Machine Learning from Data

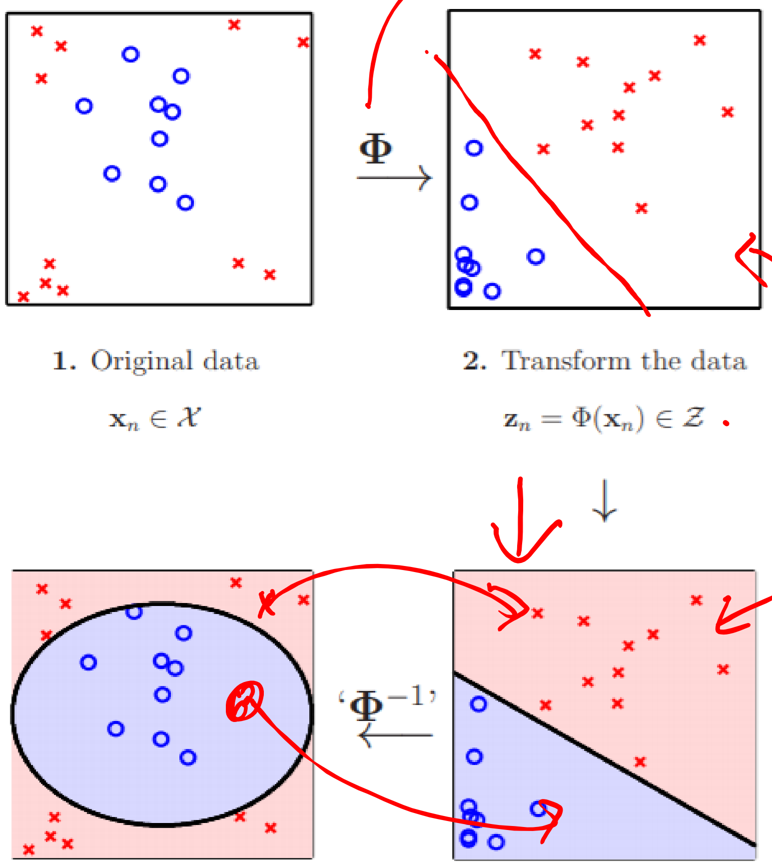
Lecture 11: Spring 2021

Today's Lecture

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

Non-Linear Transforms

function



1. Original data

$\mathbf{x}_n \in \mathcal{X}$

2. Transform the data

$\mathbf{z}_n = \Phi(\mathbf{x}_n) \in \mathcal{Z}$

3. Separate data in $\mathcal{Z}$-space

$\tilde{g}(\mathbf{z}) = \text{sign}(\tilde{\mathbf{w}}^{\mathsf{T}}\mathbf{z})$

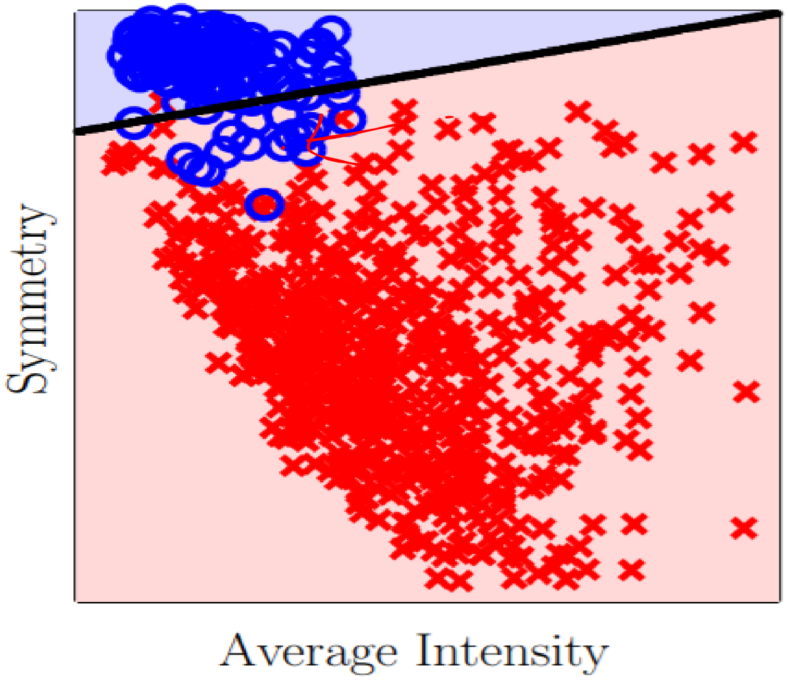
4. Classify in $\mathcal{X}$-space

$g(\mathbf{x}) = \tilde{g}(\Phi(\mathbf{x})) = \text{sign}(\tilde{\mathbf{w}}^{\mathsf{T}}\Phi(\mathbf{x}))$

