
Approximation vs Generalization

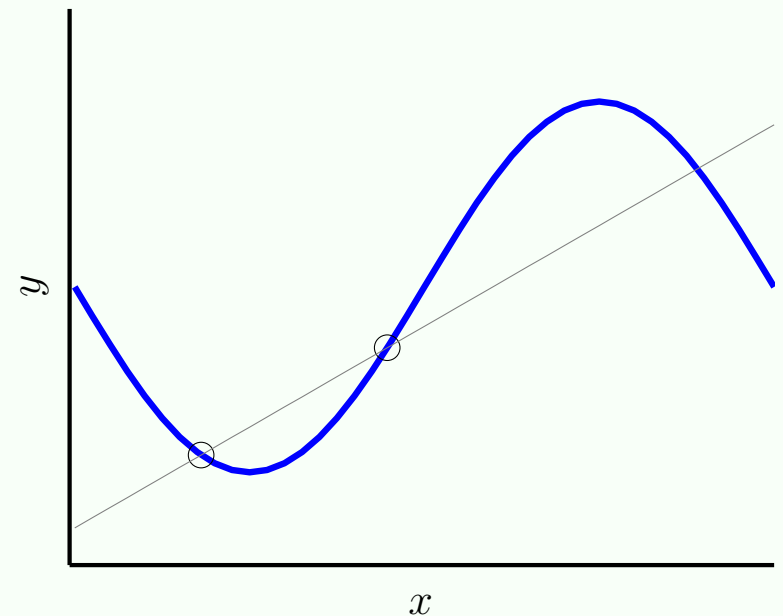
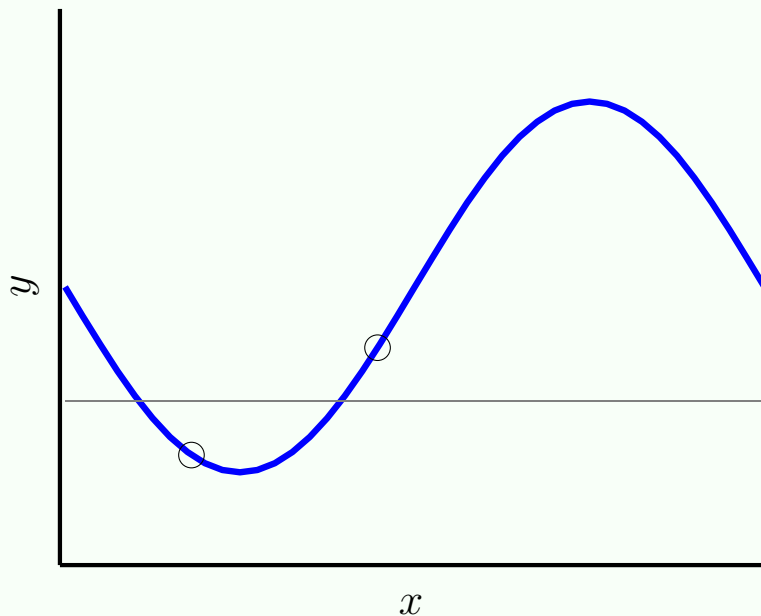
Chapter 2.3, 4

The bias-variance decomposition

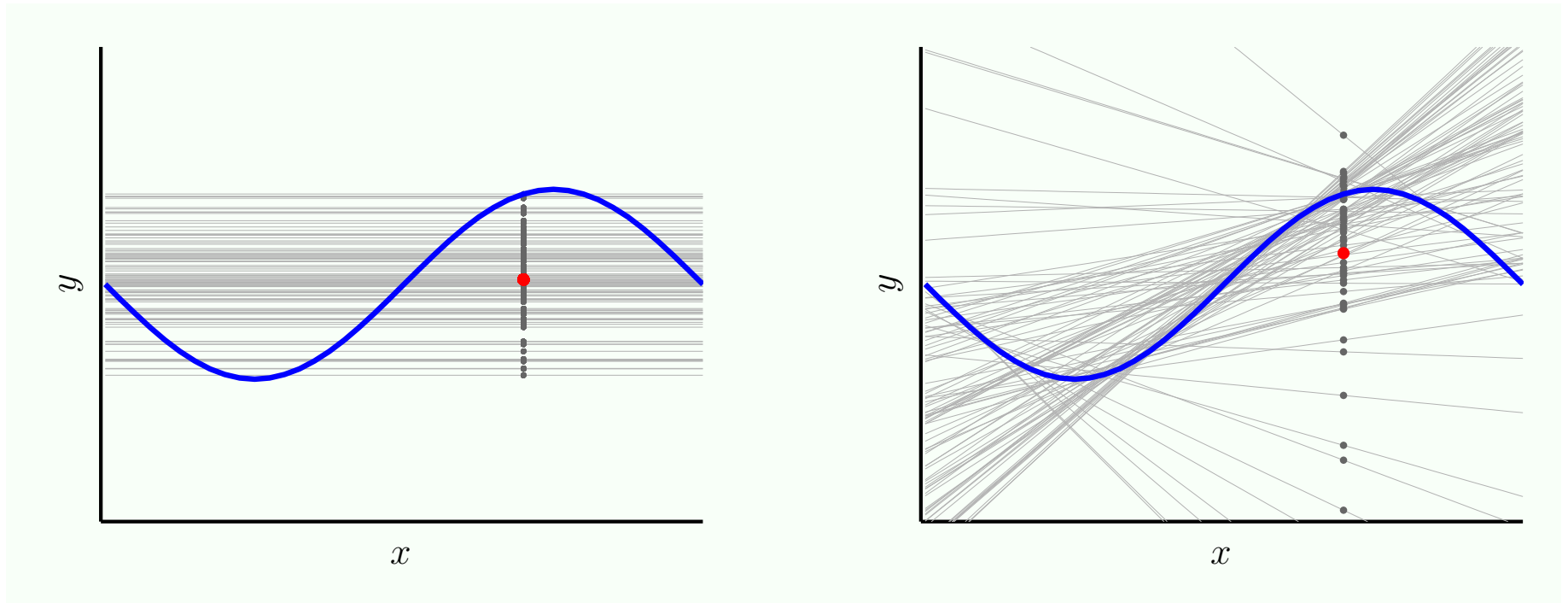
Consider a simple learning problem: two data points and two hypothesis sets.

