn classifier $f_{D \setminus j}$ with $D \setminus j$ dataset
- > Estimate true error as squared error on predicting y_j :
 - Unbiased estimate of error $\text{error}_{\text{true}}(f_{D \setminus j})!$
- > LOO cross validation: Average over all data points j :
 - For each data point you leave out, learn a new classifier $f_{D \setminus j}$
 - Estimate error as:

$$\text{error}_{LOO} = \frac{1}{n} \sum_{j=1}^n (y_j - f_{D \setminus j}(x_j))^2$$

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
 - Usually pessimistic, though – learning with less data typically gives worse answer
- > LOO is almost unbiased! Use LOO error for model selection!!!
 - E.g., picking degree

Computational cost of LOO

- > **Suppose you have 100,000 data points**
- > **You implemented a great version of your learning algorithm**
 - **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, \dots, D_k

> For each i

– Learn classifier $f_{D \setminus D_i}$ using data point not in D_i

– Estimate error of $f_{D \setminus D_i}$ on validation set D_i :

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

1	2	3	4	5
Train	Train	Validation	Train	Train

Use k -fold cross validation

> Randomly divide training data into k equal parts

- D_1, \dots, D_k

1	2	3	4	5
Train	Train	Validation	Train	Train

> For each i

- Learn classifier $f_{D \setminus D_i}$ using data point not in D_i
- Estimate error of $f_{D \setminus D_i}$ on validation set D_i :

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

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

$$\text{error}_{k\text{-fold}} = \frac{1}{k} \sum_{i=1}^k \text{error}_{\mathcal{D}_i}$$

> k -fold cross validation properties:

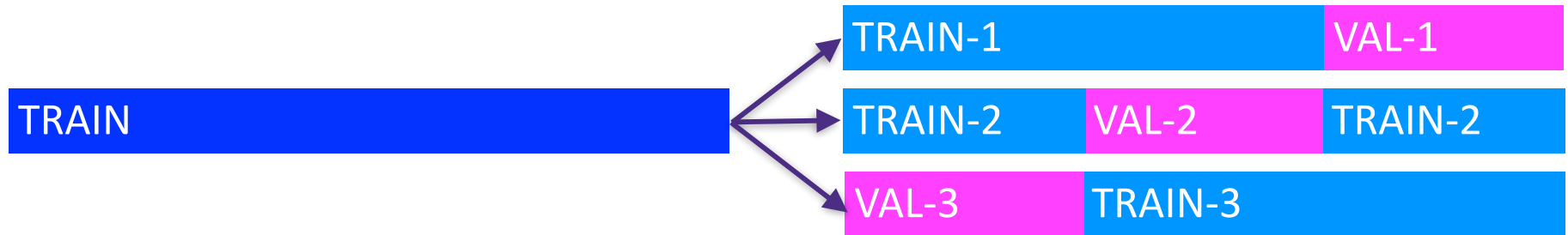
- Much faster to compute than LOO
- More (pessimistically) biased – using much less data, only $n(k-1)/k$
- Usually, $k = 10$

Recap

- > Given a dataset, begin by splitting into



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



- > Model assessment: Use **TEST** to assess the accuracy of the model you output
 - **Never ever ever ever ever train or choose 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

$$\frac{|\sum_{i=1}^n x_{i,j} y_i|}{\sqrt{\sum_{i=1}^n x_{i,j}^2}}$$

- > 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?

Recap

- > Learning is...
 - Collect some data
 - > E.g., housing info and sale price
 - Randomly split dataset into **TRAIN**, **VAL**, and **TEST**
 - > E.g., **80%**, **10%**, and **10%**, respectively
 - Choose a hypothesis class or model
 - > E.g., **linear with non-linear transformations**
 - Choose a loss function
 - > E.g., least squares **on TRAIN**
 - Choose an optimization procedure
 - > E.g., set derivative to zero to obtain estimator, **cross-validation on VAL to pick num. features**
- > Justifying the accuracy of the estimate
 - > E.g., report **TEST error**

Bias-Variance Tradeoff

Optimal Prediction

Goal: Predict $Y \in \mathbb{R}$ given $X \in \mathbb{R}^d$ if $(X, Y) \sim P_{XY}$

Find function η that minimizes

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

(Hint: for any x , $\eta(x) = c_x$ where c_x minimizes $\mathbb{E}_{Y|X}[(Y - c_x)^2 | X = x]$)

$$\frac{\partial}{\partial c} \mathbb{E}_{Y|X} [(Y - c)^2 | X = x]$$

$$= \mathbb{E}_{Y|X} [2(Y - c)(-1) | X = x]$$

$$= -2 \left(\mathbb{E}_{Y|X} [Y | X = x] - c \right) = 0$$

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

Optimal Prediction

Goal: Predict $Y \in \mathbb{R}^d$ given $X \in \mathbb{R}^d$ if $(X, Y) \sim P_{XY}$

Find function η that minimizes

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

(Hint: for any x , $\eta(x) = c_x$ where c_x minimizes $\mathbb{E}_{Y|X}[(Y - c_x)^2 | X = x]$)

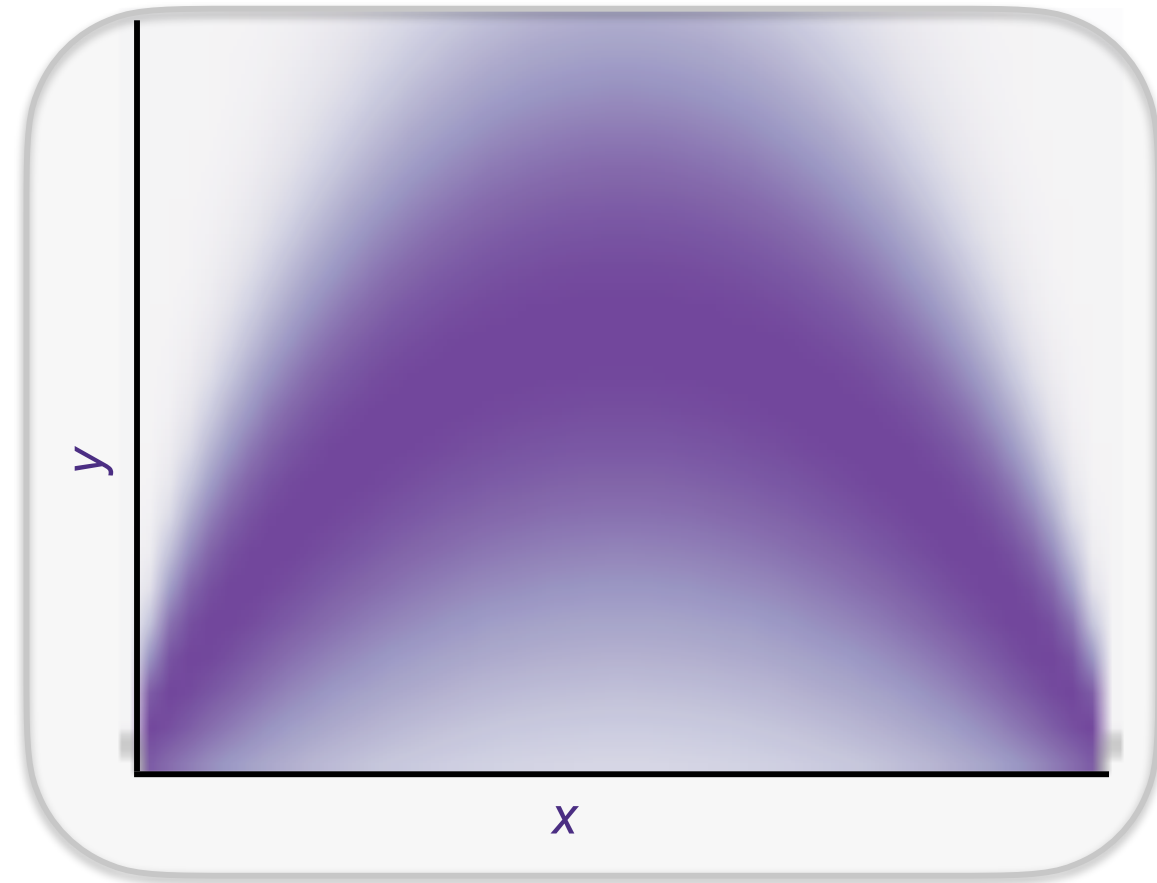
$$\begin{aligned} 0 &= \frac{d}{dc_x} \mathbb{E}_{Y|X}[(Y - c_x)^2 | X = x] \\ &= \mathbb{E}_{Y|X} \left[\frac{d}{dc_x} (Y - c_x)^2 | X = x \right] \\ &= \mathbb{E}_{Y|X}[-2(Y - c_x) | X = x] = -2\mathbb{E}_{Y|X}[Y | X = x] + 2c_x \end{aligned}$$

Squared Error Optimal Predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y | X = x]$