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

Digits Data

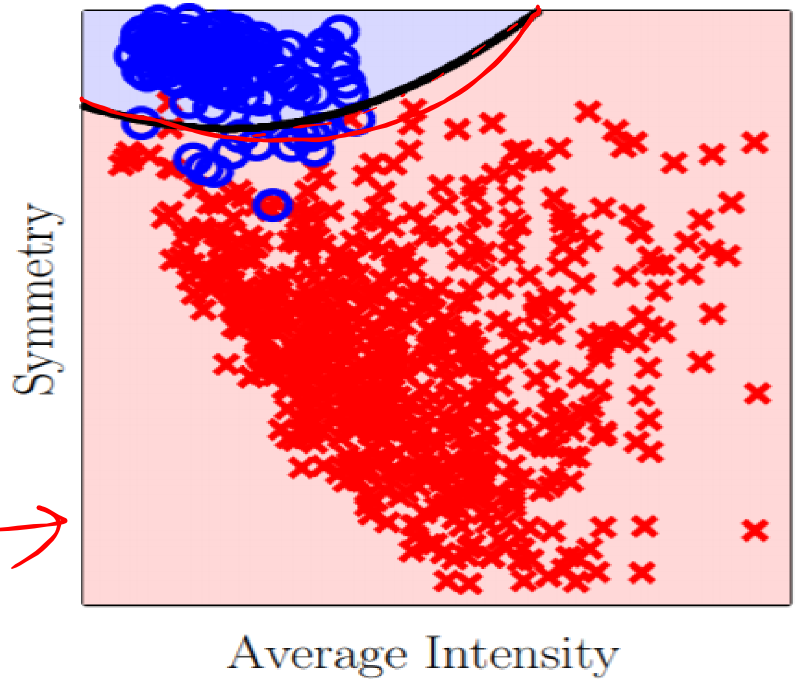
10%



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 13th

Unfortunate events.

Overfitting $\xrightarrow{\text{lead to}}$ Opposite effect.

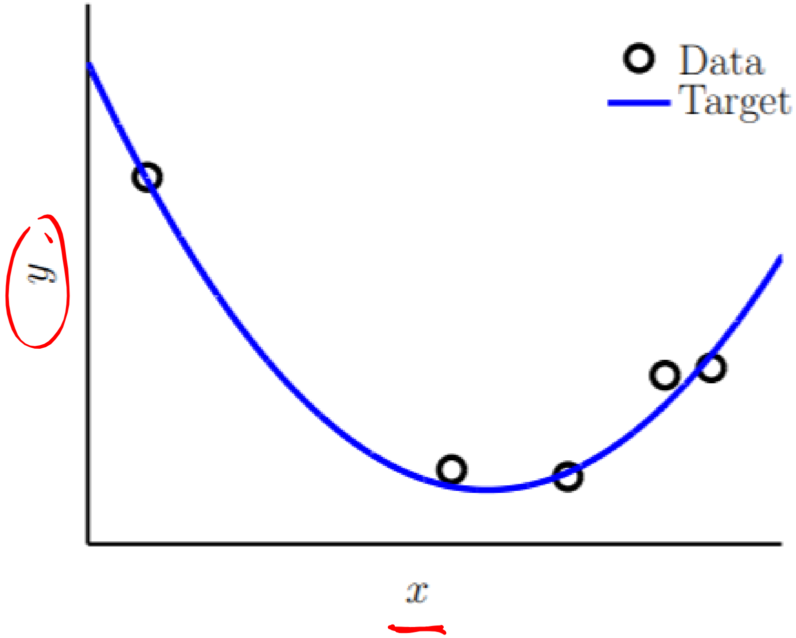
Illustration of Overfitting

Quadratic f

5 data points

A *little* noise (measurement error)

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



Overfitting Example

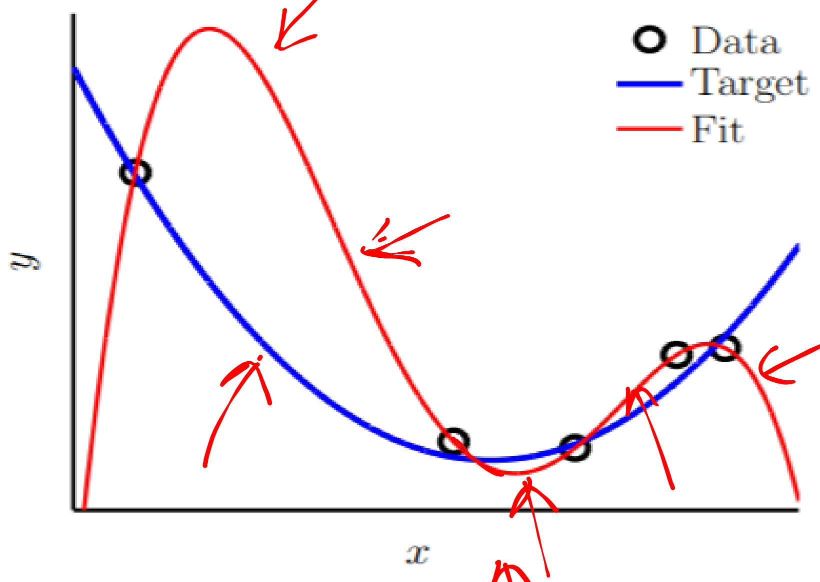
4th order polynomial

Quadratic f

5 data points

A *little* noise (measurement error)

