

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

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

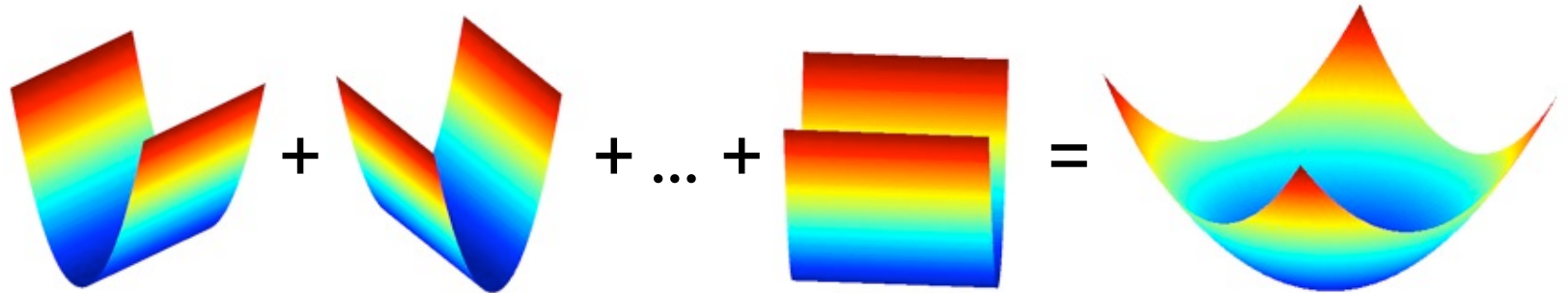
In general:  $= \arg \min_w w^T (\mathbf{X}^T \mathbf{X}) w - 2y^T \mathbf{X}w$

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



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

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”



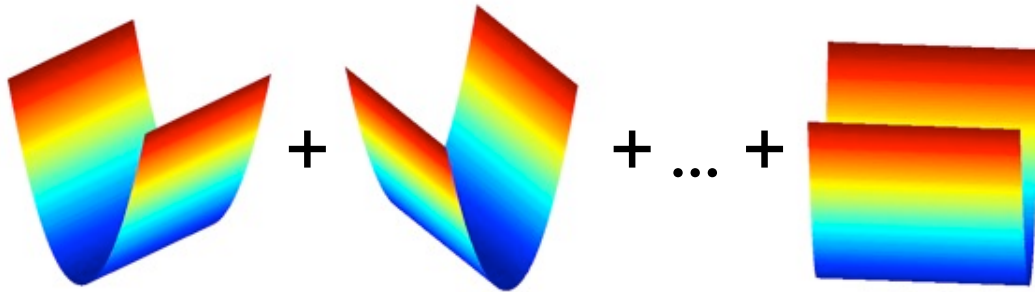
**Regularization imposes “simpler” solutions by a “complexity” penalty**

# Ridge Regression

---

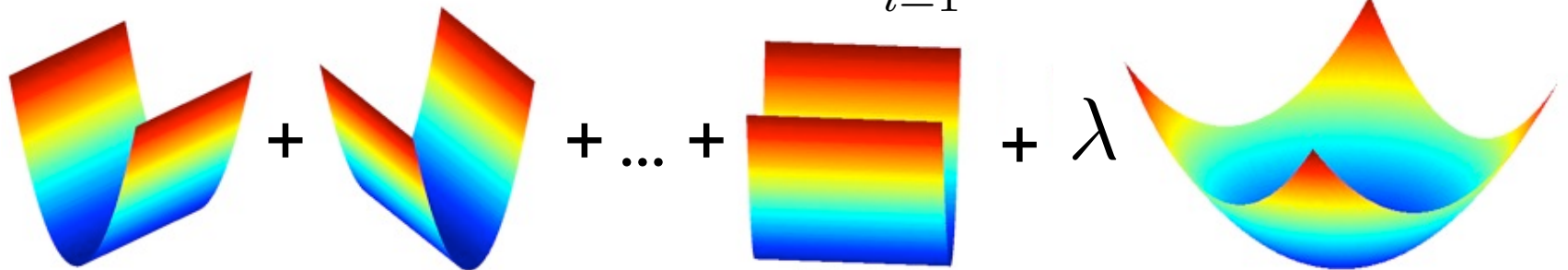
- Old Least squares objective:

