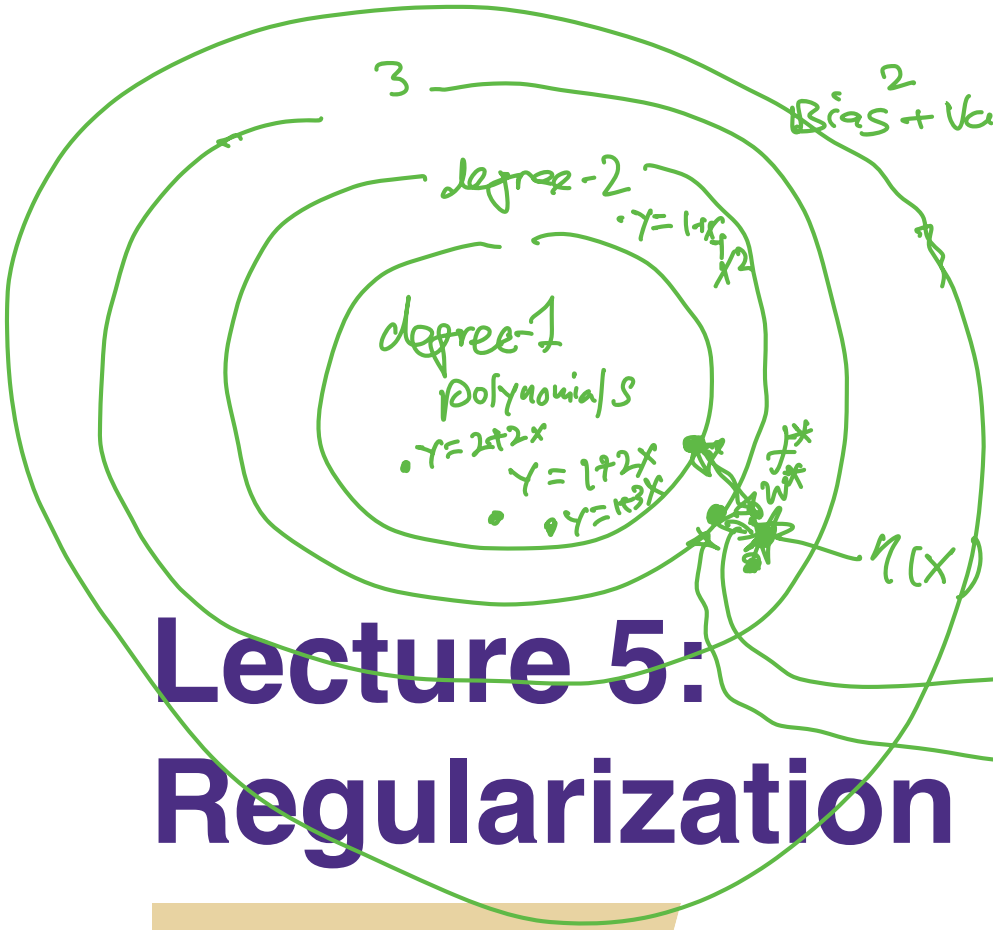
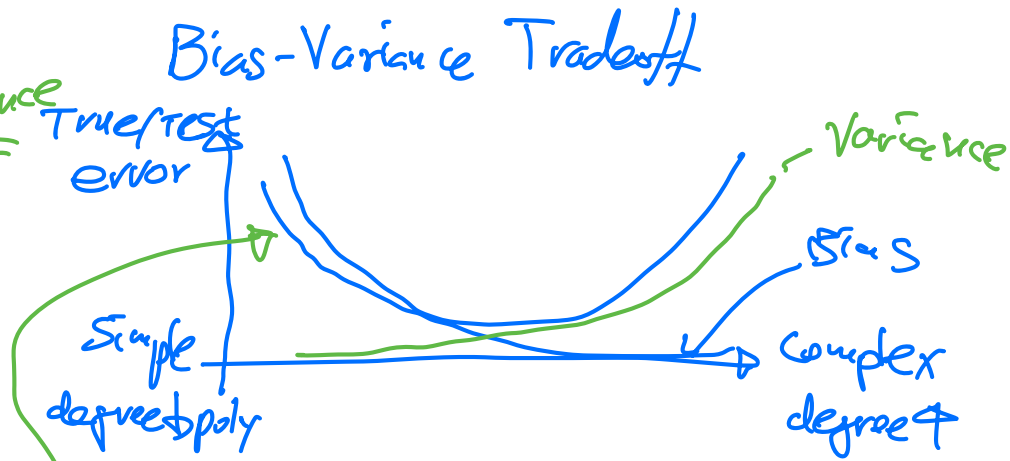




$$\text{Bias}^2 = \mathbb{E}\left[\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}(x)]\right)^2\right]$$

True Predictor — Example Predictor

$$\text{Error}^2 = \mathbb{E}\left[\left(\eta(x) - \hat{f}(x)\right)^2\right]$$

$2 + x + x^2$ — $2 + x$

$\mathbb{E}_x[\quad x^2]$

Lecture 5: Regularization

- how to avoid overfitting.

- Sewoong's OH today 2:00-3:00
- Pang Wei's office hour postponed (check on Ed)
- No exceptions on exam date/time.

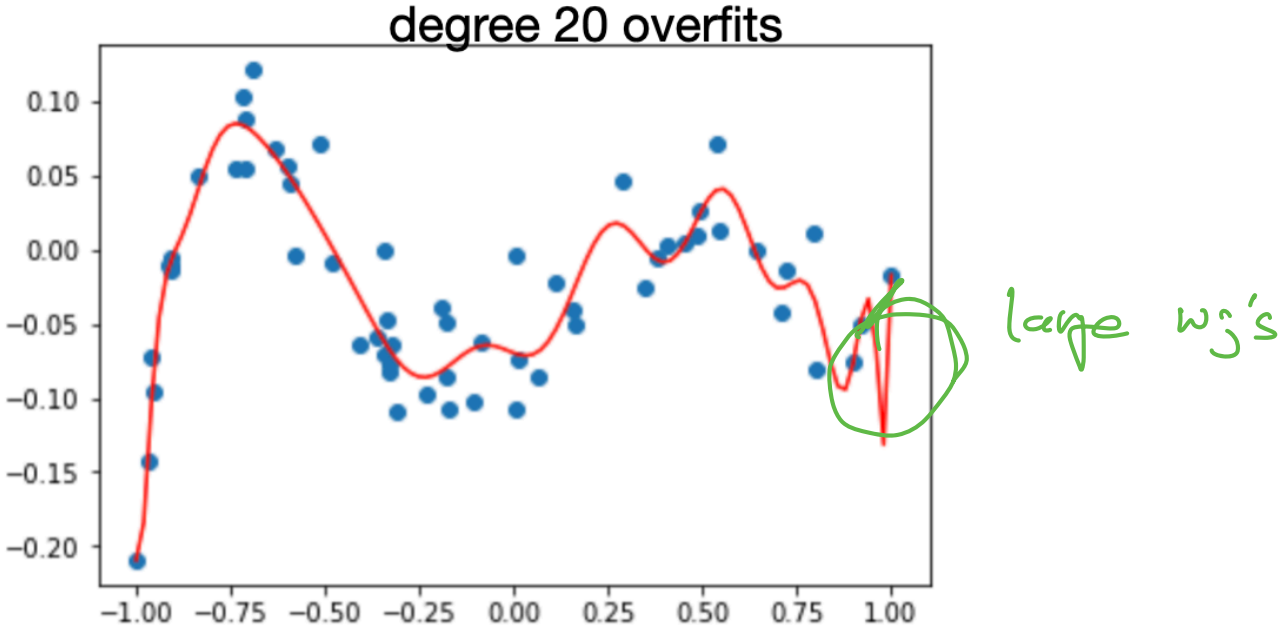
Regularization.

