Lecture 11: Overfitting

Presentation · July 2019

CITATIONS
0

READS
163

1 author:

Alaa Tharwat
Fachhochschule Bielefeld
120 PUBLICATIONS **6,195** CITATIONS

SEE PROFILE

Lecture 11: Overfitting

Alaa Tharwat

Lecture 11: Overfitting

- Review of Lecture 10
- Definition of overfitting
- The role of noise
- Deterministic noise
- Solutions for overfitting problem

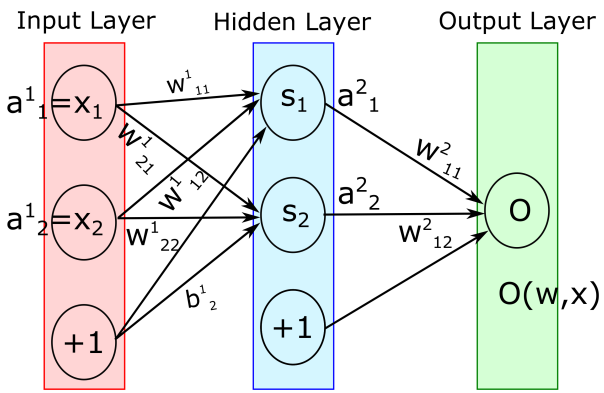
Lecture 11: Overfitting

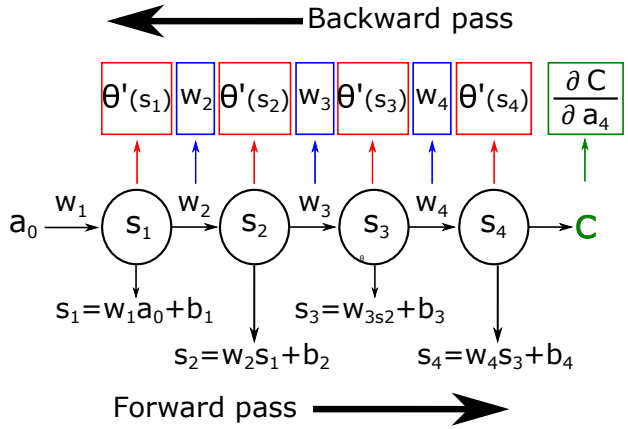
- Review of Lecture 10
- Definition of overfitting
- The role of noise
- Deterministic noise
- Solutions for overfitting problem

$$w_{ij}^{l} = \begin{cases} 1 \leq l \leq L & layers \\ 0 \leq i \leq d^{(l-1)} & inputs \\ 1 \leq j \leq d^{(l)} & outputs \end{cases}$$

$$\delta_i^{(l-1)} = \frac{\partial e(\mathbf{w})}{\partial s_i^{(l-1)}} = \sum_{j=1}^{d(l)} \frac{\partial e(\mathbf{w})}{\partial s_j^{(l)}} \times \frac{\partial s_j^{(l)}}{\partial x_i^{(l-1)}} \times \frac{\partial x_i^{(l-1)}}{\partial s_i^{(l-1)}}$$

$$= \sum_{j=1}^{d(l)} \delta_j^{(l)} \times w_{ij}^{(l)} \times \theta'(s_i^{(l-1)}) = (1 - (x_i^{(l-1)})^2) \sum_{j=1}^{d(l)} w_{ij}^{(l)} \delta_j^{(l)}$$