Statistical Learning

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

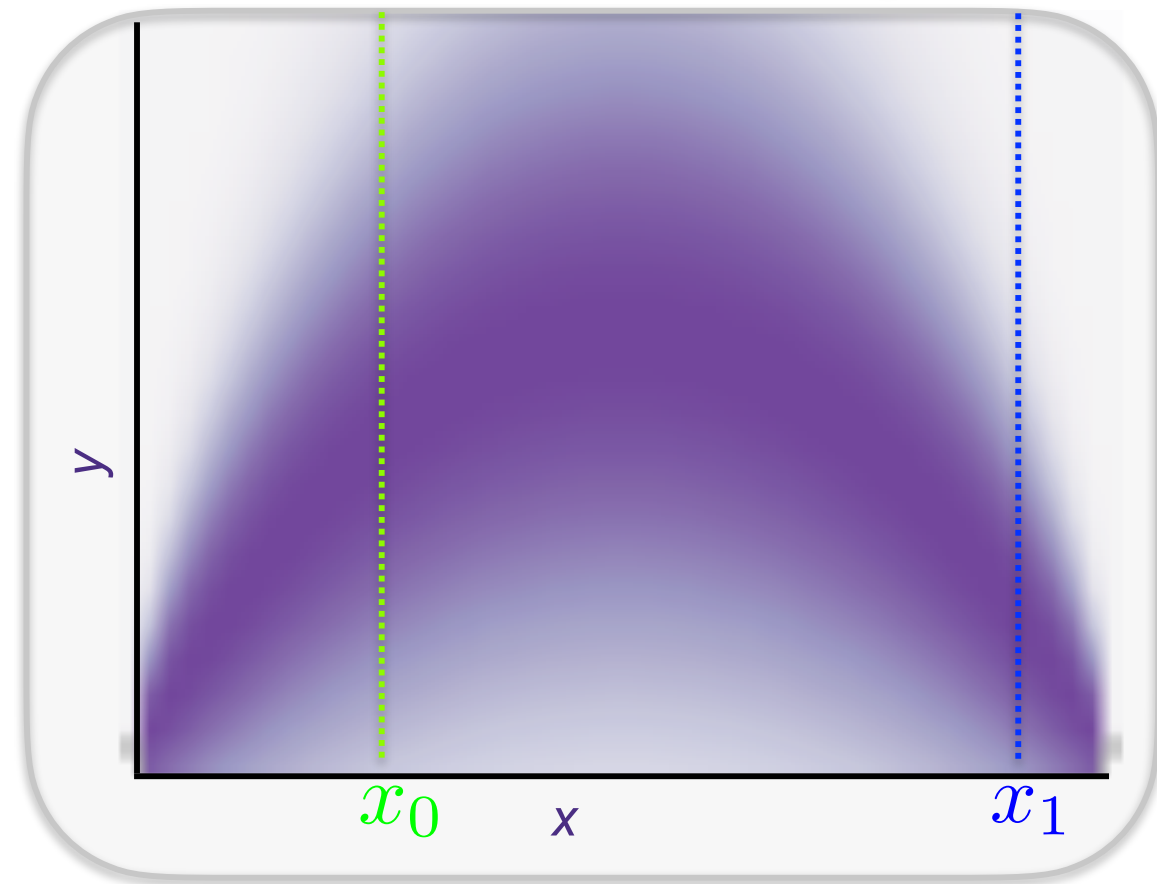
$$P_{XY}(X = x, Y = y)$$



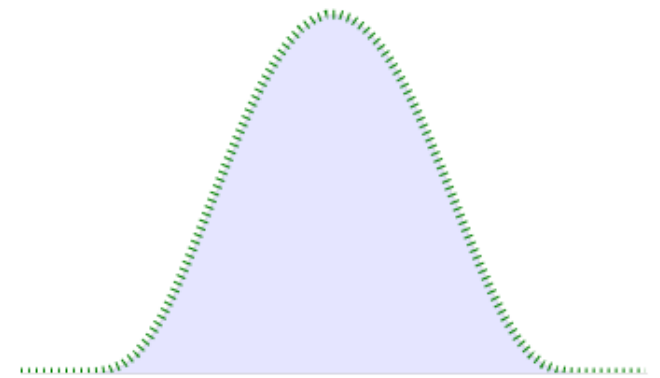
Statistical Learning

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

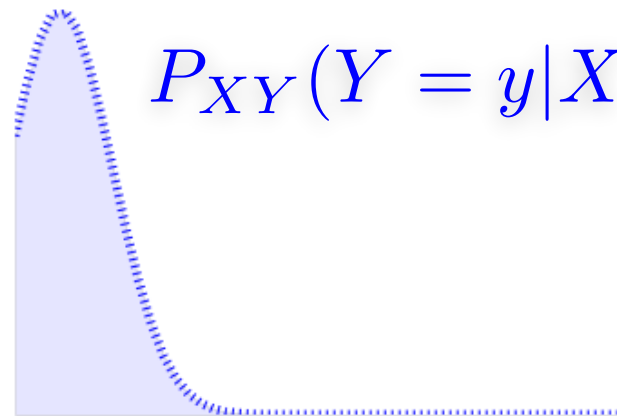
$$P_{XY}(X = x, Y = y)$$



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



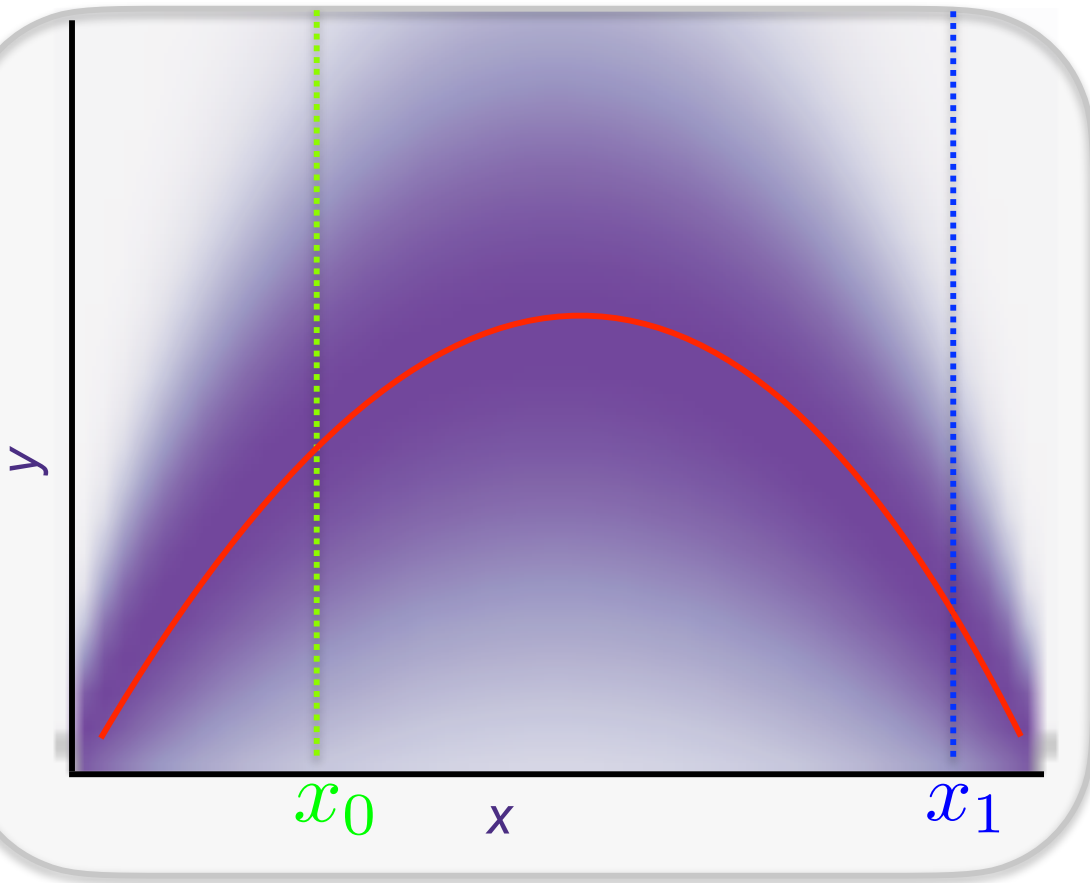
$$P_{XY}(Y = y | X = x_1)$$



Statistical Learning

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

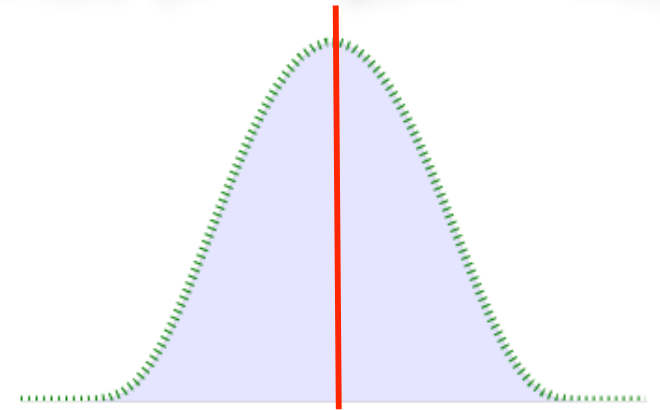
$$P_{XY}(X = x, Y = y)$$



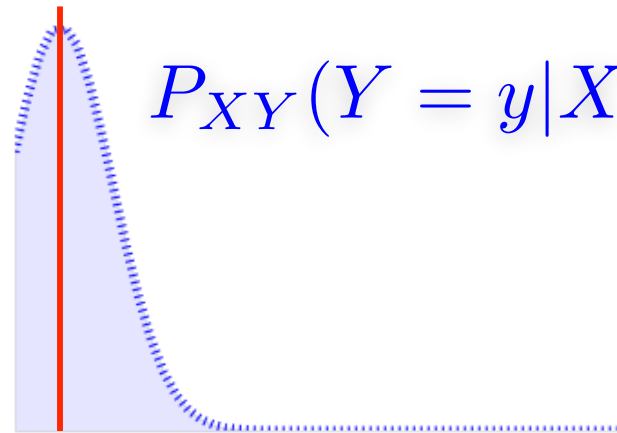
Ideally, we want to find:

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

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



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

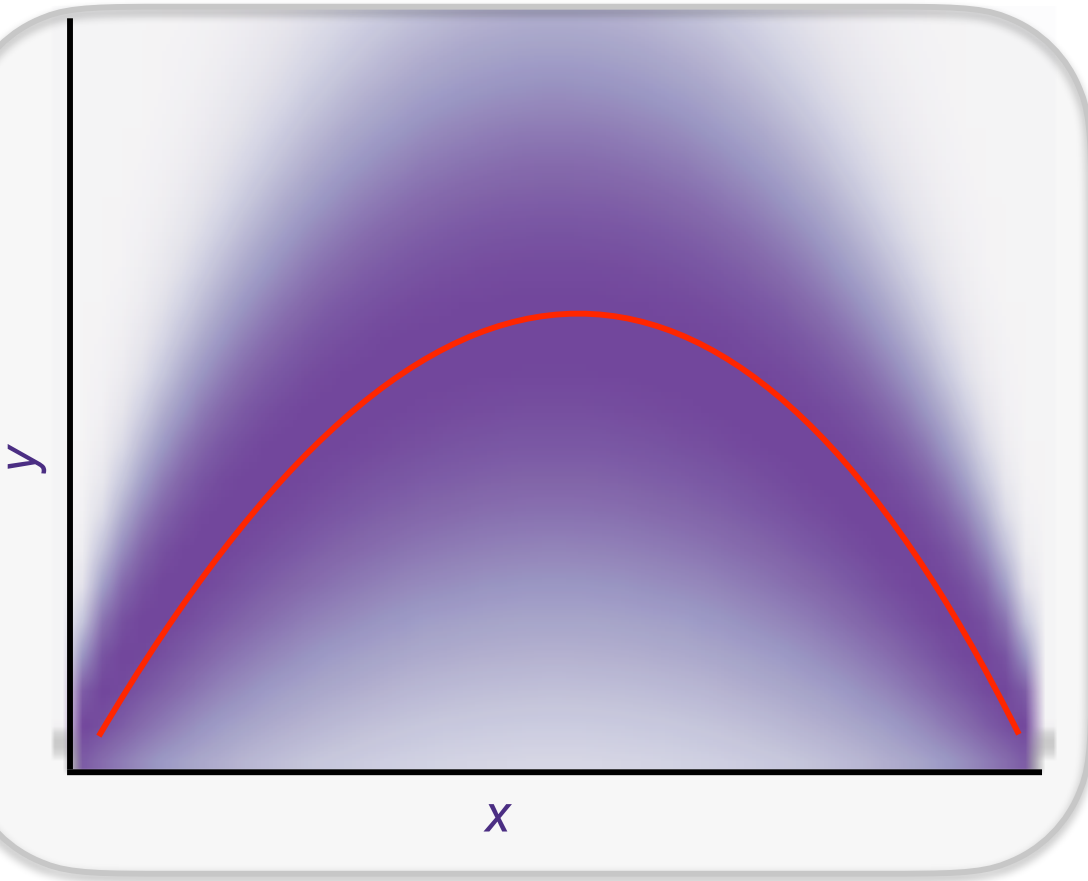


Statistical Learning

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

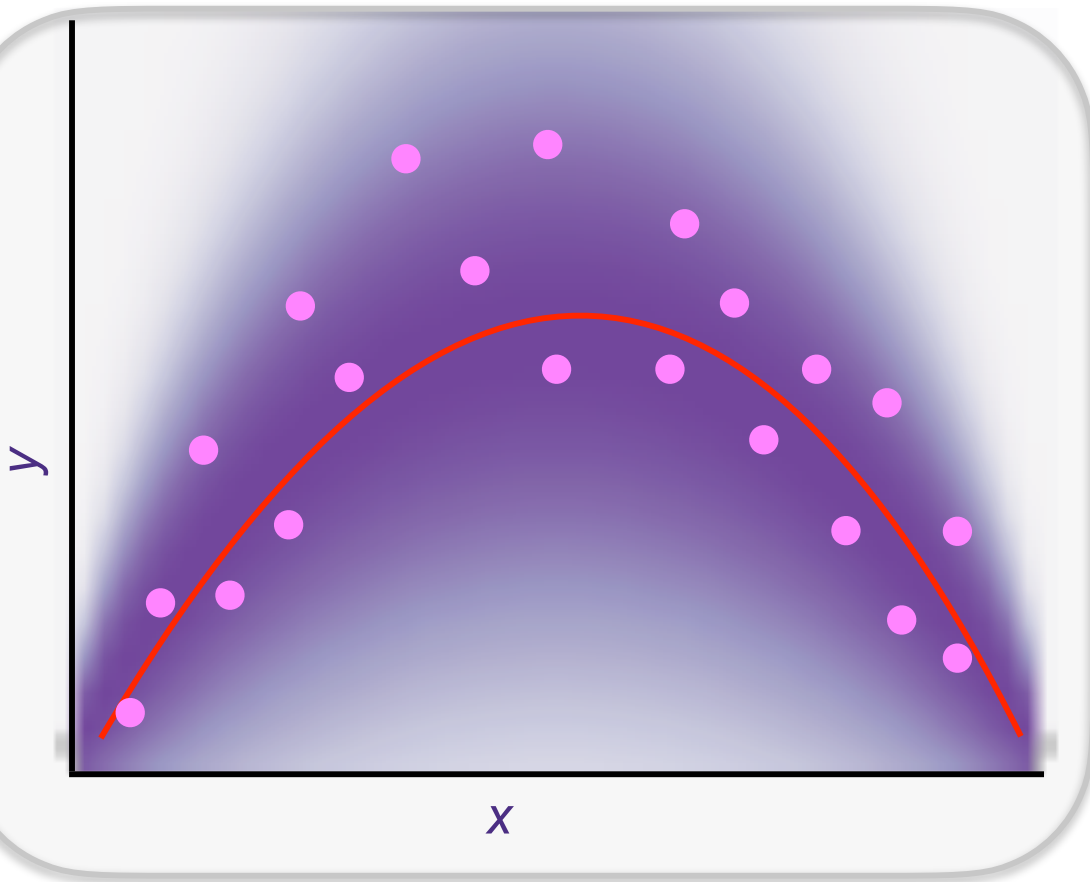
Ideally, we want to find:

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



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:

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

$$\mathcal{F} = \left\{ f: \mathbb{R}^d \rightarrow \mathbb{R} : f(x) = w^\top x, w \in \mathbb{R}^d \right\}$$

Statistical Learning

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} [(y_i - f(x_i))^2] = \mathbb{E} [(y_1 - f(x_1))^2]$$

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

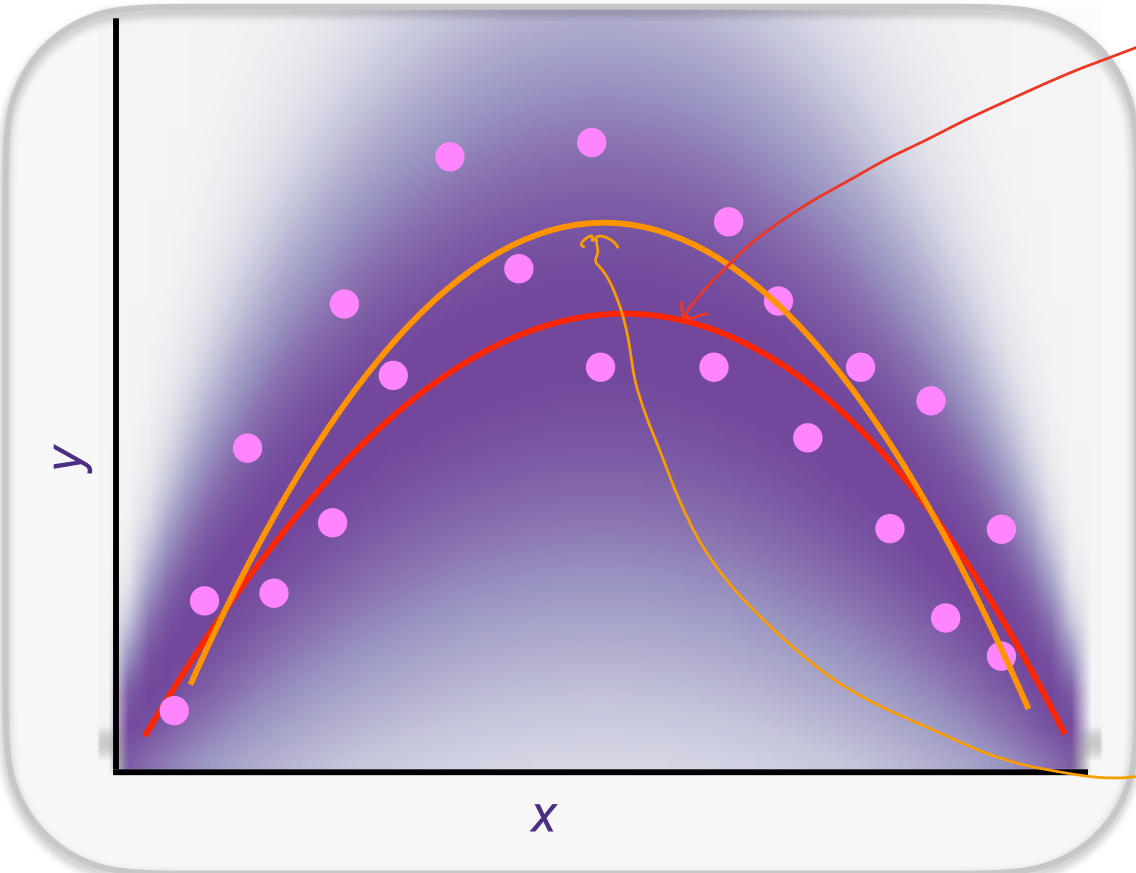
Ideally, we want to find:

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

But we only have samples:
 $(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY}$ for $i = 1, \dots, n$

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

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

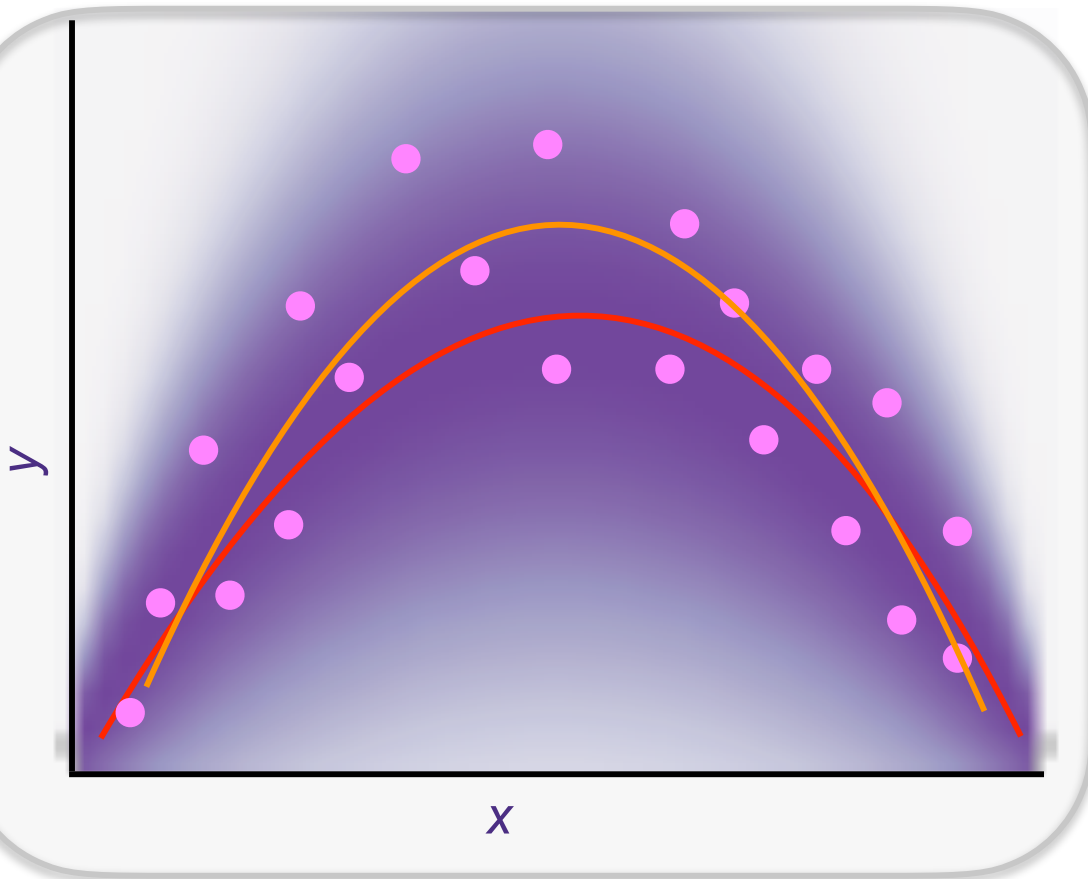


fix any f ,
then

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \right] = \mathbb{E} [(Y - f(x))^2] = \mathbb{E}_x \left[\mathbb{E}_{Y|X} [(Y - f(x))^2 | X=x] \right]$$

