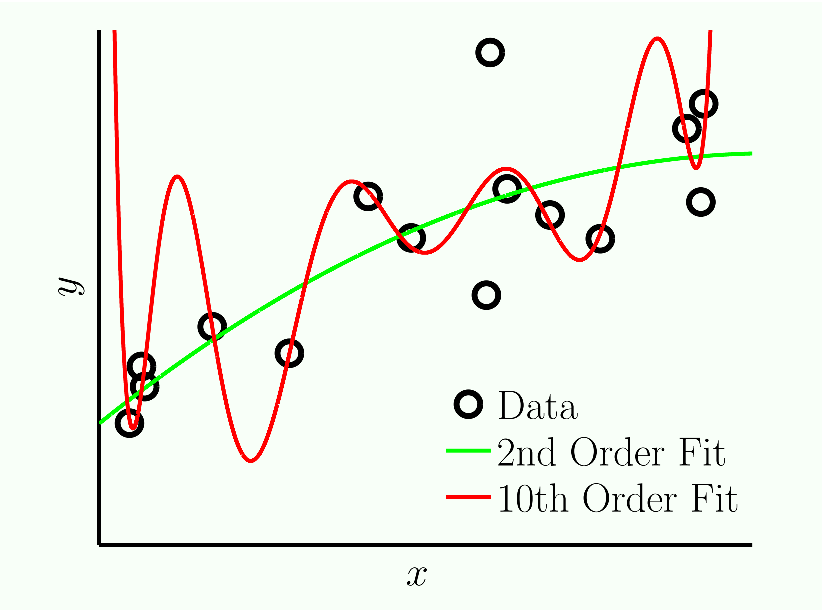
# Approximation vs Generalization

## LFD Sections 2.3, 4.1



# Assignment 2 FAQ

Does an update of the alphas/weight vector of the adatron occur regardless of whether an example is misclassified?

- Yes!

That brings up another question: when do we stop?

- After a fixed number of iterations (use the same bound you use for the perceptron).

The alpha coefficients of the adatron explode. What should I do?

- Put an upper bound on the magnitude of the alphas

What's a good value for the learning rate?

- That requires some experimentation.

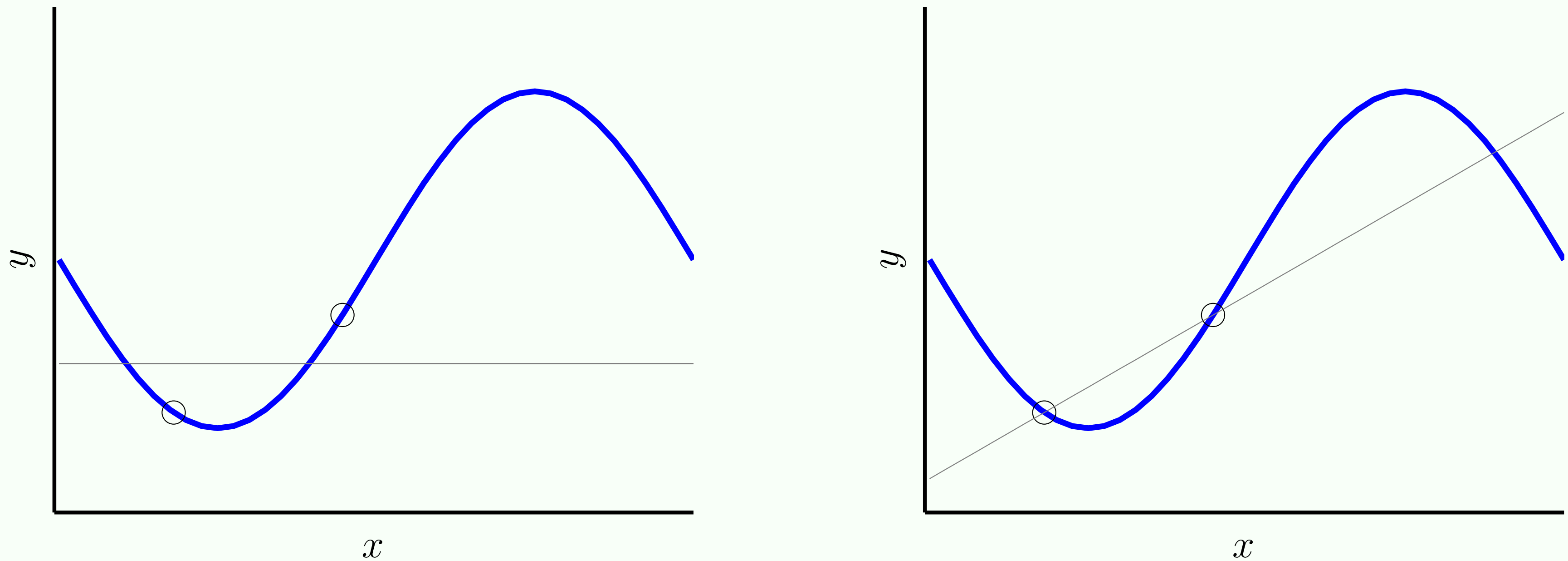
The adatron takes a long time to run

- The instructor suggests a speedup where the weight vector is not computed from scratch after each update.

