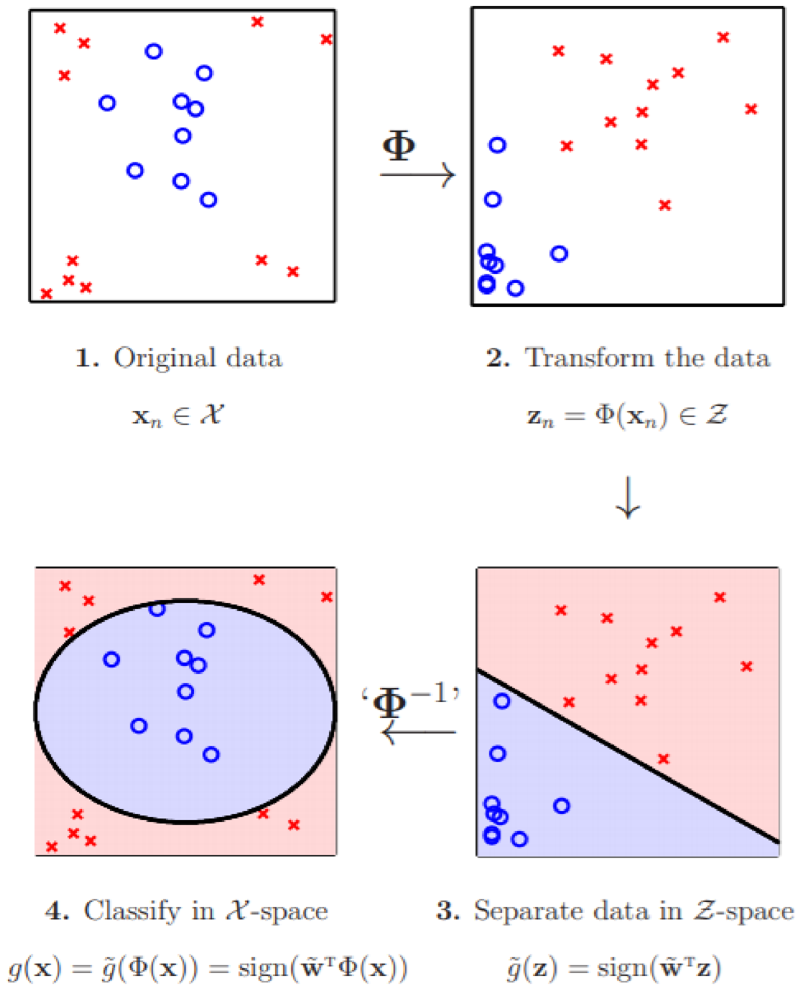
Machine Learning from Data

Lecture 11: Spring 2021

Today’s Lecture

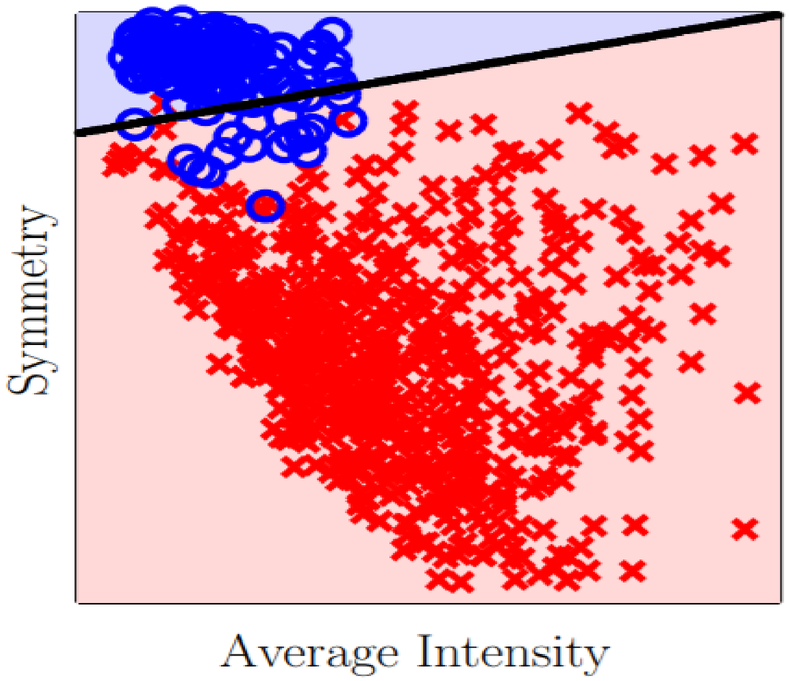
- Overfitting
  - What is overfitting?
  - When does it occur?
  - Stochastic Vs. Deterministic Noise

