FIT5201 - Data analysis algorithms

Module 2: Linear Models for Regression

Part B:

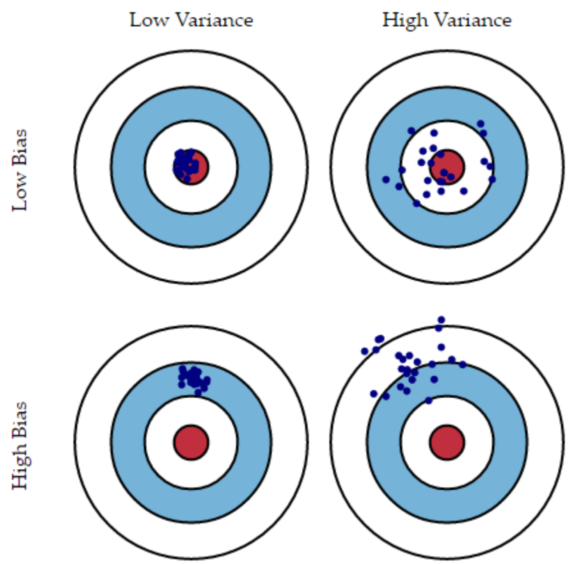
- **Bias-Variance Analysis**

Module 2: Linear Models for Regression

- Module Objectives
  - **Provide a deep understanding of linear regression models**
  - **Part B (Week 4):**
    - Bias-Variance **Analysis (Understanding)**
    - Bias-Variance in Regression (**Tutorial**)

Part B

Bias-Variance Analysis



http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias and Variance

- How to understand Bias and Variance in machine learning
  - Bias: indicate the accuracy of the models
  - Variance: indicate how consistent the models are
- Why do we need to understand Bias and Variance?
  - Diagnose model performance
  - Avoid the mistakes of over-fitting & under-fitting
    - Regularization?

Bias and Variance

Conceptual Definition

Conceptual Definition: Bias

Bias Error

Difference between the expected (or average) prediction of our model and the correct value of the true model (one we are trying to predict)

Why "expected" or "average" prediction of our model?

Recall the management of "uncertainty" for frequentists:

you need to repeat your model learning process many times:

- Each time you create a new model with new training set.
- You will have N different models whose predictions are various.

Bootstrap for Quantifying Uncertainty

- Imagine
  - We only have D, and our goal is to fit a model with parameter $\boldsymbol{w}$ to the given dataset.
  - We found $\boldsymbol{w}$ that maximises the probability of observation D.
  - We wonder if $\boldsymbol{w}$ would change if we have an alternate dataset D'
  - If we do this exercise for several alternate datasets, then we will a *distribution* over estimates for $\boldsymbol{w}$.
  - If this distribution is higher, the more uncertainty on $\boldsymbol{w}$.

Now, the problem is how to get the alternate datasets?

Bootstrap Example

- A bootstrap sample is a random sample conducted with replacement
  1. Randomly select an observation from the original data
  2. Write it down
  3. Put it back (i.e. Any observation can be selected more than once)

Repeat these steps 1-3 N times; N is the number of observations in the original sample

Conceptual Definition: Bias

Bias Error

Difference between the expected (or average) prediction of our model and the correct value of the true model (one we are trying to predict)

Why "expected" or "average" prediction of our model?

Recall the management of "uncertainty" for frequentists:

you need to repeat your model learning process many times:

- Each time you create a new model with new training set.
- You will have N different models whose predictions are various.

What does Bias measure?

How far off in general multiple models' predictions are from the correct value. (the tendency to consistently learn the same wrong thing)

Conceptual Definition: Variance

Variance Error

Variability of a model prediction for a given data point.

What does Variance measure?

Again, imagine you repeat your model learning process many times.

The variance is how inconsistent are the predictions from one another, over different training sets, not whether they are accurate or not.

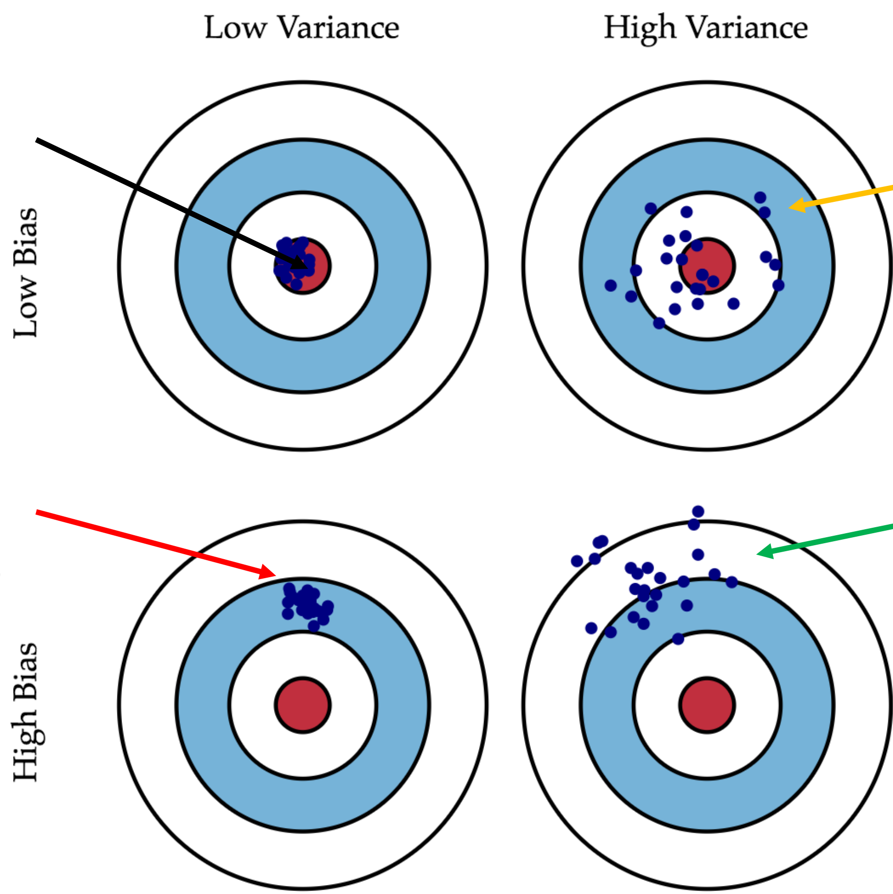
Bias and Variance

Graphical Definition

Graphical Definition: Bias & Variance

The true model

Scatter of hits (High Variance)



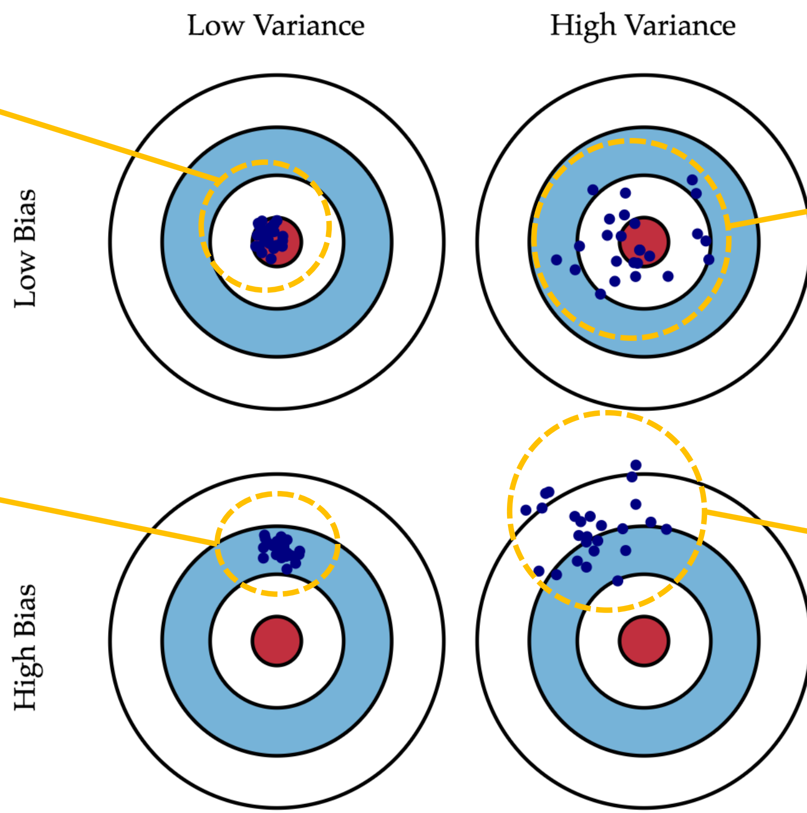
Predictions get worse (High Bias)

High Bias and High Variance

Graphical representation of bias and variance (http://scott.fortmann-roe.com/docs/BiasVariance.html)

Graphical Definition: Bias & Variance

Excellent model



Average of predictions is close to the centre, but much variations exist

predictions are similar to each other, but away from the centre

Bad model

The prediction errors are characterised by the combination of Bias & Variance

Bias and Variance

Model Complexity vs Bias-Variance

Mathematical Definition: Bias & Variance

Settings

A training set $D \coloneqq \{(x_n, t_n)\}_{n=1}^{N}$

$p(x)$: the input data points x are generated according to this distribution

$h(x)$: the target value t are generated according to this function

$y(x)$: linear basis function model we learned from D

Generalization Error

$$\varepsilon_{h,p}(y) \coloneqq \int [y(x) - h(x)]^2 p(x) dx$$

If can be computed, done!

