

CSE 446

Regularization & Sparsity

Natasha Jaques



Bias-Variance Tradeoff

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

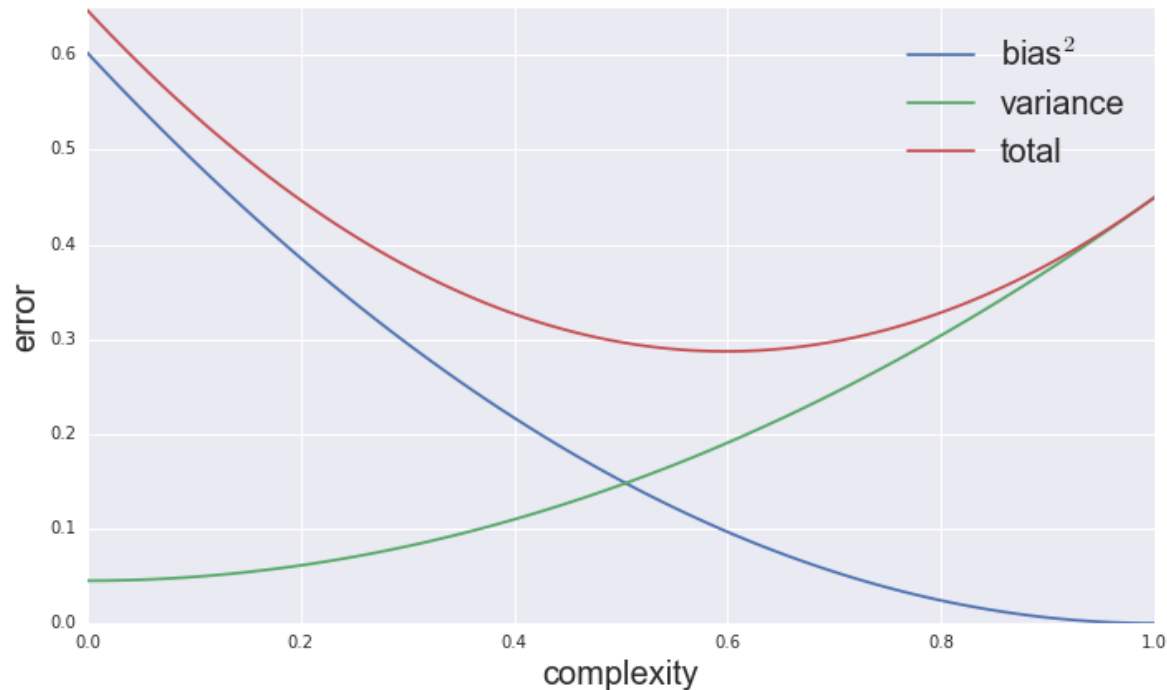
True generalization error

irreducible error

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

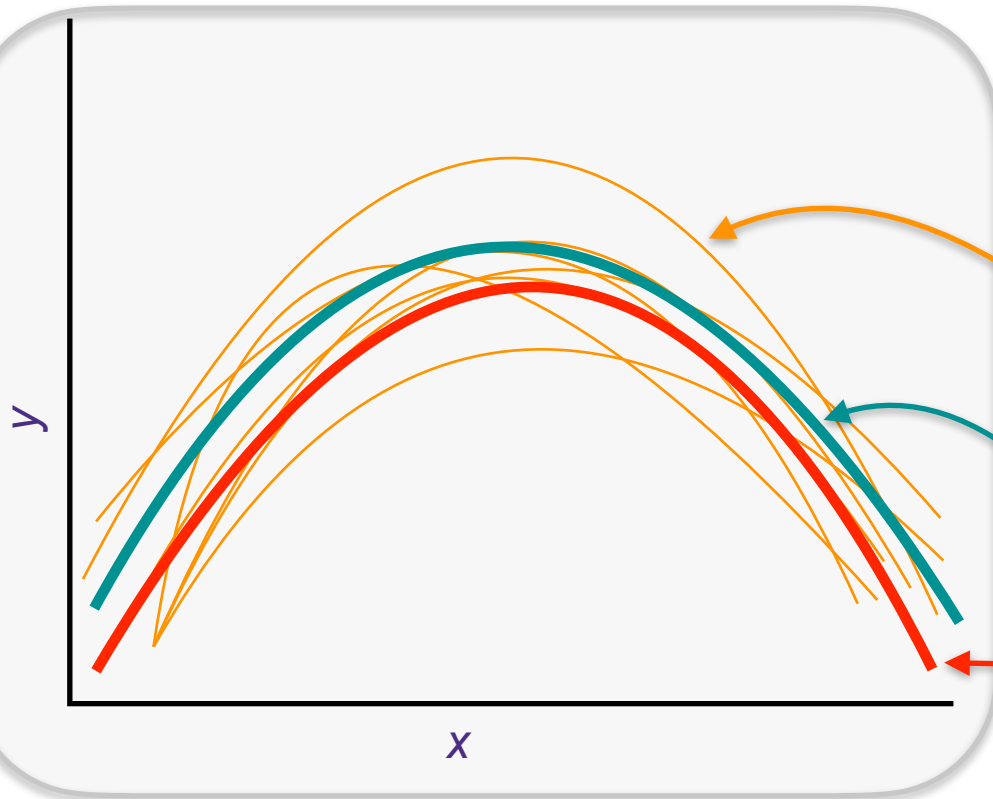
biased squared

variance



Bias-Variance Tradeoff

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



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

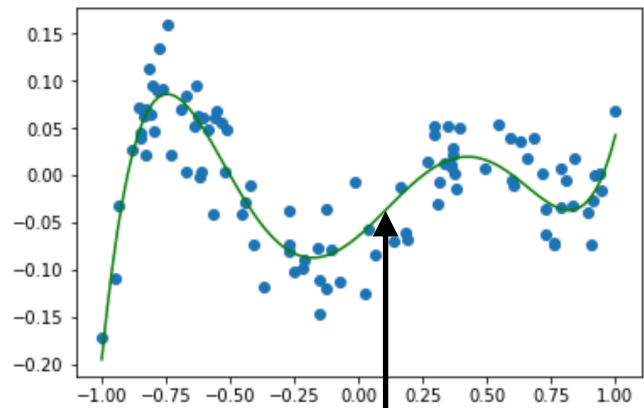
$$\mathbb{E}_{\mathcal{D}}[\hat{f}(x)]$$

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

Test error vs. model complexity

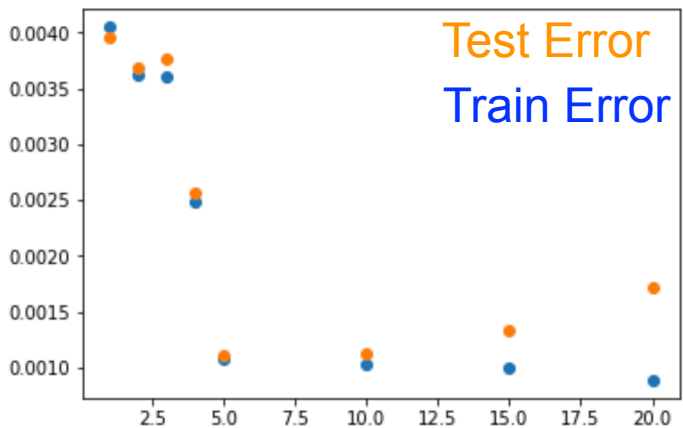
Simple model: # bias
Model complexity is below the complexity of $\eta(x)$

