Bias and Variance: A Recap

SUMMARY OF PREV. LECT.

- How well $g \in \mathcal{H}$ approximate f ? $\rightarrow$ BIAS
- How to find best $g \in \mathcal{H}$? $\rightarrow$ VARIANCE

Average hypothesis $\bar{g}(x) \in \mathcal{H}$, $\bar{g}(x) = \mathbb{E}_{\mathcal{D}}\left[g^{\mathcal{D}}(x)\right]$

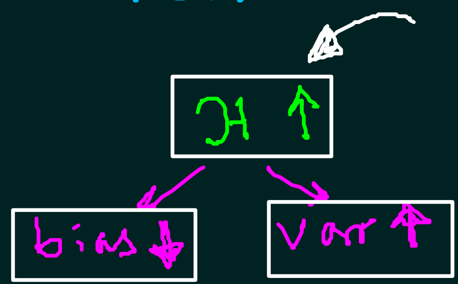
$$\left[\underbrace{\mathbb{E}_x\left[f(x)\right]}_{} = \underbrace{\int_a^b x \cdot f(x)\, dx}_{}\right]$$

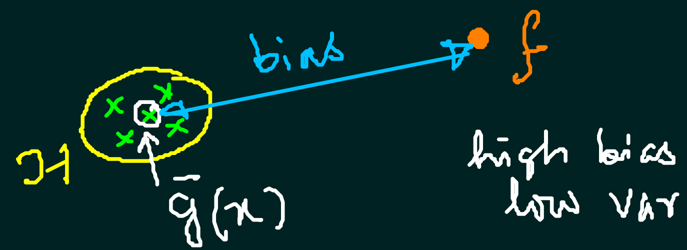
$$\mathbb{E}_{\mathcal{D}}\left[E_{out}(g)\right] = \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(x) - f(x)\right)^2\right]$$

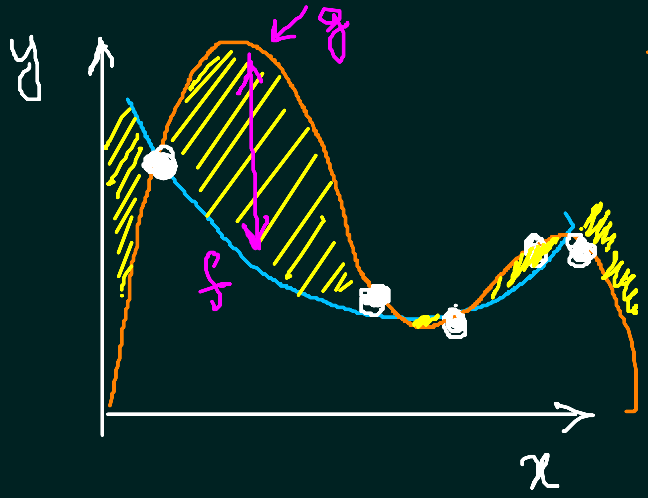
$$= \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(x) - \bar{g}(x)\right)^2\right]}_{\text{var}(x)} + \underbrace{\left(\bar{g}(x) - f(x)\right)^2}_{\text{bias}(x)}$$

Hence,

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







→ $E_{in} \approx 0$, 5 points $\Rightarrow \mathcal{H} = \{4^{th}$ order polynomial$\}$

↳ E_{out} → bad !!