# The bias-variance decomposition

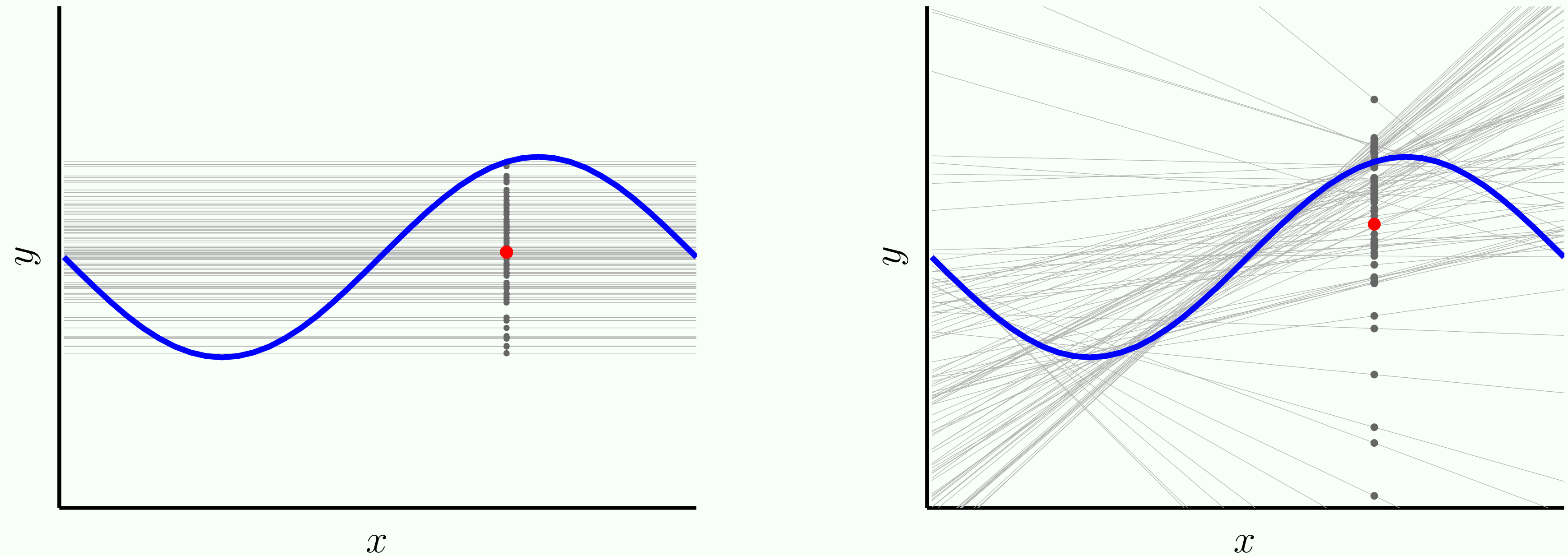
Consider a simple learning problem: two data points and two hypothesis sets.

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




Section 2.3

# 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:

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

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

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

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

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

# The bias-variance decomposition

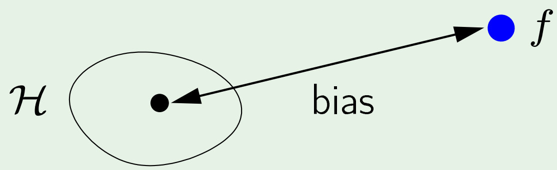
$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\mathsf{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\mathsf{bias}(\mathbf{x})}$$

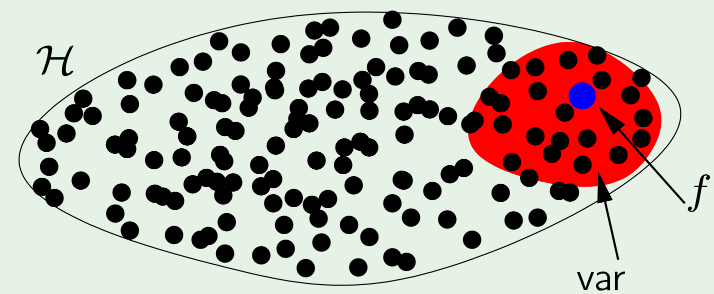
Finally, we get:

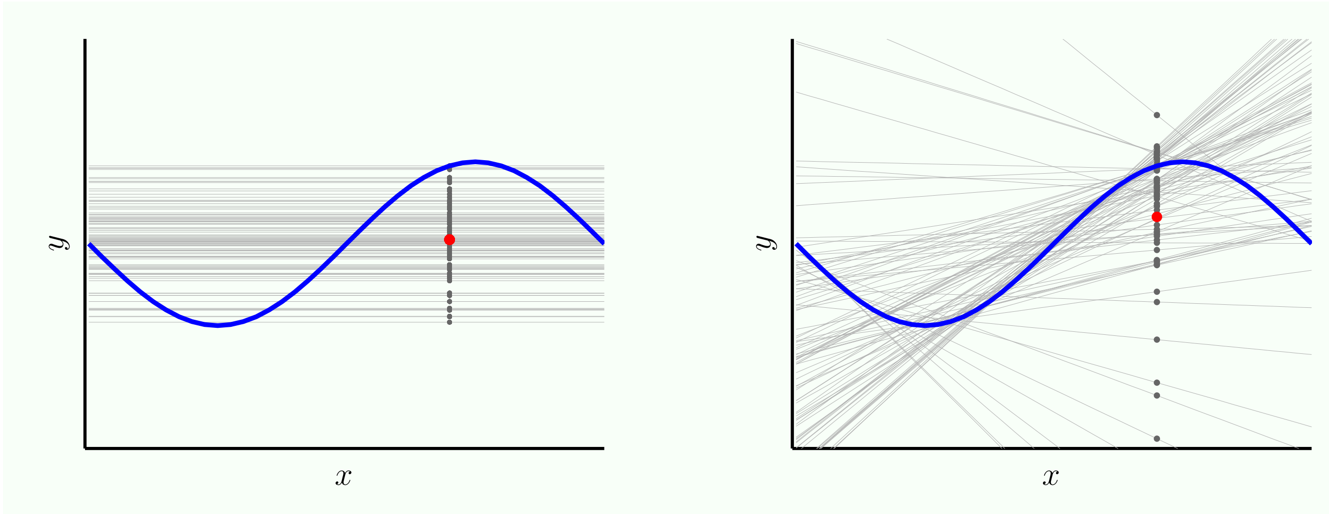
$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \\ &= \mathbb{E}_{\mathbf{x}}[\mathsf{bias}(\mathbf{x}) + \mathsf{var}(\mathbf{x})]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