$$\mathcal{H}_0 : h(x) = b$$

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



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} .

The bias-variance decomposition

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

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

To abstract away the dependence on a given dataset:

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

And let's focus on

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

To evaluate

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

We consider the "average hypothesis"

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

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

The bias-variance decomposition

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

Finally, we get:

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

The bias-variance decomposition

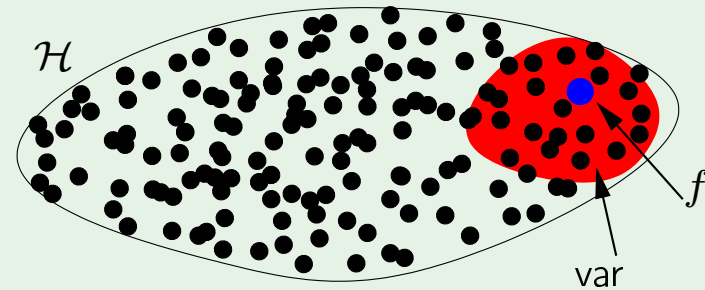
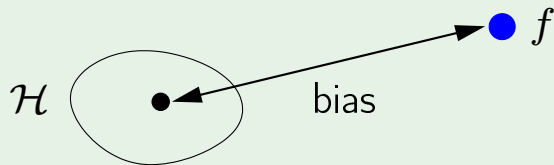
In learning there is a tradeoff:

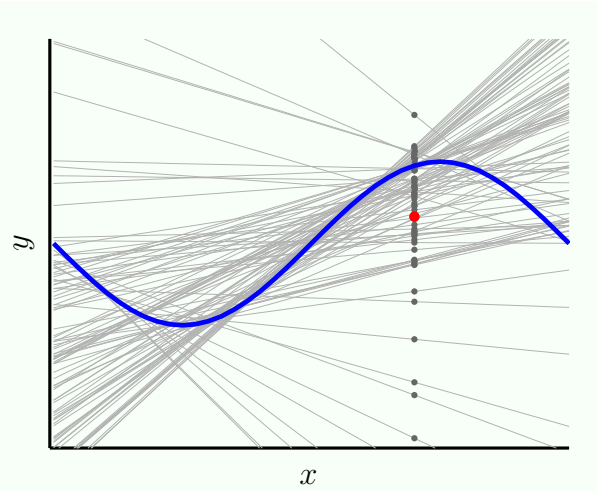
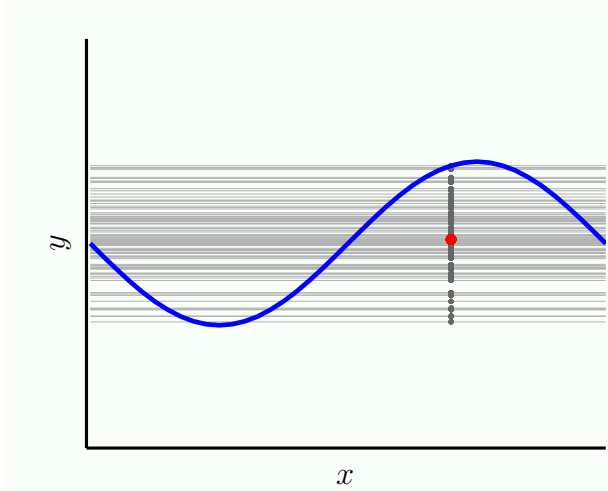
- ✧ How well can learning approximate the target function
- ✧ How close can we get to that approximation with a finite dataset.