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



- 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

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

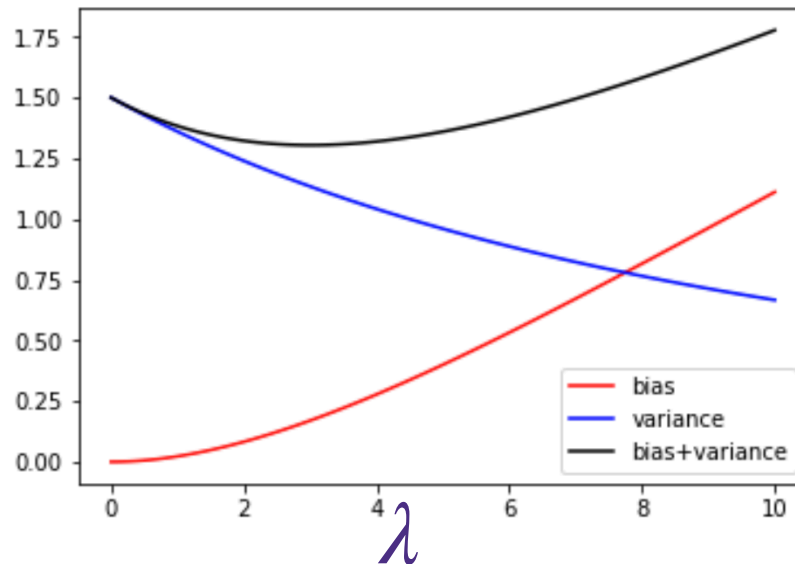
- Assume:  $\mathbf{X}^T \mathbf{X} = nI$  and  $\mathbf{y} = \mathbf{X}w + \epsilon$   $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

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  $\lambda$
- Larger  $\lambda$
- Smaller  $\lambda$
- 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:**

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

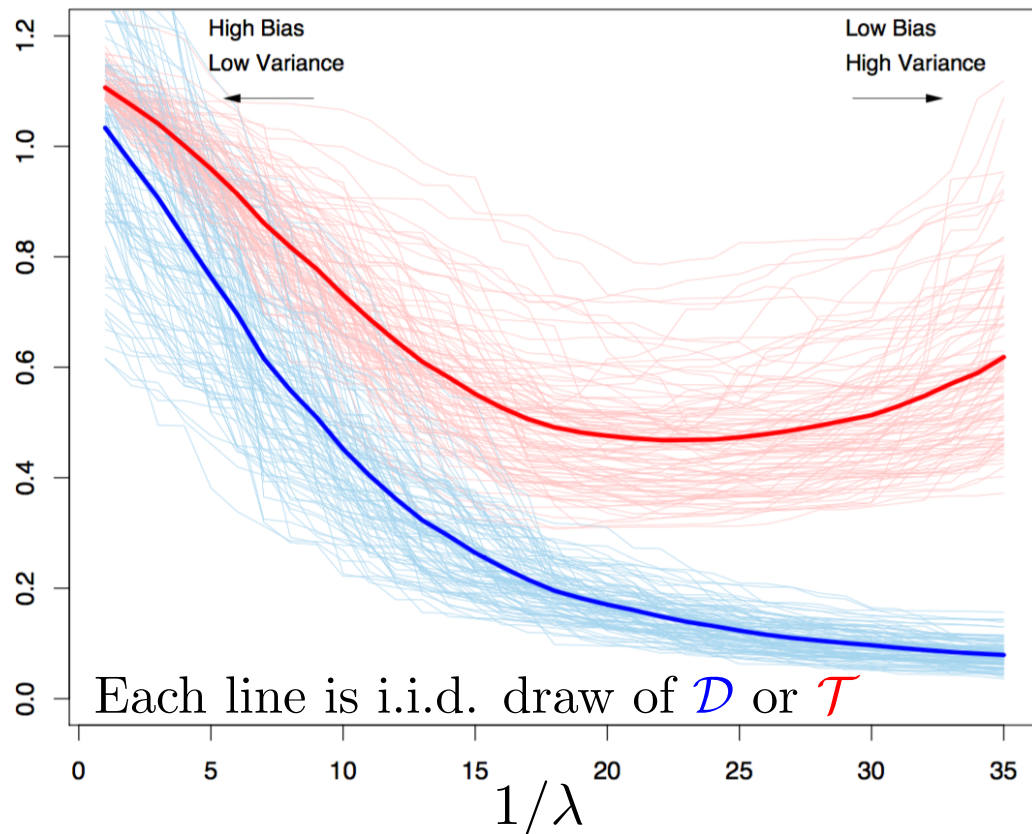
$$\frac{1}{|\mathcal{T}|} \sum_{(x_i, y_i) \in \mathcal{T}} (y_i - x_i^T \hat{w}_{\mathcal{D},ridge}^{(\lambda)})^2$$