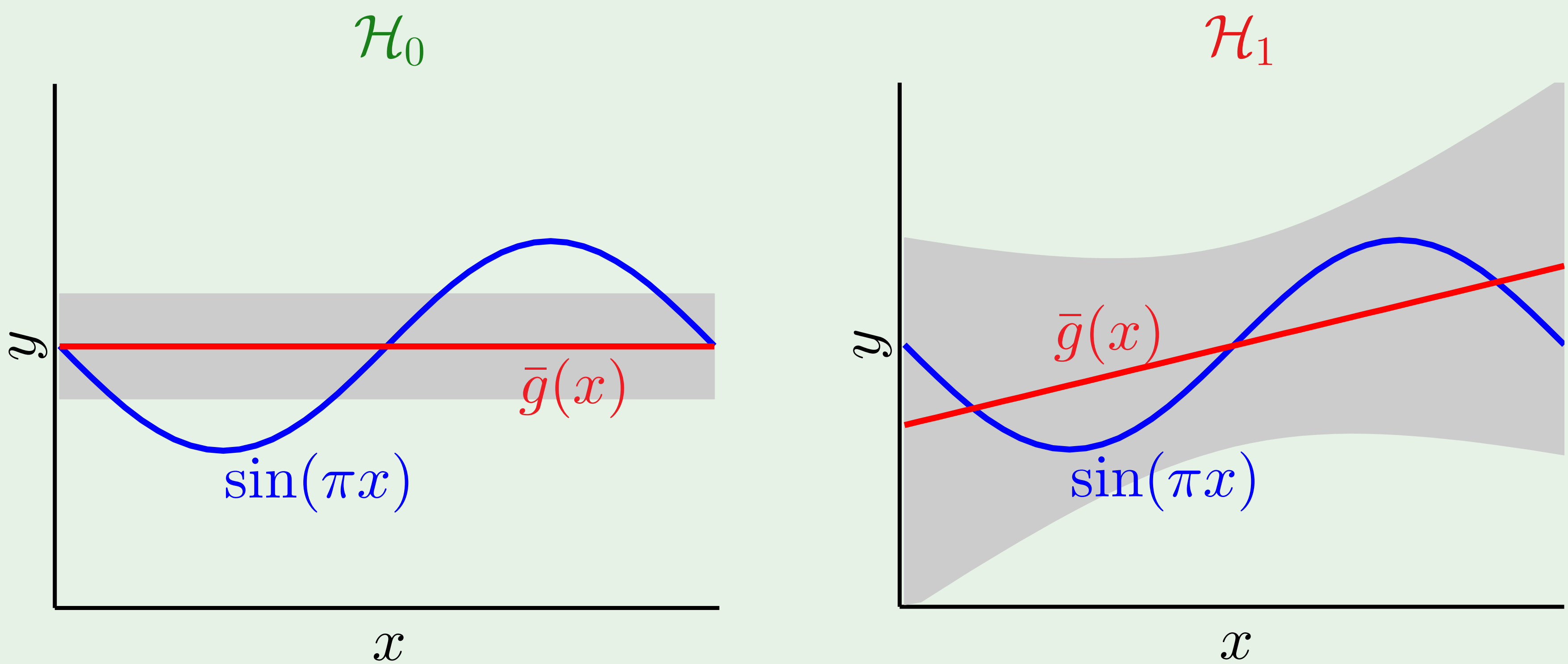
# The tradeoff between bias and variance