The tradeoff

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

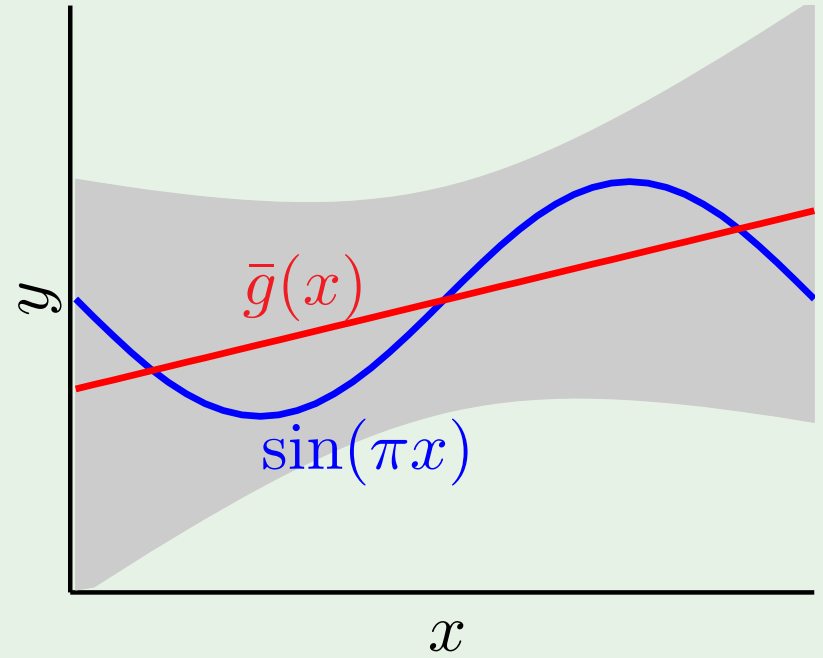
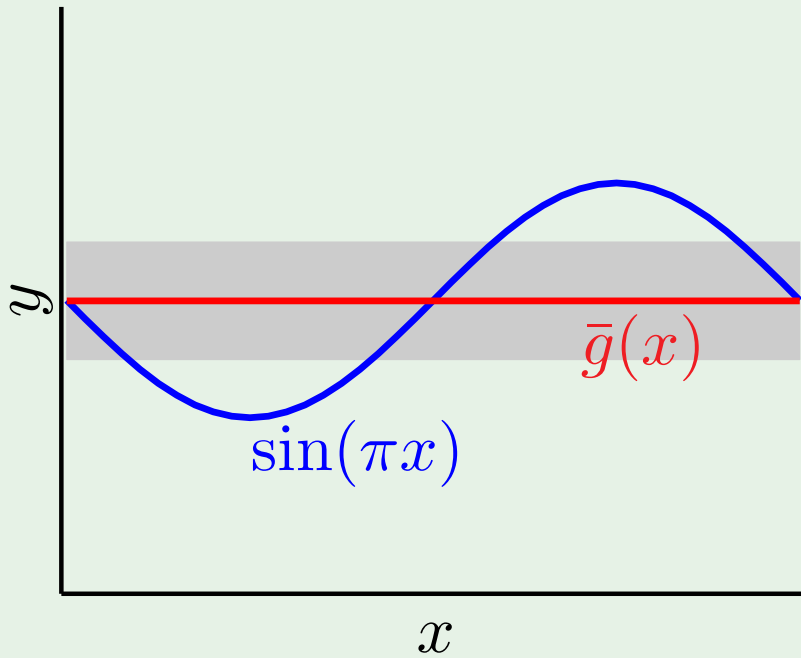
$$\text{var} = \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{\mathcal{D}} \left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2 \right] \right]$$





\mathcal{H}_0

\mathcal{H}_1



bias = **0.50**

var = **0.25**

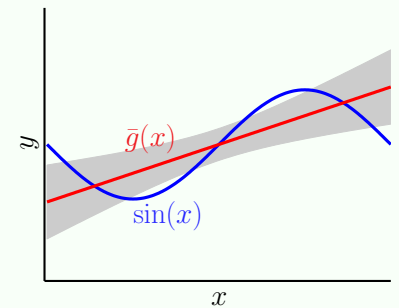
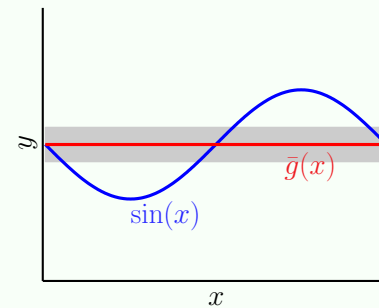
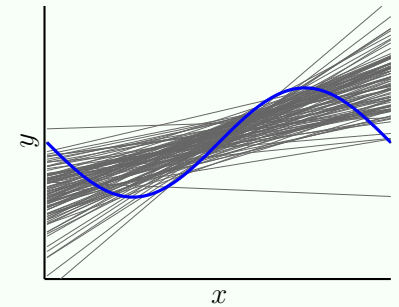
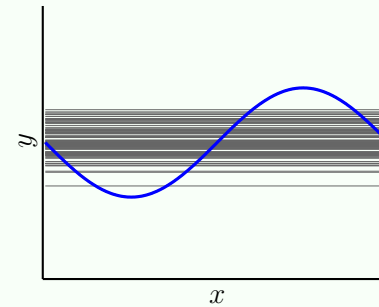
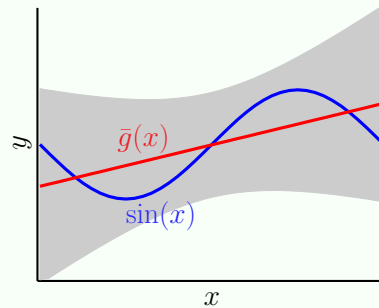
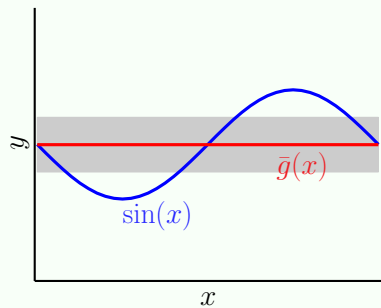
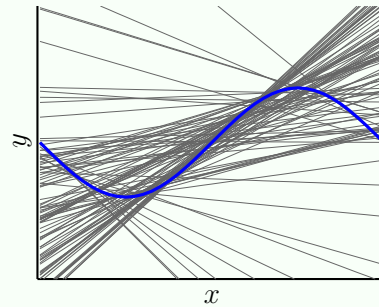
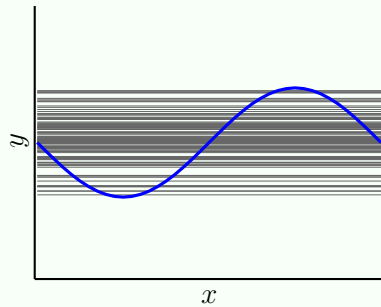
bias = **0.21**

var = **1.69**

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

two data points

five data points



\mathcal{H}_0

bias = 0.50;
var = 0.25.