Complex model:
Fits noise in train data, diverging from $\eta(x)$

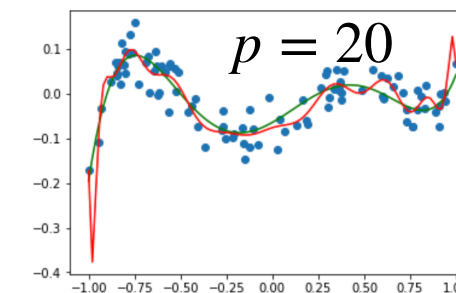
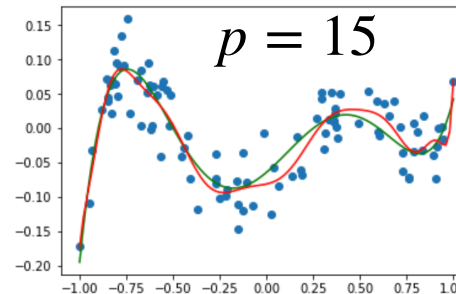
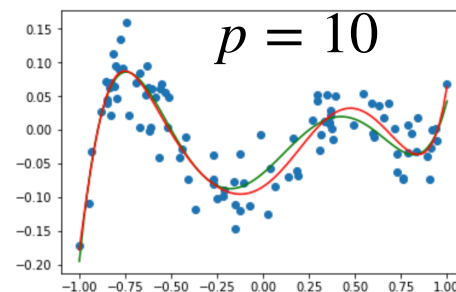
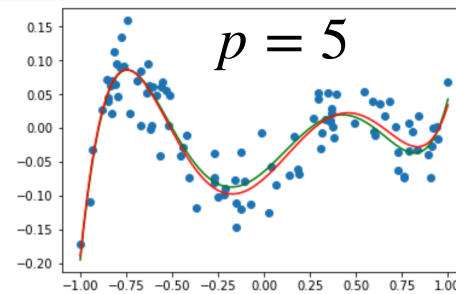
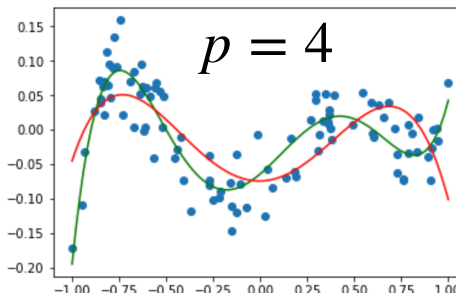
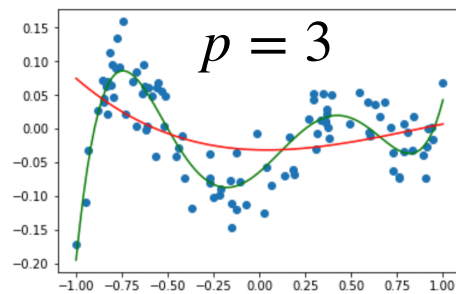
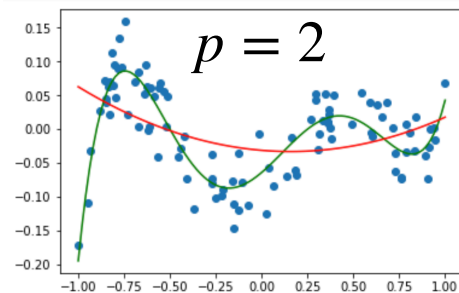
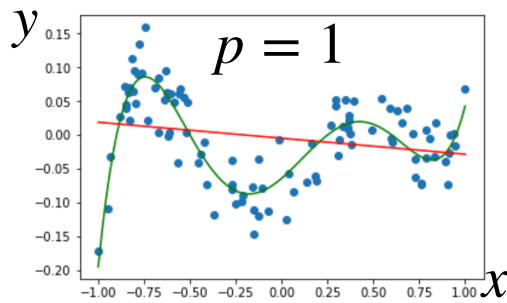


Optimal predictor $\eta(x)$ is degree-5 polynomial

Error

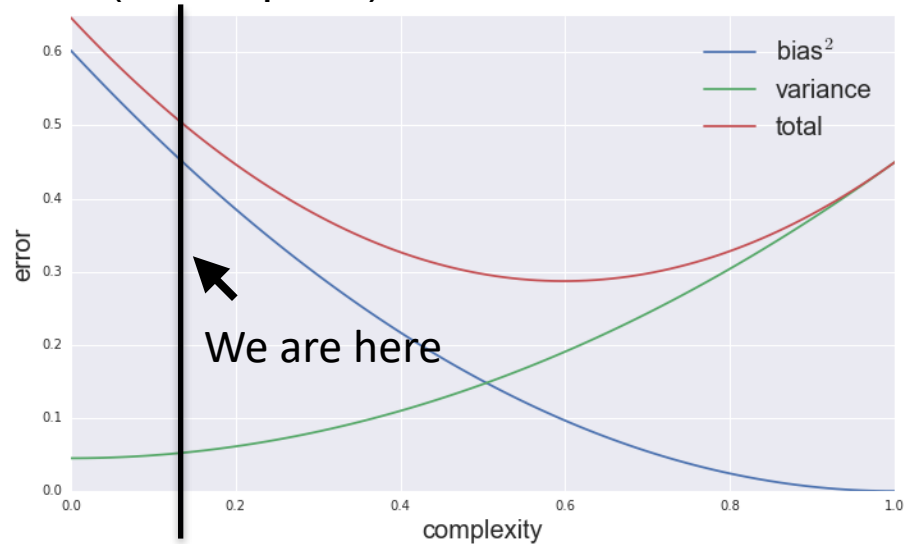


degree p of the polynomial regression



Recap: Bias-variance tradeoff with simple model

(Conceptual) bias variance tradeoff



- When model **complexity is low** (lower than the optimal predictor $\eta(x)$)

- Bias² of our predictor,

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

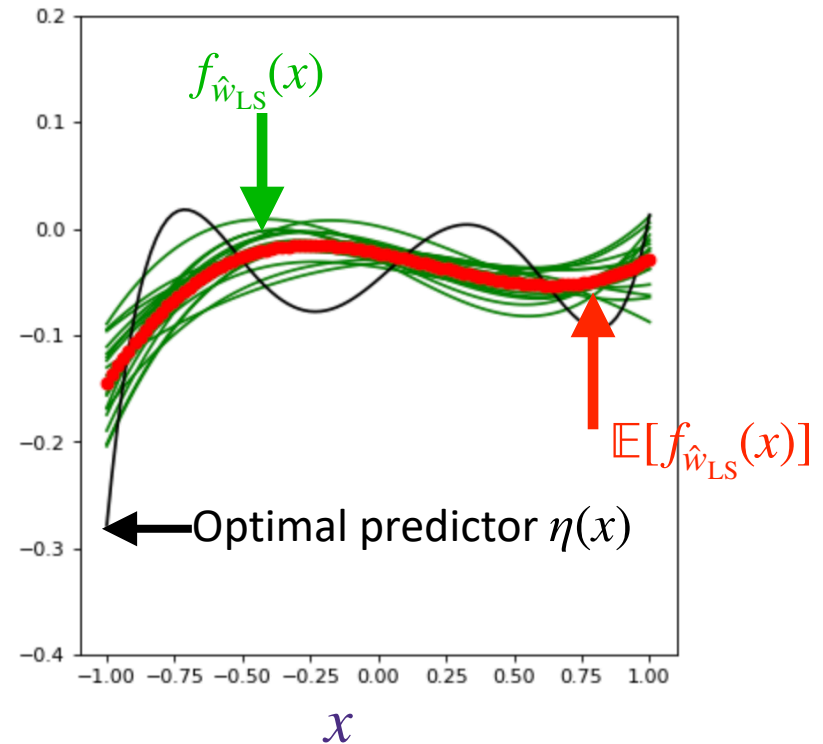
Large

- Variance of our predictor,

$$\mathbb{E}_{\mathcal{D}} \left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x) \right)^2 \right]$$