Ridge-regression

How to avoid overfitting

Sensitivity: how much prediction changes as we change the input

- 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



Sensitivity: how much prediction changes as we change the input

- 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
- And optimize for small loss and small sensitivity, by adding a ***regularizer*** in the objective (assume no offset for now)

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

regularizer

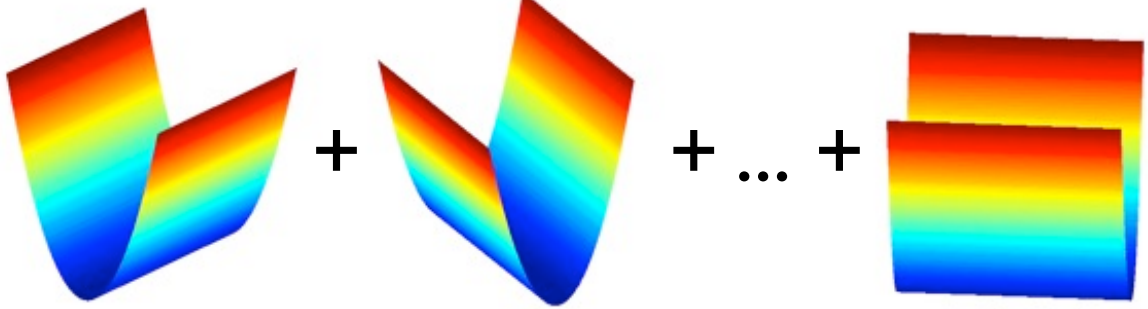
regularization coefficient

Ridge Regression

- (Original) Least squares objective:

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

$f_i(w)$



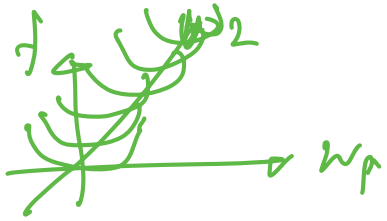
1-direction: strictly Quadratic

d-1-directions: flat.

e.g., $f(w_1, w_2) = (1 - [1,0]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix})^2 + (2 - [0,1]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix})^2$

$f_1(w_1,w_2)$ $f_2(w_1,w_2)$

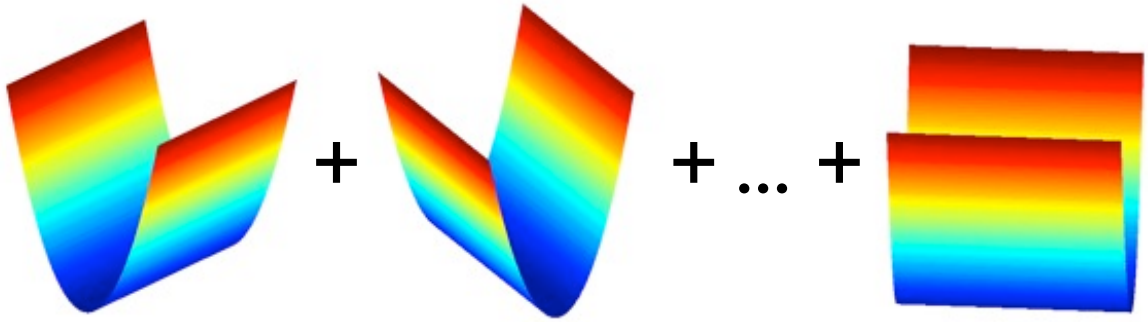
$= (1-w_1)^2$ $= (2-w_2)^2$



Ridge Regression

- (Original) Least squares objective:

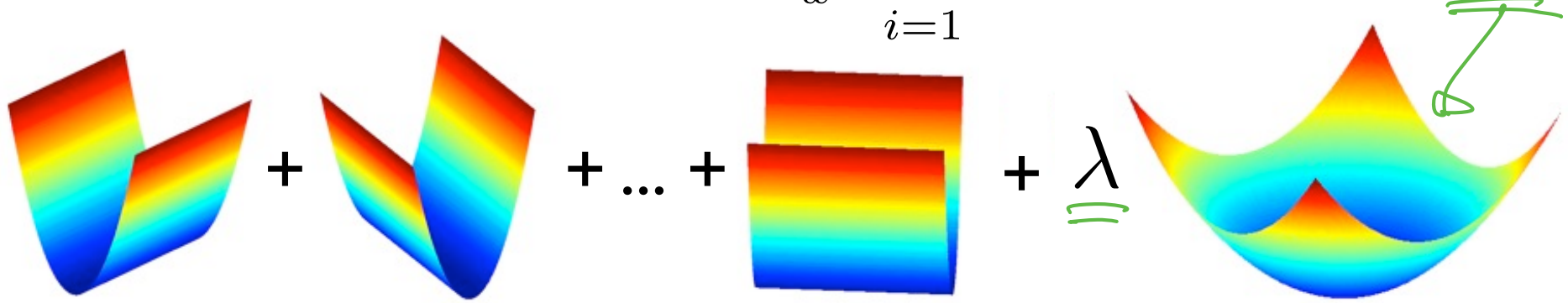
$$\hat{w}_{\text{MLE}} = \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2$$



- Ridge Regression objective:

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

$(w_1^2 + w_2^2 + \ldots)$



Minimizing the Ridge Regression Objective

$$\hat{w}_{\text{ridge}} = \arg\min_{w} \|y - Xw\|_2^2 + \lambda \|w\|_2^2 = w^T w$$

$$f(w) = (y - Xw)^T (y - Xw) + \lambda \cdot w^T \cdot I_{d \times d} \cdot w$$

$$\nabla_w f = -2X^T (y - Xw) + 2\lambda \cdot I_{d \times d} \cdot w = 0$$

$$X^T X w + \lambda \cdot I \cdot w = X^T y$$

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

added.

B is symmetric

Scalar derivative			vector gradient		
$f(x)$	$\rightarrow$	$\frac{df}{dx}$	$f(\mathbf{x})$	$\rightarrow$	$\nabla_x f(x)$
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	$\rightarrow$	$2bx$	$\mathbf{x}^T\mathbf{B}\mathbf{x}$	$\rightarrow$	$2\mathbf{B}\mathbf{x}$

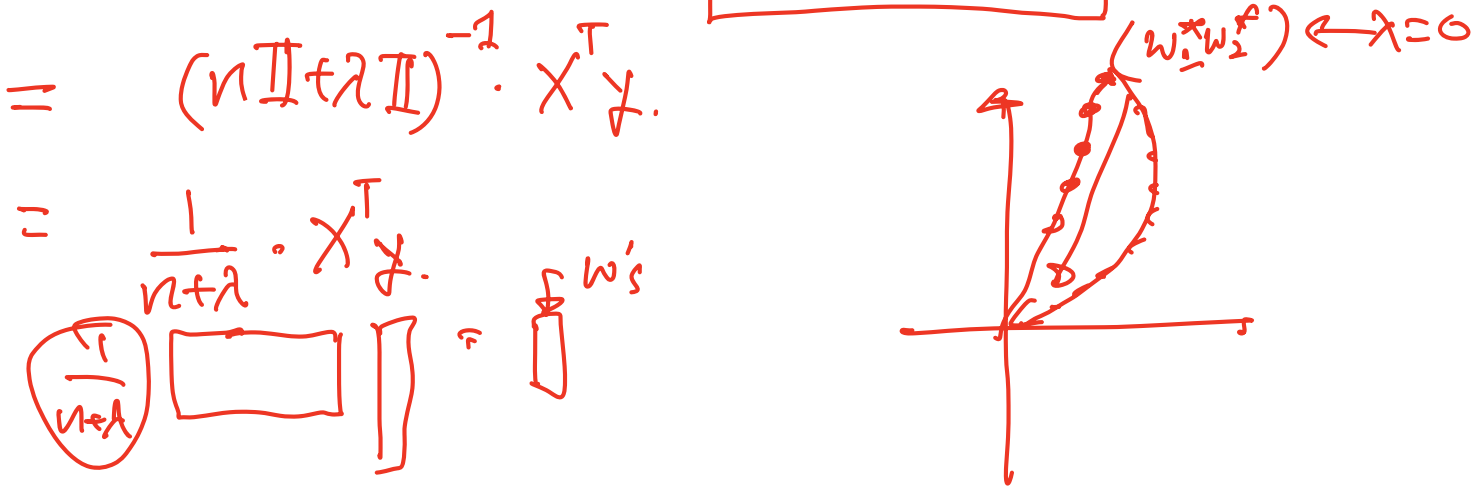
Primer.