$E_{\text{out}} = 0.75$ ✓

\mathcal{H}_1

bias = 0.21;
var = 1.69.

$E_{\text{out}} = 1.90$

\mathcal{H}_0

bias = 0.50;
var = 0.1.

$E_{\text{out}} = 0.6$

\mathcal{H}_1

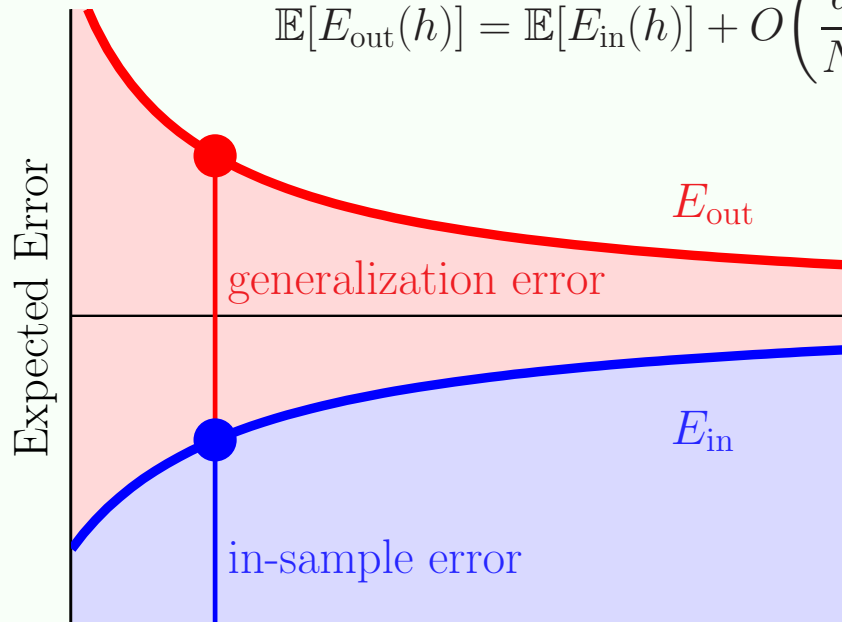
bias = 0.21;
var = 0.21.

$E_{\text{out}} = 0.42$ ✓

Two views of out-of-sample error

VC Analysis

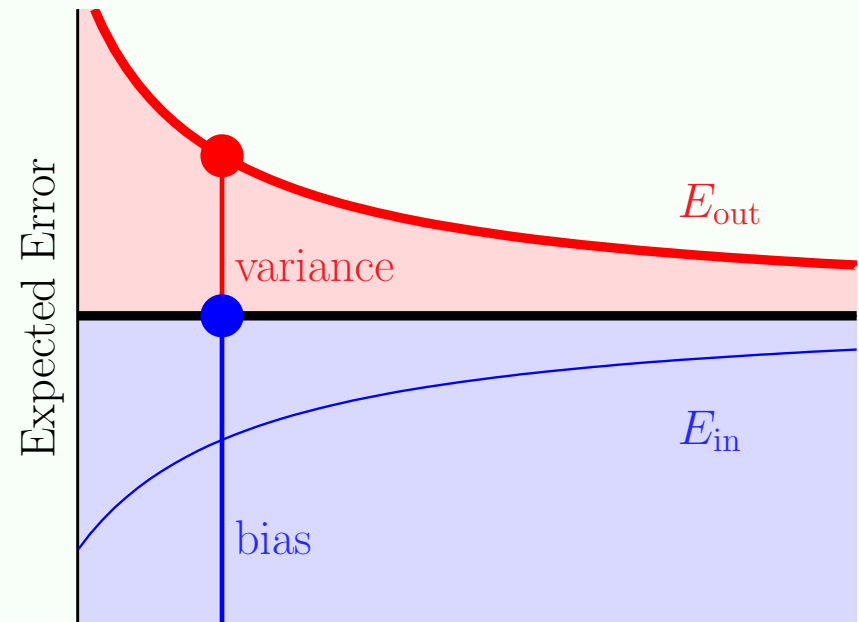
$$\mathbb{E}[E_{\text{out}}(h)] = \mathbb{E}[E_{\text{in}}(h)] + O\left(\frac{d}{N}\right)$$



