CSE 417T: Introduction to Machine Learning

Lecture 10: Overfitting

Henry Chai

09/27/18

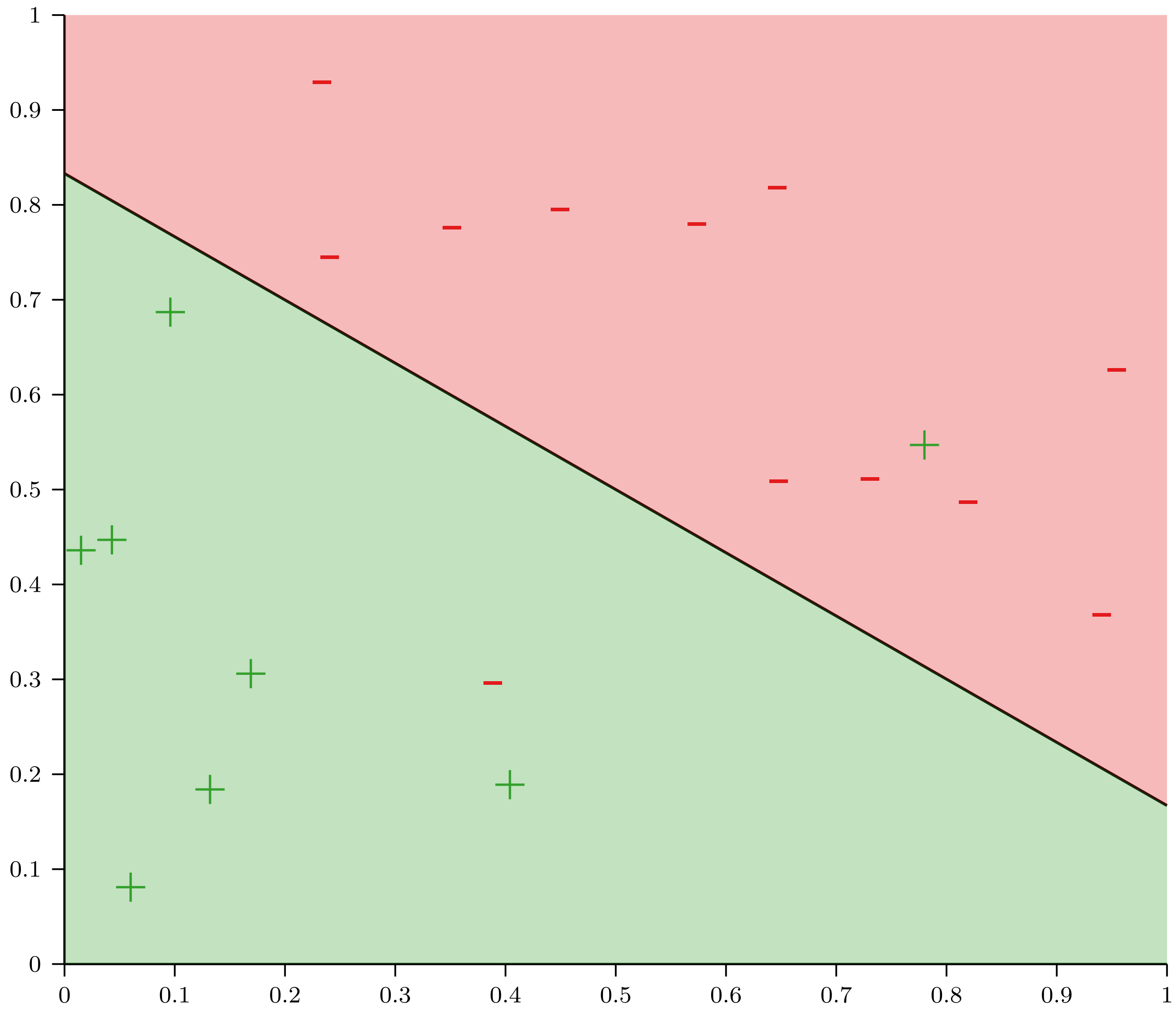
Recall

- Decide on a transformation $\Phi: \mathcal{X} \to \mathcal{Z}$
- Convert $\mathcal{D} = \{(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_n}, y_n)\}$ to $\widetilde{\mathcal{D}} = \{(\Phi(\overrightarrow{x_1}) = \overrightarrow{z_1}, y_1), \ldots, (\Phi(\overrightarrow{x_n}) = \overrightarrow{z_n}, y_n)\}$
- Fit a linear model using $\widetilde{\mathcal{D}}$, $\tilde{g}(\vec{z})$
- Return the corresponding predictor in the original space: $g(\vec{x}) = \tilde{g}\left(\Phi(\vec{x})\right)$