Shrinkage Properties

$$\hat{w}_{\text{ridge}} = \arg\min_{w} \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

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

To get some intuition, suppose input X satisfies $X^T X = n\mathbf{I}_{d\times d}$,

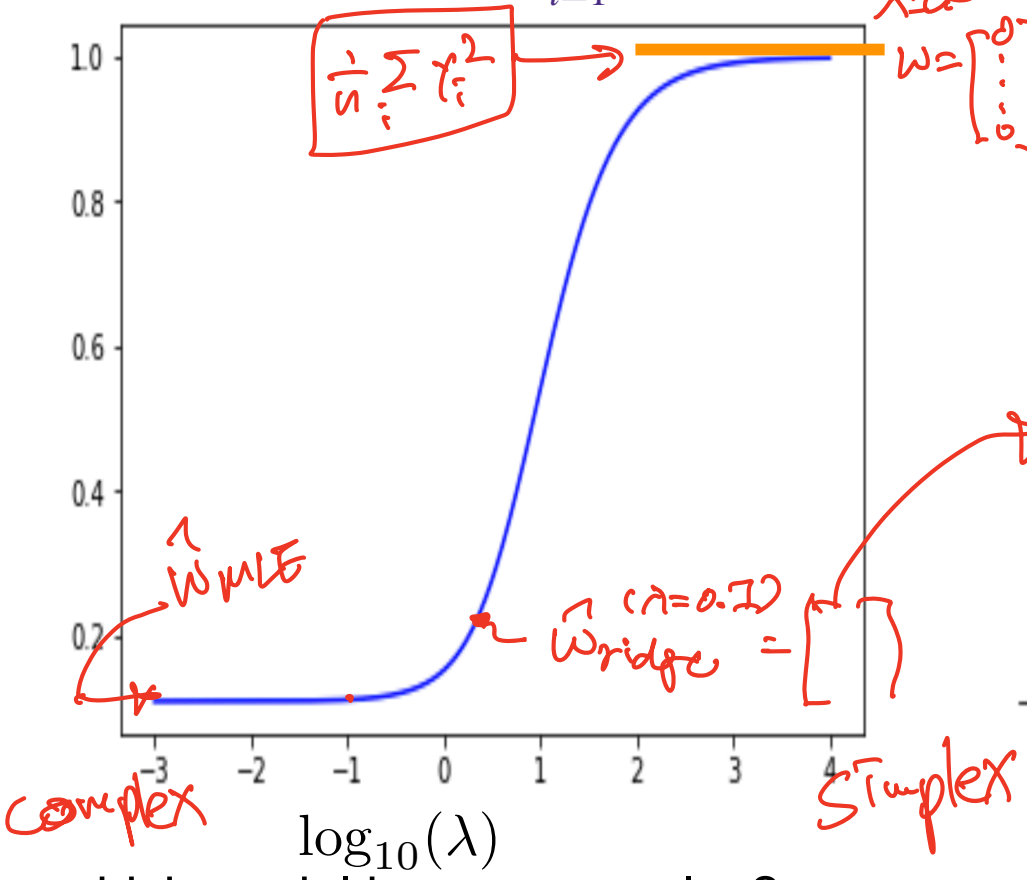


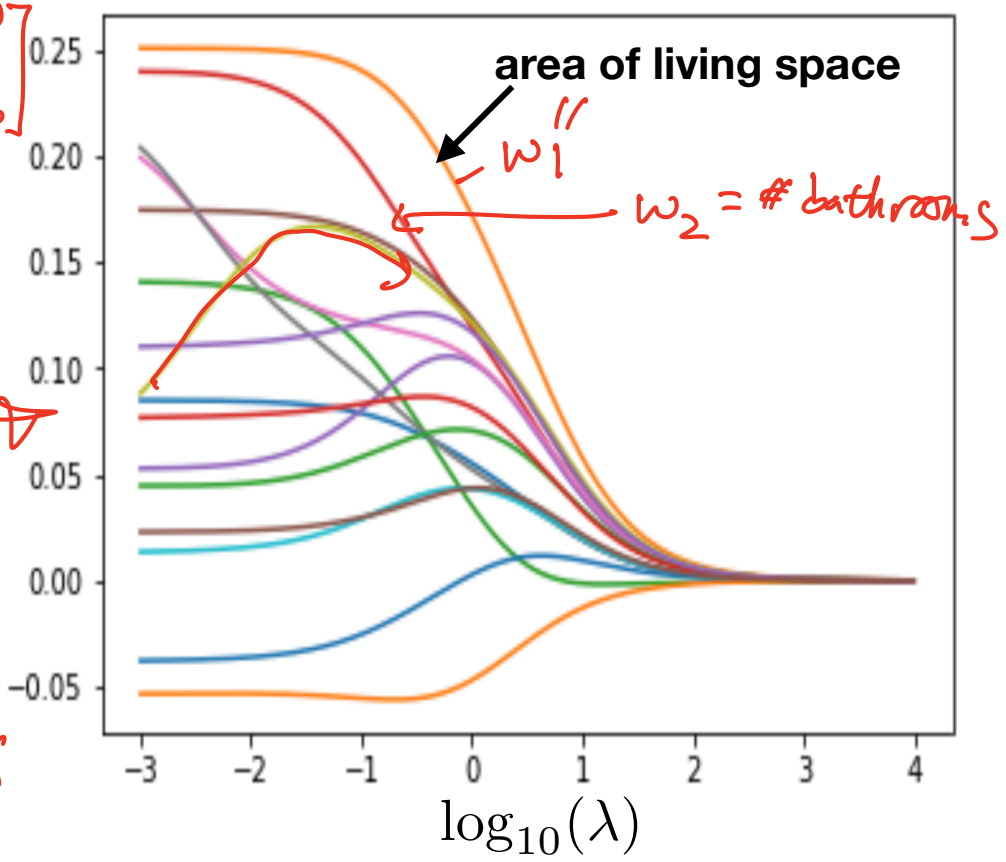
- When $\lambda = 0$, this recovers the MLE estimate, as a special case
- This defines a family of models hyper-parametrized by λ
- Large λ means more regularization and simpler model
- Small λ means less regularization and more complex model