Non-Linear Transforms

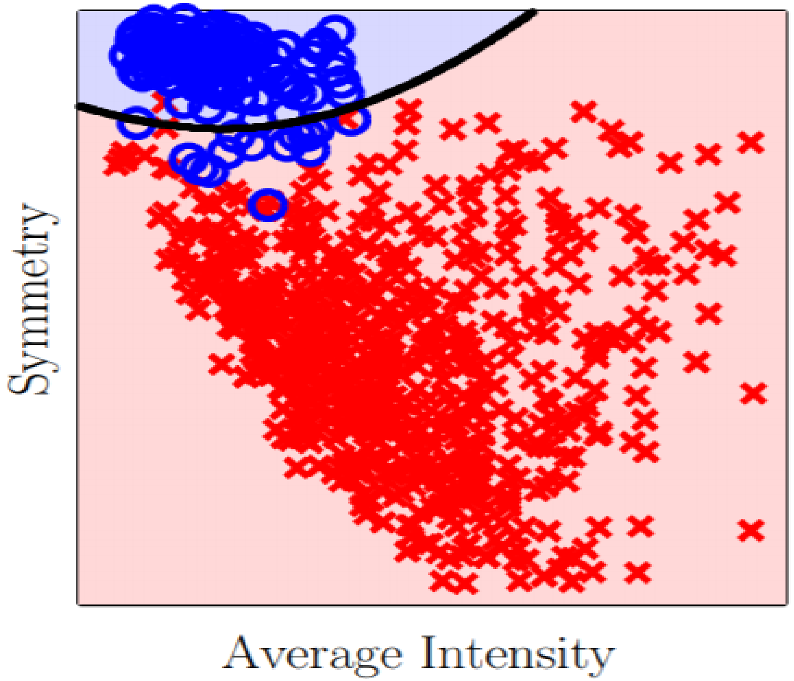


<u>$\mathcal{X}$-space is $\mathbb{R}^d$</u>	<u>$\mathcal{Z}$-space is $\mathbb{R}^{\tilde{d}}$</u>
$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$	$\mathbf{z} = \Phi(\mathbf{x}) = \begin{bmatrix} 1 \\ \Phi_1(\mathbf{x}) \\ \vdots \\ \Phi_{\tilde{d}}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 1 \\ z_1 \\ \vdots \\ z_{\tilde{d}} \end{bmatrix}$
$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$	$\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$
$y_1, y_2, \ldots, y_N$	$y_1, y_2, \ldots, y_N$
no weights	$\tilde{\mathbf{w}} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{\tilde{d}} \end{bmatrix}$
$d_{\text{VC}} = d + 1$	$\boldsymbol{d_{\text{VC}} = d + 1}$
$g(\mathbf{x}) = \text{sign}(\tilde{\mathbf{w}}^\mathsf{T}\Phi(\mathbf{x}))$	

Digits Data



Linear model
$E_{\text{in}} = 2.13\%$
$E_{\text{out}} = 2.38\%$



3rd order polynomial model
$E_{\text{in}} = 1.75\%$
$E_{\text{out}} = 1.87\%$

Humans Overfit (Superstitions)