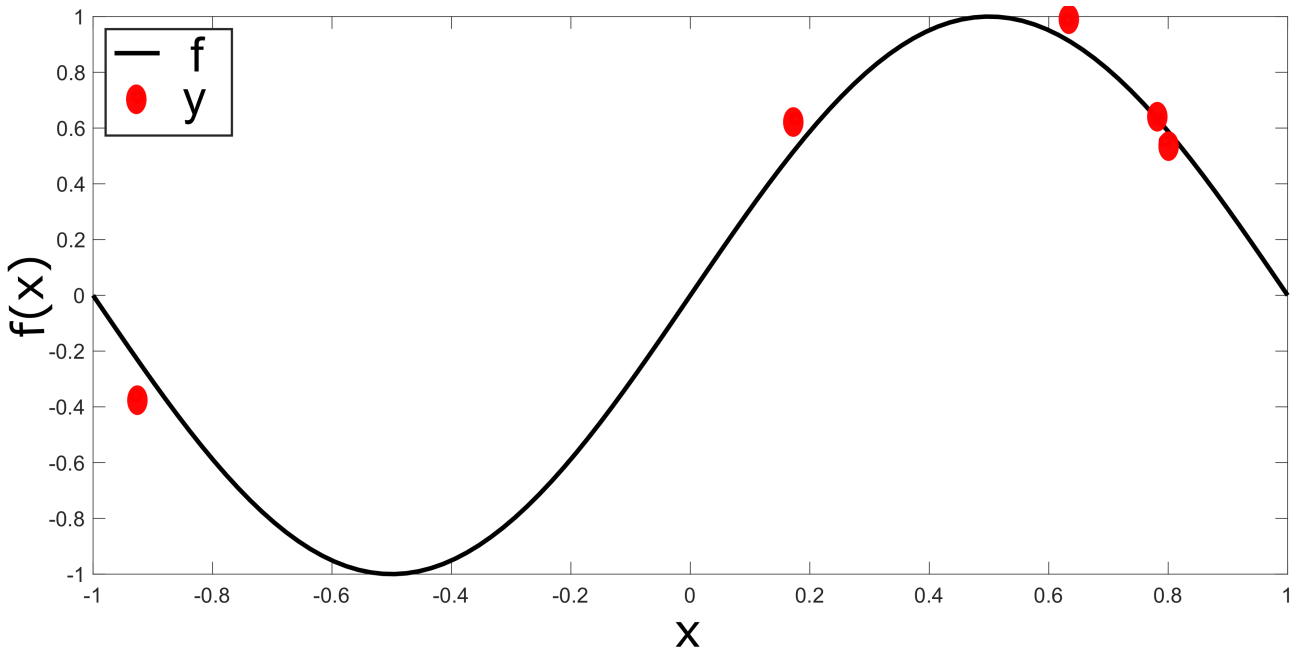




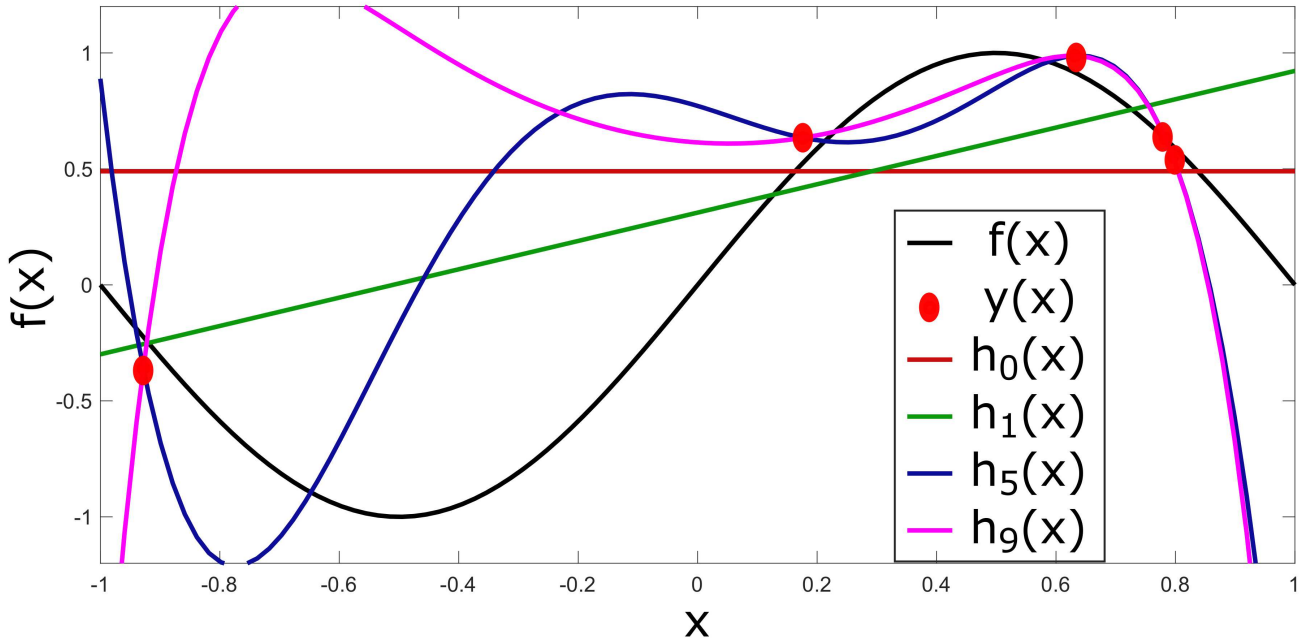
Lecture 11: Overfitting

- Review of Lecture 10
- Definition of overfitting
- The role of noise
- Deterministic noise
- Solutions for overfitting problem

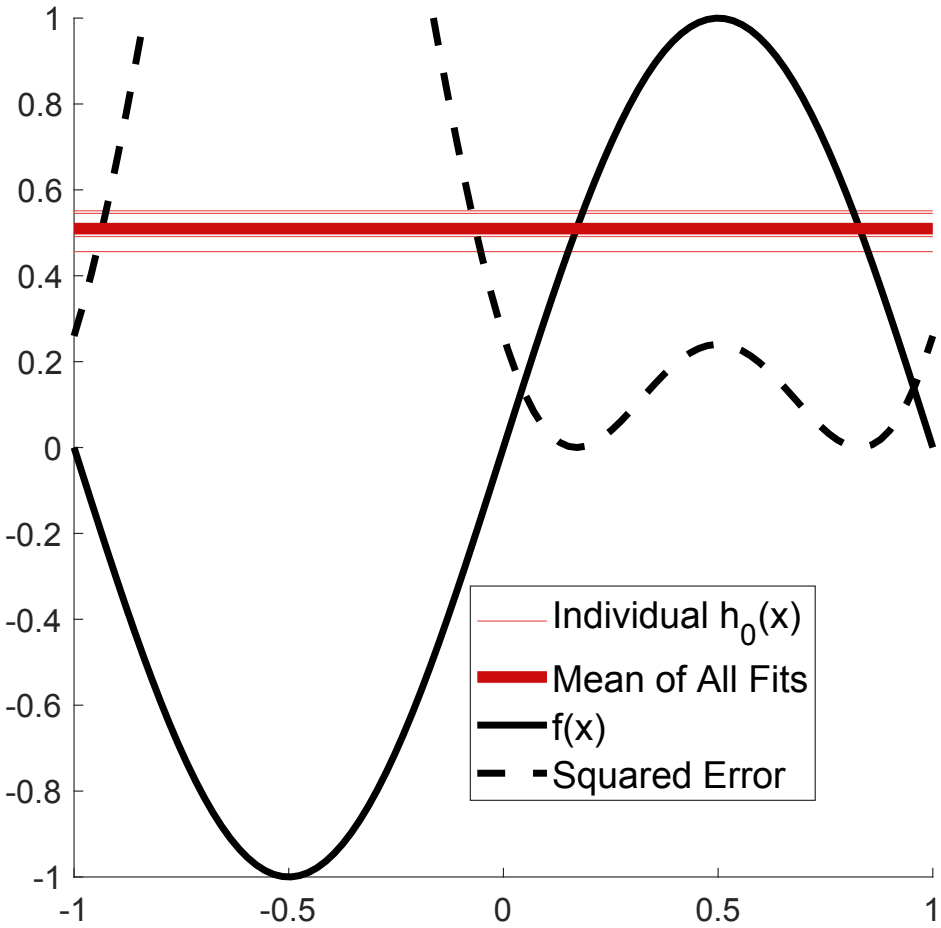
- Given, 5 samples (noisy samples) and simple target function ($f(x) = sin(\pi x)$) and $y(x) = f(x) + \epsilon$, where $\epsilon = 0.1$



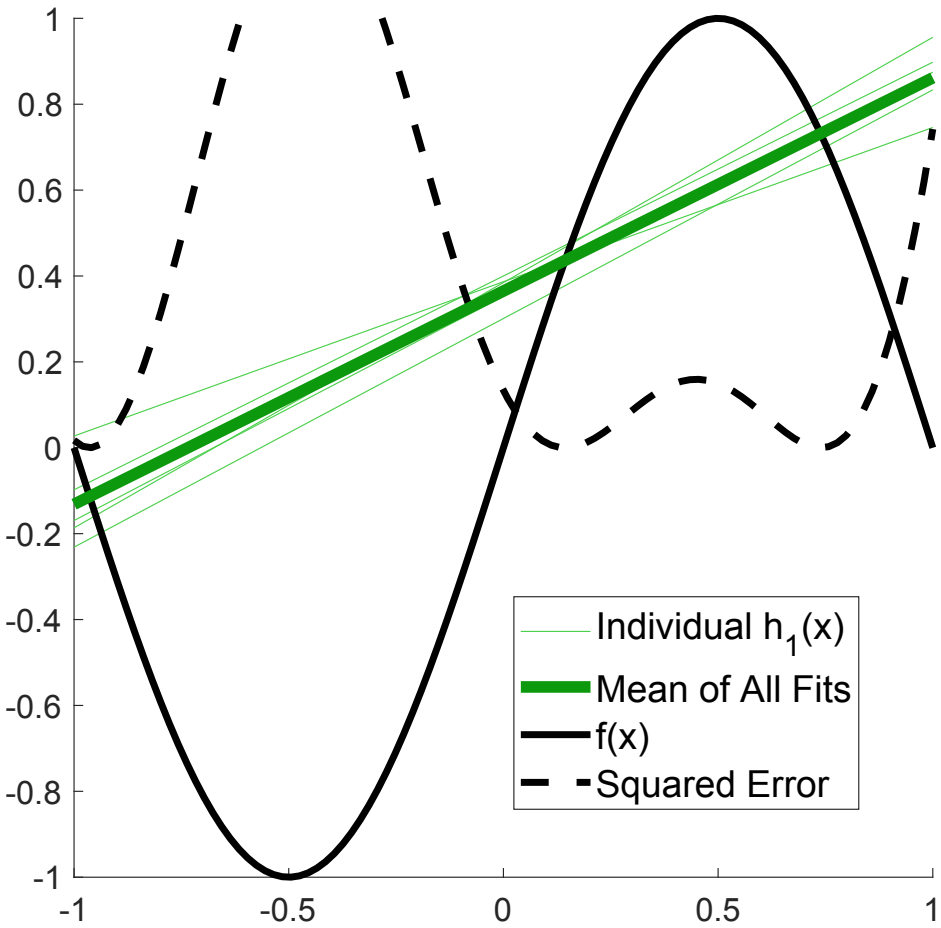
- Given four models $h_0(x)$ is a constant, $h_1(x)$ is a linear model, $h_5(x)$ quadratic model with degree five, and $h_9(x)$ quadratic with degree nine