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

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

Housing price predictor $w_i's$

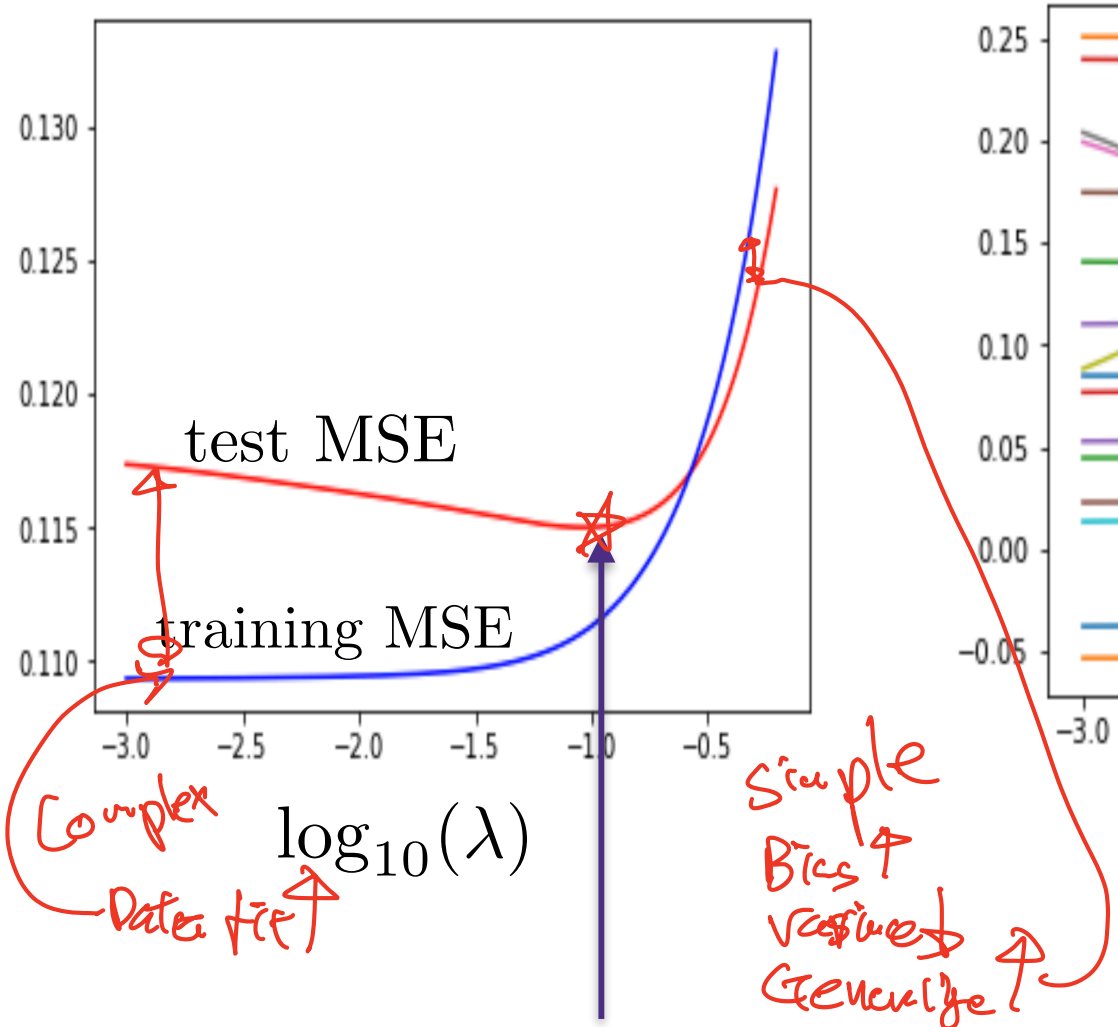




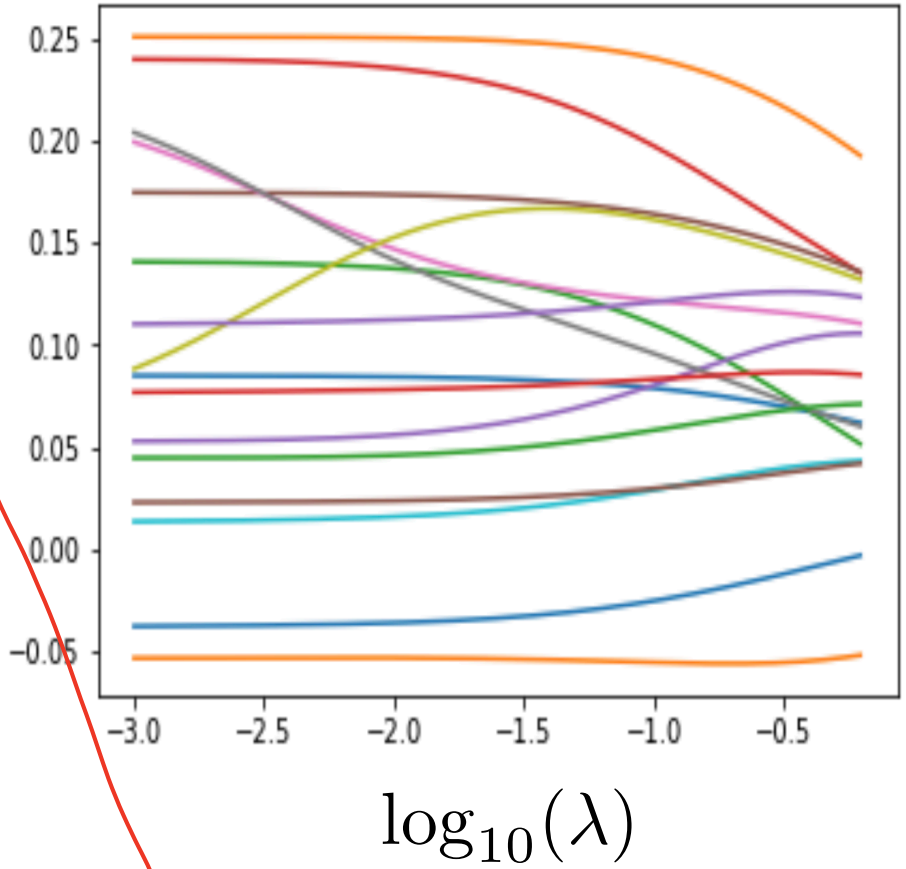
which model is more complex?

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



Housing price predictor $w_i's$



- as we increase λ, 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

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

$y_i = x_i^T w + \varepsilon_i$ ← Train sample
$y = x^T w + \varepsilon$ ← new sample
$y = w^T x = \eta(x)$

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

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

irreducible error

learning error

$(x, y) \sim P_{XY}$

$= \mathbb{E}[(x^T w + \varepsilon - x^T w)^2]$

$= \sigma^2$

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

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

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

$\mathbb{E}\left[\left(h(x) - \mathbb{E}[x^T\hat{w}_{\text{ridge}}|x]\right)^2|x\right]$ + $\mathbb{E}\left[\left(\mathbb{E}[x^T\hat{w}_{\text{ridge}}|x] - x^T\hat{w}_{\text{ridge}}\right)^2|x\right]$

Bias^2 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

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

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

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

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

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

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

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

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

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

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

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

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

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

Irreduc. Error — Bias-squared — Variance

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

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

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: $x_i \sim P_X, \quad \mathbf{y} = \mathbf{X}w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$
- The true error at a sample with feature x is

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

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

(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

n, λ

The missing calculation from previous slide

Claim: $(x^T w - \mathbb{E}_{\mathcal{D}_{\text{train}}}[x^T \hat{w}_{\text{ridge}} \,|\, x])^2 = \frac{\lambda^2}{(n+\lambda)^2}(w^T x)^2$

proof: $(x^T w - \mathbb{E}_{\mathcal{D}_{\text{train}}}[x^T \hat{w}_{\text{ridge}} \,|\, x])^2 = \left(x^T w - \mathbb{E}\left[\frac{n}{n+\lambda} x^T w + \frac{1}{n+\lambda} x^T X \epsilon\right]\right)^2$

using $\mathbb{E}[\epsilon] = 0$

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

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

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

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

The missing calculation from previous slide

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

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

$$= \mathbb{E}\left[\left(\mathbb{E}\left[\frac{n}{n+\lambda}x^T w + \frac{1}{n+\lambda}x^T X^T \epsilon\right] - x^T\hat{w}_{\text{ridge}}\right)^2\right]$$

using $\mathbb{E}[\epsilon] = 0$

$$= \mathbb{E}\left[\left(\frac{n}{n+\lambda}x^T w - x^T\hat{w}_{\text{ridge}}\right)^2\right]$$

$$= \mathbb{E}\left[\left(\frac{1}{n+\lambda}x^T X^T \epsilon\right)^2\right]$$

$$= \frac{1}{(n+\lambda)^2}\mathbb{E}\left[x^T X^T \epsilon\epsilon^T X x\right]$$

using $\mathbb{E}[\epsilon\epsilon^T] = \sigma^2\mathbf{I}$.

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

using $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$.

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

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

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

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

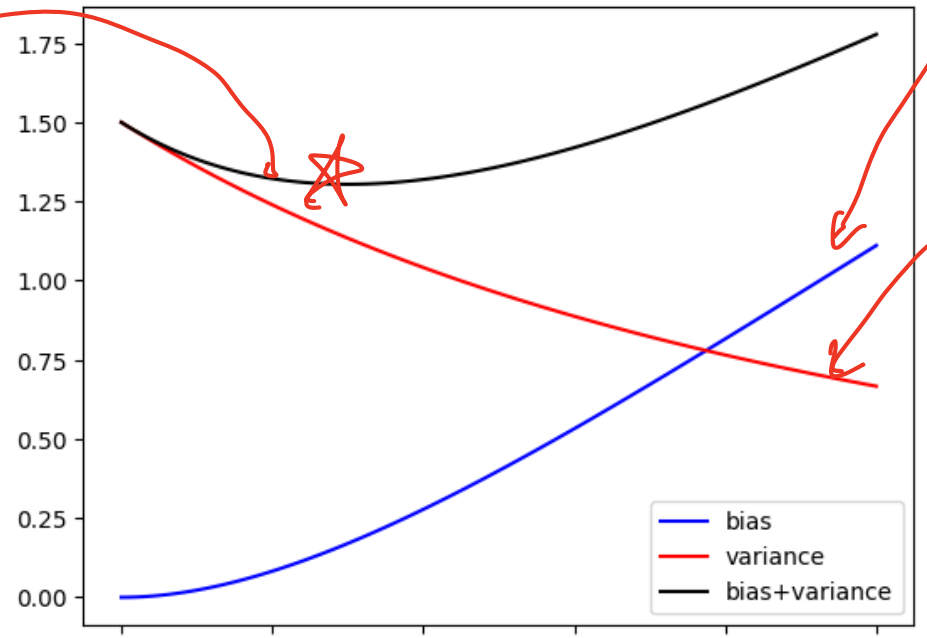
Bias-Variance Properties

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

- Ridge regressor: $\hat{w}_{\text{ridge}} = \arg\min_w \|y - Xw\|_2^2 + \lambda\|w\|_2^2$
- True error

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

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



Variance ↑

as $\lambda \to 0$,

$\hat{w}_{\text{ridge}} \to \hat{w}_{\text{MLE}}$

Bias ↑

as $\lambda \to \infty$

$\hat{w}_{\text{ridge}} \to 0$

What you need to know…

LASSO

$||w||_2^2$ $||w||_1$

- 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 b, because b does not contribute to sensitivity of the input.

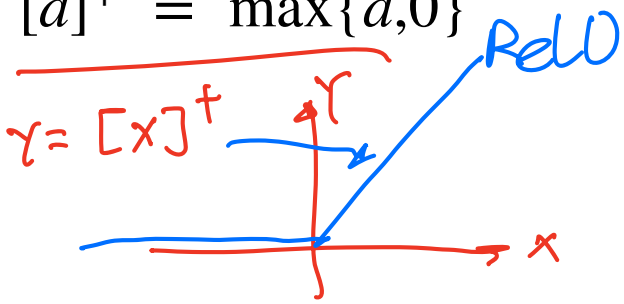
[Homework 1 Problem A5 implements Ridge regression on MNIST.]

