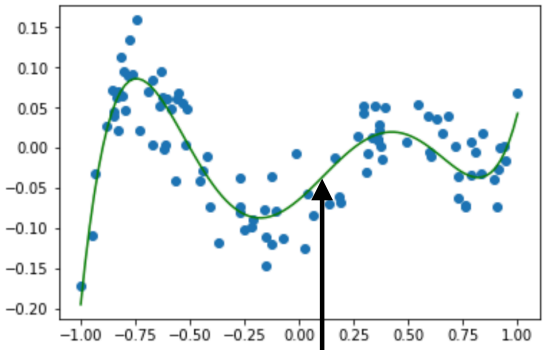
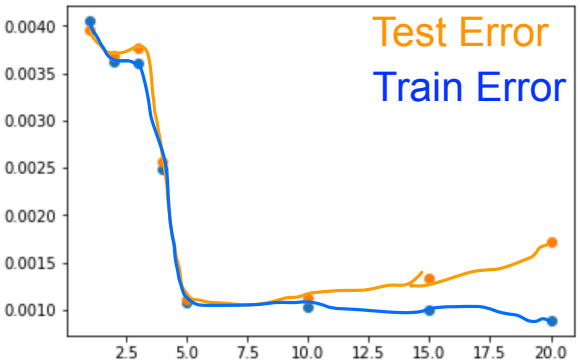
Lecture 6: Bias-Variance Tradeoff (continued)

Test error vs. model complexity



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

Error

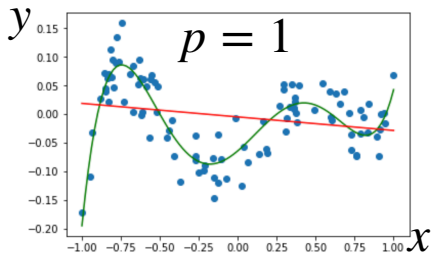


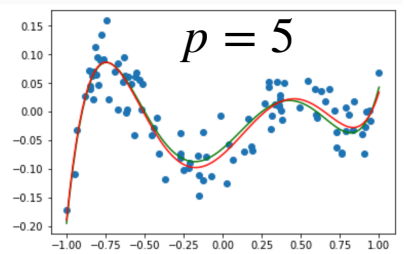
degree p of the polynomial regression

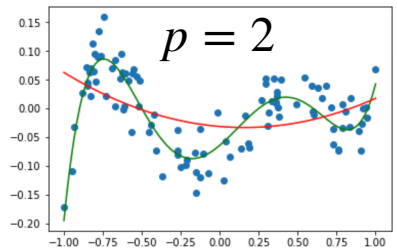
demo4_tradeoff.ipynb

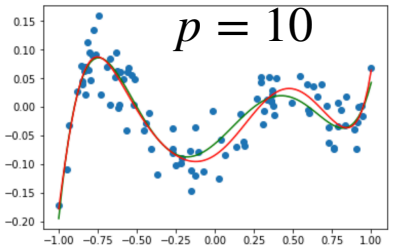
Simple model:
Model complexity is below the complexity of $\eta(x)$

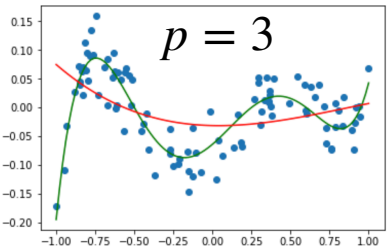
Complex model:

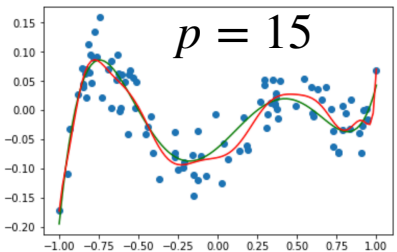


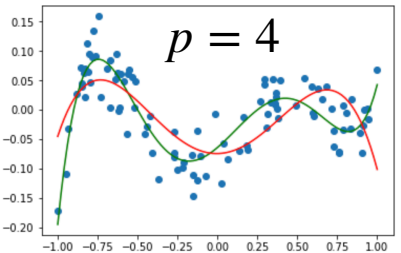


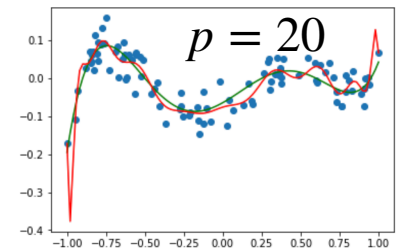




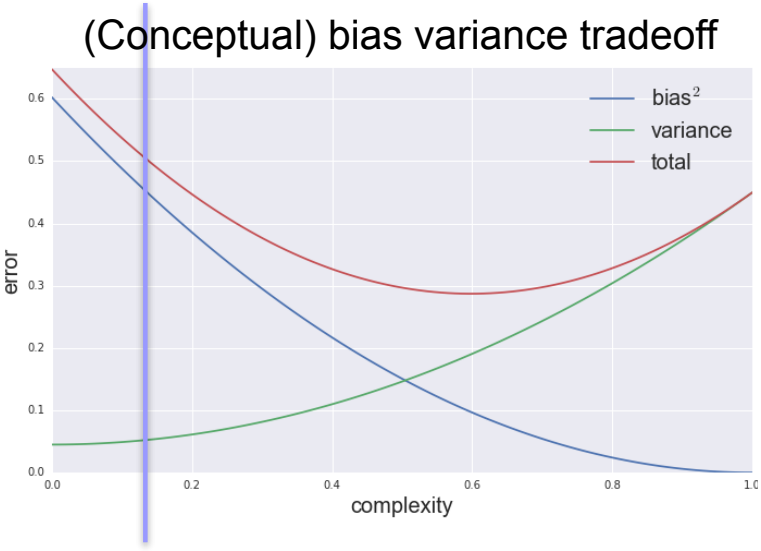


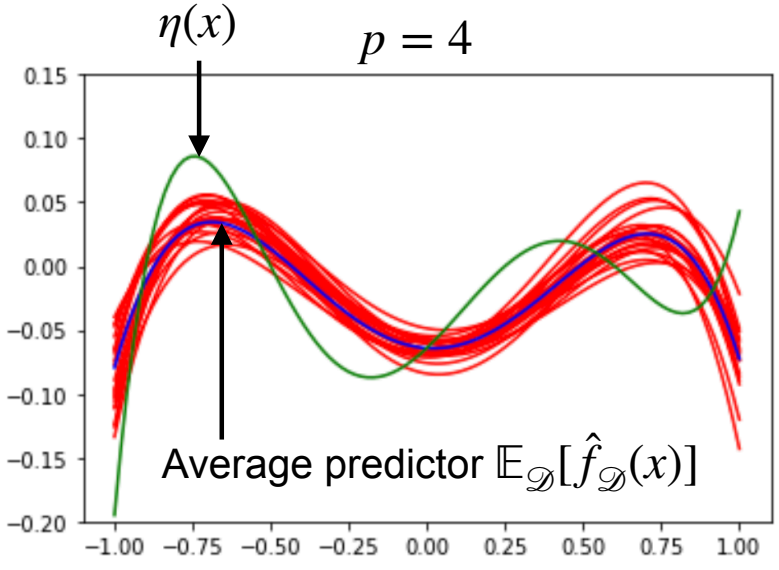






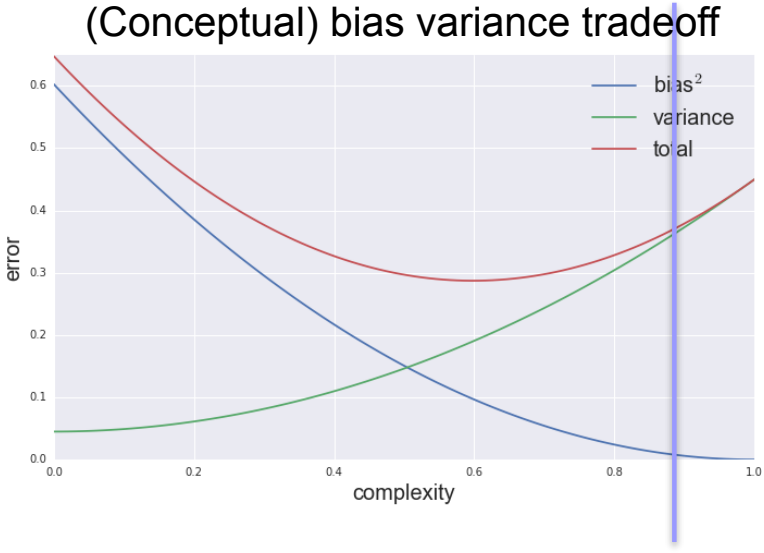
Recap: Bias-variance tradeoff with simple model

