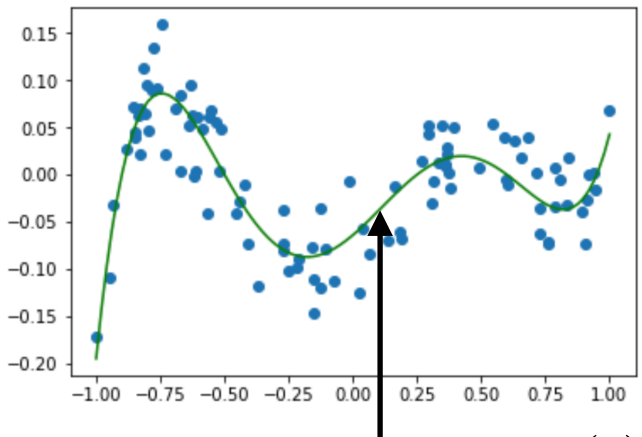
# More bias-variance and Ridge regression/ regularization

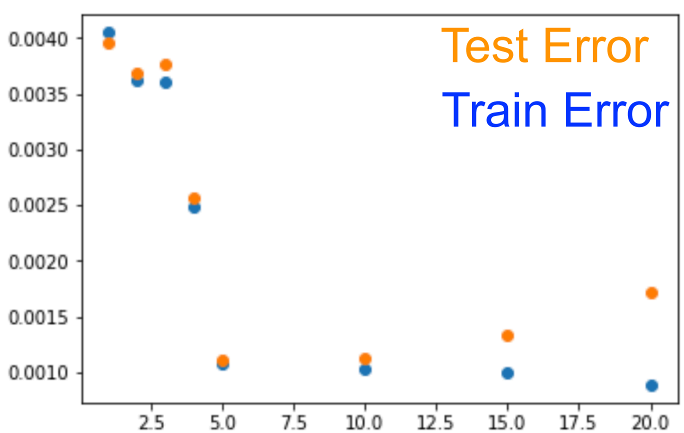
UNIVERSITY *of* WASHINGTON

# Test error vs. model complexity



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

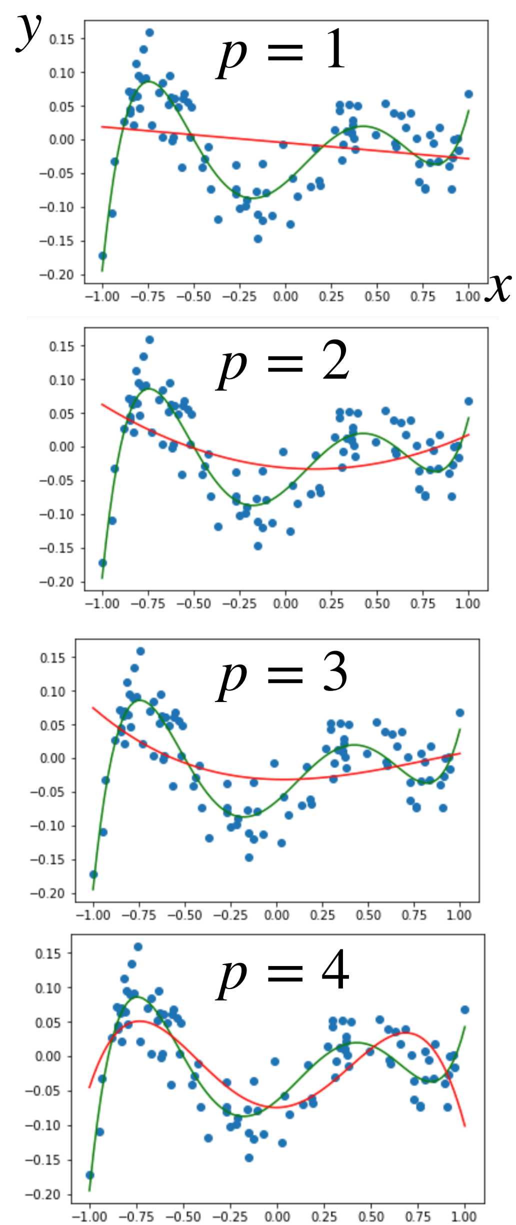
