

Regularization

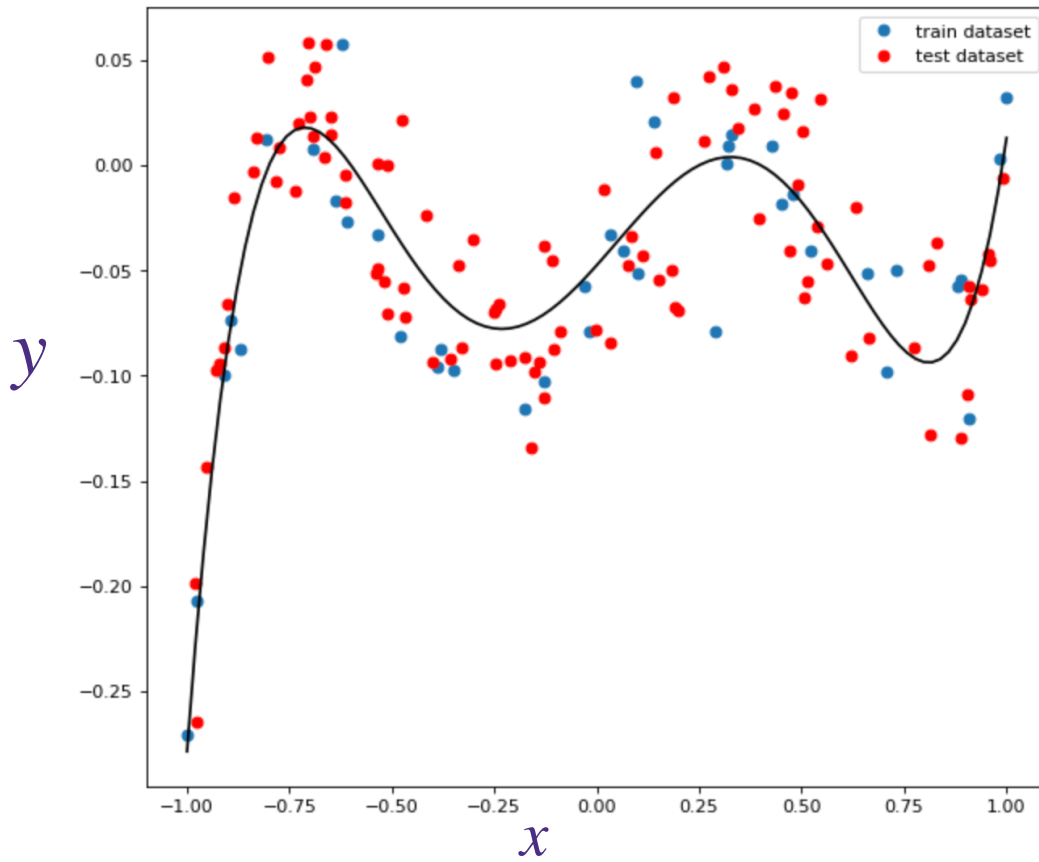


Recap: bias-variance tradeoff

- Consider 100 training examples and 100 test examples
i.i.d. drawn from degree-5 polynomial features

$$x_i \sim \text{Uniform}[-1, 1], y_i \sim f_{w^*}(x_i) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$f_w(x_i) = b^* + w_1^* x_i + w_2^* (x_i)^2 + w_3^* (x_i)^3 + w_4^* (x_i)^4 + w_5^* (x_i)^5$$