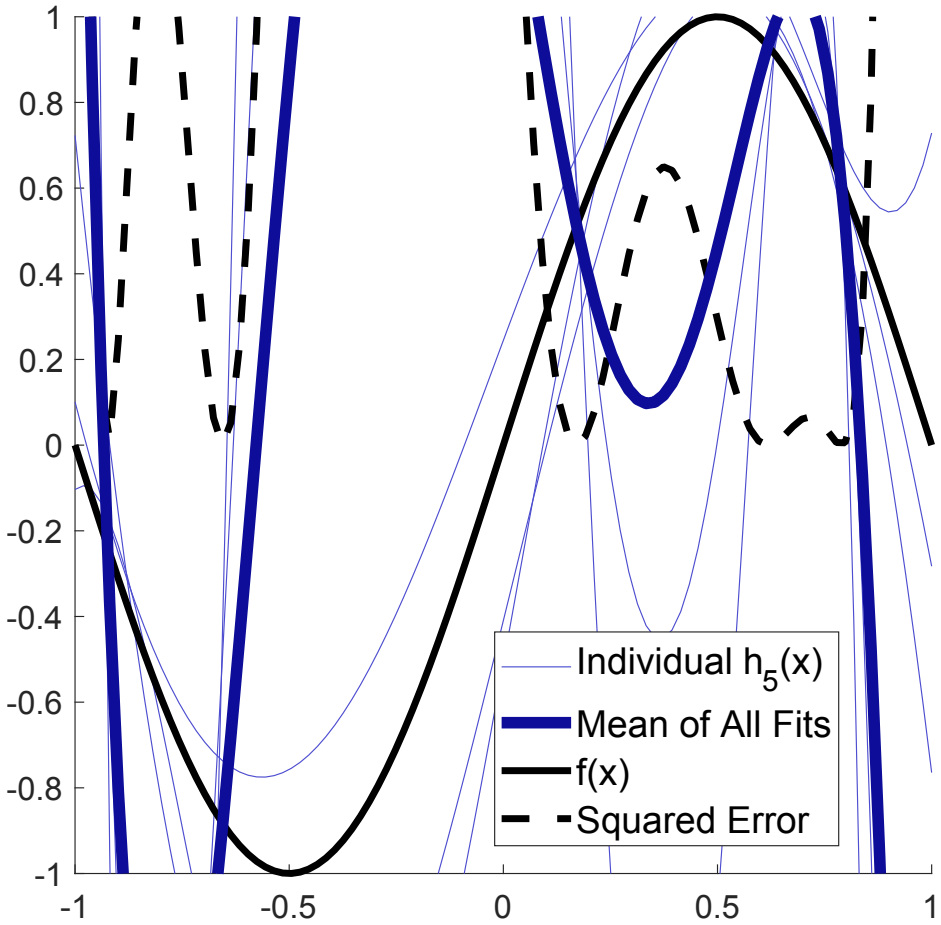
- Constant model h_0 with five different training data $(D_1, D_2, \ldots, D_5)$ each has five samples



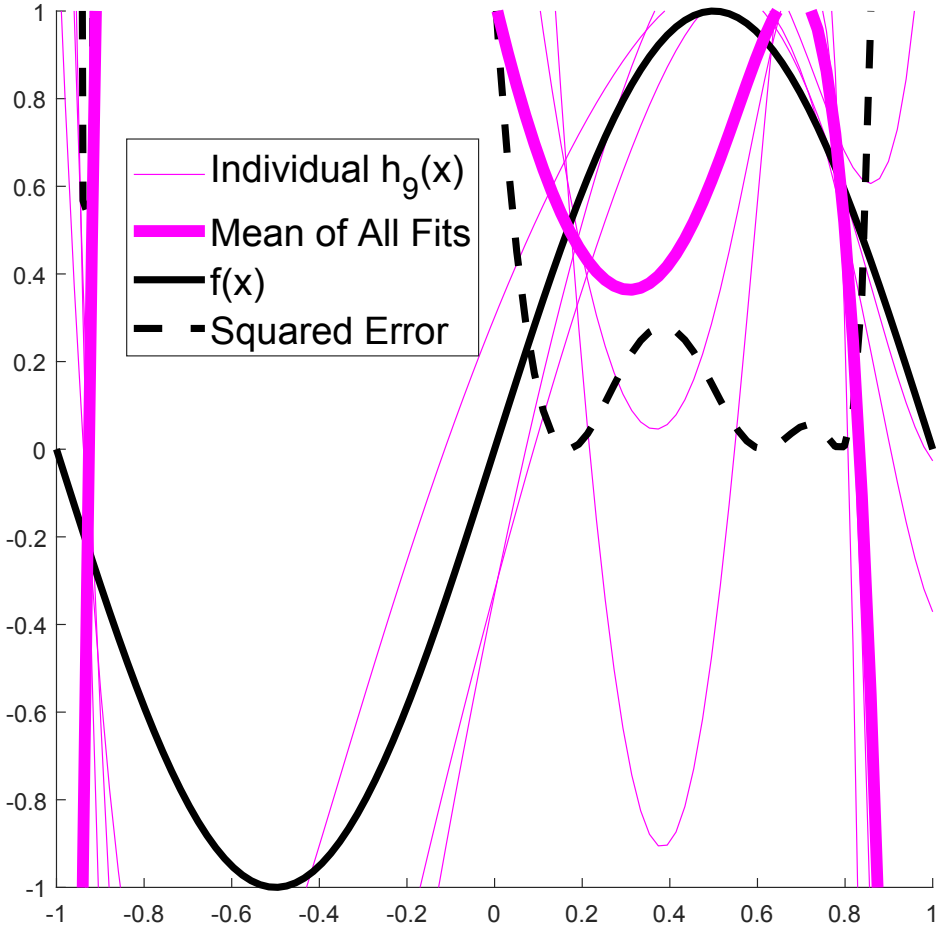
- Linear model h_1 with five different training data $(D_1, D_2, \ldots, D_5)$ each has five samples

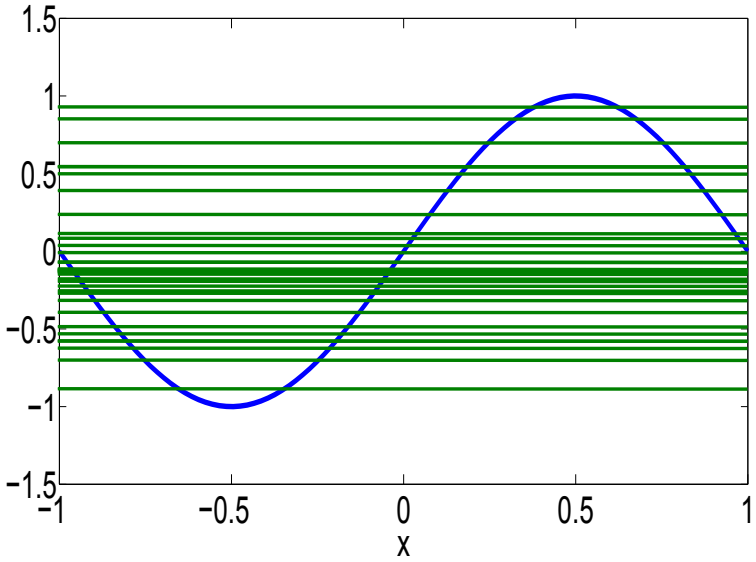


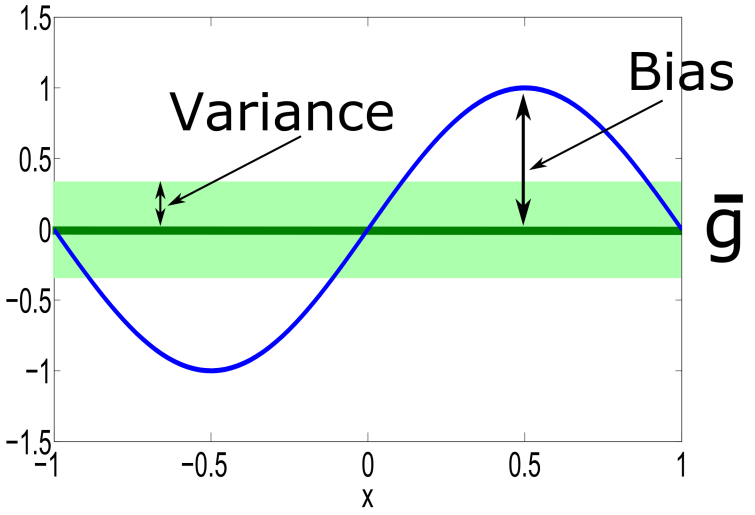
- 5^{th}-degree Quadratic/polynomial model h_5 with five different training data $(D_1, D_2, \ldots, D_5)$ each has five samples