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:**

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

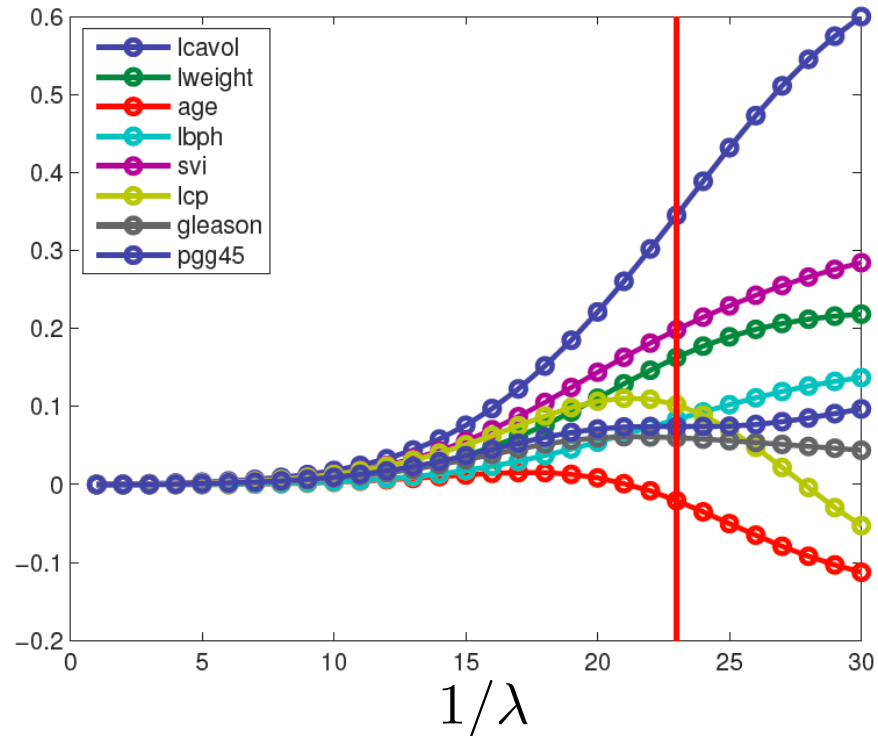
**TEST error:**

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

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

