- Recorded on Jan 17th Monday
- Replaces lecture on Wednesday 19th

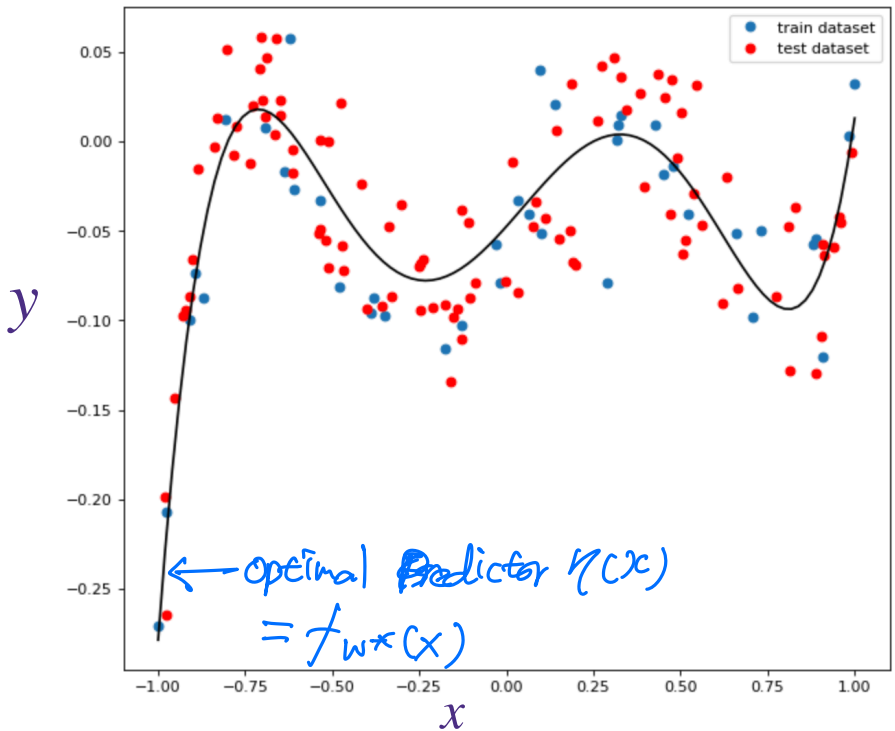
# Lecture 7: Regularization

- How to overcome overfitting.

# Recap: bias-variance tradeoff

- Consider 40 training examples and 100 test examples i.i.d.drawn from degree-5 polynomial features
  $x_i \sim \text{Uniform}[-1,1]$, $y_i \sim f_{w*}(x_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$f_{w*}(x_i) = b^* + w_1^* x_i + w_2^* (x_i)^2 + w_3^* (x_i)^3 + w_4^* (x_i)^4 + w_5^* (x_i)^5$$




This is a linear model with features

$$h(x_i) = (x_i, (x_i)^2, (x_i)^3, (x_i)^4, (x_i)^5)$$

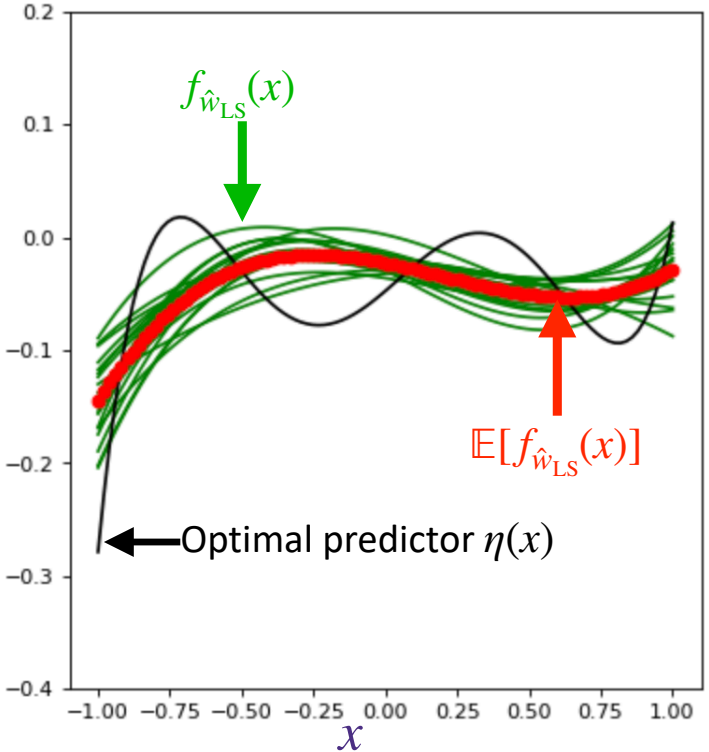
$$f_w(x_i) = h(x_i)^T w + b$$

$$\widehat{w}_{\text{LS}} = \arg \min_{b\in\mathbb{R}, w\in\mathbb{R}^5} \sum_{i=1}^{N} \left( y_i - (h_w(x_i) + b) \right)^2$$