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






bias $= \mathbf{0.50}$ var $= \mathbf{0.25}$

bias $= \mathbf{0.21}$ var $= \mathbf{1.69}$

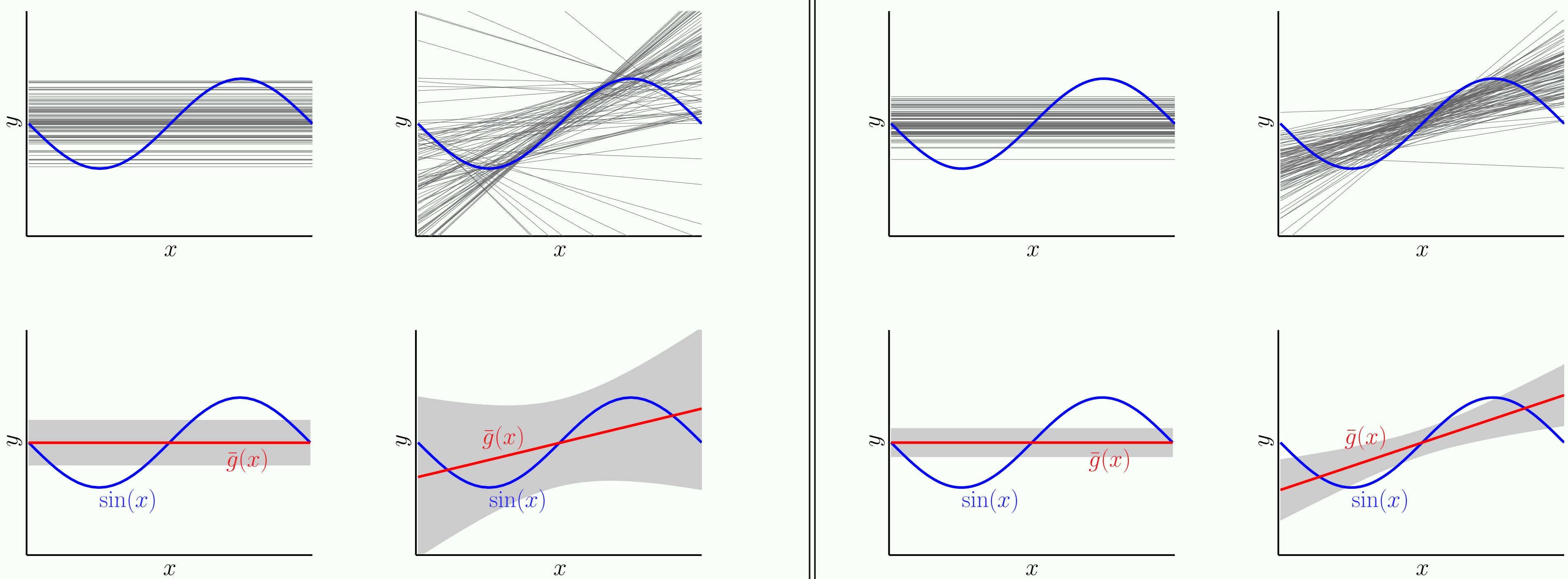
# The bias-variance decomposition

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.

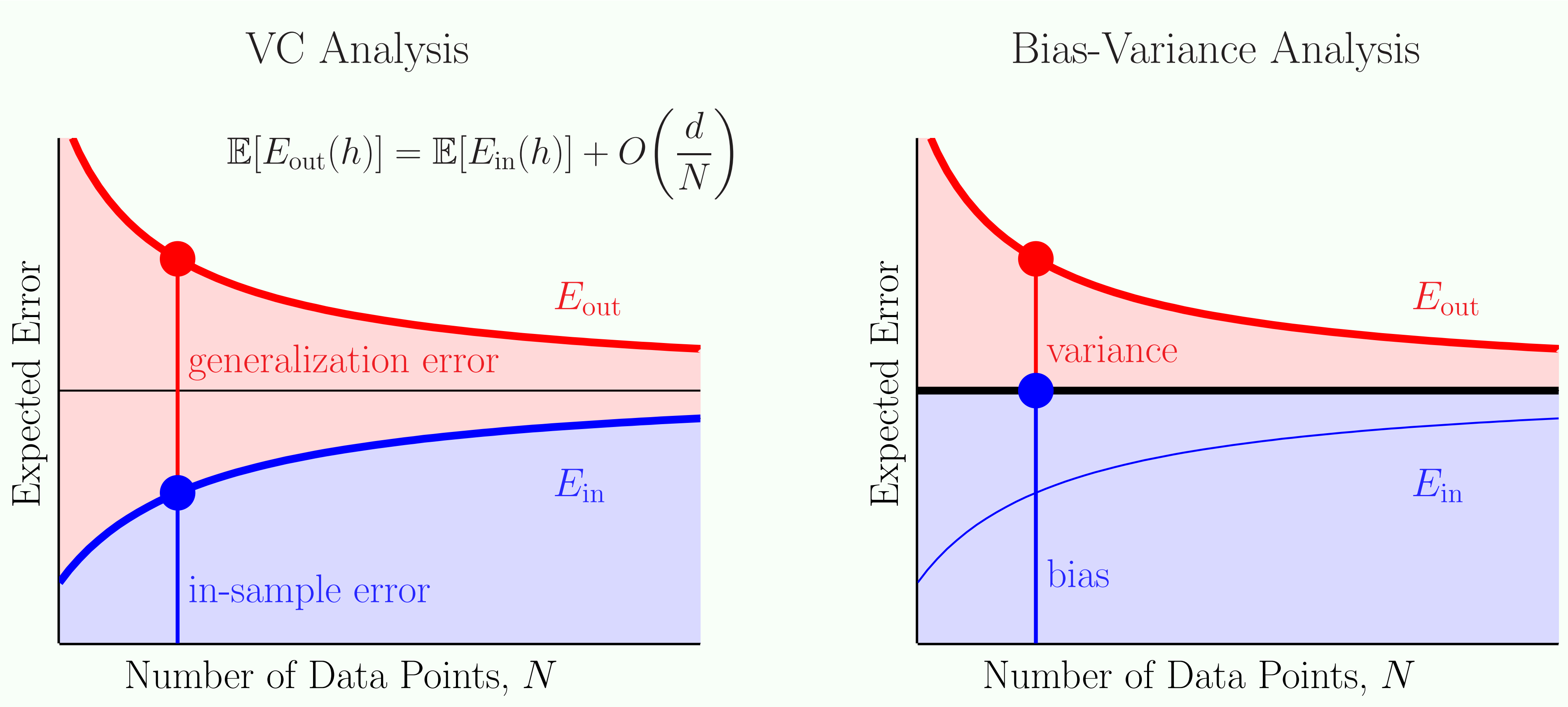
# Match model complexity to the amount of data not the complexity of the target function

## two data points / five data points




| two data points: $\mathcal{H}_0$ | two data points: $\mathcal{H}_1$ | five data points: $\mathcal{H}_0$ | five data points: $\mathcal{H}_1$ |
|---|---|---|---|
| bias $= 0.50$; | bias $= 0.21$; | bias $= 0.50$; | bias $= 0.21$; |
| var $= 0.25$. | var $= 1.69$. | var $= 0.1$. | var $= 0.21$. |
| $E_{\text{out}} = 0.75$ ✓ | $E_{\text{out}} = 1.90$ | $E_{\text{out}} = 0.6$ | $E_{\text{out}} = 0.42$ ✓ |