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

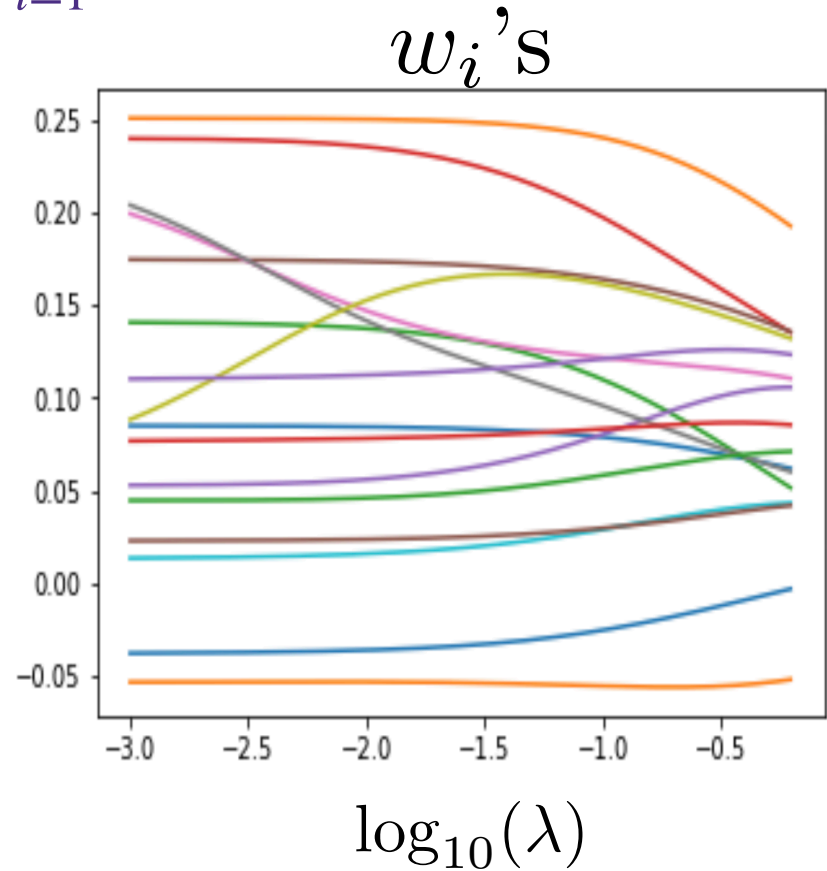
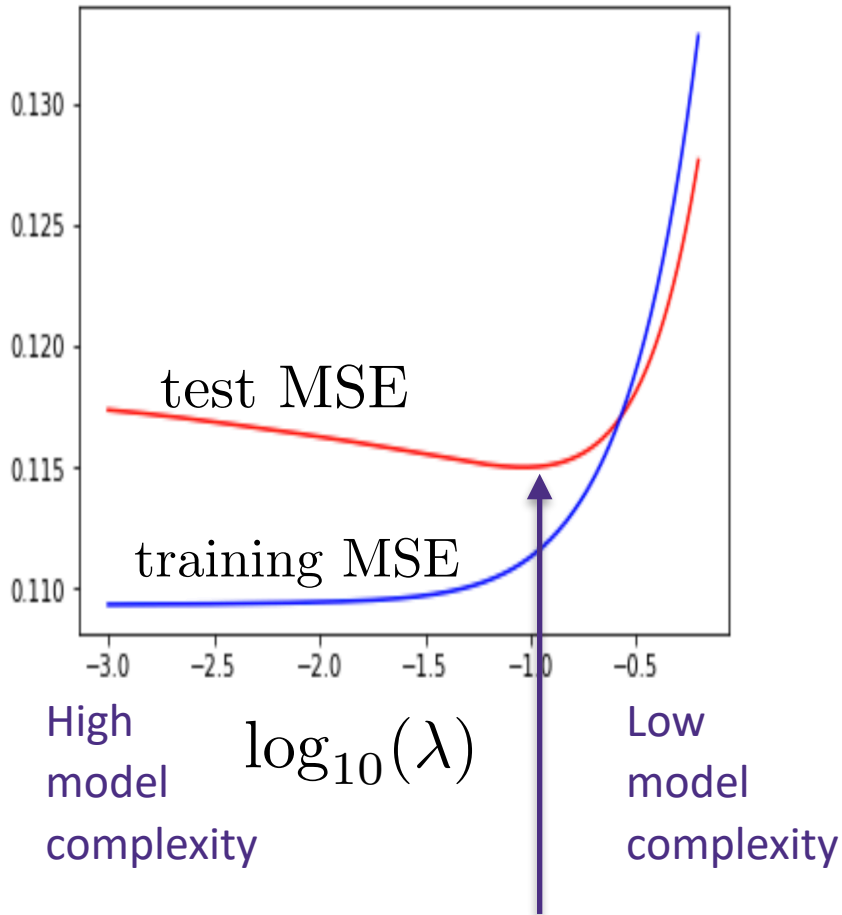
**Ridge regression:** minimize  $\sum_{i=1}^n (w^T x_i + b - y_i)^2 + \lambda \|w\|_2^2$



