Validation

Sewoong Oh

CSE446
University of Washington

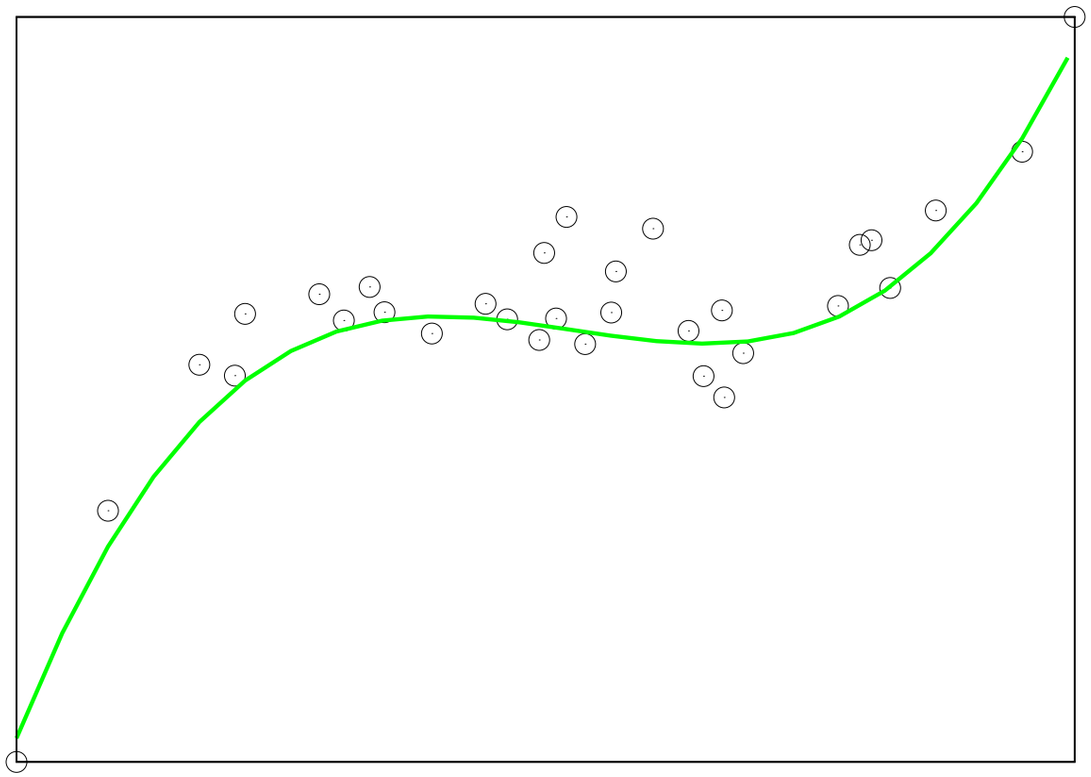
Generalization:
how do we know which model is better in predicting unseen data?

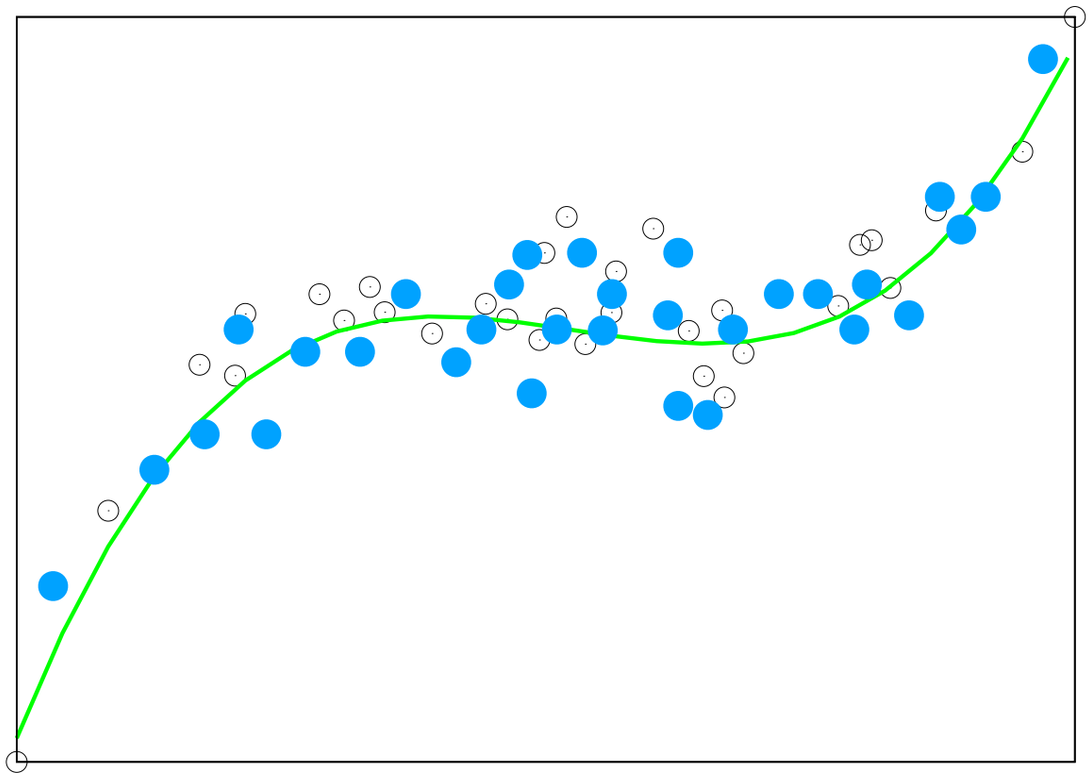
Generalization

- we say a predictor **generalizes** if it performs well on unseen data
- formal mathematical definition involves probabilistic assumptions
- first, we study practical methods for assessing generalization

In-sample and out-of-sample data

- the data used to train a predictor is **training data** or **in-sample data**
- we want the predictor to work on **out-of-sample data**
- we say a predictor **fails to generalize** if it does not perform well on out-of-sample data





- **train** a cubic predictor on 32 (in-sample) white circles: MSE 174
- **predict** *y* for 30 (out-of-sample) blue circles: MSE 192
- conclude this predictor generalizes, as in-sample MSE $\simeq$ out-of-sample MSE

Out-of-sample Validation

- a way to mimic how the predictor performs on unseen data
- key idea: divide the data into two set for **training** and **testing**

- **training set** used to construct (“train”) the predictor
- **test set** used to evaluate the predictor

- we assume that test set is similar to unseen data
- test set should never be used in training

Out-of-sample Validation

- given a single dataset $S = \{(x_i, y_i)\}_{i=1}^{n}$
- we split the dataset into two: training set and test set
- selection of data train/test should be done randomly (80/20 or 90/10 are common)
- we use **training error** (i.e. empirical risk on training dataset) for optimization (or finding the model)

$$\text{minimize} \quad \mathscr{L}_{\text{train}}(w) = \frac{1}{|S_{\text{train}}|} \sum_{i \in S_{\text{train}}} \ell(f(x_i), y_i)$$

- we use **test error** (i.e. empirical risk on test dataset) for **validation**, checking if the model behaves as expected

$$\mathscr{L}_{\text{test}}(w) = \frac{1}{|S_{\text{test}}|} \sum_{i \in S_{\text{test}}} \ell(f(x_i), y_i)$$

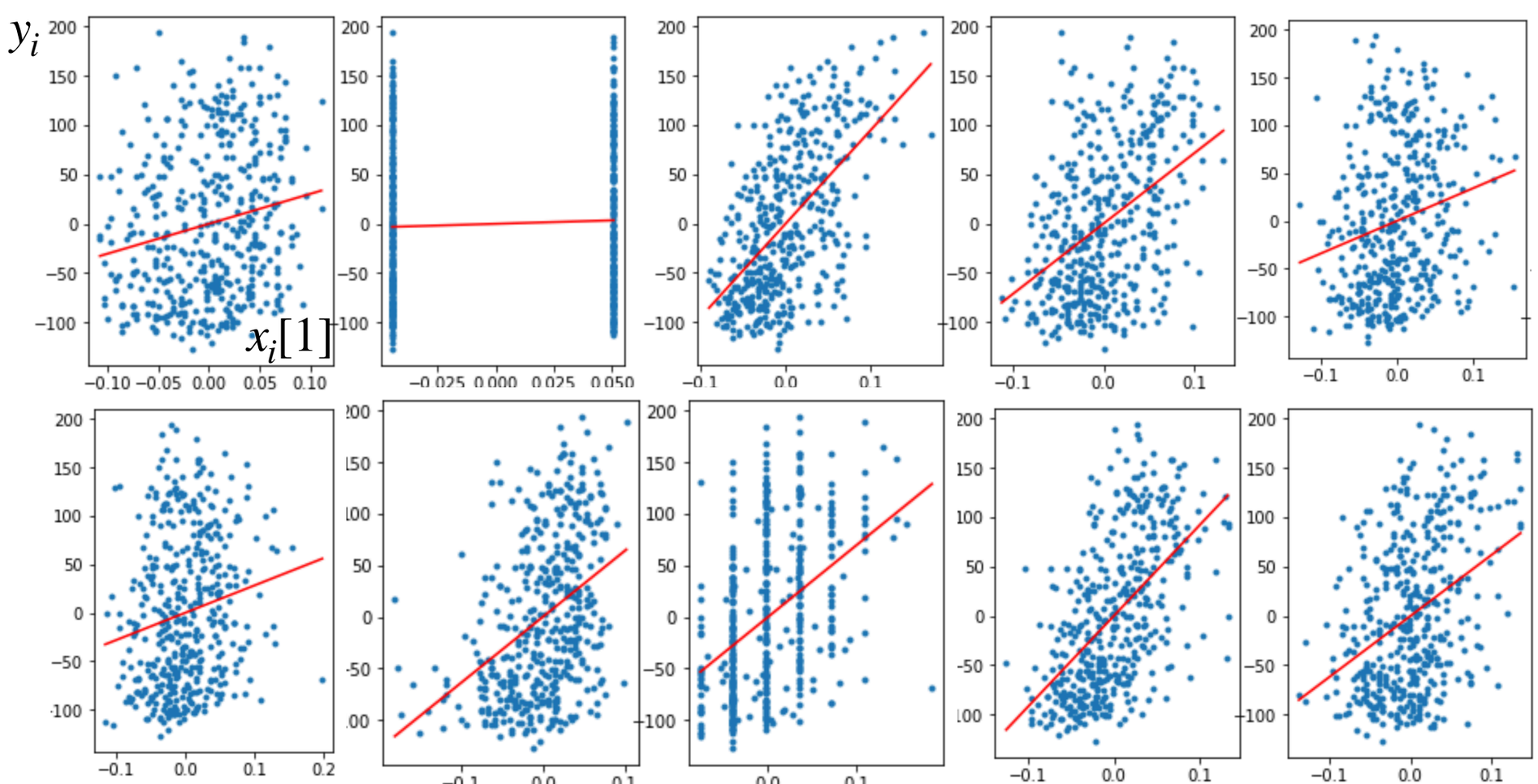
- we say a model or predictor is **overfit** if

$$\mathscr{L}_{\text{train}} \ll \mathscr{L}_{\text{test}}$$

	small training error	large training error
small test error	generalizes well performs well	possible, but unlikely
large test error	fails to generalize	generalizes well performs poorly

Choosing a predictor

- validation is useful in choosing a predictor
- typically, one trains multiple candidate predictors and chooses the predictor that has the smallest test error
- Example: Diabetes
  - 10 explanatory variables
  - from 442 patients
  - we use half for train and half for validation



Example: Diabetes

Features	Train MSE	Test MSE
All	2640	3224
S5 and BMI	3004	3453
S5	3869	4227
BMI	3540	4277
S4 and S3	4251	5302
S4	4278	5409
S3	4607	5419
None	5524	6352

- **test MSE is the primary criteria for model selection**
- Using only 2 features (S5 and BMI), one can get very close to the prediction performance of using all features
- Combining S3 and S4 does not give any performance gain