→ Overfitting & bad generalisation

$\boxed{\text{NOISY DATA}}$

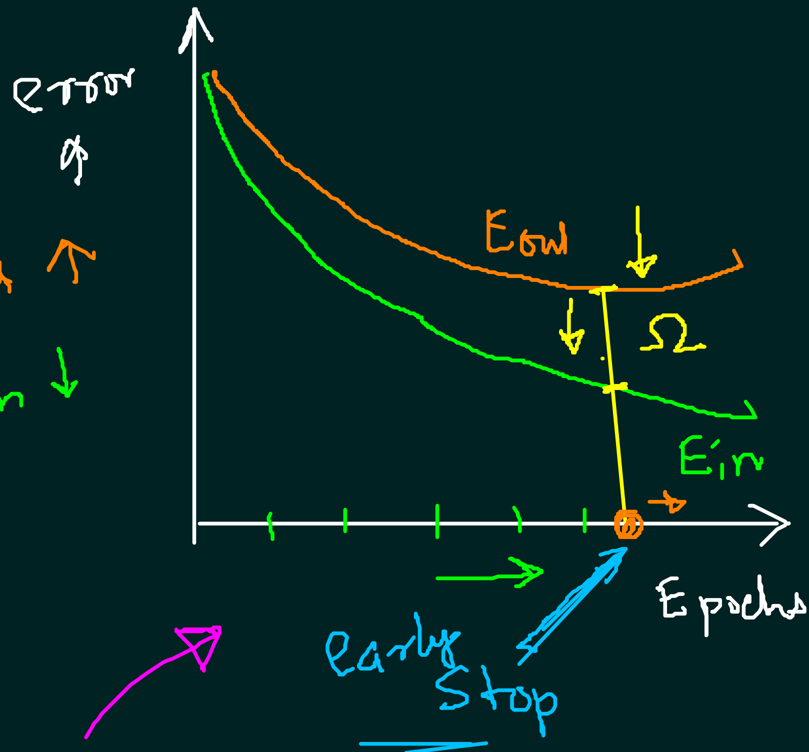
→ hypothesis

fit the $f(x)$ target

fit the $\approx \epsilon(x)$ noise

$E_{out} \uparrow$

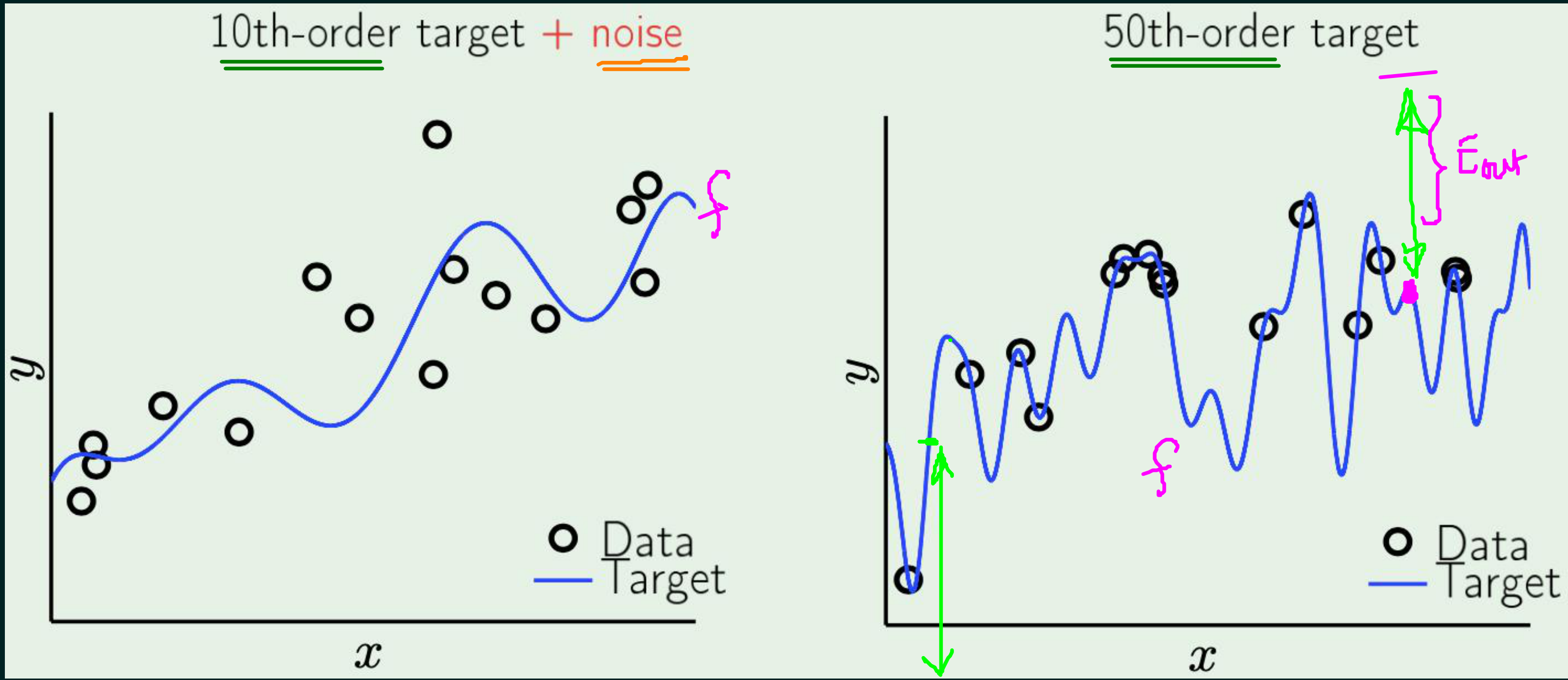
$E_{in} \downarrow$

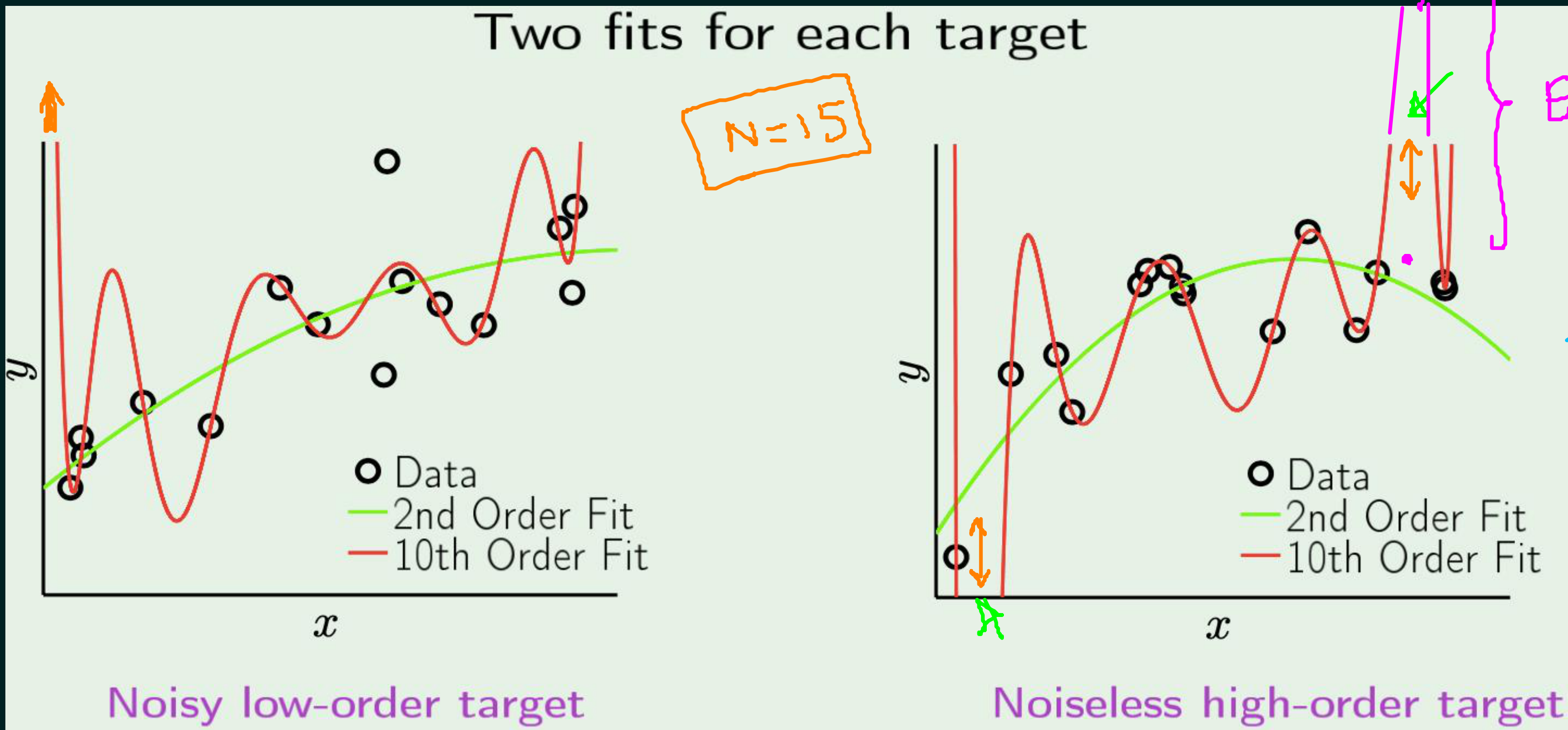


