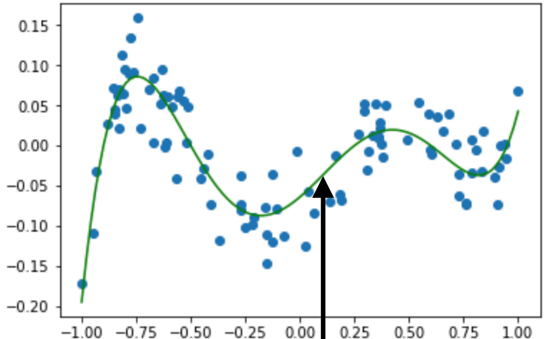
* Monday No class
* Thai

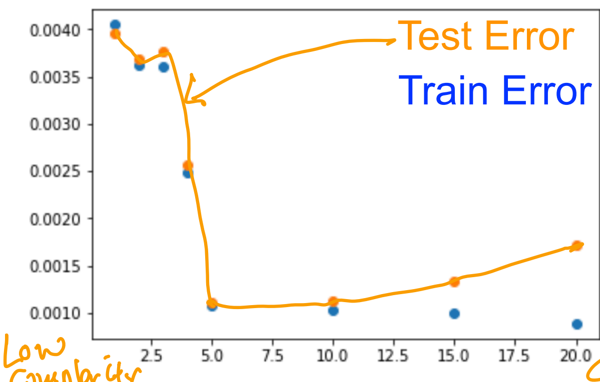
Lecture 6: Bias-Variance Tradeoff (continued)

Test error vs. model complexity



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

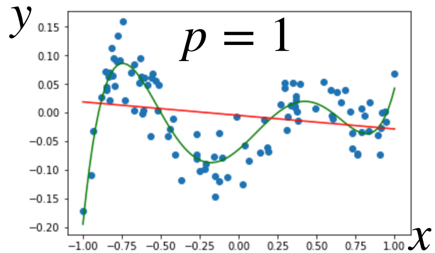
Error

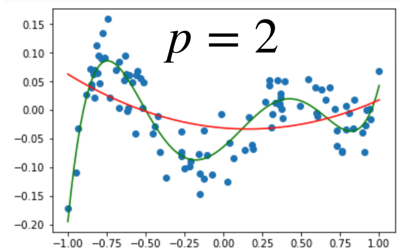


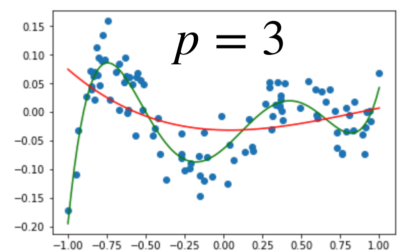
degree p of the polynomial regression

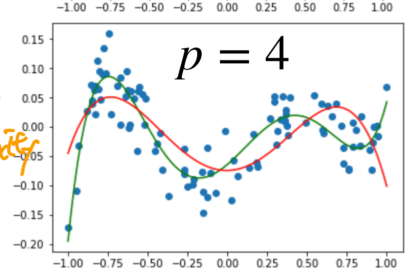
demo4_tradeoff.ipynb

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

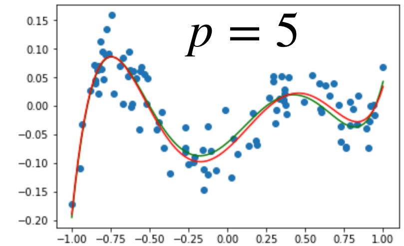


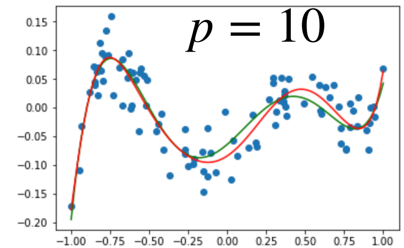


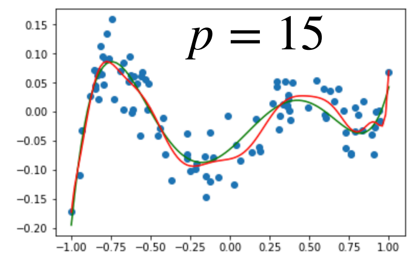


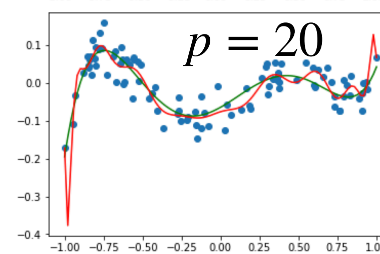


Complex model:

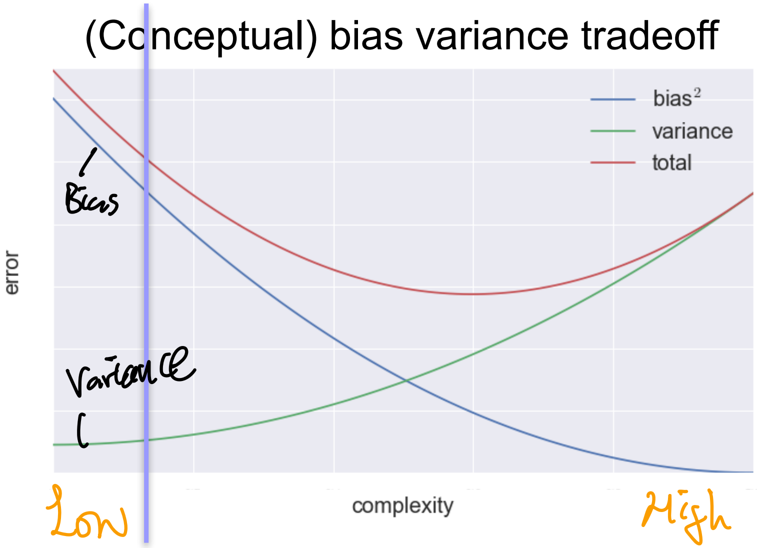


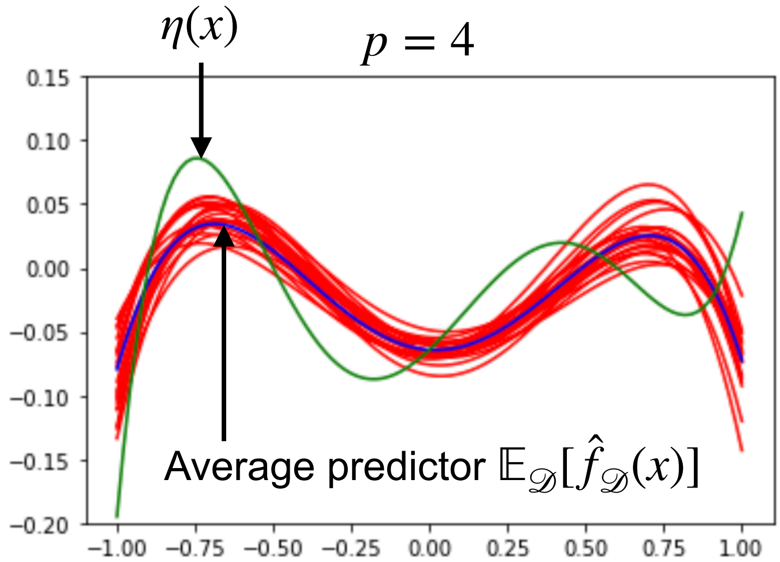






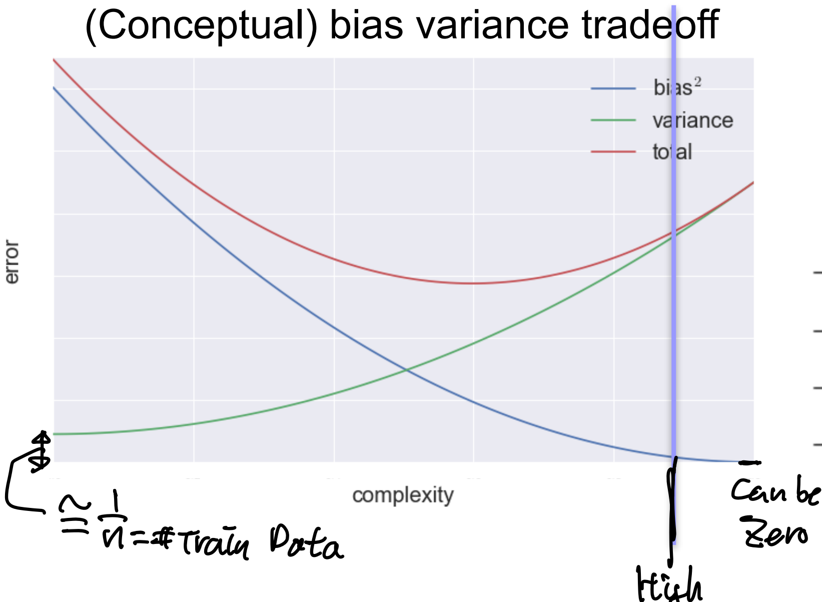
Recap: Bias-variance tradeoff with simple model

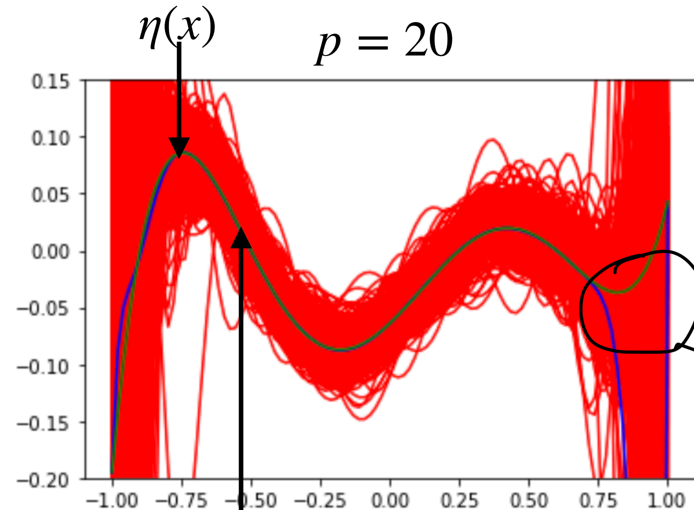




- When model **complexity is low** (lower than the optimal predictor $\eta(x)$)
  - Bias^2 of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2$, is large
  - Variance of our predictor, $\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]$, is small
  - If we have more samples, then
    - Bias does not change
    - Variance decreases
    - Because Variance is already small, overall test error does not change much