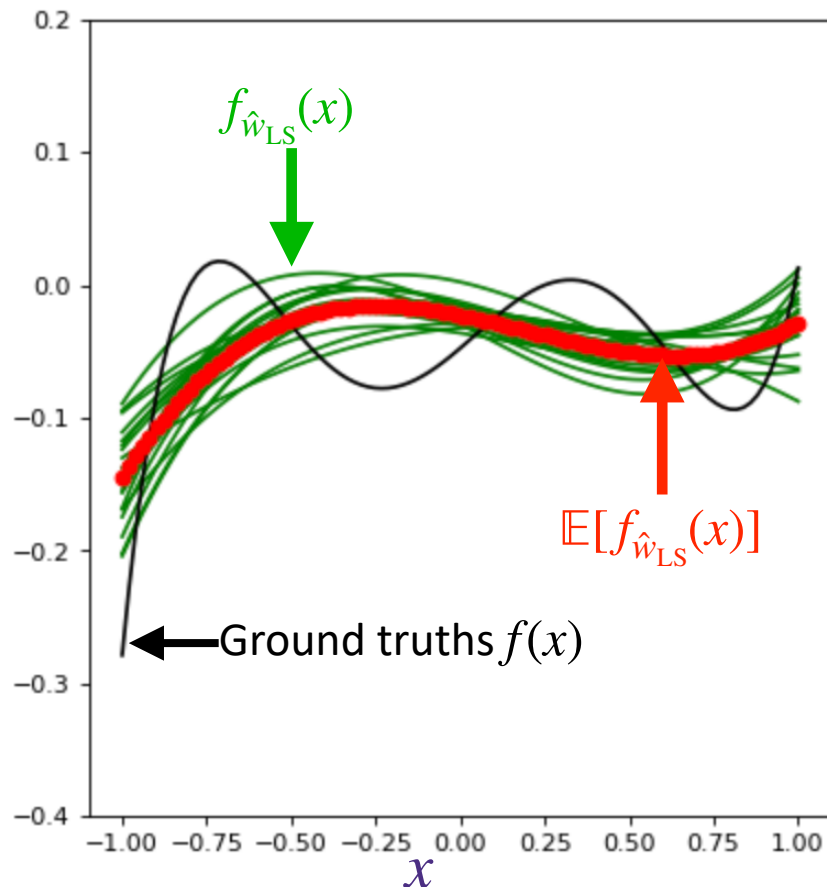


This is a linear model with features
 $h(x_i) = (x_i, (x_i)^2, (x_i)^3, (x_i)^4, (x_i)^5)$

Recap: bias-variance tradeoff

With degree-3 polynomials, we underfit

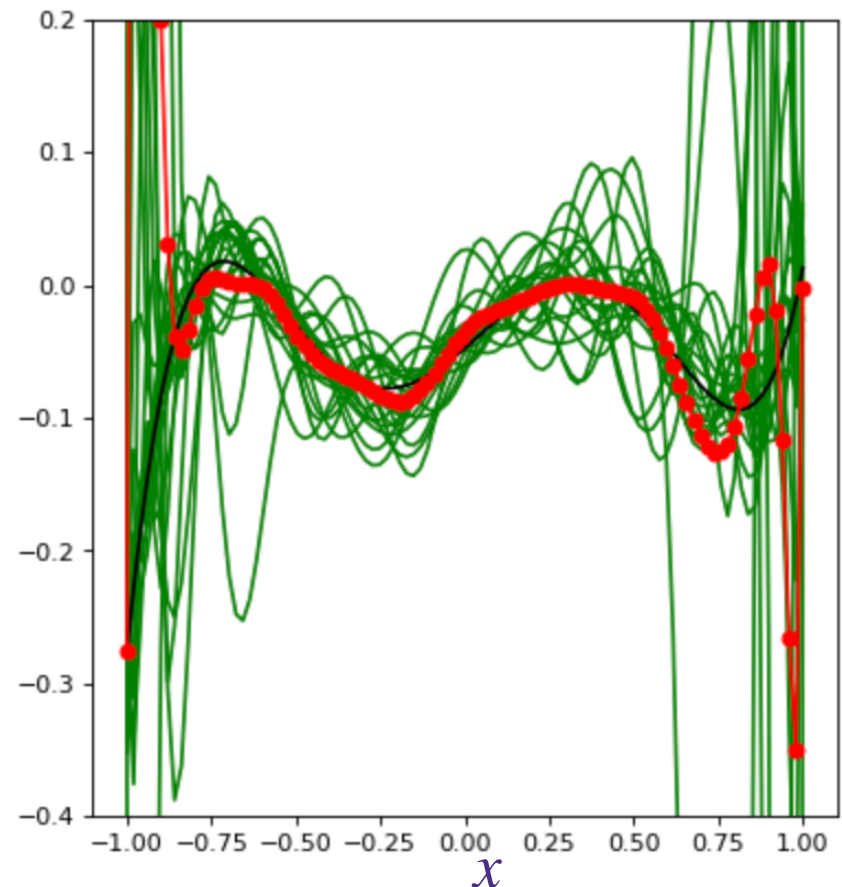
$$\hat{f}_{\hat{w}_{LS}}(x)$$



current train error = 0.0036791644380554187
current test error = 0.0037962529988410953

With degree-20 polynomials, we overfit

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



0.0005421686349568773
0.14210029429557927

Sensitivity: how to detect overfitting

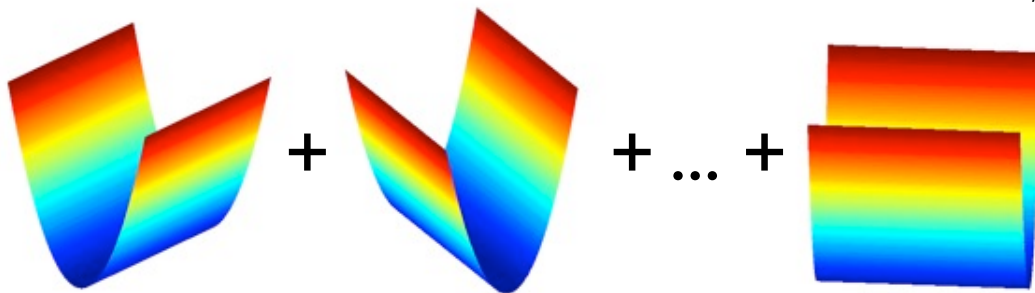
- For a linear model,
$$y \simeq b + w_1x_1 + w_2x_2 + \cdots + w_dx_d$$
if $|w_j|$ is large then the prediction is sensitive to small changes in x_j
- Large sensitivity leads to overfitting and poor generalization, and equivalently models that overfit tend to have large weights
- Note that b is a constant 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
- And optimize for small loss and small sensitivity, by adding a **regularizer** in the objective (assume no offset for now)