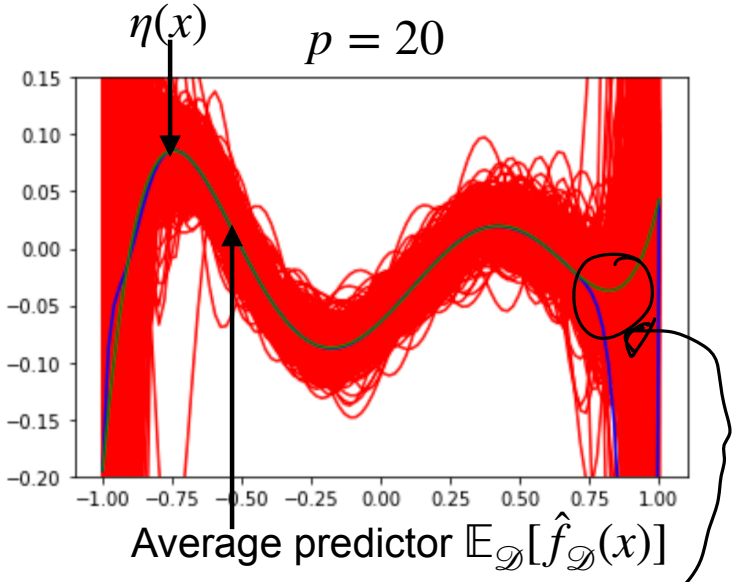




- When model **complexity is low** (lower than the optimal predictor $\eta(x)$)
  - Bias^2 of our predictor, $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2$, is large
  - Variance of our predictor, $\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]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

Recap: Bias-variance tradeoff with simple model

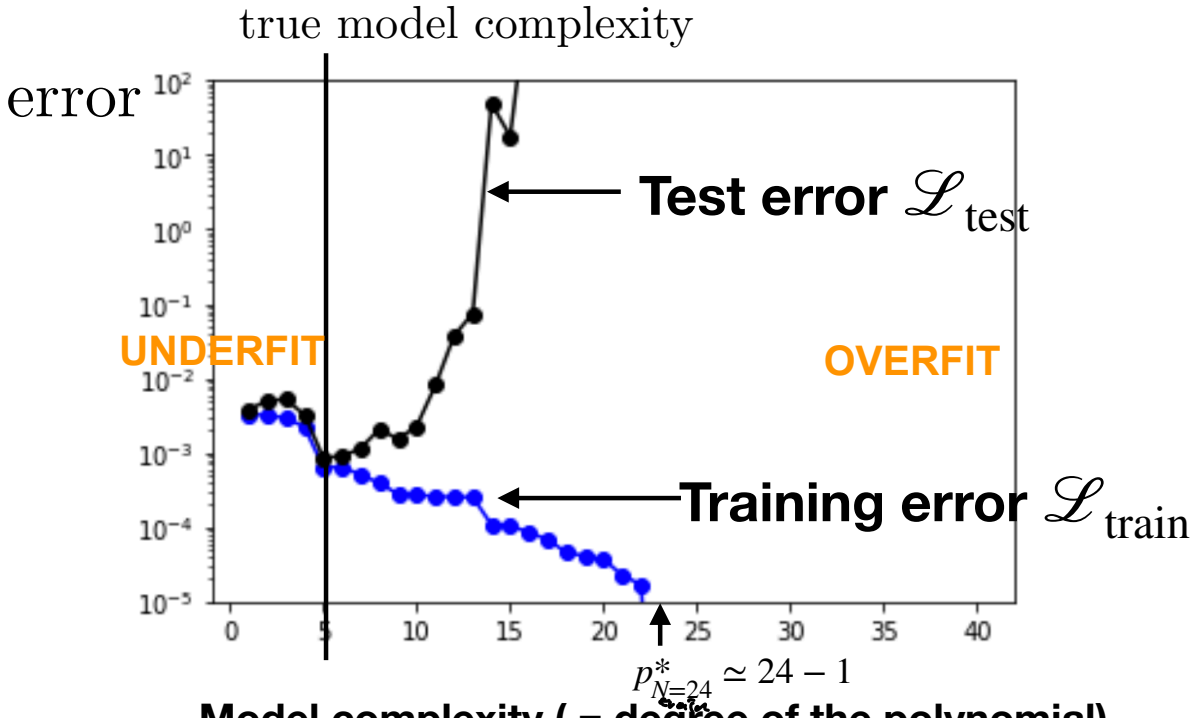




This part will also watch if I took more instances.