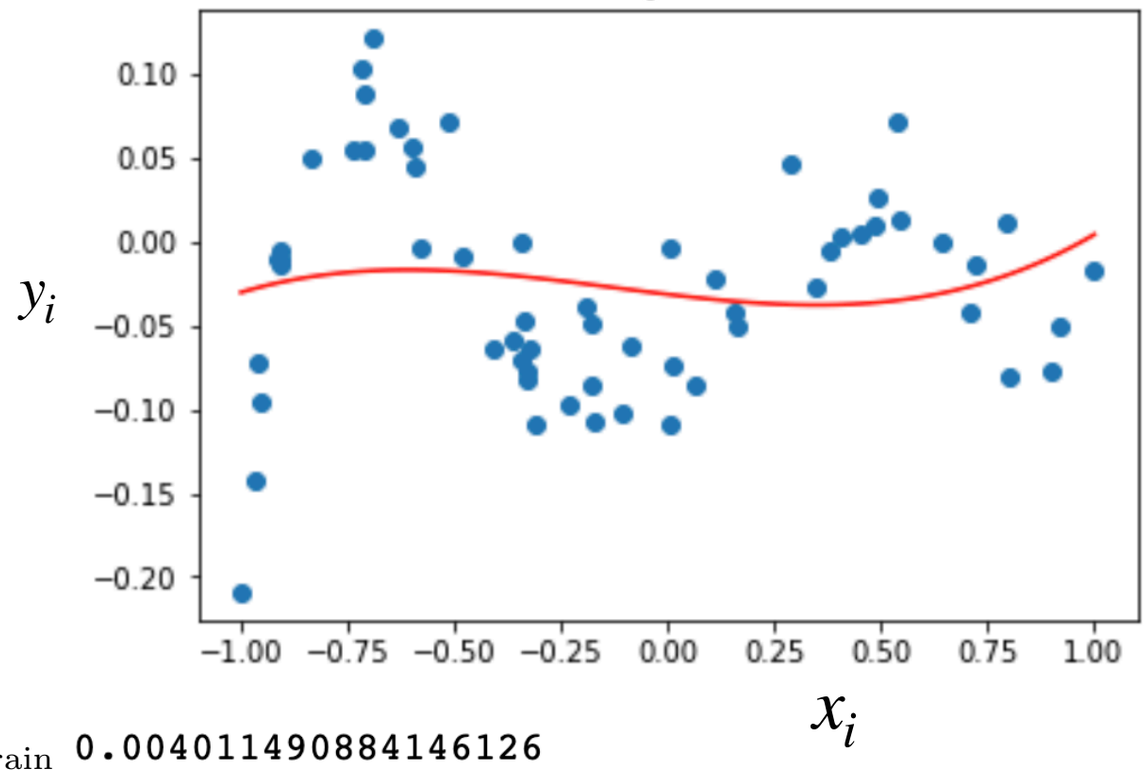
Overfitting

- a model that fits the training data well but performs poorly on test data suffers from **overfitting**
- overfitting happens if we use a model with high **model complexity**
- for example, for linear regression with polynomial features