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



$E_{in} = 0$

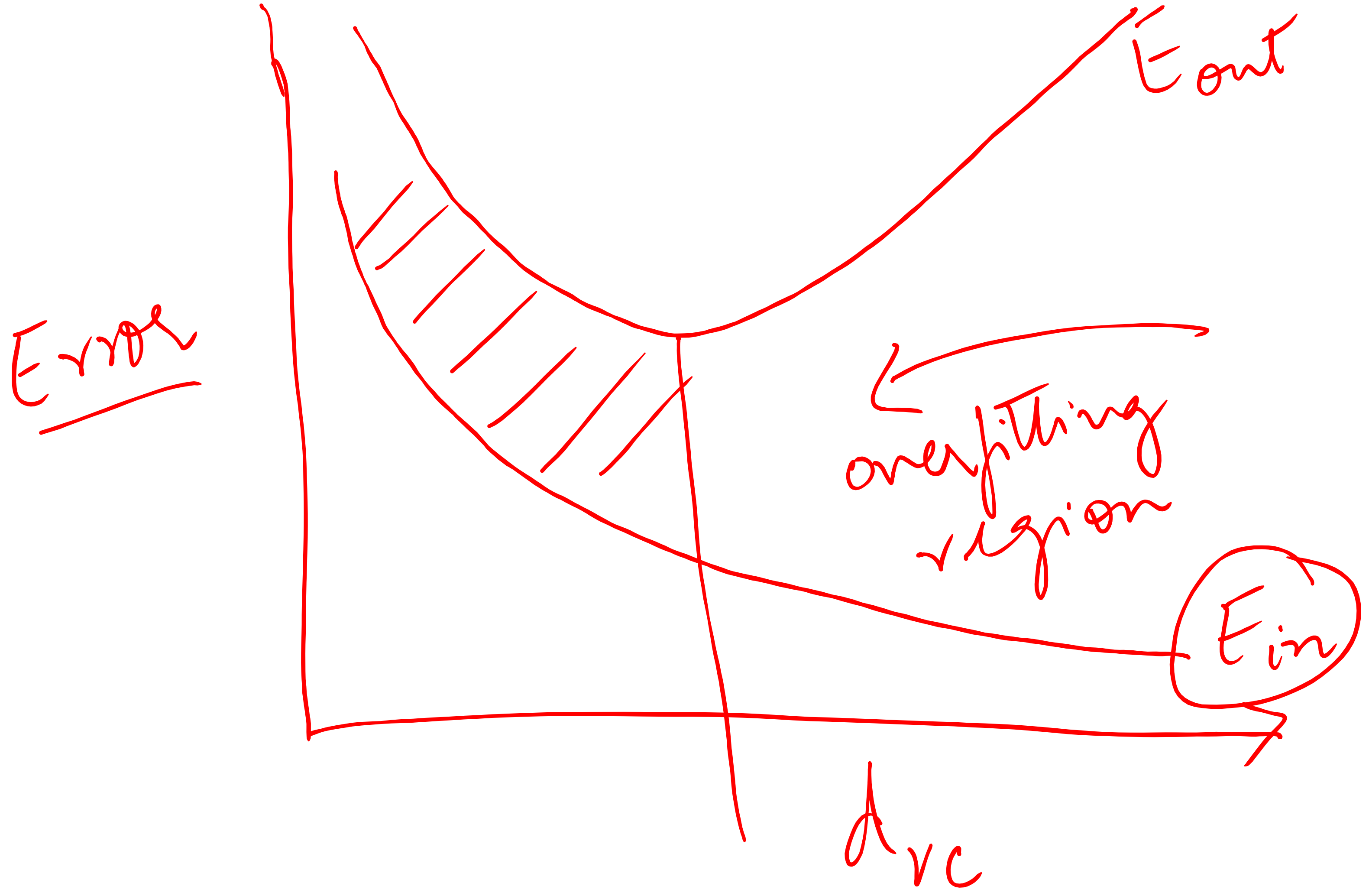
Classic overfitting: simple target with excessively complex $\mathcal{H}$.

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

Define Overfitting

- Fitting the data more than is warranted. → bad generalization

→ process of picking a $\mathcal{H}$ with lower E_{in} resulting in a higher E_{out}.