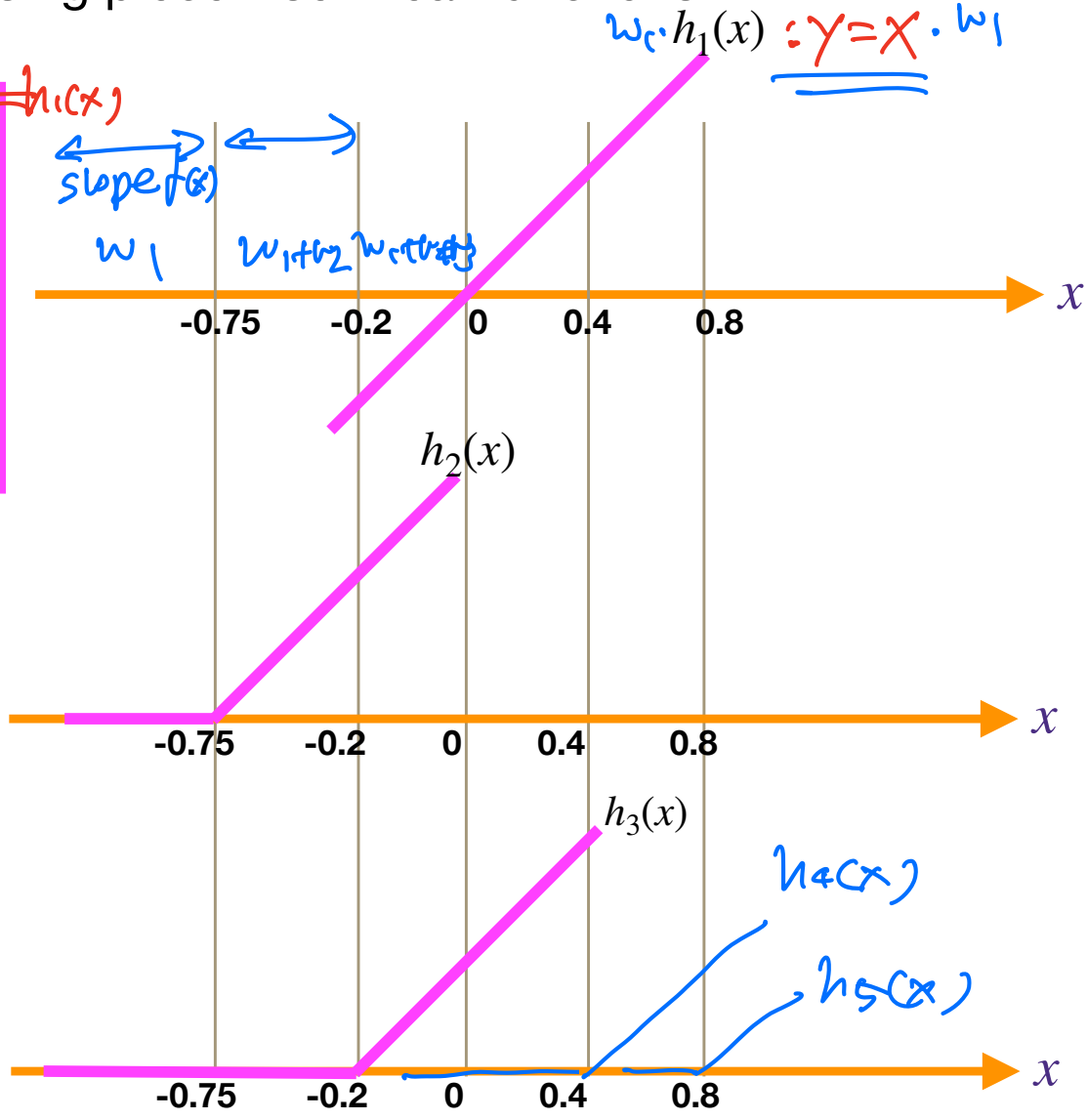
Example: piecewise linear fit

- 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

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

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



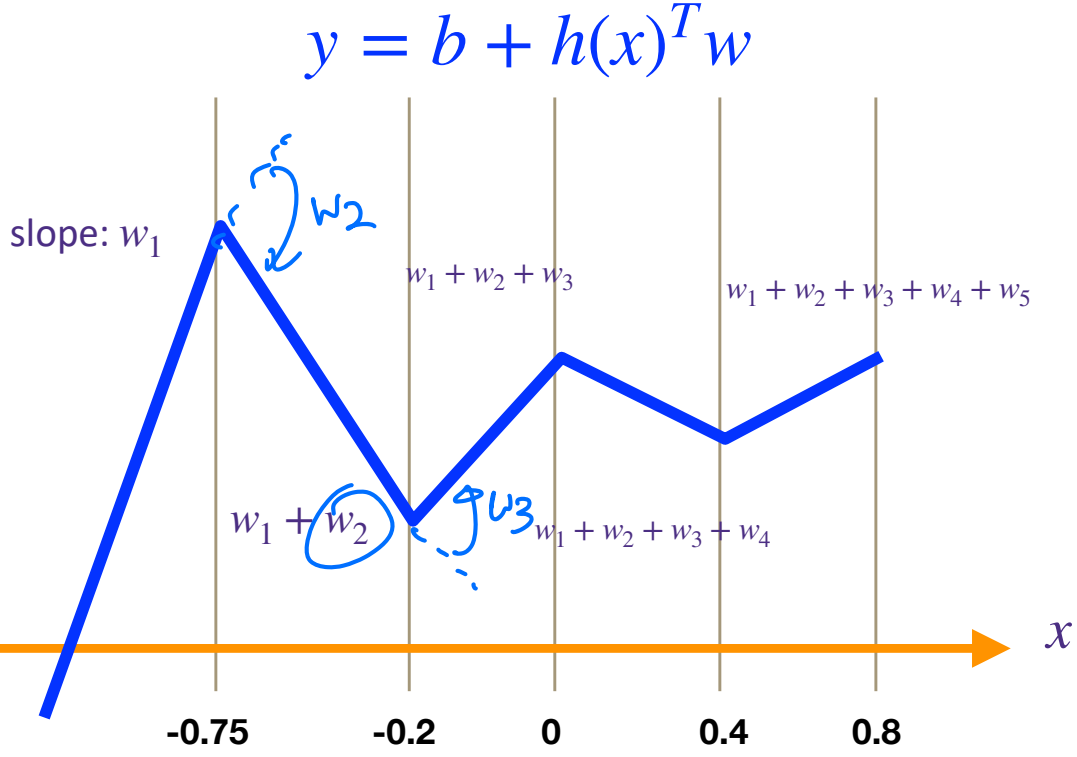


Example: piecewise linear fit

- 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

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

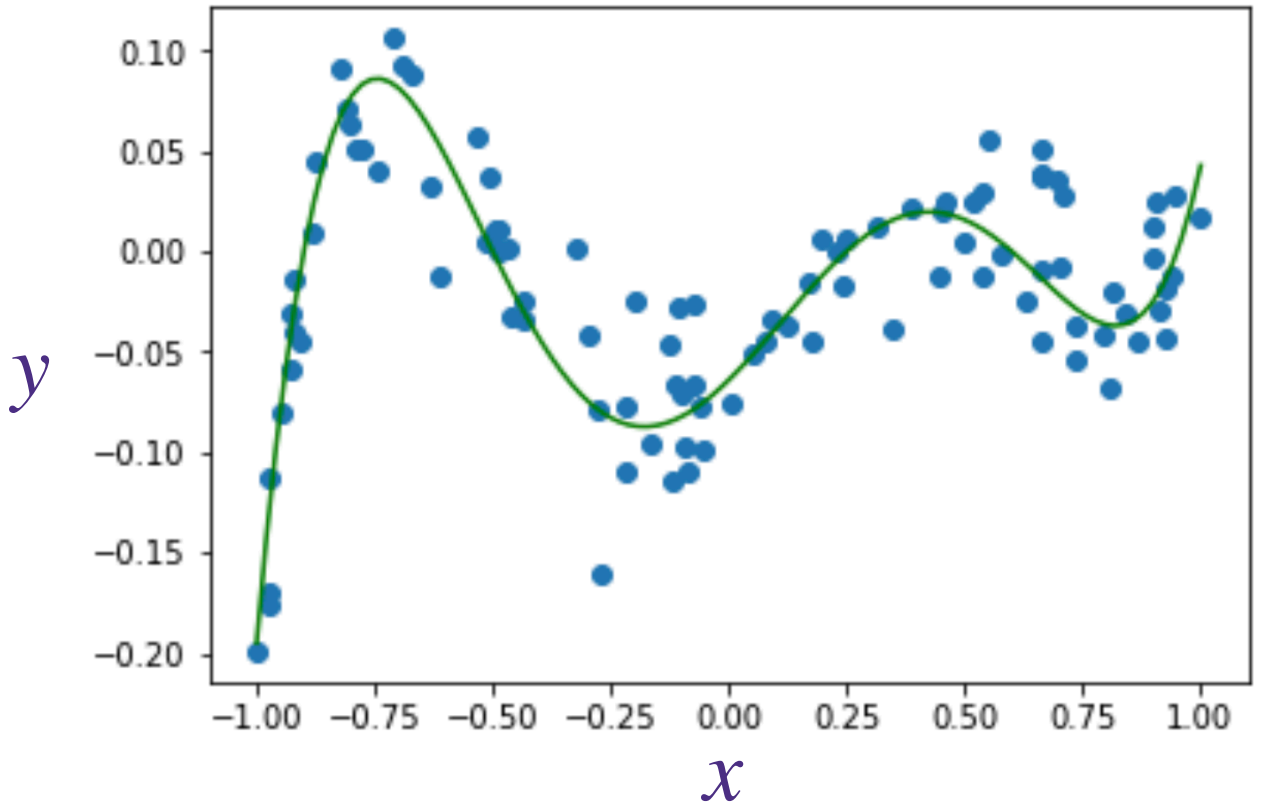
$[a]^+ \triangleq \max\{a, 0\}$



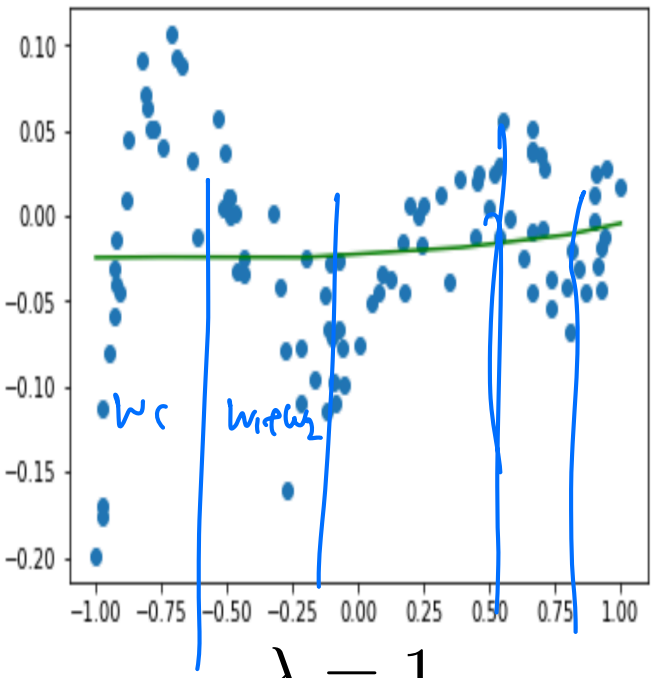
the weights capture the change in the slopes