Error



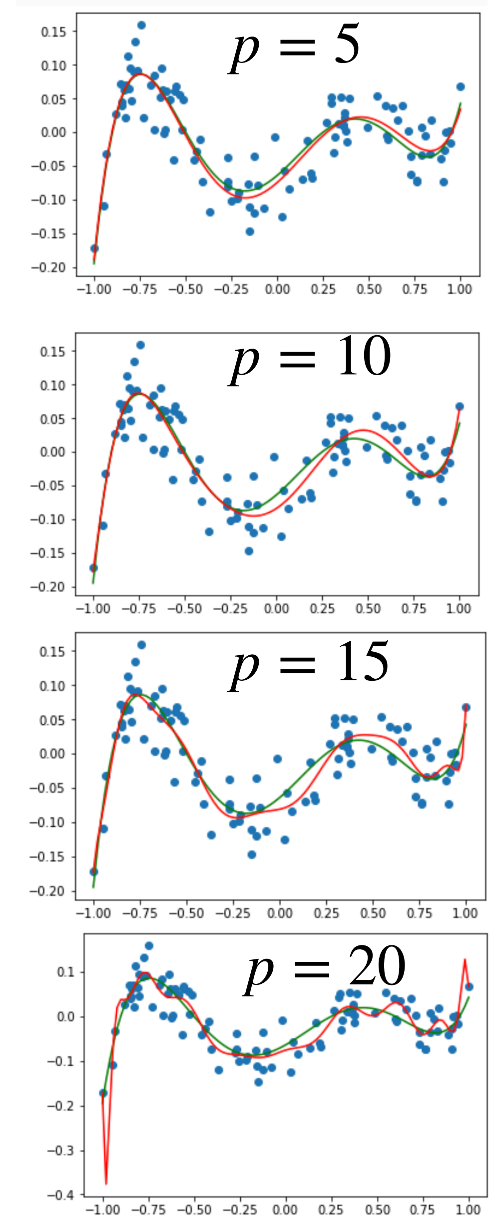

degree $p$ of the polynomial regression

demo4_tradeoff.ipynb

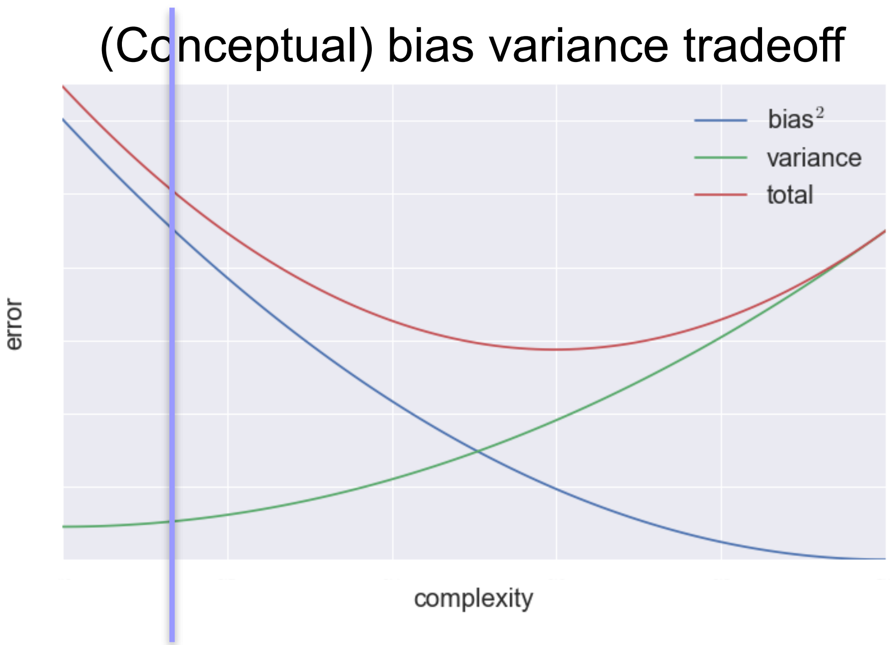
Simple model:
Model complexity is below the complexity of $\eta(x)$

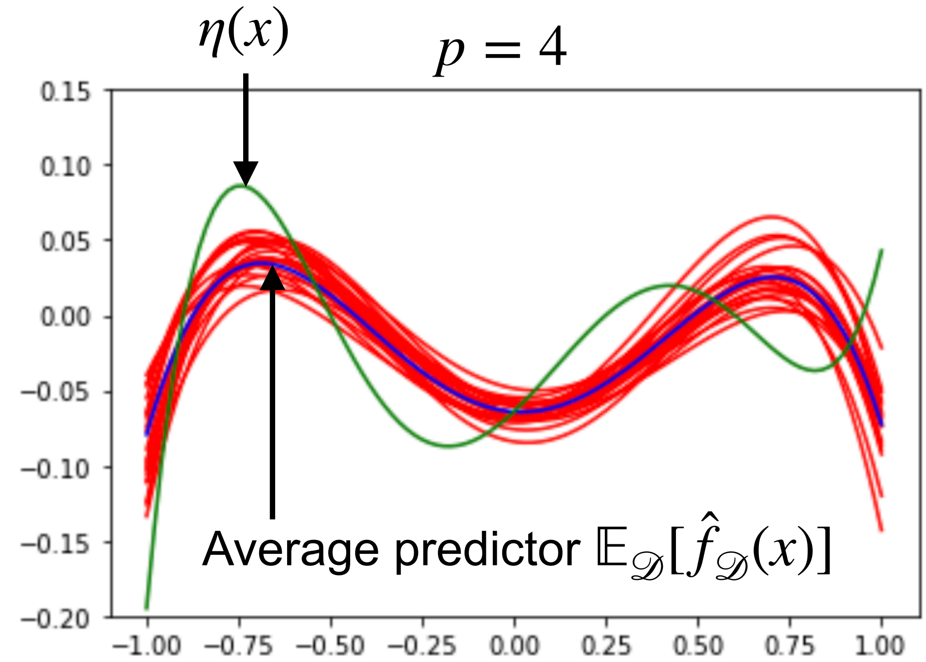



Complex model:



# Recap: Bias-variance tradeoff with simple model

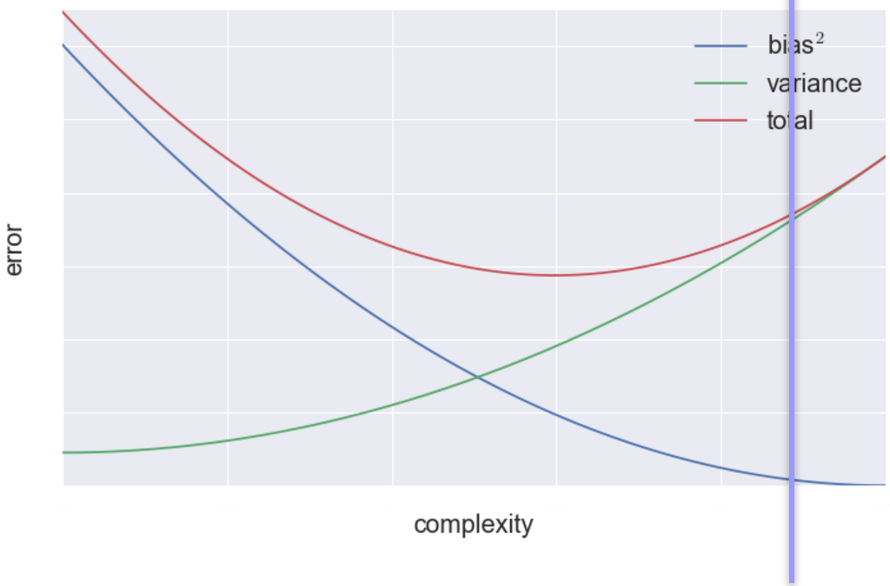


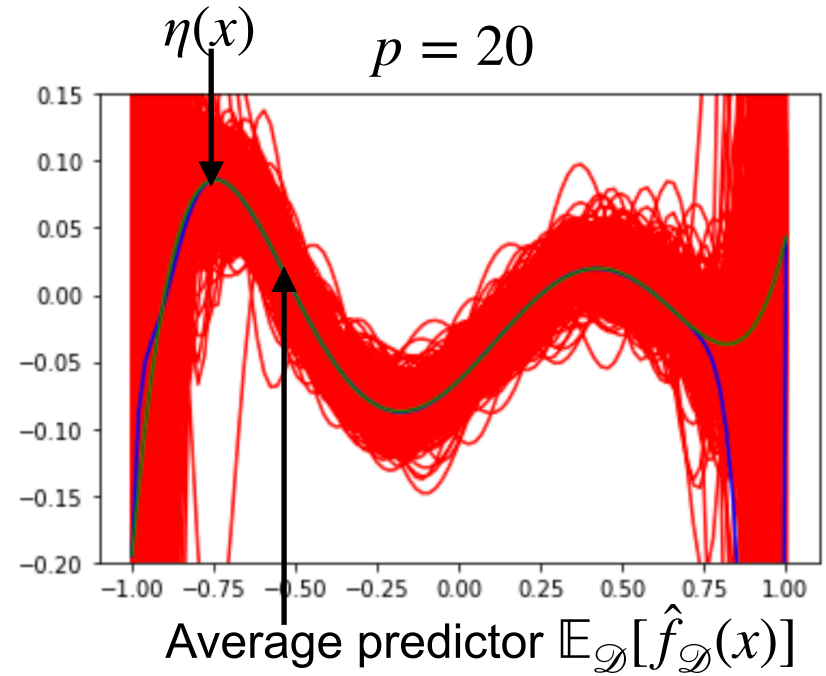




- When model **complexity is low** (lower than the optimal predictor $\eta(x)$)
  - Bias$^2$ of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2$, is large
  - Variance of our predictor, $\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]$, is small
  - If we have more samples, then
    - Bias
    - Variance
    - Because Variance is already small, overall test error