Error
Eout
overfitting region
Ein
dvc

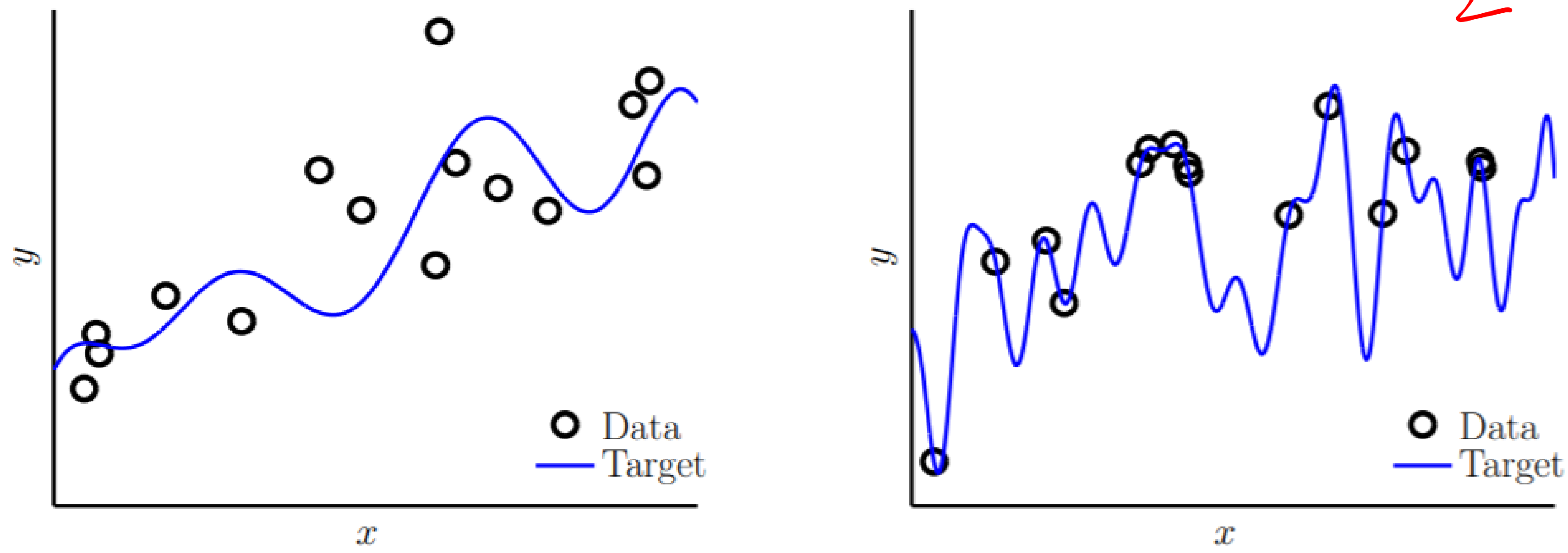
Case Study

Choice of H-Sets: $\mathcal{H}_2$ (2nd order polynomial)

$\mathcal{H}_{10}$: (10th order)

$\mathcal{H}_2$ $[1, x] \xrightarrow{\Phi} [1, x, x^2]$

$\mathcal{H}_{10}$ $[1, x] \xrightarrow{\Phi} [1, x, x^2 \ldots x^{10}]$



10th order f with noise.

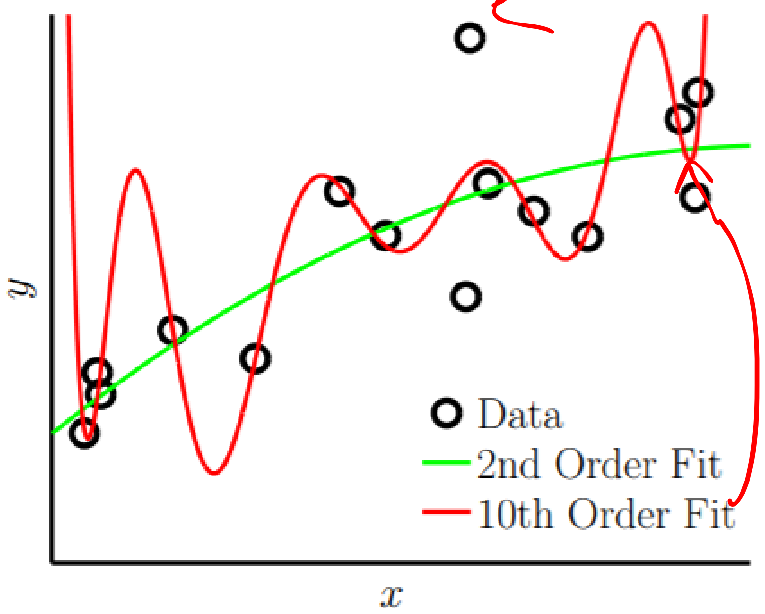
(a)

50th order f with no noise.

(b)

	f : 10th order with noise	f : 50th order no noise
simple H_2	noise ✗	No hope
complex H_{10}	✓ ✗	✓ ✗

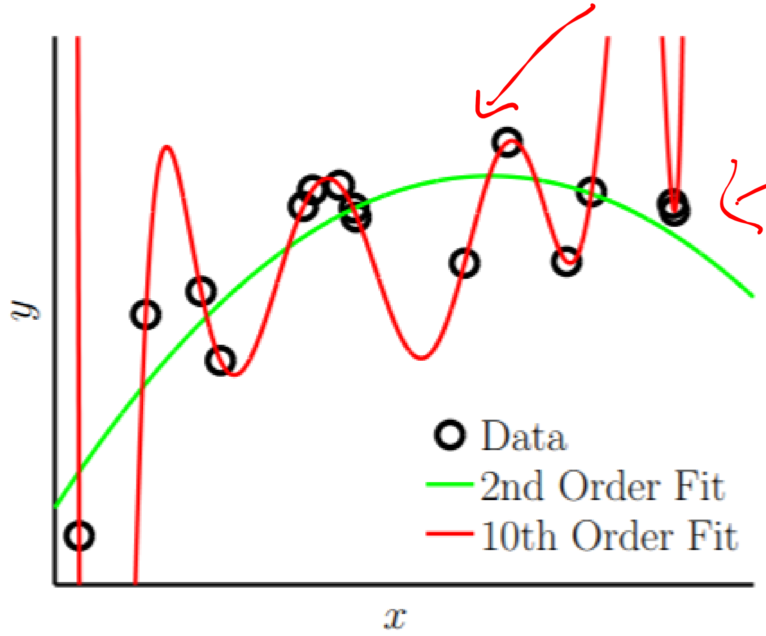
2nd order vs. 10th order polynomial



(a)