$$\hat{y} \;=\; f(x) \;=\; w_0 + w_1 x + w_2 x^2 + \cdots + w_p x^p$$

- N = 60 data points, and $p \in \{3, 4, 5, 20\}$

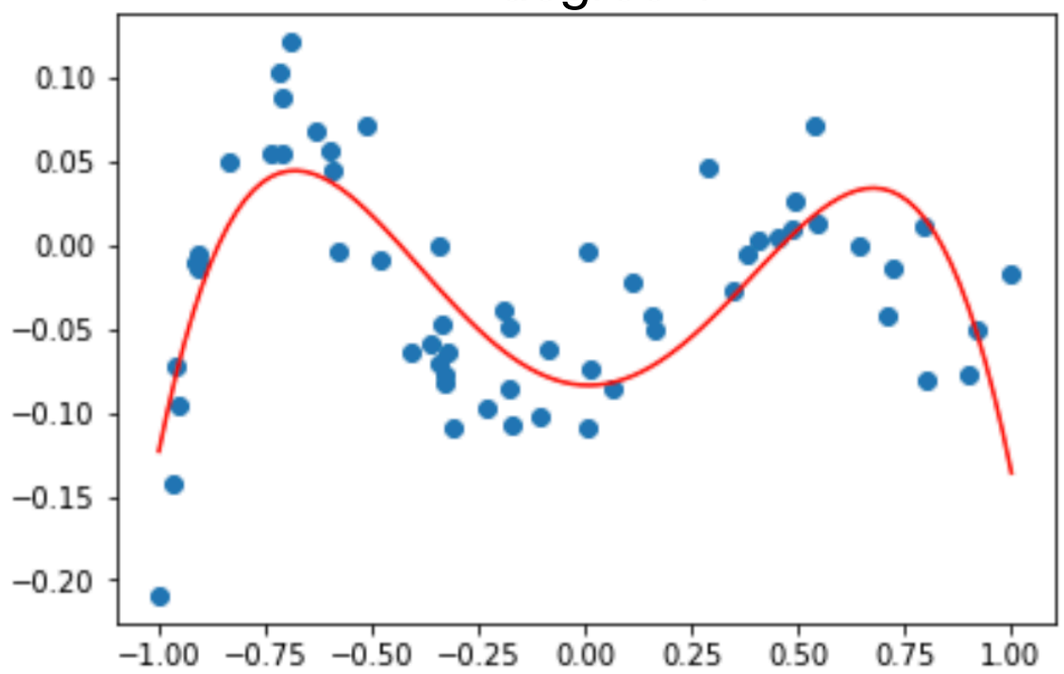
degree 3



$\text{MSE}_{\text{train}}$ `0.004011490884146126`

MSE_{test} `0.003831010290504173`

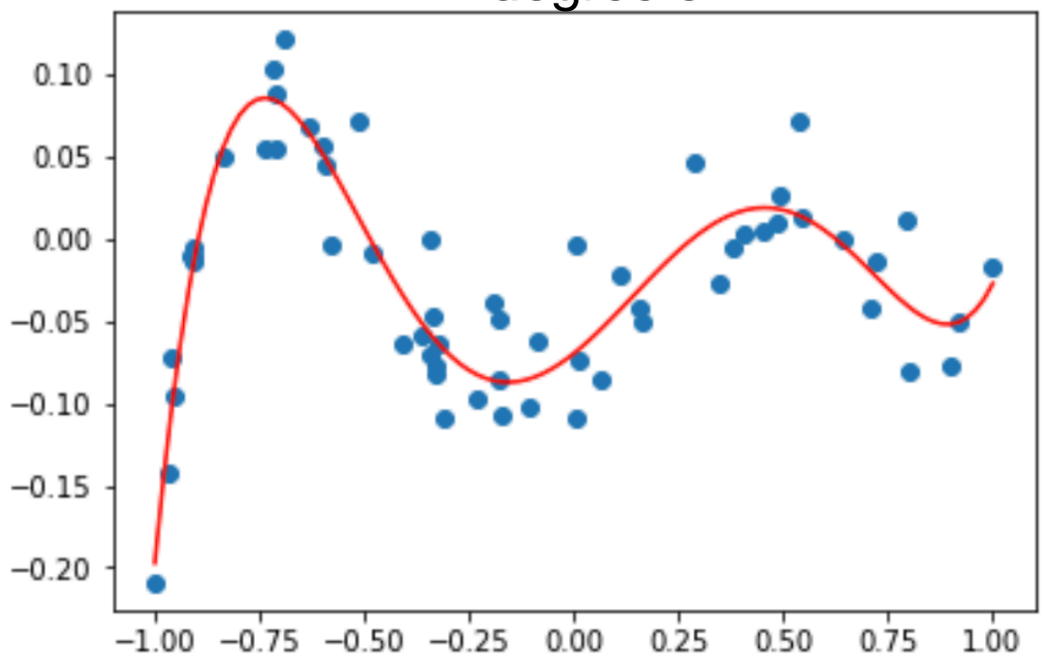
degree 4



`0.0019177229761741974`

`0.002200492720447942`

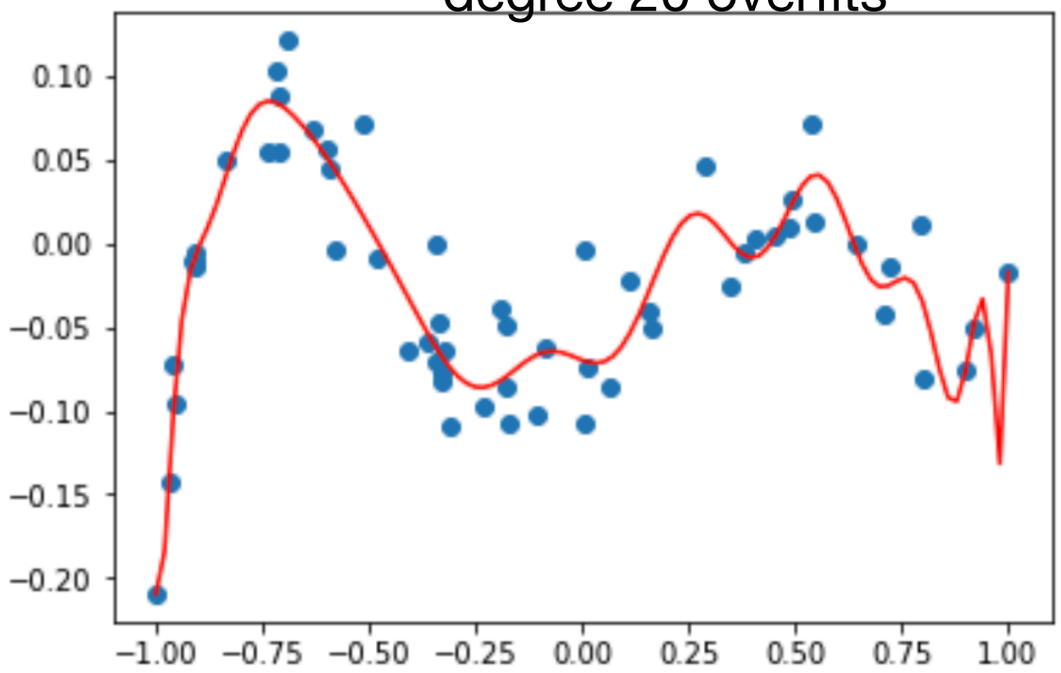
degree 5



`0.0007426853089970962`

`0.0010239915404334238`

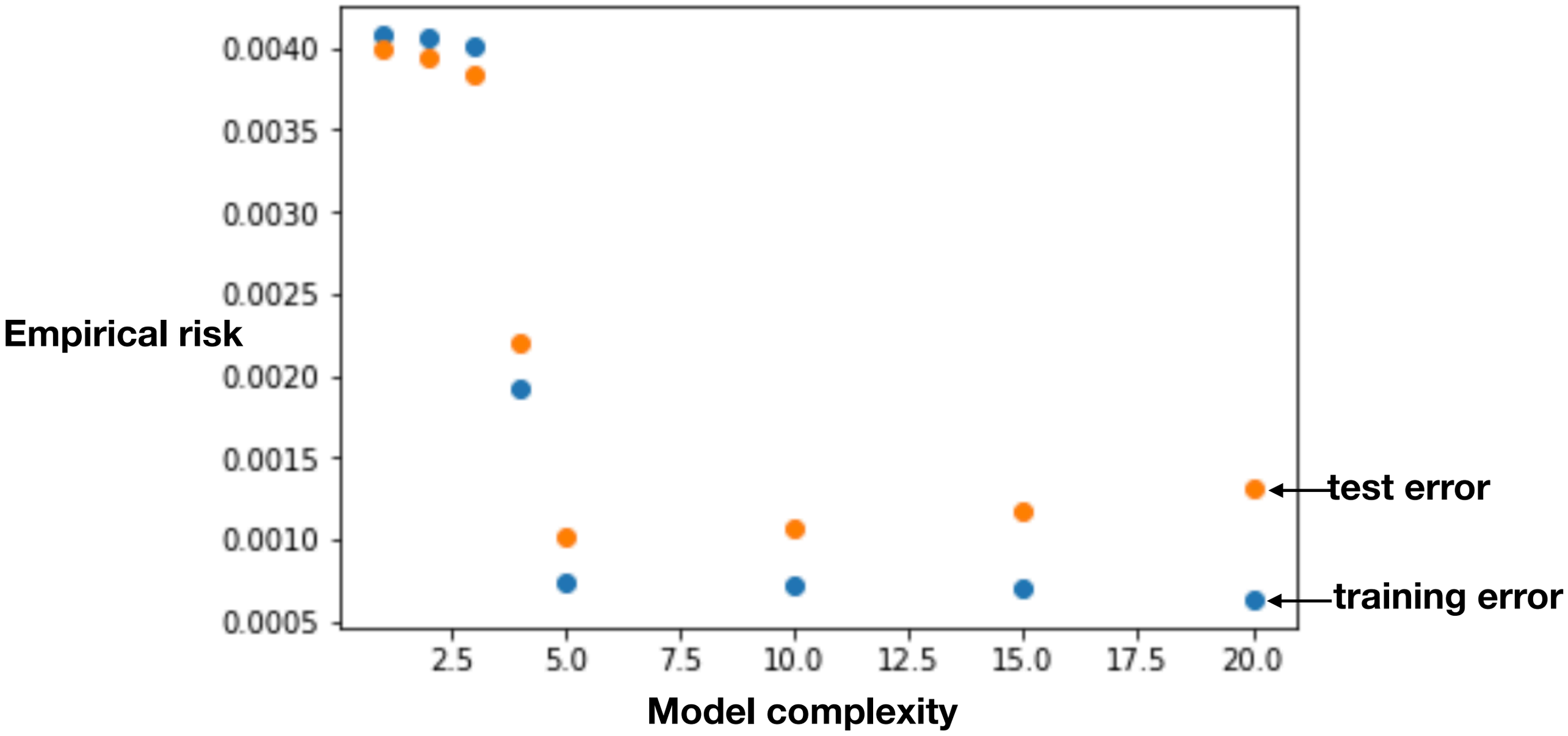
degree 20 overfits



`0.0006350545819906561`

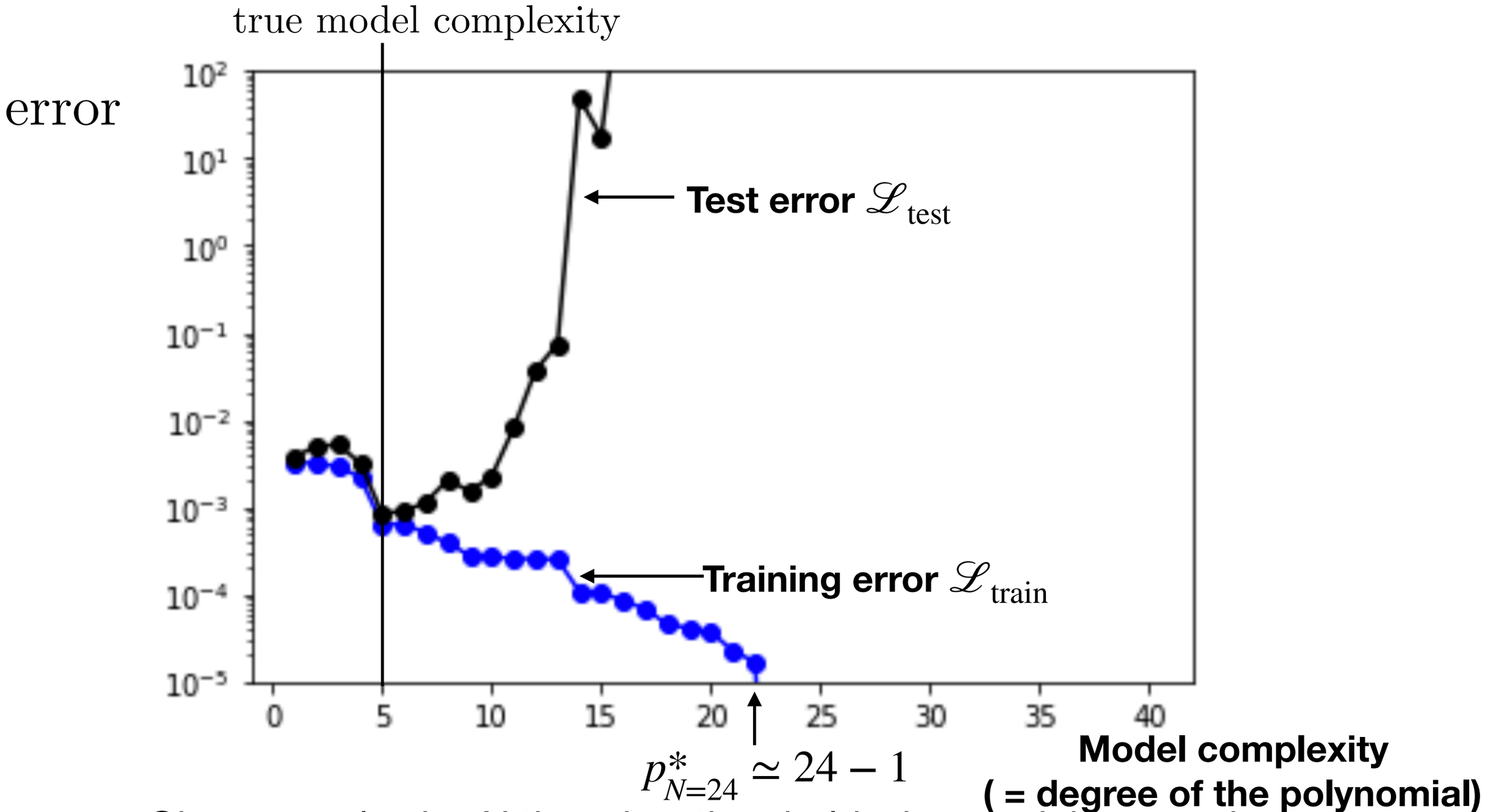
`0.0013118370515986446`

How does one choose which model to use?



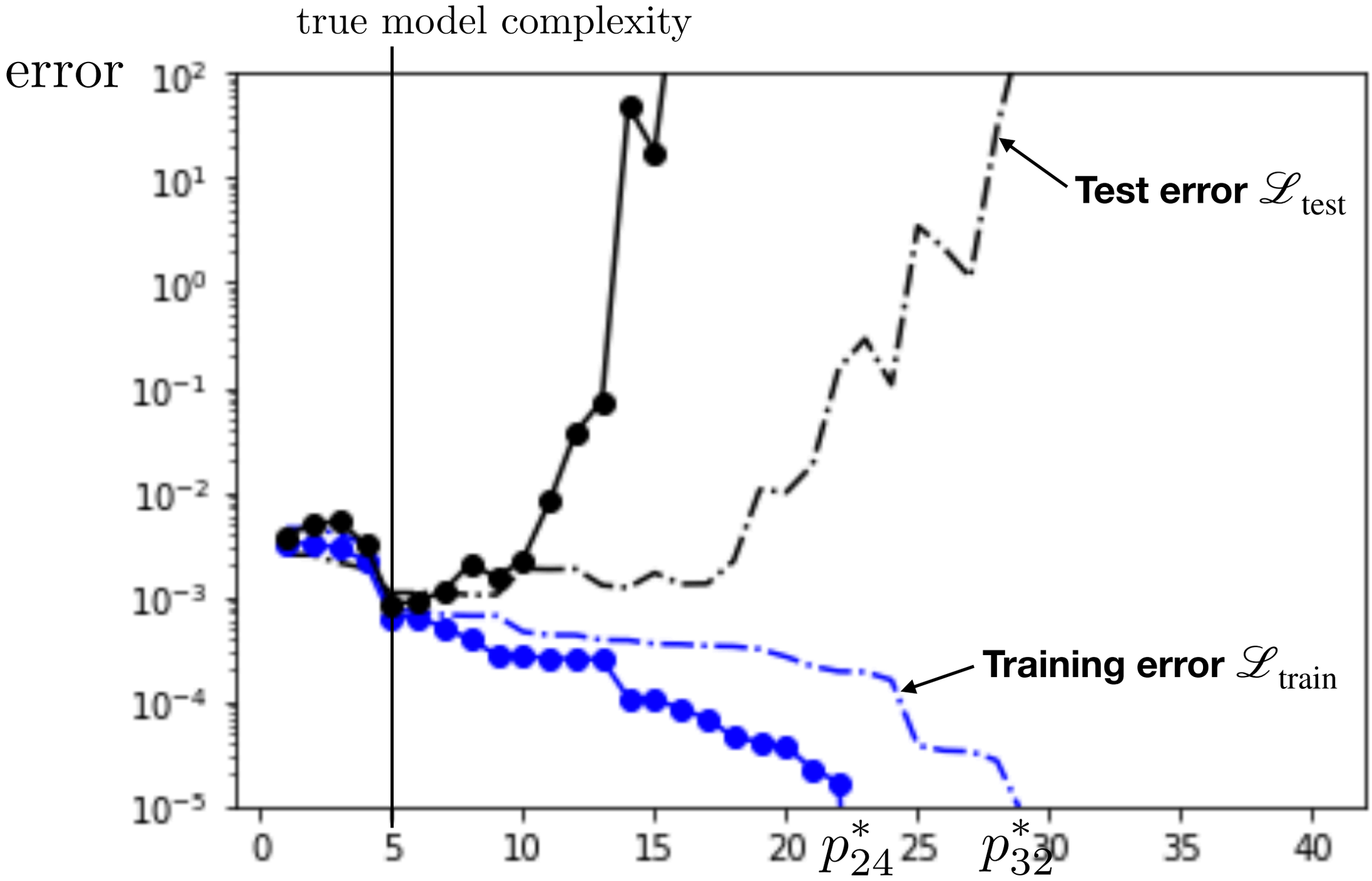
- first use 60 data points to train and 60 data points to test
- then choose degree 5 as per the above test error
- now re-train on all 120 data points with degree 5 polynomial model

- let us first fix sample size ***N=30***, collect one dataset of size N, randomly shuffle the dataset, 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



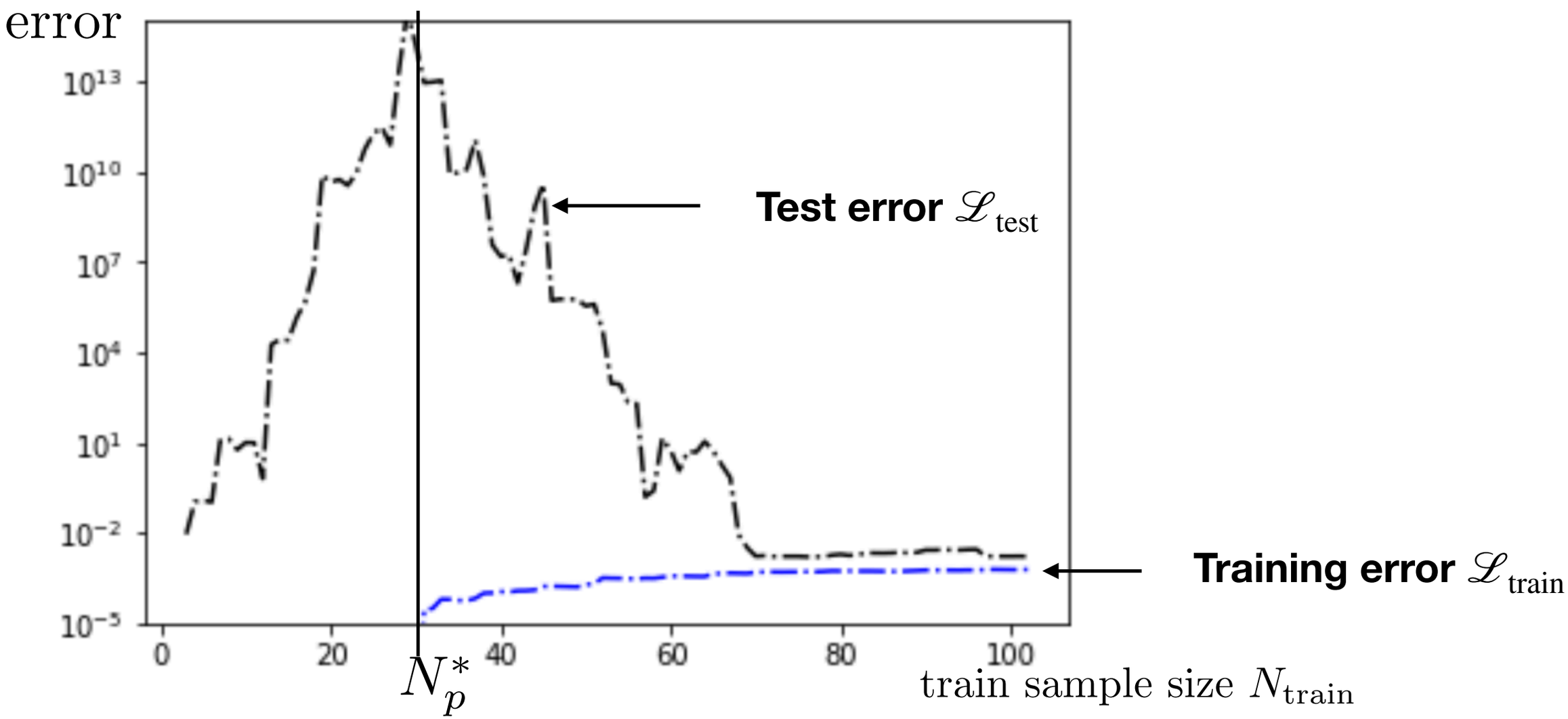
- Given sample size N there is a threshold where training error is zero
- Training error is **always** monotonically non-increasing
- Test error has a trend of going down and then up, but fluctuates

