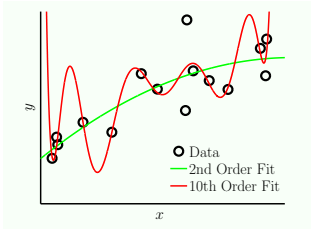
Approximation vs Generalization

Sections 2.3, 4.1 in LFD

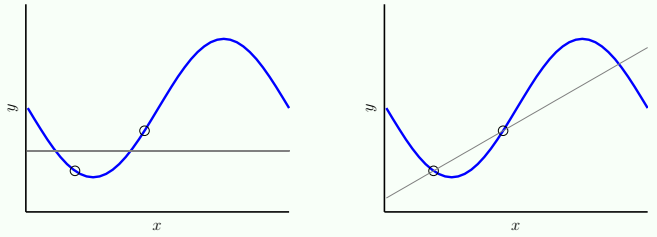


1

The bias-variance decomposition

Consider a simple learning problem: two data points and two hypothesis sets.

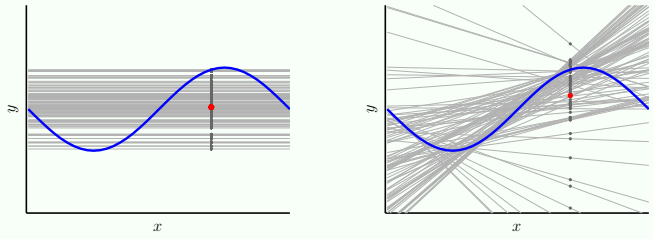
$$\mathcal{H}_0: \quad h(x) = b$$
$$\mathcal{H}_1: \quad h(x) = ax + b$$



Section 2.3

2

Repeating many times…



For each data set $\mathcal{D}$, you get a different $g^{\mathcal{D}}$.

So, for a fixed $\mathbf{x}$, $g^{\mathcal{D}}(\mathbf{x})$ is random value, depending on $\mathcal{D}$.

3

The bias-variance decomposition

Let's consider an out-of-sample error based on a squared error measure:

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

To abstract away the dependence on a given dataset:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]\end{aligned}$$

And let's focus on

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

4

The bias-variance decomposition

To evaluate $\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$

We consider the "average hypothesis" $\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right]$

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}) + \bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2 + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2 + 2\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)\right]$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right] + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2$$

5

The bias-variance decomposition

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\text{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\text{bias}(\mathbf{x})}$$

Finally, we get:

$$\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]$$

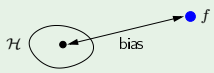
$$= \mathbb{E}_{\mathbf{x}}[\text{bias}(\mathbf{x}) + \text{var}(\mathbf{x})]$$

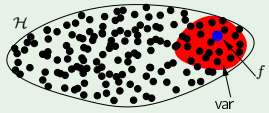
$$= \text{bias} + \text{var}$$

6

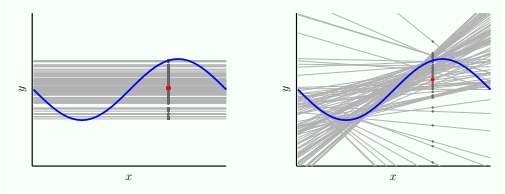
The tradeoff between bias and variance

$$\text{bias} = \mathbb{E}_{\mathbf{x}}\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] \qquad \text{var} = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]\right]$$





7



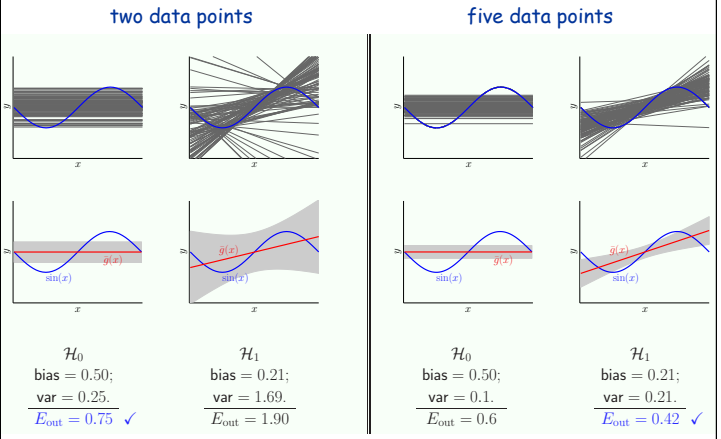
The bias-variance decomposition

In learning there is a tradeoff:

- How well can learning approximate the target function
- How close can we get to that approximation with a finite dataset.

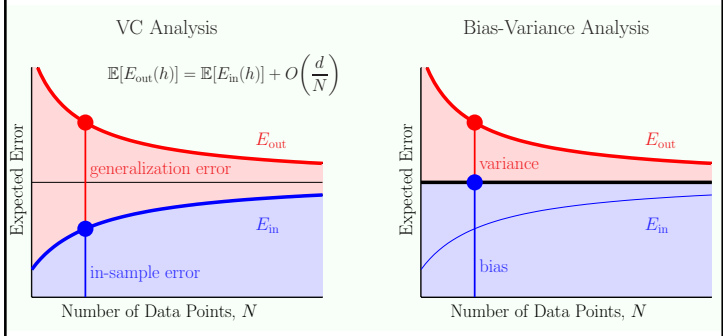
9

Match model complexity to the amount of data not the complexity of the target function



10

Two views of out-of-sample error



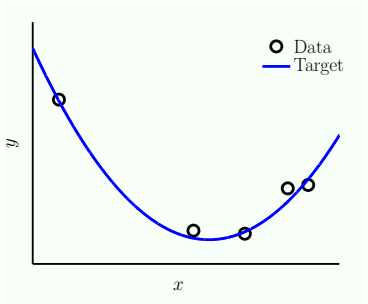
The choice of hypothesis needs to strike a balance between approximating f on the training data and generalizing on new data.