Number of Data Points, N

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

Bias-Variance Analysis

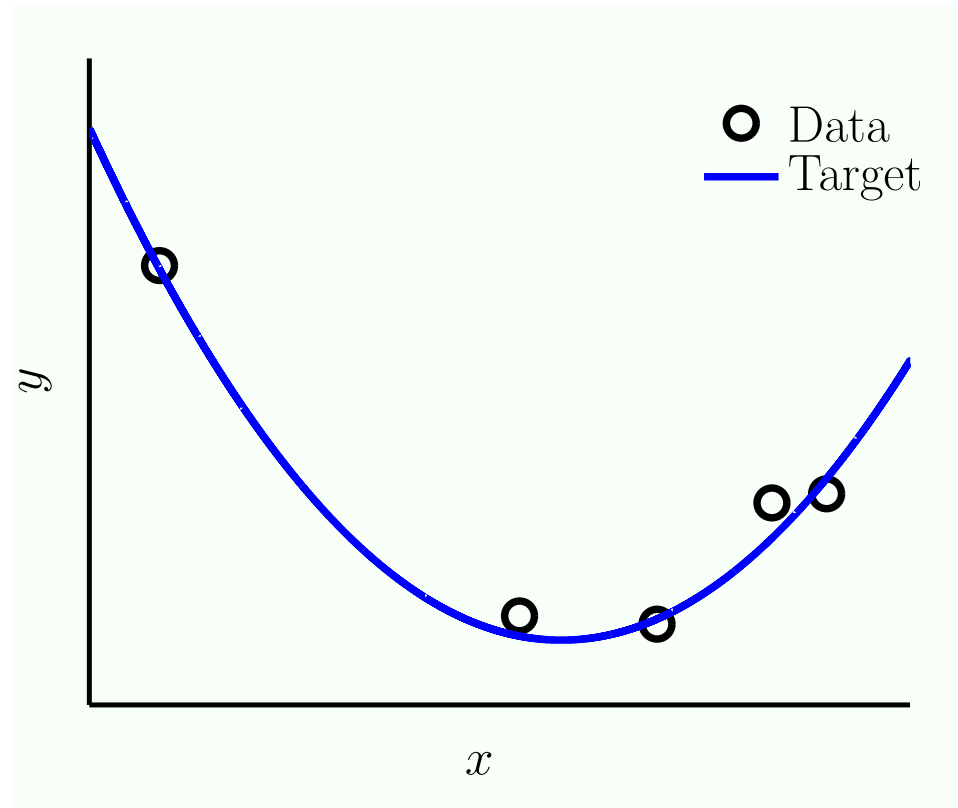


Number of Data Points, N

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

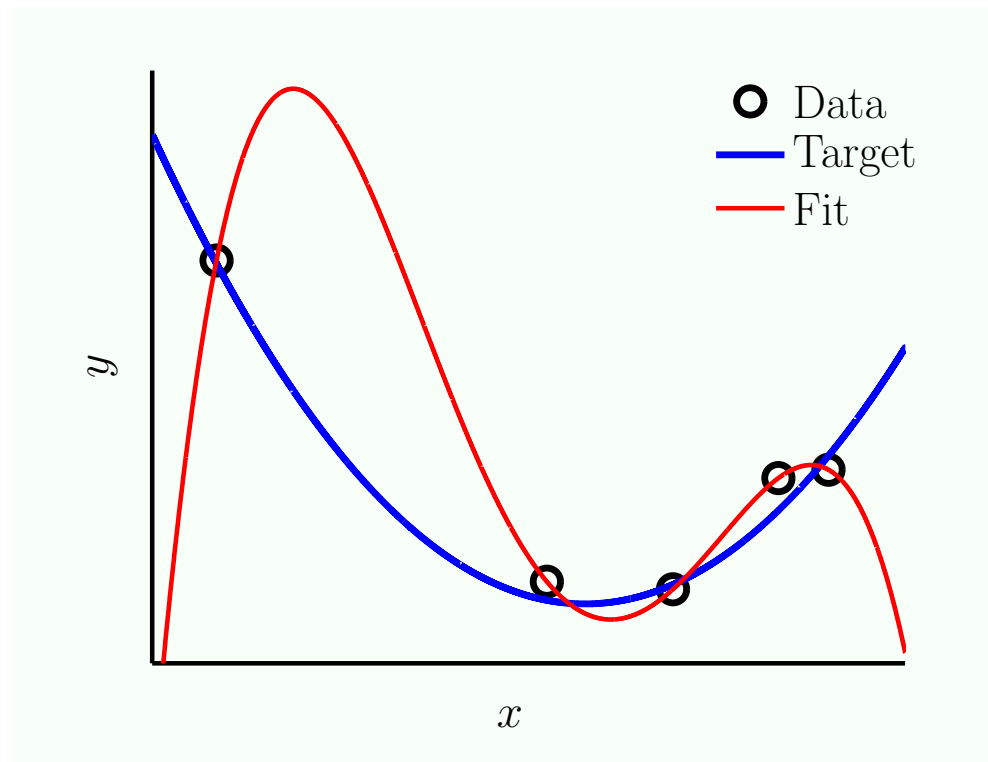
What is overfitting

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



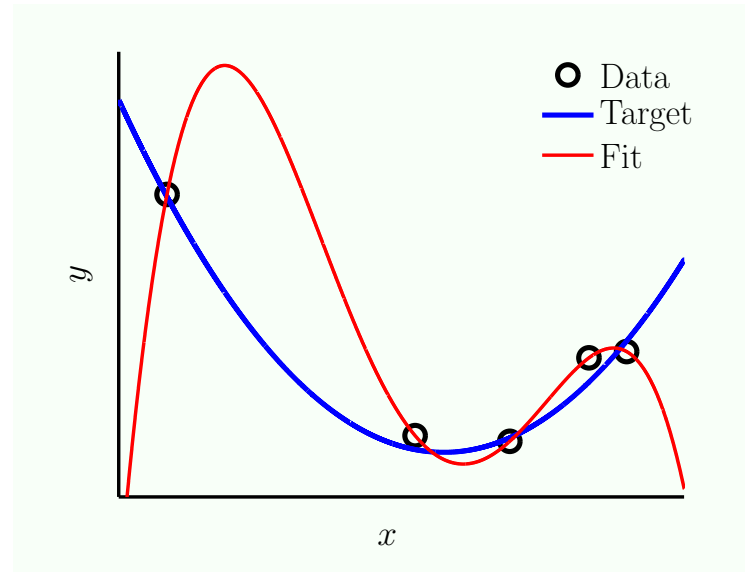
What is overfitting