Small

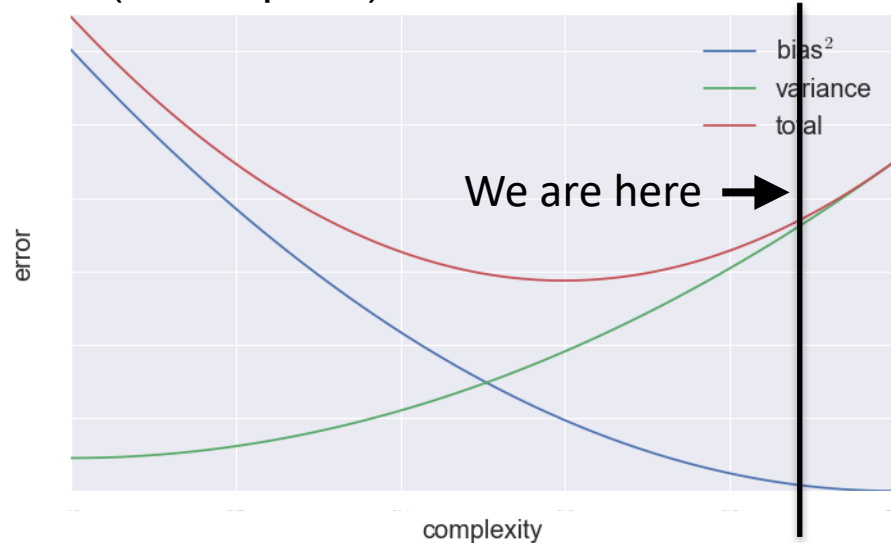
With degree-3 polynomials, we underfit



- If we have more samples (larger n), then
 - What happens to bias? # stays the same
 - What happens to variance? # decreases
 - What happens to overall test error?
 - # Stays high because bias was dominating our error term

Recap: Bias-variance tradeoff with complex model

(Conceptual) bias variance tradeoff



- When model **complexity is high** (higher than the optimal predictor $\eta(x)$)

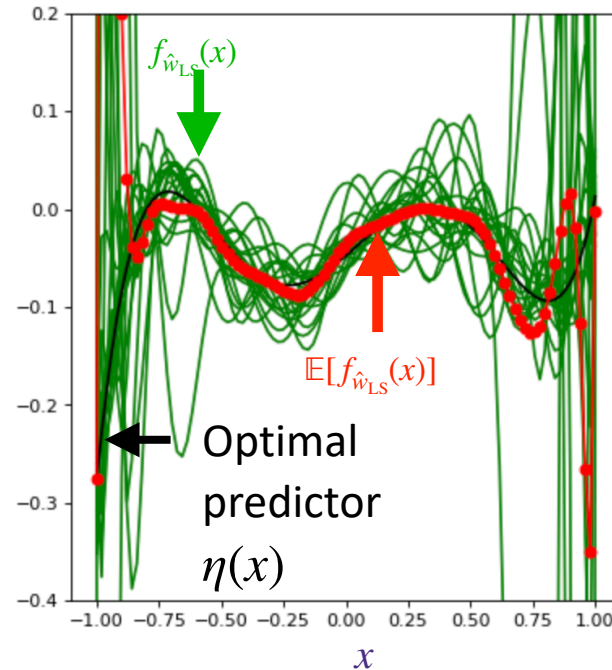
- Bias² of our predictor,
 $(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)])^2$

Low

- Variance of our predictor,
 $\mathbb{E}_{\mathcal{D}} [(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x))^2]$

High

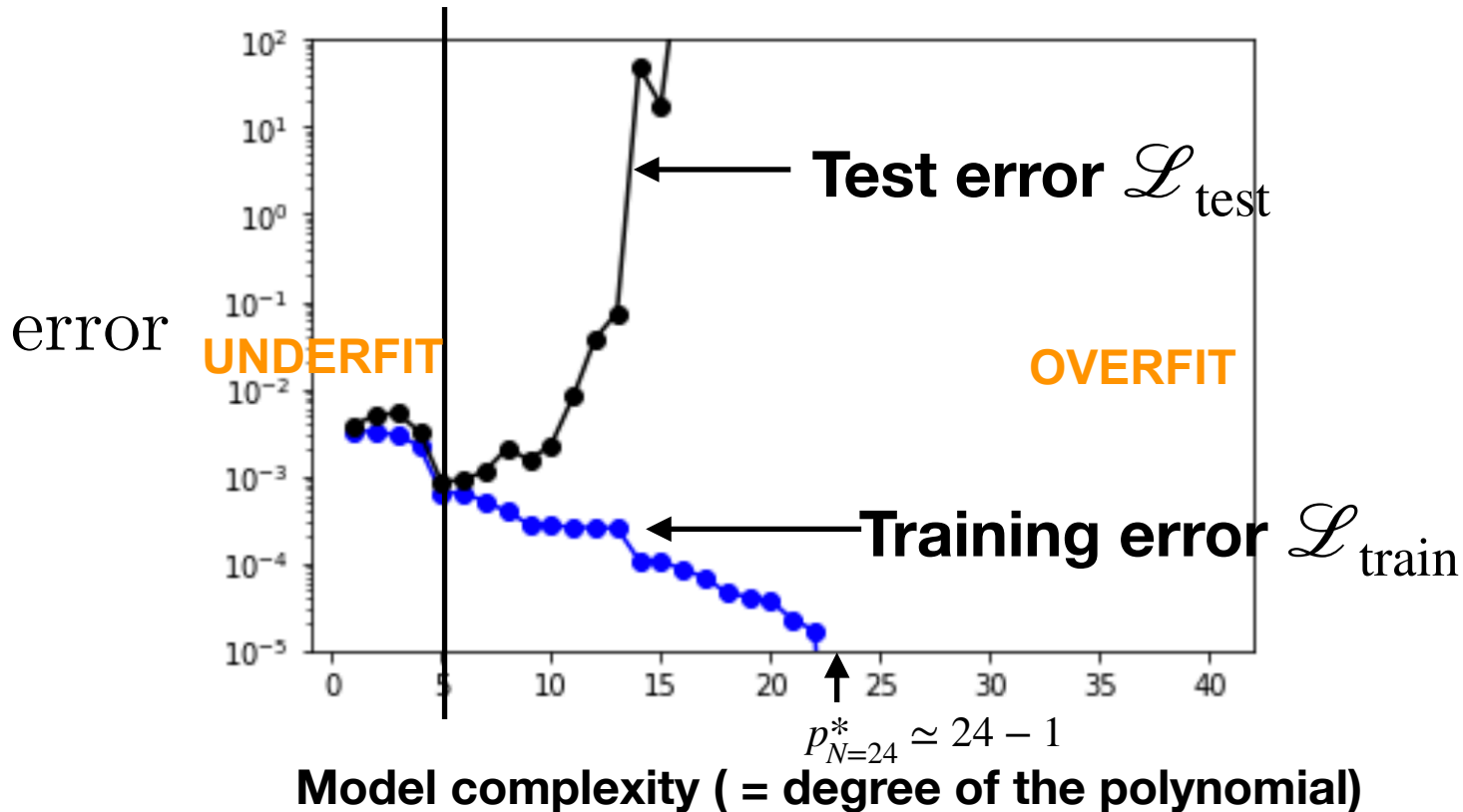
With degree-20 polynomials, we overfit



- If we have more samples (larger n), then
 - What happens to bias? # stays the same
 - What happens to variance? # decreases
 - What happens to overall test error? # decreases