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



- The threshold moves right
- Training error tends to increase: more points need to fit
- Test error tends to decrease: overfitting happens later

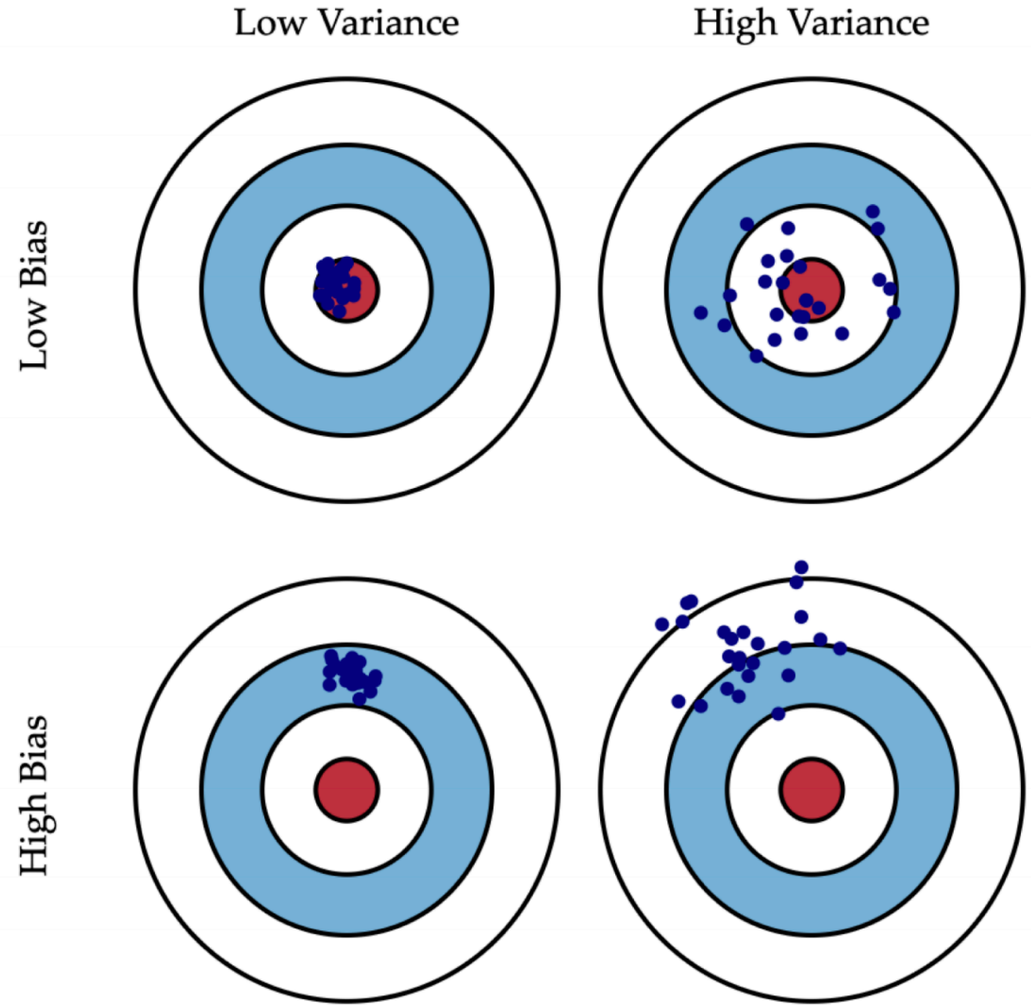
- let us now fix predictor model complexity ***p=30***, collect multiple datasets by starting with 3 samples and adding one sample at a time to the training set, but keeping a large enough test set fixed
- then we run multiple validations and plot the computed MSEs for all values of train sample size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



- There is a threshold below which training error is zero (extreme overfit)
- Below this threshold, test error is meaningless, as there are multiple predictors with zero training error
- Test error tends to decrease
- Training error tends to increase

why do they meet?

From practice to theory



Courtesy of Scott Fortmann-Roe

Notations

- the model is specified by the distribution of data $p_{x,y}$ for the paired examples (x_i, y_i)
- we denote our predictor by $f_{S_{\text{train}}}(x)$ to emphasize that our predictor depends on the training data $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$ of size n , each coming i.i.d. from distribution $p_{x,y}$
- we denote the test data by $S_{\text{test}} = \{(x_i, y_i)\}_{i=n+1}^{n+m}$ of size m, also i.i.d. from $p_{x,y}$
- when we take expectation of a function of random variables, we use the following notation to indicate what we are taking expectation of:

  $$\mathbb{E}_{(x,y)\sim p_{x,y}}[F(x, y)]$$

  indicating that the pair (x, y) is drawn from $p_{x,y}$
- we will simplify the subscript, whenever it is clear from the context, for example we might write

  $$\mathbb{E}_{p_{x,y}}[F(x, y)]$$

  or even

  $$\mathbb{E}[F(x, y)]$$

Expected test error

- goal of training a predictor is to get test error small, defined as the empirical risk on the test set

$$\mathscr{L}_{\text{test}} = \frac{1}{|S_{\text{test}}|} \sum_{i \in S_{\text{test}}} (f_{S_{\text{train}}}(x_i) - y_i)^2$$

because this is a surrogate of the expected error on (randomly chosen) unseen data

- the expected mean squared error (true error) on a new data (x, y) is defined as

$$\begin{aligned}
\mathscr{L}_{\text{true}} &= \mathbb{E}_{S_{\text{test}} \sim p_{x,y}^m, S_{\text{train}} \sim p_{x,y}^n}[\mathscr{L}_{\text{test}}] \\
&= \mathbb{E}_{S_{\text{test}} \sim p_{x,y}^m, S_{\text{train}} \sim p_{x,y}^n}\left[\frac{1}{m} \sum_{i=n+1}^{n+m} (f_{S_{\text{train}}}(x_i) - y_i)^2\right] \\
&= \mathbb{E}_{(x,y) \sim p_{x,y}, S_{\text{train}} \sim p_{x,y}^n}[(f_{S_{\text{train}}}(x) - y)^2]
\end{aligned}$$

where the last line follows from the i.i.d. assumption, and this true error is what we really care about, and hope to minimize

- for simplicity, we will write

$$\mathscr{L}_{\text{true}} = \mathbb{E}_{p_{x,y}, S_{\text{train}}}[(f_{S_{\text{train}}}(x) - y)^2]$$

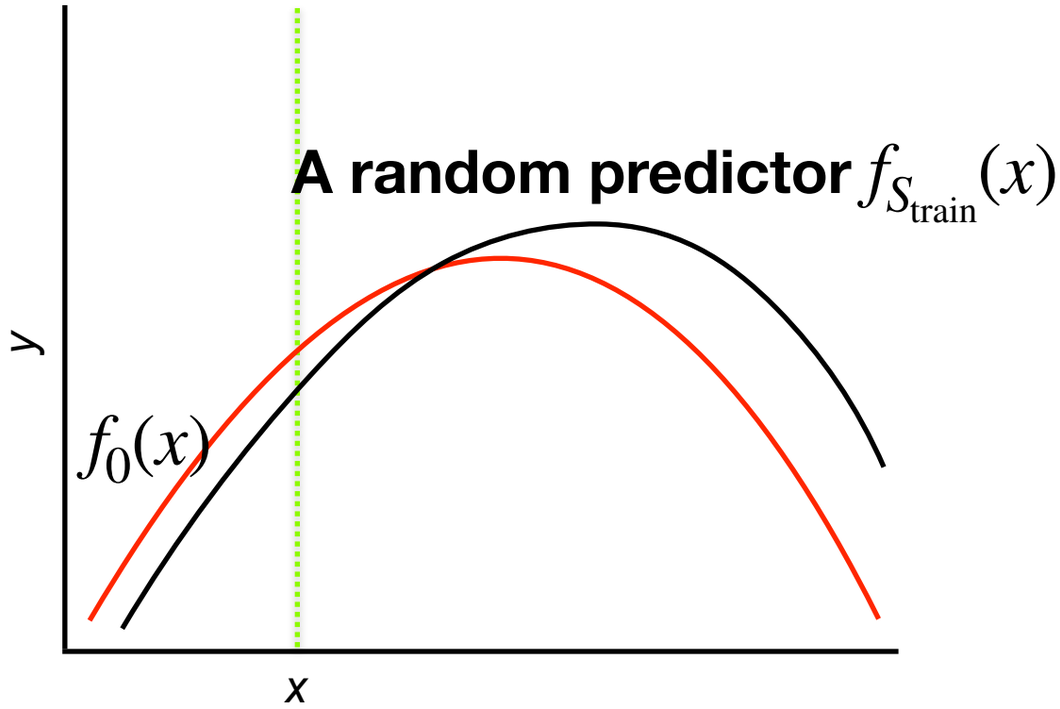
- we will decompose this expected test error, to identify three sources of error

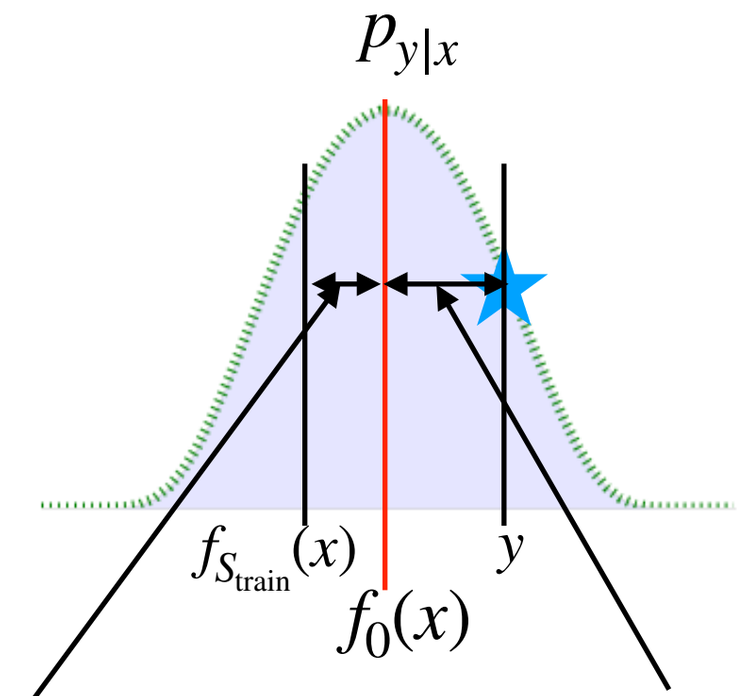
demo4_tradeoff

Canonical model

- recall the law of total expectation (or tower rule):
$$\mathbb{E}_{p_{x,y}}[F(x,y)] = \mathbb{E}_{p_x}\left[\mathbb{E}_{p_{y|x}}[F(x,y)\,|\,x]\right],$$
for any function $F(x,y)$ and any joint distribution $p_{x,y}$
- we focus on analyzing the conditional expectation
$$\mathbb{E}_{p_{y|x},S_{\text{train}}}[(f_{S_{\text{train}}}(x)-y)^2\,|\,x]$$
as by the law of total expectation (or tower rule), we have
$$\mathbb{E}_{p_{x,y},S_{\text{train}}}[(f_{S_{\text{train}}}(x)-y)^2] = \mathbb{E}_{p_x}\left[\mathbb{E}_{p_{y|x},S_{\text{train}}}[(f_{S_{\text{train}}}(x)-y)^2\,|\,x]\right]$$
- the bias-variance tradeoff we show for the conditional expectation, will imply a similar result on the joint expectation $p_{x,y}$ by simply taking the expectation with respect to p_x to the resulting formula (we do this in equation (1) on slide 21)
- this implies that we only need to specify the conditional distribution $p_{y|x}$ to proceed with the analysis, and we will focus on the **canonical model** where
$$y = f_0(x) + \varepsilon\ ,$$
where ε is drawn from $N(0,\sigma^2)$, zero mean Gaussian with variance σ^2
- note that this use of canonical model is without loss of generality, as it the same analysis can be used to capture bias-variance tradeoff for any $p_{x,y}$ but with heavier notations