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

Ridge Regression

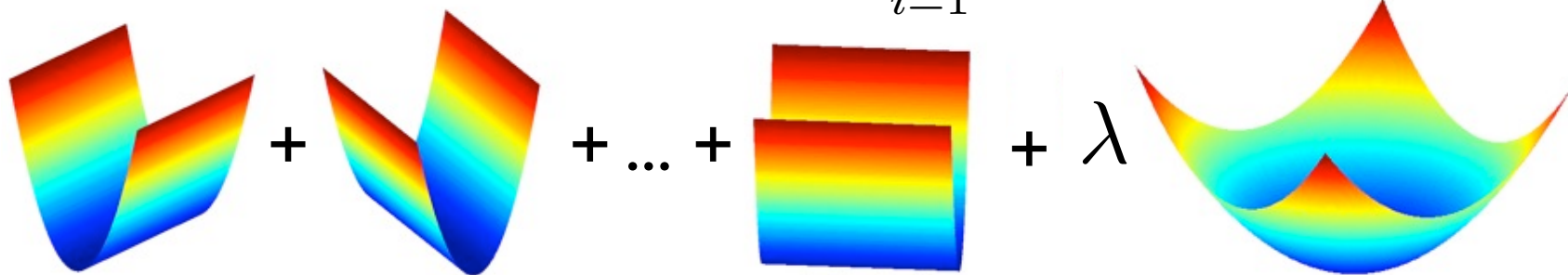
- (Original) Least squares objective:

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



- 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

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

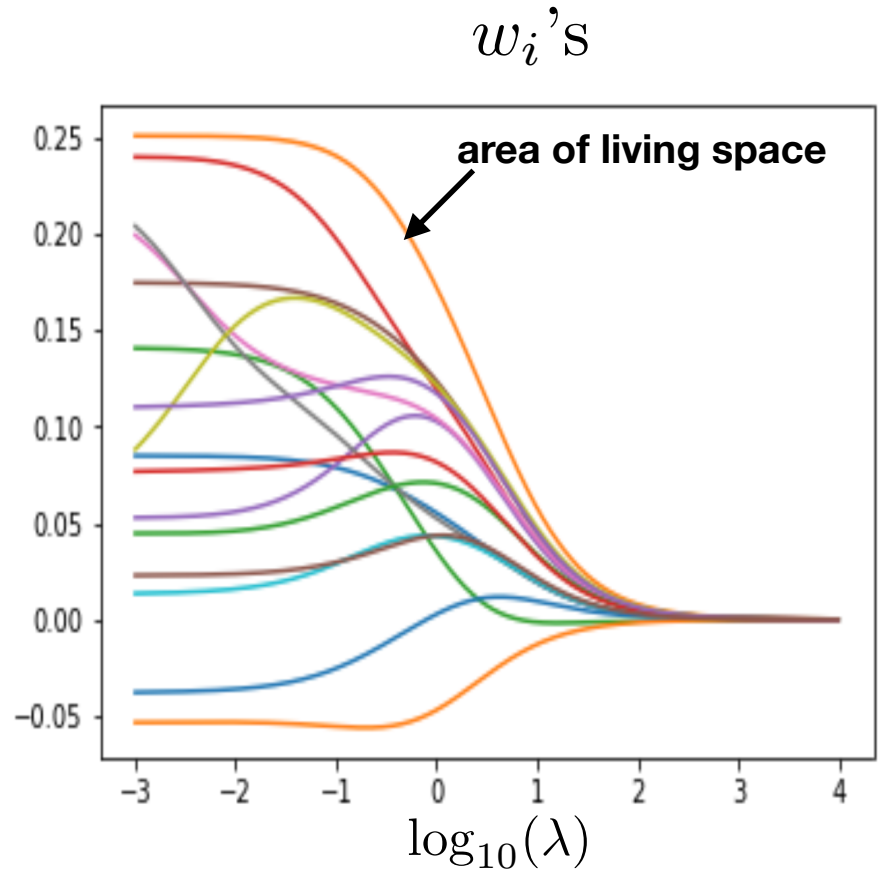
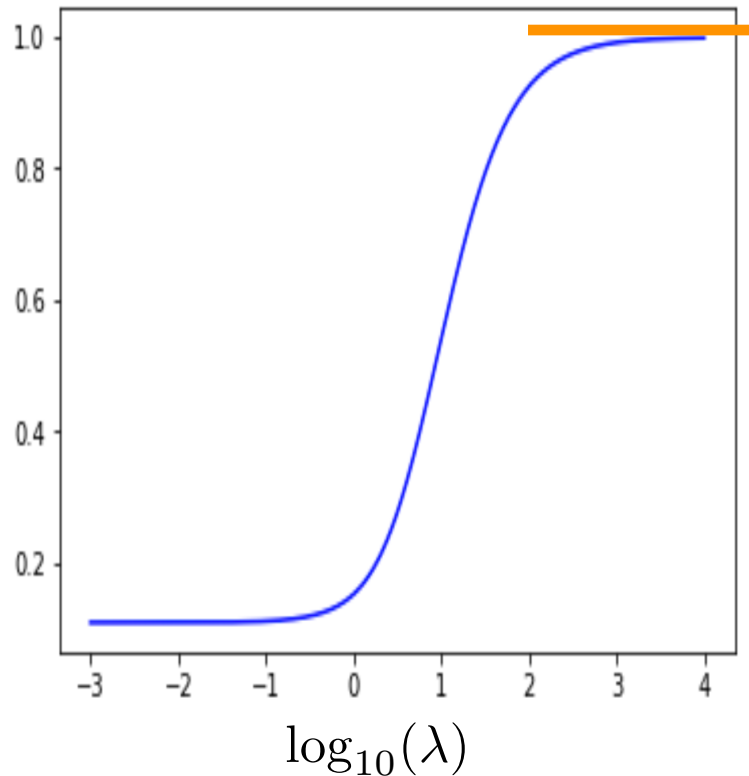
Shrinkage Properties