Pick a hypothesis that can fit the data (low bias) and not behave wildly (low variance)

11

What is overfitting

Assume a quadratic target function and a sample of 5 noisy data points:

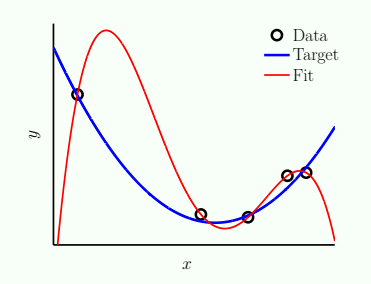


Chapter 4

12

What is overfitting

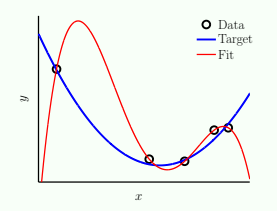
Let's fit this data with a degree 4 polynomial:



13

What is overfitting

Let's fit this data with a degree 4 polynomial:



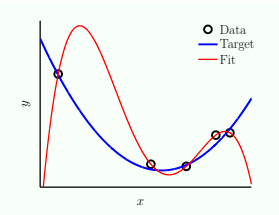
Overfitting: fitting the data more than is warranted.

E_{in} is small, and yet E_{out} is large

14

What is overfitting

Let's fit this data with a degree 4 polynomial:

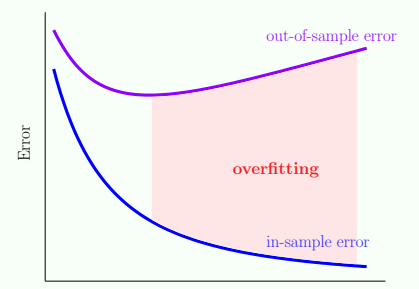


Observations:
We are overfitting the data: $E_{in} = 0$, E_{out} large

The noise did us in!

15

What is overfitting

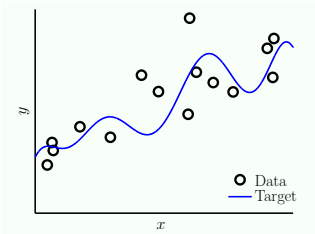


Overfitting: fitting the data more than is warranted.
In other words – using a model that is more complex than is necessary.

16

What is overfitting

Let's look at another example:

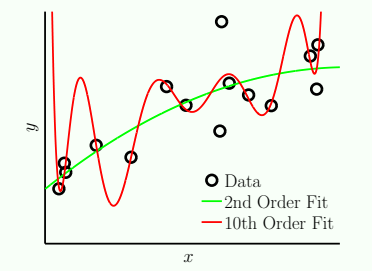


10th order f with noise.

17

What is overfitting

Let's compare fitting the data with 2nd degree and 10th degree polynomials:

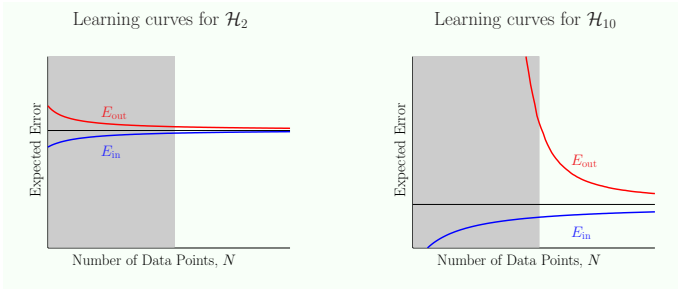


Although the data is generated with a 10th degree polynomial, the quadratic fit is better!

18

Which hypothesis?

The choice of hypothesis space depends on the number of available data points:



- High complexity hypothesis set: better chance of approximating the target function
- Low complexity hypothesis set: better chance of getting low out-of-sample error

19

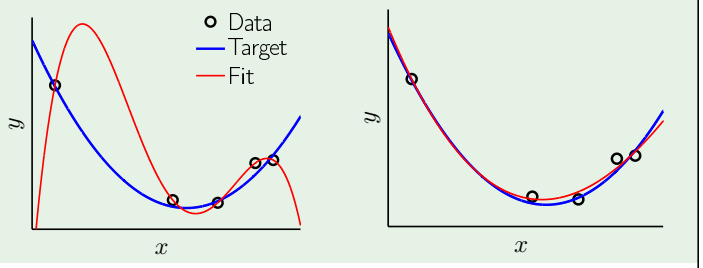
Factors that lead to overfitting

- Small number of data points
- Amount of noise
- Complexity of the target function
- Complexity of the hypothesis set

20

Regularization

The cure for overfitting - regularization



Without regularization

With regularization

21