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:

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

and are restricted to a function class (e.g., linear)

so we compute:

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

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

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:
 $(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY}$ for $i = 1, \dots, n$

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

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

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

Statistical Learning

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

Ideally, we want to find:

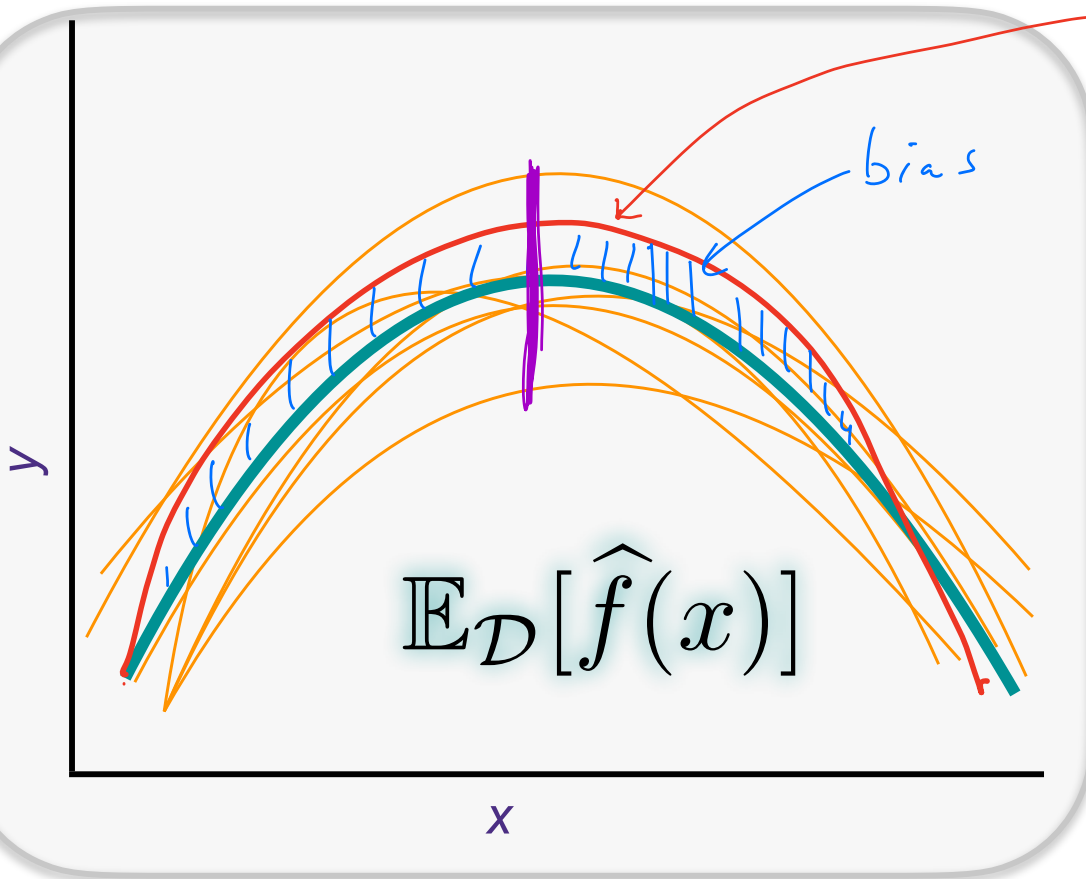
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:
 $(x_i, y_i) \stackrel{i.i.d.}{\sim} P_{XY}$ for $i = 1, \dots, n$

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

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

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



Bias-Variance Tradeoff

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

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

$$(a+b)^2 = a^2 + 2ab + b^2$$

Bias-Variance Tradeoff

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

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

$$\begin{aligned} \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] &= \mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(\underbrace{Y - \eta(x)}_a + \underbrace{\eta(x) - \hat{f}_{\mathcal{D}}(x)}_b)^2] | X = x] \\ &= \mathbb{E}_{Y|X} \left[\mathbb{E}_{\mathcal{D}} \left[(Y - \eta(x))^2 + \underbrace{2(Y - \eta(x))(\eta(x) - \hat{f}_{\mathcal{D}}(x))}_{2 \cdot b} + (\eta(x) - \hat{f}_{\mathcal{D}}(x))^2 \right] \right] \\ &= \mathbb{E}_{Y|X} \left[(Y - \eta(x))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2 \right] \end{aligned}$$

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

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

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

Bias-Variance Tradeoff

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

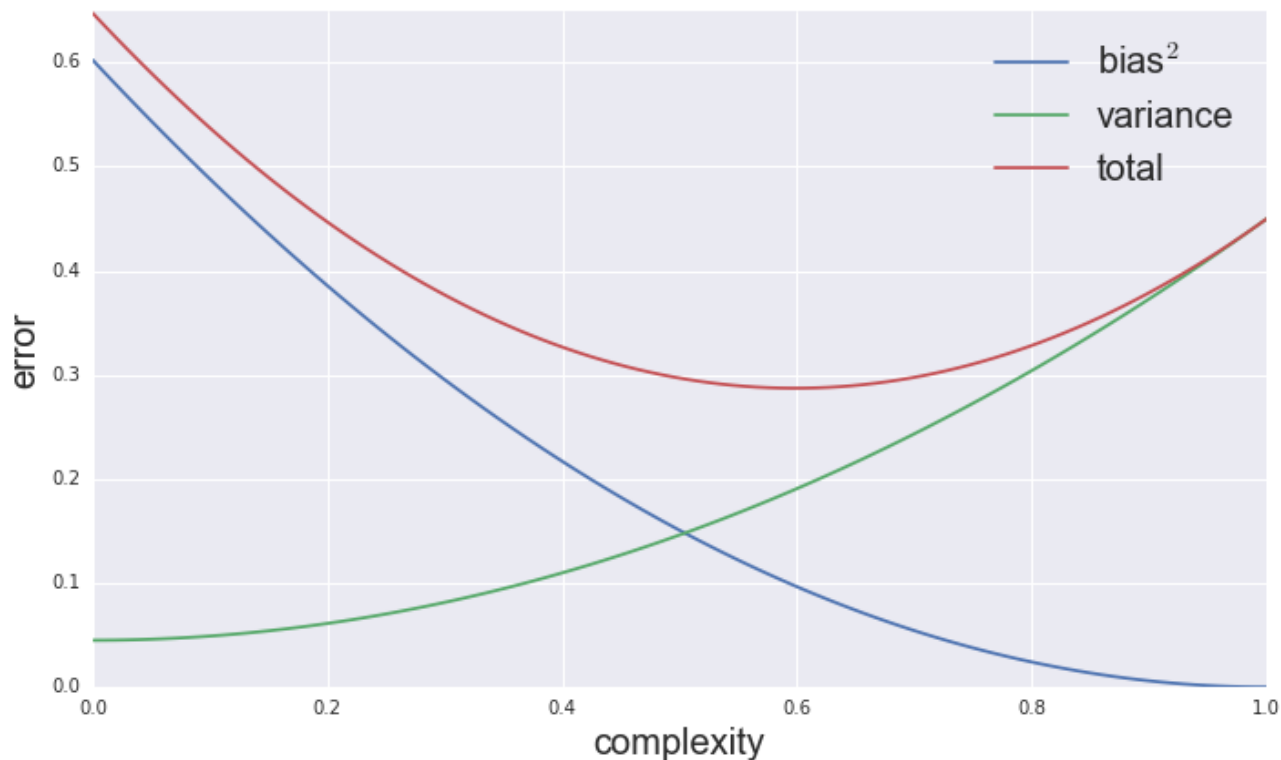
$$\begin{aligned} \underline{\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2]} &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= \mathbb{E}_{\mathcal{D}}[(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2 + 2(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)) \\ &\quad + (\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2] \\ &= \underline{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2} + \underline{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]} \end{aligned}$$

biased squared

variance

Bias-Variance Tradeoff

$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \hat{f}_{\mathcal{D}}(x))^2] | X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} + \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2}_{\text{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}}$$



Bias-Variance demo

Ridge Regression

Regularization in Linear Regression

Recall Least Squares: $\hat{w}_{LS} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$

$$= \arg \min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)$$

when $(\mathbf{X}^T \mathbf{X})^{-1}$ exists.... $= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

Regularization in Linear Regression

Recall Least Squares: $\hat{w}_{LS} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$

$$= \arg \min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)$$

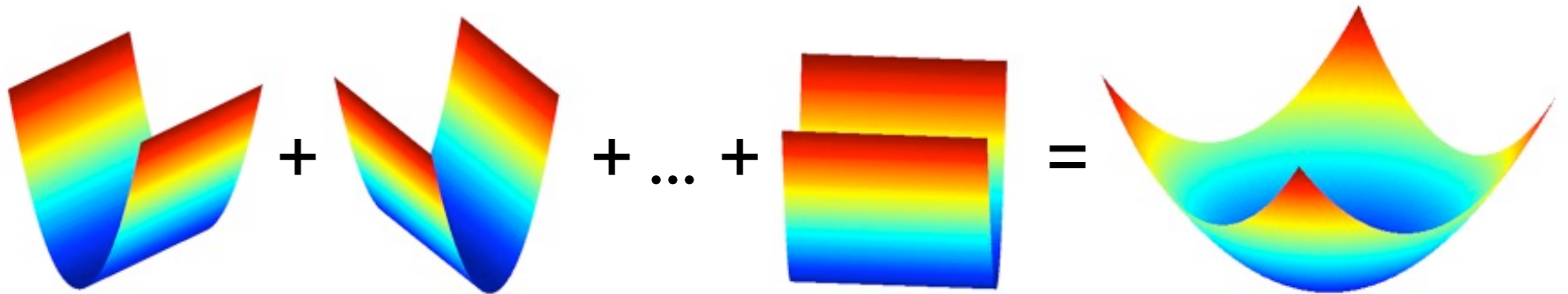
In general: $= \arg \min_w w^T (\mathbf{X}^T \mathbf{X}) w - 2y^T \mathbf{X}w$

Regularization in Linear Regression

Recall Least Squares: $\hat{w}_{LS} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$

$$= \arg \min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)$$

In general: $= \arg \min_w w^T (\mathbf{X}^T \mathbf{X})w - 2y^T \mathbf{X}w$



$$(y_1 - x_1^T w)^2 + (y_2 - x_2^T w)^2 + \dots + (y_n - x_n^T w)^2 = \sum_{i=1}^n (y_i - x_i^T w)^2$$

What if $x_i \in \mathbb{R}^d$ and $d > n$?