But we do not know $h(x)$ and $p(x)$

Mathematical Definition: Bias & Variance

Settings

A training set $D := \{(x_n, t_n)\}_{n=1}^{N}$

$p(x)$: the input data points x are generated according to this distribution

$h(x)$: the target value t are generated according to this function

$y(x)$: linear basis function model we learned from D

Generalization Error

$$\varepsilon_{h,p}(y) := \int [y(x) - h(x)]^2 p(x) dx$$

If can be computed, done!

But we do not know $h(x)$ and $p(x)$

Any idea? Removing these unknown items

Mathematical Definition: Bias & Variance

Bootstrap Idea

- Remove $p(x)$
- $y(x)$ is actually $y(x, \boldsymbol{w})$
- Frequentist: estimate $\boldsymbol{w}$ based on D, which is independently drawn from the underlying distribution $p(x)$.
- Bootstrap: simulate $p(x)$ using one data set D
- Suppose that we had a large number of data sets, for any given data set D, we can learn a model $y(x, D)$

Generalization Error on one data set

$$[y(x, D) - h(x)]^2$$

Mathematical Definition: Bias & Variance

Take expectations over the ensemble of data sets

$$E_D[\{y(x;D)-h(x)\}^2]$$

$$= E_D[\{y(x;D)-E_D[y(x;D)]+E_D[y(x;D)]-h(x)\}^2]$$

$$= E_D[\{y(x;D)-E_D[y(x;D)]\}^2+\{E_D[y(x;D)]-h(x)\}^2$$
$$+2\{y(x;D)-E_D[y(x;D)]\}\{E_D[y(x;D)]-h(x)\}]$$

$$= E_D[\{y(x;D)-E_D[y(x;D)]\}^2]+E_D[\{E_D[y(x;D)]-h(x)\}^2]$$

$$= \underbrace{E_D[\{y(x;D)-E_D[y(x;D)]\}^2]}_{variance}+\underbrace{\{E_D[y(x;D)]-h(x)\}^2}_{bias^2}$$

$$generalisation\ error = bias^2 + variance$$

Bias-Variance & Model Complexity

$$\boldsymbol{generalisation\ error = bias^2 + variance}$$

Our goal: minimize the generalization error

Trade-off: very flexible models having low bias and high variance, and relatively rigid models having high bias and low variance

Bias and Variance

Examples of Model Complexity

Examples: controlled by regularisation parameter

Model complexity: controlled by regularization parameter

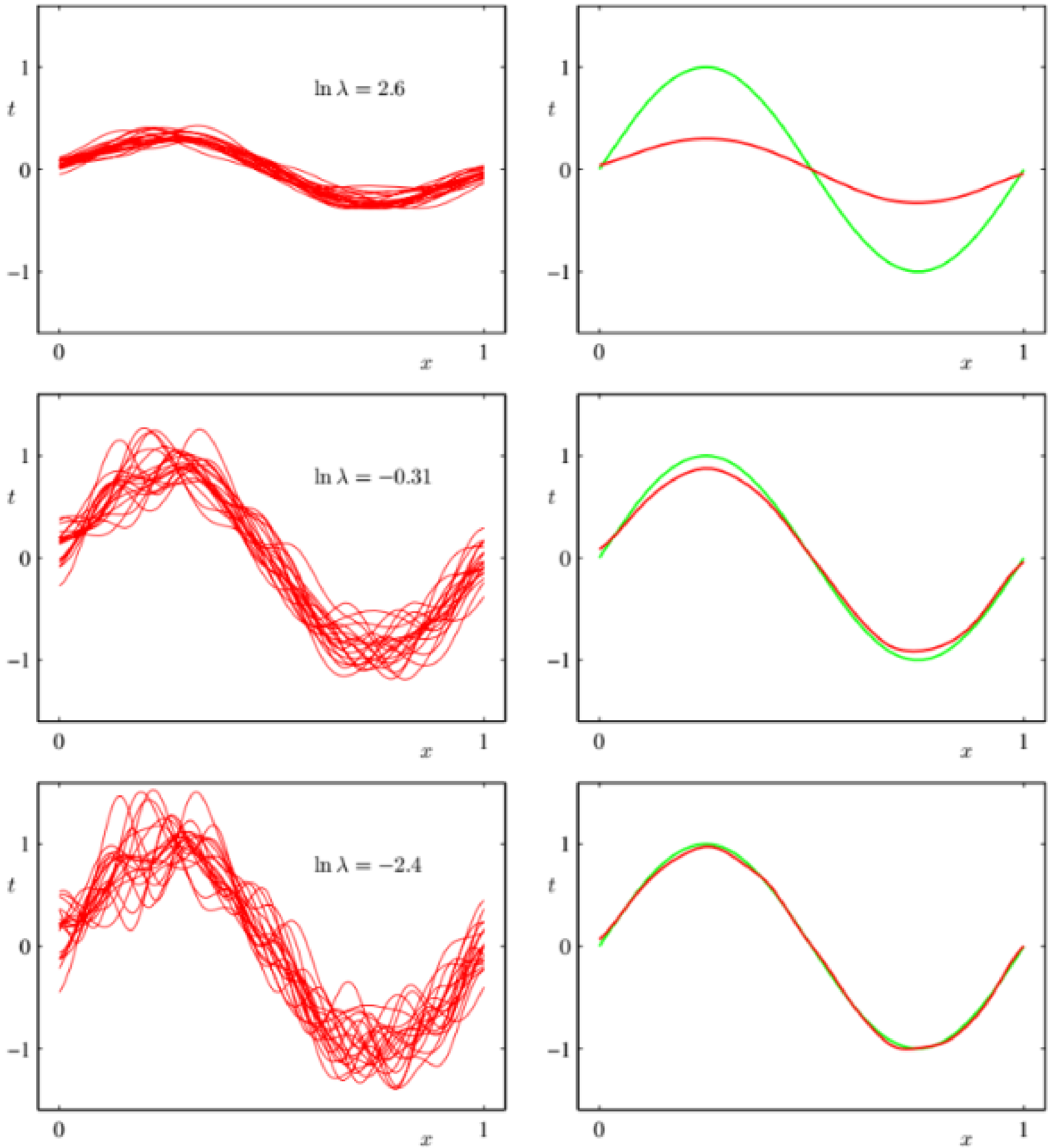
Underlying function: $h = \sin(2\pi x)$

Number of datasets: 100 (indexed by l)

Size of each dataset: 25

Model: 24 Gaussian basis functions

Regularization function: Ridge



smaller value of regularization parameter
-> smaller weight on penalizing complexity of model
-> model more complex
-> smaller bias and larger variance