→ Overfitting: $g_1(x) \rightarrow g_2(x)$

$E_{in} \downarrow$ ✓ and $E_{out} \uparrow$ ✗

g_2 overfits





$\mathcal{H}_2 = \{ \ldots\ldots \}$

$\mathcal{H}_3$

$\mathcal{H}_{10} = \{ \ldots\ldots \}$

$\mathcal{H}_7$

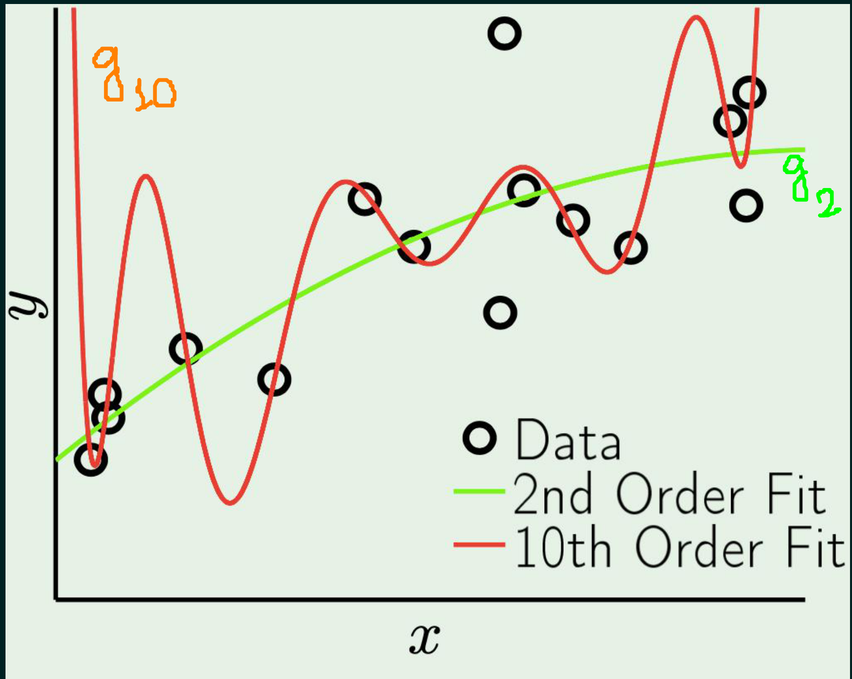
Perf. of Noisy

	2nd	10th
Ein	0.5	0.034
Eout	0.127	9.00