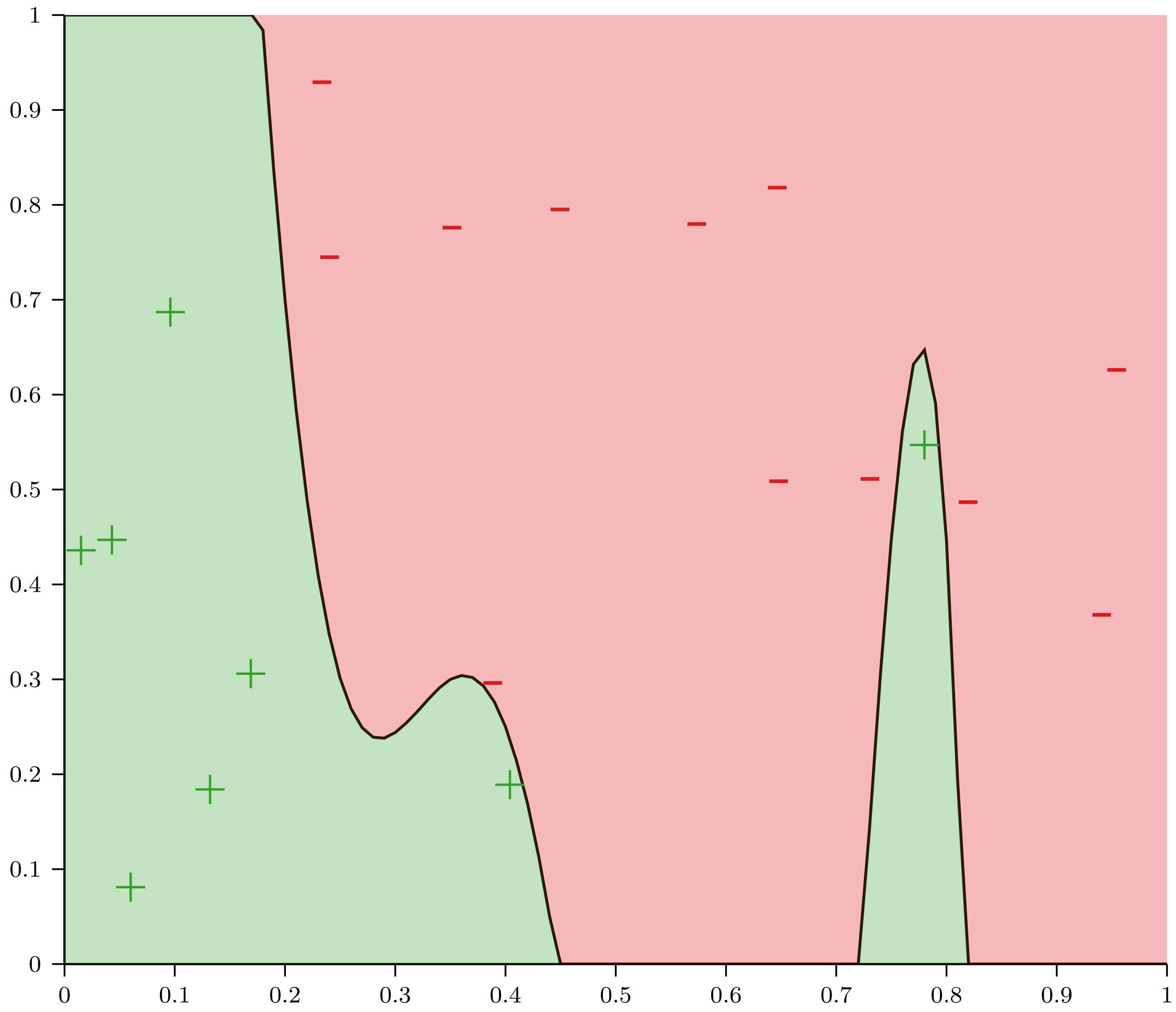
Nonlinear Transforms for Approving Credit

- Input: $[x_1 = \text{age}, x_2 = \text{income}, x_3 = \text{credit score}]$
- How close is this person to some optimal age, a^*?
- What is this person's income scaled by age?
- Having a very low credit score is more significant than having a very high credit score.
- Transformation: $\Phi(\vec{x}) = \left[x_1, x_2, x_3, |a^* - x_1|, \frac{x_2}{x_1}, \sqrt{x_3}\right]$

Linear Models



Nonlinear Models?



Tradeoffs

	Low-Dimensional Input Space	**High-Dimensional Input Space**
E_{in}	High	Low
Generalization	Good	Bad

Overfitting

Overfitting

- Overfitting is fitting the training data “more than is warranted”
- Fitting noise rather than signal