Example: piecewise linear fit

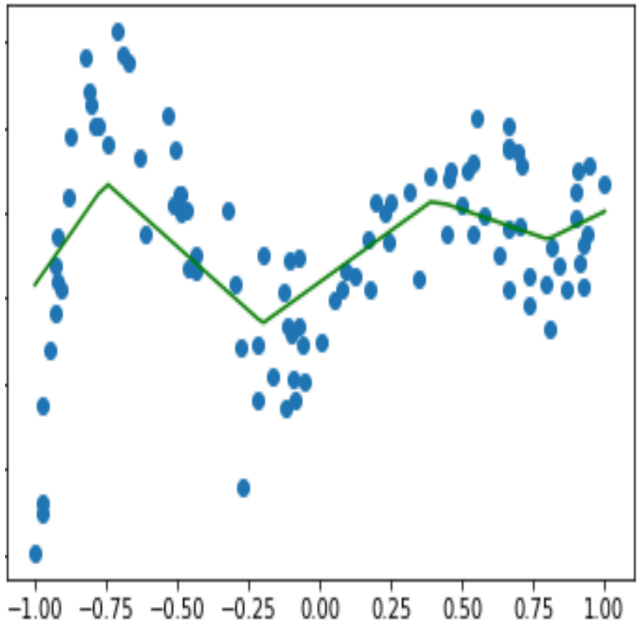
- we fit a linear model:
  $f(x) = b + w_1 h_1(x) + w_2 h_2(x) + w_3 h_3(x) + w_4 h_4(x) + w_5 h_5(x)$
- with a specific choice of features using piecewise linear functions



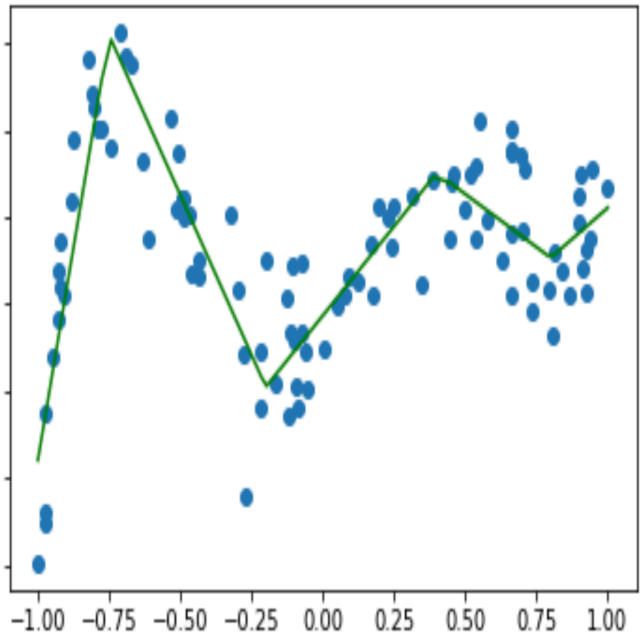
Example: piecewise linear fit (ridge regression)



$\lambda = 1$



$\lambda = 0.005$



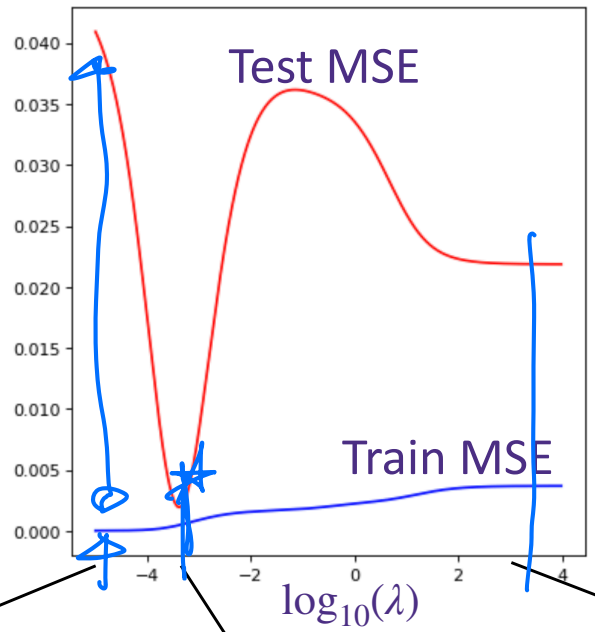
$\lambda = 0.000001$

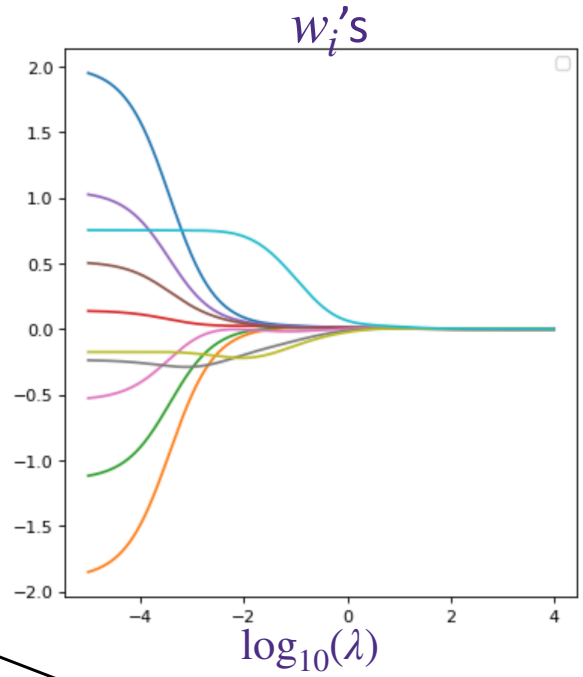
$\lambda \approx 0$

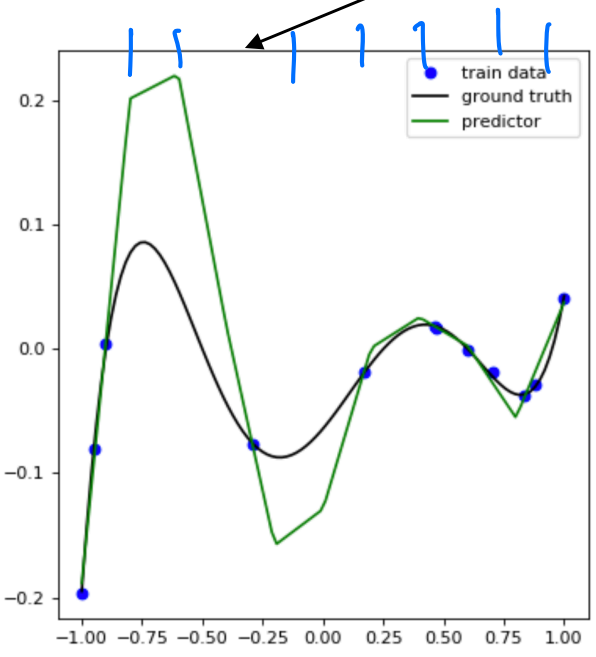
We do not observe overfitting, as d=5 << n=100