Examples: controlled by regularisation parameter

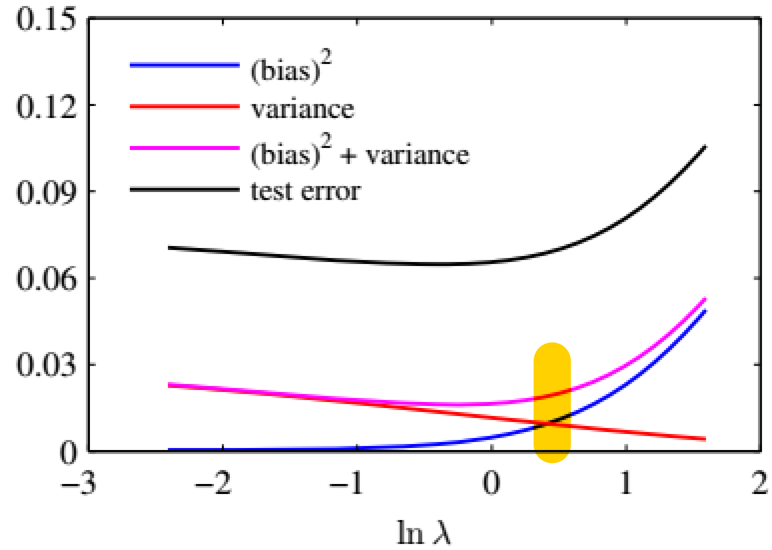
Quantitative analysis

$\bar{y}(x) := \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$

$$\text{bias}^2 := \frac{1}{N}\sum_{n=1}^{N} [\bar{y}(x_n) - h(x_n)]^2$$

$$\text{variance} := \frac{1}{N}\sum_{n=1}^{N} \frac{1}{L}\sum_{l=1}^{L} \left[y^{(l)}(x_n) - \bar{y}(x_n)\right]^2$$

$$\text{test error} := \frac{1}{L}\sum_{l=1}^{L} \frac{1}{N}\sum_{n=1}^{N} \left[y^{(l)}(x_n) - h(x_n)\right]^2$$



Examples: controlled by regularisation parameter

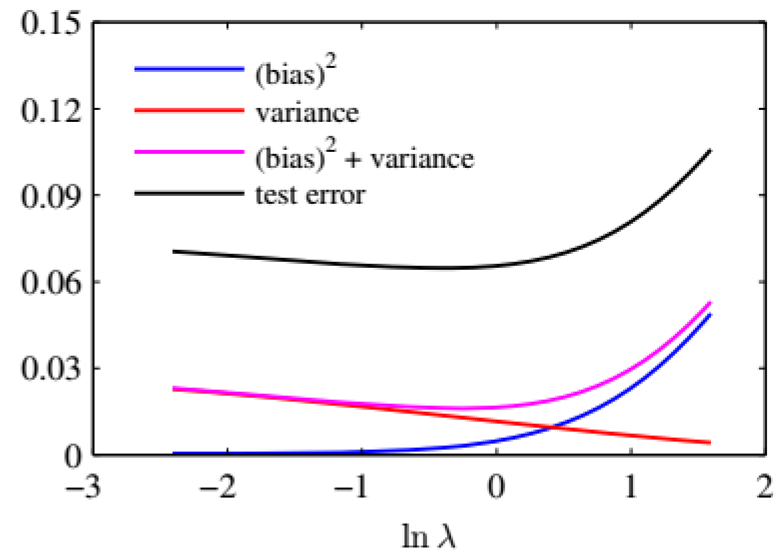
Quantitative analysis

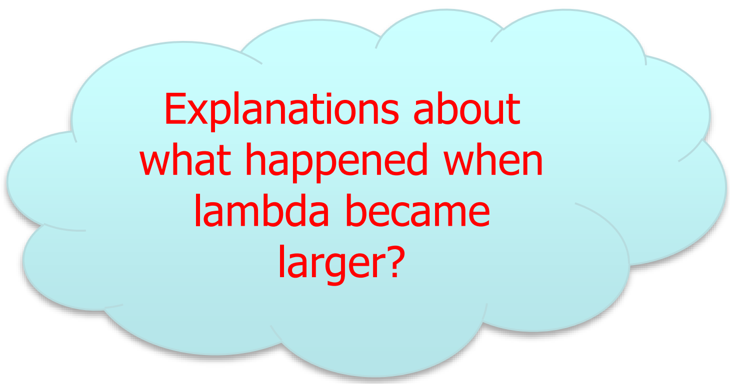
$\bar{y}(x) := \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$

$$\text{bias}^2 := \frac{1}{N}\sum_{n=1}^{N}\left[\bar{y}(x_n) - h(x_n)\right]^2$$

$$\text{variance} := \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left[y^{(l)}(x_n) - \bar{y}(x_n)\right]^2$$

$$\text{test error} := \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{n=1}^{N}\left[y^{(l)}(x_n) - h(x_n)\right]^2$$





Examples: controlled by regularisation parameter

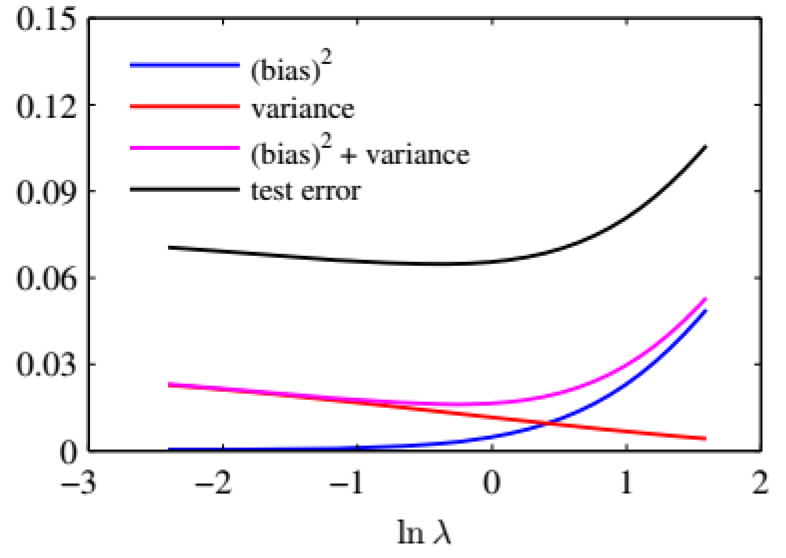
Quantitative analysis

$\bar{y}(x) := \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$

$$\text{bias}^2 := \frac{1}{N}\sum_{n=1}^{N}\left[\bar{y}(x_n) - h(x_n)\right]^2$$

$$\text{variance} := \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left[y^{(l)}(x_n) - \bar{y}(x_n)\right]^2$$

$$\text{test error} := \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{n=1}^{N}\left[y^{(l)}(x_n) - h(x_n)\right]^2$$



1. As lambda becomes larger (model becomes simpler), bias increases while variance decreases
2. The generalization error derived from frequentist view is quit similar to the test error

Examples: controlled by regularisation parameter

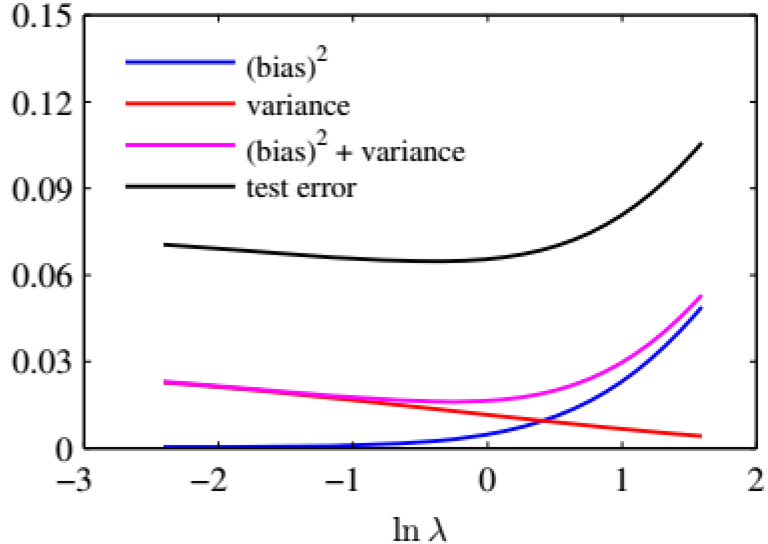
Quantitative analysis

$\bar{y}(x) := \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$

$$\text{bias}^2 := \frac{1}{N}\sum_{n=1}^{N}\left[\bar{y}(x_n) - h(x_n)\right]^2$$

$$\text{variance} := \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left[y^{(l)}(x_n) - \bar{y}(x_n)\right]^2$$

$$\text{test error} := \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{n=1}^{N}\left[y^{(l)}(x_n) - h(x_n)\right]^2$$



How to find the "sweet spot"?