- Fear of Friday the 13^{th}

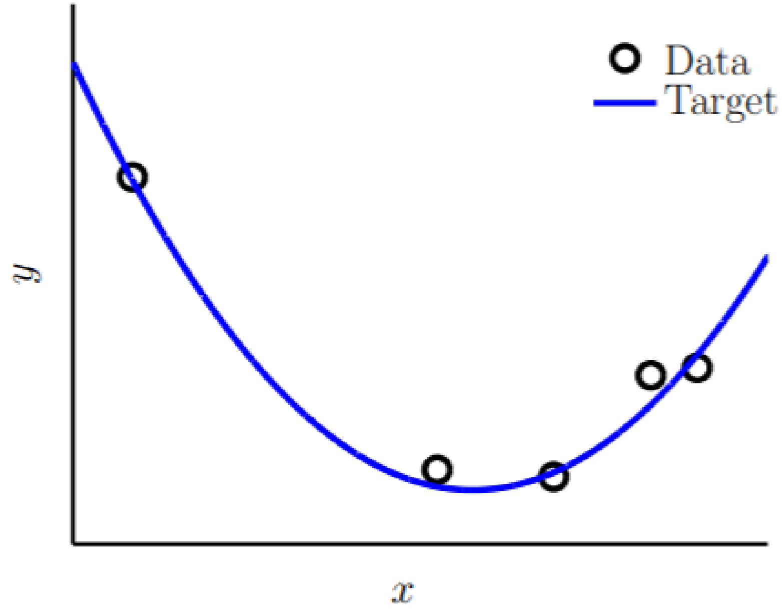
Illustration of Overfitting

Quadratic f

5 data points

A *little* noise (measurement error)

5 data points $\rightarrow$ 4th order polynomial fit



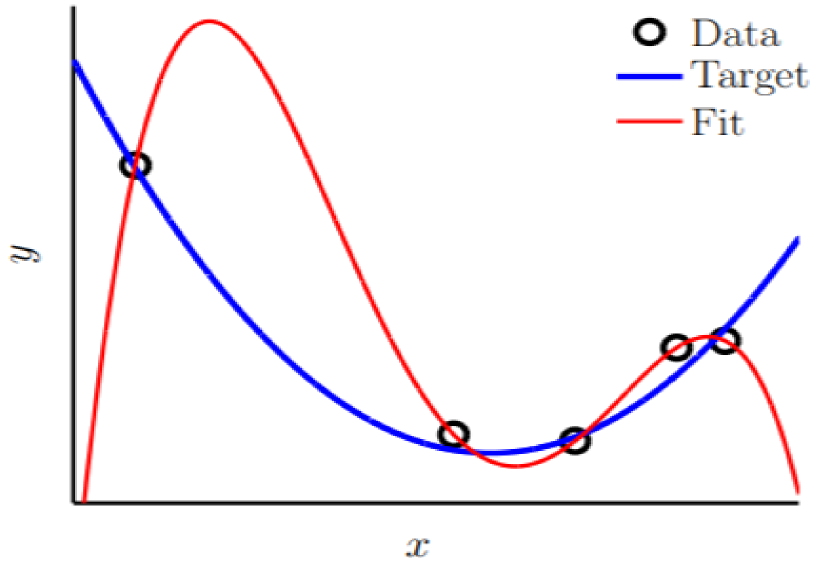
Overfitting Example

Quadratic f

5 data points

A *little* noise (measurement error)