Optimal model complexity depends on dataset size

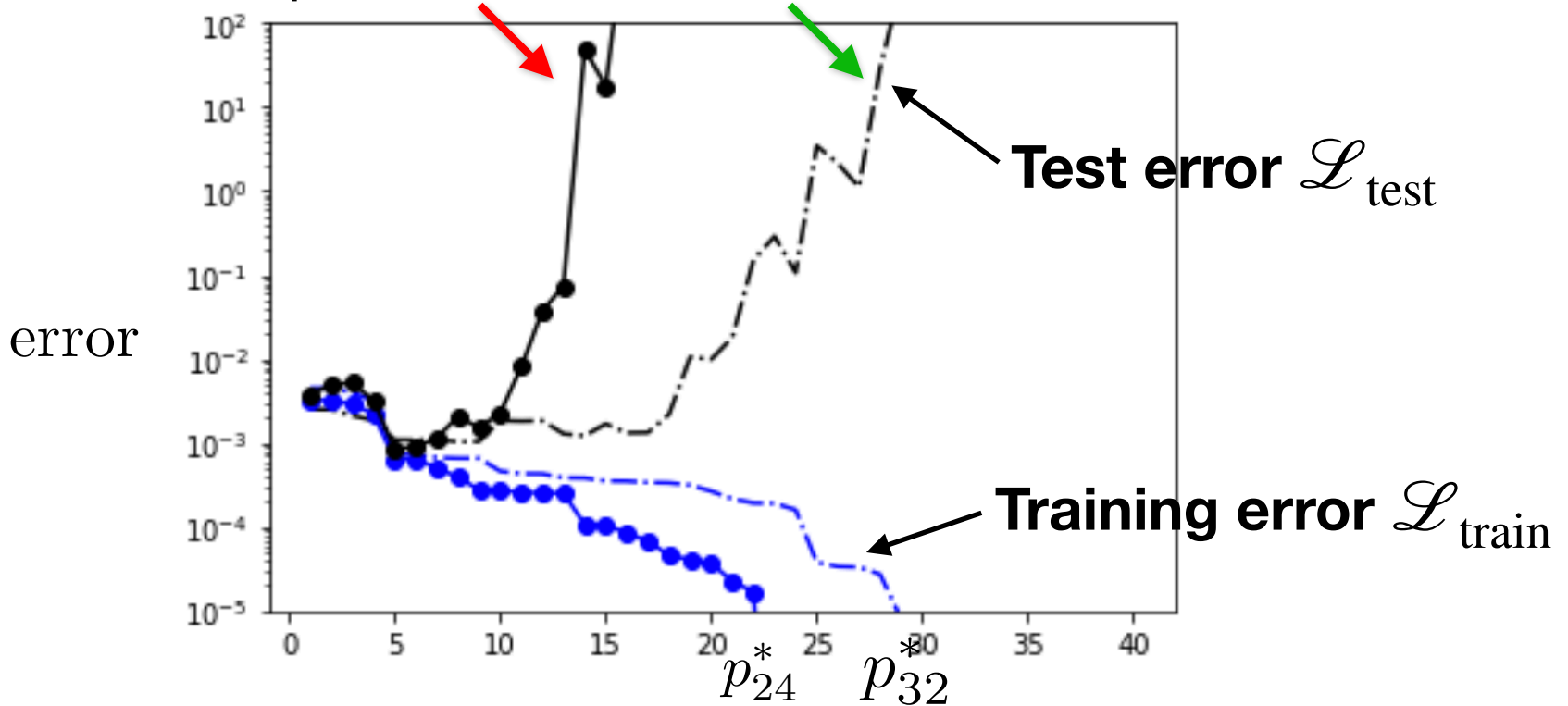
- Assume $N=30$



- Given sample size N there is a threshold, p_N^* , where training error is zero
- Training error is **always** monotonically non-increasing
- Test error has a trend of going down and then up, but fluctuates

Variance decreases with more data, letting you fit more complex models

- Now compare $N=30$ case to $N=40$

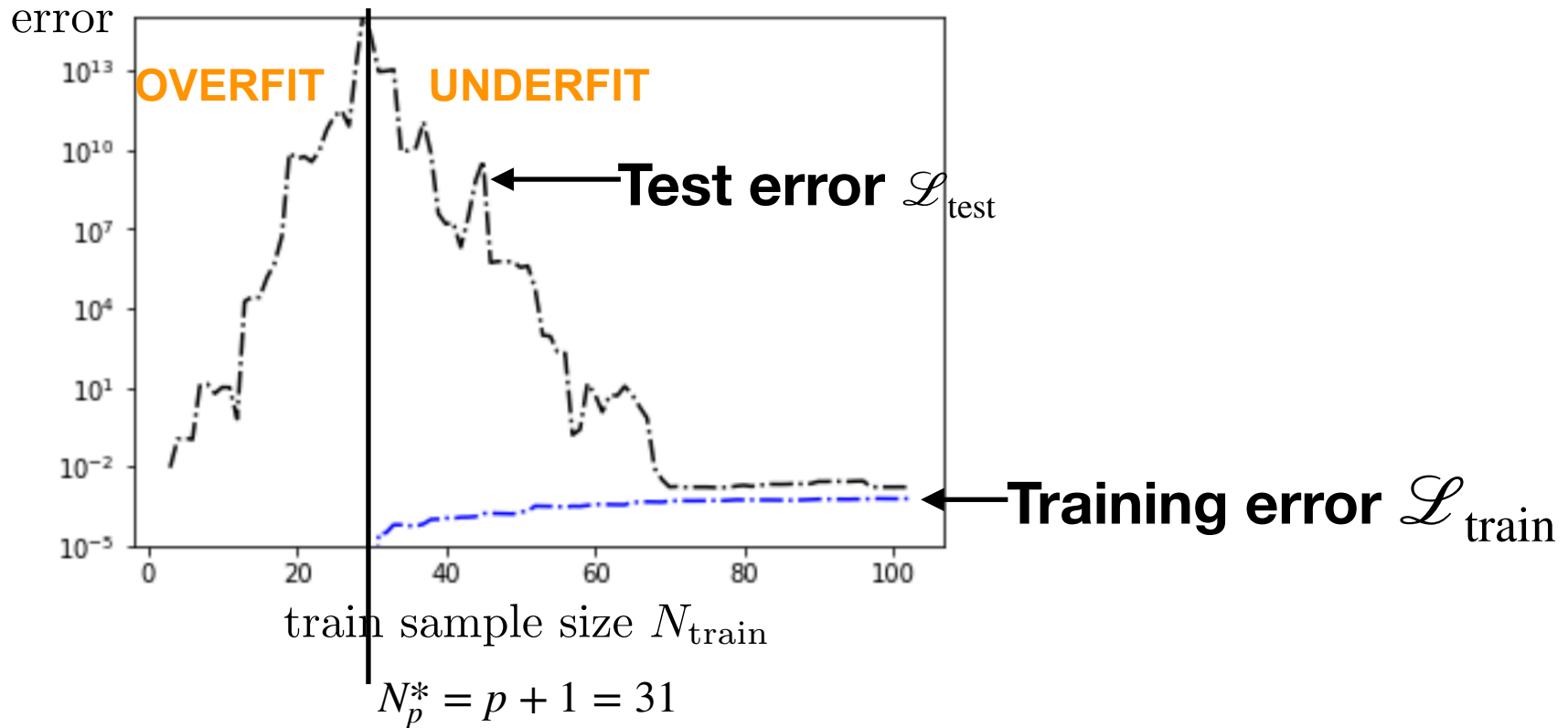


Model complexity (= degree of the polynomial)

- The threshold, p_N^* , moves right as dataset size increases**
- Training error tends to increase, because more points need to fit
- Test error tends to decrease, because Variance decreases