- model: $y_i = f_0(x_i) + \varepsilon_i$ for both training and test samples





$(f_{S_{\text{train}}} - f_0(x))$: **error due to model mismatch**

$(f_0(x) - y)$: **error due to randomness in the sampling process**

- we will analyze the conditional expectation of the true error $\mathbb{E}_{p_{y|x}, S_{\text{train}}}[(f_{S_{\text{train}}}(x) - y)^2 \mid x]$

Bias-variance tradeoff

- the conditional true error can be written as

$$\mathbb{E}_{p_{y|x},S_{\text{train}}}[(f_{S_{\text{train}}}(x)-y)^2\,|\,x] \;=\; \mathbb{E}\Big[\big(\underbrace{(f_{S_{\text{train}}}(x)-f_0(x))}_{A}-\underbrace{(y-f_0(x))}_{B}\big)^2\,|\,x\Big]$$

$$= \underbrace{\mathbb{E}_{S_{\text{train}}}[(f_{S_{\text{train}}}(x)-f_0(x))^2\,|\,x]}_{\text{learning error}\geq 0} + \underbrace{\mathbb{E}_{p_{y|x}}[(f_0(x)-y)^2\,|\,x]}_{\text{irreducible error}=\sigma^2}$$

this follows from the fact that $\mathbb{E}[(A-B)^2]=\mathbb{E}[A^2]-2\mathbb{E}[AB]+\mathbb{E}[B^2]$ and the noise is zero mean, i.e.

$$\mathbb{E}[AB] \;=\; \mathbb{E}[(f_{S_{\text{train}}}(x)-f_0(x))(y-f_0(x))\,|\,x] = \mathbb{E}[f_{S_{\text{train}}}(x)-f_0(x)\,|\,x]\underbrace{\mathbb{E}[\varepsilon\,|\,x]}_{=0} = 0$$

- **irreducible error**
  - is due to the inherent noise in the samples, and is impossible to get rid of
  - does not depend on our predictor $f(x)$
  - Is a lower bound on achievable expected test error
- **learning error**
  - is due to the randomness (and limited sample size) in the training data
  - Further decomposed into **bias** and **variance**

- the learning error can be further decomposed (with a similar trick) as

$$\mathbb{E}_{S_{\text{train}}}[(f_{S_{\text{train}}}(x) - f_0(x))^2 \,|\, x] = \mathbb{E}\Big[\big((f_{S_{\text{train}}}(x) - \mathbb{E}[f_{S_{\text{train}}}(x) \,|\, x]) - (f_0(x) - \mathbb{E}[f_{S_{\text{train}}}(x) \,|\, x])\big)^2 \,|\, x\Big]$$

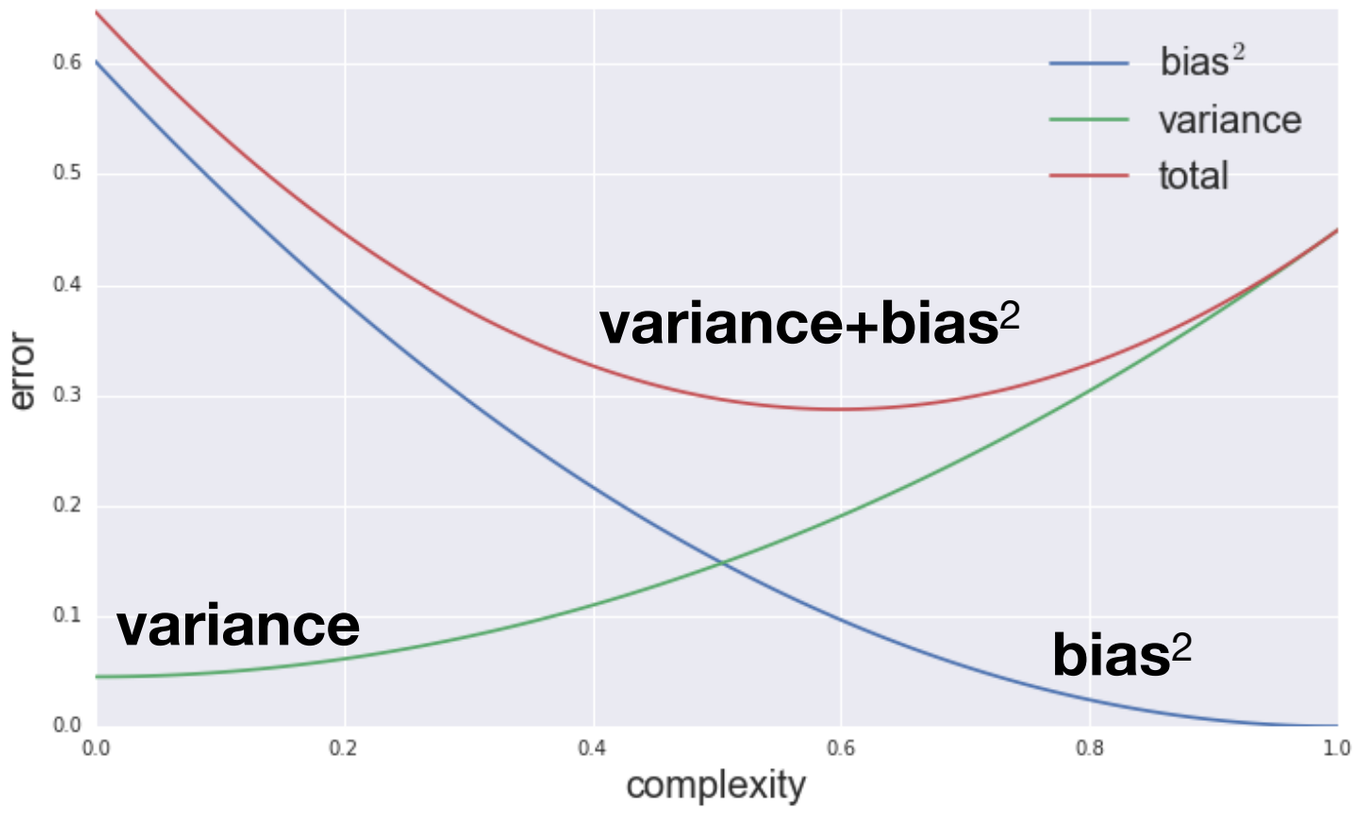
$$= \underbrace{\mathbb{E}\Big[(f_{S_{\text{train}}}(x) - \mathbb{E}[f_{S_{\text{train}}}(x) \,|\, x])^2 \,|\, x\Big]}_{\text{Variance}} + \underbrace{\big(f_0(x) - \mathbb{E}[f_{S_{\text{train}}}(x) \,|\, x]\big)^2}_{\text{Bias}^2}$$

this follows from $\mathbb{E}[(f_{S_{\text{train}}}(x) - \mathbb{E}[f_{S_{\text{train}}}(x) \,|\, x]) \,|\, x] = 0$

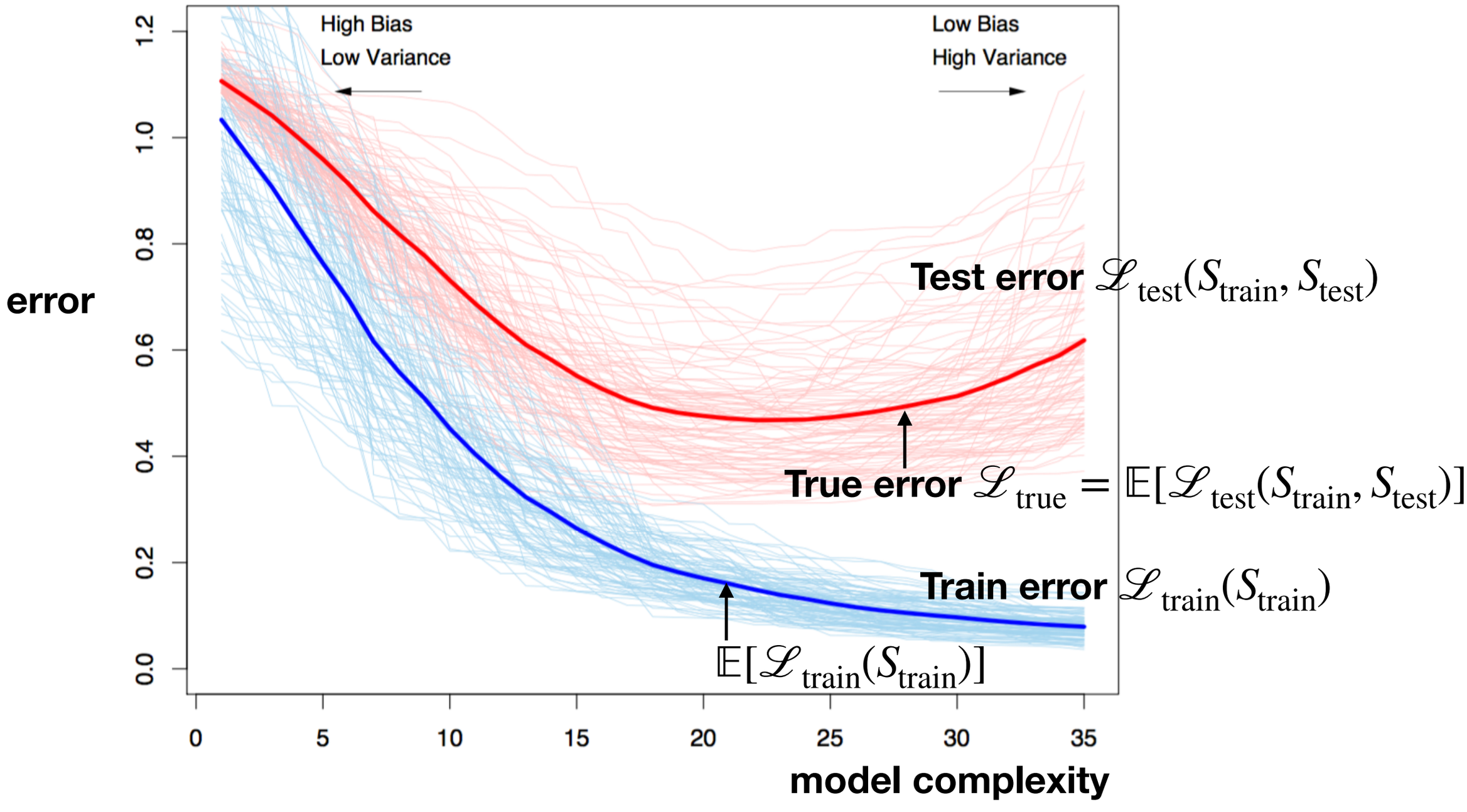
- this theoretical analysis explains the behavior of true error $\mathscr{L}_{\text{true}}$

$$\mathscr{L}_{\text{true}} = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\mathbb{E}\Big[(f_{S_{\text{train}}}(x) - \mathbb{E}[f_{S_{\text{train}}}(x) \,|\, x])^2\Big]}_{\text{(expected) Variance}} + \underbrace{\mathbb{E}_{p_x}\Big[\big(f_0(x) - \mathbb{E}[f_{S_{\text{train}}}(x) \,|\, x]\big)^2\Big]}_{\text{(expected) Bias}^2} \tag{1}$$