Recap: Bias-variance tradeoff with ~~simple~~ Complex model



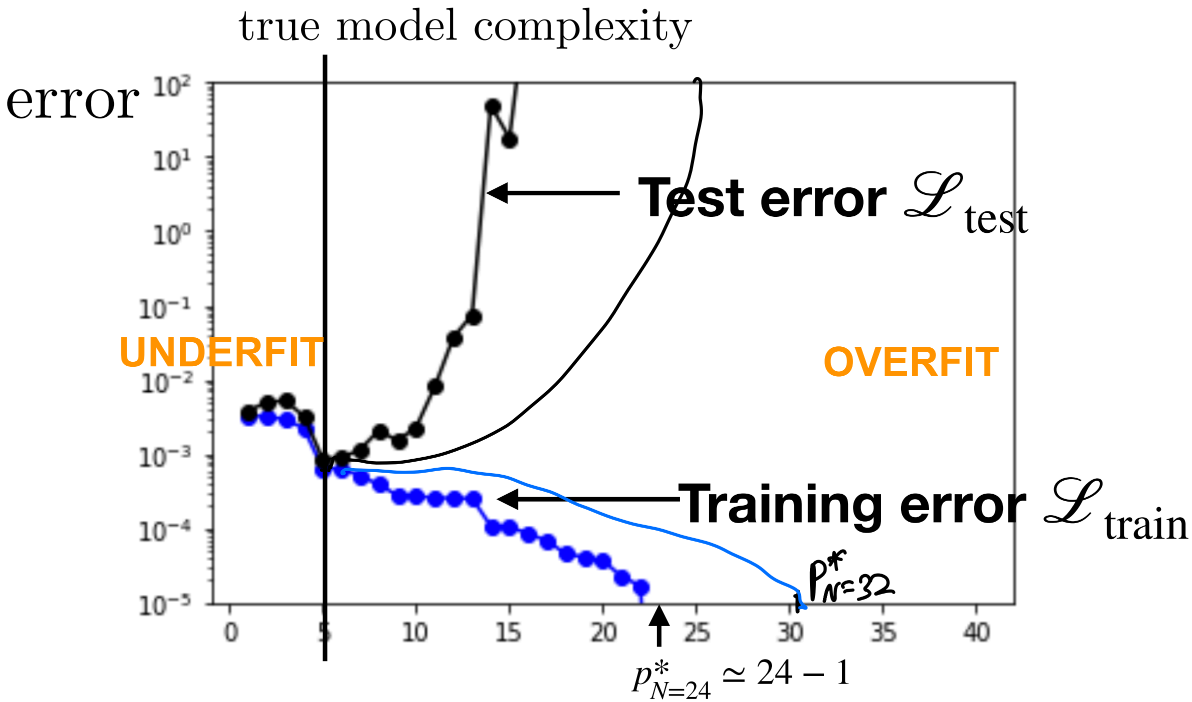


Average predictor $\mathbb{E}_{\mathscr{D}}[\hat{f}_{\mathscr{D}}(x)]$

If you run it with MANY Instances this Gap disappears.

- When model complexity is high (higher than the optimal predictor $\eta(x)$)
  - Bias of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathscr{D}}[\hat{f}_{\mathscr{D}}(x)]\right)^2$, is small
  - Variance of our predictor, $\mathbb{E}_{\mathscr{D}}\left[\left(\mathbb{E}_{\mathscr{D}}[\hat{f}_{\mathscr{D}}(x)] - \hat{f}_{\mathscr{D}}(x)\right)^2\right]$, is large
  - If we have more samples, then
    - Bias does not change
    - Variance goes down
    - Because Variance is dominating, overall test error goes down if more samples