# Two views of out-of-sample error

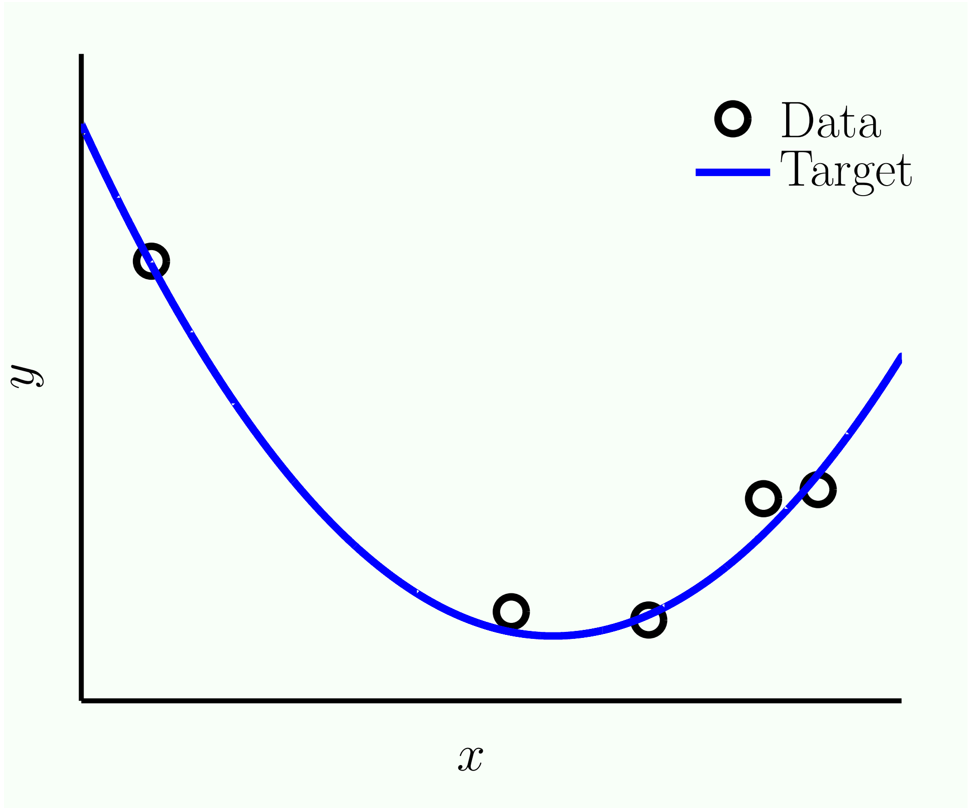



The choice of hypothesis needs to strike a balance between approximating f on the training data and generalizing on new data.