- Whether we condition on x or not when referring to **variance** and **bias** should be clear from the context
- **bias**
  - measures how the predictor is mismatched with the true model in expectation
- **variance**
  - measures how the predictor varies each time with a new training datasets



- recall the test error is an unbiased estimator of the true error, i.e. $\mathscr{L}_{\text{true}} = \mathbb{E}[\mathscr{L}_{\text{test}}]$
- and theory explains **true error**, and hence expected behavior of the (random) **test error**



Simple Gaussian example

- model: $y_i = w_1 x_i[1] + w_2 x_i[2] + 0 \cdot x_i[3] + \varepsilon_i$
- $x_i[1], x_i[2], x_i[3], \varepsilon_i \sim$ i.i.d. Gaussian $N(0, \sigma^2)$
- training data $\{(x_i, y_i)\}_{i=1}^n$
- let data matrix be $\mathbf{X} = \begin{bmatrix} x_1[1] & x_1[2] & x_1[3] \\ \vdots & \vdots & \vdots \\ x_n[1] & x_n[2] & x_n[3] \end{bmatrix} \in \mathbb{R}^{n \times 3}$
- $\mathbf{X}[:, \mathbf{1}]$ denotes the first column of $\mathbf{X}$

Simple Gaussian example

- model: $y_i = w_1 x_i[1] + w_2 x_i[2] + 0 \cdot x_i[3] + \varepsilon_i$

- **Example 1 (simple predictor):**
- consider a **simple** model of fitting only the first feature

$$\hat{y} = \hat{w}_1 x[1]$$

- linear least squares gives

$$\hat{w}_1 = (\mathbf{X}[:,\mathbf{1}]^T \mathbf{X}[:,\mathbf{1}])^{-1} \mathbf{X}[:,\mathbf{1}]^T \mathbf{y}$$

with $\mathbf{X}[:,\mathbf{1}] = \begin{bmatrix} x_1[1] \\ \vdots \\ x_n[1] \end{bmatrix}$, and $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = w_1 \mathbf{X}[:,\mathbf{1}] + w_2 \mathbf{X}[:,\mathbf{2}] + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\triangleq \varepsilon}$

cancels each other for the first term

- plugging in $\mathbf{y}$ in the equation, we get

$$\begin{aligned} \hat{w}_1 &= (\mathbf{X}[:,\mathbf{1}]^T \mathbf{X}[:,\mathbf{1}])^{-1} \mathbf{X}[:,\mathbf{1}]^T (\mathbf{X}[:,\mathbf{1}] w_1 + \mathbf{X}[:,\mathbf{2}] w_2 + \varepsilon) \\ &= w_1 + (\mathbf{X}[:,\mathbf{1}]^T \mathbf{X}[:,\mathbf{1}])^{-1} \mathbf{X}[:,\mathbf{1}]^T (w_2 \mathbf{X}[:,\mathbf{2}] + \varepsilon) \end{aligned}$$

Simple Gaussian example

- for large enough n,

$$\mathbf{X}[:,\mathbf{1}]^T\mathbf{X}[:,\mathbf{1}] = \sum_{i=1}^{n} x_i[1]^2 \;\simeq\; n\sigma^2,$$

by law of large numbers, and we will use this approximation to simplify the formula:

$$\hat{w}_1 \;=\; w_1 + \frac{1}{n\sigma^2}\mathbf{X}[:,\mathbf{1}]^T(w_2\mathbf{X}[:,\mathbf{2}] + \varepsilon) \quad \text{and}$$

$$\mathbb{E}[\hat{w}_1] \;=\; w_1$$

where we used the fact that $\mathbf{X}[:,\mathbf{1}], \mathbf{X}[:,\mathbf{2}], \varepsilon \in \mathbb{R}^n$ are independent zero-mean vectors

- we are ready to compute the **bias** and **variance**
- first, conditioned on $x = (x[1], x[2], x[3])$

$$\text{bias}^2 \;=\; (\mathbb{E}[f(x)] - f_0(x))^2 \;=\; \Big(w_1 x[1] - (w_1 x[1] + w_2 x[2])\Big)^2 \;=\; (w_2)^2 x[2]^2\,,$$

and taking expectation over p_x

$$\mathbb{E}_{p_x}[\text{bias}^2] \;=\; (w_2)^2\sigma^2$$

- note that
  - this does not decrease with (training) sample size n
  - this is due to not including $x[2]$ in our prediction, in other words, using a too simple predictor

Simple Gaussian example

- Now for the variance, conditioned on $x = (x[1], x[2], x[3])$,

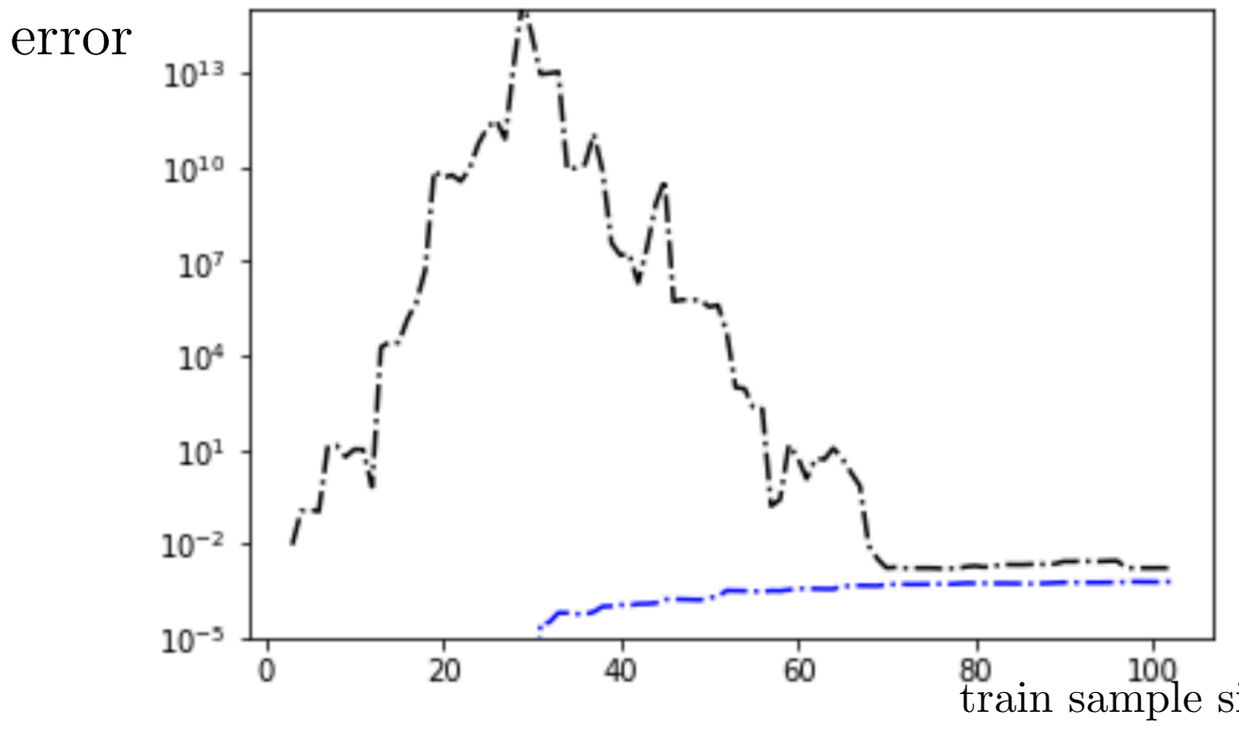
  since we have $\hat{w}_1 = w_1 + \frac{1}{n\sigma^2}\mathbf{X}[:,\mathbf{1}]^T(w_2\mathbf{X}[:,\mathbf{2}] + \varepsilon)$

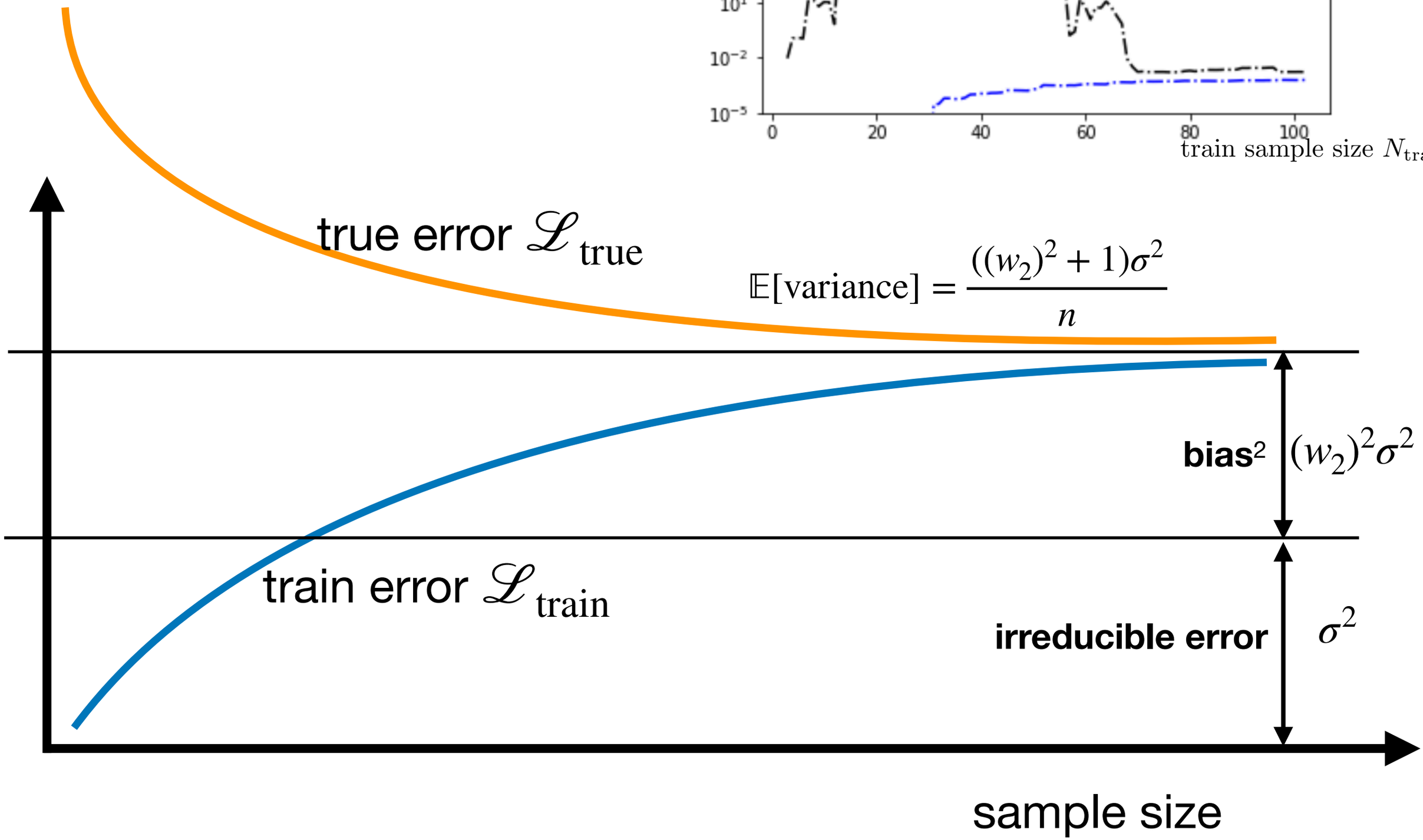
  and $\mathbb{E}[\hat{w}_1] = w_1$

- $$\text{variance} = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2 \,|\, x]$$
$$= \mathbb{E}\left[\left(\frac{1}{n\sigma^2}\mathbf{X}[:,\mathbf{1}]^T(w_2\mathbf{X}[:,\mathbf{2}] + \varepsilon)x[1]\right)^2 \,|\, x\right]$$
$$= \frac{x[1]^2}{n^2\sigma^4}\left(\mathbb{E}\left[\left(\sum_{i=1}^{n} w_2 x_i[1]x_i[2] + x_i[1]\varepsilon_i\right)^2\right]\right)$$
$$= \frac{x[1]^2}{n^2\sigma^4}\left(\mathbb{E}\left[\left(\sum_{i=1}^{n}(w_2 x_i[1]x_i[2])^2 + \sum_{i=1}^{n}(x_i[1]\varepsilon_i)^2 + \underbrace{2\sum_{i=1}^{n}(w_2 x_i[1]^2 x_i[2])}_{\text{zero mean}} + \underbrace{\sum_{i\neq j=1}^{n}(w_2 x_i[1]x_i[2] + x_i[1]\varepsilon_i)(w_2 x_j[1]x_j[2] + x_j[1]\varepsilon_j)}_{\text{zero mean}}\right]\right)$$
$$= \frac{x[1]^2}{n^2\sigma^4} n\left((w_2)^2\sigma^4 + \sigma^4\right) = \frac{((w_2)^2 + 1)x[1]^2}{n}$$