- 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 does not change
    - Variance decreases
    - Because Variance is dominating, overall test error decreases

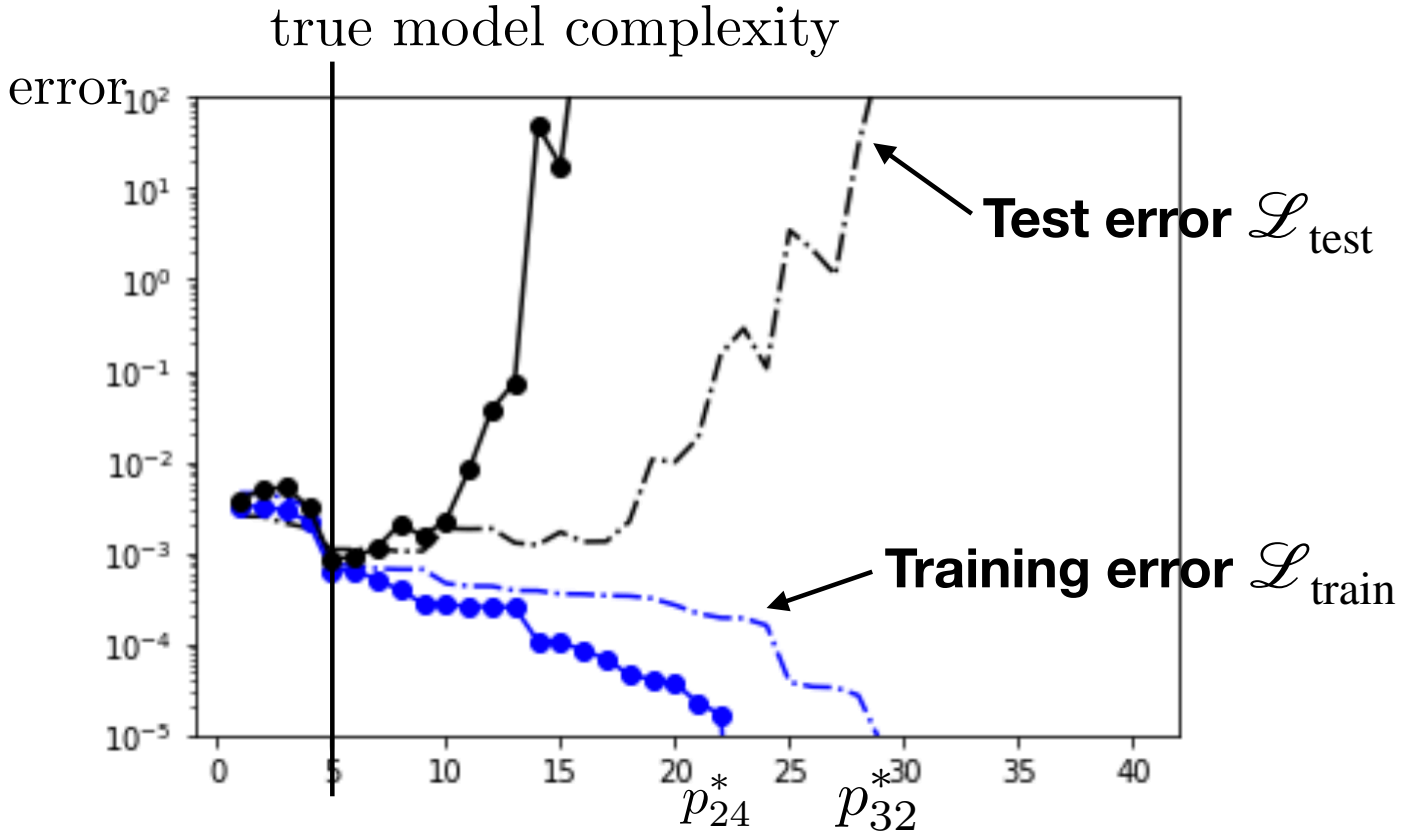
- let us first fix sample size ***N=30***, collect one dataset of size N i.i.d. from a distribution, and fix one training set S_{train} and test set S_{test} via 80/20 split
- then we run multiple validations and plot the computed MSEs for all values of ***p*** that we are interested in