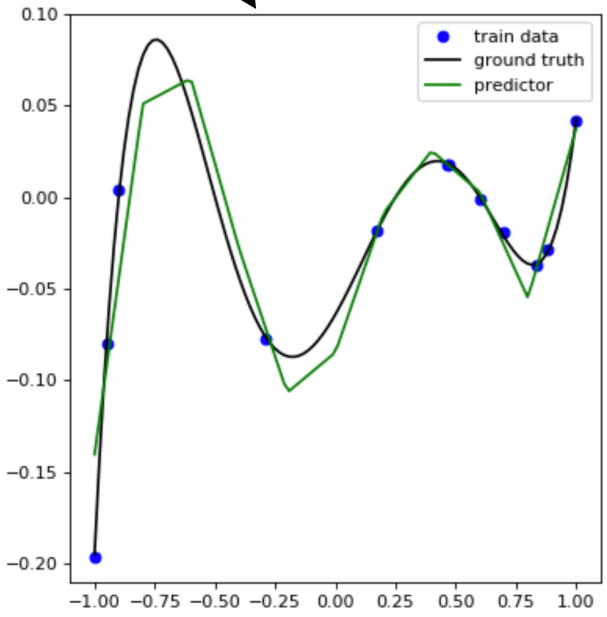
Piecewise linear with $w \in \mathbb{R}^{10}$ and n=11 samples

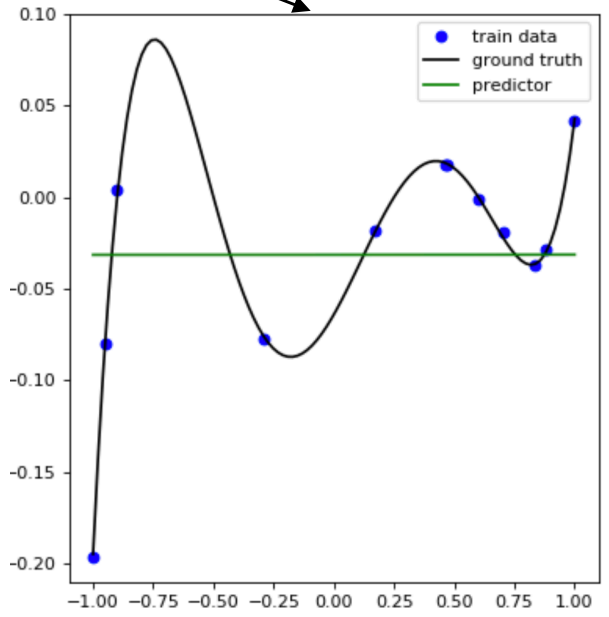
When we only have $n = 11$ samples and $10 = d$ dimensional features, we do need regularization to mitigate overfitting,









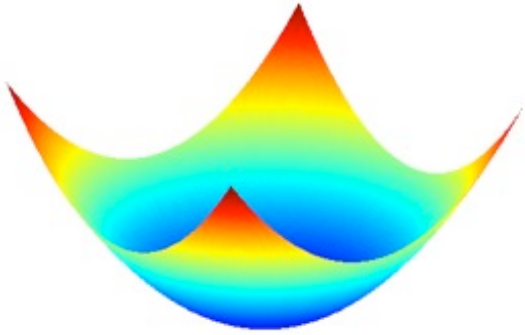


What do we do if d<n?

$$\hat{w}_{\text{MLE}} = \arg\min_{w} \|y - X^T w\|^2$$

$$\hat{w}_{\text{MLE}} = (X^T X)^{-1} X^T y$$

d<n → not invertible



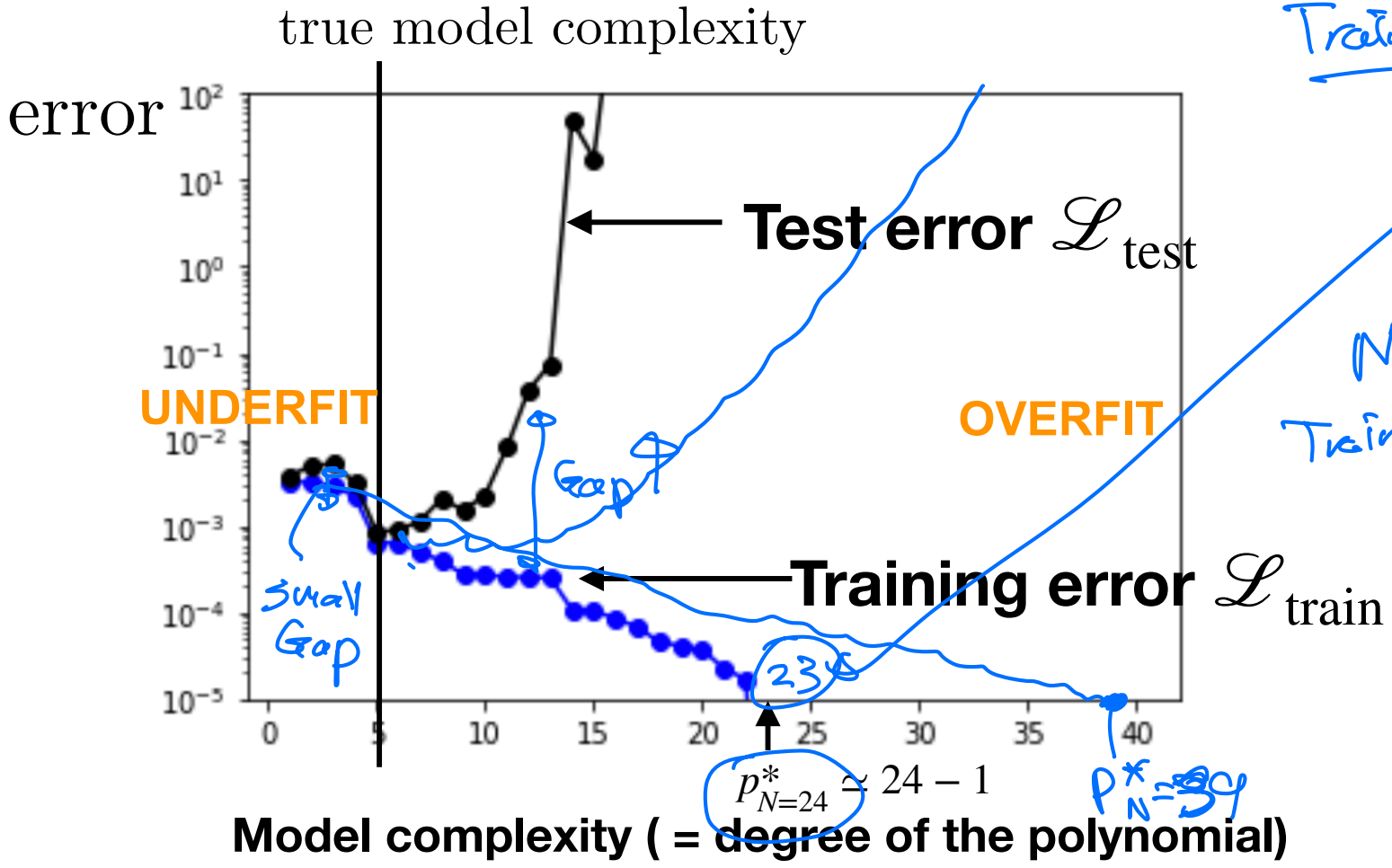
$$\hat{w}_{\text{Ridge}} = \arg\min_{w} \|y - X^T w\|^2 + \lambda \|w\|^2$$