simple noisy target

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

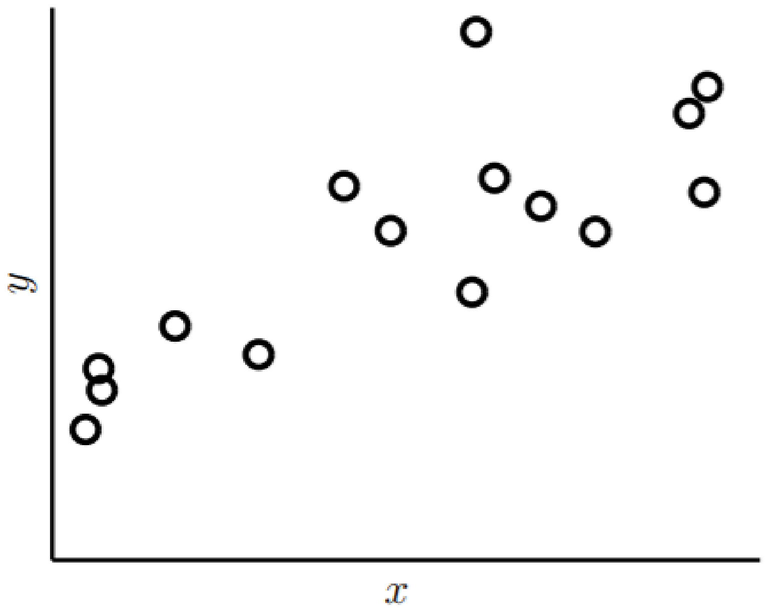


50th order

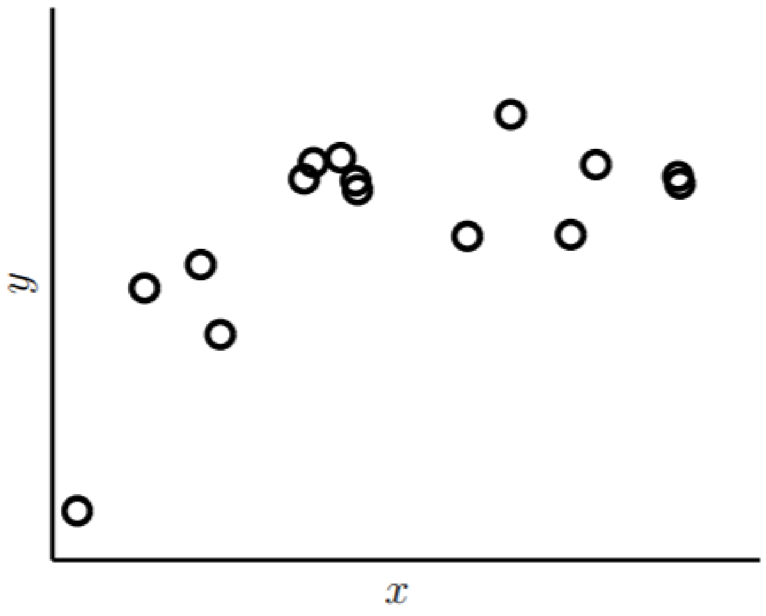
(b)

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

- $\mathcal{H}_2 : E_{in}(\mathcal{H}_2)$, $\mathcal{H}_{10} : E_{in}(\mathcal{H}_{10})$

$E_{in}(\mathcal{H}_2) \geq E_{in}(\mathcal{H}_{10})$

If $E_{out}(\mathcal{H}_{10}) > E_{out}(\mathcal{H}_2)$

g_{10} g_2

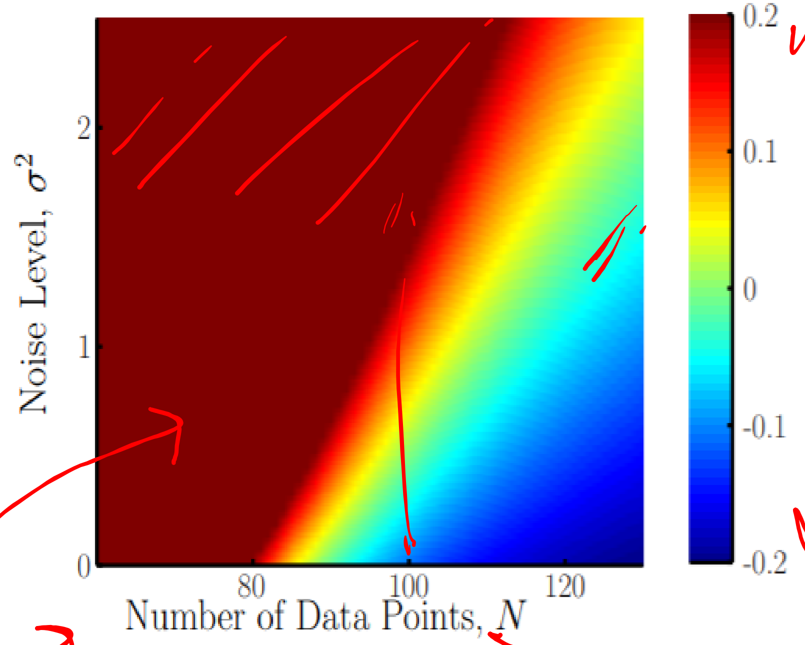
Overfit measure : $\underbrace{E_{out}(\mathcal{H}_{10}) - E_{out}(\mathcal{H}_2)}$ 6^2