# Recap: Bias-variance tradeoff with simple model

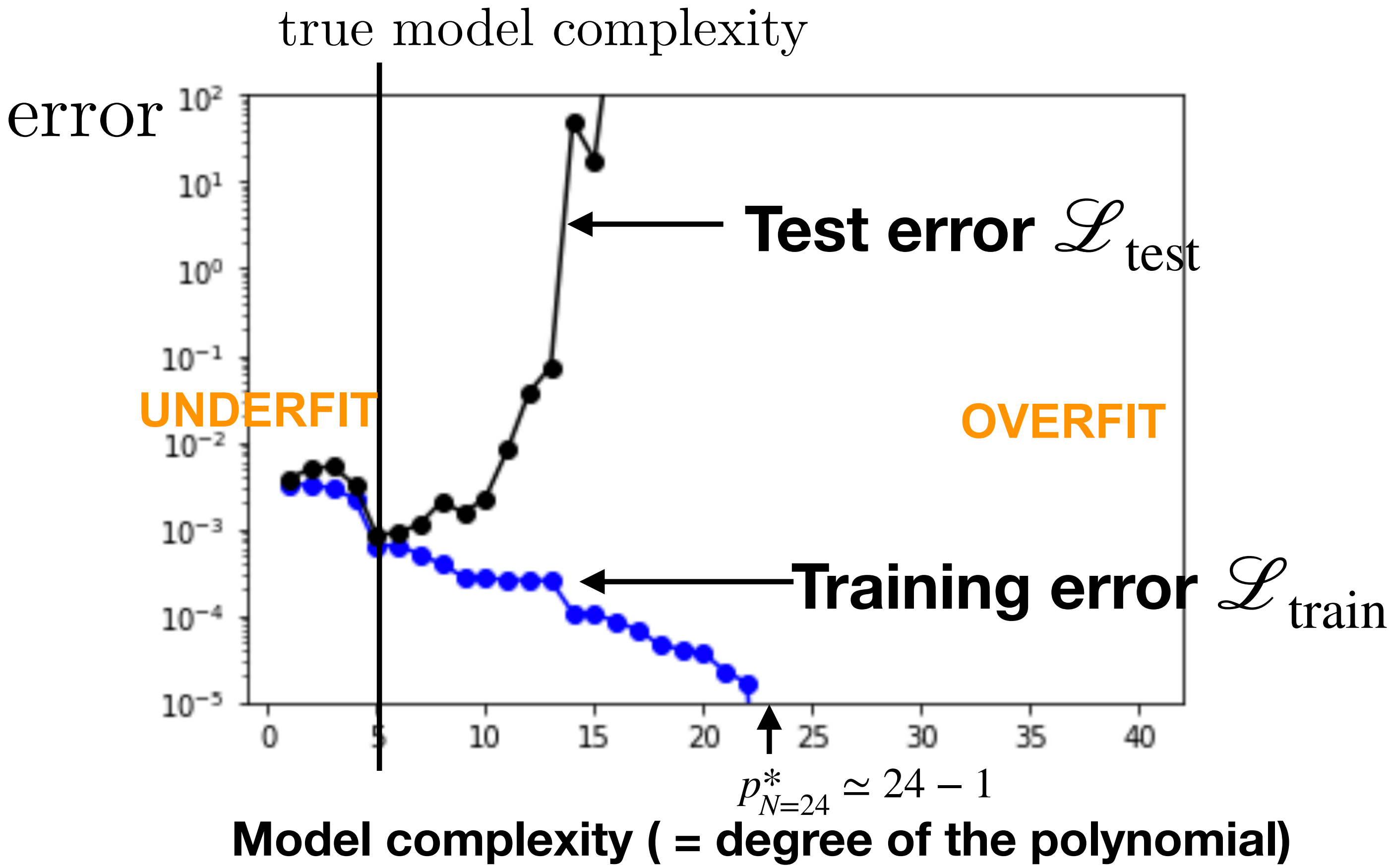
(Conceptual) bias variance tradeoff



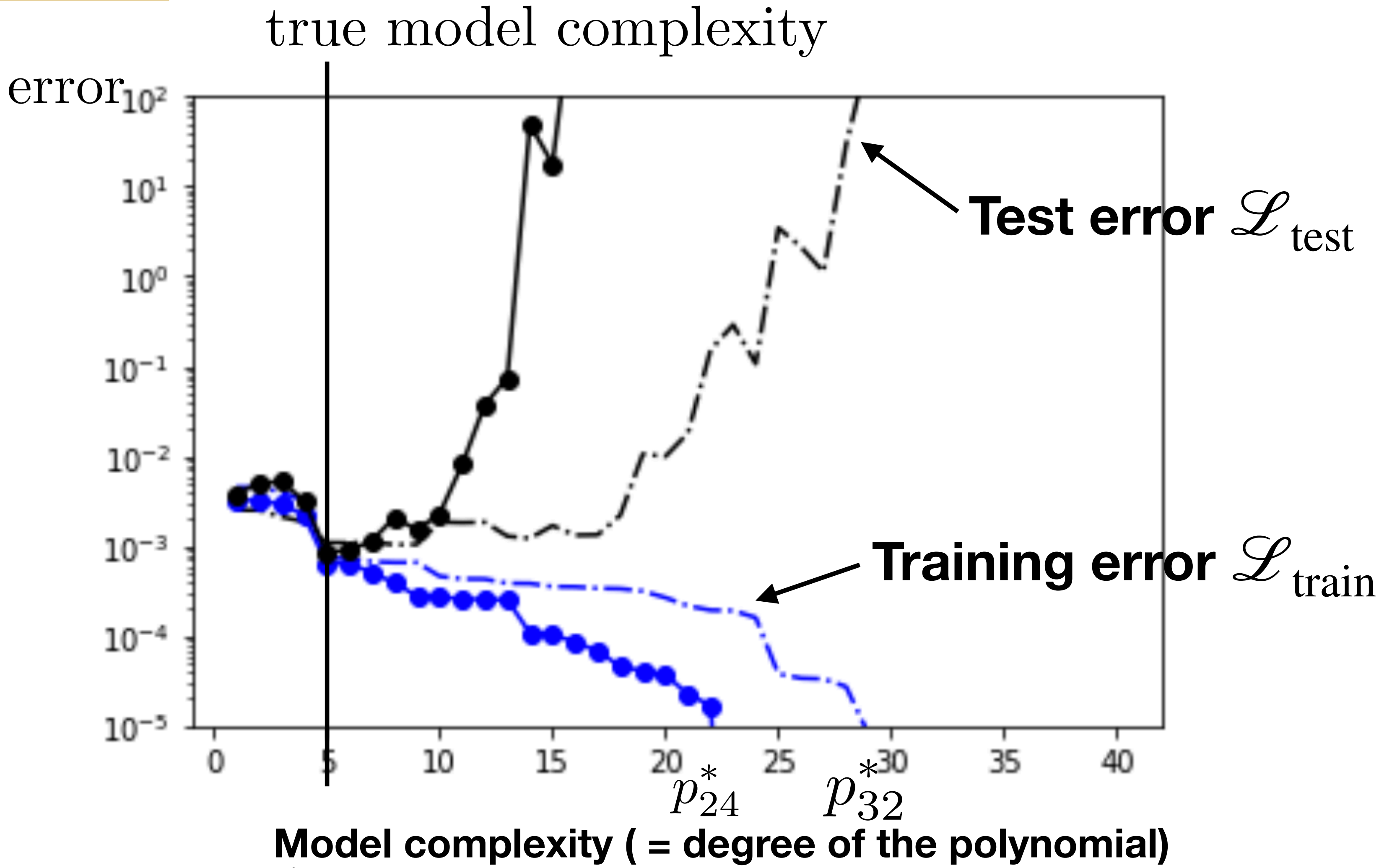

- When model complexity is high (higher than the optimal predictor $\eta(x)$)
  - Bias of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2$, is small
  - Variance of our predictor, $\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]$, is large
  - If we have more samples, then
    - Bias
    - Variance
    - Because Variance is dominating, overall test error

- 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 $S_{\text{train}}$ and test set $S_{\text{test}}$ via 80/20 split
- then we run multiple validations and plot the computed MSEs for all values of ***p*** that we are interested in