# Recap: bias-variance tradeoff

With degree-3 polynomials, we underfit

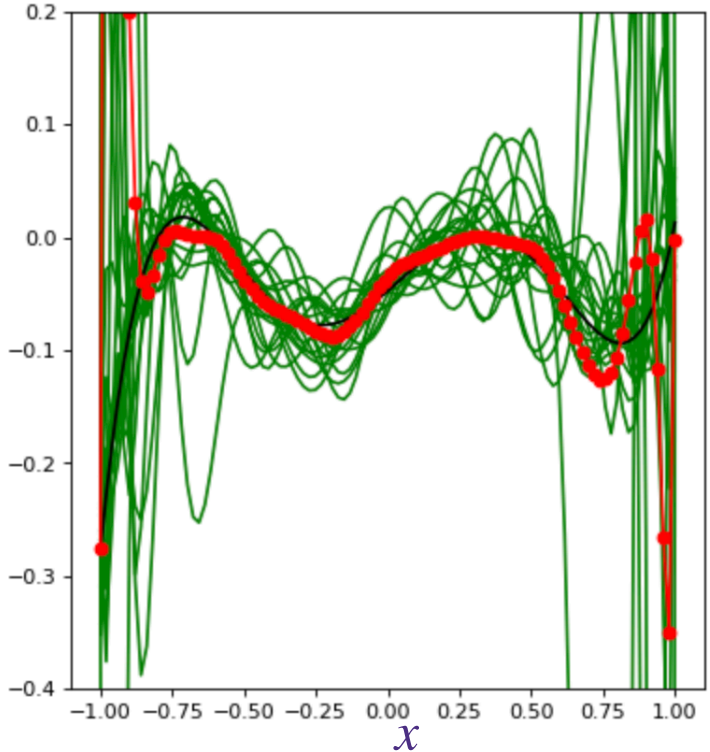
$\hat{f}_{\hat{w}_{\mathrm{LS}}}(x)$




current train error = 0.0036791644380554187
current test error = 0.0037962529988410953

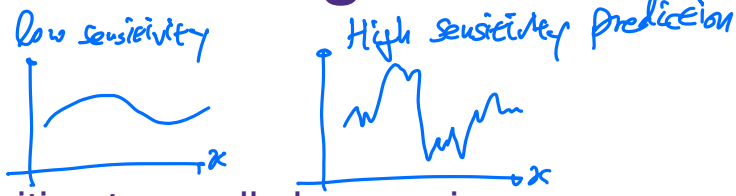
With degree-20 polynomials, we overfit

$\hat{f}_{\hat{w}_{\mathrm{LS}}}(x)$




0.0005421686349568773
0.14210029429557927

# Sensitivity: how to detect overfitting




- For a linear model,
  $y \simeq b + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$
  if $|w_j|$ is large then the prediction is sensitive to small changes in $x_j$
- Large sensitivity leads to overfitting and poor generalization, and equivalently models that overfit tend to have large weights
- Note that $b$ is a constant and hence there is no sensitivity for the offset $b$
- In **Ridge Regression,** we use a regularizer $\|w\|_2^2$ to measure and control the sensitivity of the predictor
  $= w_1^2 + w_2^2 + \cdots + w_d^2$
- And optimize for small loss and small sensitivity, by adding a **regularizer** in the objective (assume no offset for now) with **regularization coefficient** $\lambda > 0$

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

- The regularization encourages solution $w$ with smaller norm $\|w\|_2^2$, hence encouraging less overfitting
- Larger $\lambda$ means more regularization
- The first term encourages fitting the training data

# Minimizing the Ridge Regression Objective

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

$$L(w) = \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

$$\nabla_w L(w) = -2X^T(y - Xw) + 2\lambda \cdot \underbrace{I \cdot w}_{=w}$$

$$= -2\left(X^T y - X^T X w - I\lambda \cdot w\right)$$

$$= -2\left(X^T y - (X^T X + \lambda \cdot I) \cdot w\right)\Big|_{w=\hat{w}_{ridge}} = 0$$

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

# Shrinkage Properties

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

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

, $\hat{w}_{LS} = (X^TX)^{-1} \cdot X^Ty$

For example, if $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$, then

$\hat{w}_{ridge}^{(\lambda)} = \frac{1}{n+\lambda} \cdot X^Ty$.