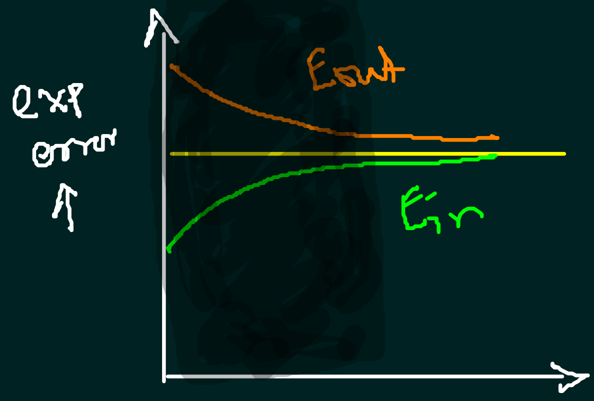
Overfit by $\mathcal{H}_{10}$

Perf. of Noiseless

	2nd	10th
Ein	0.029	10^{-5}
Eout	0.12	7680



Learning a 10th-order target

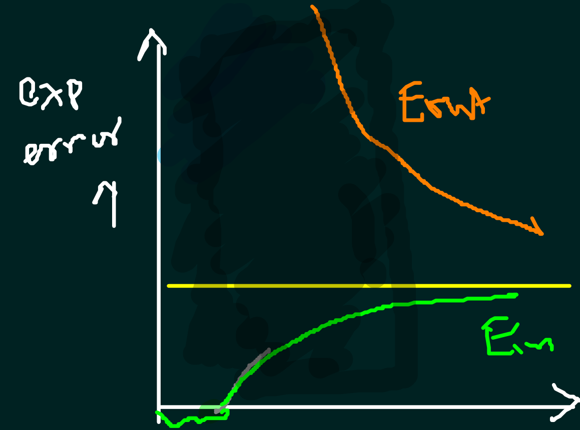


→ N

Simple Model

$\mathcal{H}_2$

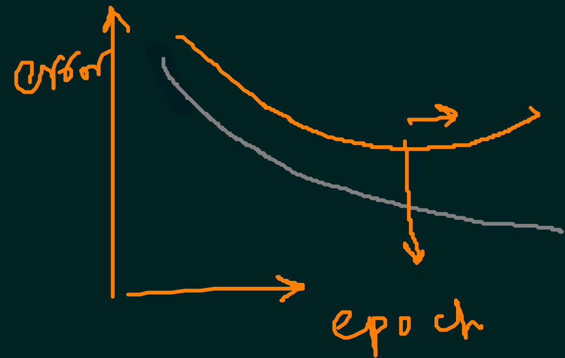
Stochastic Noise