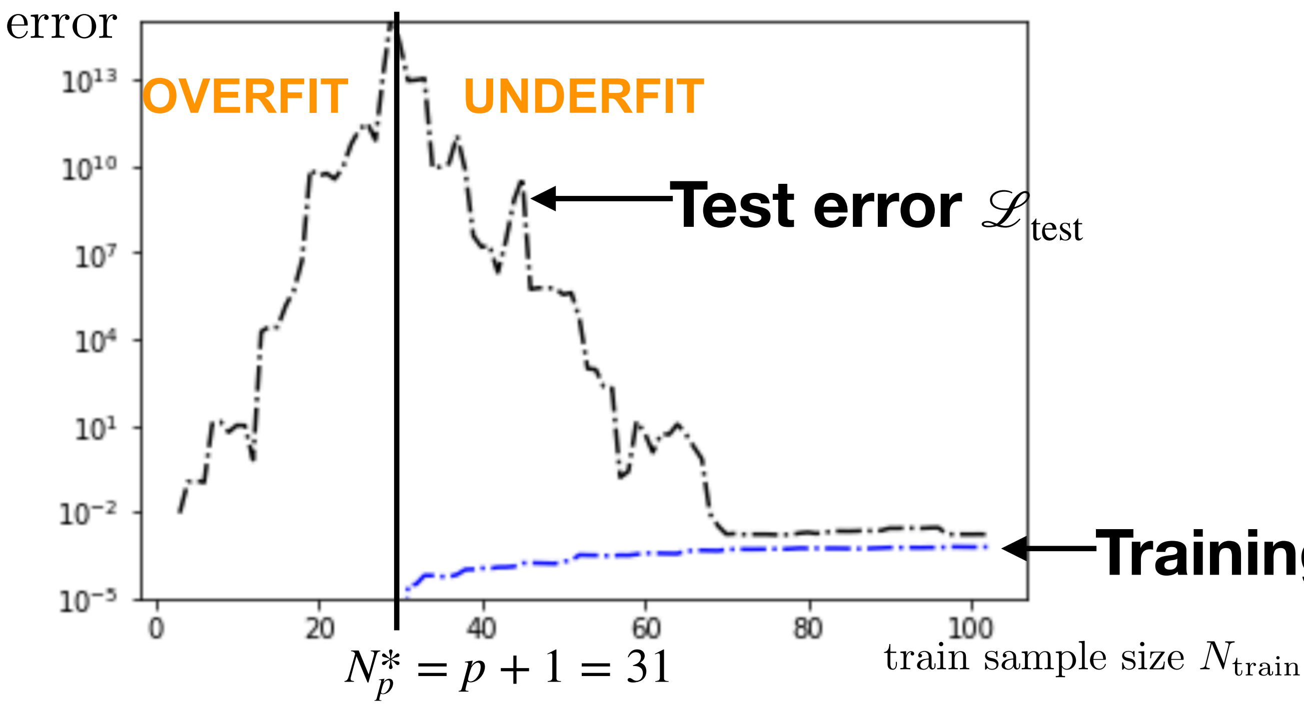

- Given sample size N there is a threshold, $p^*_N$, where training error is zero
- Training error is **always** monotonically non-increasing
- Test error has a trend of going down and then up, but fluctuates

- 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

lecture2_polynomialfit.ipynb

# Bias-variance tradeoff for linear models

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

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

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

$$=$$

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

$$\hat{f}_{\mathcal{D}}(x) = x^T\widehat{w}_{\text{MLE}} =$$

# Bias-variance tradeoff for linear models

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

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

$$\widehat{w}_{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w^* + \epsilon)$$

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

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

$$\hat{f}_{\mathcal{D}}(x) = x^T \widehat{w}_{\text{MLE}} = x^T w^* + x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

- Irreducible error: $\mathbb{E}_{X,Y}[(Y - \eta(x))^2 \,|\, X = x] =$
- Bias squared: $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2 =$
  (is independent of the sample size!)

# Bias-variance tradeoff for linear models

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

$$\widehat{w}_{\text{MLE}} = w^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\eta(x) = x^T w^*$$

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

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

# Bias-variance tradeoff for linear models

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

$$\widehat{w}_{\text{MLE}} = w^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\eta(x) = x^T w^*$$

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

- Variance: $\mathbb{E}_{\mathcal{D}}\Big[\,\big(\hat{f}_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\,\big)^2\,\Big] = \mathbb{E}_{\mathcal{D}}[x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x]$

$$= \sigma^2\,\mathbb{E}_{\mathcal{D}}[x^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x]$$

$$= \sigma^2\,x^T\mathbb{E}_{\mathcal{D}}[(\mathbf{X}^T\mathbf{X})^{-1}]x$$

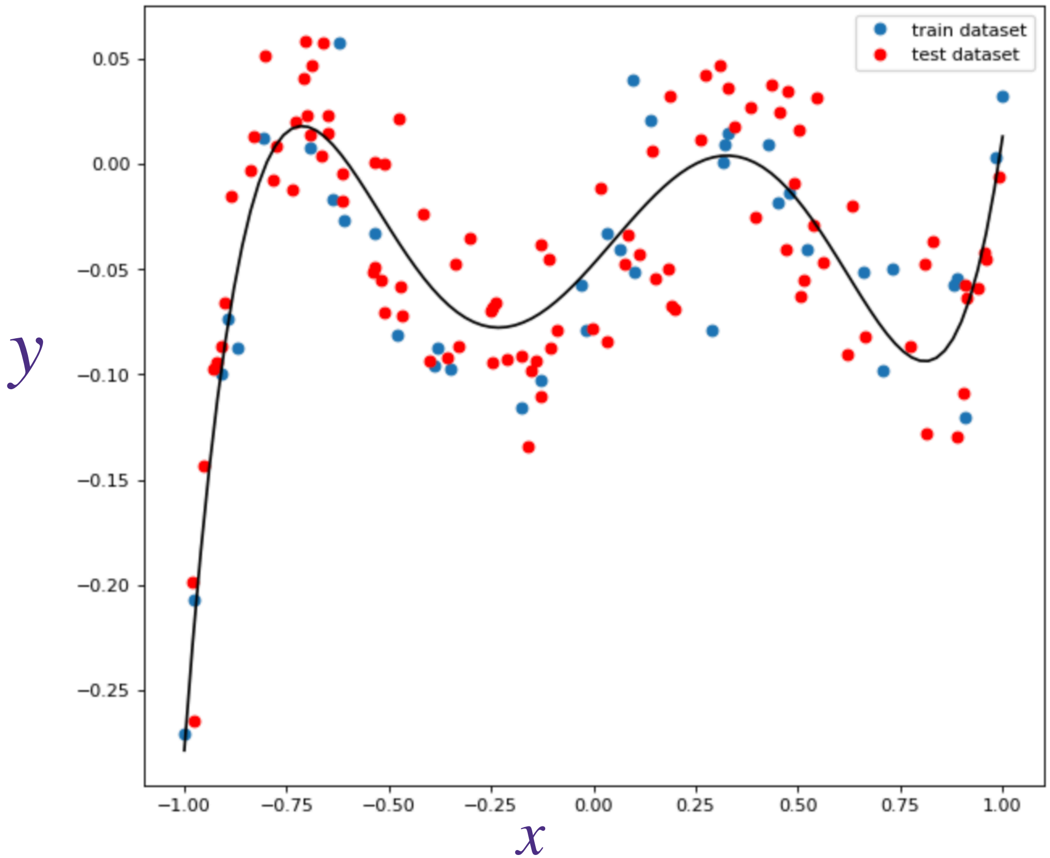
- To analyze this, let's assume that $X_i \sim \mathcal{N}(0,\mathbf{I})$ and number of samples, $n$, is large enough such that $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$ with high probability and $\mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}] \simeq \frac{1}{n}\mathbf{I}$, then
  - Variance is $\frac{\sigma^2 x^T x}{n}$, and decreases with increasing sample size $n$