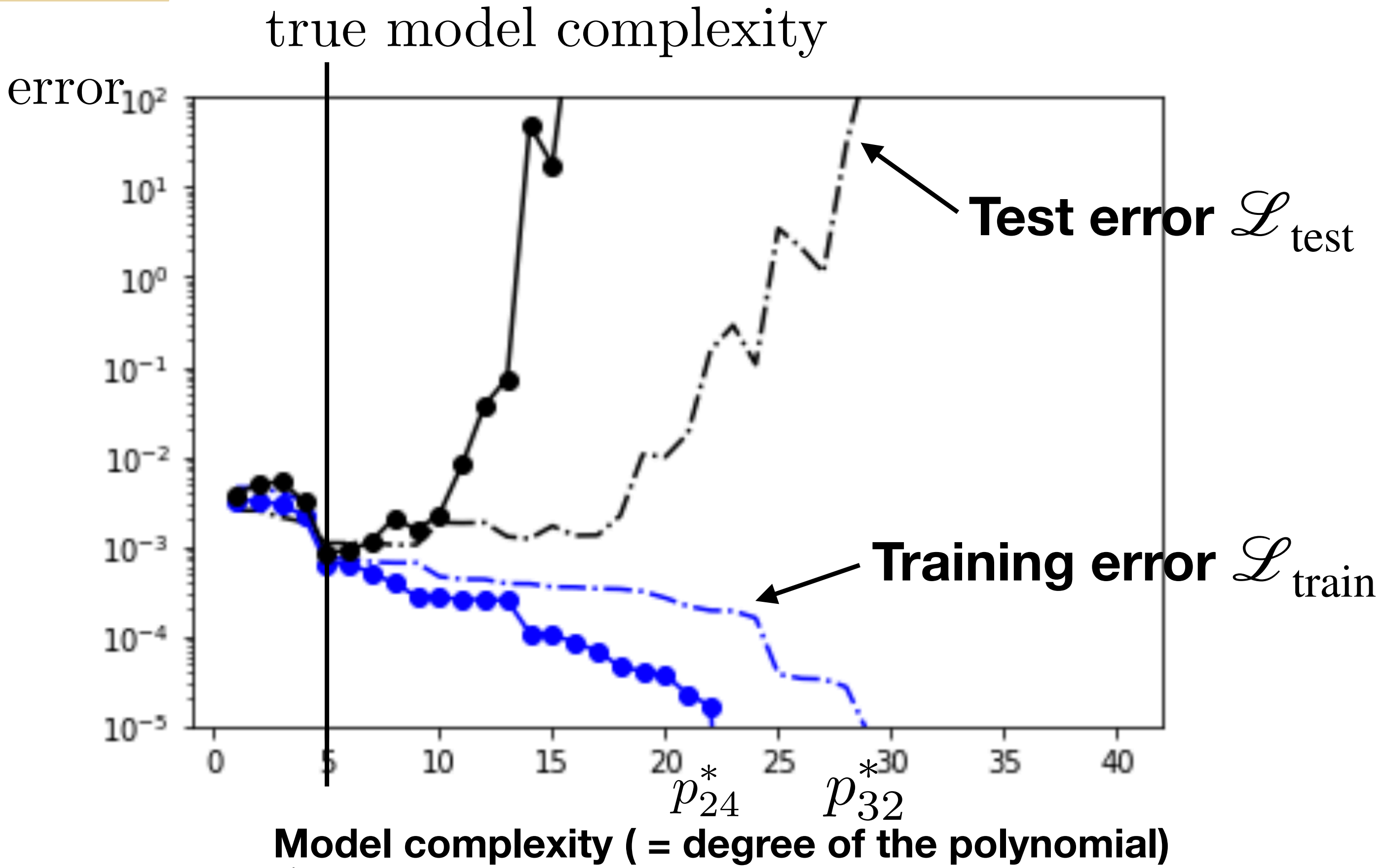
- let us first fix sample size ***N=30***, collect one dataset of size N i.i.d. from a distribution, and fix one training set S_{train} and test set S_{test} via 80/20 split
  $\to 24$ $\to 6$
- then we run multiple validations and plot the computed MSEs for all values of ***p*** that we are interested in