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

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

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

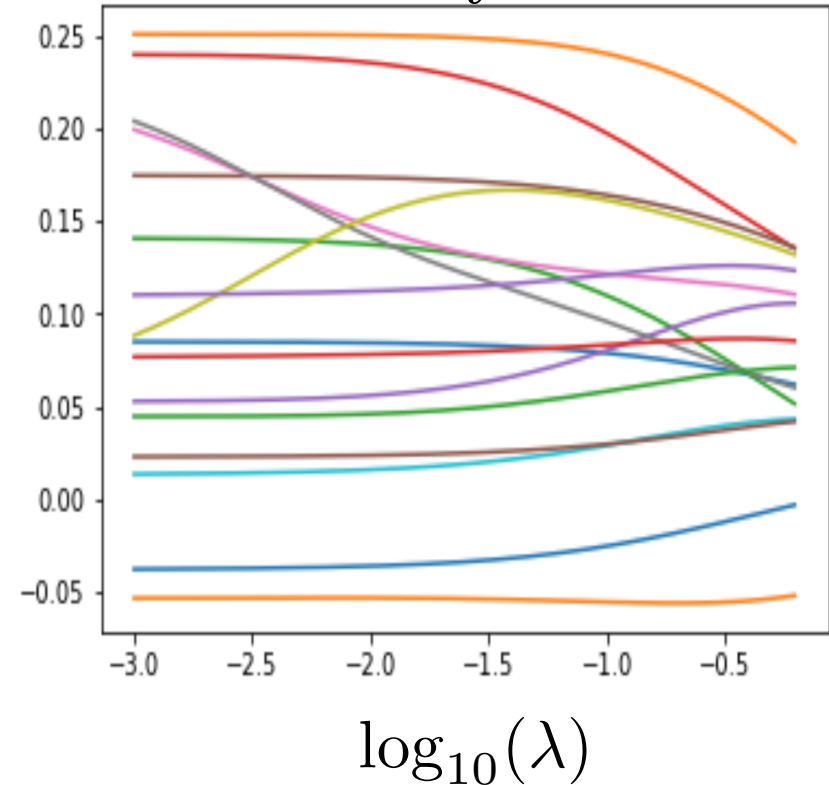
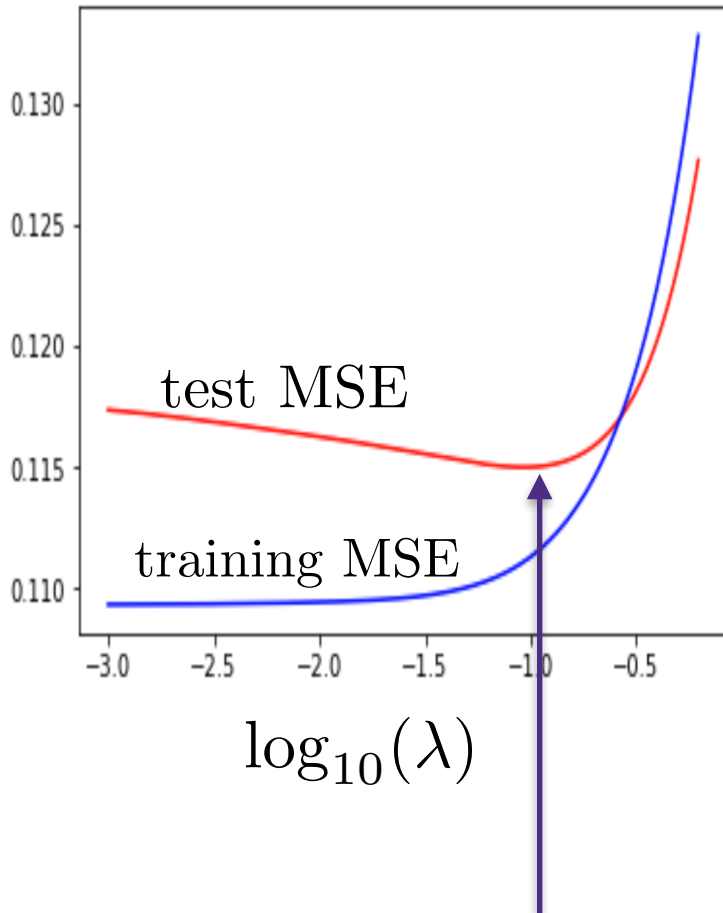
training MSE $\frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \hat{w}_{\text{ridge}}^{(\lambda)})^2$