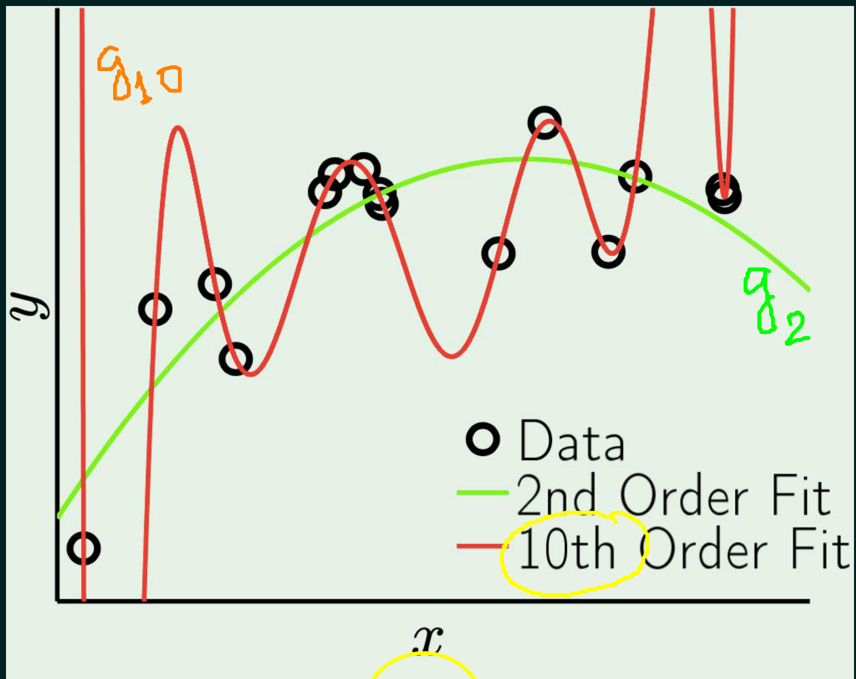


→ N

Complicated Model

$\mathcal{H}_{10}$



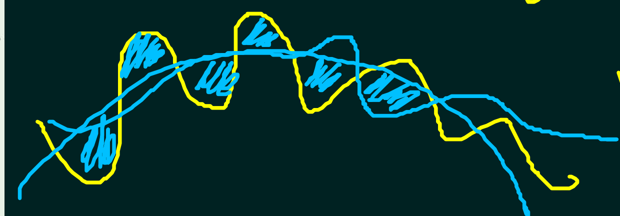


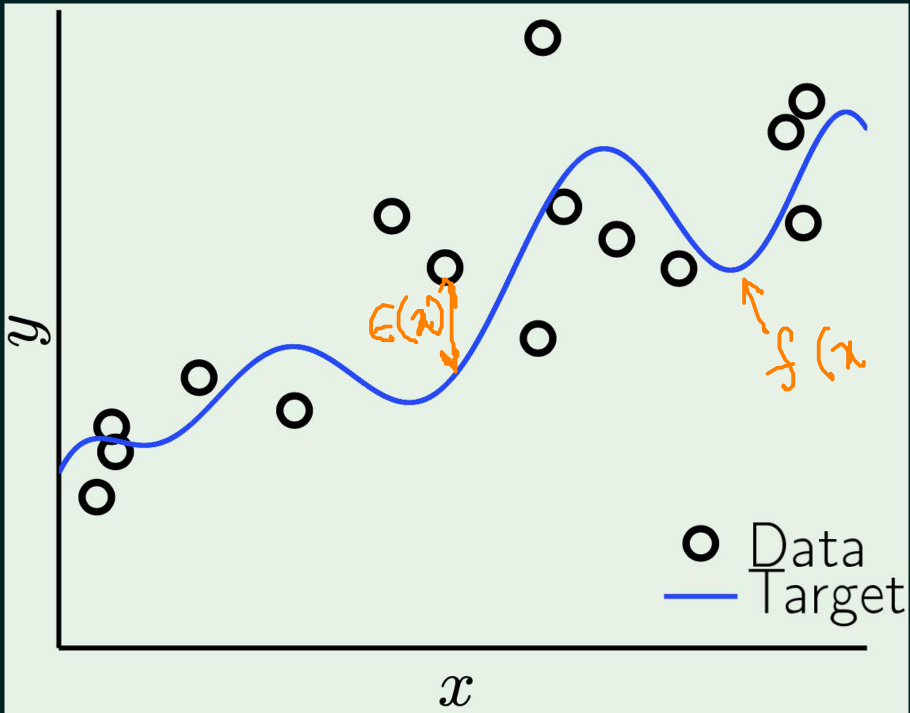
Learning a 50th-order target

Deterministic Noise

Is it really NO noise?

↳ trying to fit higher order poly by approx. with lower order hyp