Variance decreases with more data, letting you fit more complex models

- Choose model complexity $p=30$, vary dataset size n



- There is a threshold, N_p^* , below which training error is zero (extreme overfit)
- Above the threshold, test error tends to decrease
- Training error tends to increase (harder to fit so much data with simple model)

Regularization helps avoid overfitting

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

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

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

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

Sensitivity increases overfitting

- For a linear model,
$$y \simeq b + w_1x_1 + w_2x_2 + \dots + w_dx_d$$
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

Never regularize 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)

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

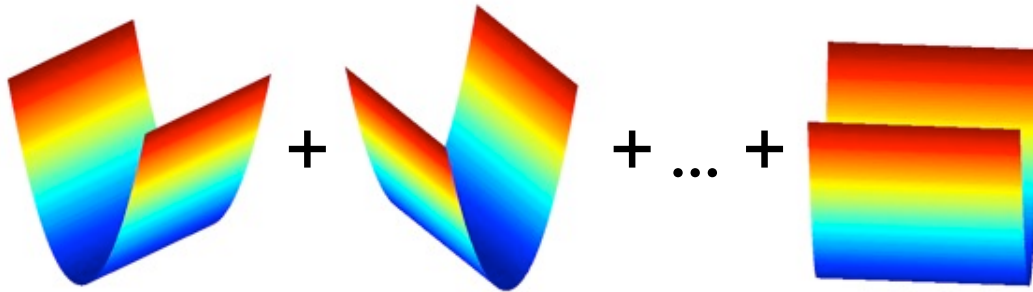

Regularize to make weights smaller / less sensitive. Multiple ways to do that. We'll start with L2 norm

Ridge Regression

if $d > n$

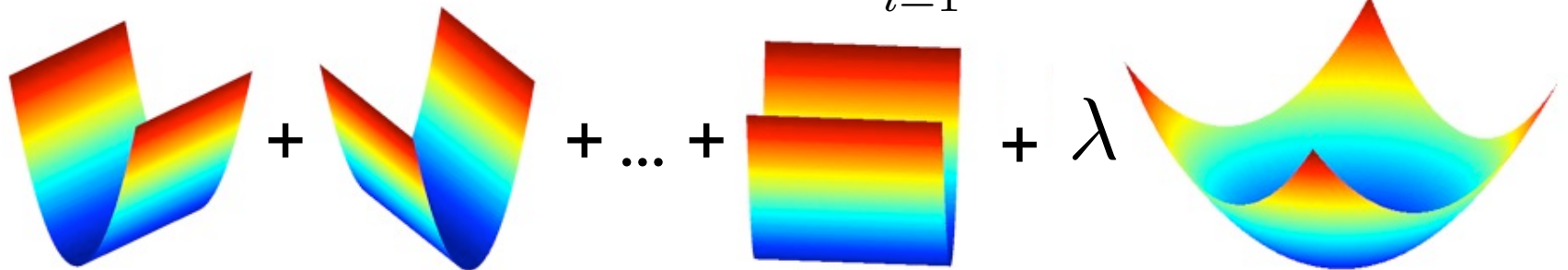
- Old Least squares objective:

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



- 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

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

$w \in \mathbb{R}^{d \times 1}$
 $X \in \mathbb{R}^{n \times d}$

More identities

$$\|w\|_p = (|w_1|^p + \dots + |w_d|^p)^{\frac{1}{p}}$$

$$I^d = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \ddots & \\ & & 1 \end{bmatrix}$$

$$\arg \min_w \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

$$\arg \min_w (Xw - y)^T (Xw - y) + \lambda w^T w$$

From previous derivation

$$\nabla_w [-2y^T Xw + w^T X^T Xw + \lambda w^T w] = 0$$

$$-2X^T y + 2X^T Xw + 2\lambda w = 0$$

$$X^T Xw + \lambda w = X^T y$$

$$(X^T X + \lambda I)w = X^T y$$

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

Scalar derivative	Vector derivative
$f(x) \rightarrow \frac{df}{dx}$	$f(\mathbf{x}) \rightarrow \frac{df}{d\mathbf{x}}$
$bx \rightarrow b$	$\mathbf{x}^T \mathbf{B} \rightarrow \mathbf{B}$
$bx \rightarrow b$	$\mathbf{x}^T \mathbf{b} \rightarrow \mathbf{b}$
$x^2 \rightarrow 2x$	$\mathbf{x}^T \mathbf{x} \rightarrow 2\mathbf{x}$
$bx^2 \rightarrow 2bx$	$\mathbf{x}^T \mathbf{B} \mathbf{x} \rightarrow 2\mathbf{B} \mathbf{x}$

Shrinkage Properties

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

$$w \in \mathbb{R}^{d \times 1}$$

$$X \in \mathbb{R}^{n \times d}$$

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

What problem does this solve?