$\hat{w}_{LS} = \frac{1}{n} X^Ty$

$\hat{w}_{ridge}^{(\lambda)} = \frac{n}{n+\lambda} \cdot \hat{w}_{LS}$

shrinks $\hat{w}_{LS}$.

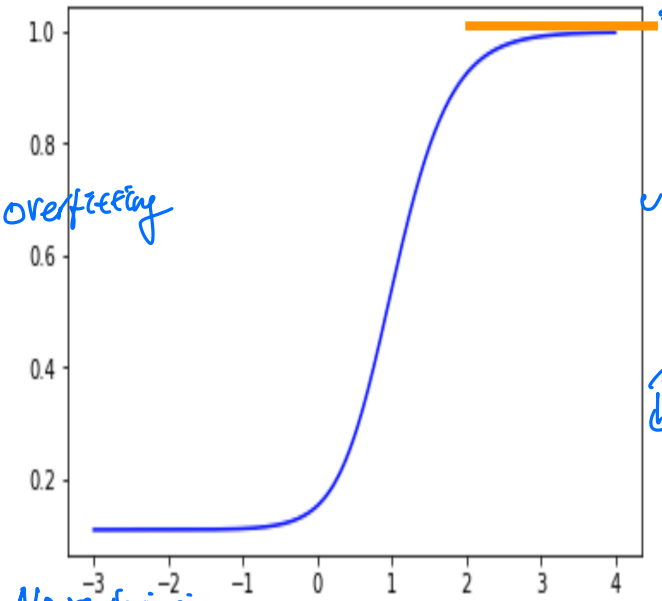
Larger $\lambda$ shrinks more.

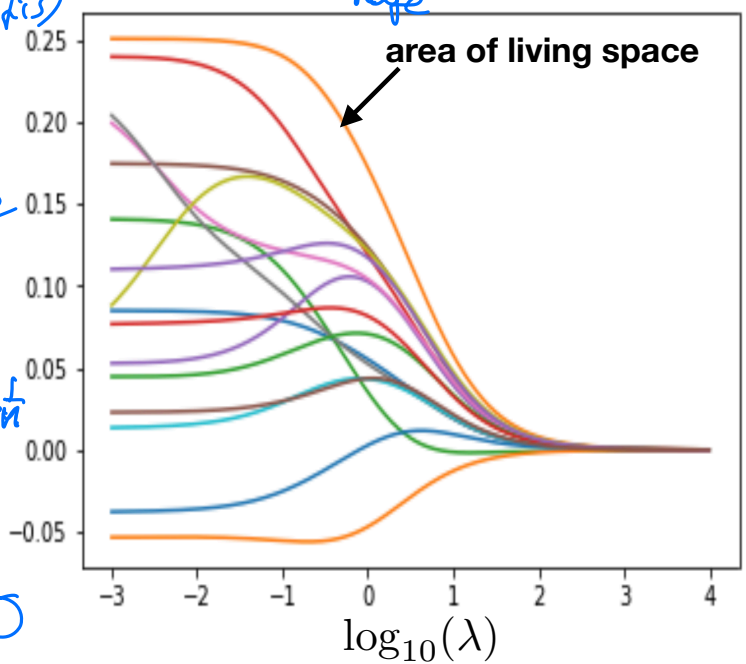
Similar shrinking effect for general $\mathbf{X}^T\mathbf{X}$, which we do not go into details in class (come to my OH if interested).

