

## Module 8.2 : Train error vs Test error

- Consider a new point  $(x, y)$  which was not seen during training

- Consider a new point  $(x, y)$  which was not seen during training
- If we use the model  $\hat{f}(x)$  to predict the value of  $y$  then the mean square error is given by

$$E[(y - \hat{f}(x))^2]$$

(average square error in predicting  $y$  for many such unseen points)

- We can show that

$$\begin{aligned} E[(y - \hat{f}(x))^2] &= Bias^2 \\ &+ Variance \\ &+ \sigma^2 \text{ (irreducible error)} \end{aligned}$$

- Consider a new point  $(x, y)$  which was not seen during training
- If we use the model  $\hat{f}(x)$  to predict the value of  $y$  then the mean square error is given by

$$E[(y - \hat{f}(x))^2]$$

(average square error in predicting  $y$  for many such unseen points)

- We can show that

$$\begin{aligned} E[(y - \hat{f}(x))^2] &= \textit{Bias}^2 \\ &+ \textit{Variance} \\ &+ \sigma^2 \text{ (irreducible error)} \end{aligned}$$

- [See proof here](#)

- Consider a new point  $(x, y)$  which was not seen during training
- If we use the model  $\hat{f}(x)$  to predict the value of  $y$  then the mean square error is given by

$$E[(y - \hat{f}(x))^2]$$

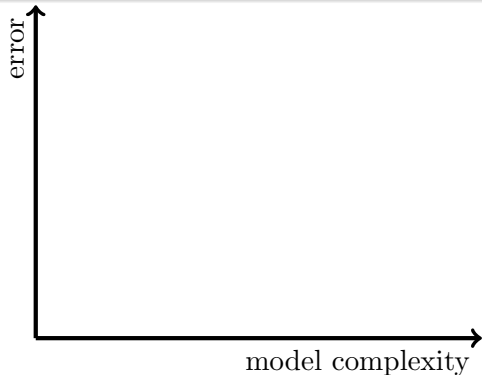
(average square error in predicting  $y$  for many such unseen points)

- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$

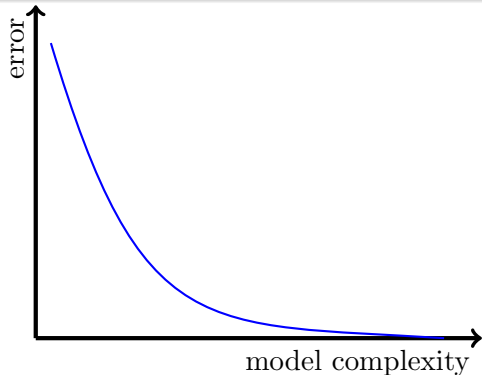
- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training

- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:  
*train<sub>err</sub>* (say, mean square error)  
*test<sub>err</sub>* (say, mean square error)

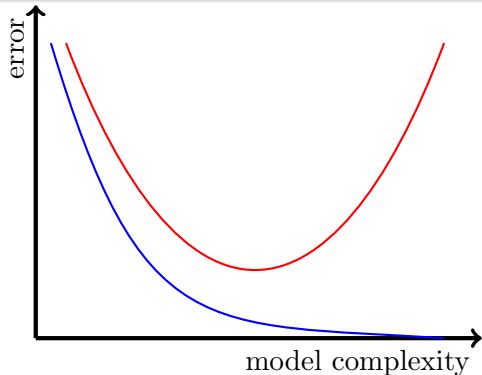




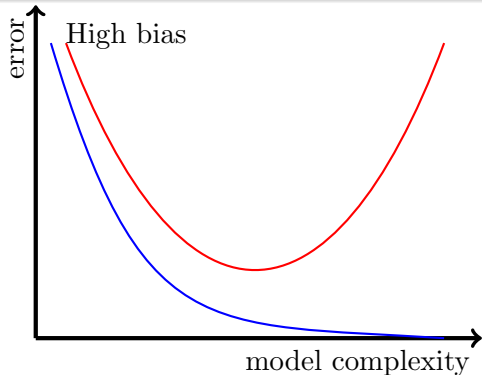
- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:  
*train<sub>err</sub>* (say, mean square error)  
*test<sub>err</sub>* (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