From  
Kevin Murphy  
textbook

> Typical approach: select  $\lambda$  using cross validation, up next

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



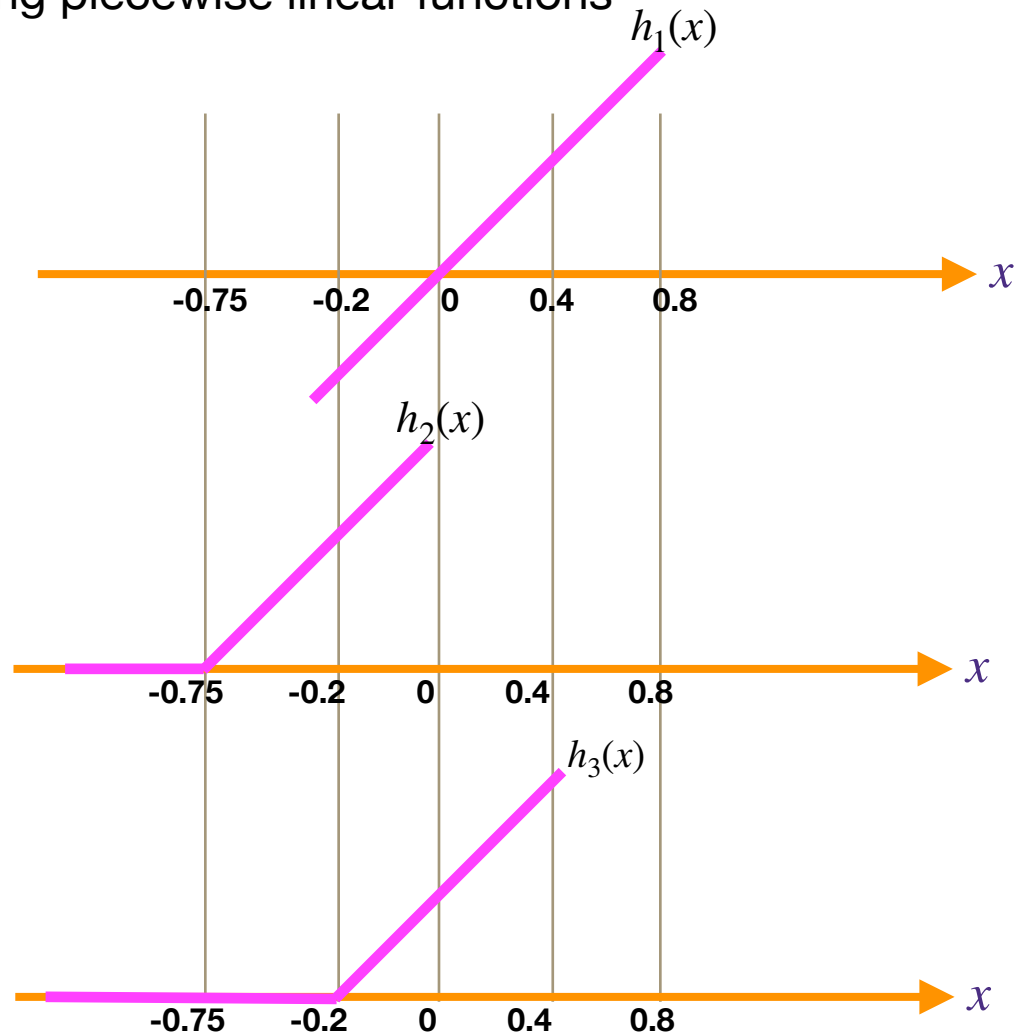
- this gain in test MSE comes from shrinking  $w$ 's to get a less sensitive predictor (which in turn reduces the variance)