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

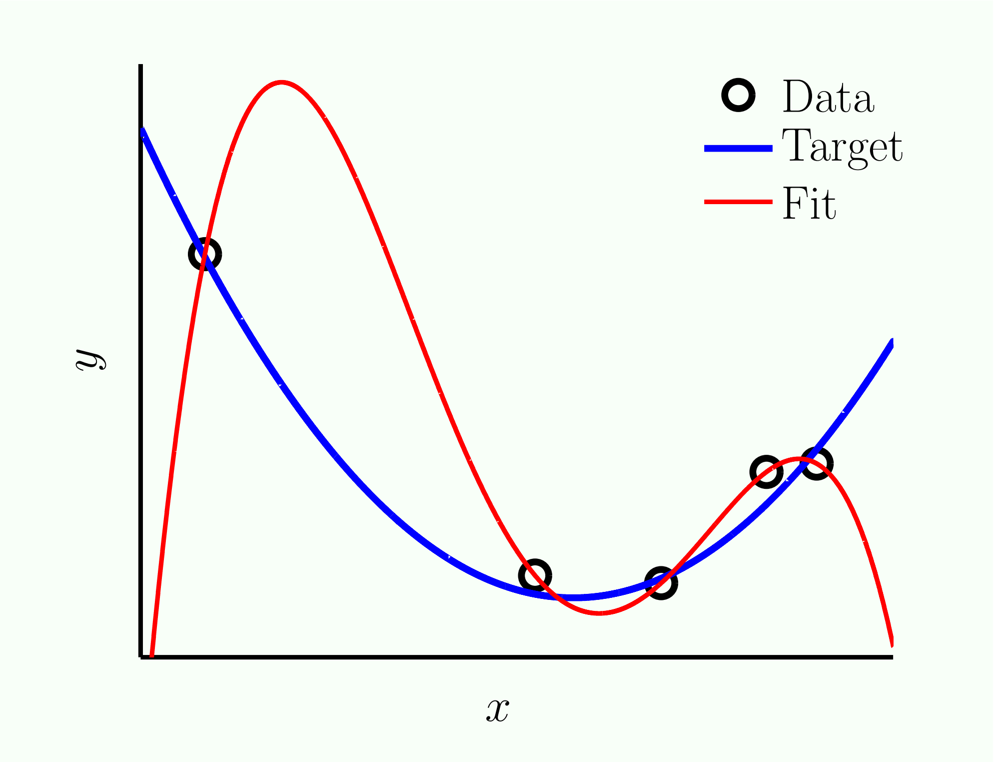
# What is overfitting

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



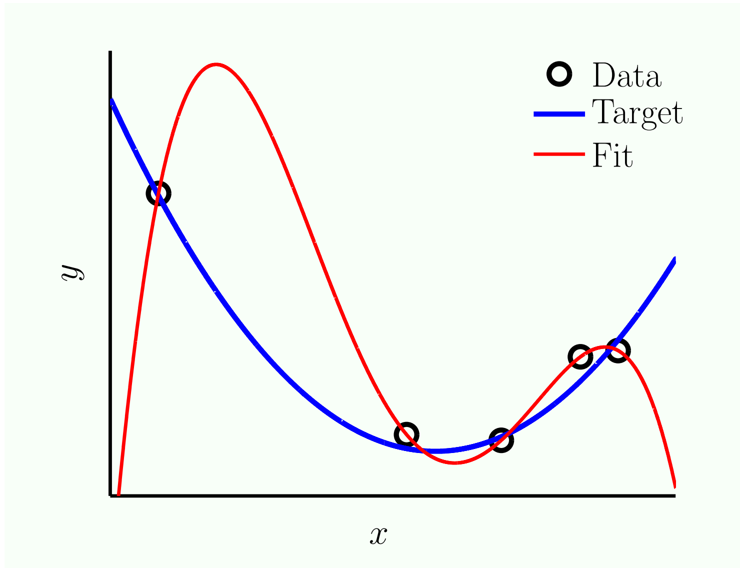
# What is overfitting

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



# What is overfitting

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

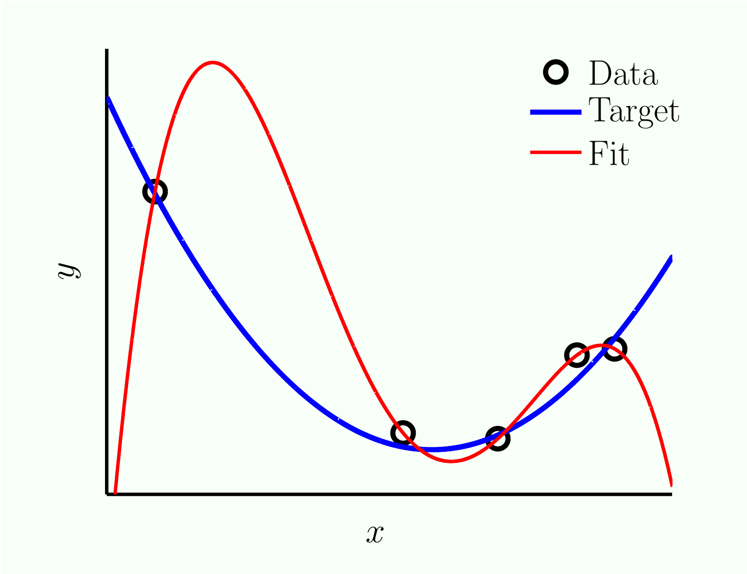



Overfitting: fitting the data more than is warranted.

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

# What is overfitting

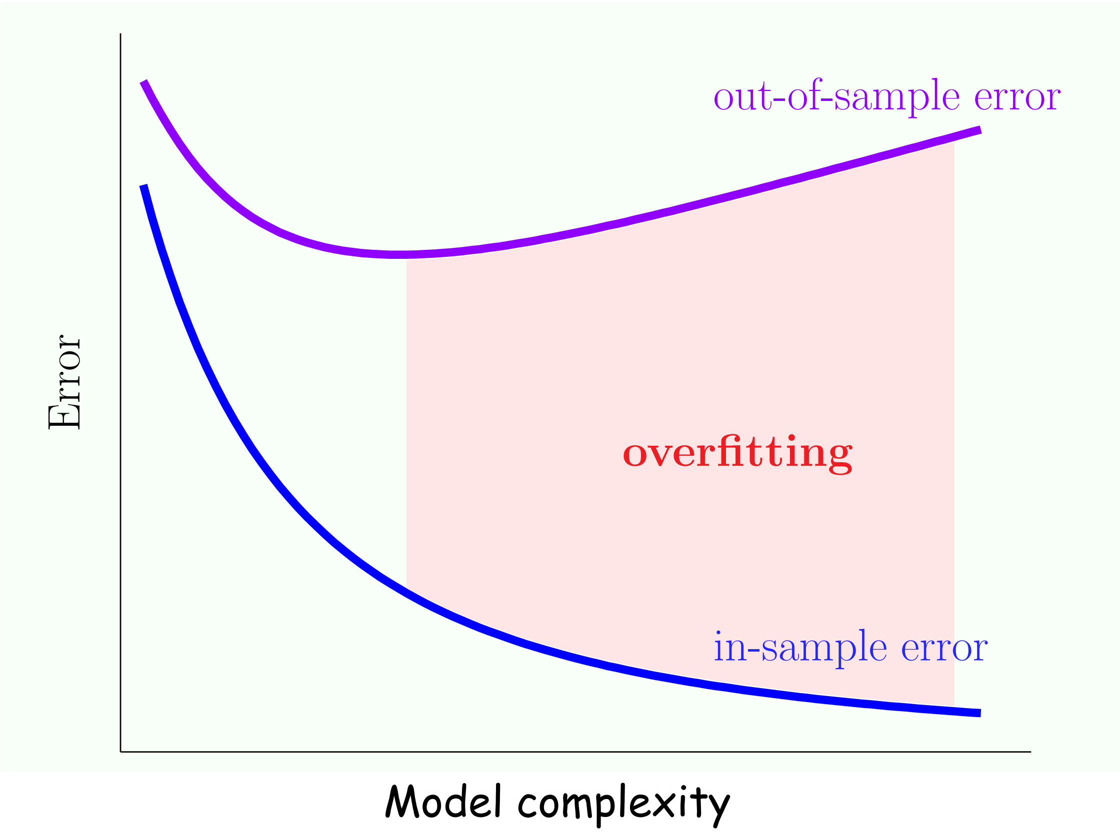
Let's fit this data with a degree 4 polynomial:




Observations:

- ✓ We are overfitting the data: $E_{in} = 0$, $E_{out}$ large
- ✓ The noise did us in!