Regularization in Linear Regression

Recall Least Squares: $\hat{w}_{LS} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$

When $x_i \in \mathbb{R}^d$ and $d > n$ the objective function is flat in some directions:



Regularization in Linear Regression

Recall Least Squares: $\hat{w}_{LS} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$

When $x_i \in \mathbb{R}^d$ and $d > n$ the objective function is flat in some directions:

Implies optimal solution is *not unique* and unstable due to lack of curvature:

- small changes in training data result in large changes in solution
- often the *magnitudes* of w are “very large”

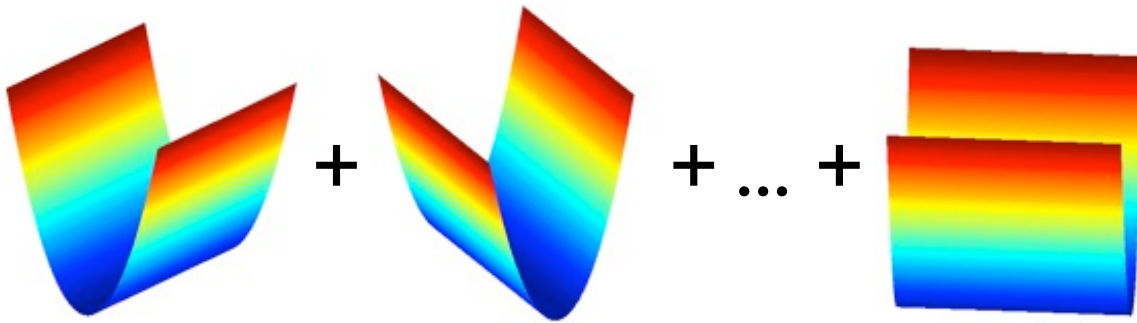


Regularization imposes “simpler” solutions by a “complexity” penalty

Ridge Regression

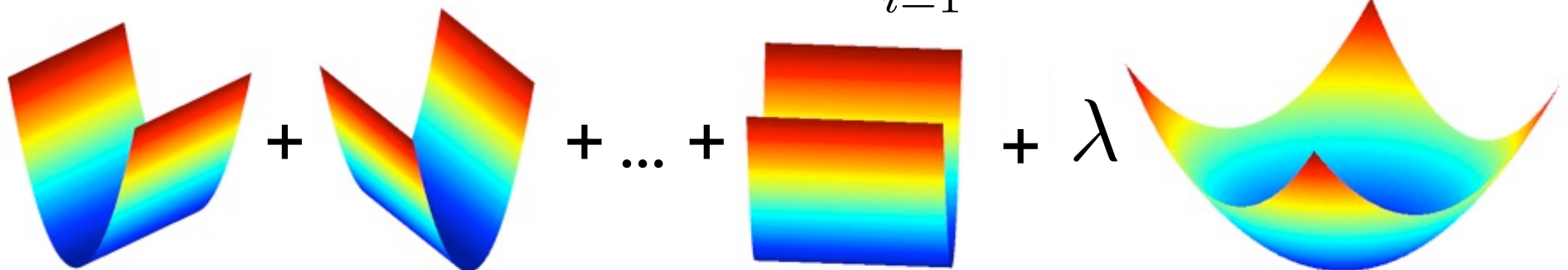
- Old Least squares objective:

$$\hat{w}_{LS} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$



- Ridge Regression objective:

$$\hat{w}_{ridge} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2 + \lambda ||w||_2^2$$



Minimizing the Ridge Regression Objective

$$\hat{w}_{ridge} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2 + \lambda \|w\|_2^2$$

Shrinkage Properties

$$\begin{aligned}\hat{w}_{ridge} &= \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2 + \lambda \|w\|_2^2 \\ &= (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y}\end{aligned}$$

Bias-Variance Properties

$$\hat{w}_{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y}$$

- **Assume:** $\mathbf{X}^T \mathbf{X} = nI$ **and** $\mathbf{y} = \mathbf{X}w + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x, \text{train}}[(Y - x^T \hat{w}_{ridge})^2 | X = x]$?

Bias-Variance Properties

$$\hat{w}_{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y}$$

- **Assume:** $\mathbf{X}^T \mathbf{X} = nI$ **and** $\mathbf{y} = \mathbf{X}w + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x, \text{train}}[(Y - x^T \hat{w}_{ridge})^2 | X = x]$?

$$\begin{aligned} & \mathbb{E}_{Y|X, \mathcal{D}}[(Y - x^T \hat{w}_{ridge})^2 | X = x] \\ &= \underbrace{\mathbb{E}_{Y|X}[(Y - \mathbb{E}_{Y|X}[Y|X = x])^2 | X = x]}_{\text{Irreducible Error}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{Y|X}[Y|X = x] - x^T \hat{w}_{ridge})^2]}_{\text{Learning Error}} \end{aligned}$$

Bias-Variance Properties

$$\hat{w}_{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y}$$

- Assume: $\mathbf{X}^T \mathbf{X} = nI$ and $\mathbf{y} = \mathbf{X}w + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x, \text{train}}[(Y - x^T \hat{w}_{ridge})^2 | X = x]$?

$$\begin{aligned} & \mathbb{E}_{Y|X, \mathcal{D}}[(Y - x^T \hat{w}_{ridge})^2 | X = x] \\ &= \mathbb{E}_{Y|X}[(Y - \mathbb{E}_{Y|X}[Y|X = x])^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{Y|X}[Y|X = x] - x^T \hat{w}_{ridge})^2] \\ &= \mathbb{E}_{Y|X}[(Y - x^T w)^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(x^T w - x^T \hat{w}_{ridge})^2] \\ &= \underbrace{\sigma^2}_{\text{Irreduc. Error}} + \underbrace{(x^T w - \mathbb{E}_{\mathcal{D}}[x^T \hat{w}_{ridge}])^2}_{\text{Bias-squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[x^T \hat{w}_{ridge}] - x^T \hat{w}_{ridge})^2]}_{\text{Variance}} \end{aligned}$$

Bias-Variance Properties

$$\hat{w}_{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y}$$

- Assume: $\mathbf{X}^T \mathbf{X} = nI$ and $\mathbf{y} = \mathbf{X}w + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x, \text{train}}[(Y - x^T \hat{w}_{ridge})^2 | X = x]$?

$$\begin{aligned} & \mathbb{E}_{Y|X, \mathcal{D}}[(Y - x^T \hat{w}_{ridge})^2 | X = x] \\ &= \mathbb{E}_{Y|X}[(Y - \mathbb{E}_{Y|X}[Y|X = x])^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{Y|X}[Y|X = x] - x^T \hat{w}_{ridge})^2] \\ &= \mathbb{E}_{Y|X}[(Y - x^T w)^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(x^T w - x^T \hat{w}_{ridge})^2] \\ &= \underbrace{\sigma^2}_{\text{Irreduc. Error}} + \underbrace{(x^T w - \mathbb{E}_{\mathcal{D}}[x^T \hat{w}_{ridge}])^2}_{\text{Bias-squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[x^T \hat{w}_{ridge}] - x^T \hat{w}_{ridge})^2]}_{\text{Variance}} \end{aligned}$$

$$\begin{aligned} \hat{w}_{ridge} &= (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T (\mathbf{X}w + \epsilon) \\ &= \frac{n}{n + \lambda} w + \frac{1}{n + \lambda} \mathbf{X}^T \epsilon \end{aligned}$$

Bias-Variance Properties

$$\hat{w}_{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y}$$

- Assume: $\mathbf{X}^T \mathbf{X} = nI$ and $\mathbf{y} = \mathbf{X}w + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x, \text{train}}[(Y - x^T \hat{w}_{ridge})^2 | X = x]$?

$$\begin{aligned} & \mathbb{E}_{Y|X, \mathcal{D}}[(Y - x^T \hat{w}_{ridge})^2 | X = x] \\ &= \mathbb{E}_{Y|X}[(Y - \mathbb{E}_{Y|X}[Y|X = x])^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{Y|X}[Y|X = x] - x^T \hat{w}_{ridge})^2] \\ &= \mathbb{E}_{Y|X}[(Y - x^T w)^2 | X = x] + \mathbb{E}_{\mathcal{D}}[(x^T w - x^T \hat{w}_{ridge})^2] \\ &= \sigma^2 + (x^T w - \mathbb{E}_{\mathcal{D}}[x^T \hat{w}_{ridge}])^2 + \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[x^T \hat{w}_{ridge}] - x^T \hat{w}_{ridge})^2] \\ &= \sigma^2 + \frac{\lambda^2}{(n + \lambda)^2} (w^T x)^2 + \frac{\sigma^2 n}{(n + \lambda)^2} \|x\|_2^2 \end{aligned}$$

(verify at home)

Irreduc. Error

Bias-squared

Variance

Bias-Variance Properties

$$\hat{w}_{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y}$$

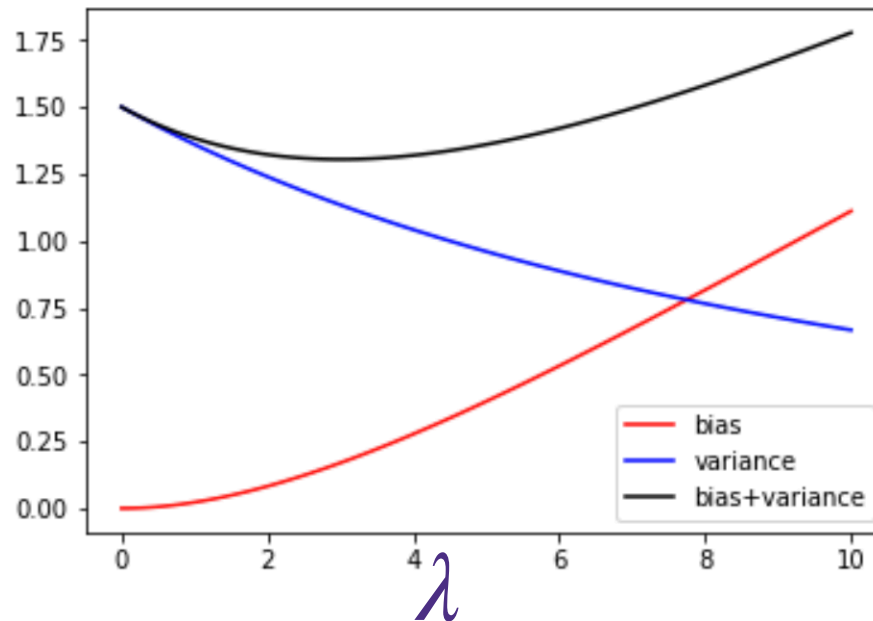
- Assume: $\mathbf{X}^T \mathbf{X} = nI$ and $\mathbf{y} = \mathbf{X}w + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

If $x \in \mathbb{R}^d$ and $Y \sim \mathcal{N}(x^T w, \sigma^2)$, what is $\mathbb{E}_{Y|x, \text{train}}[(Y - x^T \hat{w}_{ridge})^2 | X = x]$?