# 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

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

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

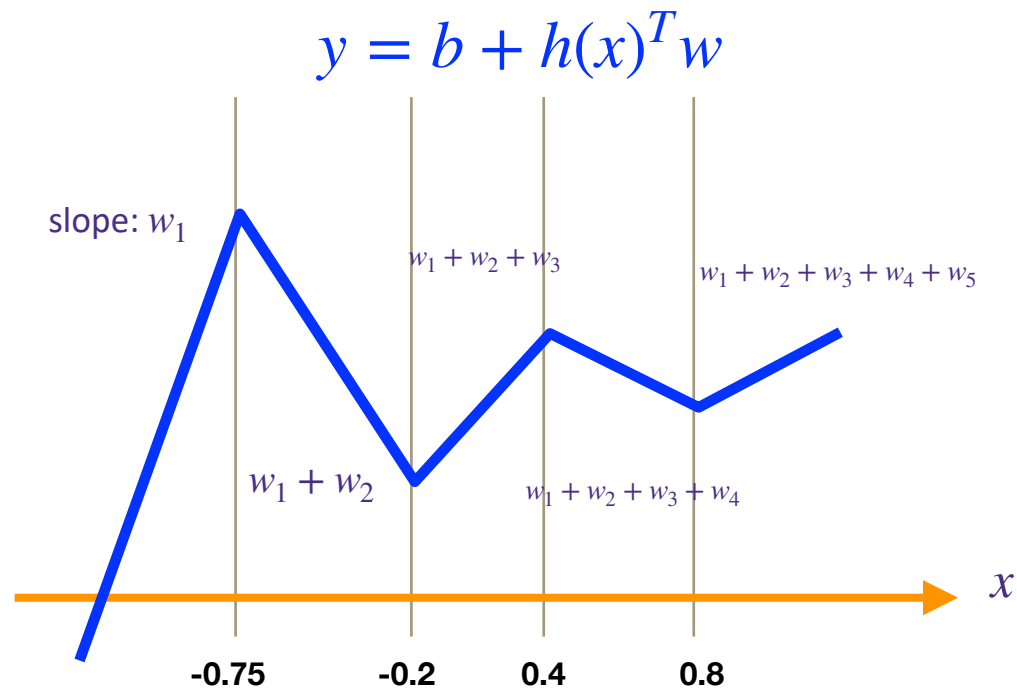


# Example: piecewise linear fit

- 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