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

w_i 's



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

Bias-Variance Properties

- Recall: $\hat{\mathbf{w}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X$, $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- The true error at a sample with feature x is
$$\mathbb{E}_{y, \mathcal{D}_{\text{train}} | x} [(y - x^T \hat{\mathbf{w}}_{\text{ridge}})^2 | x]$$

Bias-Variance Properties

- Recall: $\hat{\mathbf{w}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X$, $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- The true error at a sample with feature x is

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

Bias-Variance Properties

- Recall: $\hat{\mathbf{w}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X$, $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- The true error at a sample with feature x is

$$\begin{aligned} & \mathbb{E}_{y, \mathcal{D}_{\text{train}} | x} [(y - x^T \hat{\mathbf{w}}_{\text{ridge}})^2 | x] \\ &= \mathbb{E}_{y|x} [(y - \mathbb{E}[y | x])^2 | x] + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(\mathbb{E}[y | x] - x^T \hat{\mathbf{w}}_{\text{ridge}})^2 | x] \\ &= \mathbb{E}_{y|x} [(y - x^T \mathbf{w})^2 | x] + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(x^T \mathbf{w} - x^T \hat{\mathbf{w}}_{\text{ridge}})^2 | x] \end{aligned}$$

Bias-Variance Properties

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X$, $\mathbf{y} = \mathbf{X}w + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- The true error at a sample with feature x is

$$\mathbb{E}_{y, \mathcal{D}_{\text{train}} | x} [(y - x^T \hat{w}_{\text{ridge}})^2 | x]$$

$$= \mathbb{E}_{y|x} [(y - \mathbb{E}[y | x])^2 | x] + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(\mathbb{E}[y | x] - x^T \hat{w}_{\text{ridge}})^2 | x]$$

$$= \mathbb{E}_{y|x} [(y - x^T w)^2 | x] + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(x^T w - x^T \hat{w}_{\text{ridge}})^2 | x]$$

$$= \underbrace{\sigma^2}_{\text{Irreduc. Error}} + \underbrace{(x^T w - \mathbb{E}_{\mathcal{D}_{\text{train}}} [x^T \hat{w}_{\text{ridge}} | x])^2}_{\text{Bias-squared}} + \underbrace{\mathbb{E}_{\mathcal{D}_{\text{train}}} [(\mathbb{E}_{\tilde{\mathcal{D}}_{\text{train}}} [x^T \hat{w}_{\text{ridge}} | x] - x^T \hat{w}_{\text{ridge}})^2 | x]}_{\text{Variance}}$$

Irreduc. Error

Bias-squared

Variance

Bias-Variance Properties

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X$, $\mathbf{y} = \mathbf{X}w + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- The true error at a sample with feature x is

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

Suppose $\mathbf{X}^T \mathbf{X} = n \mathbf{I}$, then $\hat{w}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T (\mathbf{X}w + \epsilon)$

$$= \frac{n}{n + \lambda} w + \frac{1}{n + \lambda} \mathbf{X}^T \epsilon$$

Bias-Variance Properties

Suppose $\mathbf{X}^T \mathbf{X} = n\mathbf{I}$, then

$$\hat{w}_{\text{ridge}} = \frac{n}{n + \lambda} w + \frac{1}{n + \lambda} \mathbf{X}^T \epsilon$$

- Recall: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$
- To analyze bias-variance tradeoff, we need to assume probabilistic generative model: $x_i \sim P_X$, $\mathbf{y} = \mathbf{X}w + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- The true error at a sample with feature x is

$$\mathbb{E}_{y, \mathcal{D}_{\text{train}} | x} [(y - x^T \hat{w}_{\text{ridge}})^2 | x]$$

$$= \mathbb{E}_{y|x} [(y - \mathbb{E}[y | x])^2 | x] + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(\mathbb{E}[y | x] - x^T \hat{w}_{\text{ridge}})^2 | x]$$

$$= \mathbb{E}_{y|x} [(y - x^T w)^2 | x] + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(x^T w - x^T \hat{w}_{\text{ridge}})^2 | x]$$

$$= \sigma^2 + (x^T w - \mathbb{E}_{\mathcal{D}_{\text{train}}} [x^T \hat{w}_{\text{ridge}} | x])^2 + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(\mathbb{E}_{\tilde{\mathcal{D}}_{\text{train}}} [x^T \hat{w}_{\text{ridge}} | x] - x^T \hat{w}_{\text{ridge}})^2 | x]$$