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

$$= \underbrace{\sigma^2}_{\text{Irreduc. Error}} + \underbrace{\frac{\lambda^2}{(n + \lambda)^2} (w^T x)^2}_{\text{Bias-squared}} + \underbrace{\frac{\sigma^2 n}{(n + \lambda)^2} \|x\|_2^2}_{\text{Variance}}$$

(verify at home)



$$d=10, n=20, \sigma^2 = 3.0, \|w\|_2^2 = 10$$

Ridge Regression: Effect of Regularization

$$\hat{w}_{ridge} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2 + \lambda \|w\|_2^2$$

- Solution is indexed by the regularization parameter λ
- Larger λ
- Smaller λ
- As $\lambda \rightarrow 0$, $\hat{w}_{ridge} \rightarrow$
- As $\lambda \rightarrow \infty$, $\hat{w}_{ridge} \rightarrow$

Ridge Regression: Effect of Regularization

$$\mathcal{D} \stackrel{i.i.d.}{\sim} P_{XY}$$

$$\hat{w}_{\mathcal{D},ridge}^{(\lambda)} = \arg \min_w \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - x_i^T w)^2 + \lambda \|w\|_2^2$$

TRAIN error:

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - x_i^T \hat{w}_{\mathcal{D},ridge}^{(\lambda)})^2$$

TRUE error:

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

TEST error:

$$\mathcal{T} \stackrel{i.i.d.}{\sim} P_{XY}$$

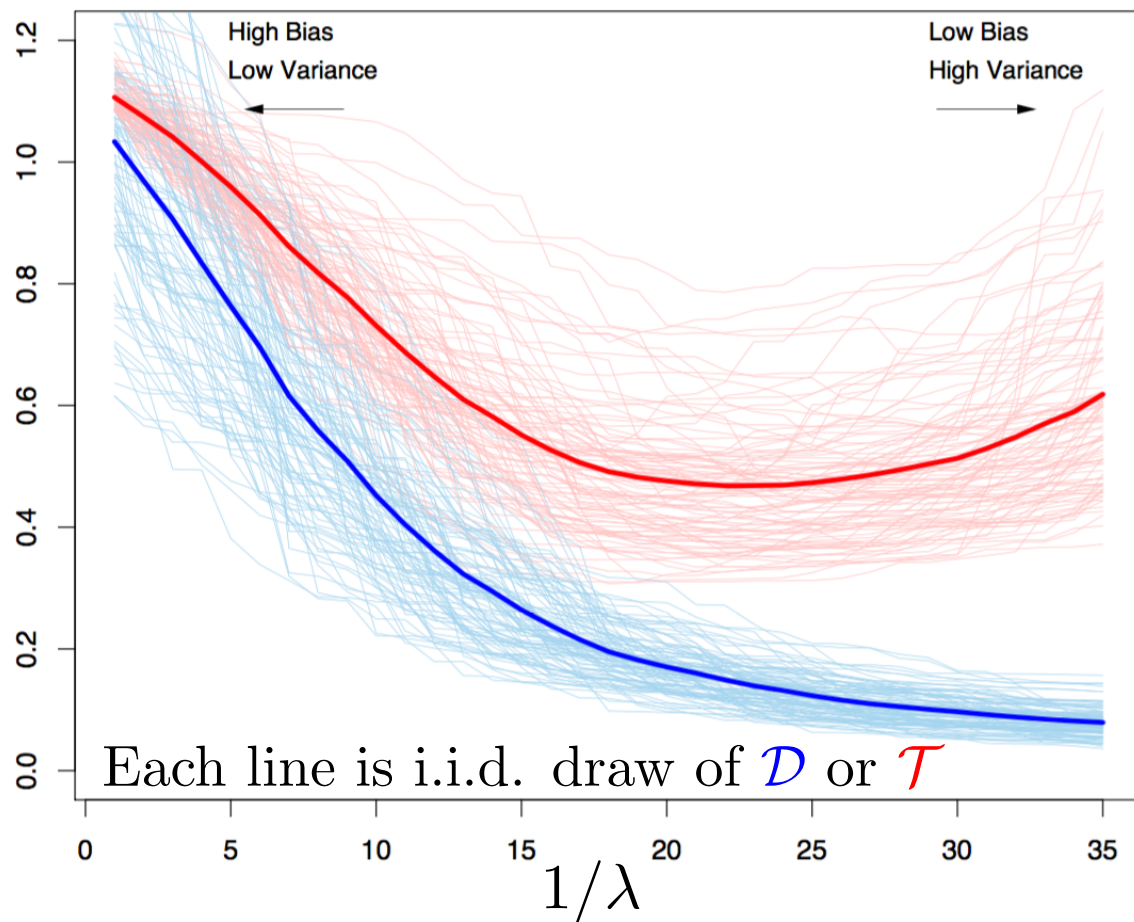
$$\frac{1}{|\mathcal{T}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - x_i^T \hat{w}_{\mathcal{D},ridge}^{(\lambda)})^2$$

Important: $\mathcal{D} \cap \mathcal{T} = \emptyset$

Ridge Regression: Effect of Regularization

$$\mathcal{D} \stackrel{i.i.d.}{\sim} P_{XY}$$

$$\hat{w}_{\mathcal{D},ridge}^{(\lambda)} = \arg \min_w \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - x_i^T w)^2 + \lambda \|w\|_2^2$$



TRAIN error:

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - x_i^T \hat{w}_{\mathcal{D},ridge}^{(\lambda)})^2$$

TRUE error:

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

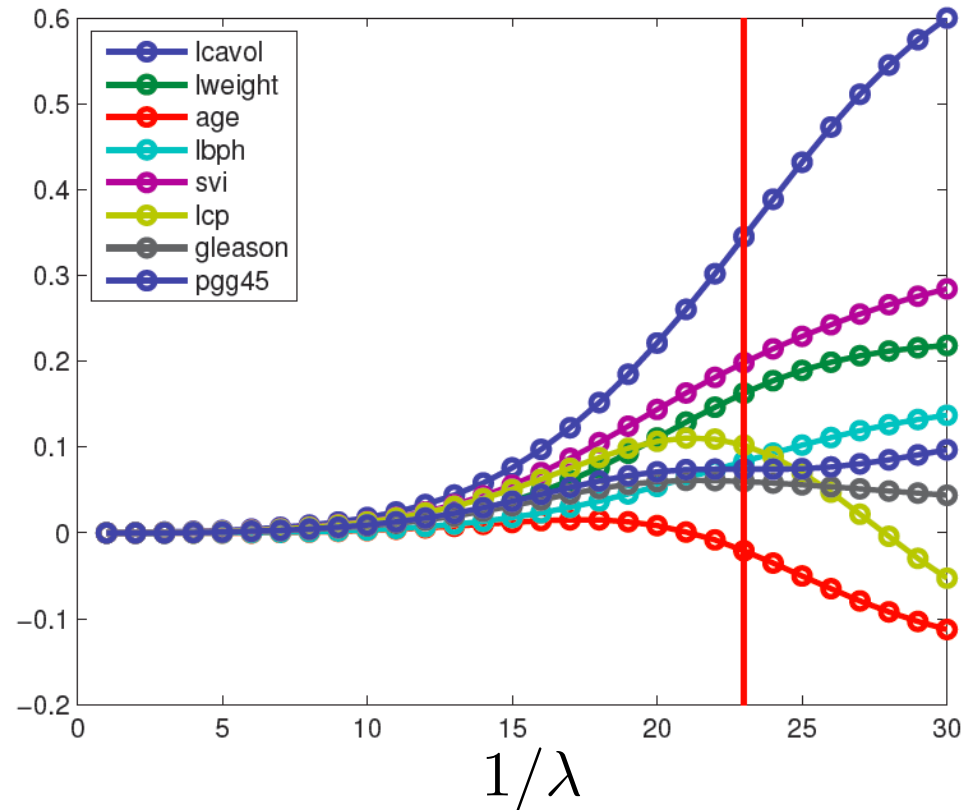
TEST error:

$$\mathcal{T} \stackrel{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{T}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - x_i^T \hat{w}_{\mathcal{D},ridge}^{(\lambda)})^2$$

Important: $\mathcal{D} \cap \mathcal{T} = \emptyset$

Ridge regression: minimize $\sum_{i=1}^n (w^T x_i + b - y_i)^2 + \lambda \|w\|_2^2$