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

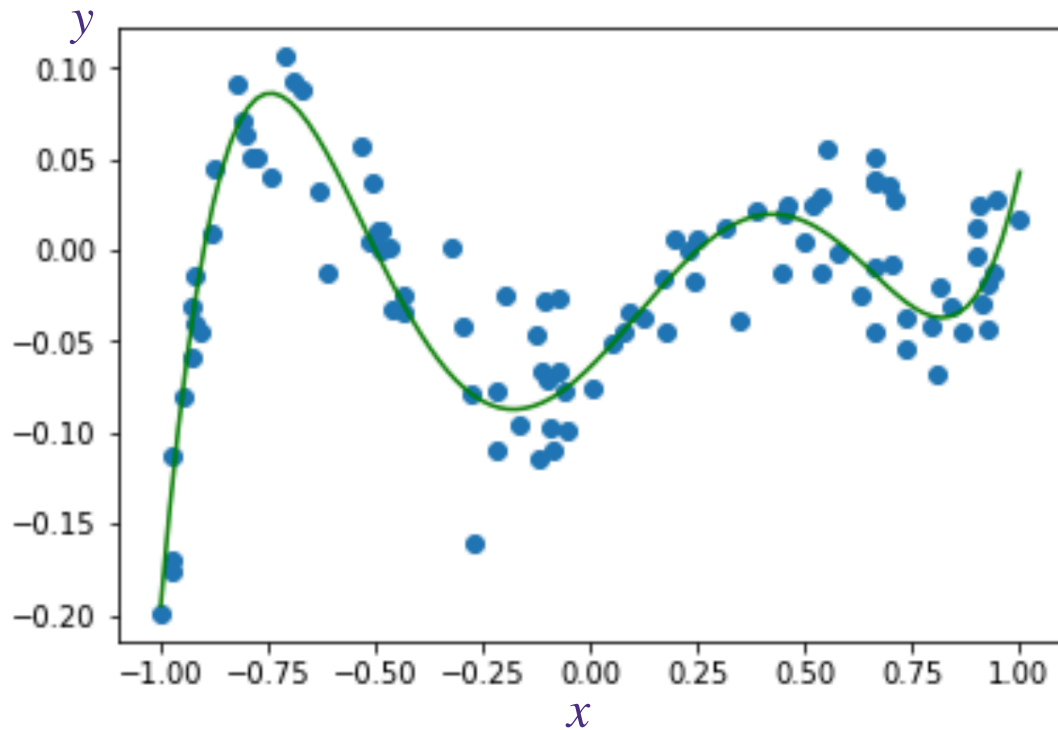
$$[a]^+ \triangleq \max\{a, 0\}$$



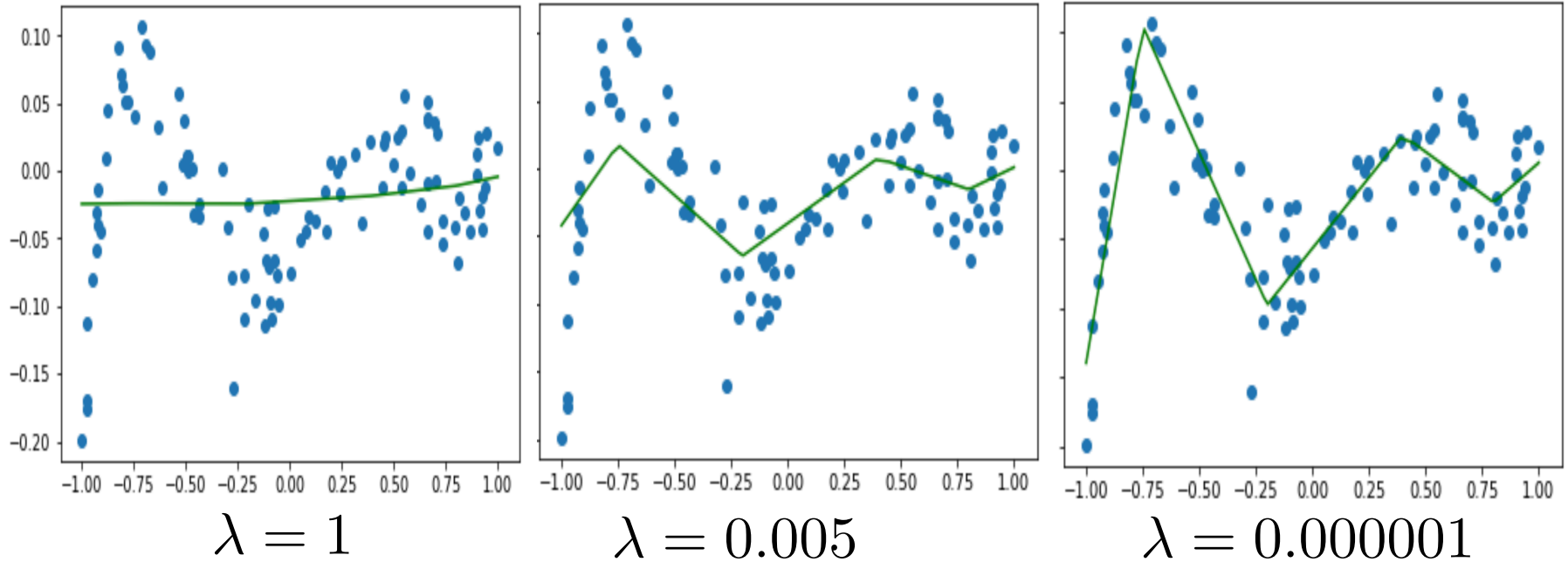
**the weights capture the change in the slopes**