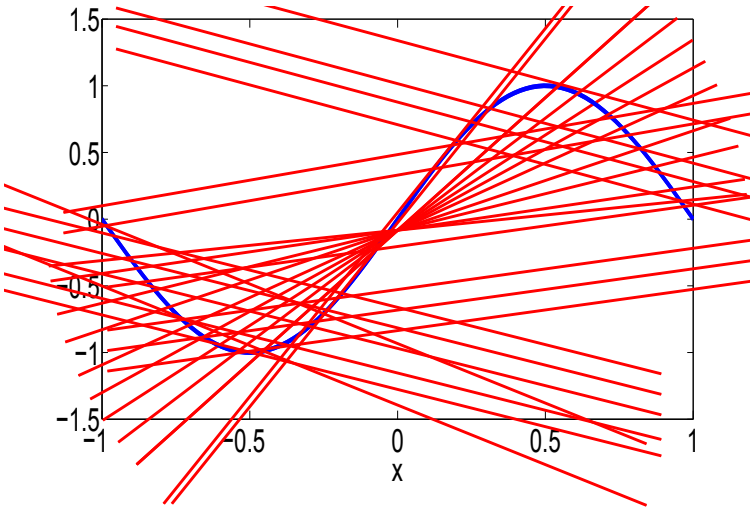


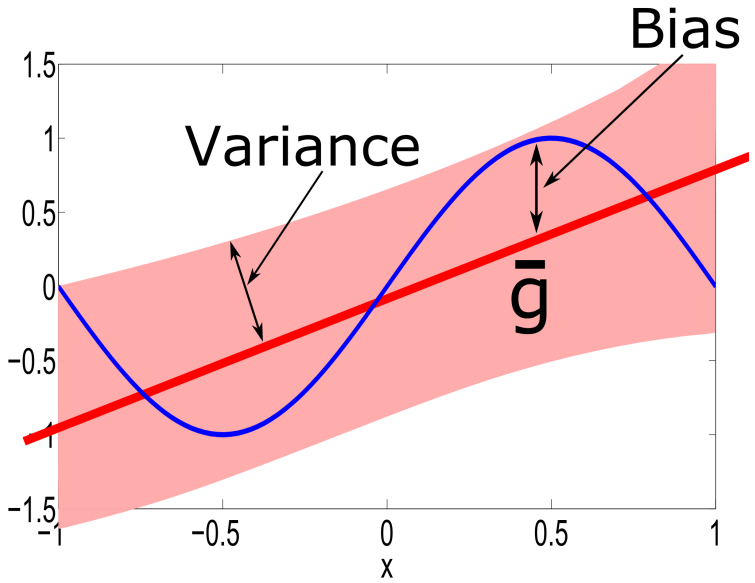
- 9^{th}-degree Quadratic/polynomial model h_9 with five different training data $(D_1, D_2, \ldots, D_5)$ each has five samples