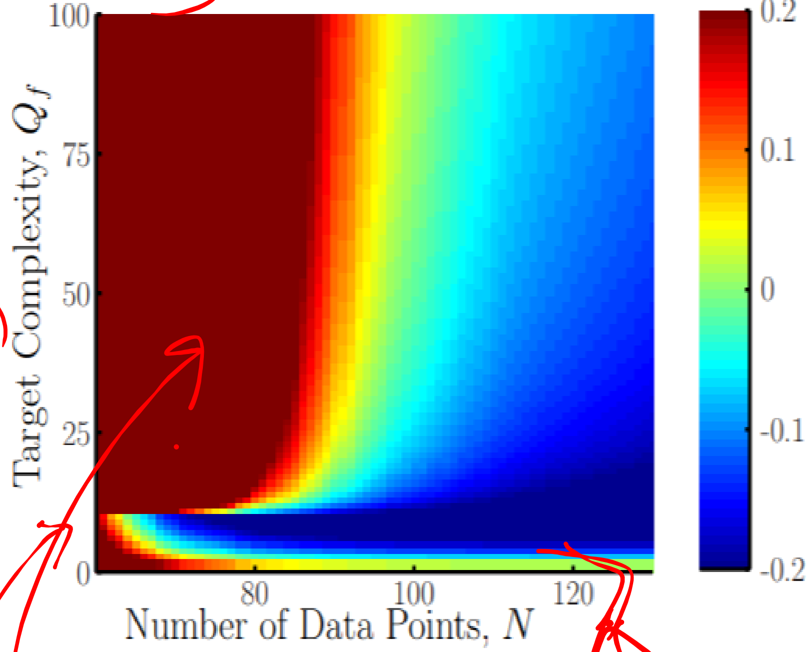
Controlled : i) Noise (measurement error) (ii) N

iii) complexity of f

(degree of poly. Q)

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

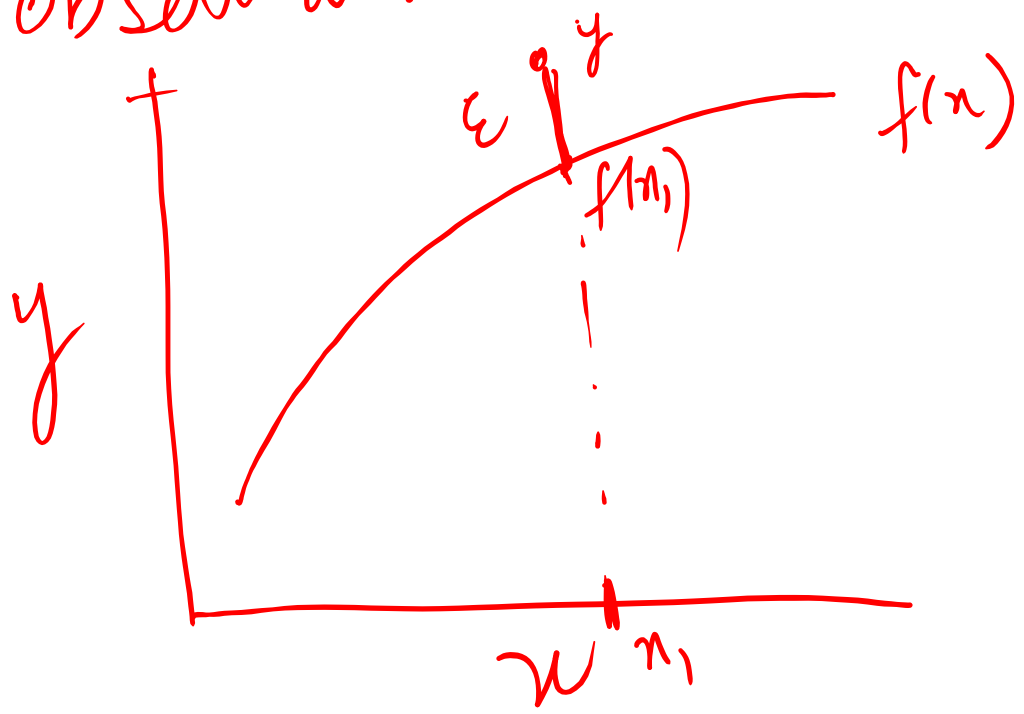
- That part of y which cannot be modelled
  ↳ observed value.

↳ 2 Sources:

i) Stochastic Noise
(measurement error)