5 data points $\rightarrow$ 4th order polynomial fit



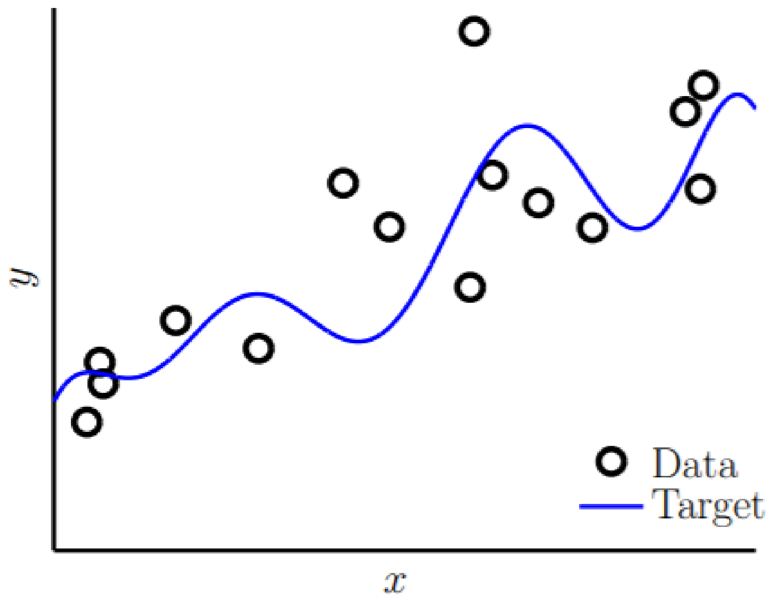
Classic overfitting: simple target with excessively complex $\mathcal{H}$.

$E_{\text{in}} \approx 0$; $E_{\text{out}} \gg 0$

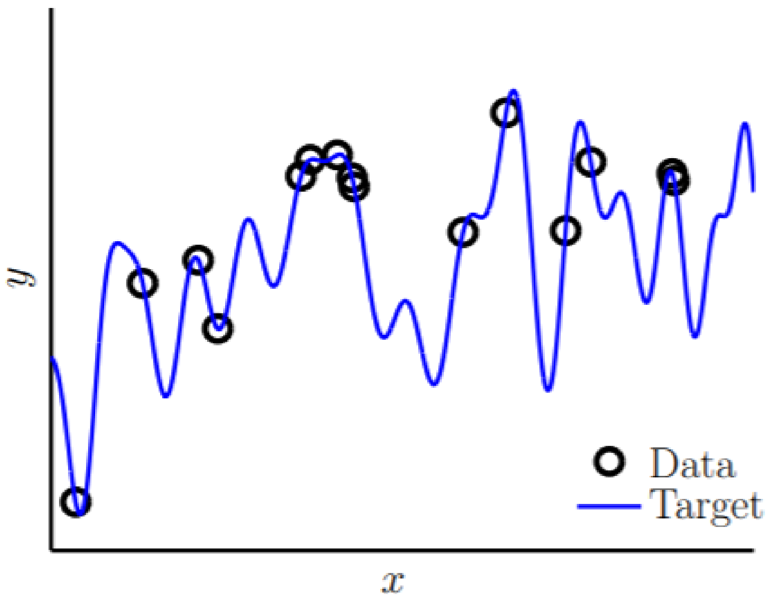
Define Overfitting

- Fitting the data more than is warranted.

Case Study

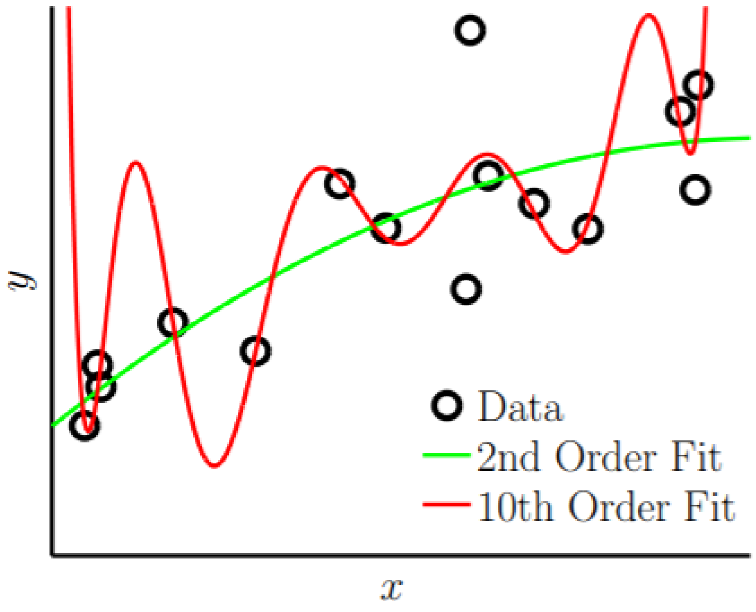


10th order f with noise.