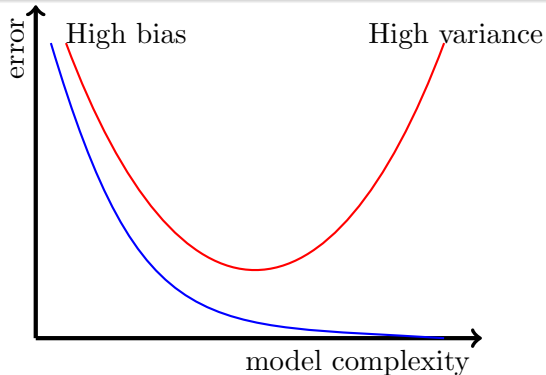
- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:
  - $train_{err}$  (say, mean square error)
  - $test_{err}$  (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



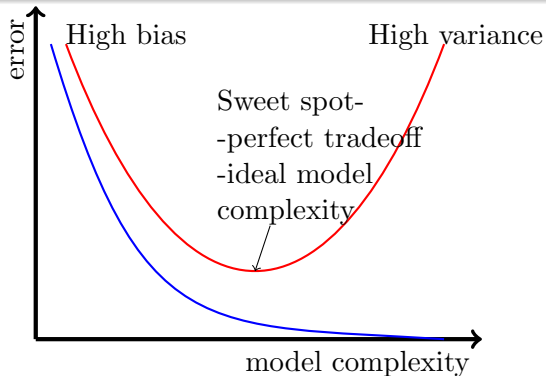
- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:  
*train<sub>err</sub>* (say, mean square error)  
*test<sub>err</sub>* (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



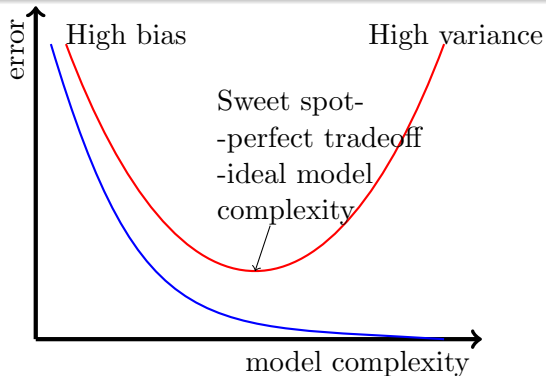
- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:  
*train<sub>err</sub>* (say, mean square error)  
*test<sub>err</sub>* (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:  
*train<sub>err</sub>* (say, mean square error)  
*test<sub>err</sub>* (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:  
*train<sub>err</sub>* (say, mean square error)  
*test<sub>err</sub>* (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure



$$\begin{aligned}
 E[(y - \hat{f}(x))^2] &= Bias^2 \\
 &+ Variance \\
 &+ \sigma^2 \text{ (irreducible error)}
 \end{aligned}$$

- The parameters of  $\hat{f}(x)$  (all  $w_i$ 's) are trained using a training set  $\{(x_i, y_i)\}_{i=1}^n$
- However, at test time we are interested in evaluating the model on a validation (unseen) set which was not used for training
- This gives rise to the following two entities of interest:  
*train<sub>err</sub>* (say, mean square error)  
*test<sub>err</sub>* (say, mean square error)
- Typically these errors exhibit the trend shown in the adjacent figure

## Intuitions developed so far

- Let there be  $n$  training points and  $m$  test (validation) points

$$train_{err} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m} \sum_{i=n+1}^{n+m} (y_i - \hat{f}(x_i))$$



## Intuitions developed so far

- Let there be  $n$  training points and  $m$  test (validation) points

$$train_{err} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m} \sum_{i=n+1}^{n+m} (y_i - \hat{f}(x_i))$$

- As the model complexity increases  $train_{err}$  becomes overly optimistic and gives us a wrong picture of how close  $\hat{f}$  is to  $f$

## Intuitions developed so far

- Let there be  $n$  training points and  $m$  test (validation) points

$$train_{err} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m} \sum_{i=n+1}^{n+m} (y_i - \hat{f}(x_i))$$