Examples: controlled by regularisation parameter

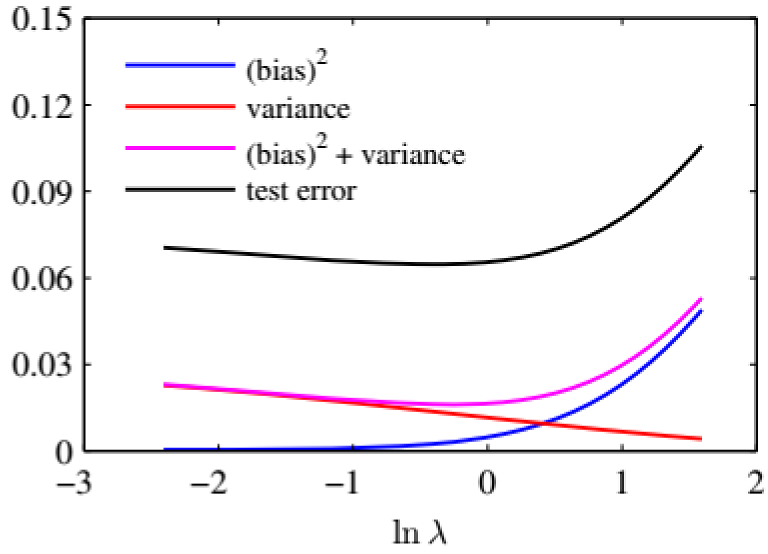
Quantitative analysis

$$\bar{y}(x) := \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

$$\text{bias}^2 := \frac{1}{N}\sum_{n=1}^{N} [\bar{y}(x_n) - h(x_n)]^2$$

$$\text{variance} := \frac{1}{N}\sum_{n=1}^{N} \frac{1}{L}\sum_{l=1}^{L} \left[y^{(l)}(x_n) - \bar{y}(x_n)\right]^2$$

$$\text{test error} := \frac{1}{L}\sum_{l=1}^{L} \frac{1}{N}\sum_{n=1}^{N} \left[y^{(l)}(x_n) - h(x_n)\right]^2$$



An indication of overfitting and underfitting?

Example: controlled by number of parameters

- Prediction goal - learn the function:

$$h(x) = \sin(2\pi x) + \epsilon,$$

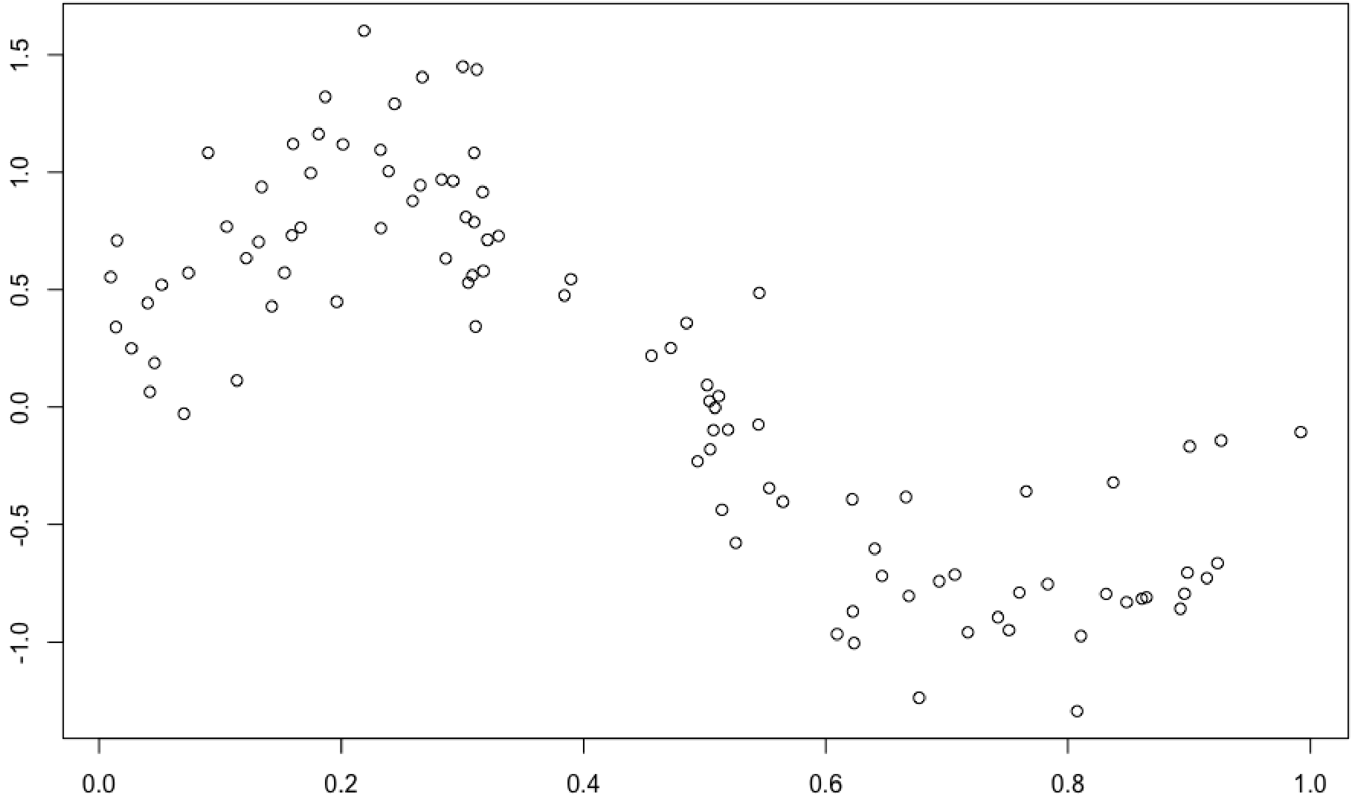
- Training data:

  100 sampled data points generated by the function h

- Prediction Models: 3 modes
  - o 0-order polynomial: $y = w_0$
  - o 1-order polynomial: $y = w_0 + w_1 x$
  - o 3-order polynomial: $y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x^3$
  - o 15-order polynomial: $y = w_0 + w_1 x_1 + \cdots + w_{15} x^{15}$

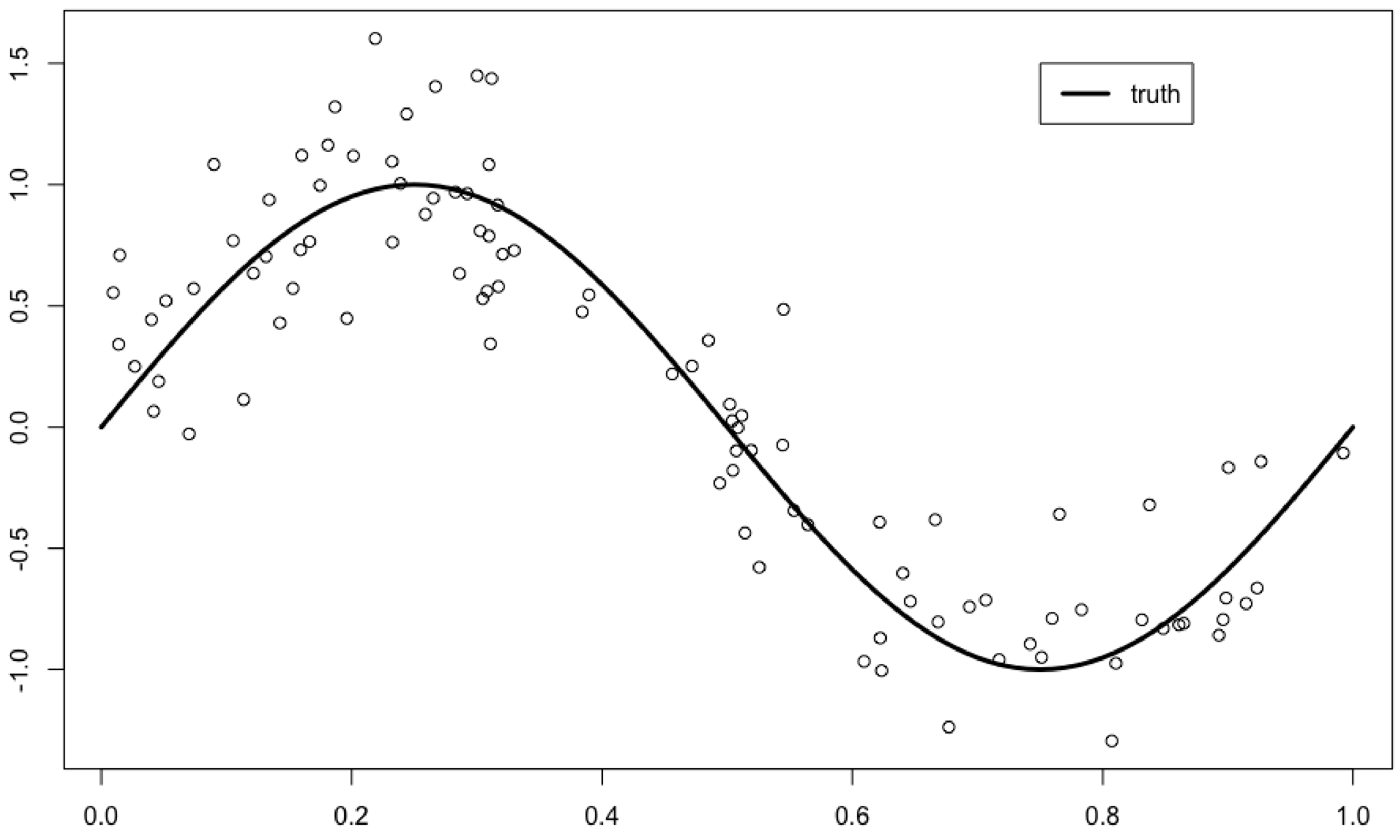
Example: controlled by number of parameters