(verify at home)

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

Irreduc. Error

Bias-squared

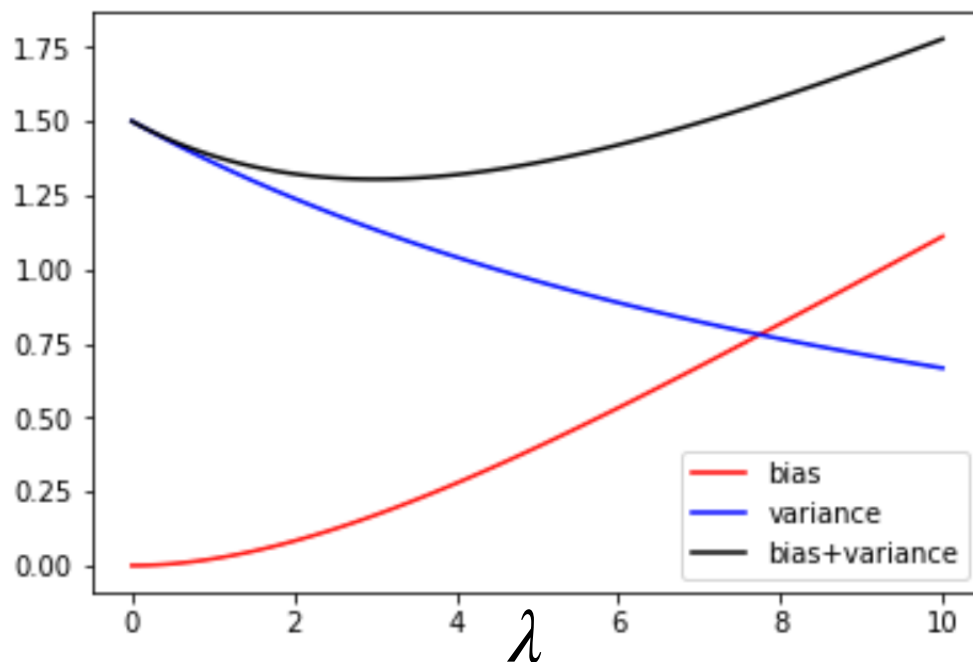
Variance

Bias-Variance Properties

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

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

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



as $\lambda \rightarrow 0$,

$$\hat{w}_{ridge} \rightarrow \hat{w}_{LS}$$

as $\lambda \rightarrow \infty$

$$\hat{w}_{ridge} \rightarrow 0$$

What you need to know...

> Regularization

- Penalizes complex models towards preferred, simpler models

> Ridge regression

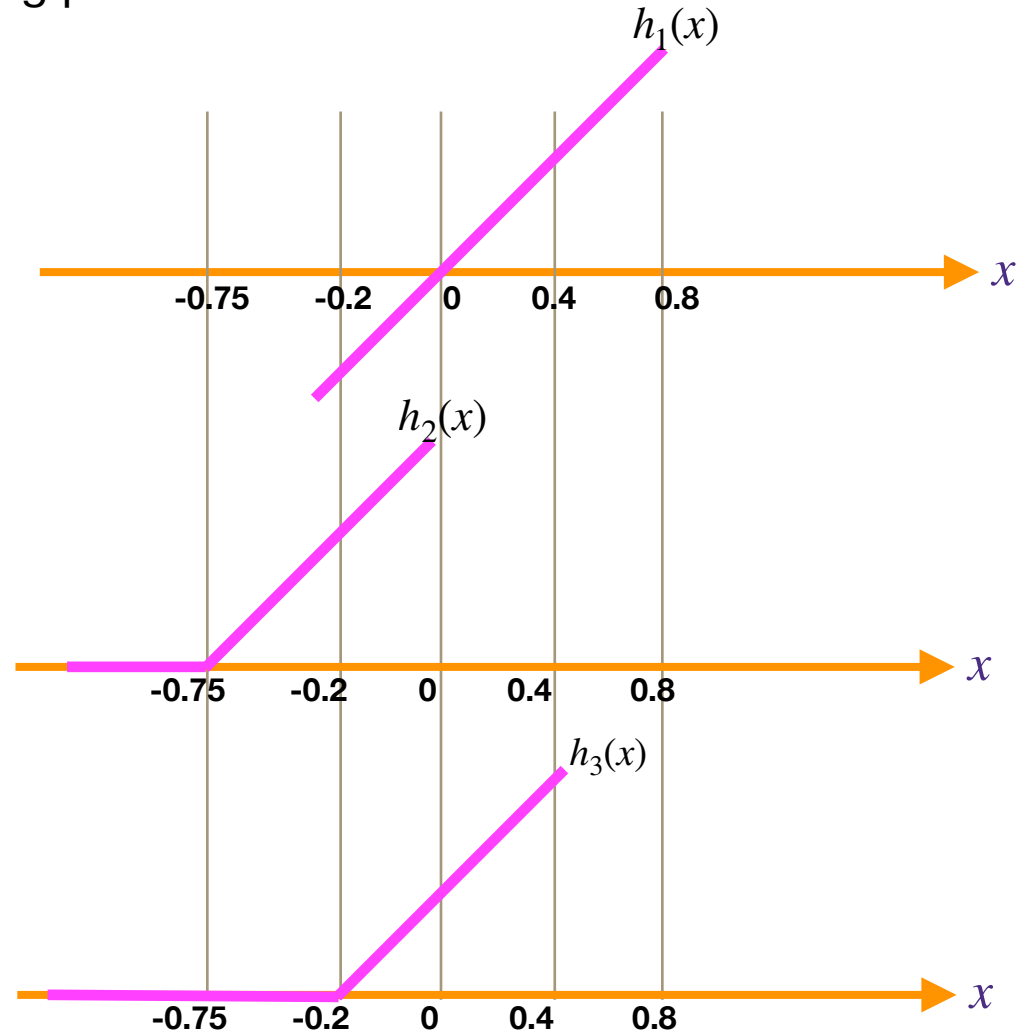
- L_2 penalized least-squares regression
- Regularization parameter trades off model complexity with training error
- Never regularize the offset!