- When $\lambda = 0$, this gives the least squares model
- This defines a family of models hyper-parametrized by $\lambda$
- Large $\lambda$ means more regularization and simpler model
- Small $\lambda$ means less regularization and more complex model

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

training MSE $\frac{1}{n} \sum_{i=1}^{n} (y_i - (x_i^T \hat{w}_{\text{ridge}}^{(\lambda)} + \hat{b}_{\text{ridge}}))^2$



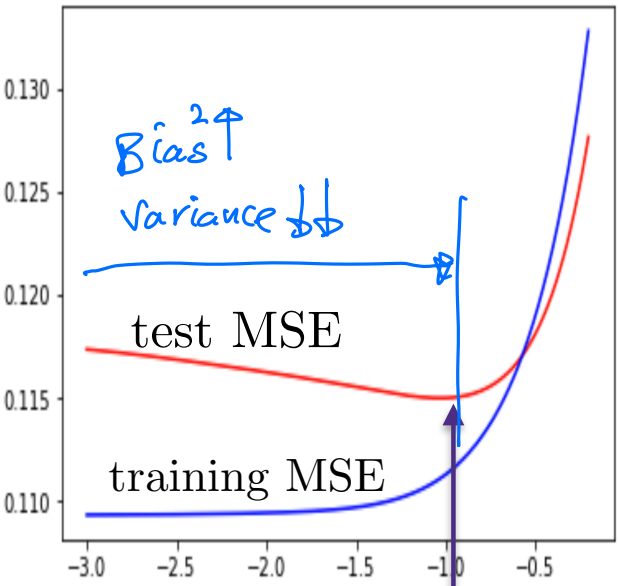



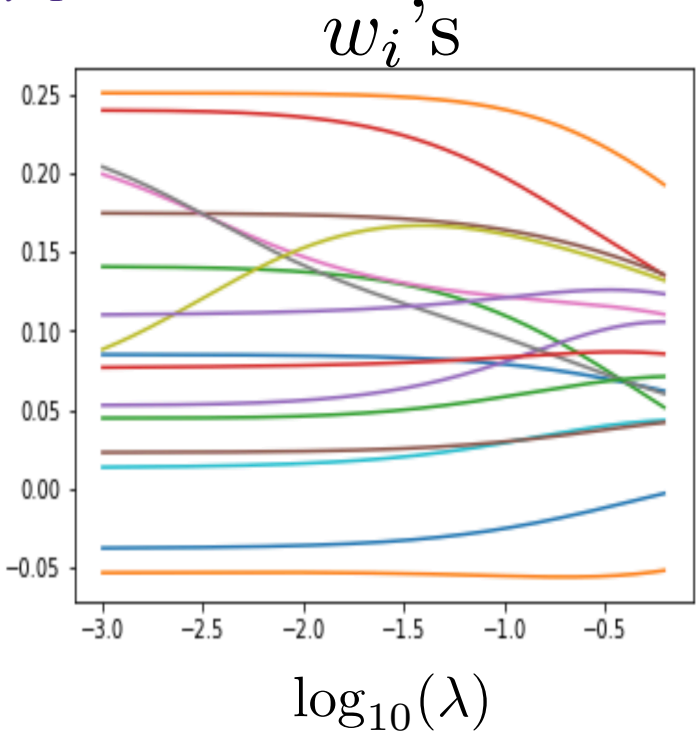

High model complexity

Low model complexity

- Left plot: leftmost training error is with no regularization: 0.1093
- Left plot: rightmost training error is variance of the training data: 0.9991
- Right plot: called **regularization path**

# 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)

# Bias-variance tradeoff for ridge regression model

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

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

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

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

$$= (X^TX+\lambda I)^{-1}(X^TX+\lambda I)w^* - \lambda I w^* + X^T\varepsilon$$

$$= w^* - (X^TX+\lambda I)^{-1}\cdot\lambda\cdot w^* + (X^TX+\lambda I)^{-1}X^T\varepsilon$$

→ Bias (λ↑ ↑) → Variance ↓

For example, if $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$, then

$$\hat{w}_{\text{ridge}} = w^* - \frac{\lambda}{n+\lambda}w^* + \frac{1}{n+\lambda}X^T\varepsilon$$

# Bias-variance tradeoff for ridge regression model

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

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

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

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

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

For example, if $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$, then

$$\hat{w}_{\text{ridge}} = \underbrace{w^* - \frac{\lambda}{n+\lambda}w^*}_{\text{estimate is shrunk by regularizer}} + \underbrace{\frac{1}{n+\lambda}\mathbf{X}^T\epsilon}_{\text{error due to noise}}$$