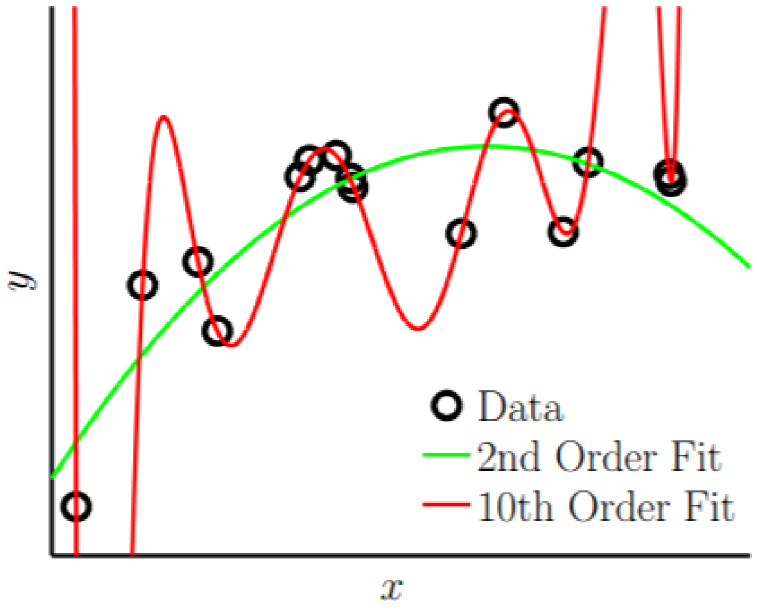
50th order f with no noise.