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

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

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

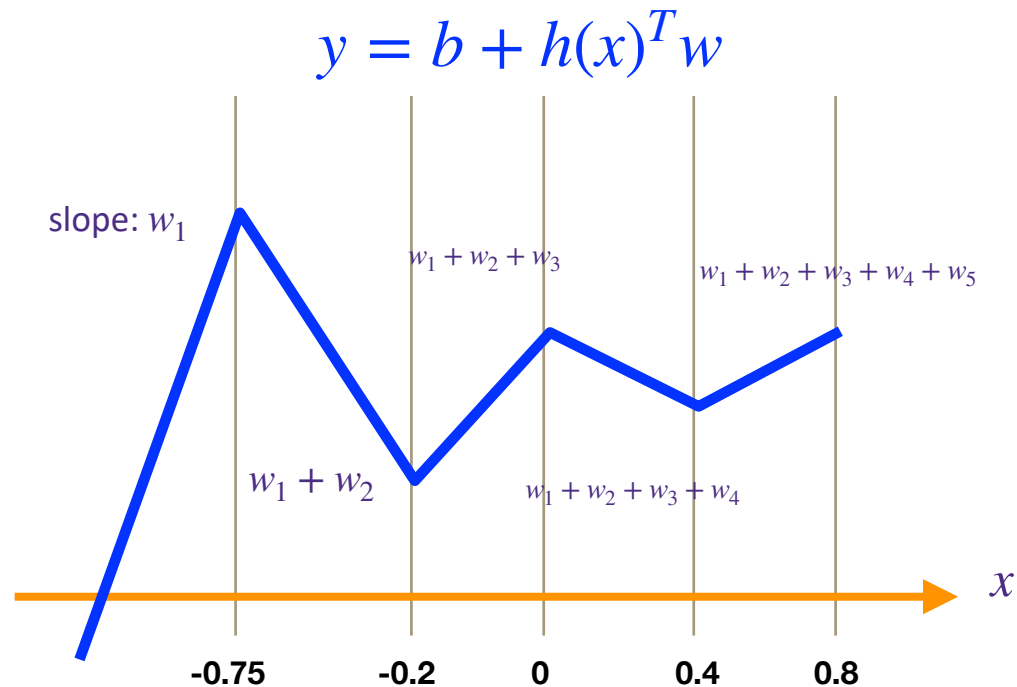


Example: piecewise linear fit

- 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

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