# Regularization

# Recap: bias-variance tradeoff

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

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



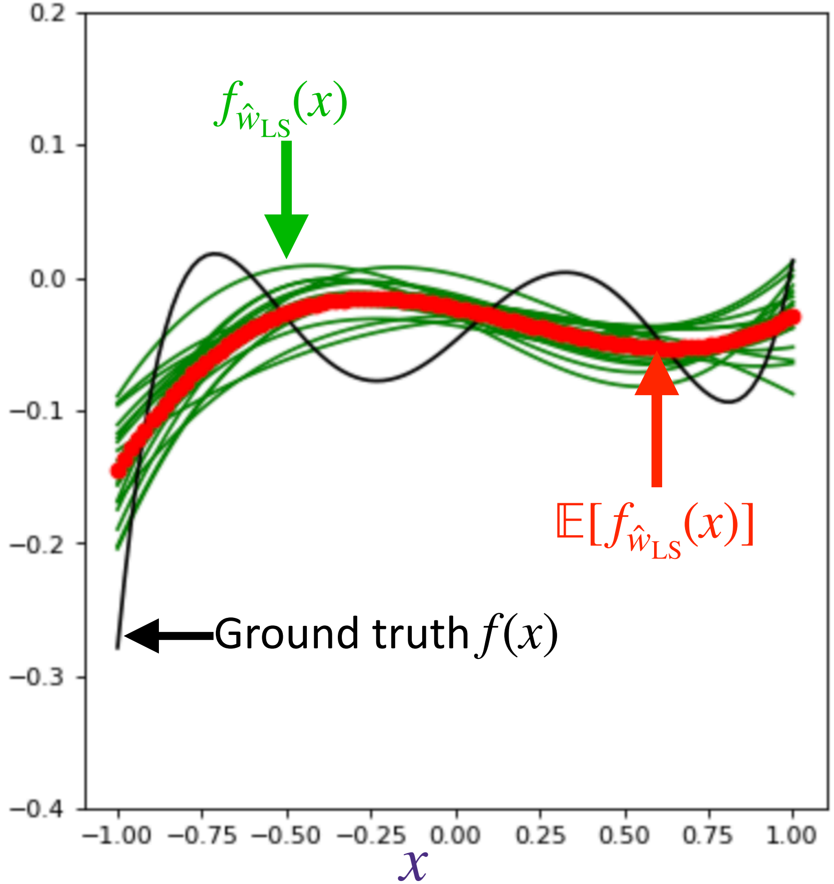

This is a linear model with features

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

# Recap: bias-variance tradeoff

With degree-3 polynomials, we underfit

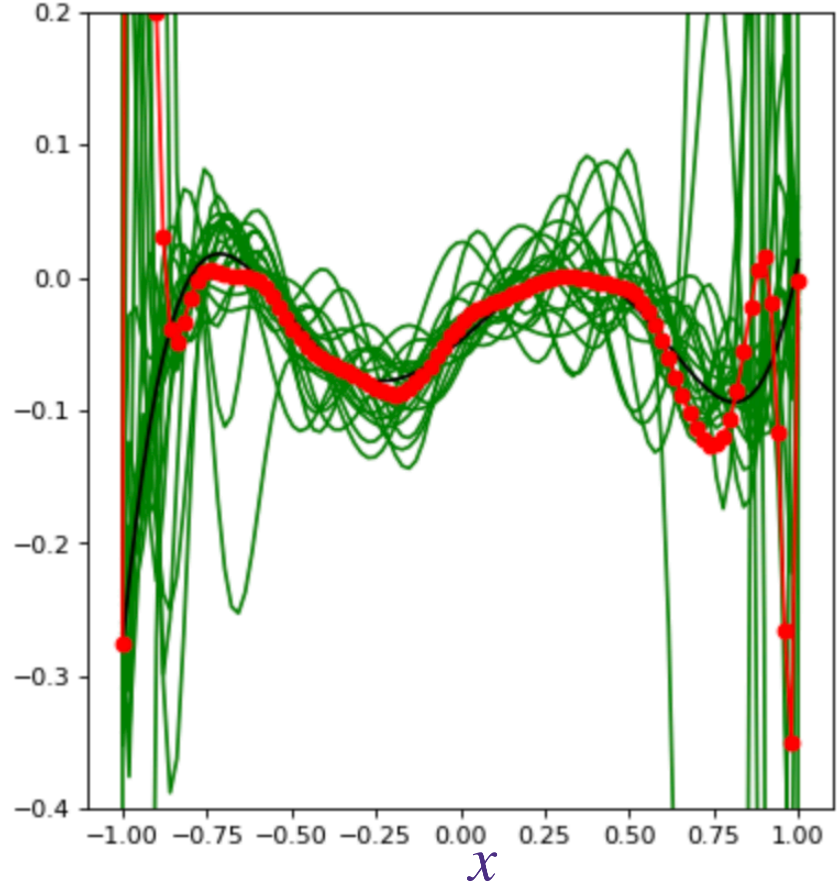
$\hat{f}_{\hat{w}_{\text{LS}}}(x)$




```
current train error = 0.0036791644380554187
current test error  = 0.0037962529988410953
```