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



What is overfitting

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



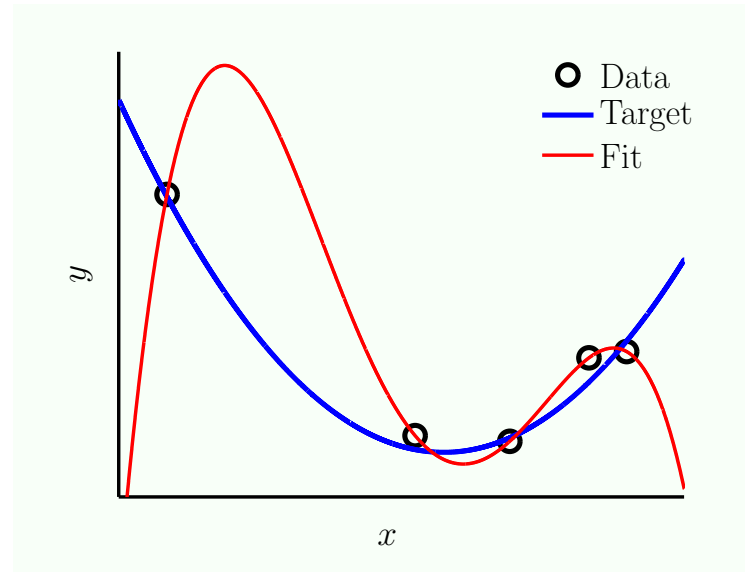
Observations:

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

The noise did us in!

What is overfitting

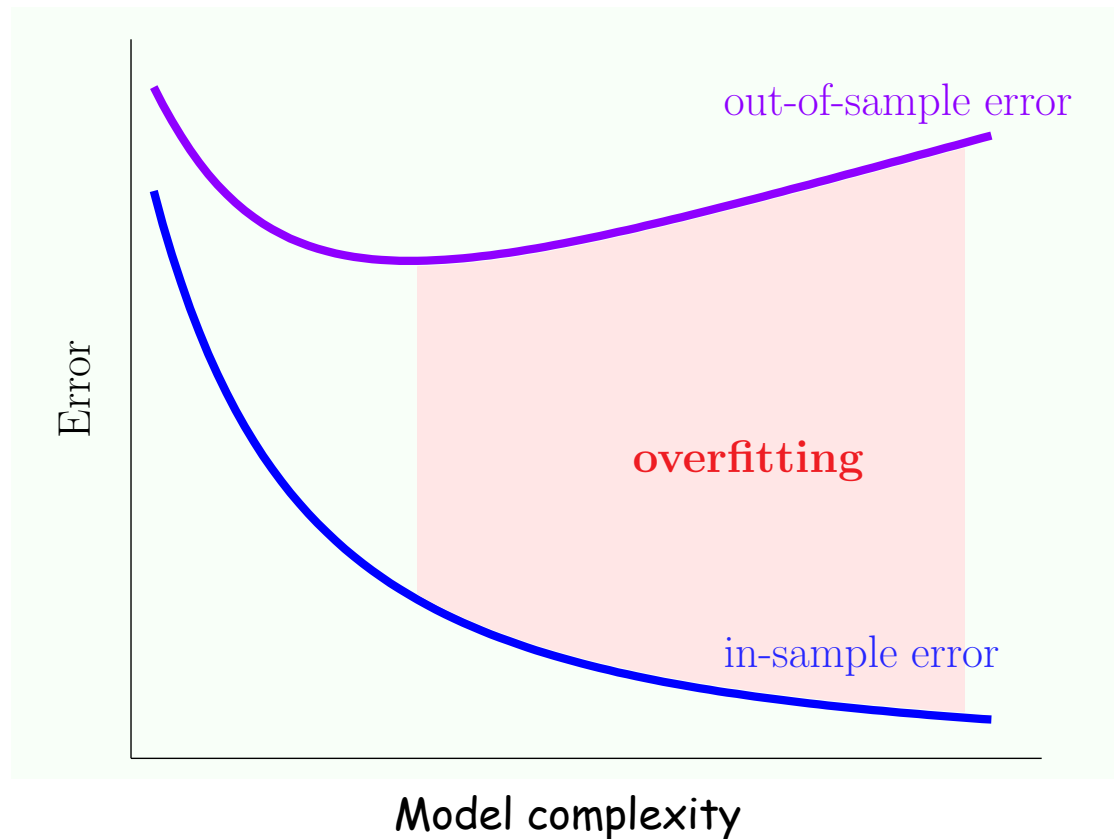
Let's fit this data with a degree 4 polynomial:



Overfitting: fitting the data more than is warranted.