Experimental Setup

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$ and $n = 20$

- f is a 10^{th}-order polynomial in x with additive Gaussian noise

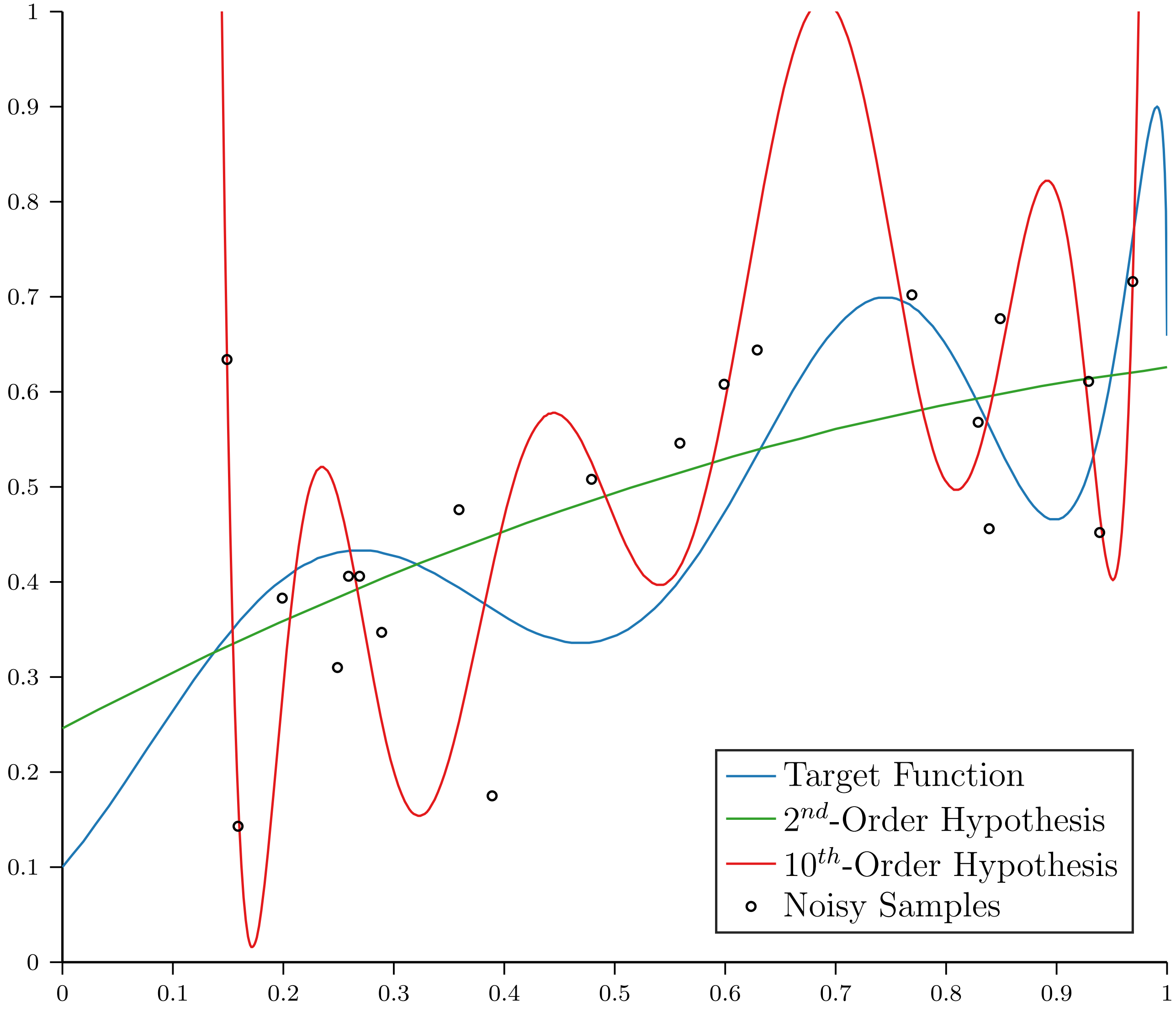
$$y = \sum_{d=0}^{10} a_d x^d + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2)$$

- $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomials
  - $\vec{z} = \Phi_2(x) = [x, x^2]$

- $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomials
  - $\vec{z} = \Phi_{10}(x) = [x, x^2, x^3, x^4, x^5, x^6, x^7, x^8, x^9, x^{10}]$