- 100 training data points



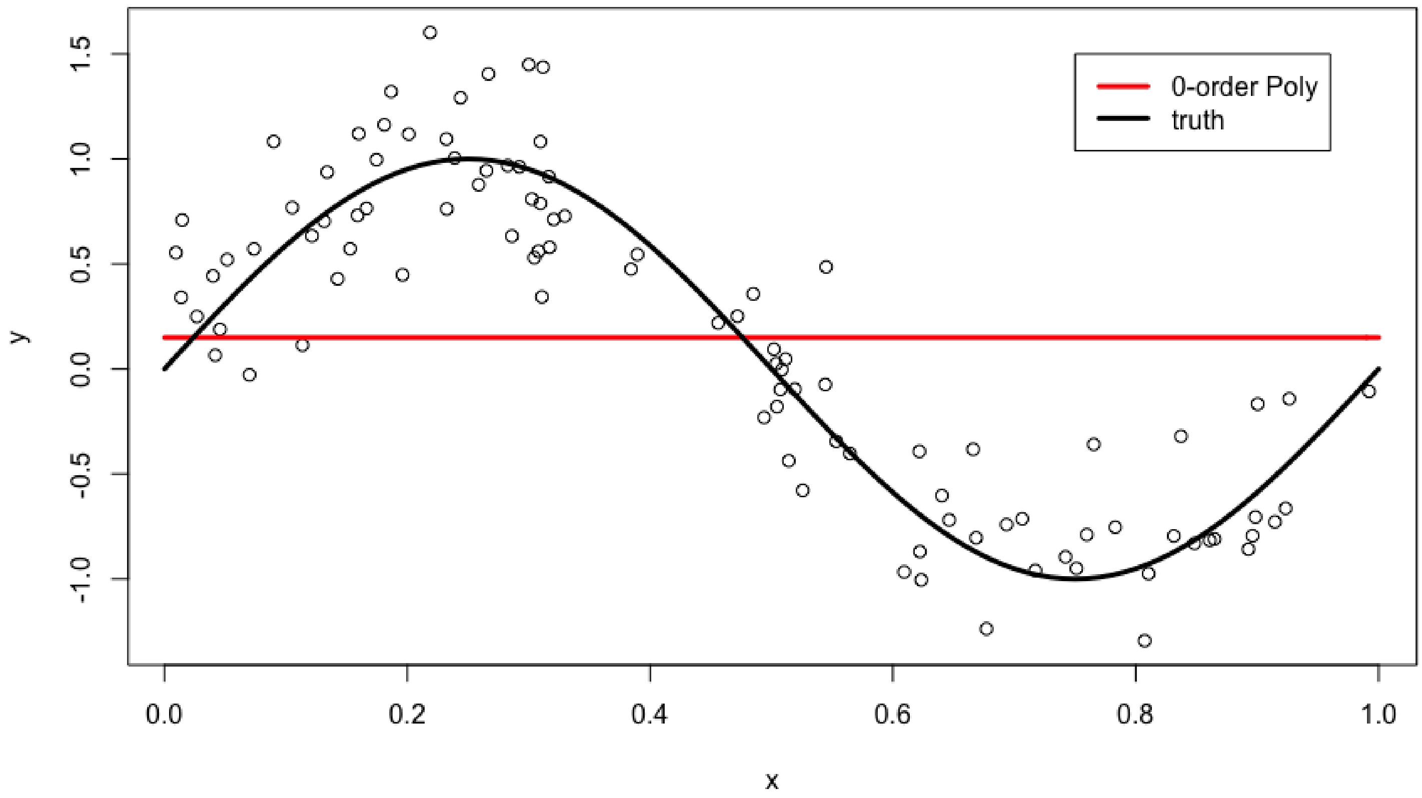
Example: controlled by number of parameters

- The true function and its curve: $h(x) = \sin(2\pi x) + \epsilon$



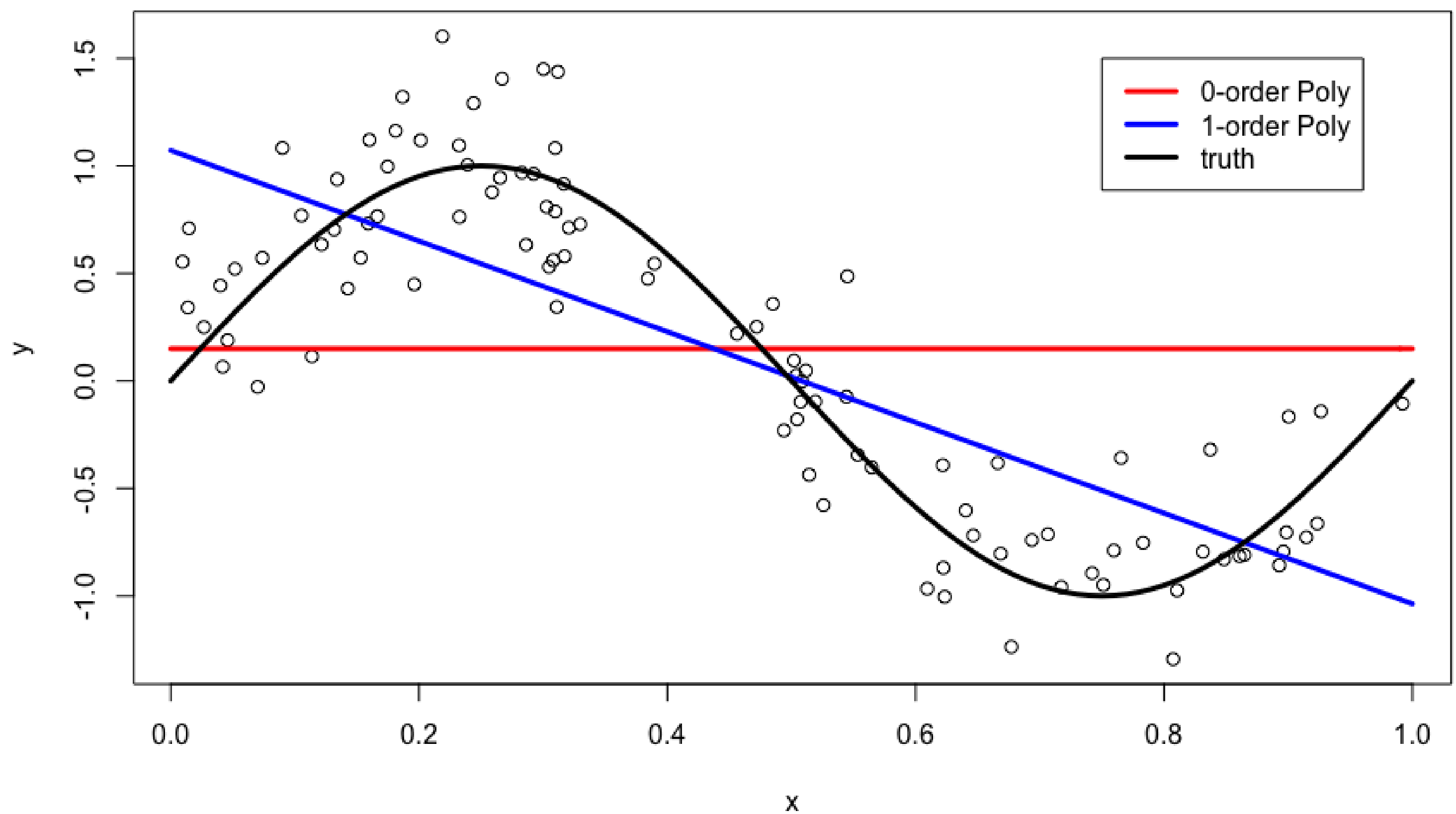
Example: controlled by number of parameters