- As the model complexity increases  $train_{err}$  becomes overly optimistic and gives us a wrong picture of how close  $\hat{f}$  is to  $f$
- The validation error gives the real picture of how close  $\hat{f}$  is to  $f$

## Intuitions developed so far

- Let there be  $n$  training points and  $m$  test (validation) points

$$train_{err} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

$$test_{err} = \frac{1}{m} \sum_{i=n+1}^{n+m} (y_i - \hat{f}(x_i))$$

- As the model complexity increases  $train_{err}$  becomes overly optimistic and gives us a wrong picture of how close  $\hat{f}$  is to  $f$
- The validation error gives the real picture of how close  $\hat{f}$  is to  $f$
- We will concretize this intuition mathematically now and eventually show how to account for the optimism in the training error

- Let  $D = \{x_i, y_i\}_{i=1}^{m+n}$ , then for any point  $(x, y)$  we have,

$$y_i = f(x_i) + \varepsilon_i$$

- Let  $D = \{x_i, y_i\}_{i=1}^{m+n}$ , then for any point  $(x, y)$  we have,

$$y_i = f(x_i) + \varepsilon_i$$

- which means that  $y_i$  is related to  $x_i$  by some true function  $f$  but there is also some noise  $\varepsilon$  in the relation

- Let  $D = \{x_i, y_i\}_{i=1}^{m+n}$ , then for any point  $(x, y)$  we have,

$$y_i = f(x_i) + \varepsilon_i$$

- which means that  $y_i$  is related to  $x_i$  by some true function  $f$  but there is also some noise  $\varepsilon$  in the relation
- For simplicity, we assume

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

- Let  $D = \{x_i, y_i\}_{i=1}^{m+n}$ , then for any point  $(x, y)$  we have,

$$y_i = f(x_i) + \varepsilon_i$$

- which means that  $y_i$  is related to  $x_i$  by some true function  $f$  but there is also some noise  $\varepsilon$  in the relation
- For simplicity, we assume

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

and of course we do not know  $f$

- Let  $D = \{x_i, y_i\}_{i=1}^{m+n}$ , then for any point  $(x, y)$  we have,

$$y_i = f(x_i) + \varepsilon_i$$

- which means that  $y_i$  is related to  $x_i$  by some true function  $f$  but there is also some noise  $\varepsilon$  in the relation
- For simplicity, we assume

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

and of course we do not know  $f$

- Further we use  $\hat{f}$  to approximate  $f$  and estimate the parameters using  $T \subset D$  such that

$$y_i = \hat{f}(x_i)$$



- Let  $D = \{x_i, y_i\}_{i=1}^{m+n}$ , then for any point  $(x, y)$  we have,

$$y_i = f(x_i) + \varepsilon_i$$

- which means that  $y_i$  is related to  $x_i$  by some true function  $f$  but there is also some noise  $\varepsilon$  in the relation
- For simplicity, we assume

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

and of course we do not know  $f$

- Further we use  $\hat{f}$  to approximate  $f$  and estimate the parameters using  $T \subset D$  such that

$$y_i = \hat{f}(x_i)$$

- We are interested in knowing

$$E[(\hat{f}(x_i) - f(x_i))^2]$$

- Let  $D = \{x_i, y_i\}_{i=1}^{m+n}$ , then for any point  $(x, y)$  we have,

$$y_i = f(x_i) + \varepsilon_i$$

- which means that  $y_i$  is related to  $x_i$  by some true function  $f$  but there is also some noise  $\varepsilon$  in the relation

- For simplicity, we assume

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

and of course we do not know  $f$

- Further we use  $\hat{f}$  to approximate  $f$  and estimate the parameters using  $T \subset D$  such that

$$y_i = \hat{f}(x_i)$$

- We are interested in knowing

$$E[(\hat{f}(x_i) - f(x_i))^2]$$

but we cannot estimate this directly because we do not know  $f$

- Let  $D = \{x_i, y_i\}_{i=1}^{m+n}$ , then for any point  $(x, y)$  we have,

$$y_i = f(x_i) + \varepsilon_i$$

- which means that  $y_i$  is related to  $x_i$  by some true function  $f$  but there is also some noise  $\varepsilon$  in the relation