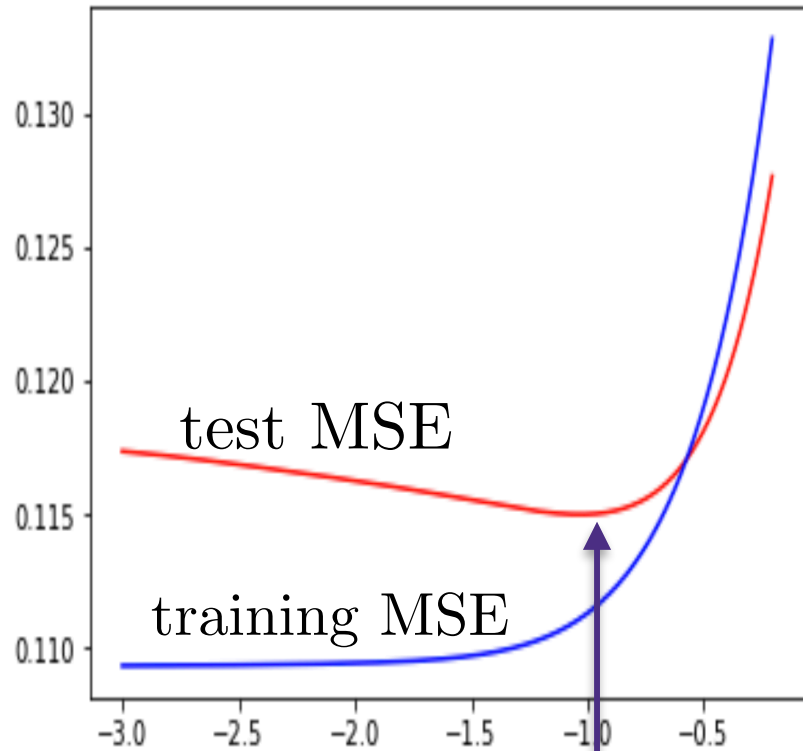


From
Kevin Murphy
textbook

> Typical approach: select λ using cross validation, up next

Ridge regression: minimize

$$\sum_{i=1}^n (w^T x_i + b - y_i)^2 + \lambda \|w\|_2^2$$

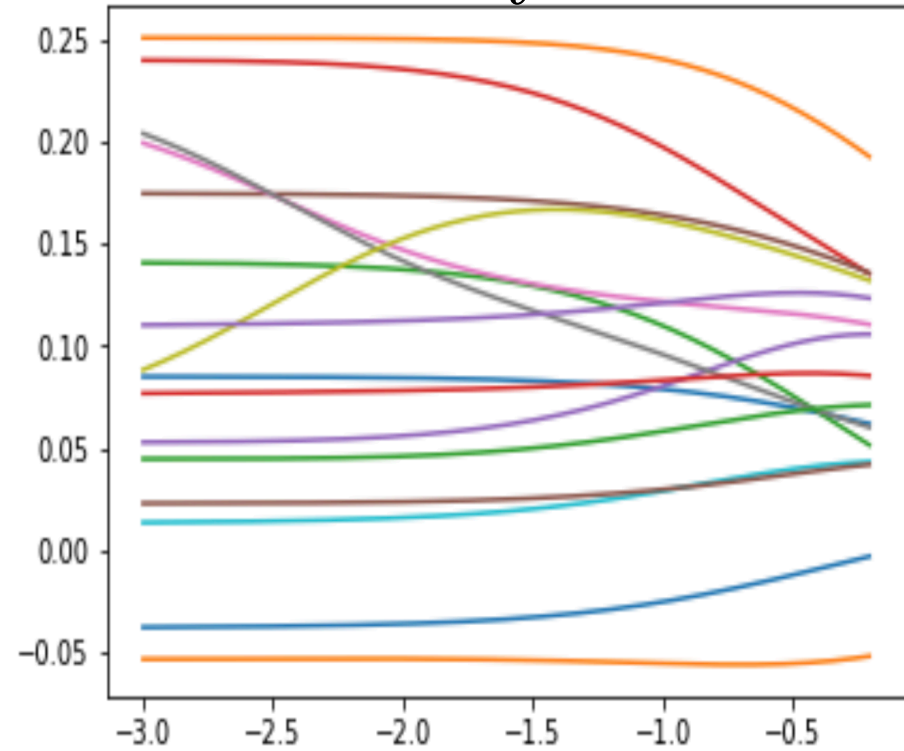


High
model
complexity

$\log_{10}(\lambda)$

Low
model
complexity

w_i 's



$\log_{10}(\lambda)$

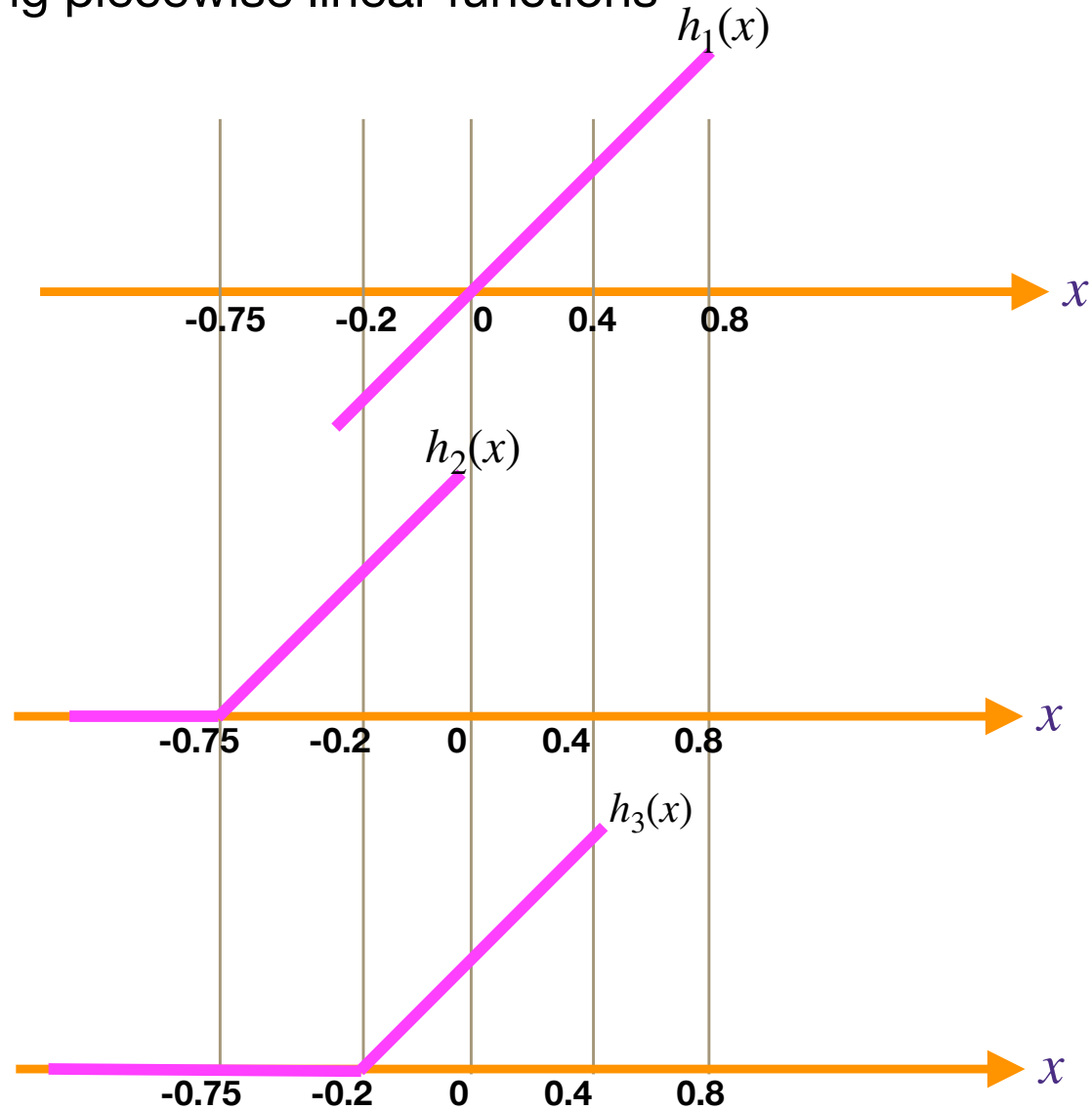
- this gain in test MSE comes from shrinking w 's to get a less sensitive predictor (which in turn reduces the variance)

Example: piecewise linear fit

- we fit a linear model for $x \in [-1, 1]$:
$$f(x) = b + w_1 h_1(x) + w_2 h_2(x) + w_3 h_3(x) + w_4 h_4(x) + w_5 h_5(x)$$
- with a specific choice of features using piecewise linear functions

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ h_3(x) \\ h_4(x) \\ h_5(x) \end{bmatrix} = \begin{bmatrix} x \\ [x + 0.75]^+ \\ [x + 0.2]^+ \\ [x - 0.4]^+ \\ [x - 0.8]^+ \end{bmatrix}$$

$$[a]^+ \triangleq \max\{a, 0\}$$

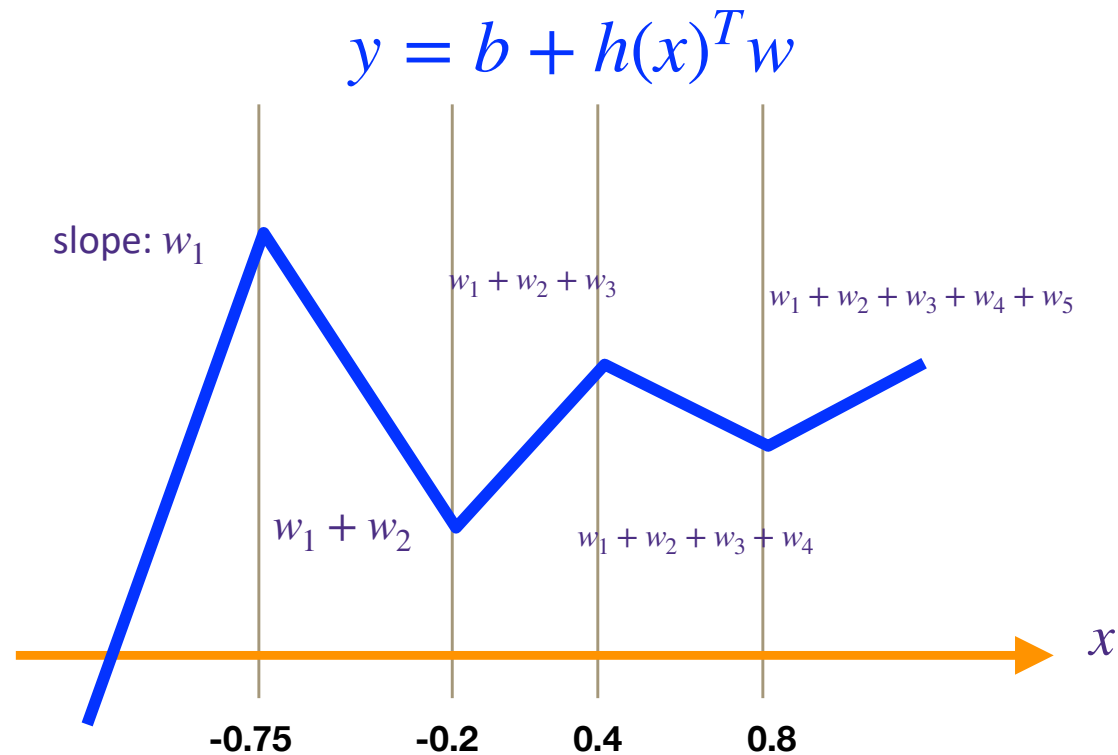


Example: piecewise linear fit

- we fit a linear model:
 $f(x) = b + w_1 h_1(x) + w_2 h_2(x) + w_3 h_3(x) + w_4 h_4(x) + w_5 h_5(x)$
- with a specific choice of features using piecewise linear functions