With degree-20 polynomials, we overfit

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




```
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 $b$

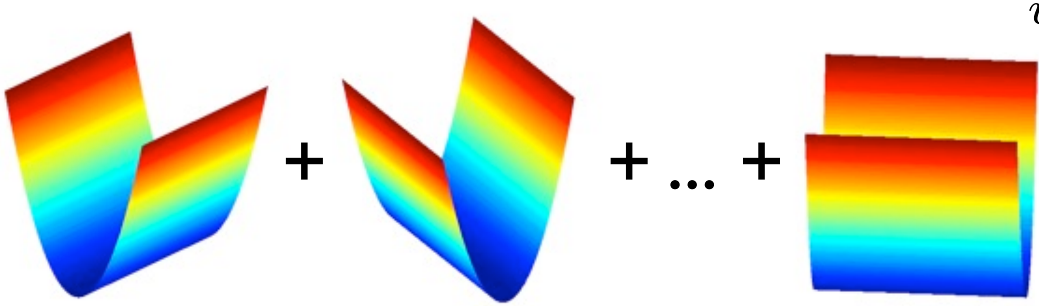
- In **Ridge Regression,** we use a regularizer $||w||_2^2$ to measure and control the sensitivity of the predictor
- And optimize for small loss and small sensitivity, by adding a **regularizer** in the objective (assume no offset for now)

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

# Ridge Regression

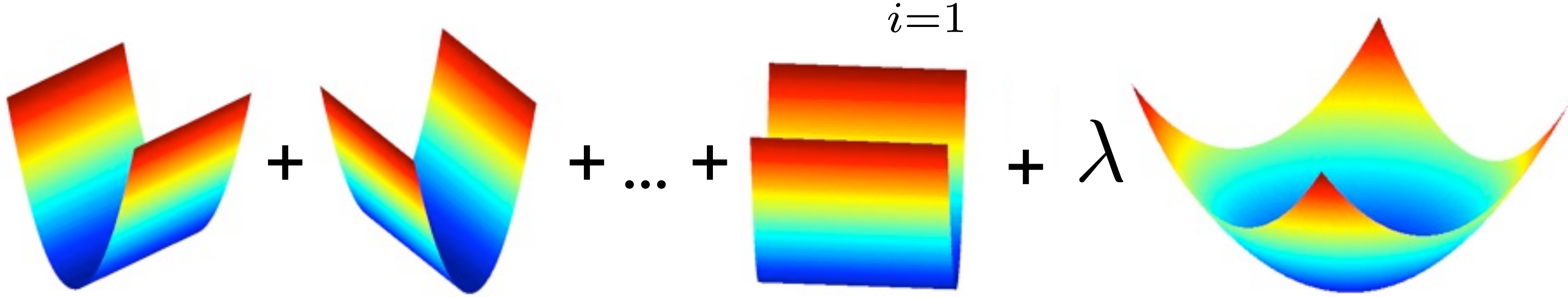
- (Original) Least squares objective:

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




- Ridge Regression objective:

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



# Minimizing the Ridge Regression Objective

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

# Shrinkage Properties

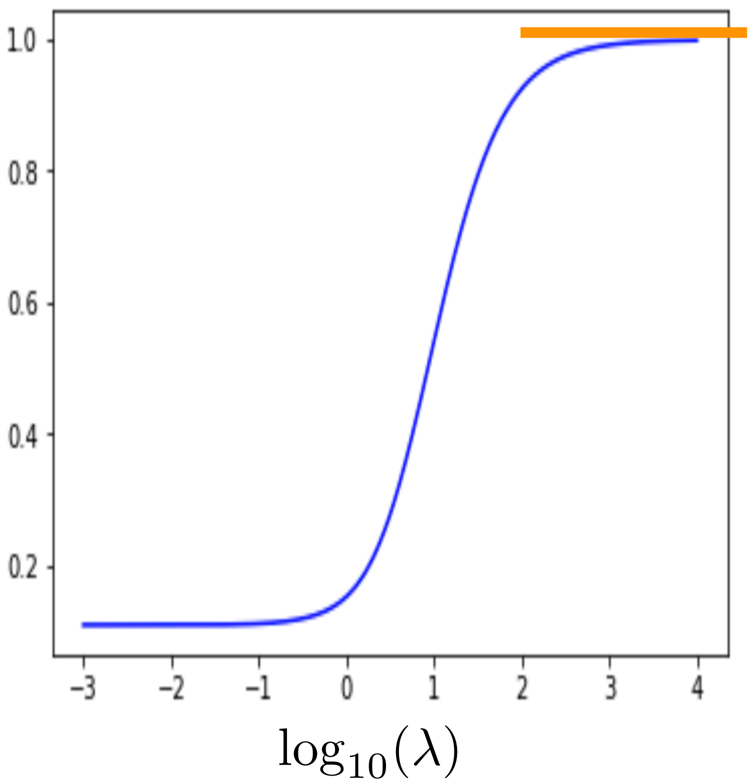
$$\widehat{w}_{ridge} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$$