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

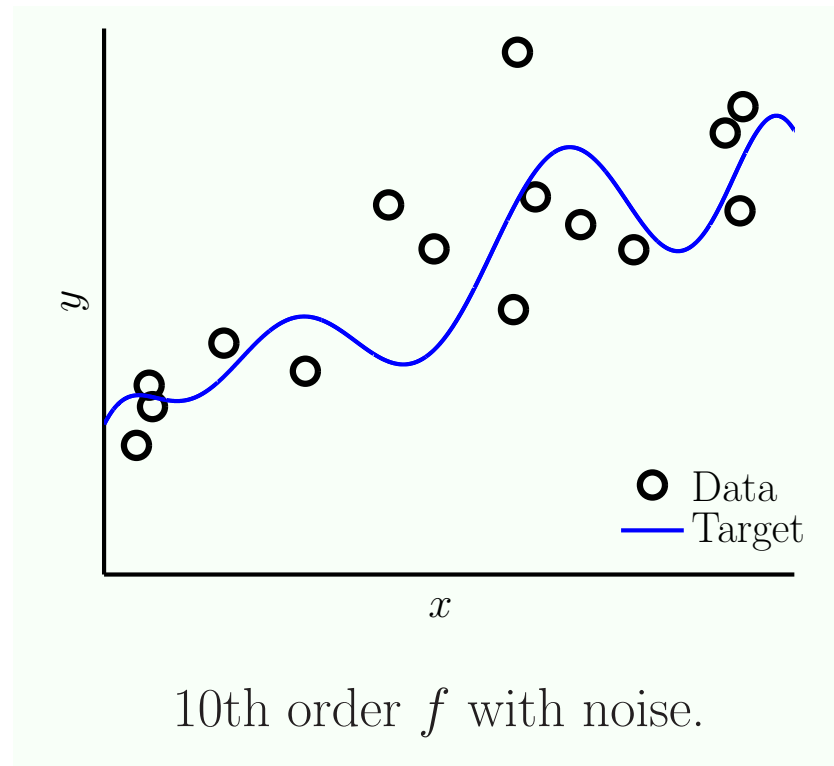
What is overfitting



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

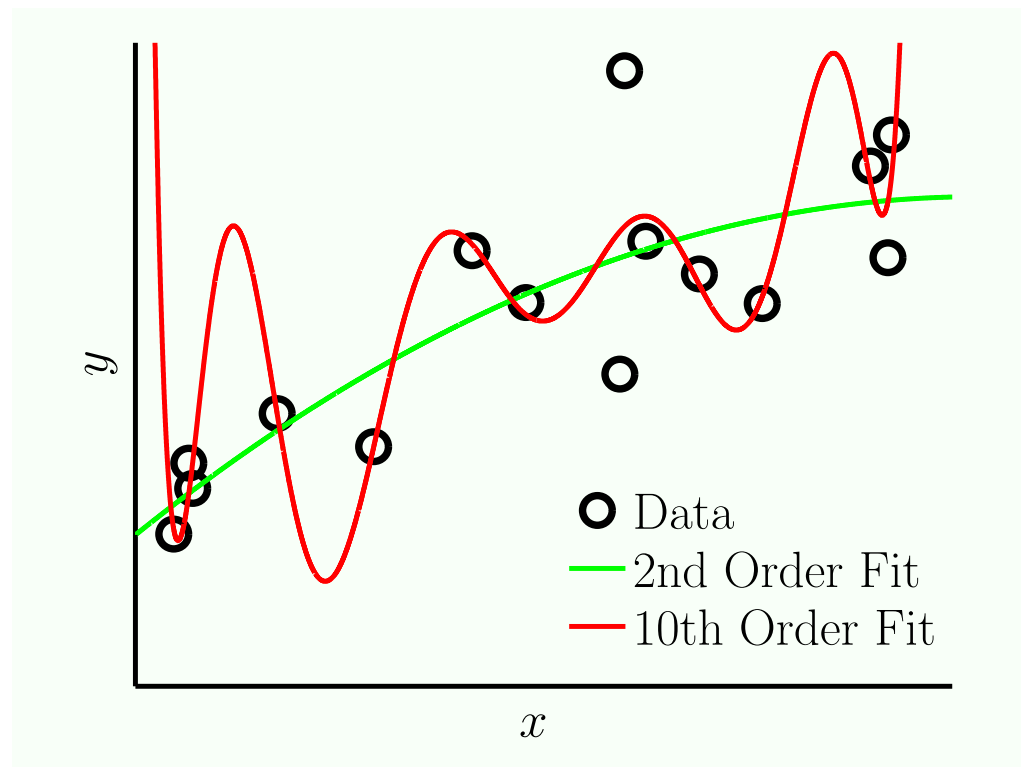
What is overfitting

Let's look at another example:



What is overfitting

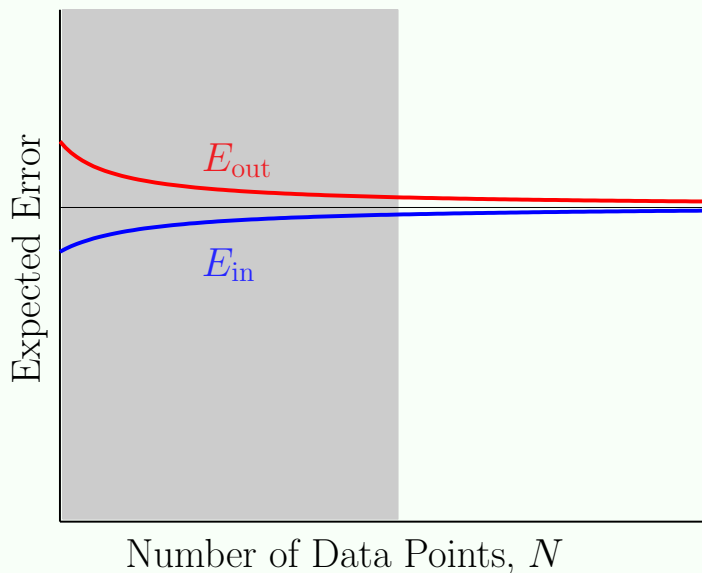
Let's compare fitting the data with 2nd degree and 10th degree polynomials:



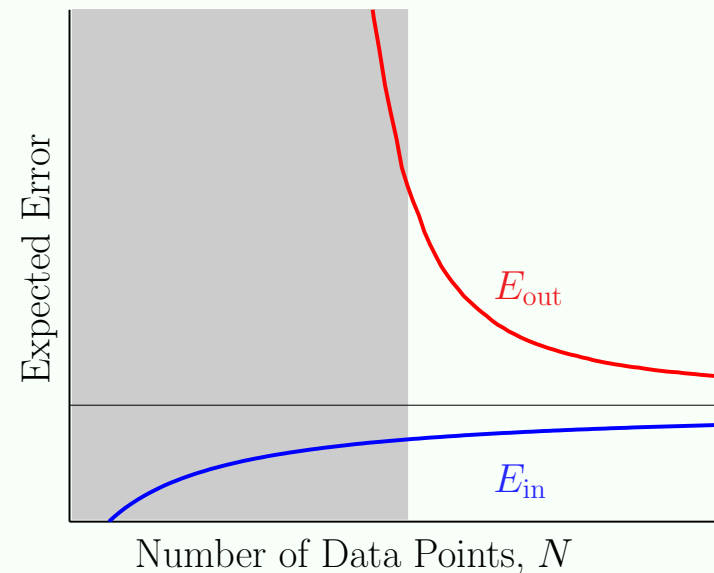
Which hypothesis?

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

Learning curves for \mathcal{H}_2



Learning curves for \mathcal{H}_{10}



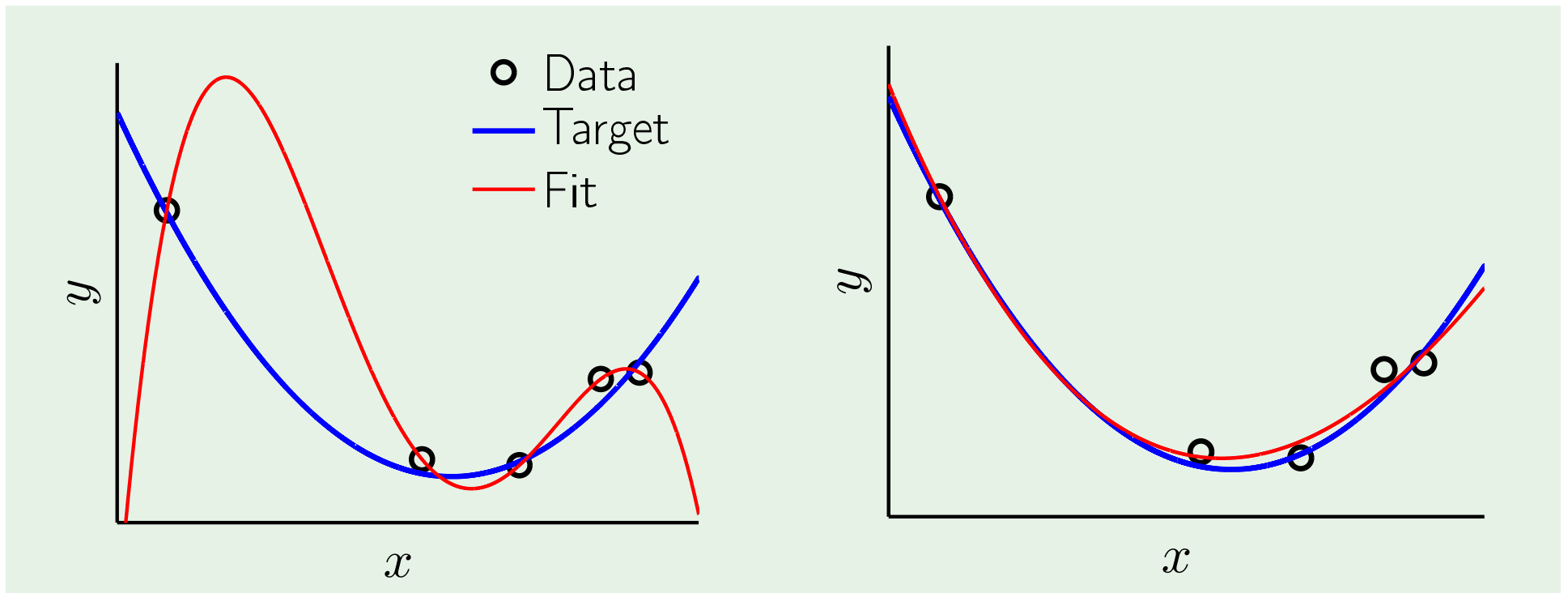
- ❖ High complexity hypothesis set: better chance of approximating the target function
- ❖ Low complexity hypothesis set: better chance of getting low out-of-sample error

Factors that lead to overfitting

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

Regularization

The cure for overfitting - regularization



Without regularization

With regularization