2^{nd} order vs. 10^{th} order polynomial



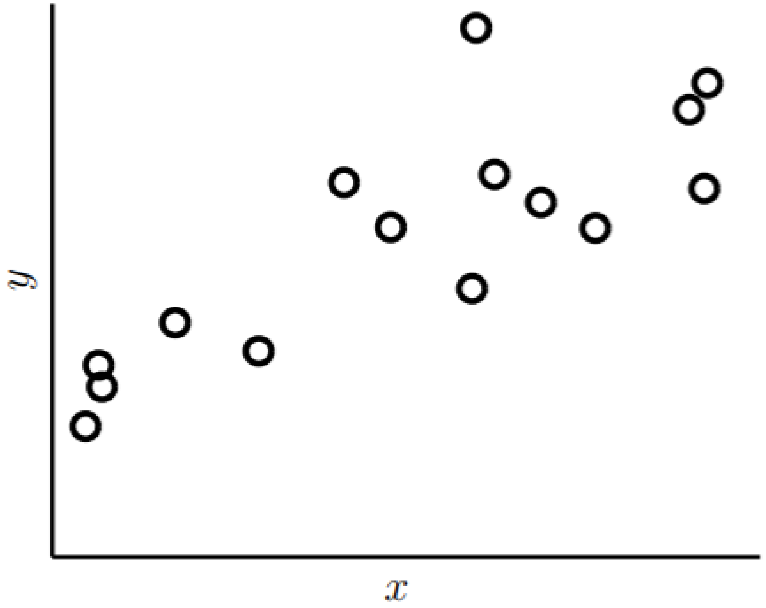
simple noisy target

	2nd Order	10th Order
E_{in}	0.050	0.034
E_{out}	0.127	**9.00**

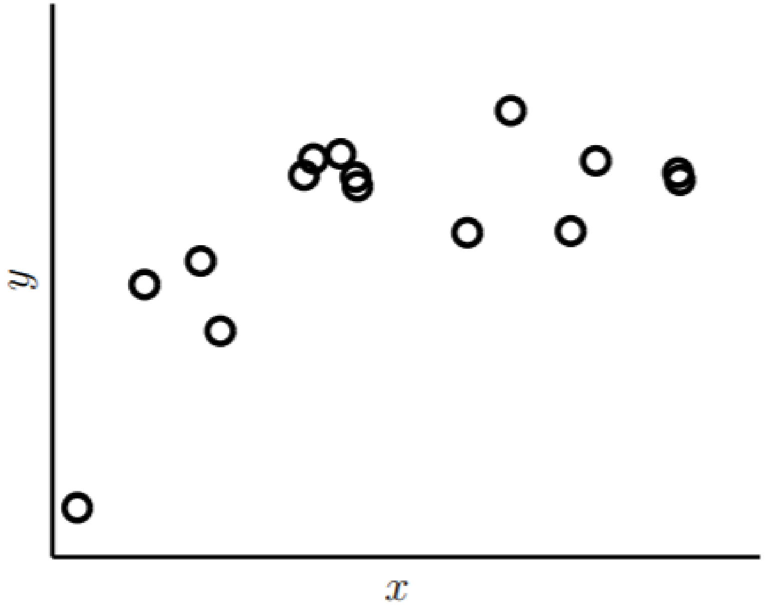


complex noiseless target

	2nd Order	10th Order
E_{in}	0.029	10^{-5}
E_{out}	0.120	**7680**



Simple f with noise.



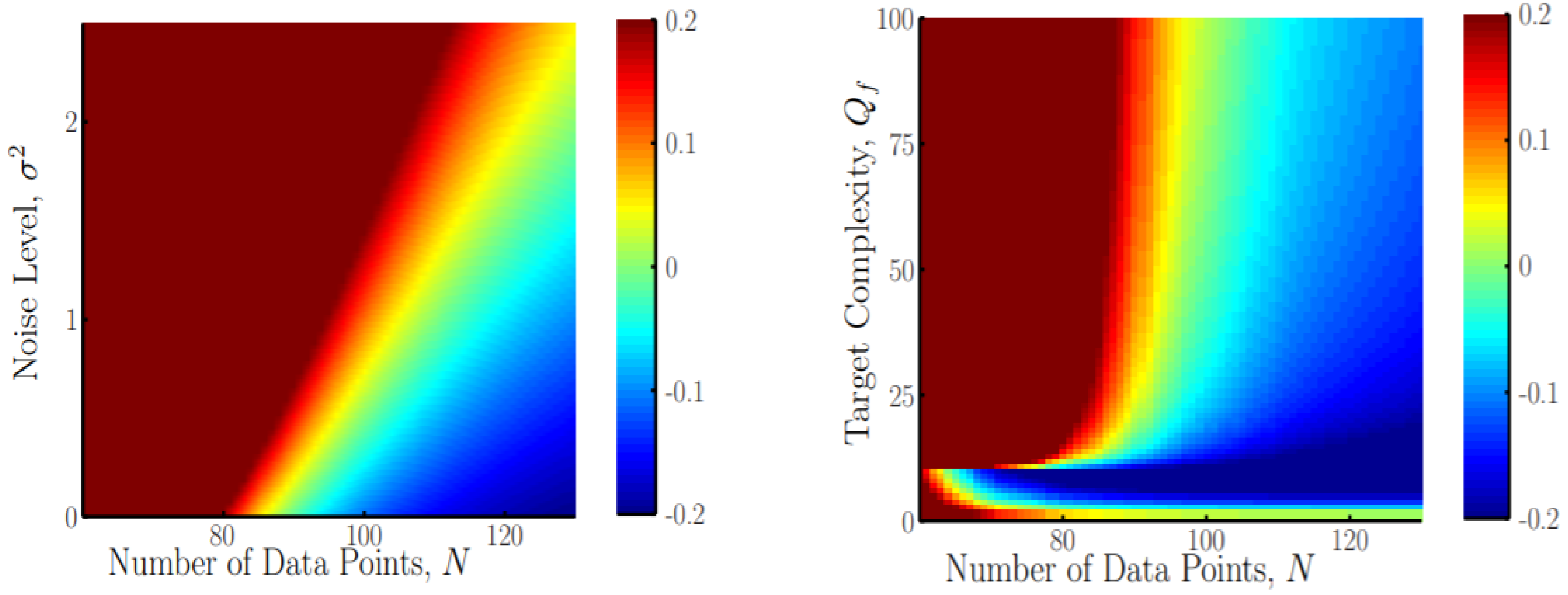
Complex f with no noise.