# 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

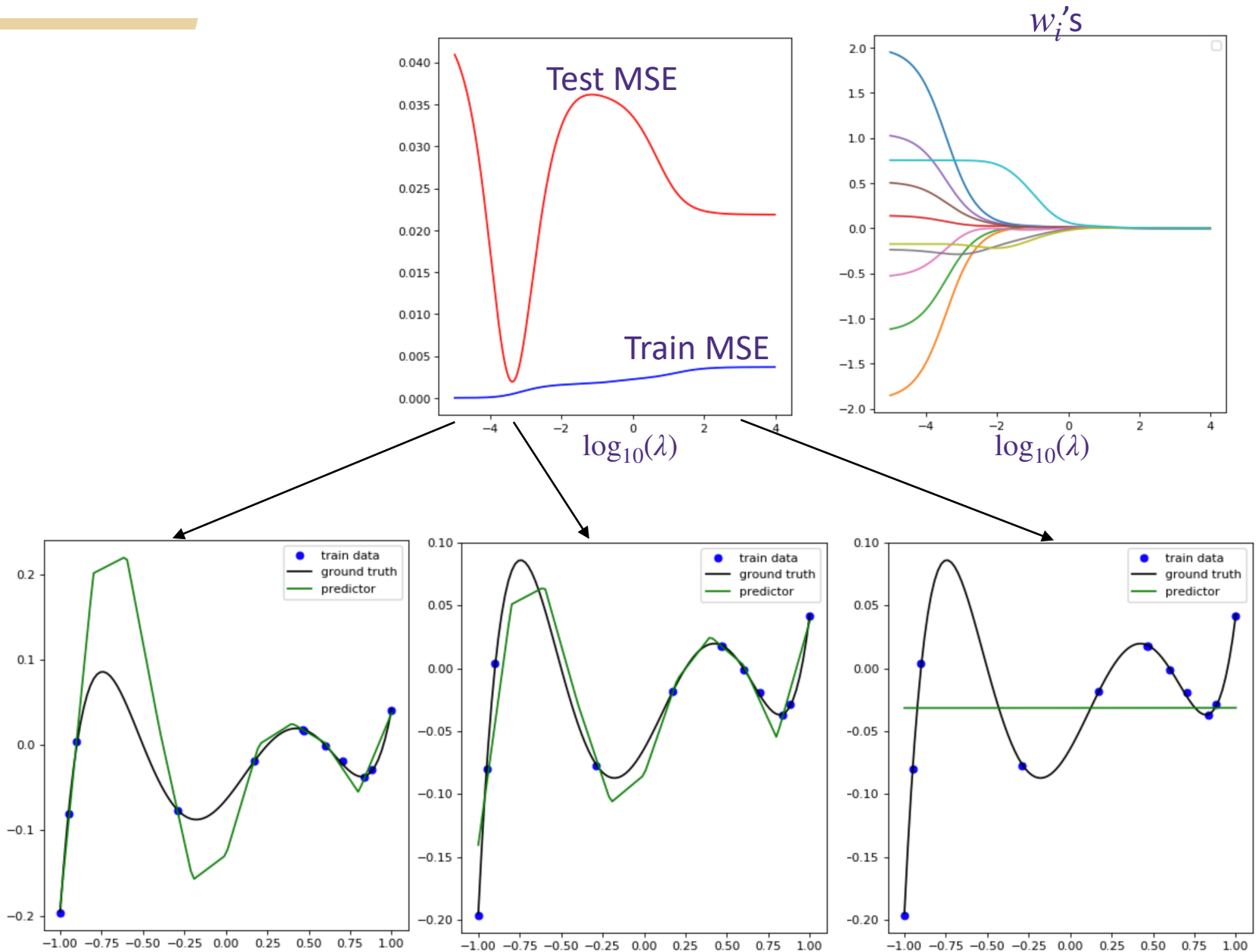


# Example: piecewise linear fit (ridge regression)



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!