Model complexity (= degree of the polynomial)

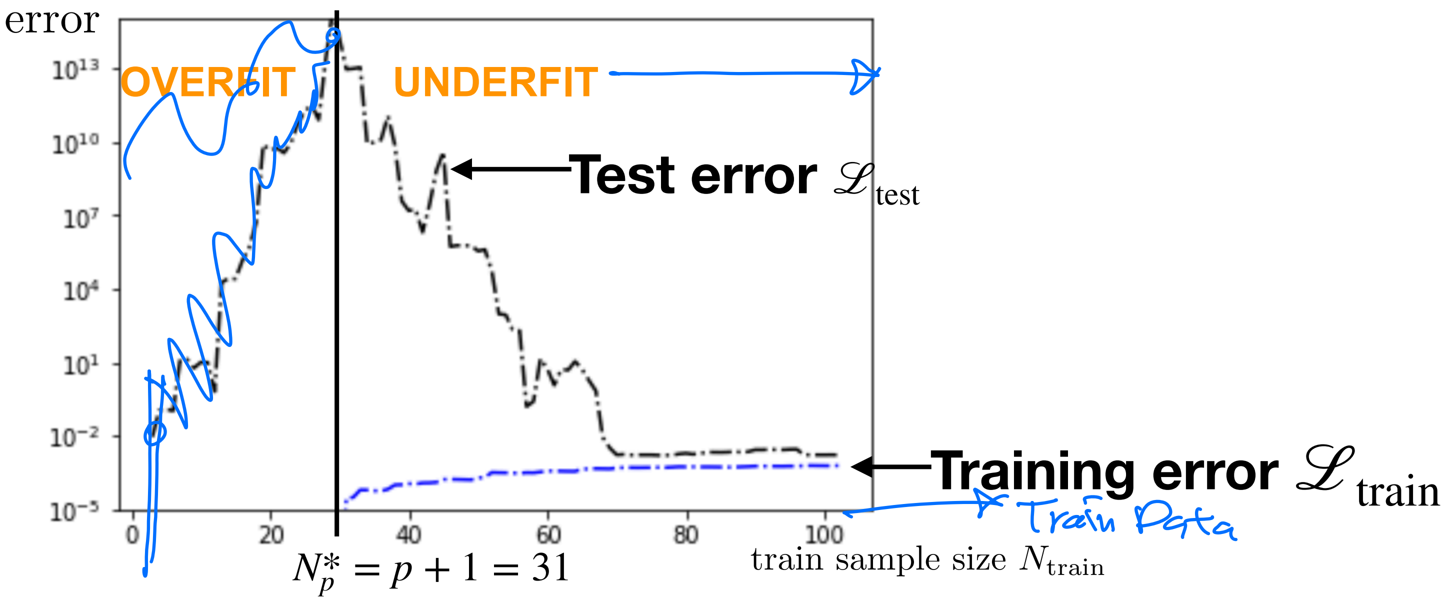
- Given sample size N there is a threshold, p^*_N, where training error is zero
- Training error is **always** monotonically non-increasing
- Test error has a trend of going down and then up, but fluctuates