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

unique

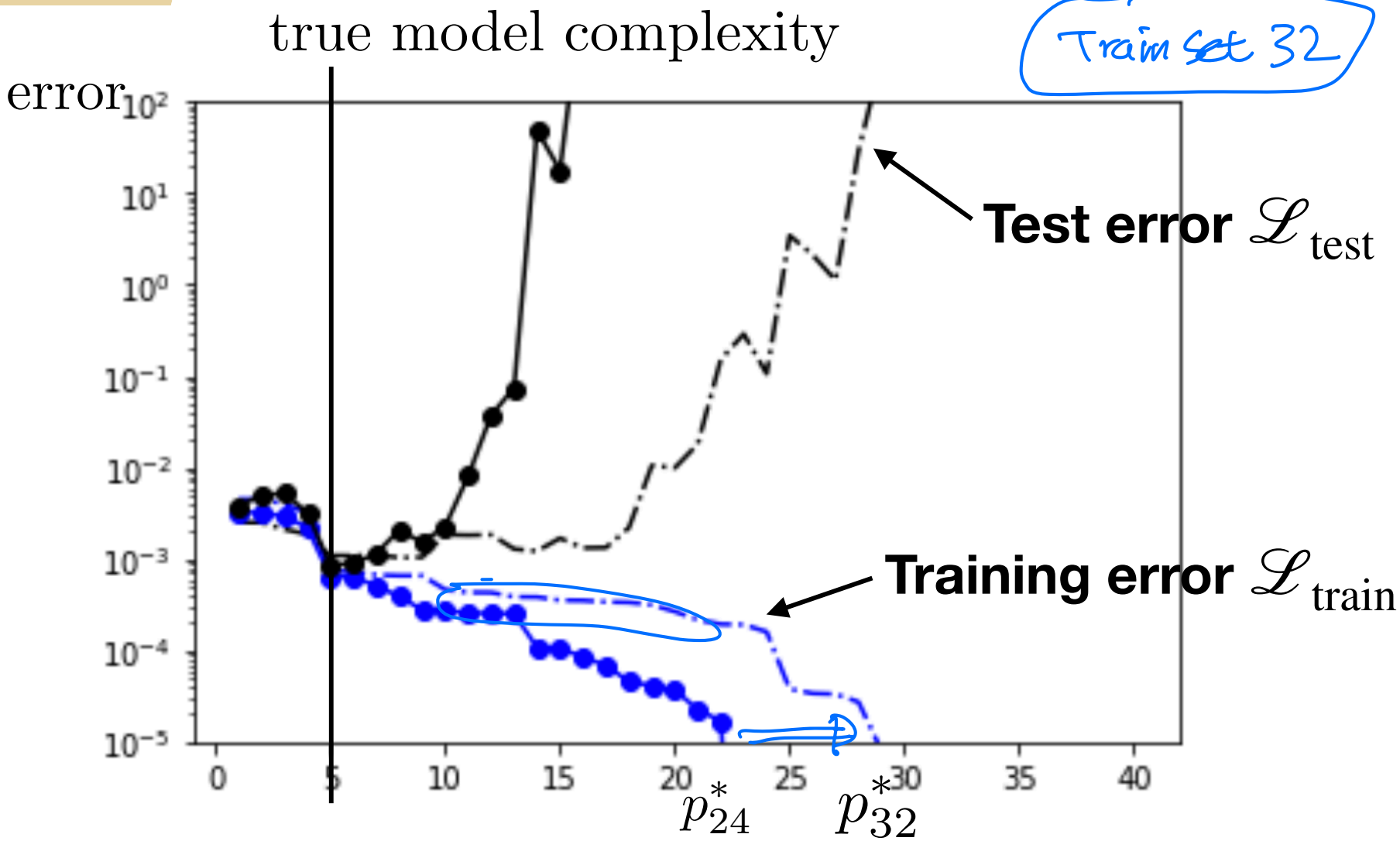
invertible

- let us first fix sample size ***N=30***, collect one dataset of size N i.i.d. from a distribution, and fix one training set S_{train} and test set S_{test} via 80/20 split
- then we run multiple validations and plot the computed MSEs for all values of ***p*** that we are interested in



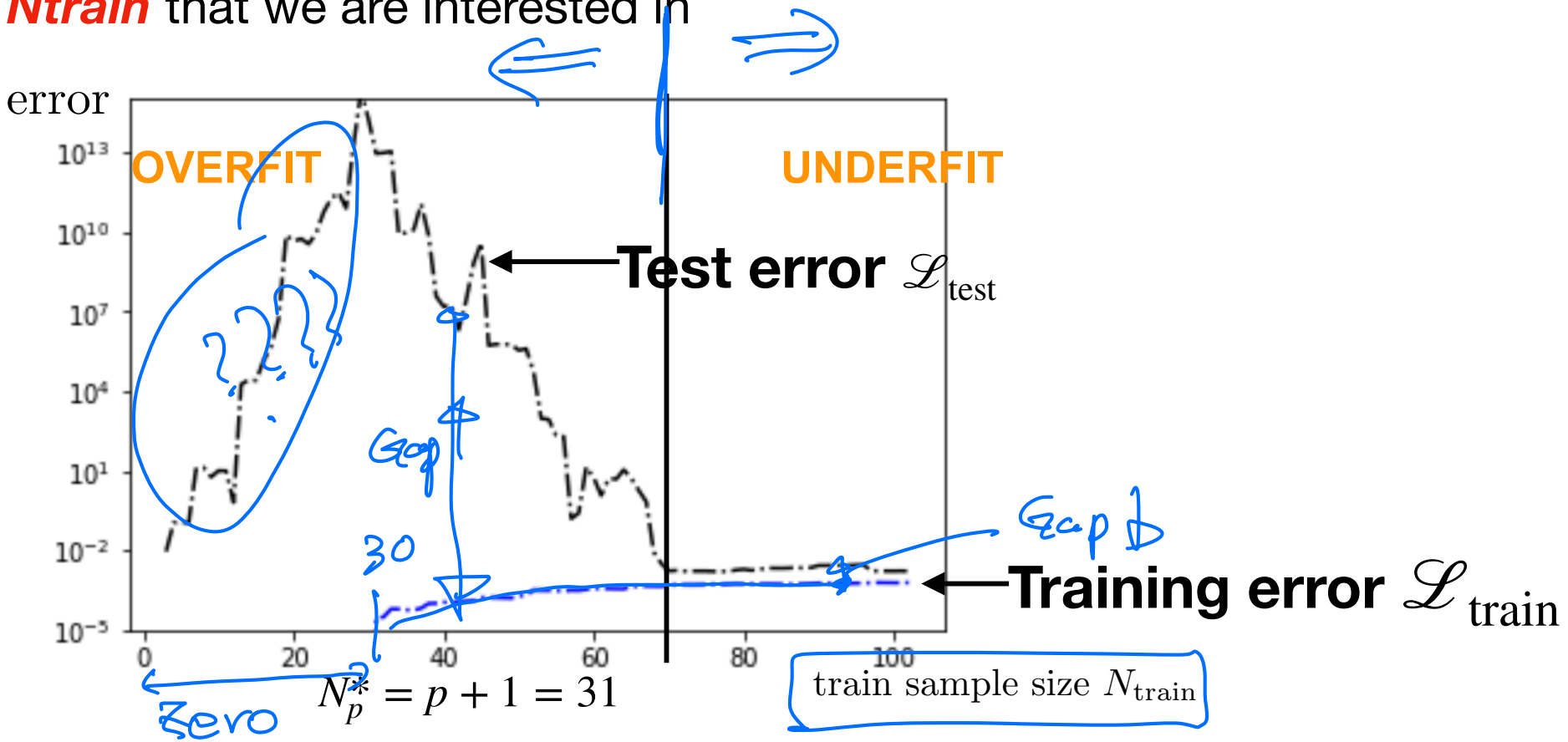
- Given sample size N there is a threshold, p^*_N, where training error is zero
- Training error is **always** monotonically non-increasing
- Test error has a trend of going down and then up, but fluctuates

- let us now repeat the process changing the sample size to **N=40** , and see how the curves change



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

- let us now fix predictor model complexity ***p=30***, collect multiple datasets by starting with 3 samples and adding one sample at a time to the training set, but keeping a large enough test set fixed
- then we plot the computed MSEs for all values of train sample size ***Ntrain*** that we are interested in



- There is a threshold, N_p^*, below which training error is zero (extreme overfit)
- Below this threshold, test error is meaningless, as we are overfitting and there are multiple predictors with zero training error some of which have very large test error
- Test error tends to decrease
- Training error tends to increase

Questions?

- Good questions on Ed Discussion
  - Will we be tested in "bias"?
    - Bias shows up in many places, and you will have to know the concept. Anything that is taught in lectures can show up in the exams.
  - Why do we use x to denote a column vector and not a row vector?
  - θ^* is the same as θ_*, it is just my writing that is not always consistent
  - What is θ^*?
    - The reason it is unnatural to think about θ^* is that it is something that does not exist in reality. Only time it exists is when you generate simulated data yourself (like in lecture notes and homework).
    - The right interpretation is that we hypothesize that nature has chosen to generate the data from a distribution, which can be written as $P(\cdot\,; \theta^*)$.
    - Whether this assumption is correct or not, we are deciding to go ahead with our MLE process.
    - That gives us some MLE estimate and corresponding distribution $P(\cdot\,; \theta^*_{\text{MLE}})$. What we do with it, and what we believe about it is up to us. (Hence you need to check your accuracy on a holdout set, which we will learn later)

[Homework 1 Problem A4 analyzes similar bias-variance tradeoffs]

Questions?