$= \left(1 - \frac{\lambda}{n+\lambda}\right)w^*$

→ larger $\lambda$ increases bias

→ larger $\lambda$ decreases variance

# Bias-variance tradeoff for ridge regression model

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

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

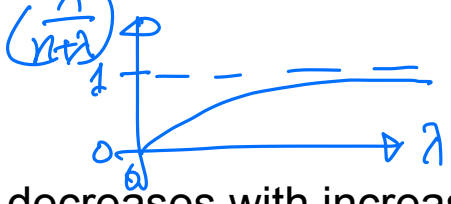
$$\eta(x) = \mathbb{E}_{Y|X}[Y \,|\, X = x] = x^T w^*$$

For example, if $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$, then

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

$\mathbb{E}$ → $x^T w^* - \frac{\lambda}{n+\lambda} x^T w^*$ ($\mathbb{E}(\epsilon) = 0$)

- Irreducible error: $\mathbb{E}_{X,Y}[(Y - \eta(x))^2 \,|\, X = x] = \sigma^2$
- Bias squared: $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2 = \left(x^T w^* - \right)^2 = \frac{\lambda^2}{(n+\lambda)^2} \cdot (x^T w^*)^2$




$$= \frac{\lambda^2}{(n+\lambda)^2} (x^T w^*)^2$$

1. Bias decreases with increasing sample size
2. Bias increases with increasing $\lambda$

# Bias-variance tradeoff for ridge regression model

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

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

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

For example, if $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$, then

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

$x^T w^* - \frac{1}{n+\lambda} x^T w^*$

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

$$= \frac{\sigma^2}{(n+\lambda)^2} \cdot x^T \mathbb{E}[X^T X] x$$

$$= n \cdot \mathbb{I}$$

$$= \frac{\sigma^2 n}{(n+\lambda)^2} \|x\|_2^2$$

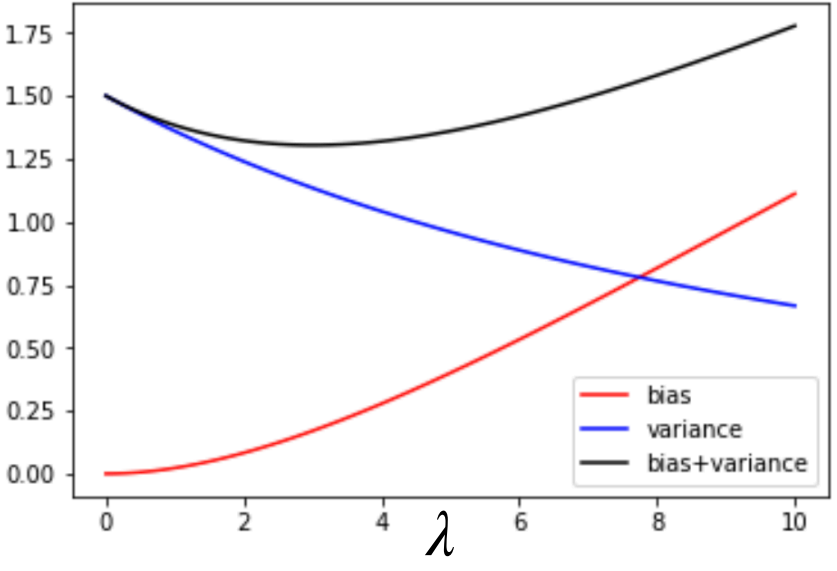
  - Variance decreases with increasing sample size
  - Variance decrease with increasing $\lambda$

# Bias-Variance Properties

- Ridge regressor: $\widehat{w}_{ridge} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$
- True error

$$\mathbb{E}_{Y|X,\mathcal{D}}[(y - x^T \hat{w}_{\text{ridge}})^2 \,|\, x] = \sigma^2 + \underbrace{\frac{\lambda^2}{(n+\lambda)^2}(w^T x)^2}_{\text{Bias-squared}} + \underbrace{\frac{\sigma^2 n}{(n+\lambda)^2}\|x\|_2^2}_{\text{Variance}}$$