# What is overfitting

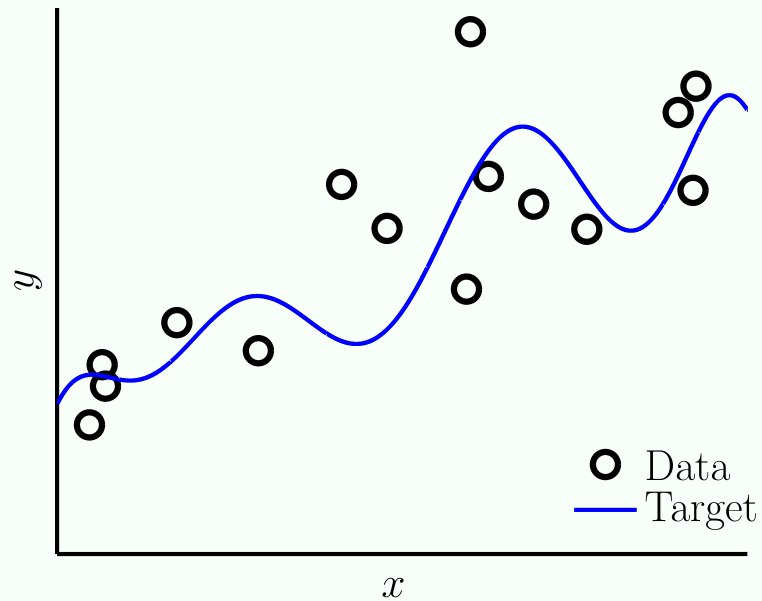



Overfitting: fitting the data more than is warranted.

In other words – using a model that is more complex than is necessary.

# What is overfitting

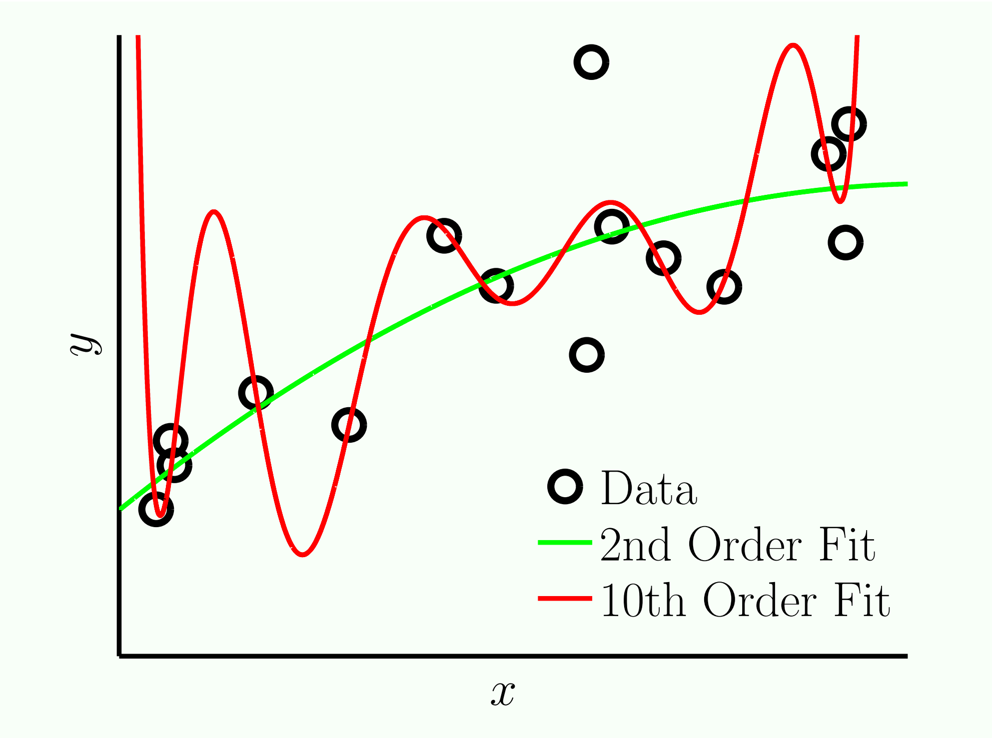
Let's look at another example:




10th order $f$ with noise.