$y = f(x) + \varepsilon$

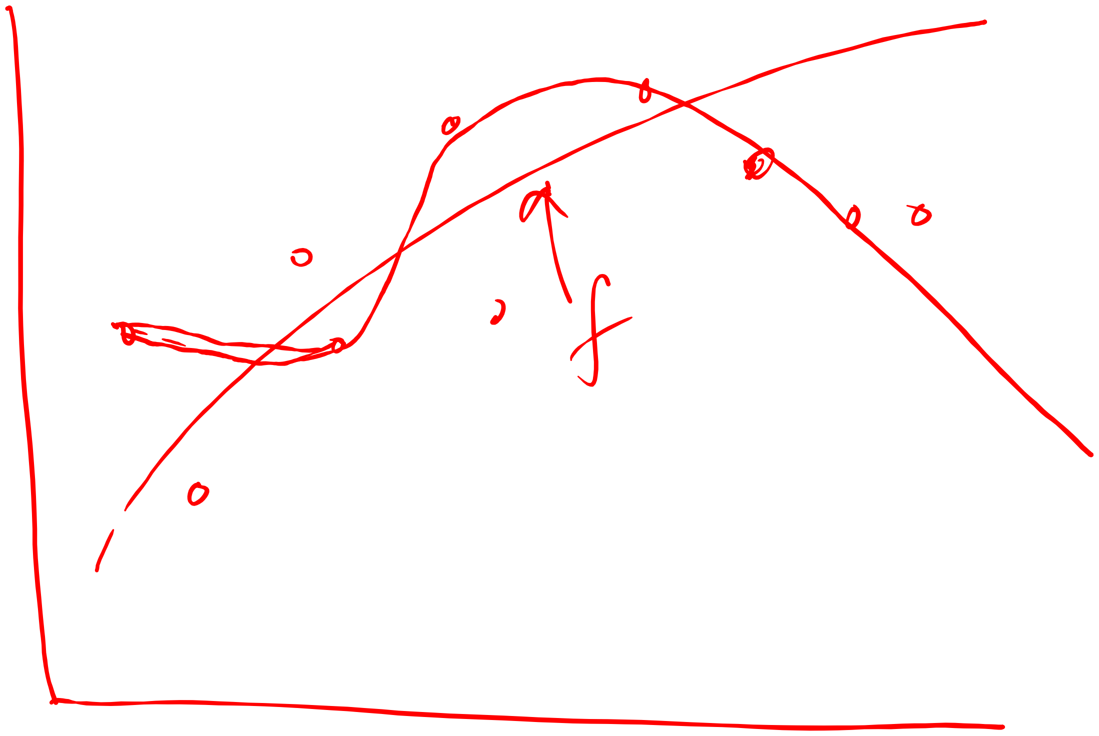
$y = f(x)$



Data with noise

→ higher chances of my model going astray.

Noise → cannot be modelled



Stochastic Noise

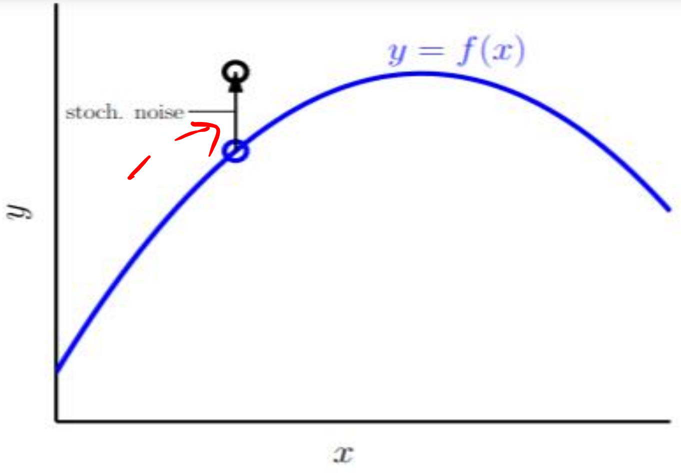
We would like to learn from O:

$$y_n = f(x_n)$$

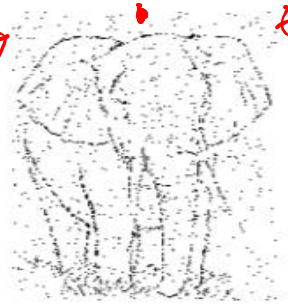
Unfortunately, we only observe O:

$$y_n = f(x_n) + \text{`stochastic noise'}$$

$\uparrow$ no one can model this



Stochastic Noise: fluctuations/measurement errors we $\boxed{\text{cannot model.}}$

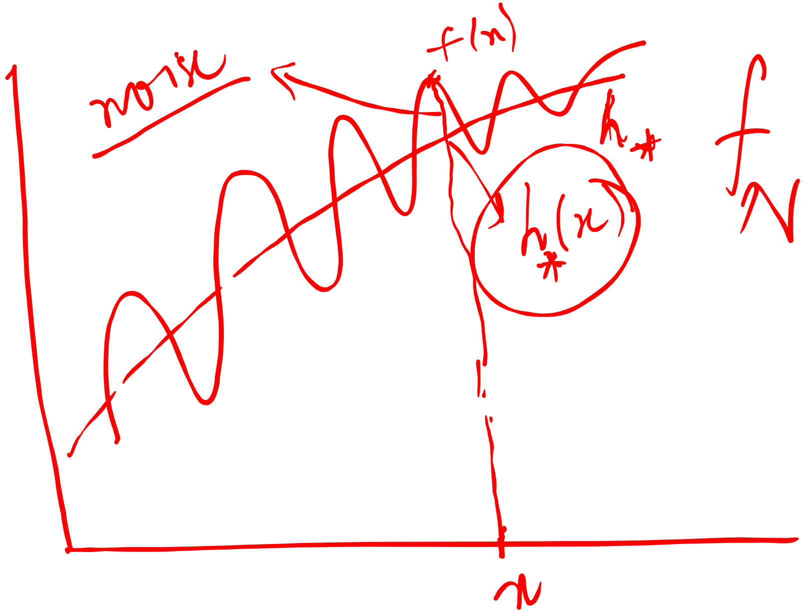


100 pictures

Deterministic Noise

→ find H-set

that part of the (f) that cannot be modelled.



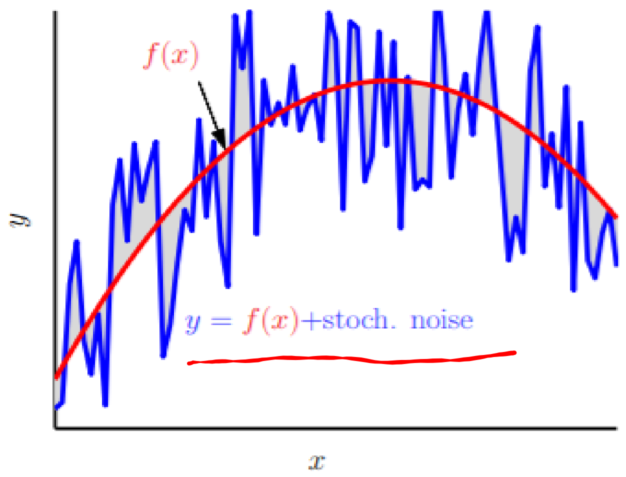
f is complex
No measurement error

Stochastic & Deterministic $\xrightarrow{\text{effect}}$ Overfit.