- and $\mathbb{E}[\text{variance}] = \frac{((w_2)^2 + 1)\sigma^2}{n}$
- this decrease with (training) sample size n

- **Analysis** explains the empirical observation on the right, on error vs. sample size





Simple Gaussian example

- **Example 2 (moderate predictor):**
- consider a **moderate** model of fitting the first two features

$$\hat{y} \;=\; \hat{w}_1 x[1] + \hat{w}_2 x[2]$$

- then, we will show that bias is smaller (in fact zero) and variance is larger, i.e.
  - $\mathbb{E}_{p_x}[\text{bias}^2] \;=\; 0$
  - $\mathbb{E}[\text{variance}] = \dfrac{2\sigma^2}{n}$
- linear least squares gives

$$\begin{bmatrix}\hat{w}_1\\ \hat{w}_2\end{bmatrix} \;=\; (\mathbf{X}[:,\mathbf{1}:\mathbf{2}]^T\mathbf{X}[:,\mathbf{1}:\mathbf{2}])^{-1}\mathbf{X}[:,\mathbf{1}:\mathbf{2}]^T\mathbf{y}$$

with $\mathbf{X}[:,\mathbf{1}:\mathbf{2}] = \begin{bmatrix} x_1[1] & x_1[2] \\ \vdots & \vdots \\ x_n[1] & x_n[2]\end{bmatrix}$, and $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n\end{bmatrix} \;=\; w_1\mathbf{X}[:,\mathbf{1}] + w_2\mathbf{X}[:,\mathbf{2}] + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n\end{bmatrix}}_{\triangleq\,\varepsilon}$

- plugging in $\mathbf{y}$ in the equation, we get

$$\begin{aligned}\begin{bmatrix}\hat{w}_1\\ \hat{w}_2\end{bmatrix} &= (\mathbf{X}[:,\mathbf{1}:\mathbf{2}]^T\mathbf{X}[:,\mathbf{1}:\mathbf{2}])^{-1}\mathbf{X}[:,\mathbf{1}:\mathbf{2}]^T(\mathbf{X}[:,\mathbf{1}:\mathbf{2}]\begin{bmatrix}w_1\\ w_2\end{bmatrix} + \varepsilon)\\ &= \begin{bmatrix}w_1\\ w_2\end{bmatrix} + (\mathbf{X}[:,\mathbf{1}:\mathbf{2}]^T\mathbf{X}[:,\mathbf{1}:\mathbf{2}])^{-1}\mathbf{X}[:,\mathbf{1}:\mathbf{2}]^T\,\varepsilon\end{aligned}$$

Simple Gaussian example

- for large enough n,

$$\mathbf{X}[:,\mathbf{1}]^T\mathbf{X}[:,\mathbf{1}] = \begin{bmatrix} \sum_{i=1}^n x_i[1]^2 & \sum_{i=1}^n x_i[1]x_i[2] \\ \sum_{i=1}^n x_i[1]x_i[2] & \sum_{i=1}^n x_i[2]^2 \end{bmatrix} \simeq n\sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

by law of large numbers, and we will use this approximation to simplify the formula:

$$\begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \frac{1}{n\sigma^2}\mathbf{X}[:,\mathbf{1}:\mathbf{2}]^T\varepsilon \quad \text{and}$$

$$\mathbb{E}\left[\begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix}\right] = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

where we used the fact that $\mathbf{X}[:,\mathbf{1}], \mathbf{X}[:,\mathbf{2}], \varepsilon \in \mathbb{R}^n$ are independent zero-mean vectors

- we are ready to compute the **bias** and **variance**
- first, conditioned on $x = (x[1], x[2], x[3])$

$$\text{bias}^2 = (\mathbb{E}[f(x)] - f_0(x))^2 = \Big(w_1x[1] + w_2x[2] - (w_1x[1] + w_2x[2]) \Big)^2 = 0\,,$$

- note that
  - this is an unbiased predictor

Simple Gaussian example

- Now for the variance, conditioned on $x = (x[1], x[2], x[3])$,

since we have $\begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \frac{1}{n\sigma^2} \mathbf{X}[:, \mathbf{1}:\mathbf{2}]^T \varepsilon$

and $\mathbb{E}\left[\begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix}\right] = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

- $\text{variance} = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2 \,|\, x]$

$$= \mathbb{E}\left[\left(\frac{1}{n\sigma^2} \varepsilon^T \mathbf{X}[:, \mathbf{1}:\mathbf{2}] \begin{bmatrix} x[1] \\ x[2] \end{bmatrix}\right)^2 | x\right]$$

$$= \frac{1}{n^2\sigma^4}\left(\mathbb{E}\left[\left(\sum_{i=1}^{n} (x_i[2]x[2] + x_i[1]x[1])\varepsilon_i\right)^2\right]\right)$$

$$= \frac{(x[1]^2 + x[2]^2)n\sigma^4}{n^2\sigma^4} = \frac{x[1]^2 + x[2]^2}{n}$$

- and $\mathbb{E}[\text{variance}] = \frac{2\sigma^2}{n}$
- this decrease with (training) sample size n

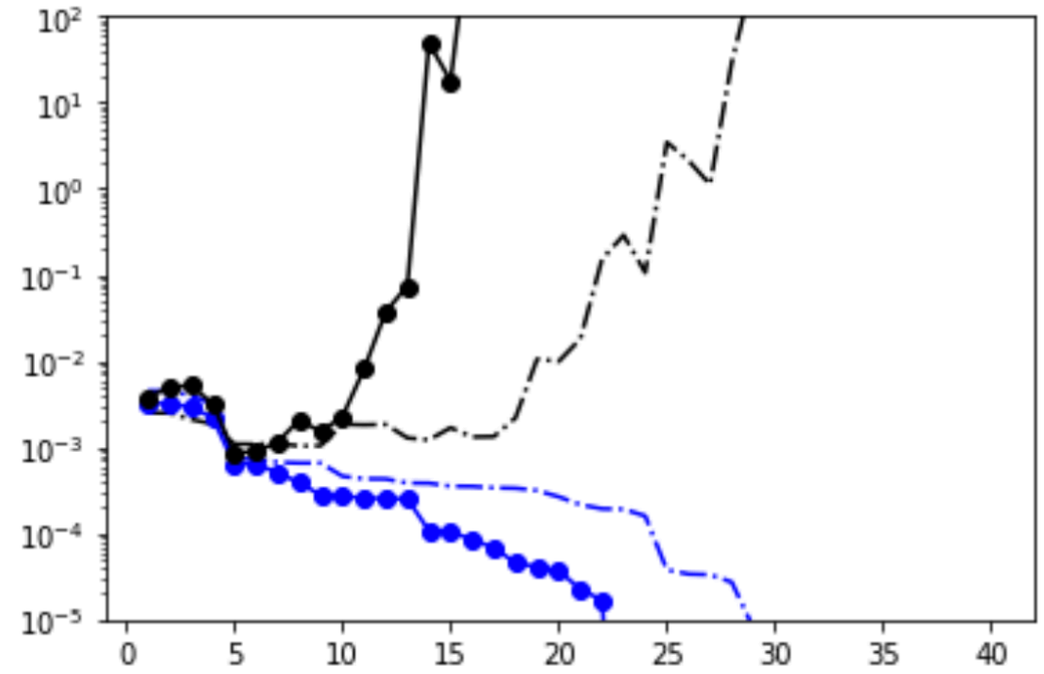
Simple Gaussian example

- **Example 3 (complex predictor):**
- consider a **complex** model of fitting

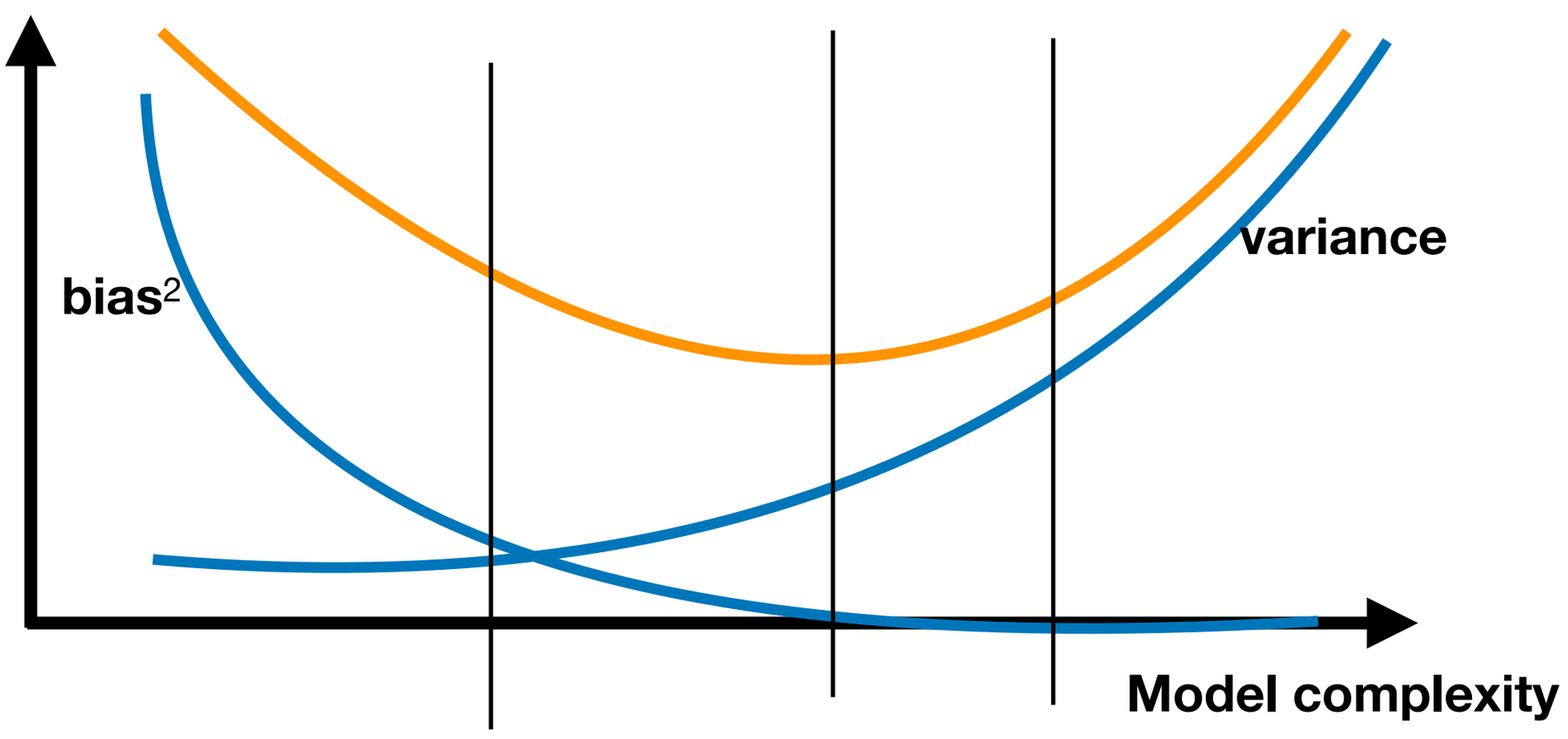
$$\hat{y} = \hat{w}_1 x[1] + \hat{w}_2 x[2] + \hat{w}_2 x[3]$$