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




Complex model
Overfitting

as $\lambda \to 0$,
$\hat{w}_{\text{ridge}} \to \hat{w}_{\text{LS}}$

simpler model
Underfitting

as $\lambda \to \infty$
$\hat{w}_{\text{ridge}} \to 0$

# What you need to know…

- Regularization
  - Penalizes complex models towards simpler models

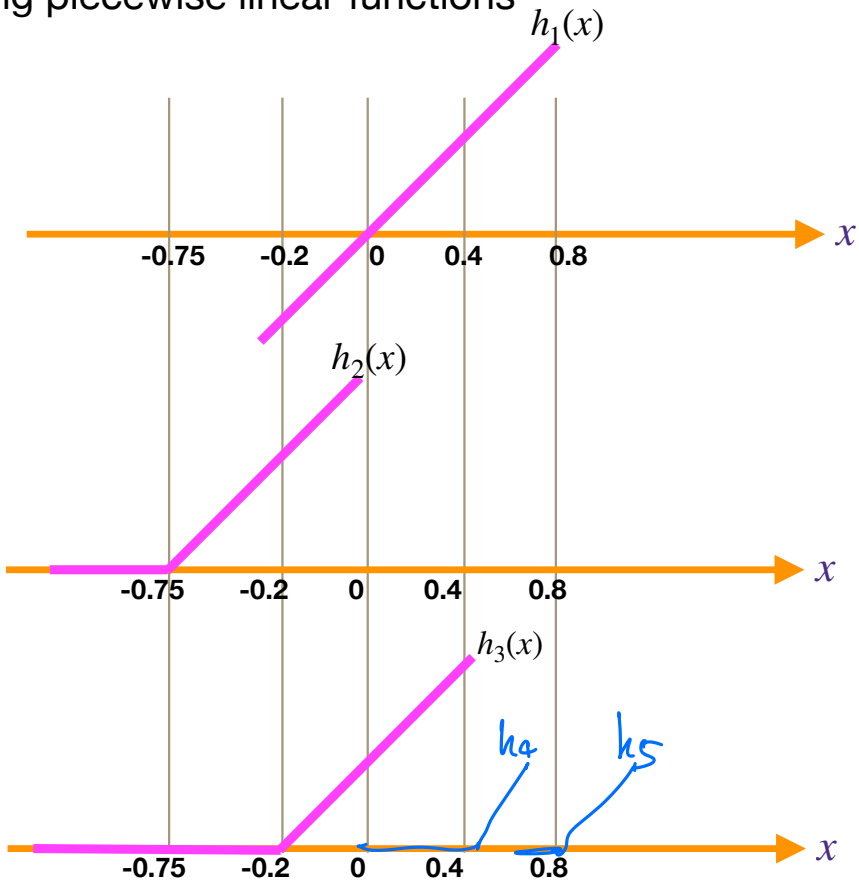
- Ridge regression
  - $L_2$ penalized least-squares regression $\lambda \cdot ||w||_2^2$
  - Regularization parameter trades off model complexity with training error
  - Never regularize the offset!