- The 0-degree polynomial model: $y = w_0$ (red line)



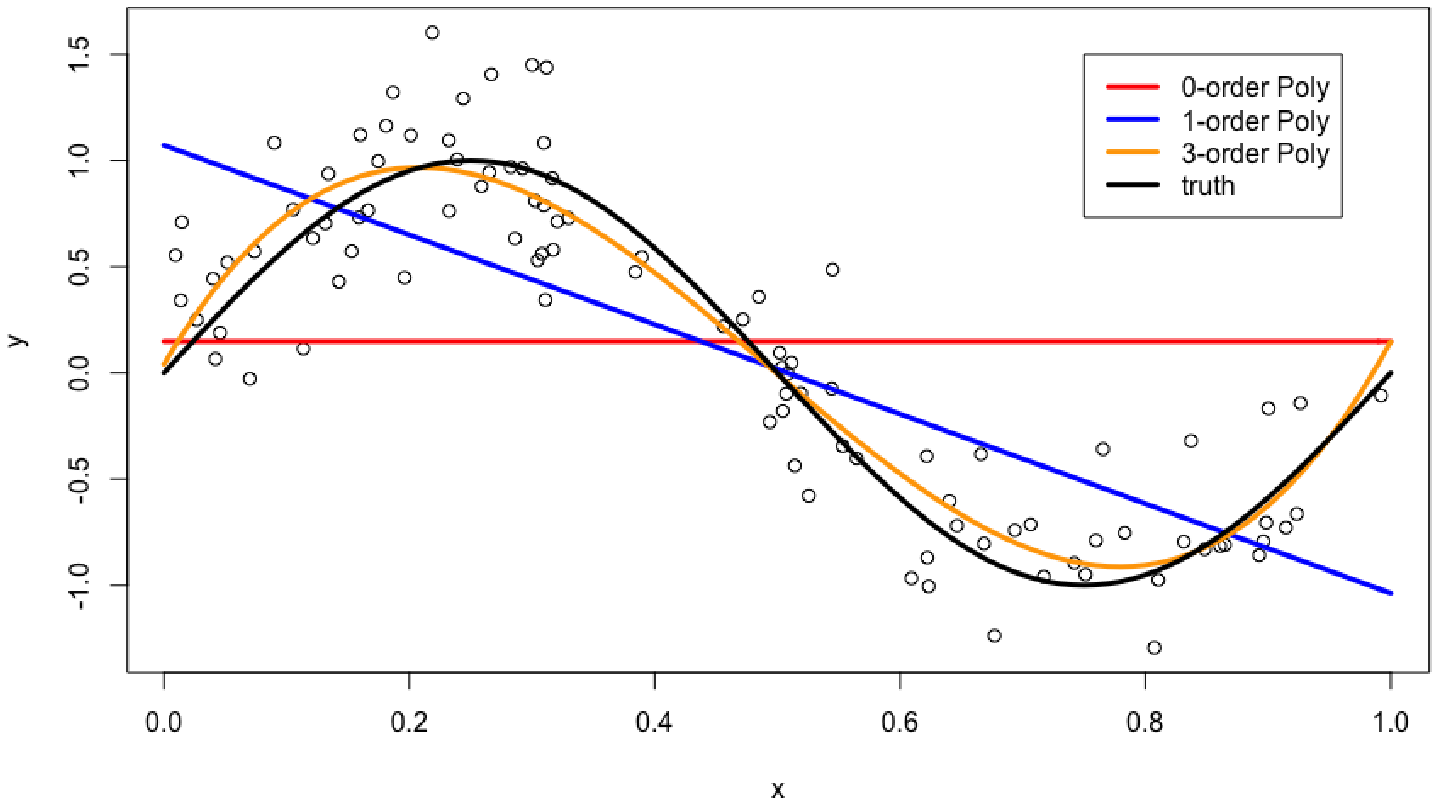
Example: controlled by number of parameters

- The 1-degree polynomial model: $y = w_0 + w_1 x$ (blue line)



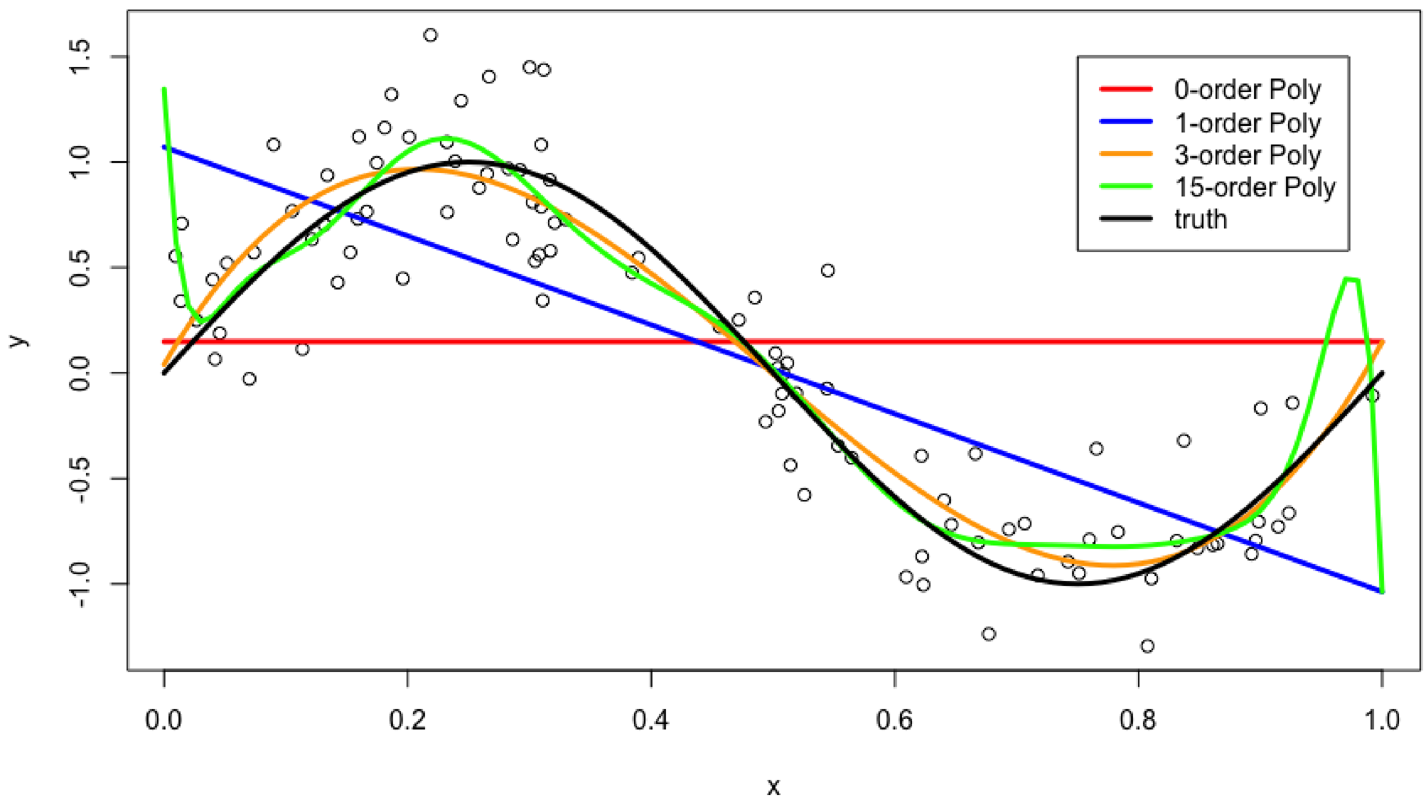
Example: controlled by number of parameters

- The 3-degree polynomial model: $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$ (orange curve)