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ h_3(x) \\ h_4(x) \\ h_5(x) \end{bmatrix} = \begin{bmatrix} x \\ [x + 0.75]^+ \\ [x + 0.2]^+ \\ [x - 0.4]^+ \\ [x - 0.8]^+ \end{bmatrix}$$

$$[a]^+ \triangleq \max\{a, 0\}$$



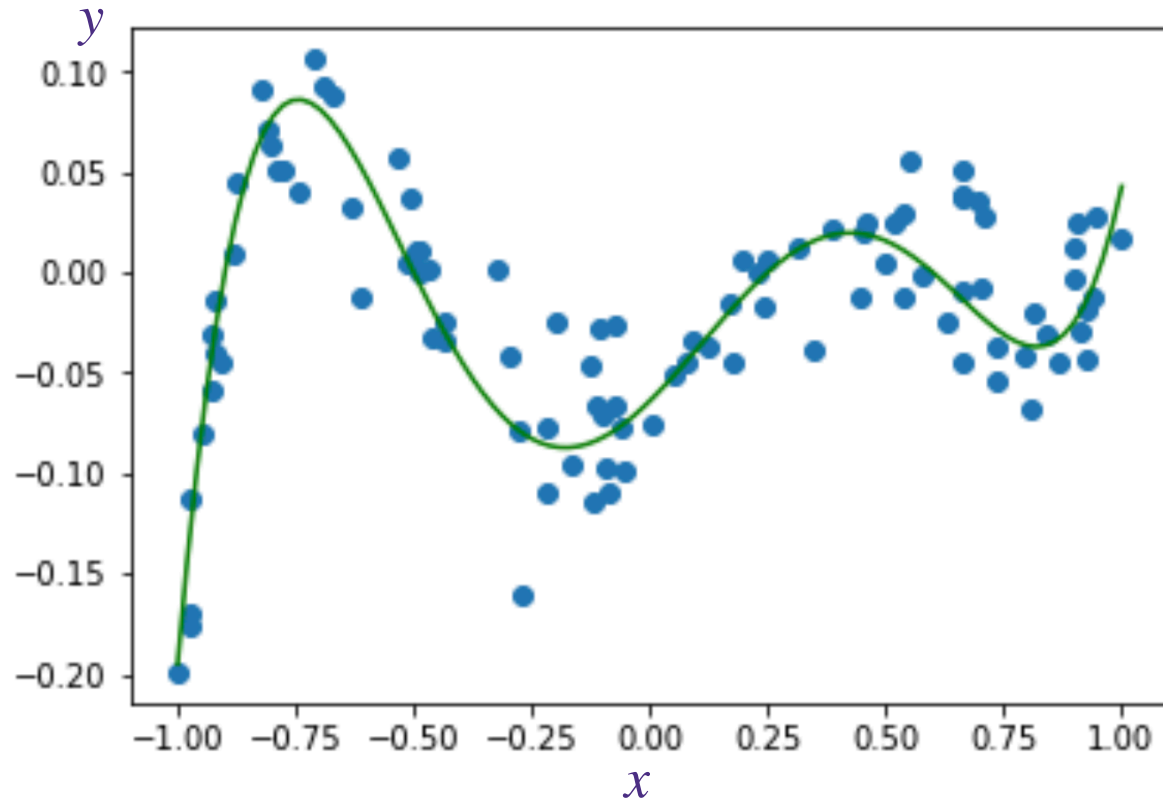
the weights capture the change in the slopes

Example: piecewise linear fit

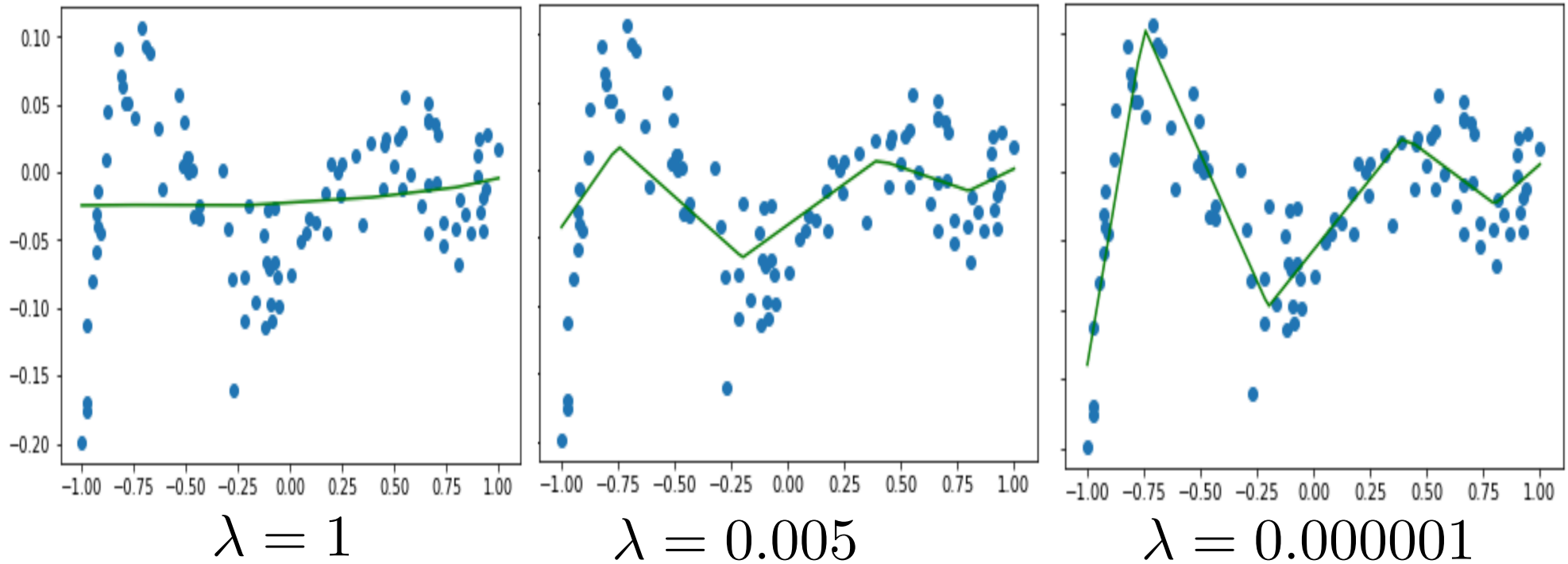
- we fit a linear model:

$$f(x) = b + w_1h_1(x) + w_2h_2(x) + w_3h_3(x) + w_4h_4(x) + w_5h_5(x)$$

- with a specific choice of features using piecewise linear functions

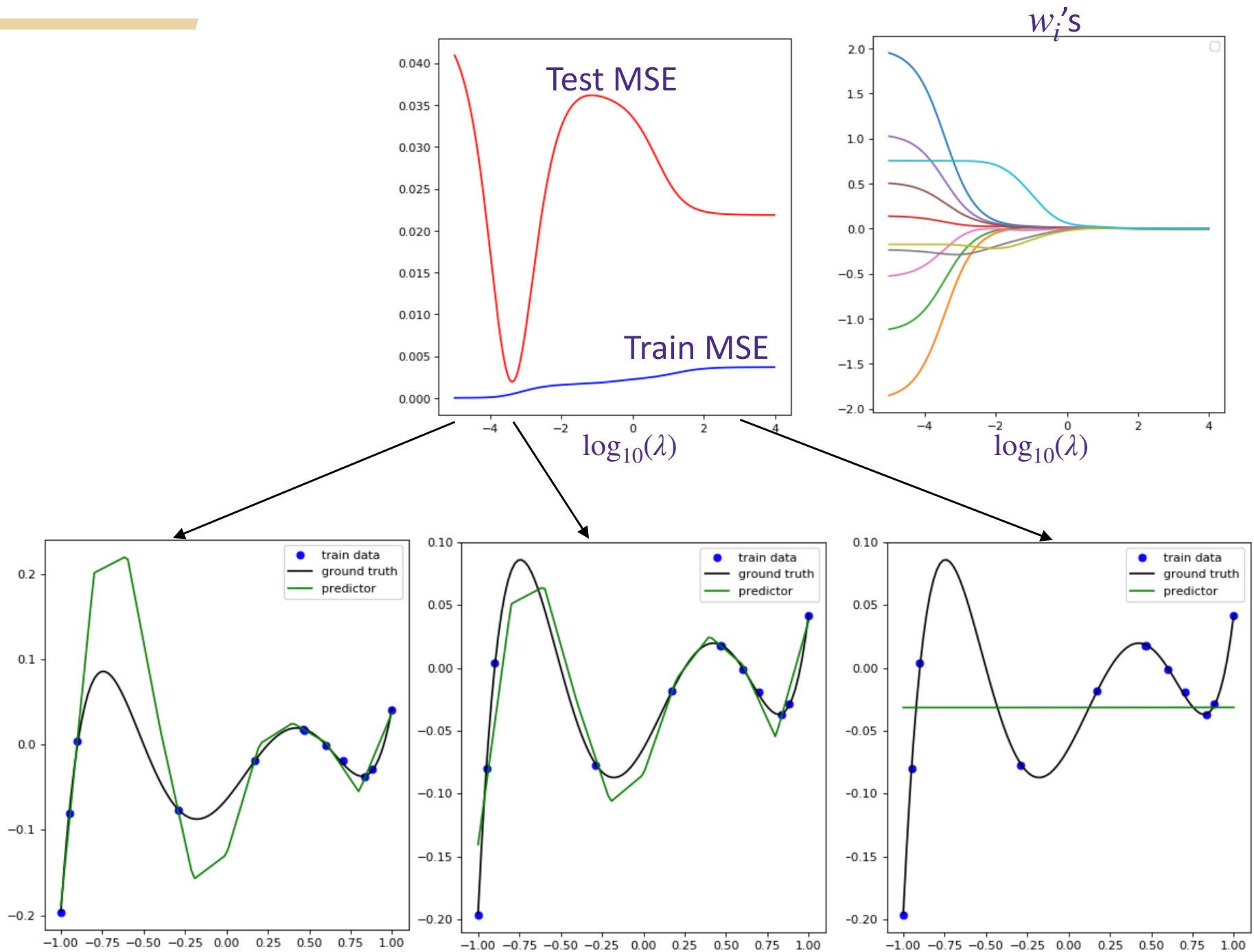


Example: piecewise linear fit (ridge regression)



We do not observe overfitting, as $d=5$ and $n=100$

Can avoid overfitting even $w \in \mathbb{R}^{10}$ and $n=11$ samples



What you need to know...

> Regularization

- Penalizes complex models towards preferred, simpler models

> Ridge regression

- L_2 penalized least-squares regression
- Regularization parameter trades off model complexity with training error
- Never regularize the offset!