# What is overfitting

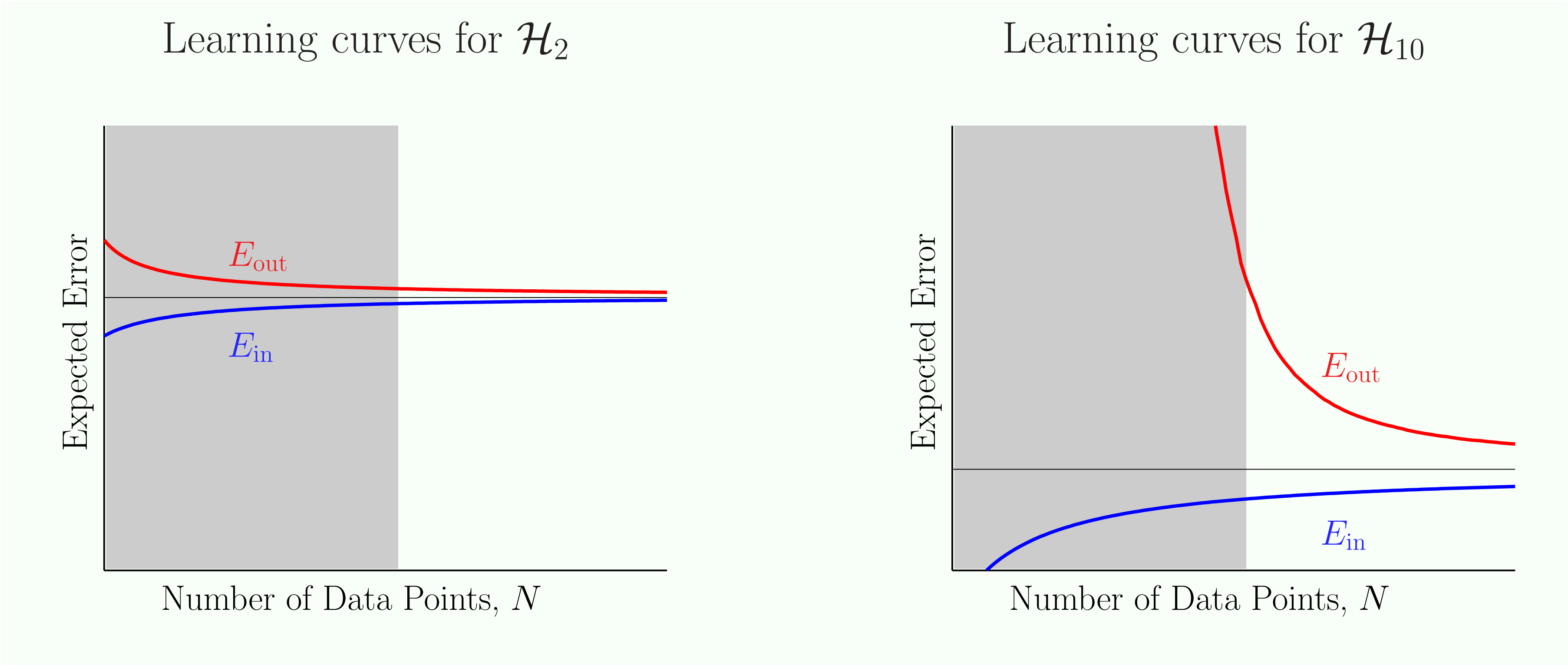
Let's compare fitting the data with 2nd degree and 10th degree polynomials:




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

# Which hypothesis?

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



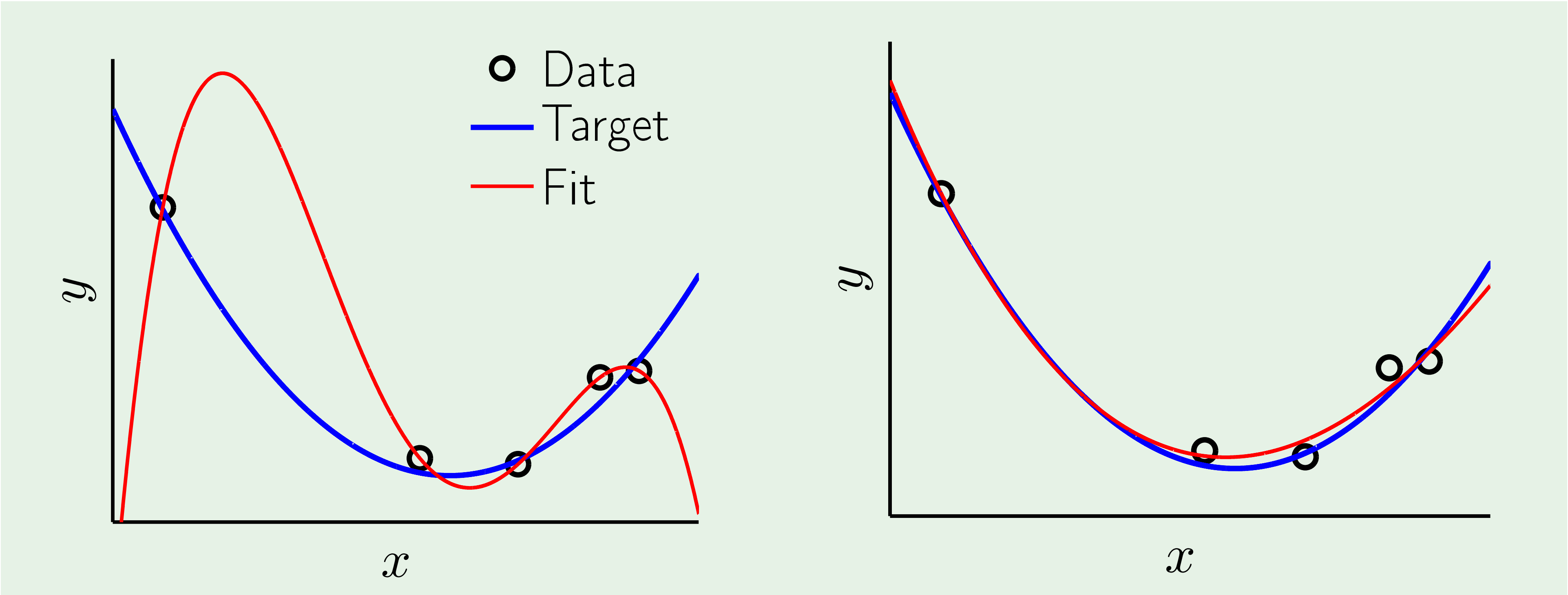

- High complexity hypothesis set: better chance of approximating the target function
- Low complexity hypothesis set: better chance of getting low out-of-sample error

# 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