- For simplicity, we assume

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

and of course we do not know  $f$

- Further we use  $\hat{f}$  to approximate  $f$  and estimate the parameters using  $T \subset D$  such that

$$y_i = \hat{f}(x_i)$$

- We are interested in knowing

$$E[(\hat{f}(x_i) - f(x_i))^2]$$

but we cannot estimate this directly because we do not know  $f$

- We will see how to estimate this empirically using the observation  $y_i$  & prediction  $\hat{y}_i$

$$E[(\hat{y}_i - y_i)^2]$$

$$E[(\hat{y}_i - y_i)^2] = E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i)$$

$$\begin{aligned} E[(\hat{y}_i - y_i)^2] &= E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i) \\ &= E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2] \end{aligned}$$

$$E[(\hat{y}_i - y_i)^2] = E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i)$$

$$= E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2]$$

$$= E[(\hat{f}(x_i) - f(x_i))^2] - 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2]$$

$$E[(\hat{y}_i - y_i)^2] = E[(\hat{f}(x_i) - f(x_i) - \varepsilon_i)^2] \quad (y_i = f(x_i) + \varepsilon_i)$$

$$= E[(\hat{f}(x_i) - f(x_i))^2 - 2\varepsilon_i(\hat{f}(x_i) - f(x_i)) + \varepsilon_i^2]$$

$$= E[(\hat{f}(x_i) - f(x_i))^2] - 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))] + E[\varepsilon_i^2]$$

$$\therefore E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$



We will take a small detour to understand how to empirically estimate an Expectation and then return to our derivation

- Suppose we have observed the goals scored( $z$ ) in  $k$  matches as  $z_1 = 2, z_2 = 1, z_3 = 0, \dots z_k = 2$

- Suppose we have observed the goals scored( $z$ ) in  $k$  matches as  $z_1 = 2, z_2 = 1, z_3 = 0, \dots z_k = 2$
- Now we can empirically estimate  $E[z]$  i.e. the expected number of goals scored as

$$E[z] = \frac{1}{k} \sum_{i=1}^k z_i$$

- Suppose we have observed the goals scored( $z$ ) in  $k$  matches as  $z_1 = 2, z_2 = 1, z_3 = 0, \dots z_k = 2$
- Now we can empirically estimate  $E[z]$  i.e. the expected number of goals scored as

$$E[z] = \frac{1}{k} \sum_{i=1}^k z_i$$

- Analogy with our derivation: We have a certain number of observations  $y_i$  & predictions  $\hat{y}_i$  using which we can estimate

$$E[(\hat{y}_i - y_i)^2] =$$

- Suppose we have observed the goals scored( $z$ ) in  $k$  matches as  $z_1 = 2, z_2 = 1, z_3 = 0, \dots z_k = 2$
- Now we can empirically estimate  $E[z]$  i.e. the expected number of goals scored as

$$E[z] = \frac{1}{k} \sum_{i=1}^k z_i$$

- Analogy with our derivation: We have a certain number of observations  $y_i$  & predictions  $\hat{y}_i$  using which we can estimate

$$E[(\hat{y}_i - y_i)^2] = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2$$

... returning back to our derivation

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations



$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}}$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} -$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}}$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance } (\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$\therefore \text{covariance}(X, Y)$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$$\therefore \text{covariance}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$$\begin{aligned} \because \text{covariance}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[(X)(Y - \mu_Y)] \text{ (if } \mu_X = E[X] = 0) \end{aligned}$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$$\begin{aligned} \because \text{covariance}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[(X)(Y - \mu_Y)] \text{ (if } \mu_X = E[X] = 0) \\ &= E[XY] - E[X\mu_Y] \end{aligned}$$



$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$$\begin{aligned} \because \text{covariance}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[(X)(Y - \mu_Y)] \text{ (if } \mu_X = E[X] = 0) \\ &= E[XY] - E[X\mu_Y] = E[XY] - \mu_Y E[X] \end{aligned}$$

$$E[(\hat{f}(x_i) - f(x_i))^2] = E[(\hat{y}_i - y_i)^2] - E[\varepsilon_i^2] + 2E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]$$

- We can empirically evaluate R.H.S using training observations or test observations