- $\mathbb{E}_{p_x}[\text{bias}^2] = 0$
- $\mathbb{E}[\text{variance}] = \dfrac{3\sigma^2}{n}$
- we skip the detailed proof here, as it is almost identical to the previous one

- this explains the observation on the bias-variance tradeoff



$\mathbb{E}_{p_x}[\text{bias}^2]$	$= (w_2)^2\sigma^2$	**0**	**0**
$\mathbb{E}[\text{variance}]$	$= \dfrac{((w_2)^2+1)\sigma^2}{n}$	$\dfrac{2\sigma^2}{n}$	$\dfrac{3\sigma^2}{n}$



* Notations

Model: $P_{XY} \sim (X_i, Y_i)$ example: $y_i = f_0(x_i) + \varepsilon_i$

Sample: $S_{train} = \{(x_i, y_i)\}_{i=1}^{n}$ $|S_{train}| = n$

$S_{test} = \{(x_i, y_i)\}_{i=n+1}^{n+m}$ $|S_{test}| = m$

Expectation: $\mathbb{E}_{(x,y)\sim P_{XY}}[F(x,y)] = \mathbb{E}_{P_{XY}}[F(x,y)] = \mathbb{E}[F(x,y)]$

$\mathbb{E}_{y\sim P_{Y|x}}[F(x,y)|x] = \mathbb{E}[F(x,y)|x]$

* Goal

(Expected) test error: $L_{test} = \frac{1}{m}\sum_{i=n+1}^{n+m}(f_{S_{train}}(x_i) - y_i)^2$

true error: $L_{true} = \mathbb{E}_{S_{test}\sim P_{XY}^m, S_{train}\sim P_{XY}^n}[L_{test}]$

$= \mathbb{E}\left[\frac{1}{m}\sum_{i=n+1}^{n+m}(f_{S_{train}}(x_i) - y_i)^2\right]$

$= \frac{1}{m}\sum_{i=n+1}^{n+m}\mathbb{E}_{(x_i,y_i)\sim P_{XY}, S_{train}}\left[(f_{S_{train}}(x_i) - y_i)^2\right]$

$$\mathbb{E}_{P_{xy},\,S_{train}}\left[\left(f_{S_{train}}(x)-Y\right)^2\right]$$

$$=\mathbb{E}_{P_x}\left[\underbrace{\mathbb{E}_{P_{y|x},\,S_{train}}\left[\left(f_{S_{train}}(x)-Y\right)^2\,\middle|\,x\right]}_{\text{Variance}+\text{Bias}^2+\sigma^2}\right]$$

model: $Y=f_0(x)+\varepsilon$

(Conditional)

true error $=\mathbb{E}_{P_{y|x},\,S_{train}}\left[\left(\underbrace{f_{S_{train}}(x)-f_0(x)}_{A}-\underbrace{\varepsilon}_{B}\right)^2\,\middle|\,x\right]$

$$=\underbrace{\mathbb{E}\left[\left(f_{S_{train}}(x)-f_0(x)\right)^2\,\middle|\,x\right]}_{A^2}+\underbrace{\mathbb{E}\left[\varepsilon^2\,|\,x\right]}_{B^2}+\underbrace{\mathbb{E}\left[-2AB\right]}_{\square\times\mathbb{E}[\varepsilon]=0}$$

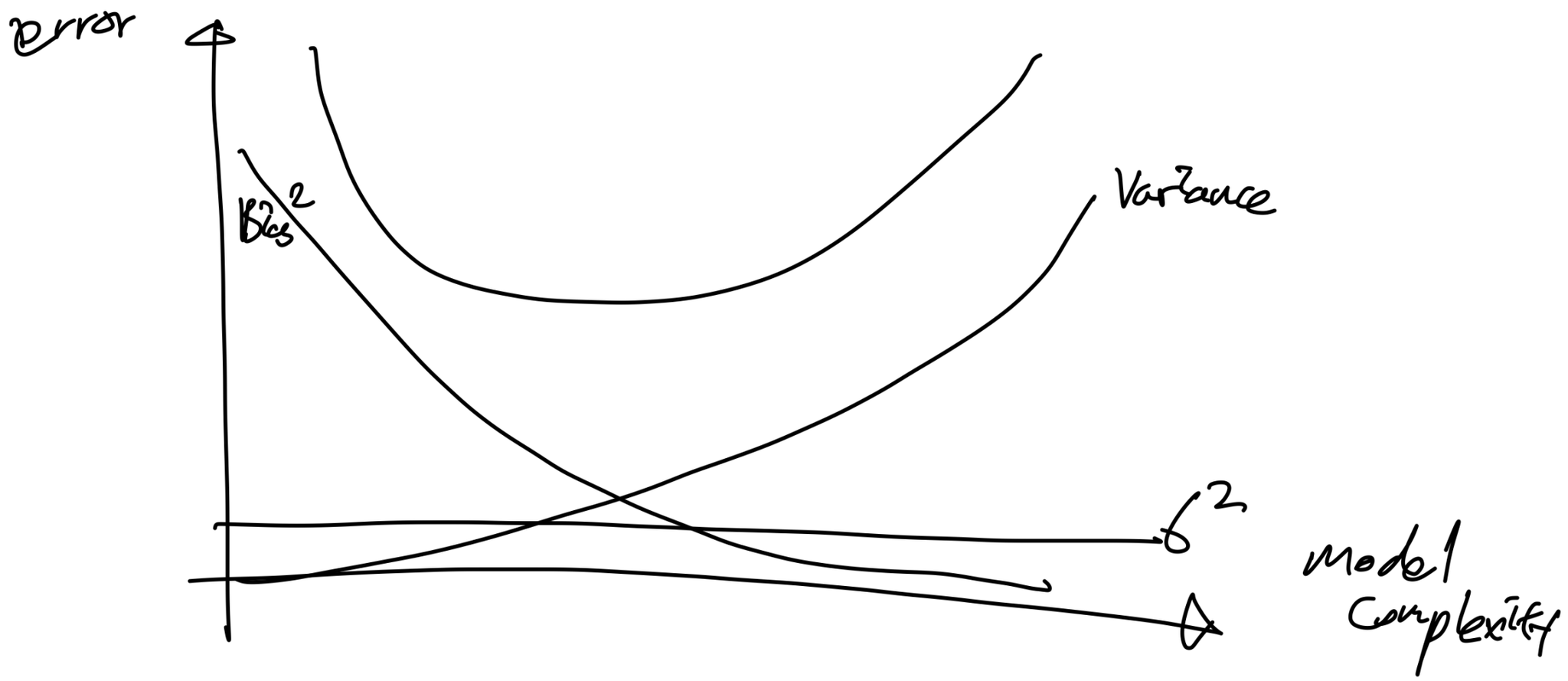
learning error ≥ 0 — irreducible error.

$$\text{learning error} = \mathbb{E}\left[(f_{strain}(x) - f_0(x))^2 \mid x \right]$$

$$= \mathbb{E}\left[\left(\underbrace{f_{strain}(x) - \mathbb{E}[f_{strain}(x) \mid x]}_{A} + \underbrace{\mathbb{E}[f_{strain}(x) \mid x] - f_0(x)}_{B} \right)^2 \mid x \right]$$

$$= \underbrace{\mathbb{E}\left[(f_{strain}(x) - \bar{f}(x))^2 \mid x \right]}_{\text{Variance.}} + \underbrace{(\bar{f}(x) - f_0(x))^2}_{\text{Bias}^2}$$

$$L_{true} = \sigma^2 + \mathbb{E}_{P_x}[\text{Variance}] + \mathbb{E}_{P_x}[\text{Bias}^2]$$



*Simple Gaussian Example

model: $y_i = w_1 x_i[1] + w_2 x_i[2] + 0 \cdot x_i[3] + \varepsilon_i$

$x_i \in \mathbb{R}^3$, $x_i[1], x_i[2], x_i[3], \varepsilon_i \sim N(0, \sigma^2)$

Train: $\{(x_i, y_i)\}_{i=1}^n$

Data Matrix $X = \begin{bmatrix} x_1[1] & x_1[2] & x_1[3] \\ \vdots & & \\ x_n[1] & x_n[2] & x_n[3] \end{bmatrix} \in \mathbb{R}^{n \times 3}$

$X[:,1]$

*Example 1: Simple Predictor

$\hat{y} = \hat{w}_1 \cdot X[1]$

$\hat{w}_1 = (X[:,1]^T X[:,1])^{-1} X[:,1]^T \cdot y$

linear least squares predictor

$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = w_1 \cdot X[:,1] + w_2 X[:,2] + \varepsilon$

$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

$= w_1 + (X[:,1]^T X[:,1])^{-1} X[:,1]^T (X[:,2] \cdot w_2 + \varepsilon)$

$\mathbb{E}[\hat{w}_1] = w_1$

*Bias: $\mathbb{E}_x\left[\left(\hat{f}(x) - f_0(x) \right)^2 \right]$

$= \mathbb{E}_x\left[\left(\mathbb{E}[\hat{w}_1] \cdot x[1] - (w_1 x[1] + w_2 x[2]) \right)^2 \right]$

$\mathbb{E}[\hat{w}_1] = w_1$

$= \mathbb{E}_x\left[w_2^2 x[2]^2 + \varepsilon^2 \right] = w_2^2 \sigma^2$ ← does not decrease with n; price we pay for using simple model

*Variance: $\hat{w}_1 = w_1 + \left(X[:,1]^T X[:,1] \right)^{-1} X[:,1]^T \left(w_2 X[:,2] + \varepsilon \right)$

$\sum_{i=1}^{n} x_i[1]^2 \cong n \cdot \mathbb{E}[x_i[1]^2] = n\sigma^2$

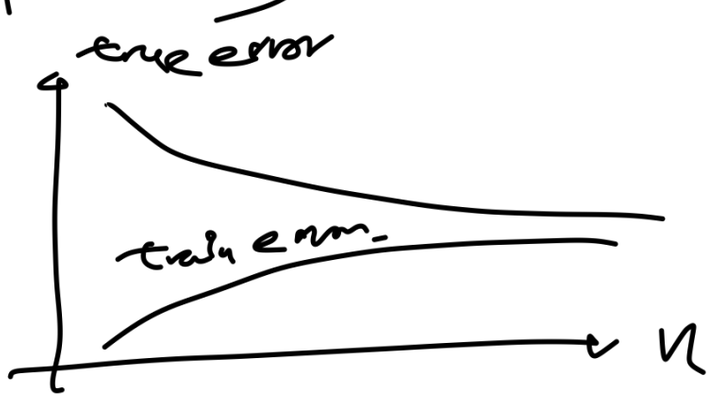
$= \mathbb{E}\left[\left(w_1 x[1] - \hat{w}_1 x[1] \right)^2 \right]$

$= \mathbb{E}\left[\frac{1}{(n\sigma^2)^2} \left(X[:,1]^T X[:,2] \cdot w_2 + X[:,1]^T \varepsilon \right)^2 x[1]^2 \right]$

$= \frac{\sigma^2}{n^2\sigma^4} \left\{ w_2^2 \mathbb{E}\left[\left(\sum_{i=1}^{n} x_i[1] x_i[2] \right)^2 \right] + \mathbb{E}\left[\left(\sum_{i=1}^{n} x_i[1] \varepsilon \right)^2 \right] \right\}$

$w_2^2 n\sigma^4$, $n\sigma^4$

$= \frac{(w_2^2 + 1)\sigma^2}{n}$ → goes down with n. → much smaller.



*Example 2: Moderate model.

$$\hat{y} = \hat{w}_1 X[1] + \hat{w}_2 X[2]$$

claim: $\mathbb{E}[bias^2] = 0$

$\mathbb{E}[variance] = \frac{2\cdot\sigma^2}{n}$

*Example 3:

$$\hat{y} = \hat{w}_1 X[1] + \hat{w}_2 X[2] + \hat{w}_3 X[3]$$

claim: $\mathbb{E}[bias^2] = 0$

$\mathbb{E}[variance] = \frac{3\sigma^2}{n}$