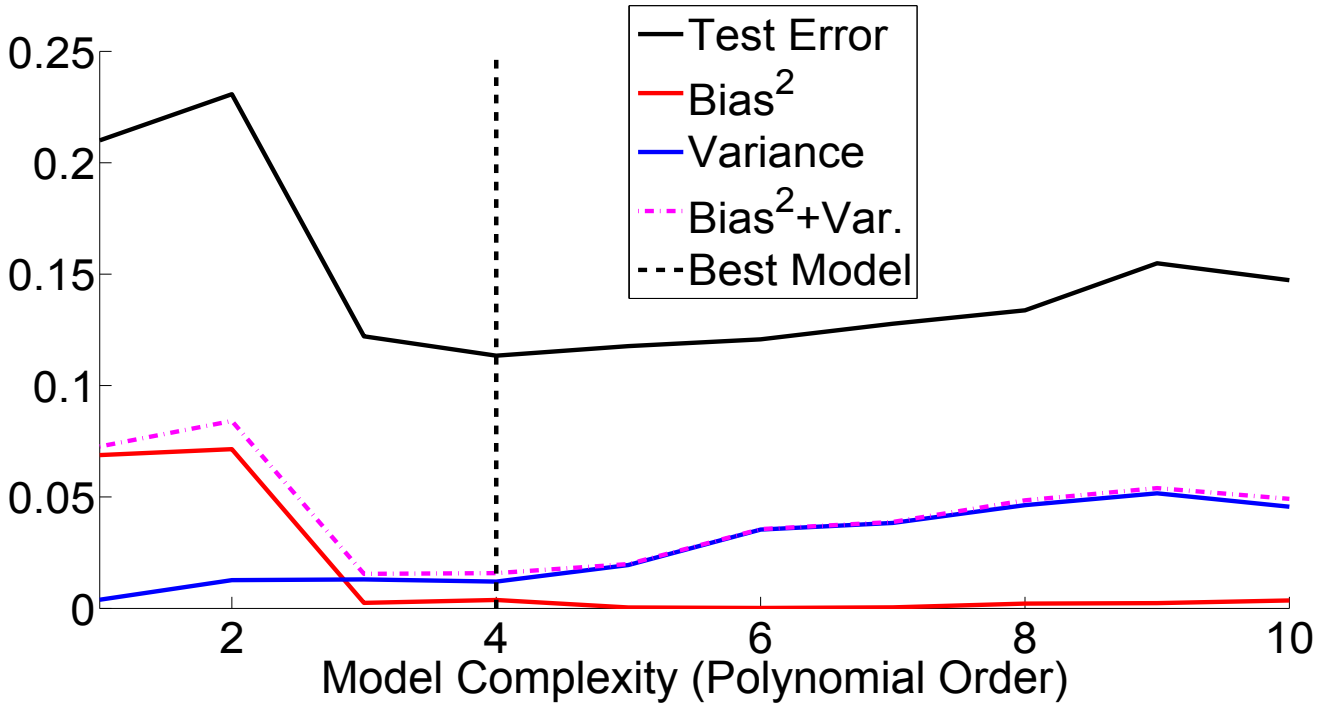

1.5
1
0.5
0
−0.5
−1
−1.5
−1
−0.5
0
0.5
1
x


Variance
Bias
ḡ
1.5
1
0.5
0
−0.5
−1
−1.5
−1
−0.5
0
0.5
1
x

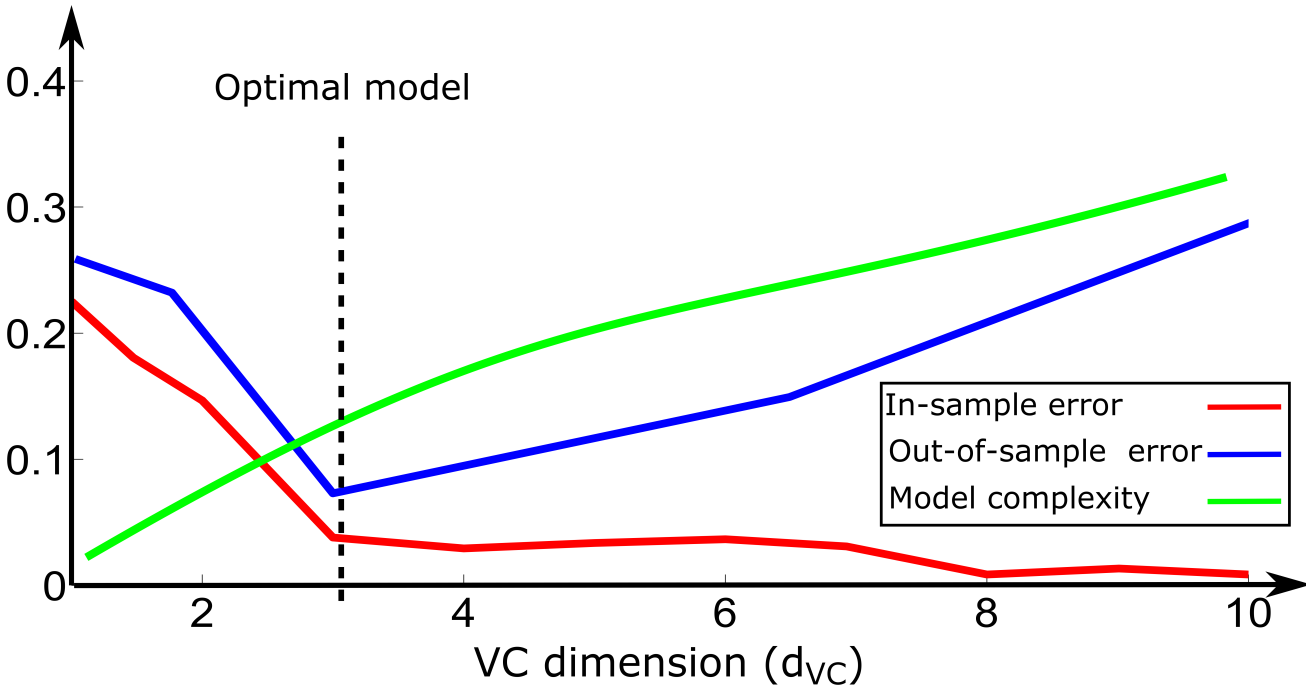

1.5
1
0.5
0
−0.5
−1
−1.5
−1
−0.5
0
0.5
1
x

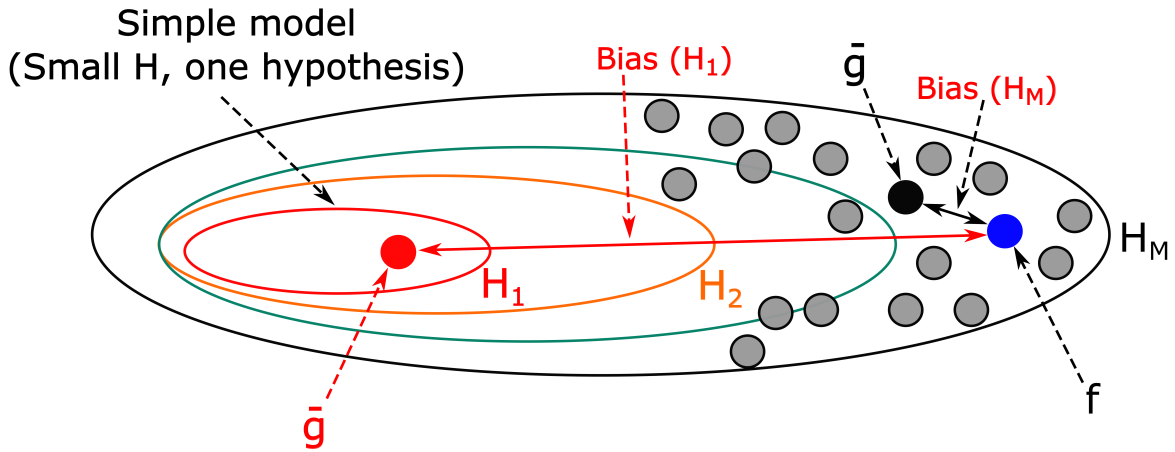

Bias
Variance
ḡ
1.5
1
0.5
0
−0.5
−1
−1.5
−1
−0.5
0
0.5
1
x

- Overfitting is a comparative term
- Overfitting: $E_{in} \downarrow$ and $E_{out} \uparrow$
- Overfitting: fitting the data more than is warranted, these data include noise or outliers and this is the problem of overfitting



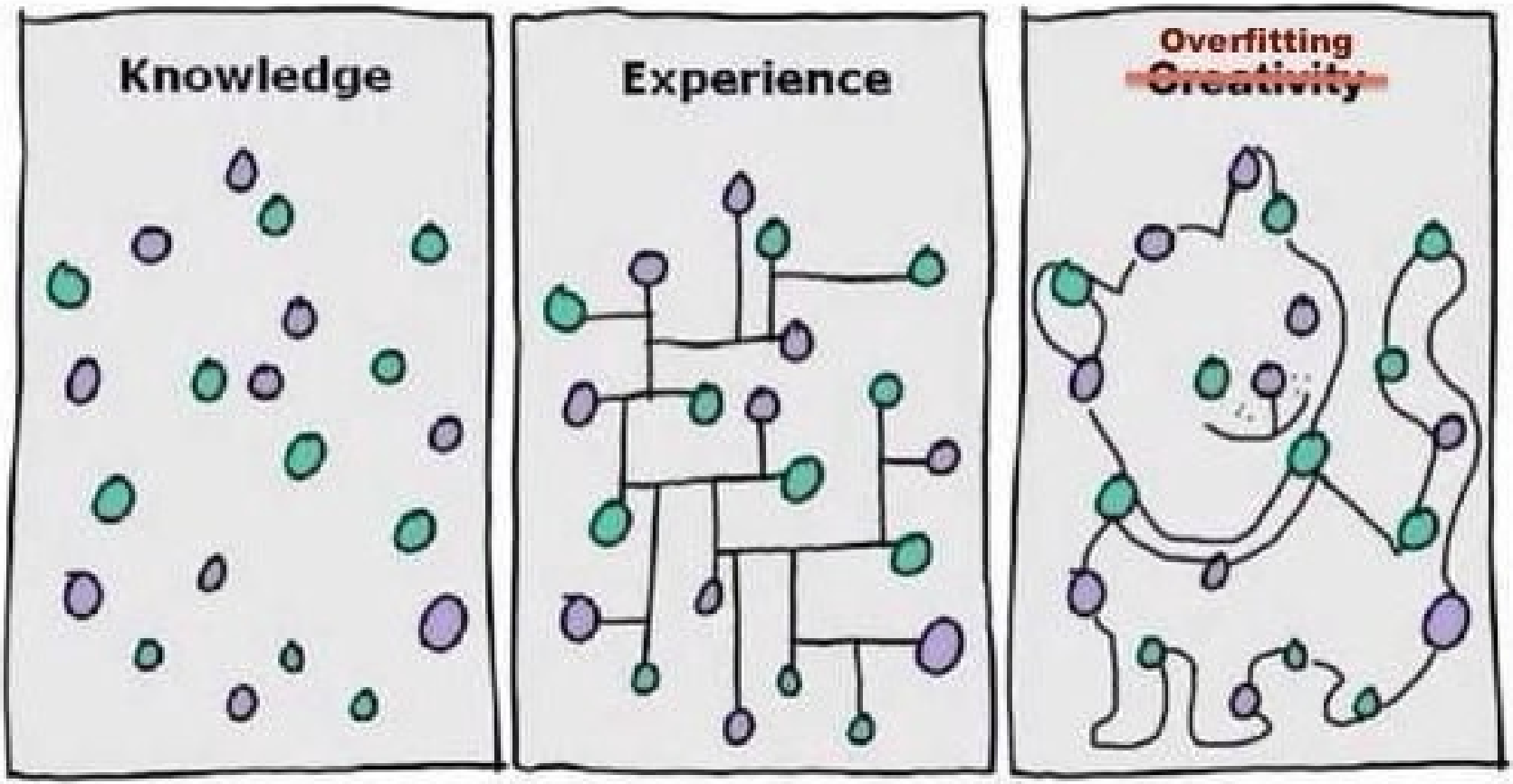
Using a model such as neural networks or SVM