Matrix is always invertible

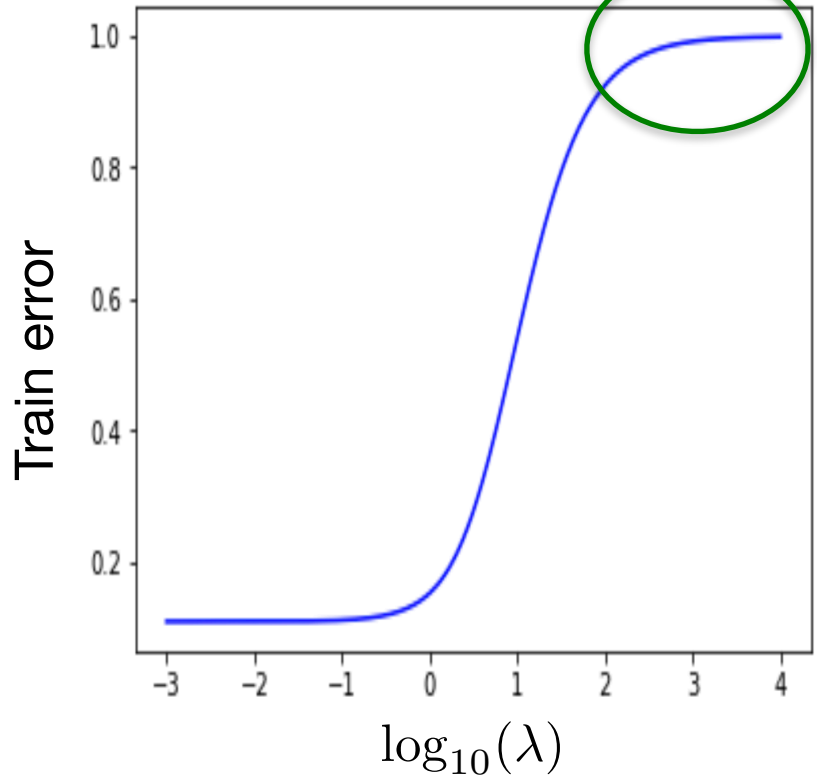
Large λ means “dividing by” larger term

- 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

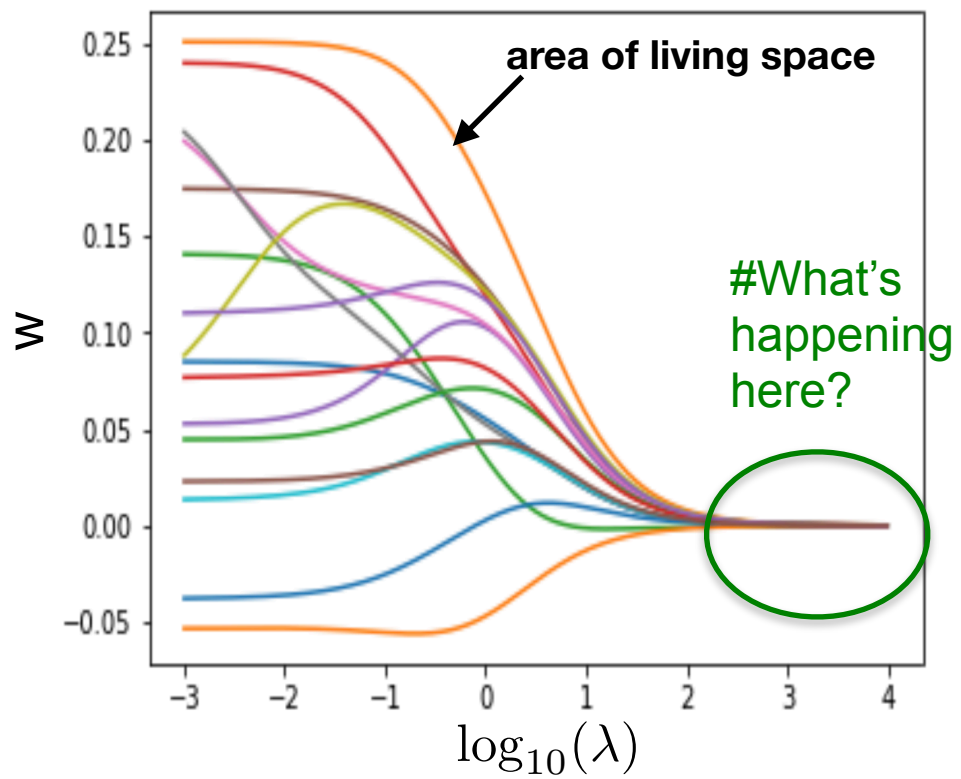
Optimal λ depends on magnitude of features

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

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



Housing price predictor w_i 's



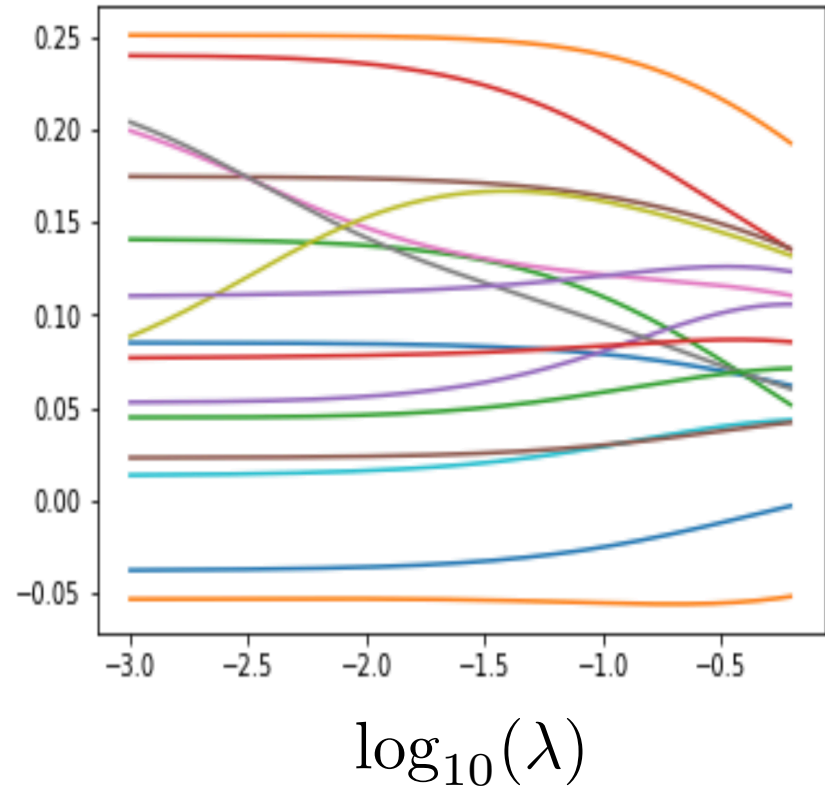
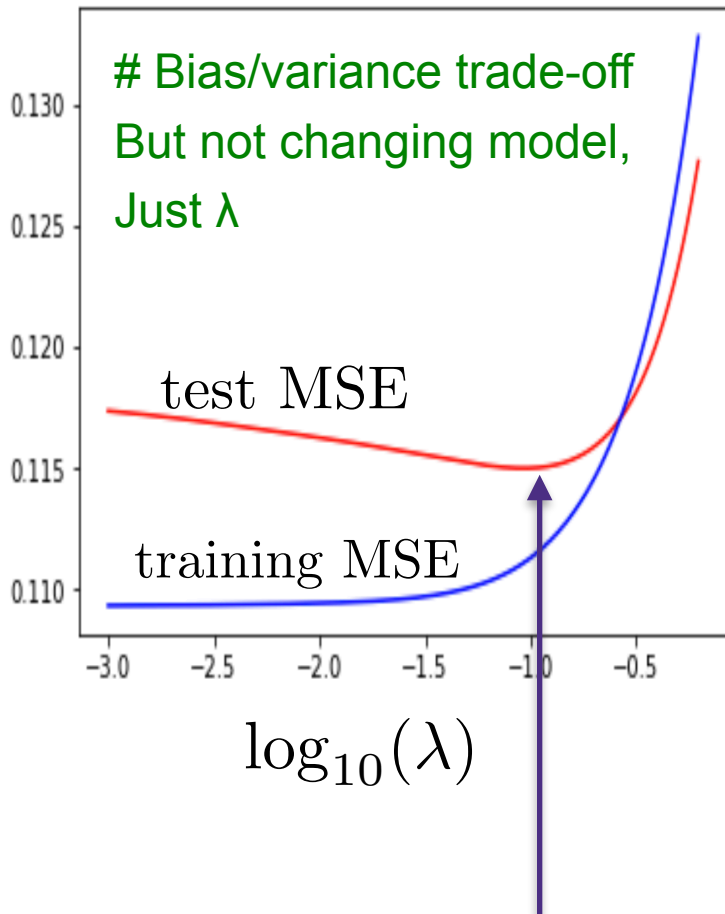
Right plot: called **regularization path**

Too much regularization introduces bias / underfitting

Best predictor (area of living space) shrinks slowest

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

Housing price predictor 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)

- this is the role of regularizer

Bias-Variance Properties

Plan: work through an example where we can directly put bias/variance in terms of λ

- 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: $x_i \sim P_X$, $\mathbf{y} = \mathbf{X}w + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ for some ground truth model parameter w
- The true error at a sample with feature x is $\mathbb{E}_{y, \mathcal{D}_{\text{train}} | x} [(y - x^T \hat{\mathbf{w}}_{\text{ridge}})^2 | x]$ # x is a test sample, $\hat{\mathbf{w}}_{\text{ridge}}$ is fit to a dataset

Bias-Variance Properties

Plan: work through an example where we can directly put bias/variance in terms of λ

- 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

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

Bias-Variance Properties

Plan: work through an example where we can directly put bias/variance in terms of λ

- 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

$$\begin{aligned} & \mathbb{E}_{y, \mathcal{D}_{\text{train}} | x} [(y - x^T \hat{w}_{\text{ridge}})^2 | x] \quad \# x \text{ is a test sample, } \hat{w}_{\text{ridge}} \text{ is fit to a dataset} \\ &= \mathbb{E}_{y|x} [(y - \mathbb{E}[y | x])^2 | x] + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(\mathbb{E}[y | x] - x^T \hat{w}_{\text{ridge}})^2 | x] \\ &= \mathbb{E}_{y|x} [(y - x^T w)^2 | x] + \mathbb{E}_{\mathcal{D}_{\text{train}}} [(x^T w - x^T \hat{w}_{\text{ridge}})^2 | x] \quad \# \text{ true } w \end{aligned}$$

Bias-Variance Properties

Plan: work through an example where we can directly put bias/variance in terms of λ