**Case 1:** Using test observations

$$\underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} = \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}$$

$$\begin{aligned} \because \text{covariance}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[(X)(Y - \mu_Y)] \text{ (if } \mu_X = E[X] = 0) \\ &= E[XY] - E[X\mu_Y] = E[XY] - \mu_Y E[X] = E[XY] \end{aligned}$$

$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
= & \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
&= \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]

$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
&= \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]

$$\therefore \varepsilon \perp (\hat{f}(x_i) - f(x_i))$$

$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
&= \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]

$$\therefore \varepsilon \perp (\hat{f}(x_i) - f(x_i))$$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] = 0$$

$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
&= \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]

$$\therefore \varepsilon \perp (\hat{f}(x_i) - f(x_i))$$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)]$$

$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
&= \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]

$$\therefore \varepsilon \perp (\hat{f}(x_i) - f(x_i))$$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] = 0 \cdot E[\hat{f}(x_i) - f(x_i)]$$



$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
&= \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]

$$\therefore \varepsilon \perp (\hat{f}(x_i) - f(x_i))$$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] = 0 \cdot E[\hat{f}(x_i) - f(x_i)] = 0$$

$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
&= \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]

$$\therefore \varepsilon \perp (\hat{f}(x_i) - f(x_i))$$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] = 0 \cdot E[\hat{f}(x_i) - f(x_i)] = 0$$

$$\therefore \text{true error} = \text{empirical test error} + \text{small constant}$$

$$\begin{aligned}
& \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\
&= \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{m} \sum_{i=n+1}^{n+m} \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))}
\end{aligned}$$

- None of the test observations participated in the estimation of  $\hat{f}(x)$  [the parameters of  $\hat{f}(x)$  were estimated only using training data]

$$\therefore \varepsilon \perp (\hat{f}(x_i) - f(x_i))$$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] = E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] = 0 \cdot E[\hat{f}(x_i) - f(x_i)] = 0$$

$$\therefore \text{true error} = \text{empirical test error} + \text{small constant}$$

- Hence, we should always use a validation set (independent of the training set) to estimate the error

## Case 2: Using training observations

$$\begin{aligned} & \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))} \end{aligned}$$

## Case 2: Using training observations

$$\begin{aligned} & \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))} \end{aligned}$$

Now,  $\varepsilon \not\perp \hat{f}(x)$  because  $\varepsilon$  was used for estimating the parameters of  $\hat{f}(x)$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))]$$

## Case 2: Using training observations

$$\begin{aligned} & \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))} \end{aligned}$$

Now,  $\varepsilon \not\perp \hat{f}(x)$  because  $\varepsilon$  was used for estimating the parameters of  $\hat{f}(x)$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))]$$

## Case 2: Using training observations

$$\begin{aligned} & \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))} \end{aligned}$$

Now,  $\varepsilon \not\perp \hat{f}(x)$  because  $\varepsilon$  was used for estimating the parameters of  $\hat{f}(x)$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] \neq E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)]$$

## Case 2: Using training observations

$$\begin{aligned} & \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))} \end{aligned}$$

Now,  $\varepsilon \not\perp \hat{f}(x)$  because  $\varepsilon$  was used for estimating the parameters of  $\hat{f}(x)$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] \neq E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] \neq 0$$

Hence, the empirical train error is smaller than the true error and does not give a true picture of the error



## Case 2: Using training observations

$$\begin{aligned} & \underbrace{E[(\hat{f}(x_i) - f(x_i))^2]}_{\text{true error}} \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}_{\text{empirical estimation of error}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}_{\text{small constant}} + 2 \underbrace{E[\varepsilon_i(\hat{f}(x_i) - f(x_i))]}_{= \text{covariance}(\varepsilon_i, \hat{f}(x_i) - f(x_i))} \end{aligned}$$

Now,  $\varepsilon \not\perp \hat{f}(x)$  because  $\varepsilon$  was used for estimating the parameters of  $\hat{f}(x)$

$$\therefore E[\varepsilon_i \cdot (\hat{f}(x_i) - f(x_i))] \neq E[\varepsilon_i] \cdot E[\hat{f}(x_i) - f(x_i)] \neq 0$$

Hence, the empirical train error is smaller than the true error and does not give a true picture of the error

But how is this related to model complexity? Let us see