- With high degree polynomials, the model is more complex, $E_{in} = 0$, BUT, E_{out} is huge $\rightarrow$ bad generalization (see model h_9) [Overfitting problem]
- With low degree polynomials, the model is simple, E_{in} is low but not zero, and E_{out} is low (compared to the complex model)
- With overfitting: $E_{in} \downarrow$ and $E_{out} \uparrow$
- The problem of overfitting occurs when we try to fit the data to the extent that the model fits the noise and outliers


"AI expert"
more layers
overfitting model


Knowledge
Experience
Overfitting
Creativity


My model with 99% accuracy
Me
Overfitting

Assume we have a simple model, $g(x) = c$, the model consists of only constant

$$E_D[(g^{(D)}(x) - f(x))^2] = E_D[(g^{(D)}(x) - \bar{g}(x) + \bar{g}(x) - f(x))^2]$$
$$= E_D[(g^{(D)}(x) - \bar{g}(x))^2] + E_D[(\bar{g}(x) - f(x))^2]$$
$$+ 2E_D[(g^{(D)}(x) - \bar{g}(x))(\bar{g}(x) - f(x))]$$
$$= \text{Variance} + \text{Bias}$$
$$+ \underbrace{2(E_D[g^{(D)}(x)\bar{g}(x) - E_D[g^{(D)}(x)f(x)] - E_D[\bar{g}(x)^2] + E_D[\bar{g}(x)f(x)])}$$

and the term $2(E_D[g^{(D)}(x)\bar{g}(x) - E_D[g^{(D)}(x)f(x)] - E_D[\bar{g}(x)^2] + E_D[\bar{g}(x)f(x)])$ can be simplified as follows:

$$2(E_D[g^{(D)}(x)\bar{g}(x) - E_D[g^{(D)}(x)f(x)] - E_D[\bar{g}(x)^2] + E_D[\bar{g}(x)f(x)])$$
$$= 2(\bar{g}(x)\bar{g}(x) - \bar{g}(x)f(x) - \bar{g}(x)^2 + \bar{g}(x)f(x)) = 0$$

because $\bar{g}(x) = g^{(D)}(x)$ as mentioned before, and $E_D[f(x)] = f(x)$ because $f(x)$ is deterministic

- Hence, with simple models

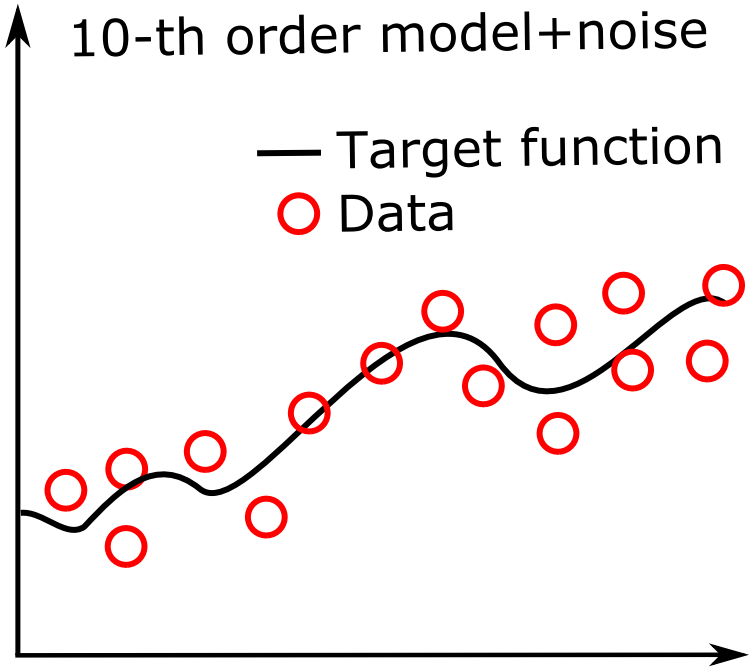
$$E_D[g^{(D)}(x) - (f(x))^2] = \text{Variance} + \text{Bias}$$

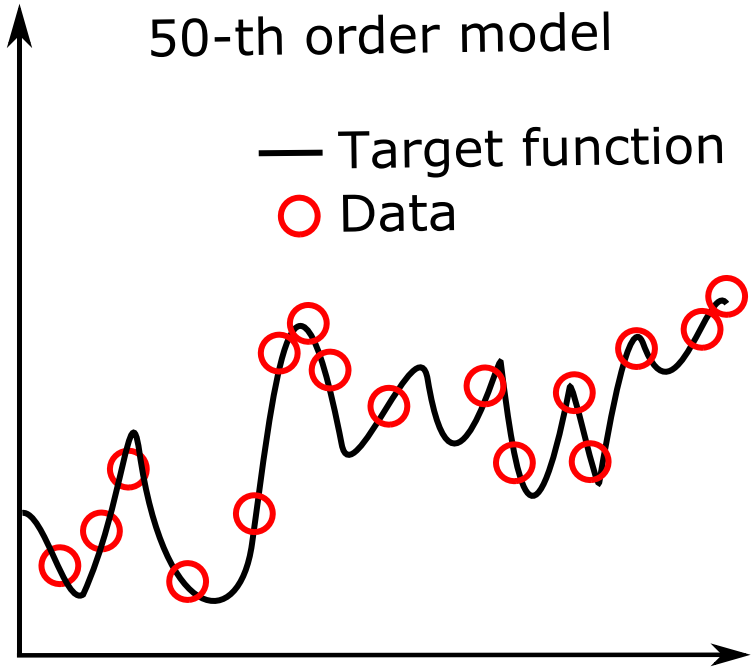
- Since the model is constant; as a consequence, the variance is small or end to be zero
- The bias error is high.
- On the other hand, given a complex model; more training data will perfectly interpolated and hence lower bias will be obtained but the variance will be high.
- The high bias indicates that the learning model cannot find the relevant relationships between the given data and the target outputs, this is called **Underfitting**.
- More complex models are able to represent the training data more accurately with low bias BUT with high variance, this is called the **Overfitting** problem.