- 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

$$\mathbb{E}_{y, \mathcal{D}_{\text{train}} | x} [(y - x^T \hat{w}_{\text{ridge}})^2 | x] \quad \# x \text{ is a test sample, } \hat{w}_{\text{ridge}} \text{ is fit to a dataset}$$

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

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

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

Irreduc. Error

Bias-squared

Variance

Bias-Variance Properties

Plan: work through an example where we can directly put bias/variance in terms of λ

- 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

$$\mathbb{E}_{y, \mathcal{D}_{\text{train}} | x} [(y - x^T \hat{w}_{\text{ridge}})^2 | x] \quad \# x \text{ is a test sample, } \hat{w}_{\text{ridge}} \text{ is fit to a dataset}$$

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

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

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

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

Not realistic but pretend we're sampling independent Gaussians

$$= \frac{n}{n + \lambda} w + \frac{1}{n + \lambda} \mathbf{X}^T \epsilon \quad \# \lambda \text{ trades off weight on data vs. noise}$$

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

- λ trades off bias + variance
- Larger $\lambda \rightarrow$
 - Bigger bias
 - Smaller variance

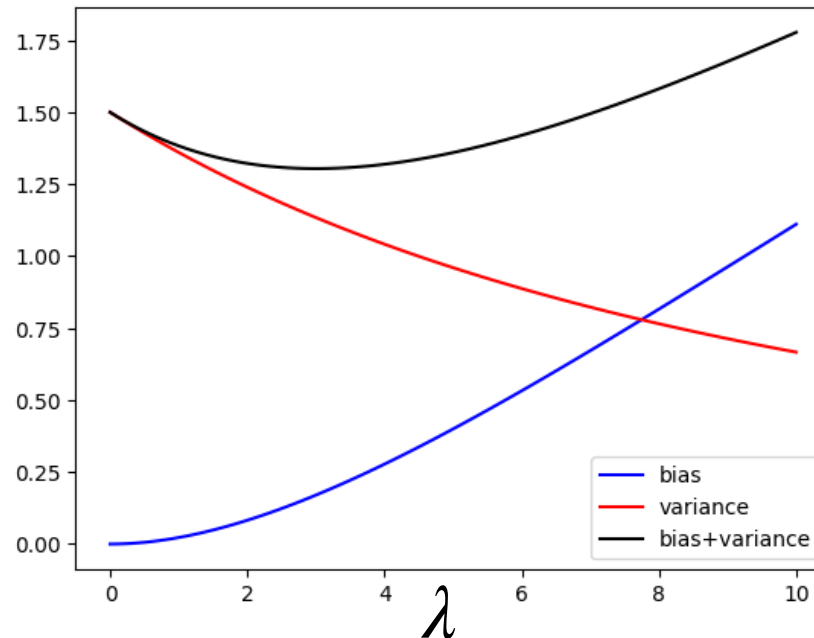
Bias-Variance Properties

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

- 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

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

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



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

- L₂ penalized least-squares regression $\lambda \|w\|_2^2$
- Regularization parameter trades off model complexity with training error
- Never regularize the offset!

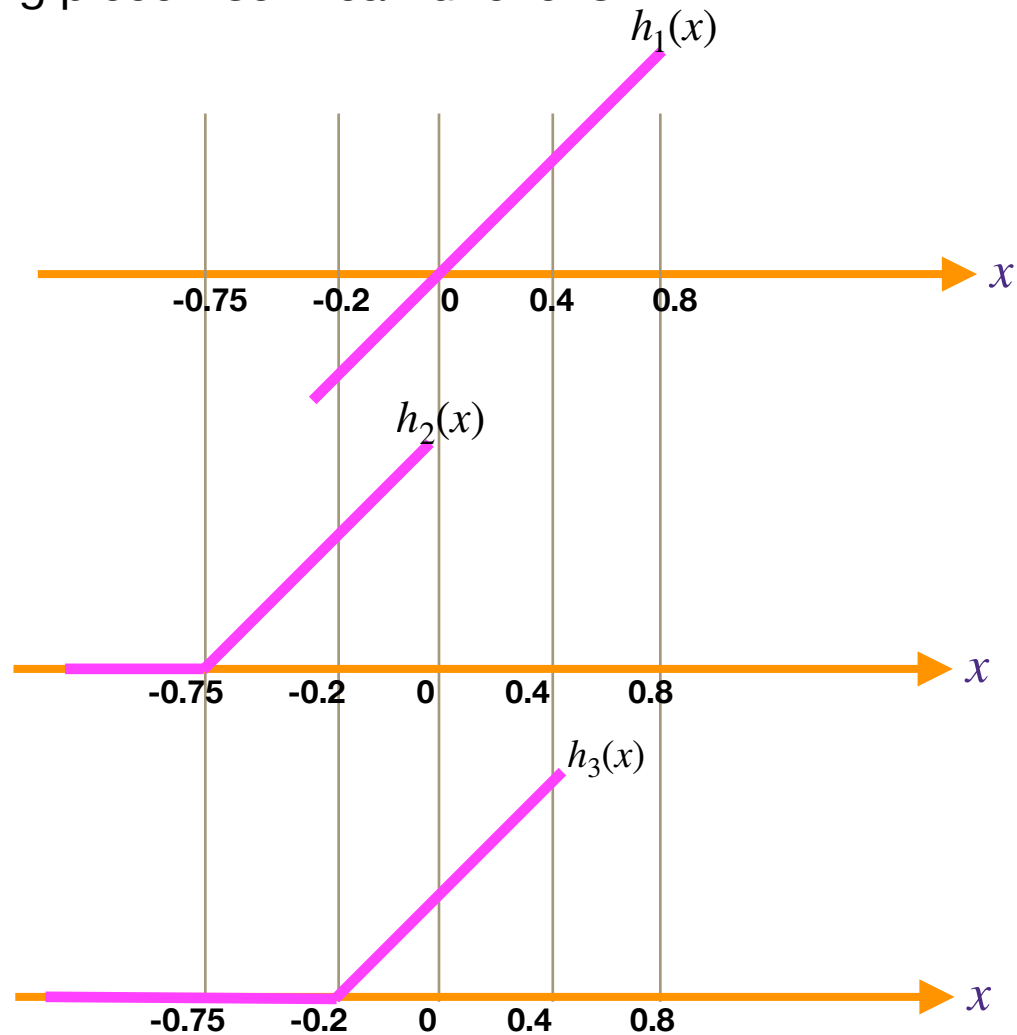
Example: piecewise linear fit

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

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

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

ReLU: rectified linear unit!