# 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

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

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

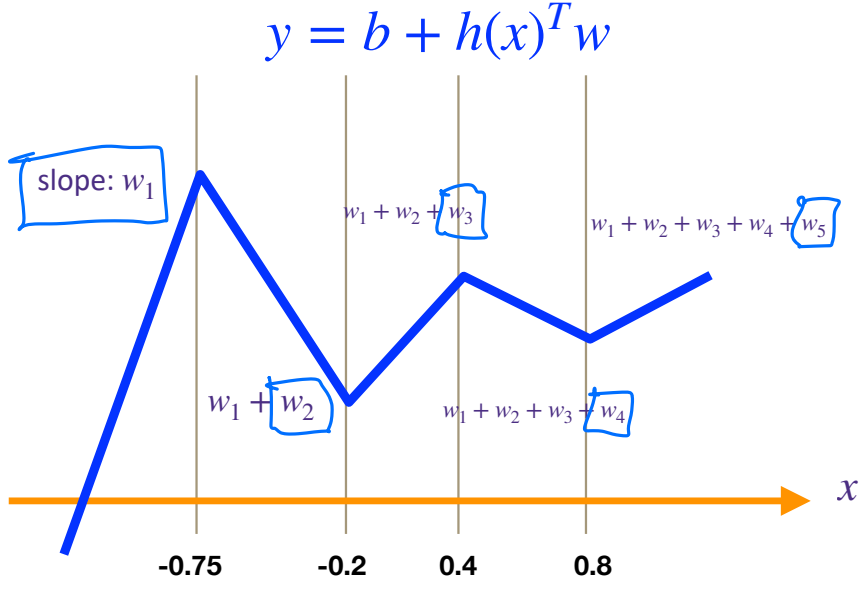


# Example: piecewise linear fit

- 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

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

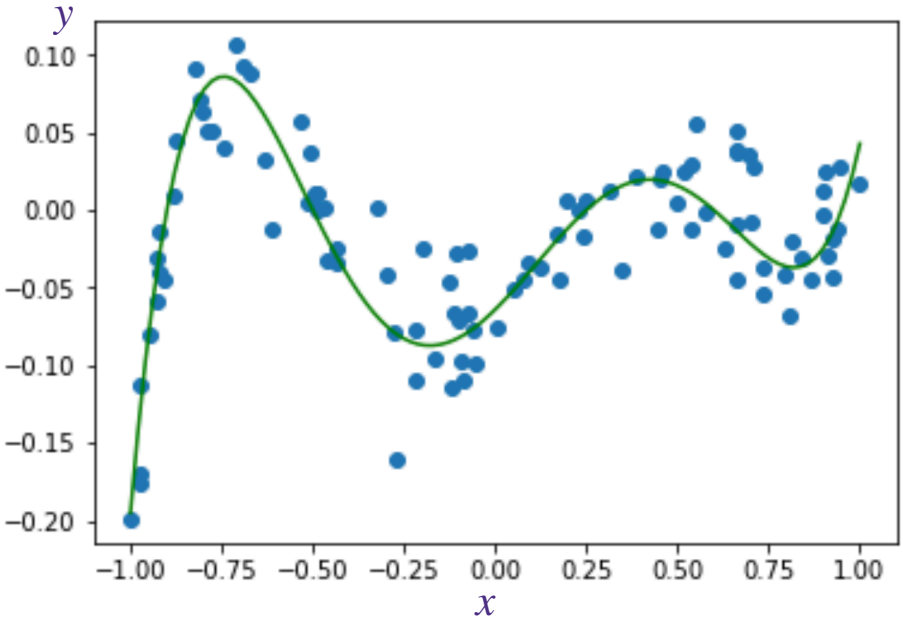
$[a]^+ \triangleq \max\{a, 0\}$




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

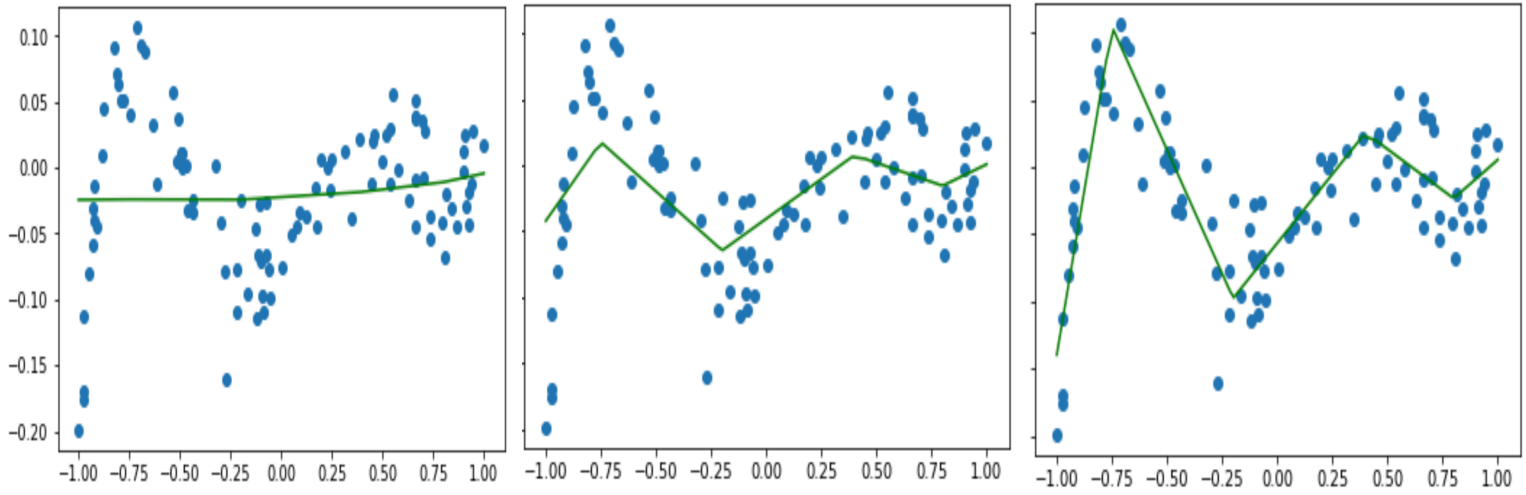
# Example: piecewise linear fit

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