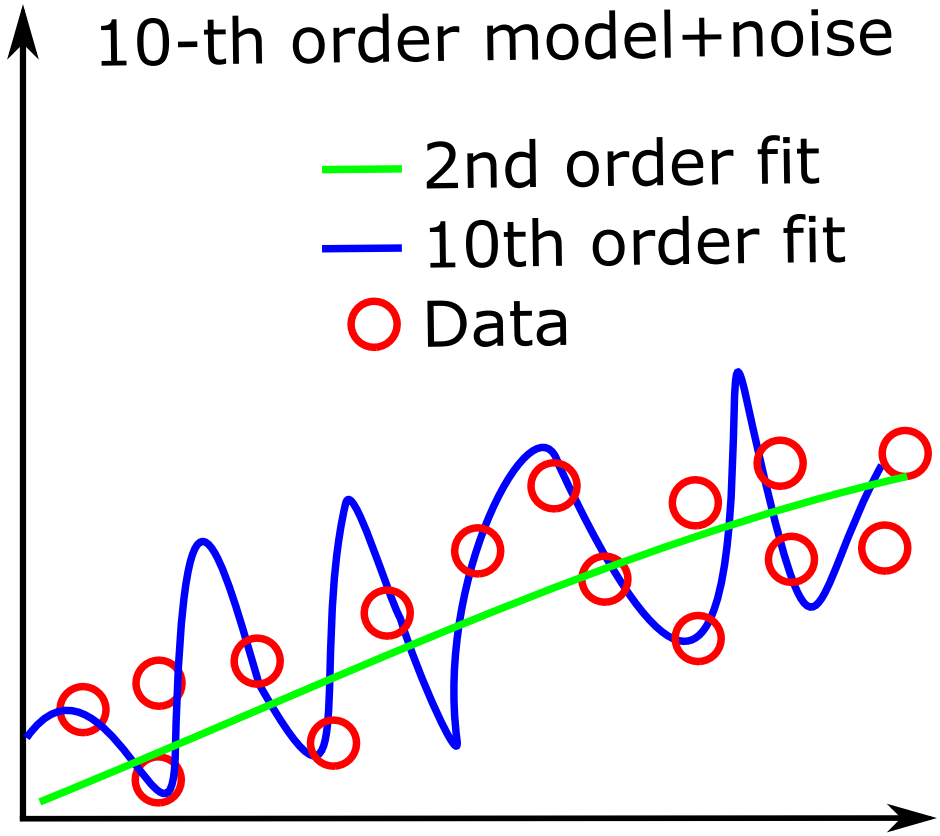
Lecture 11: Overfitting

- Review of Lecture 10
- Definition of overfitting
- The role of noise
- Deterministic noise
- Solutions for overfitting problem

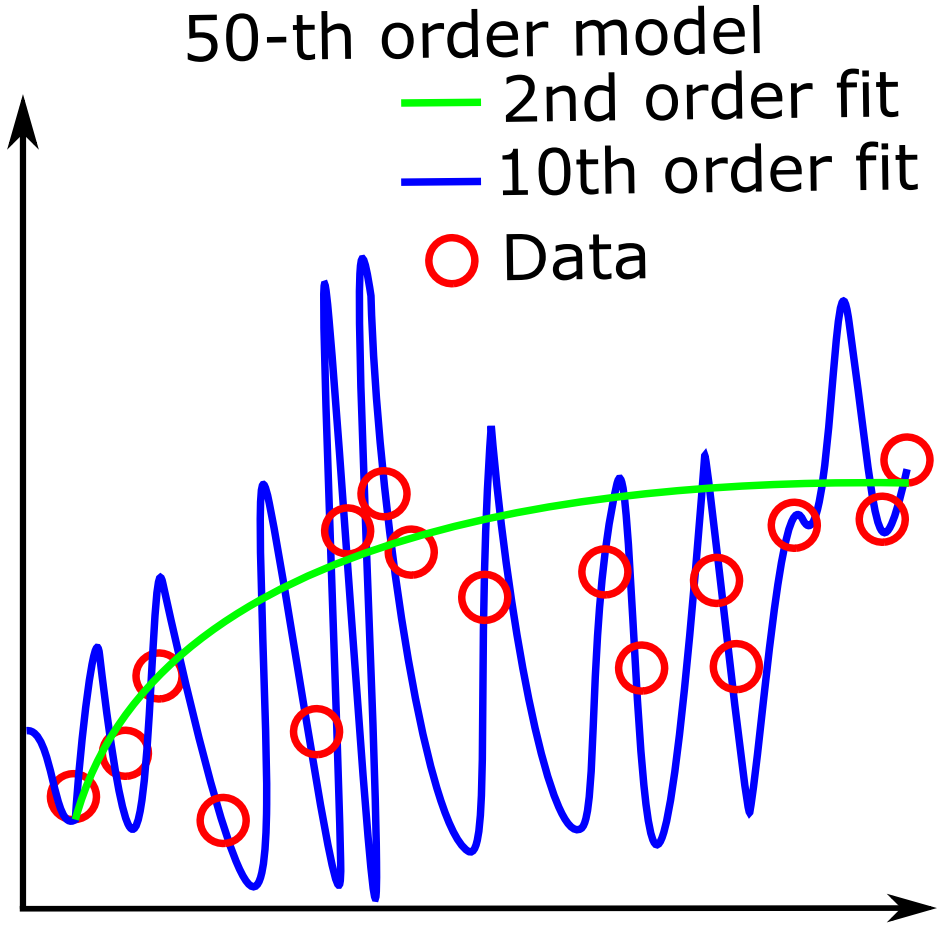
- Given two targets, simple model (2^{nd} order) with noise and complex model (10^{th} order) without noise





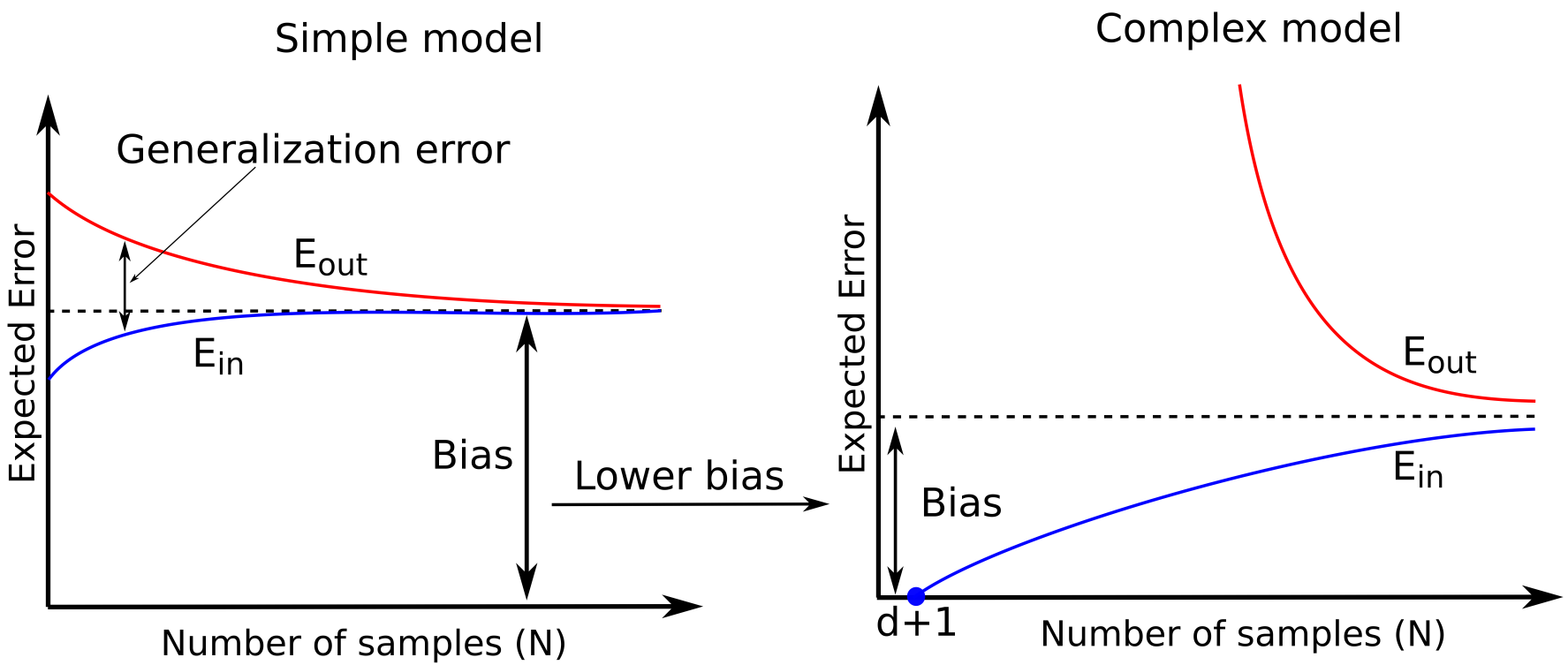


- With noisy low-order target
  - With 2^{nd} order model, $E_{in} = 0.05$ and $E_{out} = 0.127$
  - With 10^{th} order model, $E_{in} = 0.034$ and $E_{out} = 9.0$
- The model with 10^{th} order which matches the complexity of the target function lost, because the model matches the data resources than the target complexity