$$y = f(x) + \underbrace{\epsilon(x)}_{\sigma^2}$$

$$= \underbrace{\left[\sum_{q=0}^{Q_f} \underline{a_q} x^q\right]}_{\text{Normalize (unity)}} + \epsilon(x)$$

noise level $= \sigma^2$

$\boxed{N}$

Overfitting Measure:

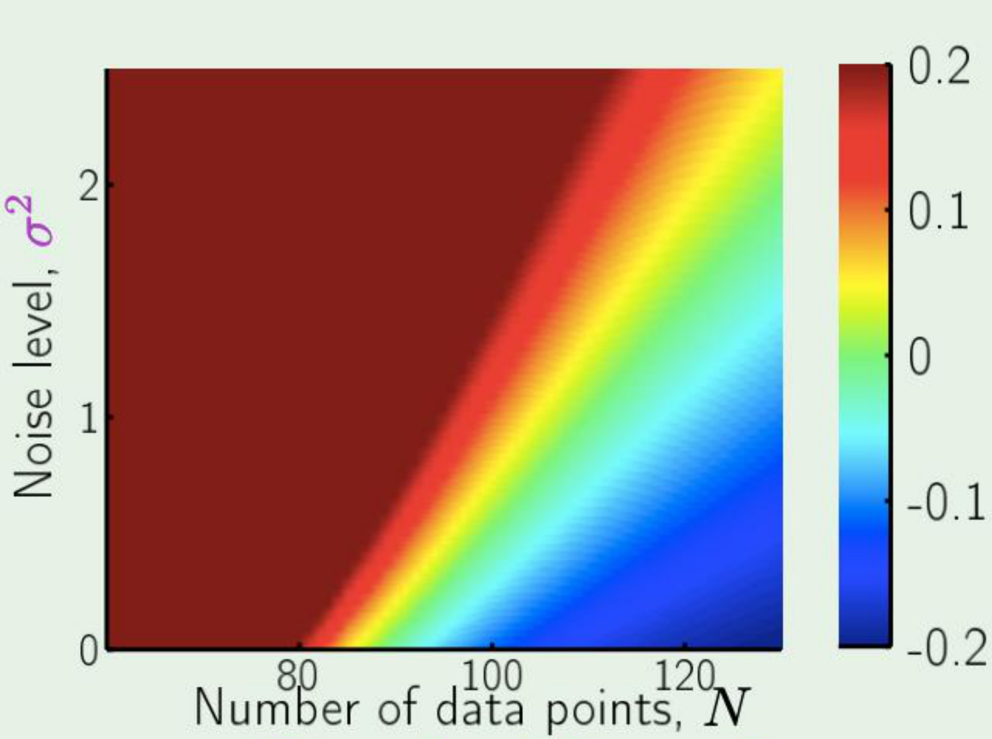
$$\left[E_{out}(g_{10}) - E_{out}(g_2)\right]$$