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

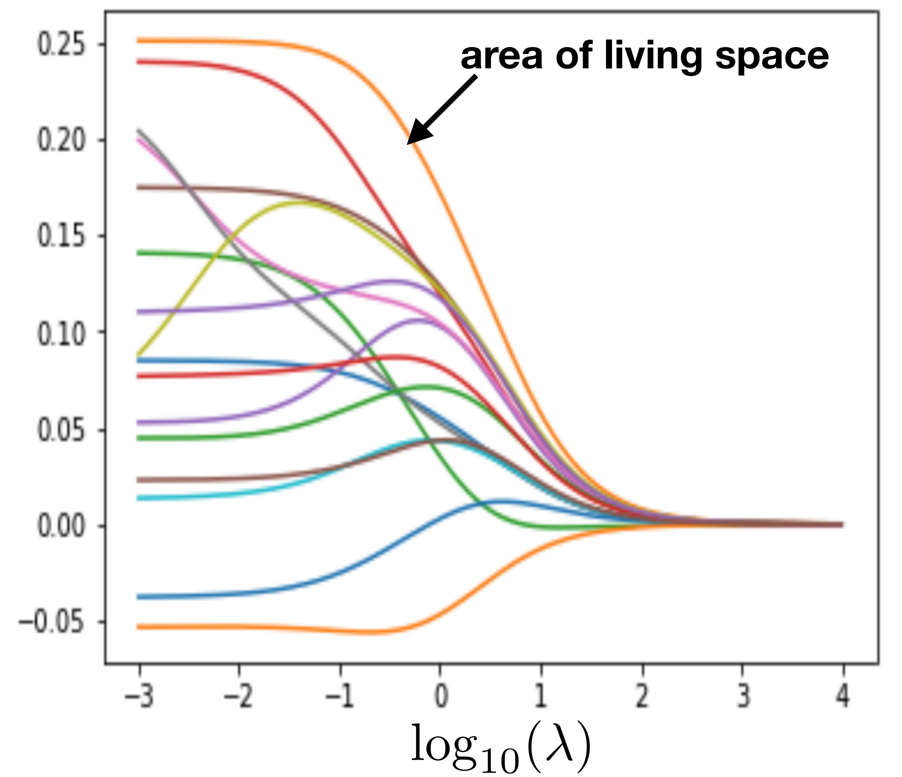
- When $\lambda = 0$, this gives the least squares model
- This defines a family of models hyper-parametrized by $\lambda$
- Large $\lambda$ means more regularization and simpler model
- Small $\lambda$ means less regularization and more complex model

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

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

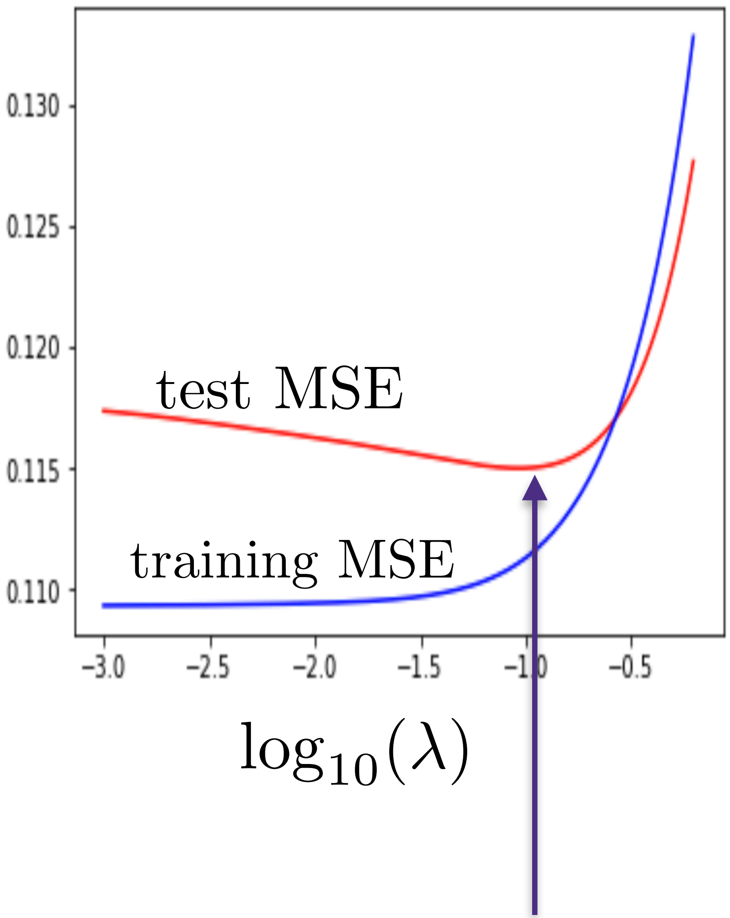


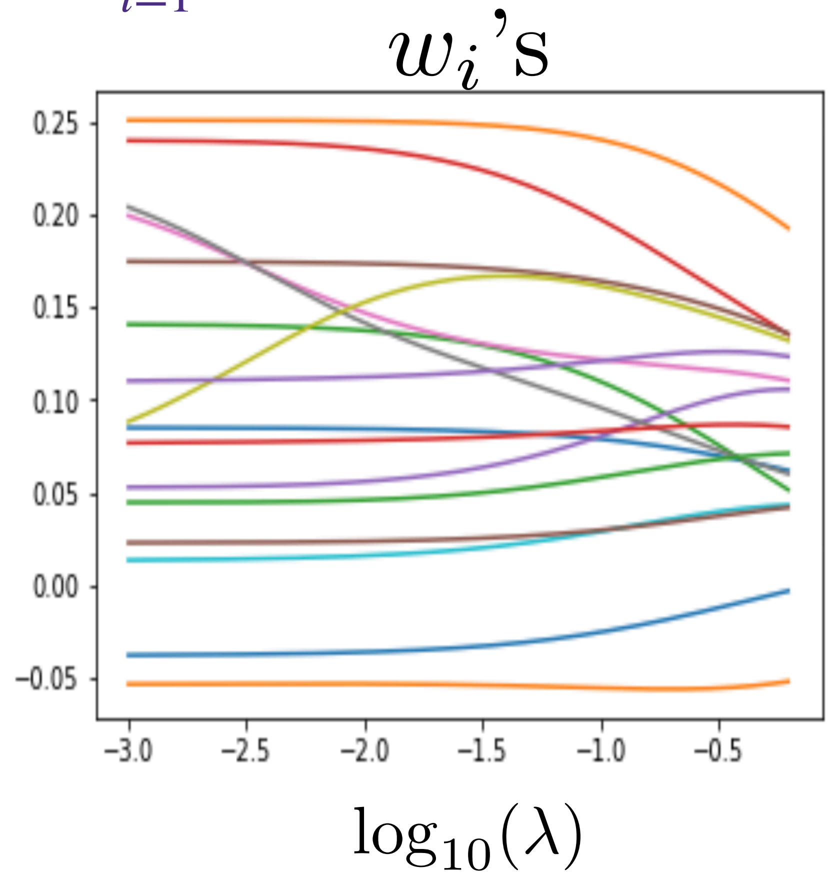

$w_i$'s




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




- 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: $\hat{w}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- 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 $\mathbb{E}_{y,\mathcal{D}_{\text{train}}|x}[(y - x^T\hat{w}_{\text{ridge}})^2 \,|\, x]$

# 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

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