Example: controlled by number of parameters

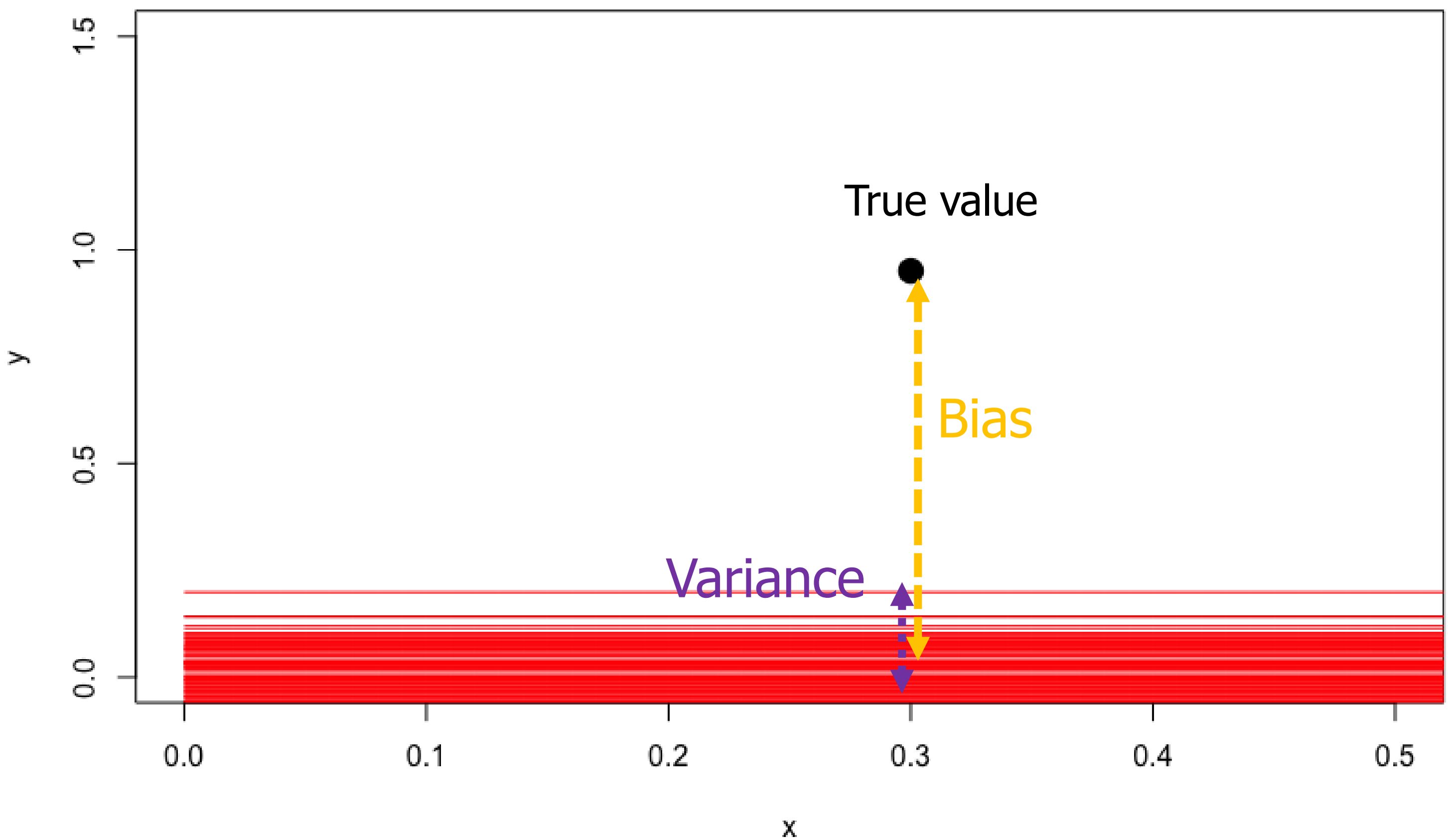
- The 15-degree polynomial model: $y = w_0 + w_1 x + \cdots + w_{15} x^{15}$ (green curve)



Bias?
Variance?

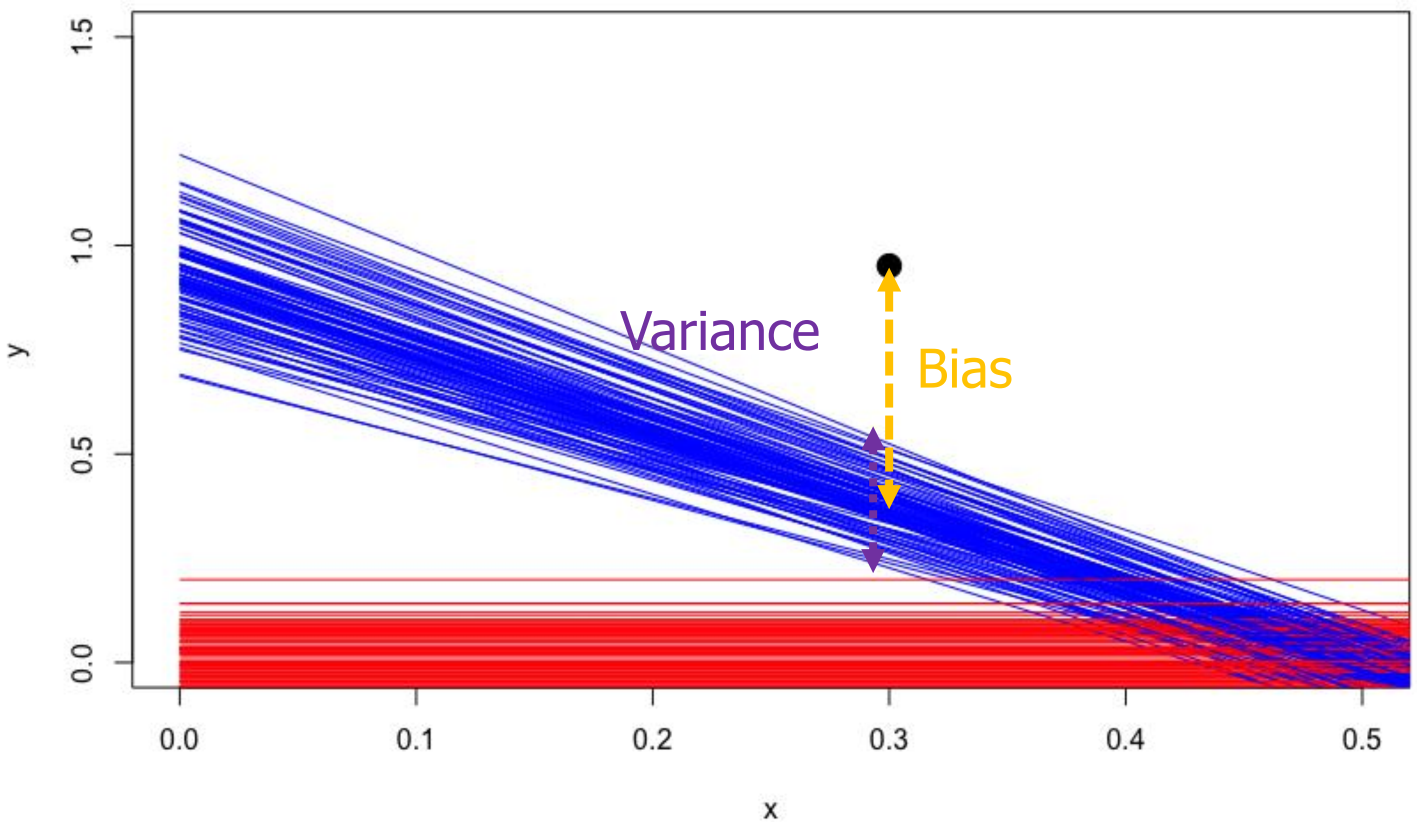
Example: controlled by number of parameters

- ❑ Estimate the bias and variance at the point $x = 0.3$
- ❑ 100-time experiments: 0-order polynomial