- 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

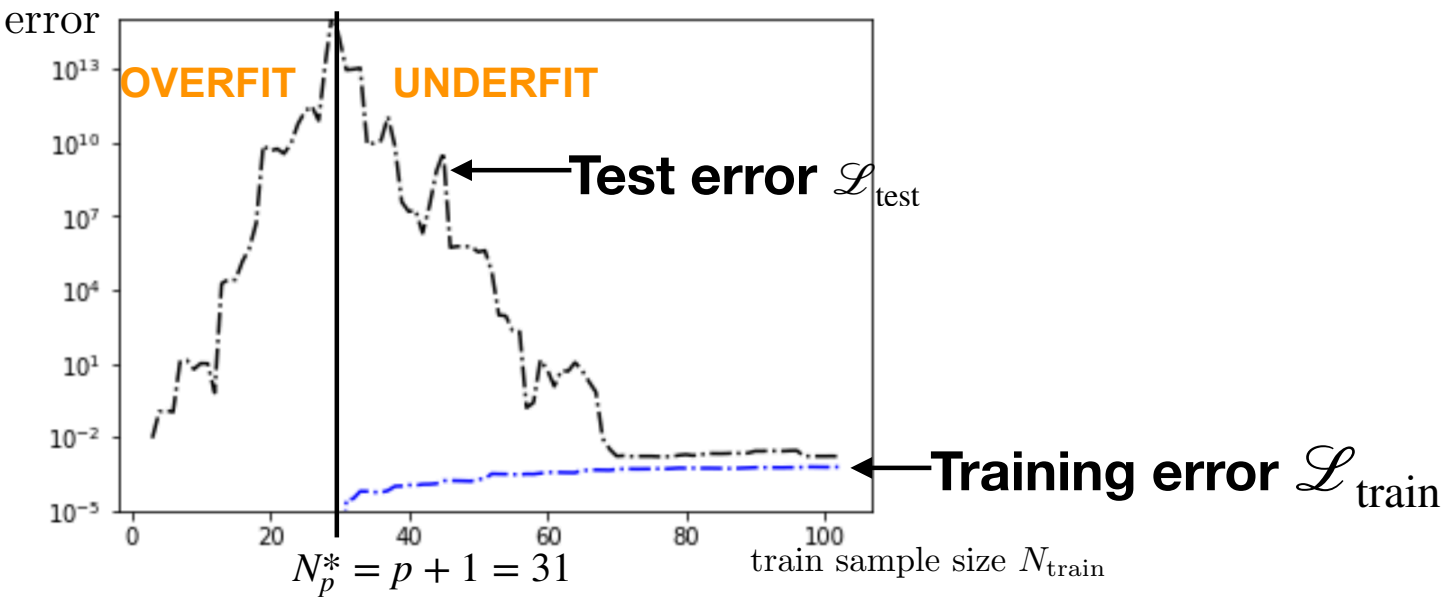
Q. what happens if N↑

- 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 because they are R.V.

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

Bold because they are vector $y \in \mathbb{R}^n$, matrix $X \in \mathbb{R}^{n \times d}$

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

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

$$\eta(x) = \mathbb{E}_{Y|X}[Y \mid X = x] = x^T w^* \quad \leftarrow Y_i = X_i^T w^* + \varepsilon_i$$

($Y \in \mathbb{R}$, $X \in \mathbb{R}^d$)

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

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: $\mathbb{E}_{X,Y}[(Y - \eta(x))^2 \,|\, X = x] =$ $\mathbb{E}\left[\left(x^T w^* + \epsilon_i - x^T w^*\right)^2\right] = \sigma^2$
- Bias squared: $\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2 = x^T w^* - \mathbb{E}\left[x^T w^* + x^T(X^TX)^{-1}X^T\varepsilon\right] = 0$ (is independent of the sample size!)

zero mean

Bias-variance tradeoff for linear models

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

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

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

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

- Variance: $\mathbb{E}_{\mathscr{D}}\left[\left(\hat{f}_{\mathscr{D}}(x) - \mathbb{E}_{\mathscr{D}}[\hat{f}_{\mathscr{D}}(x)]\right)^2\right] = \mathbb{E}\left[x^T(X^TX)^{-1}X^T\underbrace{\varepsilon\varepsilon^T}X(X^TX)^{-1}x\right]$

$\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 \cdot I$

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

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

Bias-variance tradeoff for linear models

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

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

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

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

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

Questions?