$\mathcal{H}$ **should match** ***quantity and quality of data***, **not** f

Measure Overfitting

-

Overfit Measure: $E_{\text{out}}(\mathcal{H}_{10}) - E_{\text{out}}(\mathcal{H}_2)$



Number of data points ↑	Overfitting ↓
Noise ↑	Overfitting ↑
Target complexity ↑	Overfitting ↑

Noise

•

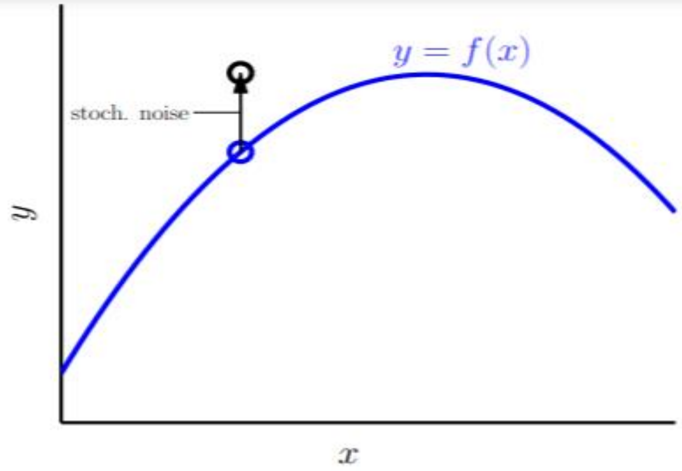
Stochastic Noise

We would like to learn from ○:

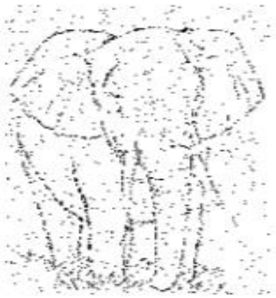
$$y_n = f(x_n)$$

Unfortunately, we only observe ○:

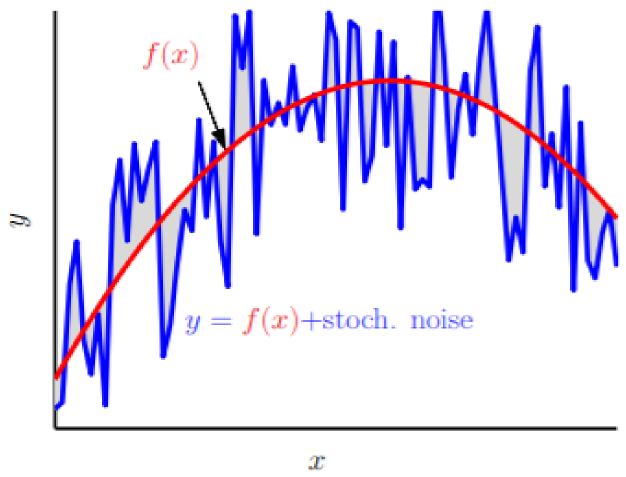
$$y_n = f(x_n) + \underset{\substack{\uparrow \\ \text{no one can model this}}}{\text{'stochastic noise'}}$$