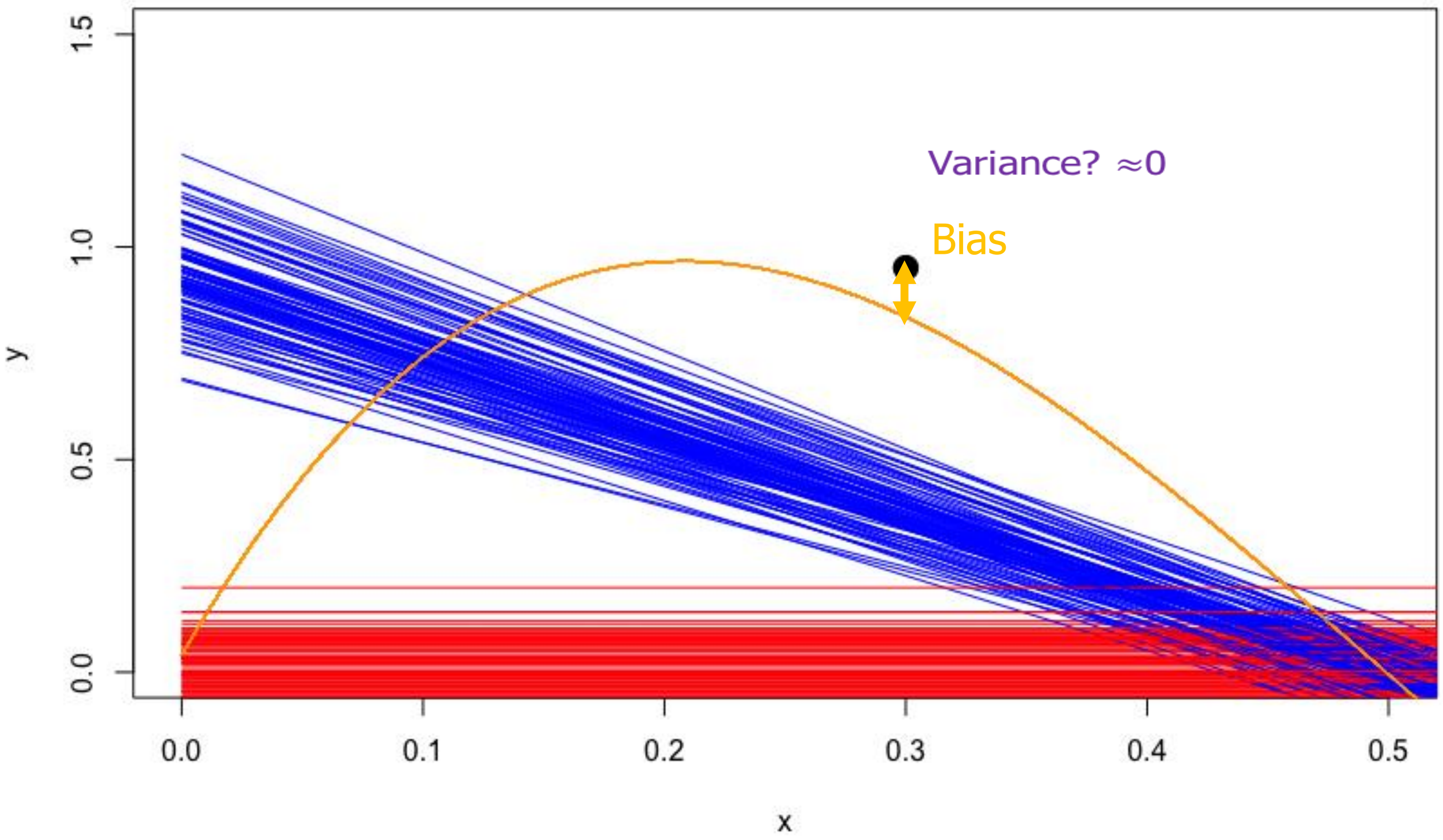
Example: controlled by number of parameters

- Estimate the bias and variance at the point $x = 0.3$
- 100-time experiments: 1-order polynomial model



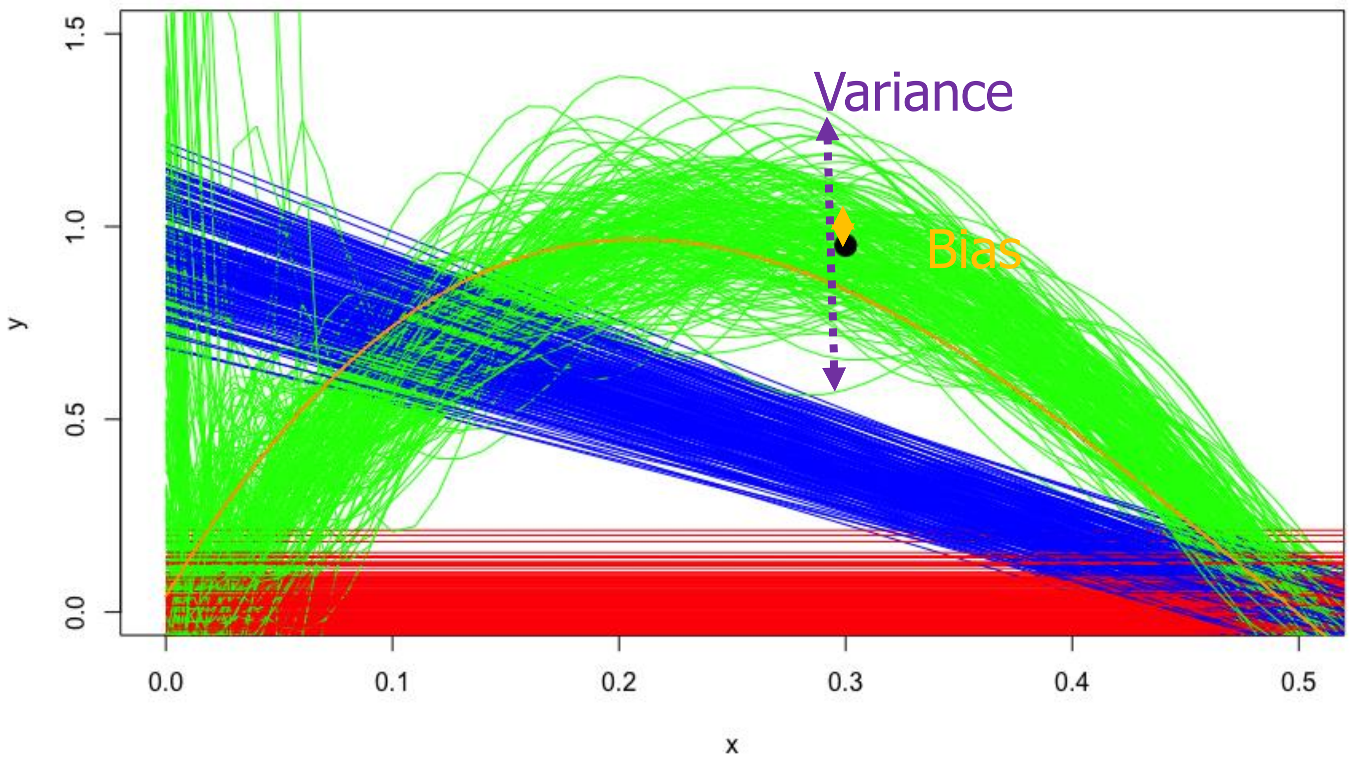
Example: controlled by number of parameters

- Estimate the bias and variance at the point $x = 0.3$
- 100-time experiments: 3-order polynomial model



Example: controlled by number of parameters

- Estimate the bias and variance at the point $x = 0.3$
- 100-time experiments: 15-order polynomial model



Comparison: controlled by number of parameters (quantitative)

Model	Bias	Variance	MSE
0-order Poly	0.9117	0.0052	1.0361
1-order Poly	0.3353	0.0039	0.4539
3-order Poly	0.0039	0.0028	0.1032
15-order Poly	0.0008	0.0121	0.1069

Worst=Blue, Best=Red

Useful Guidance in Practice

Key things to think about when managing Bias and Variance

My model has a high error on a test set. Do I need to train my model on a larger dataset?

- When your model has a high variance, you can try this.

- But if your model has a high bias, it will not fix the problem.

Useful Guidance in Practice

Key things to think about when managing Bias and Variance

Do I need to train my model with smaller sets of features (i.e. predictors)?

- When your model has a high variance, you can try this.
- But if your model has a high bias, it will not fix the problem.

Do I need to obtain new features?

- When your model has a high bias, usually this way works well.

Tutorial (Week 4)

Bias and Variance in Regression.

What will we learn in Week 5

- **Part A in Module 3**
  - Linear Models for Classification

- Online quiz has been released


THANK YOU
for your attention!