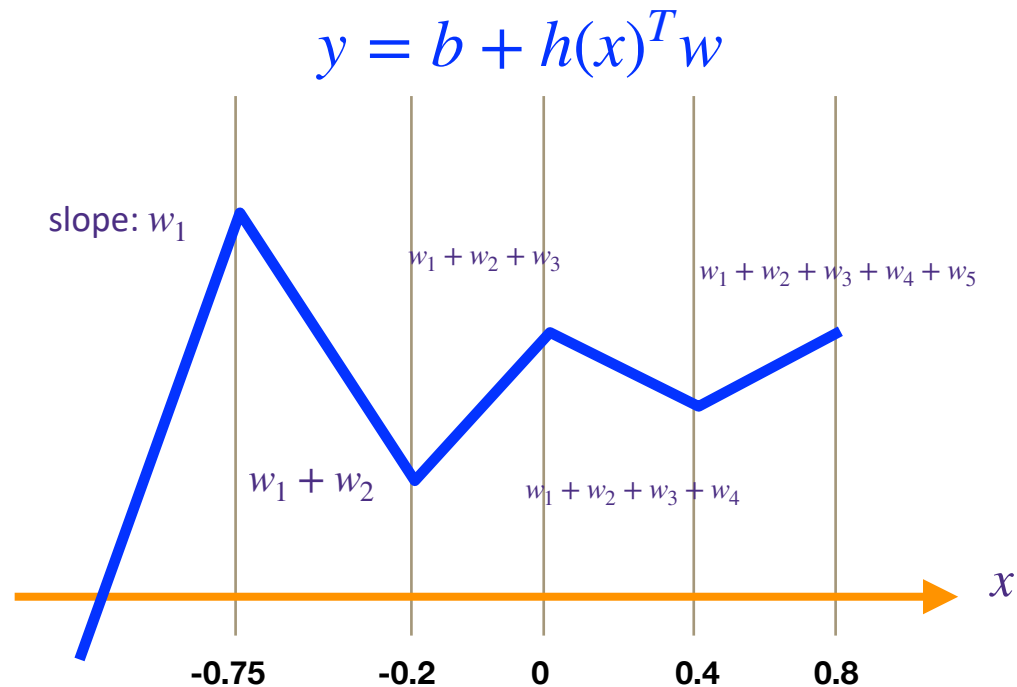


Example: piecewise linear fit

- 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

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

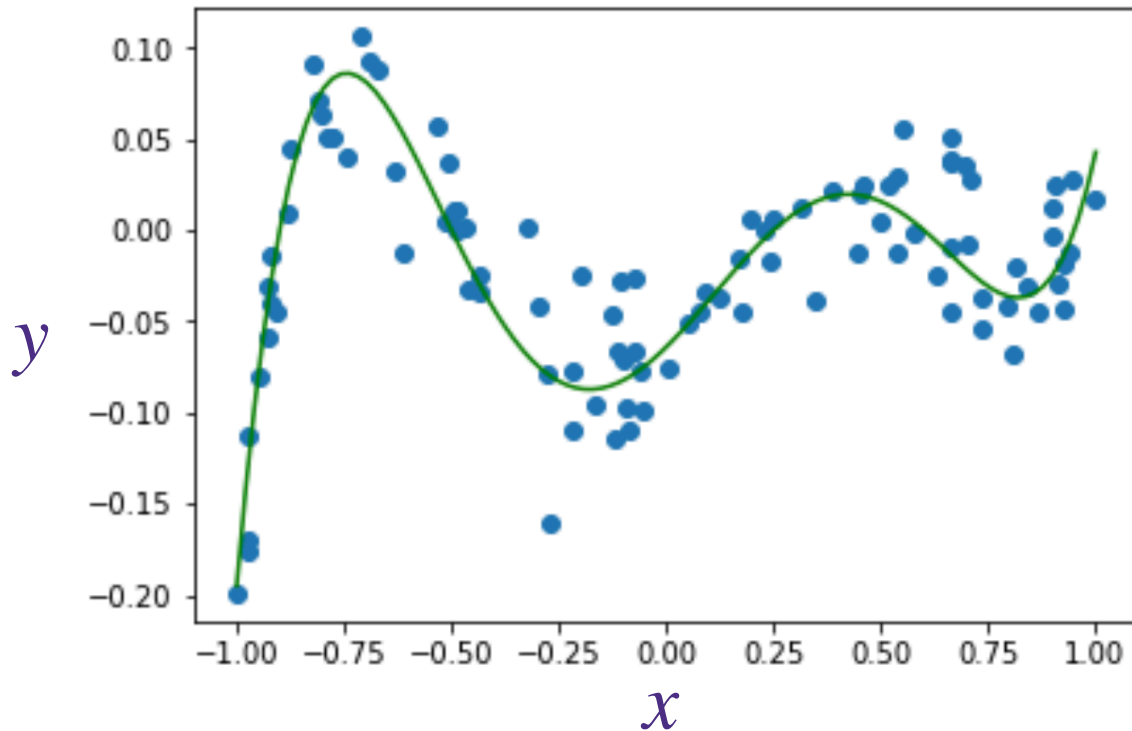
$$[a]^+ \triangleq \max\{a, 0\}$$



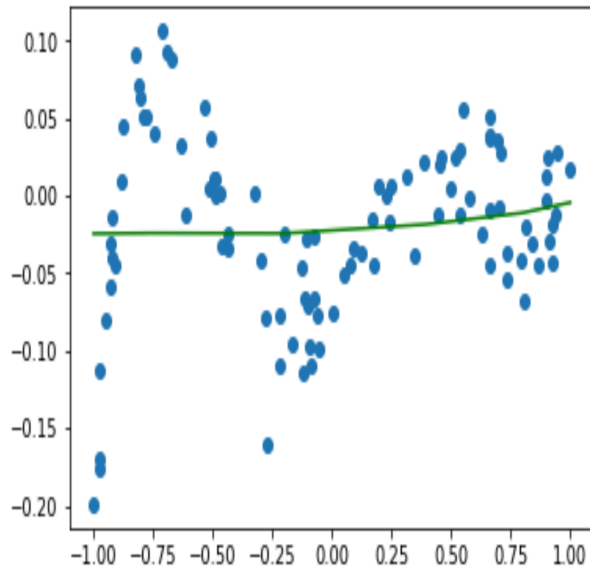
the weights capture the change in the slopes

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

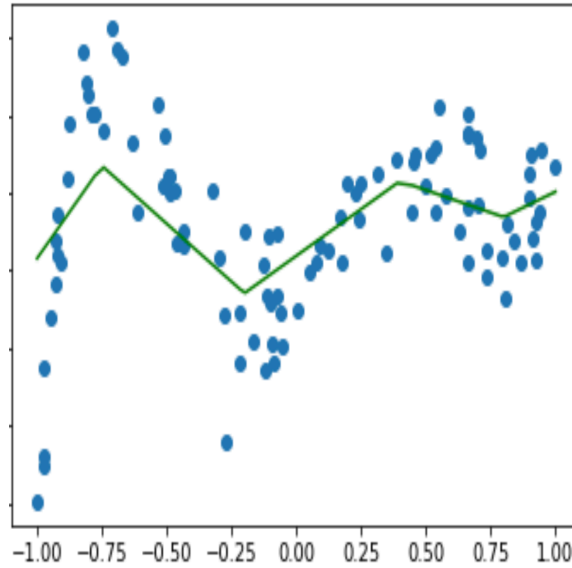


Example: piecewise linear fit (ridge regression)

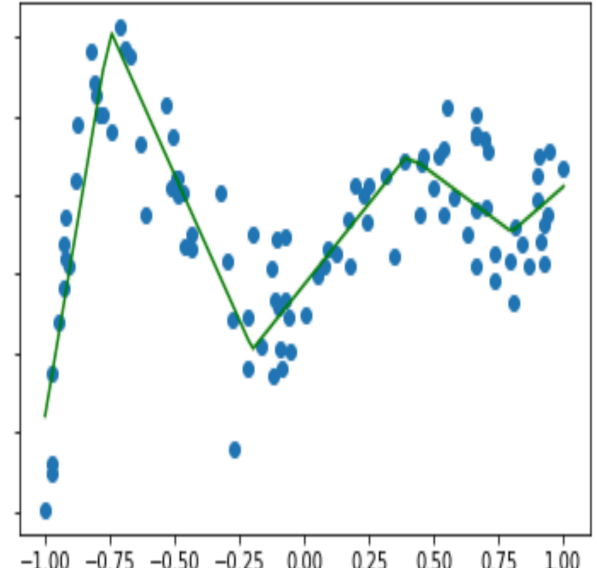


$\lambda = 1$

Large λ means
small weights / slopes
closer to a line



$\lambda = 0.005$



$\lambda = 0.000001$

Small λ means
big w / slopes

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

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