Regenerate x values $\xrightarrow{S}$ different y values
$\xrightarrow{D}$ same y values.

ii) $S \xrightarrow{\text{deficiency}} \mathcal{H}$
$D \xrightarrow[\text{a complex}]{\text{choose}} \mathcal{H}$ $\}$ $\uparrow$ Trade-off.

Stochastic Noise



source: random measurement errors

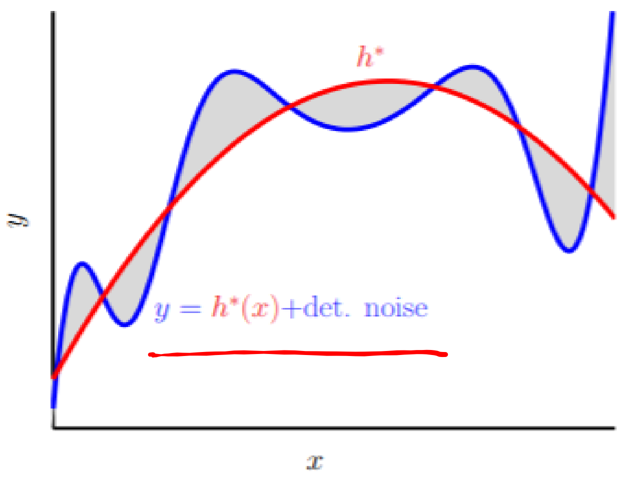
re-measure y_n

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 ○:

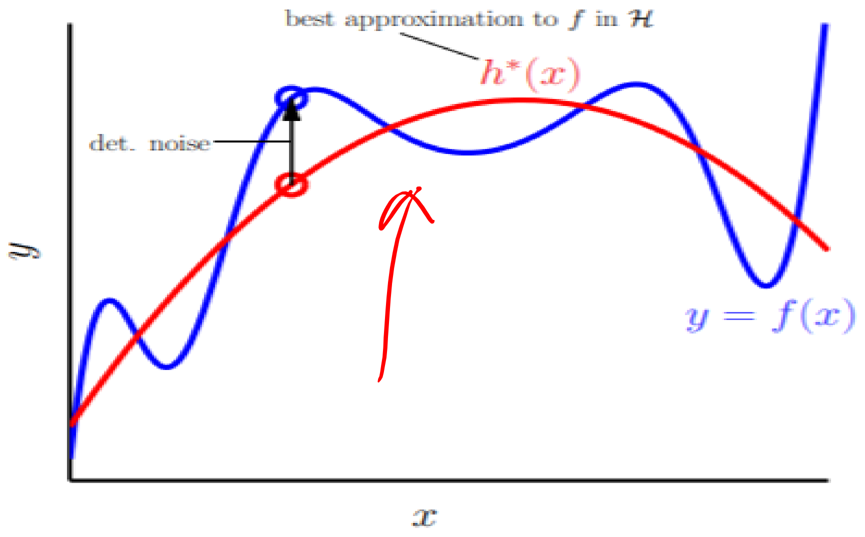
$$y_n = h^*(x_n)$$

Unfortunately, we only observe ○:

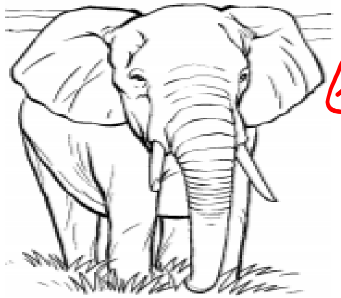
$$\begin{aligned} y_n &= f(x_n) \\ &= h^*(x_n) + \text{`deterministic noise'} \end{aligned}$$

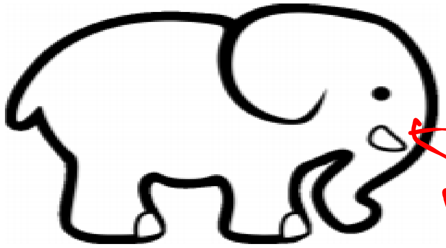
↑

$\mathcal{H}$ cannot model this



Deterministic Noise: the part of f we cannot model.





Bias Variance Analysis and Noise

- $E_D[E_{out}] = \underbrace{Bias(x)}_{(g(x)-f(x))^2} + \underbrace{Variance(x)}_{var(g(x))}$

$$E_{out}(x) = (g(x) - \overset{f(x)}{y})^2, \quad y = f(x) + \varepsilon$$

$$= (g(x) - f(x) - \varepsilon)^2$$

$$= \underbrace{(g(x) - f(x))^2}_{} - 2\varepsilon\, g(x) f(x) \overset{0}{} + \underbrace{\varepsilon^2}_{}$$

$$E_D[E_{out}(x)] =$$

$$\underline{E\left[E_{out}(x)\right]} = \underbrace{E\left(g(x)-f(x)\right)^2} + \underbrace{E\left(\varepsilon^2\right)}$$

$$= \text{Bias} + \text{Variance} + \sigma^2$$

$(x \,\&\, N)$ $\quad N \uparrow$

$$E\left[E_{out}\right] = \text{Bias} + \underbrace{\text{Variance}}_{\text{indirect impact of noise}} + \sigma^2 \longrightarrow \text{stochastic noise}$$

deterministic noise ✓

stochastic noise ✓

Model $\longrightarrow$ <u>lower bias, lower variance</u>

Summarize

Overfitting → known problem.

Noise → cannot be modelled, always there.

Can we do something about this?

→ Yes → <u>Regularization.</u>

Validation

Summary

•

Thanks!