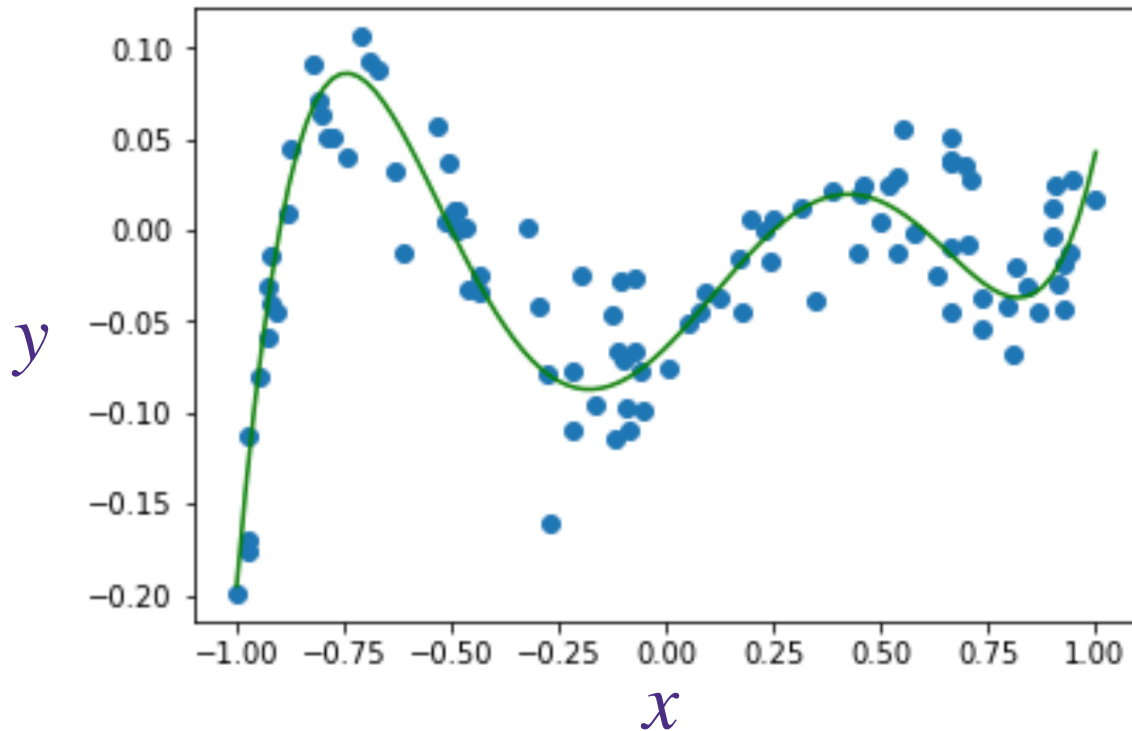
$$[a]^+ \triangleq \max\{a, 0\}$$



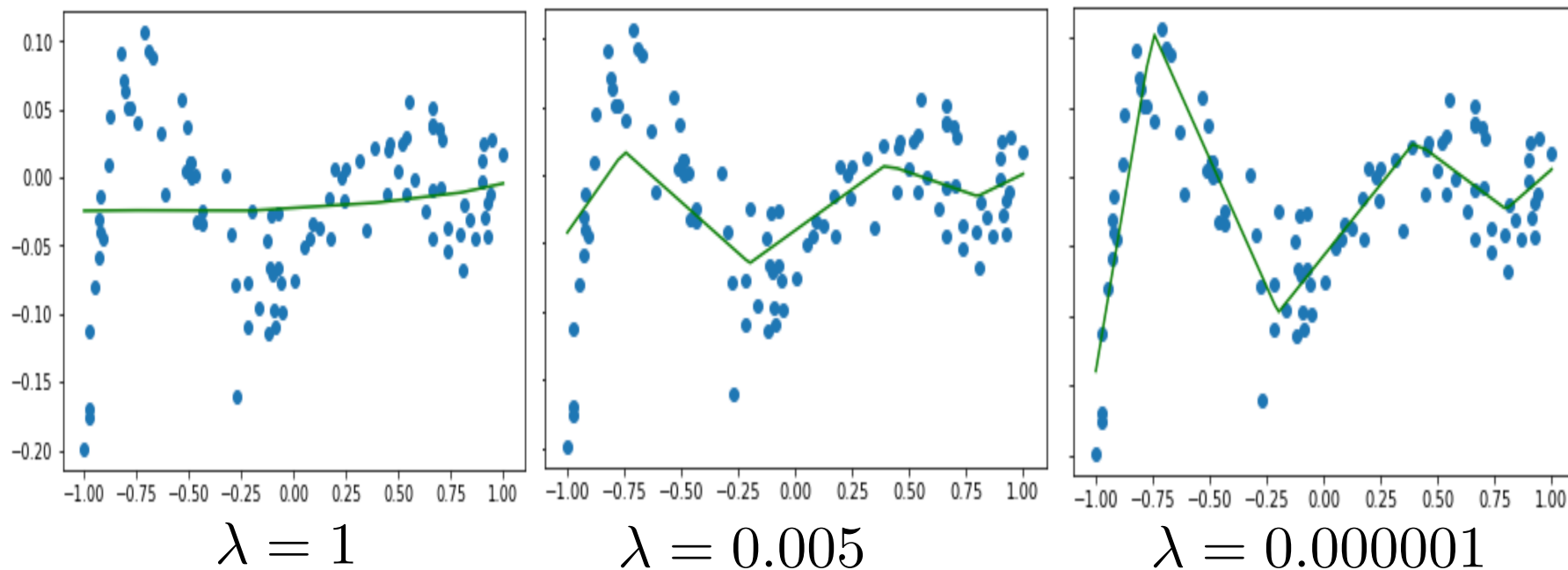
the weights capture the change in the slopes

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 $d=5 \ll n=100$

Piecewise linear with $w \in \mathbb{R}^{10}$ and $n=11$ samples