Stochastic Noise: **fluctuations/measurement errors we cannot model.**



Stochastic Noise

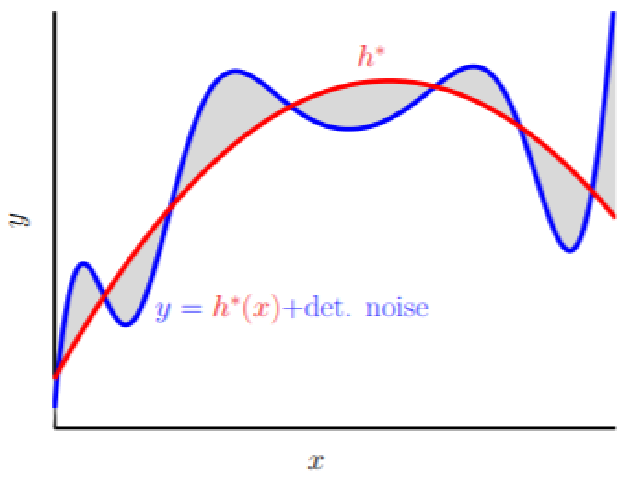


source: random measurement errors

re-measure y_n
  stochastic noise changes.

change $\mathcal{H}$
  stochastic noise the same.

Deterministic Noise



source: learner's $\mathcal{H}$ cannot model f

re-measure y_n
  deterministic noise the same.

change $\mathcal{H}$
  deterministic noise changes.

We have single $\mathcal{D}$ and fixed $\mathcal{H}$ so we cannot distinguish

Deterministic Noise

We would like to learn from ○:

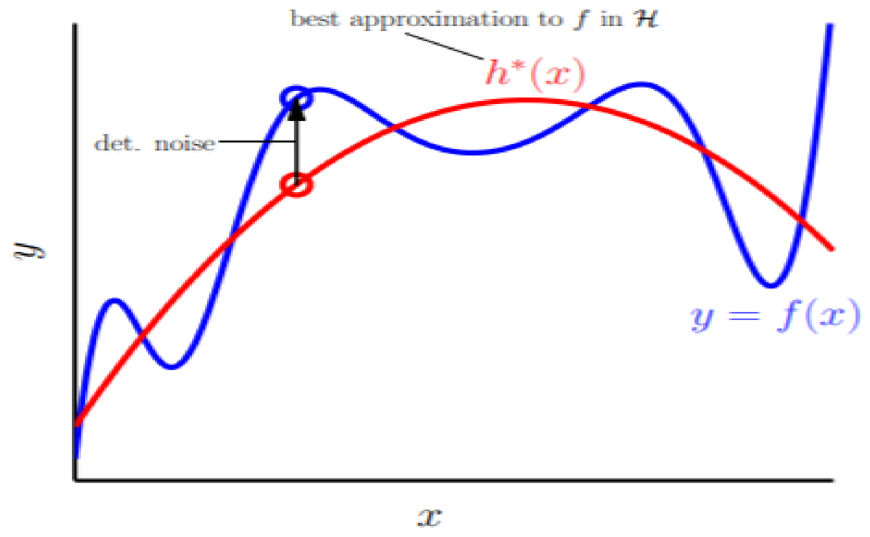
$$y_n = h^*(x_n)$$

Unfortunately, we only observe ○:

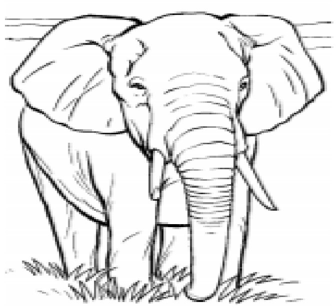
$$\begin{aligned} y_n &= f(x_n) \\ &= h^*(x_n) + \text{`deterministic noise'} \end{aligned}$$

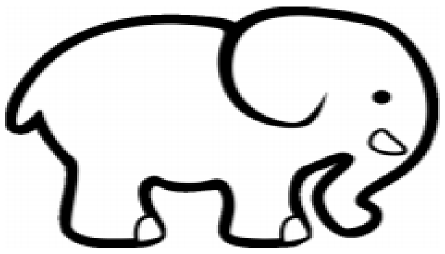
$\uparrow$

$\mathcal{H}$ cannot model this



Deterministic Noise: **the part of f we cannot model.**





Bias Variance Analysis and Noise

•

Summary

-

Thanks!