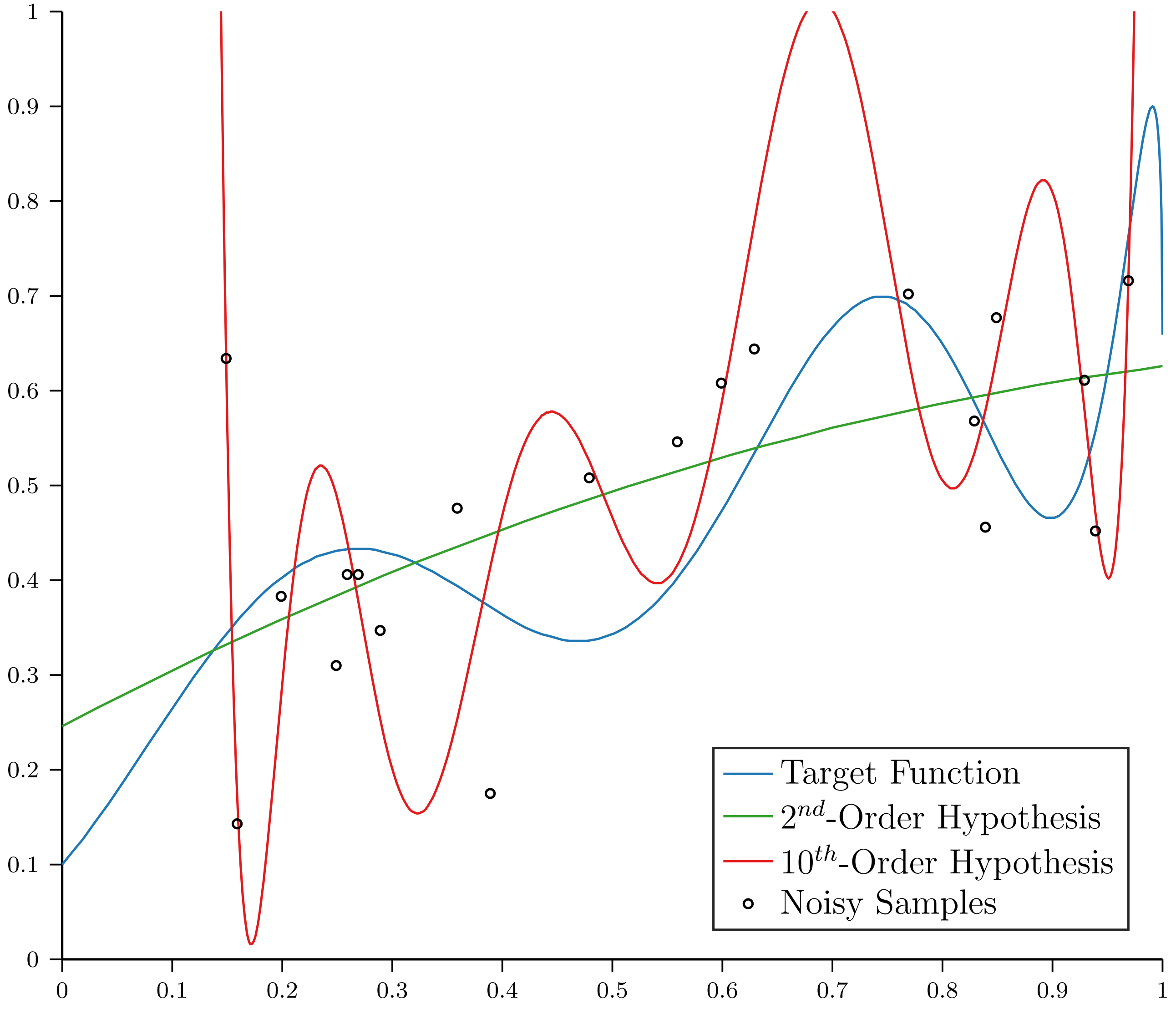
Noisy Targets

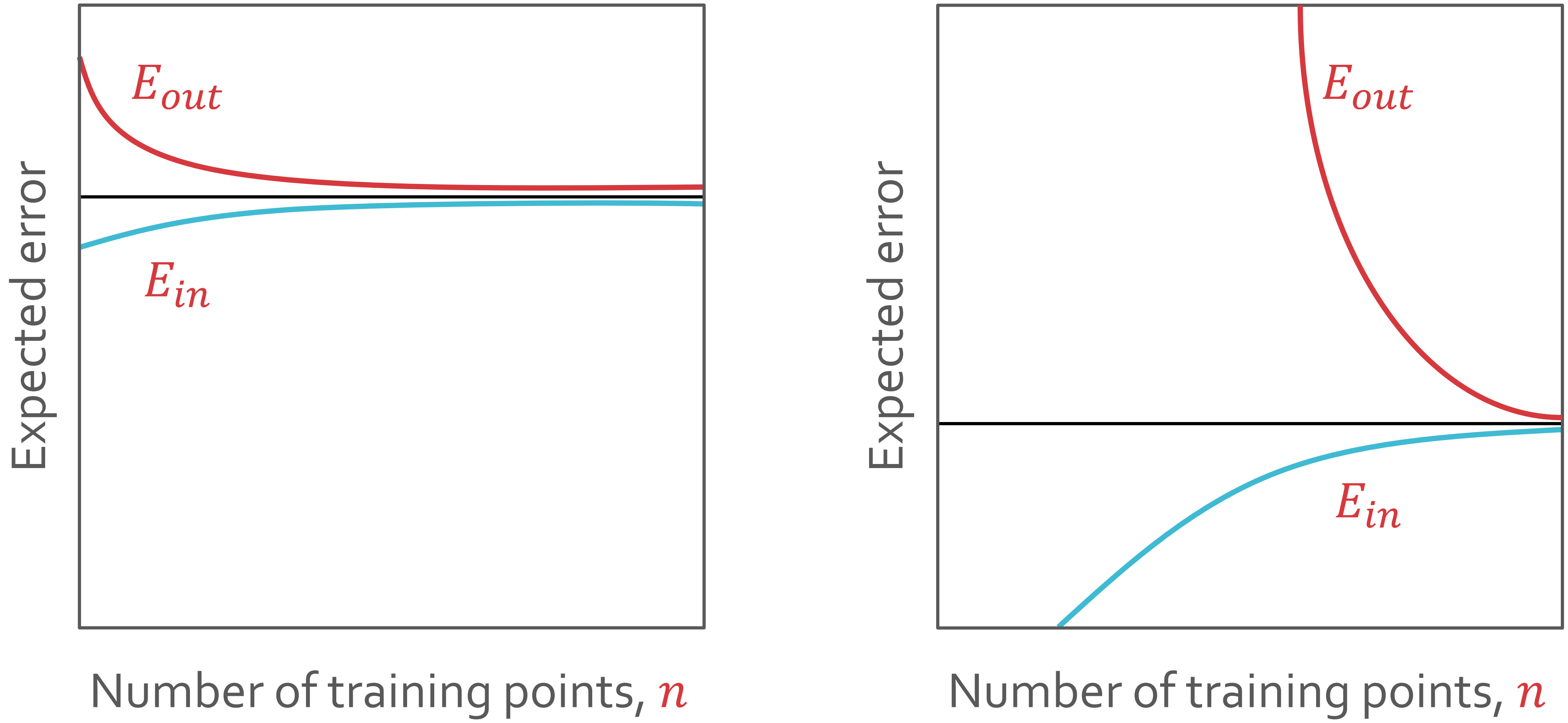
- 10-d target function with additive Gaussian noise
- $y = f(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$
- $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomial
- $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomial



Noisy Targets

	$\mathcal{H}_2$	$\mathcal{H}_{10}$
E_{in}	0.016	0.011
E_{out}	0.009	3797

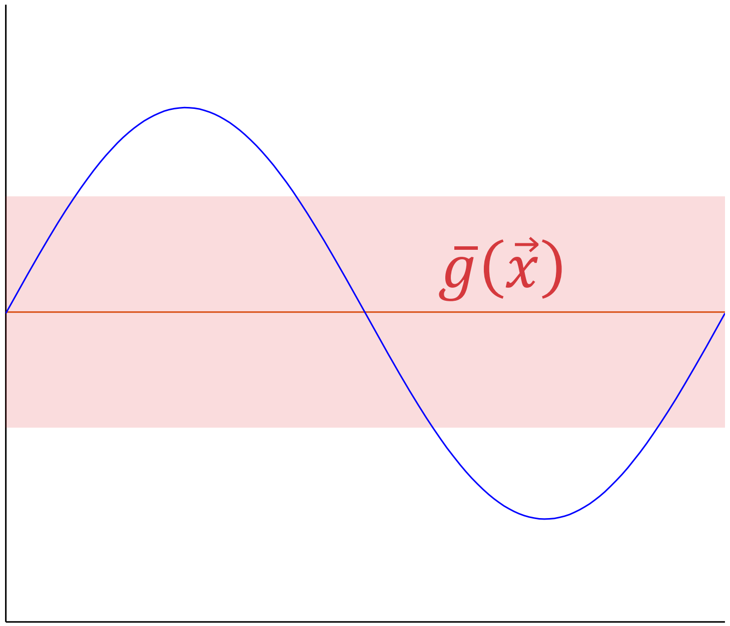




Simple model

Complex model

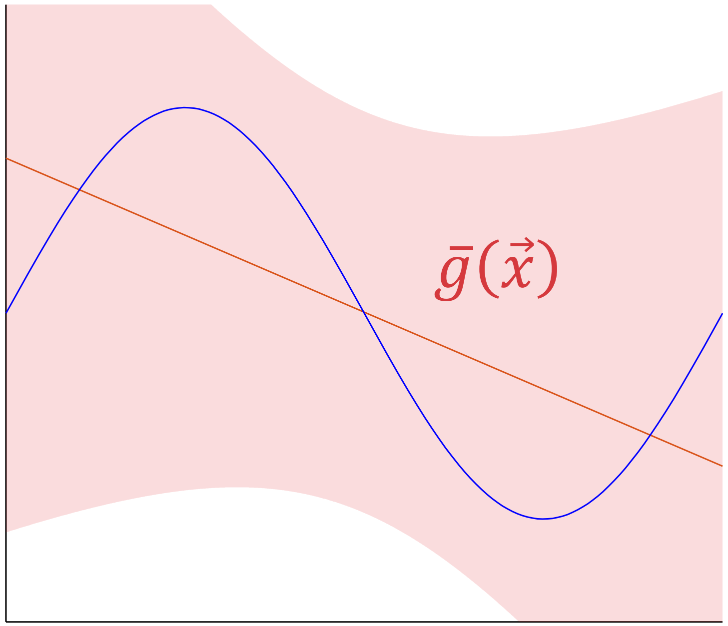
Bias-Variance Tradeoff (Example)



Bias of $\bar{g}(\vec{x}) \approx 0.50$

Variance of $g_{\mathcal{D}}(\vec{x}) \approx 0.25$

$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] \boxed{\approx 0.75}$



Bias of $\bar{g}(\vec{x}) \approx 0.21$

Variance of $g_{\mathcal{D}}(\vec{x}) \approx 1.74$

$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] \approx 1.95$

Experimental Setup

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$ and $n = 100$

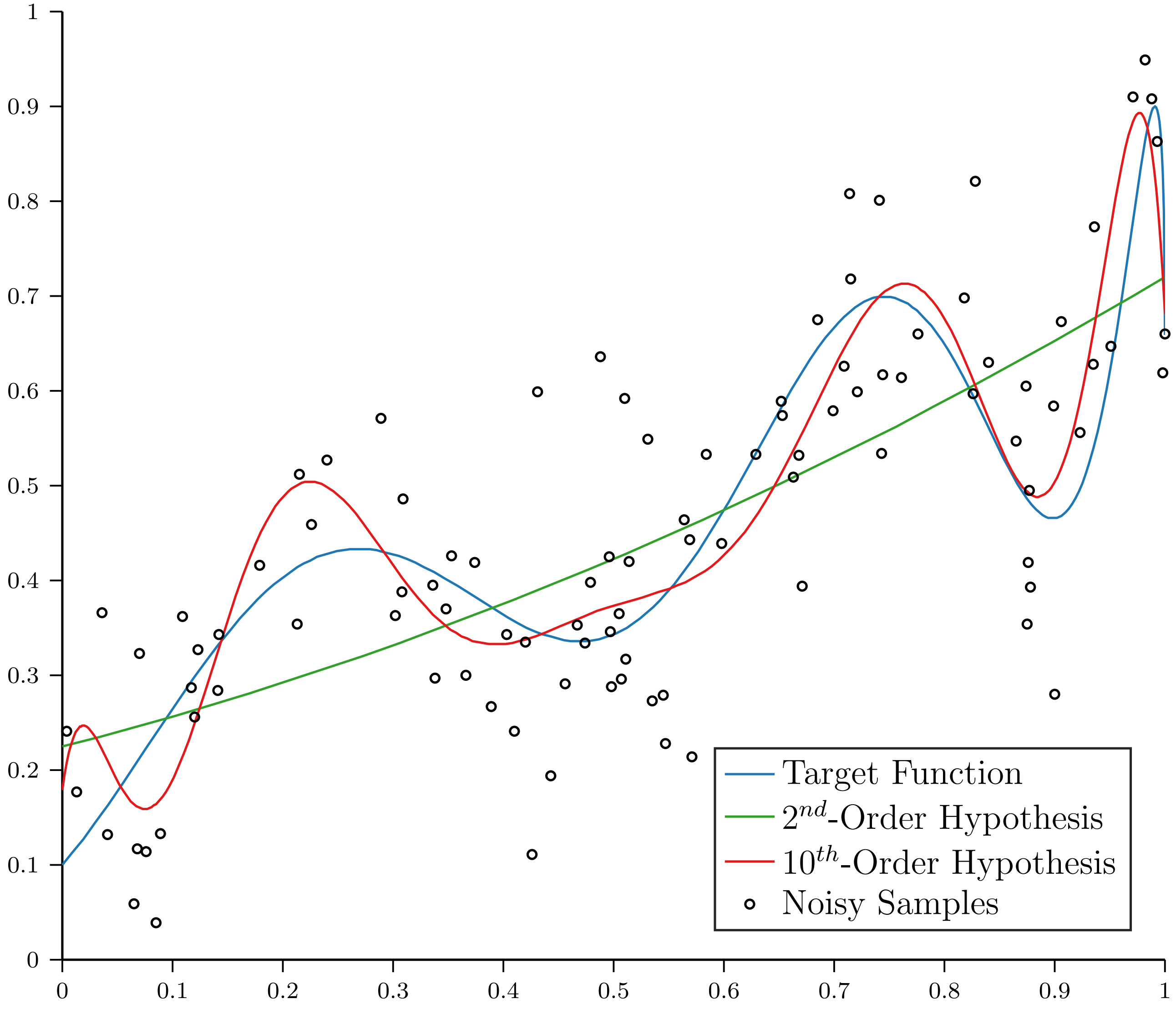
- f is a 10^{th}-order polynomial in x with additive Gaussian noise