- 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 $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$

Capital Means Random

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

Bold means they are concatenated

$$\mathbb{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbb{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$$

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

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

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

$$=$$

$$\eta(x) \overset{\Delta}{=} \mathbb{E}_{Y|X}[Y \mid X = x] = \mathbb{E}[x^Tw^* + \varepsilon_i] = x^Tw^*$$

def.

$$\hat{f}_{\mathcal{D}}(x) = x^T \widehat{w}_{\text{MLE}} = \underbrace{x^Tw^*}_{=\eta(x)} + \underbrace{x^T(X^TX)^{-1}X^T\varepsilon}_{\text{error}}$$

Bias-variance tradeoff for linear models

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

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

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

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

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

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

- Irreducible error: $\mathbb{E}_{X,Y}[(Y - \eta(x))^2 \mid X = x] = \mathbb{E}\left[\left(\cancel{x^T w^*} + \varepsilon_i - \cancel{x^T w^*}\right)^2\right] = \mathbb{E}[\varepsilon_i^2] = \sigma^2$
- Bias squared: $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2 = x^T w^* - \mathbb{E}[x^T w^* + x^T (X^T X)^{-1} X^T \underbrace{\varepsilon}_{\text{zero-mean}}]$ $= x^T w^* - x^T w^* = 0$
  (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: $\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2\right] = \mathbb{E}\left[\left(\cancel{x^T w^*} + x^T (X^T X)^{-1} X^T \epsilon - x^T w^*\right)^2\right]$

$$= \mathbb{E}\left[x^T \underbrace{(X^T X)^{-1} X^T}_{} \underbrace{\epsilon \epsilon^T}_{} X (X^T X)^{-1} x\right]$$

indep $\quad \mathbb{E}[\epsilon\epsilon^T] = \sigma^2 \cdot I$

$$= \sigma^2\, \mathbb{E}\left[x^T (X^T X)^{-1} \cancel{X^T X (X^T X)^{-1}} x\right]$$

$$= \underline{\sigma^2 \cdot x^T\, \mathbb{E}[(X^T X)^{-1}]\, x}$$

Bias-variance tradeoff for linear models

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

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

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

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

$\mathbb{E}\,\epsilon = 0$

$\mathbb{E}\,\epsilon\epsilon^T = \sigma^2 \cdot \mathbb{I}_{n\times n}$

$\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$

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

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

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

$\mathbb{R}^d$ — Train Data

- 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

$\Sigma\,[X_1\,X_2\,\cdots]\begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = \sum_{i=1}^{n} x_i x_i^T \cong \mathbb{E}[x_i x_i^T]\cdot n = n\cdot\mathbb{I}_{d\times d}$

  - Variance is $\boxed{\dfrac{\sigma^2\boxed{x^Tx}}{n}}$, and decreases with increasing sample size n

$(X^TX)^{-1} \cong \frac{1}{n}\mathbb{I}_{d\times d}$

Questions?