- With noiseless high-order target
  - With 2^{nd} order model, $E_{in} = 0.029$ and $E_{out} = 0.120$
  - With 10^{th} order model, $E_{in} \approx 0$ and $E_{out} = 7680$

- The dotted line represents the error which comes from (1) the noise and (2) limitations of the model, i.e. inability of the 2nd order to fit 10 order target
- More samples reduce the error
- The complex model lost when the number of samples is inadequate



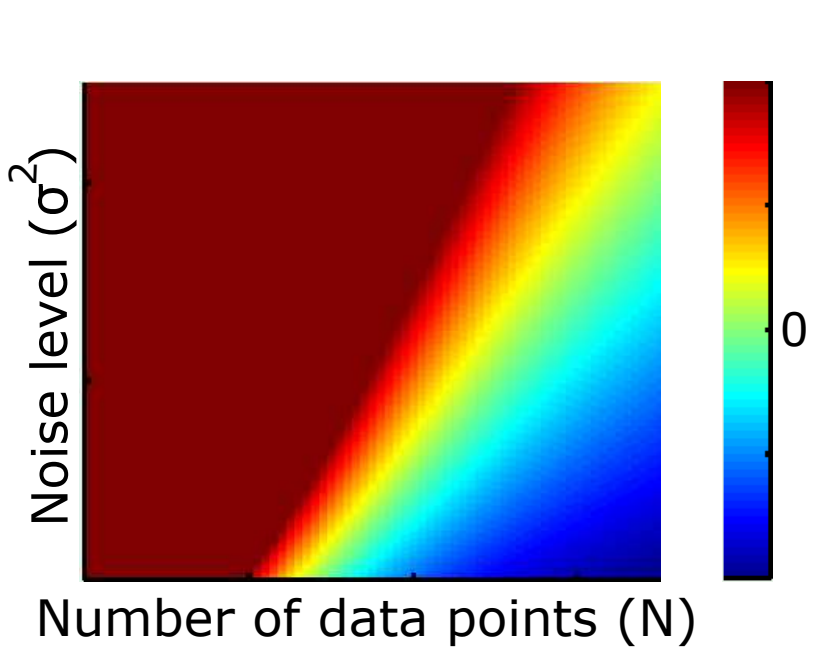
Lecture 11: Overfitting

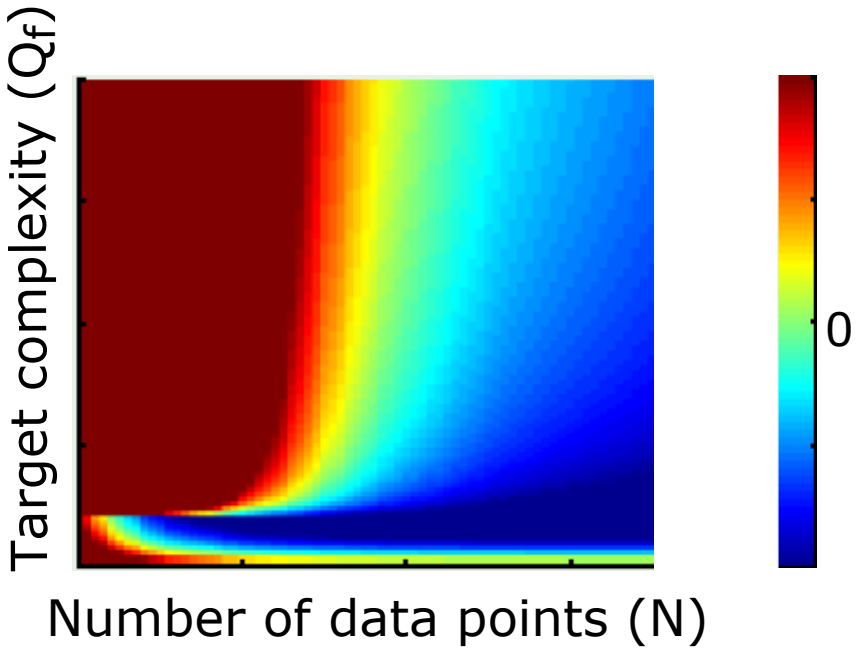
- Review of Lecture 10
- Definition of overfitting
- The role of noise
- Deterministic noise
- Solutions for overfitting problem

Impact of **noise level** and **target complexity**

- Assume we have noisy data $y(x) = f(x) + \epsilon(x)$, $E[\epsilon] = 0$ and $Var[\epsilon] = \sigma^2 \Rightarrow$ noise level
- To make the target function more complex (higher order polynomial), $y(x) = \sum_{q=0}^{Q_f} \alpha_q x^q + \epsilon(x)$, Q_f is the target complexity or order of the polynomial
- We have three parameters
  - Data size (N)
  - Noise level (σ^2)
  - Target complexity (Q_f)

- Given two models, simple model $g_2 \in H_2$ and complex $g_{10} \in H_{10}$
- Overfitting measures $E_{out}(g_{10}) - E_{out}(g_2)$





- More samples save from the overfitting
- Increasing the noise increases the overfitting (the model tries to fit noisy data)
- Increasing the target complexity increases the overfitting (the model tries to fit the generated data not the complexity of the target function)

- The noise is called **stochastic** noise, and the effect of target complexity is called **deterministic** noise

Impact of noise

- Number of samples ↑ Overfitting ↓
- Stochastic noise ↑ Overfitting ↑
- Deterministic noise ↑ Overfitting ↑

- The deterministic noise is a part of f that H cannot capture $(f(x) - h^*(x))$
- Deterministic noise is different than the stochastic noise where
  - It depends on H: for the same target function if we use more sophisticated hypotheses set, the deterministic noise will be smaller, but, stochastic noise is the same
  - Fixed for a given x, where the stochastic noise changes because it is generated randomly

$$
\begin{aligned}
E_D[E_{out}(g^{(D)})] &= E_D[(g^{(D)}(x) - f(x))^2] \\
&= E_D[(g^{(D)}(x) - \bar{g}(x))^2] + (\bar{g}(x) - f(x))^2 + E_D[\epsilon^2] \\
&= \text{Variance} + \text{Bias} + \sigma^2
\end{aligned}
$$

- The first term $(g^{(D)}(x) - \bar{g}(x))^2$ indicates how far your hypotheses set from the best possible you can get (**Variance**)
- The second term $(\bar{g}(x) - f(x))^2$ is the **Bias** (i.e. biased from the target function) and it represents how far the average/best model from the target function. Your Bias hypotheses set is biased away from the target function, Deterministic noise=bias
- The term σ^2 represents the variance of noise ($\text{Var}[\epsilon] = \sigma^2$). Thus, this term cannot be minimized and hence it is called *irreducible error*, and it is independent of the classification or regression model, this is also called Stochastic noise

Lecture 11: Overfitting

- Review of Lecture 10
- Definition of overfitting
- The role of noise
- Deterministic noise
- Solutions for overfitting problem

There are two well-known solutions for the overfitting problem

- **Regularization**: Regularization technique is used for solving the overfitting problem by adding an extra term to the cost function. This term is called the regularization term. This term is used to make a balance between minimizing the original cost function and finding small weights (Lecture 12)
- **Validation**: such as cross-validation (Lecture 13)