$$y = \sum_{d=0}^{10} a_d x^d + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2)$$

- $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomials
  - $\vec{z} = \Phi_2(x) = [x, x^2]$
- $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomials
  - $\vec{z} = \Phi_{10}(x) = [x, x^2, x^3, x^4, x^5, x^6, x^7, x^8, x^9, x^{10}]$

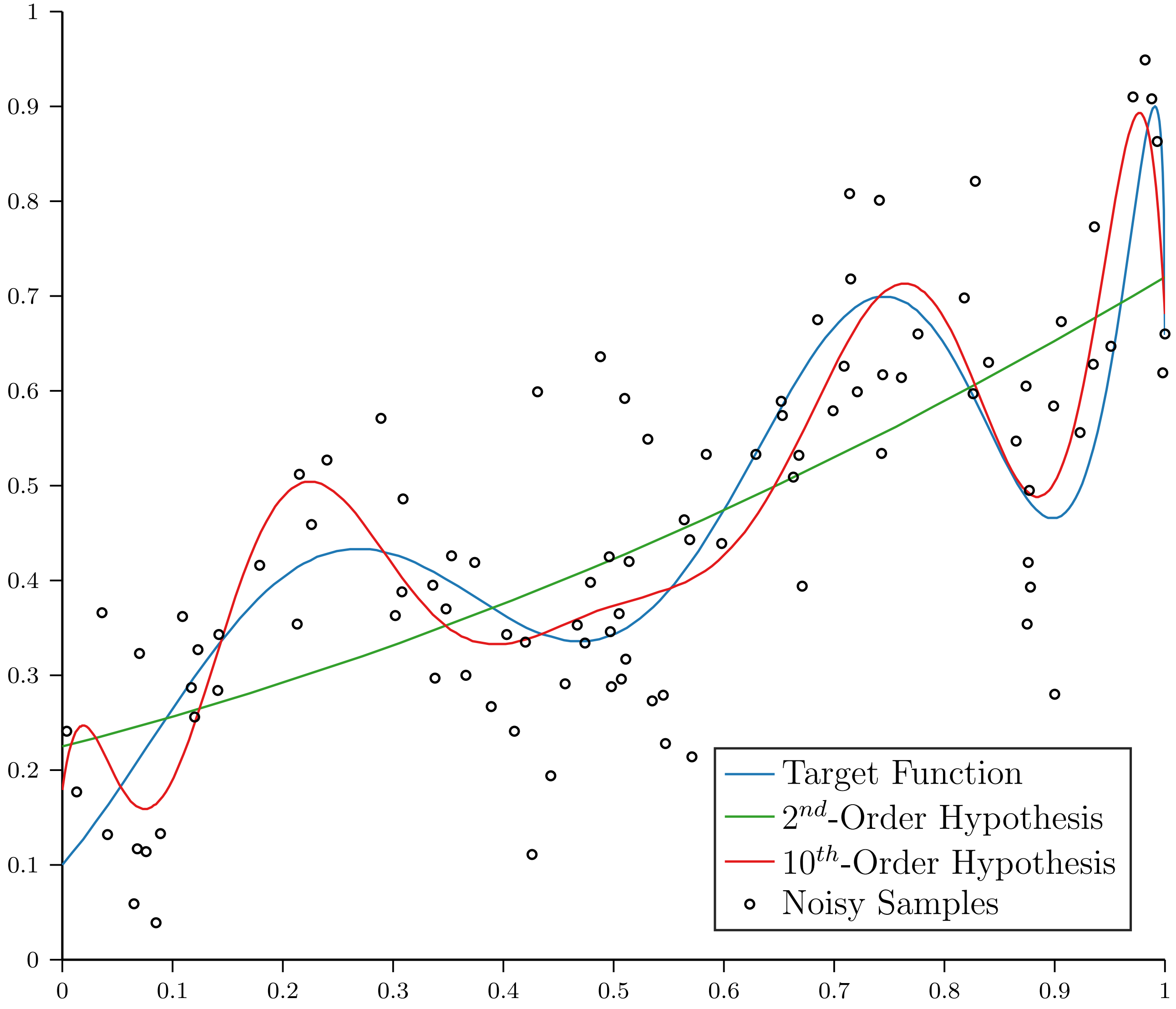
Noisy Targets

- 10-d target function with additive Gaussian noise
- $y = f(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$
- $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomial
- $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomial



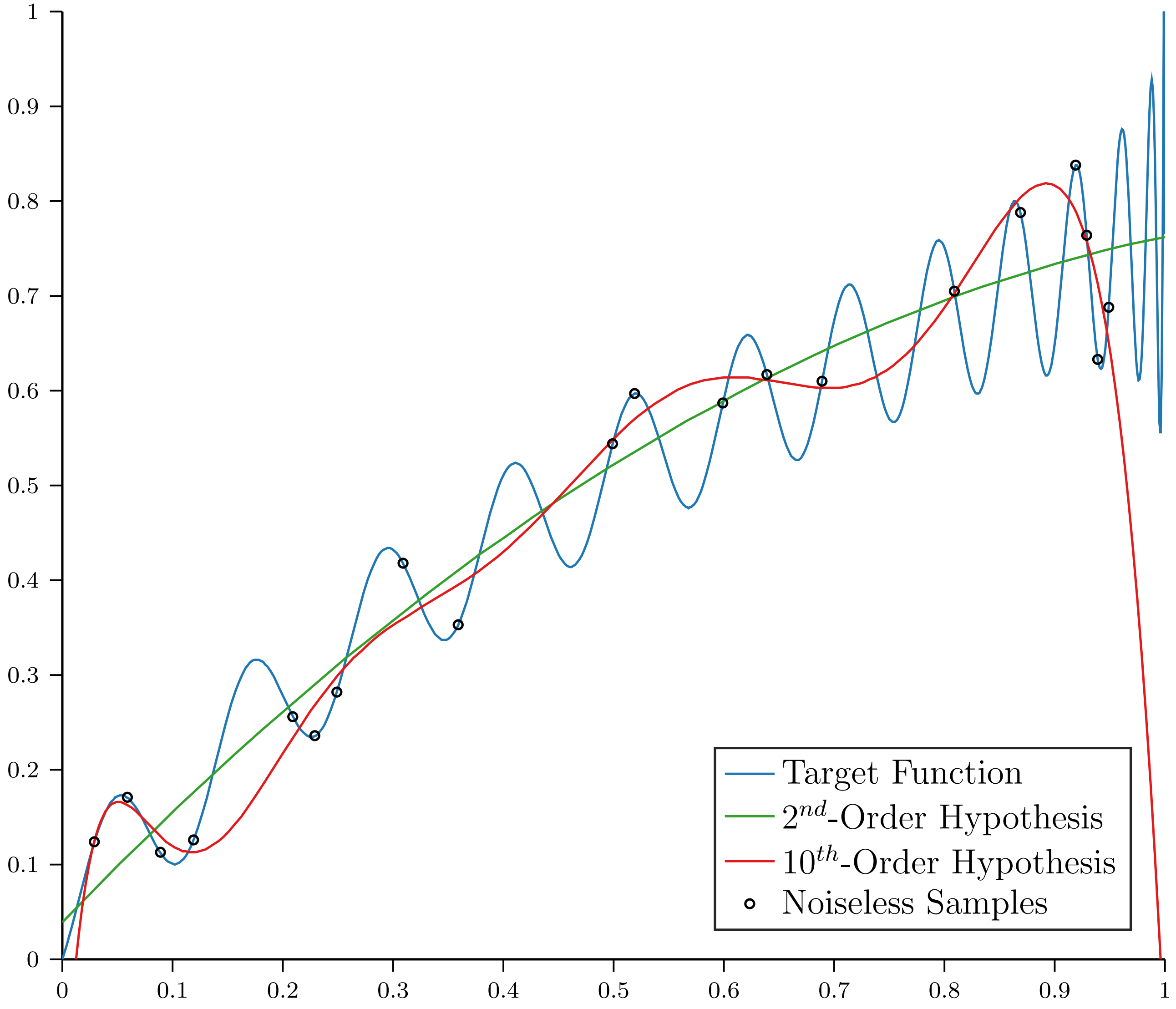
Noisy Targets

	$\mathcal{H}_2$	$\mathcal{H}_{10}$
E_{in}	0.018	0.010
E_{out}	0.009	0.003



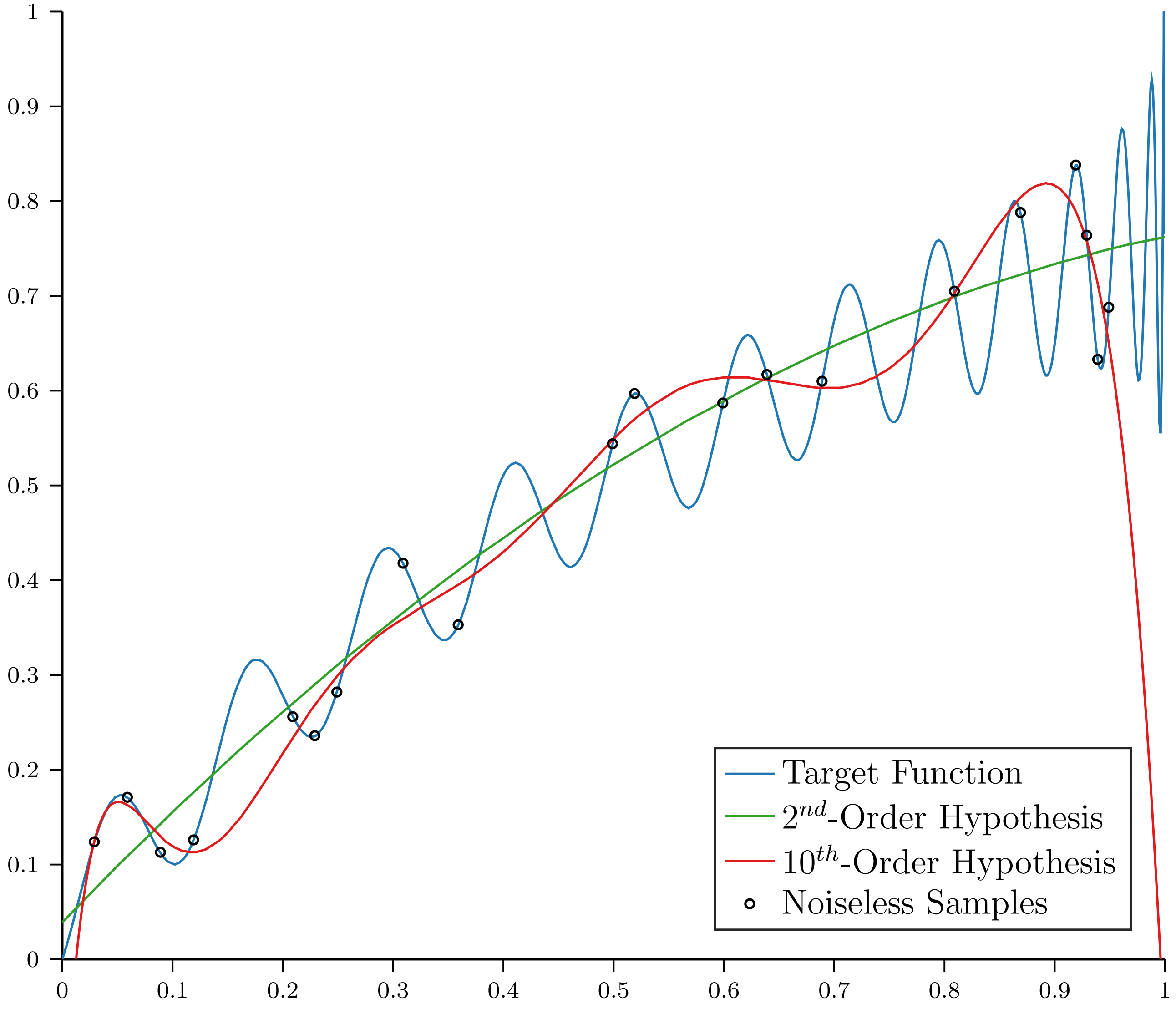
Noiseless Targets

- 50-d target function with no noise
- $y = \sum_{d=0}^{50} a_d x^d$
- $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomial
- $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomial



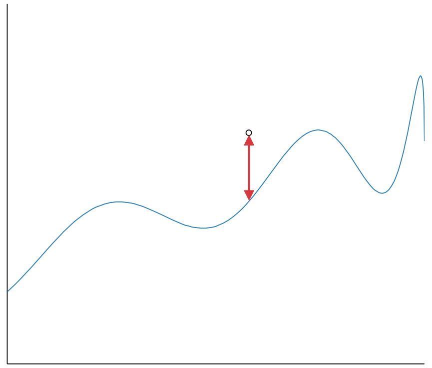
Noiseless Targets

	$\mathcal{H}_2$	$\mathcal{H}_{10}$
E_{in}	0.003	0.001
E_{out}	0.004	0.016

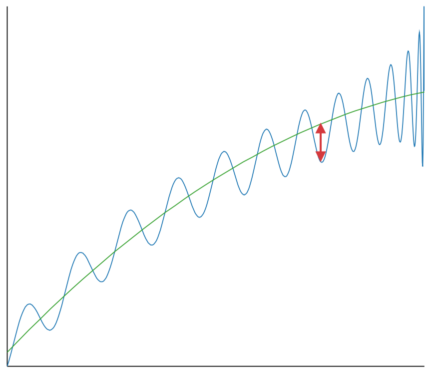


Two Types of Noise

- Stochastic noise
  - Measurement error
  - Random
  - Not affected by choice of $\mathcal{H}$



- Deterministic noise
  - Limitations of $\mathcal{H}$
  - Not random
  - Dependent on $\mathcal{H}$ and f



- **Given a single dataset $\mathcal{D}$ and a fixed $\mathcal{H}$, the two types of noise are indistinguishable**

Overfitting

	Direction	Overfitting
Number of points	↑	↓
	↓	↑
Stochastic noise	↑	↑
	↓	↓
Deterministic noise	↑	↑
	↓	↓

Bias-Variance ft. Noise

- $\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\vec{x}}[(g_{\mathcal{D}}(\vec{x}) - y)^2]\right]$ where $y = f(\vec{x}) + \epsilon$

$$\vdots$$

$$= \mathbb{E}_{\vec{x}}[\text{Variance of } g_{\mathcal{D}}(\vec{x})]$$

$$+ \mathbb{E}_{\vec{x}}[\text{Bias of } \bar{g}(\vec{x})]$$

$$+ \text{Stochastic noise}$$

Bias-Variance ft. Noise

- $\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\vec{x}}[(g_{\mathcal{D}}(\vec{x}) - y)^2]\right]$ where $y = f(\vec{x}) + \epsilon$

$$\vdots$$

$$= \mathbb{E}_{\vec{x}}[\text{Variance of } g_{\mathcal{D}}(\vec{x})]$$

$$+ \mathbb{E}_{\vec{x}}[\text{Deterministic noise of } \bar{g}(\vec{x})]$$

$$+ \text{Stochastic noise}$$