# Example: piecewise linear fit (ridge regression)

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



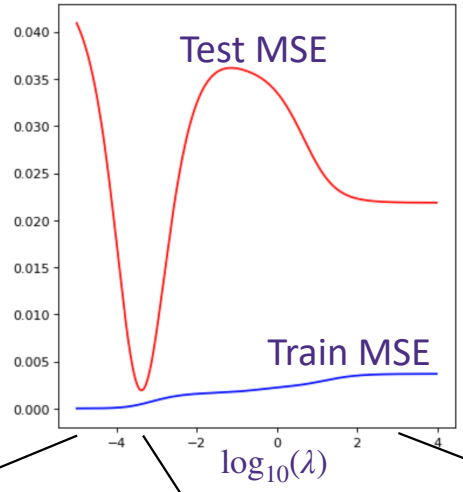

$\lambda = 1$ $\lambda = 0.005$ $\lambda = 0.000001$

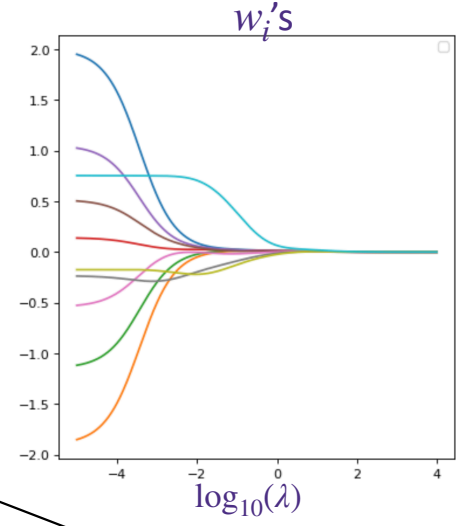
regularization ↑

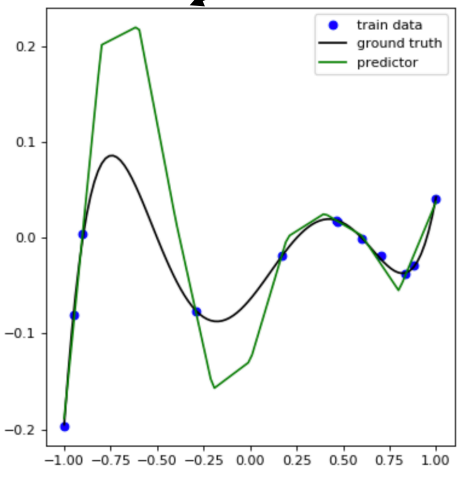
Prefer $w_1, w_2, \ldots \cong 0$

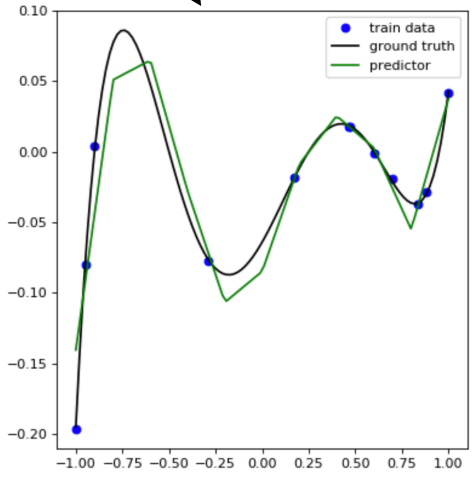
We do not observe overfitting, as d=5 and n=100

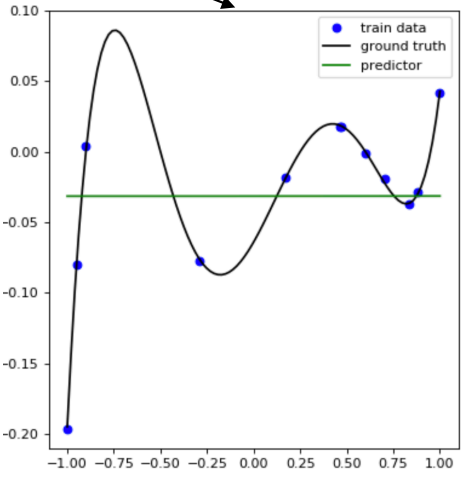
# Can avoid overfitting even $w \in \mathbb{R}^{10}$ and n=11 samples



# Questions?