$$\left[\underbrace{P_N(x) \cdot P_\sigma(x)}\right]$$

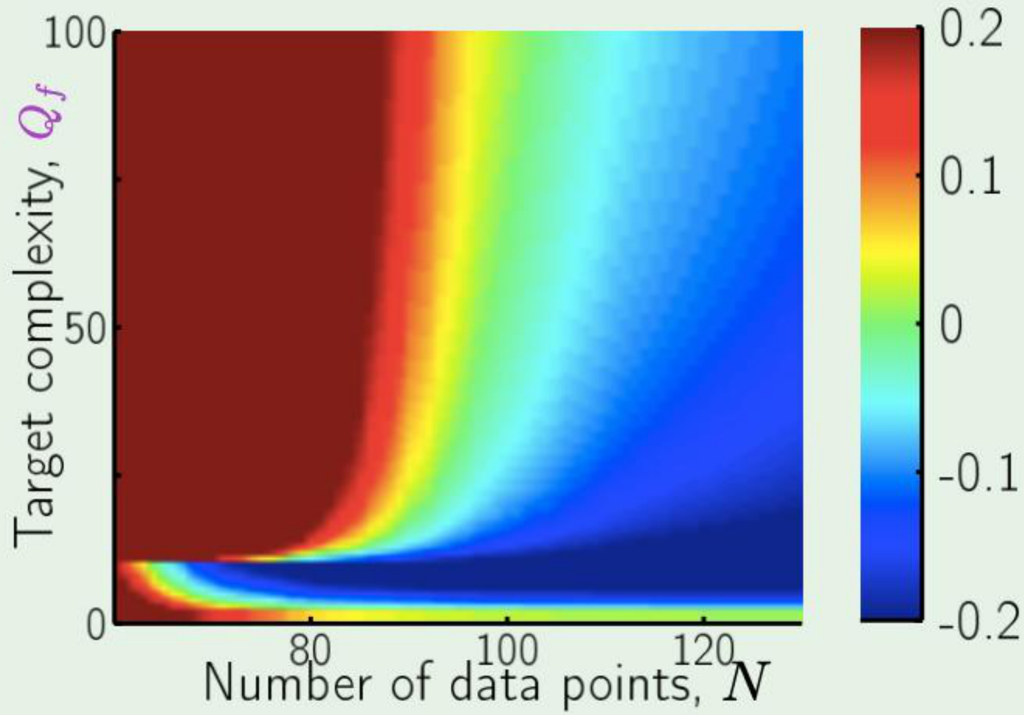
$[g_{10} \in \mathcal{H}_{10}$ and $g_2 \in \mathcal{H}_2]$

[Stochastic Noise]

Impact of "noise"



Stochastic noise



Deterministic noise

# of TE ↑ ⇒ Overfitting ↓

Stochastic Noise ↑ ⇒ overfit ↑

Deterministic Noise ↑ ⇒ overfit ↑

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

y

Assume: $y = f(x) + \epsilon(x)$

where $\mathbb{E}_{\epsilon}\left[\epsilon(x)\right] = 0$

$+\bar{g}(x) - \bar{g}(x)$

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

$$= \mathbb{E}_{\mathcal{D},\epsilon}\left[\left(g^{\mathcal{D}}(x) - \bar{g}(x)\right)^2 + \left(\bar{g}(x) - f(x)\right)^2 + \left(\epsilon(x)\right)^2 + \cancel{\text{cross terms}}\right]$$

$\int_a^b x f(x)\,dx$

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

deterministic noise

Stochastic Noise

Ex: $f(x) = x + \cancel{1}$

$E[f(x)] \int_{r_1}^{r_2} x(x+1)\,dx$

How to deal with overfitting

→ Put brakes (1%)

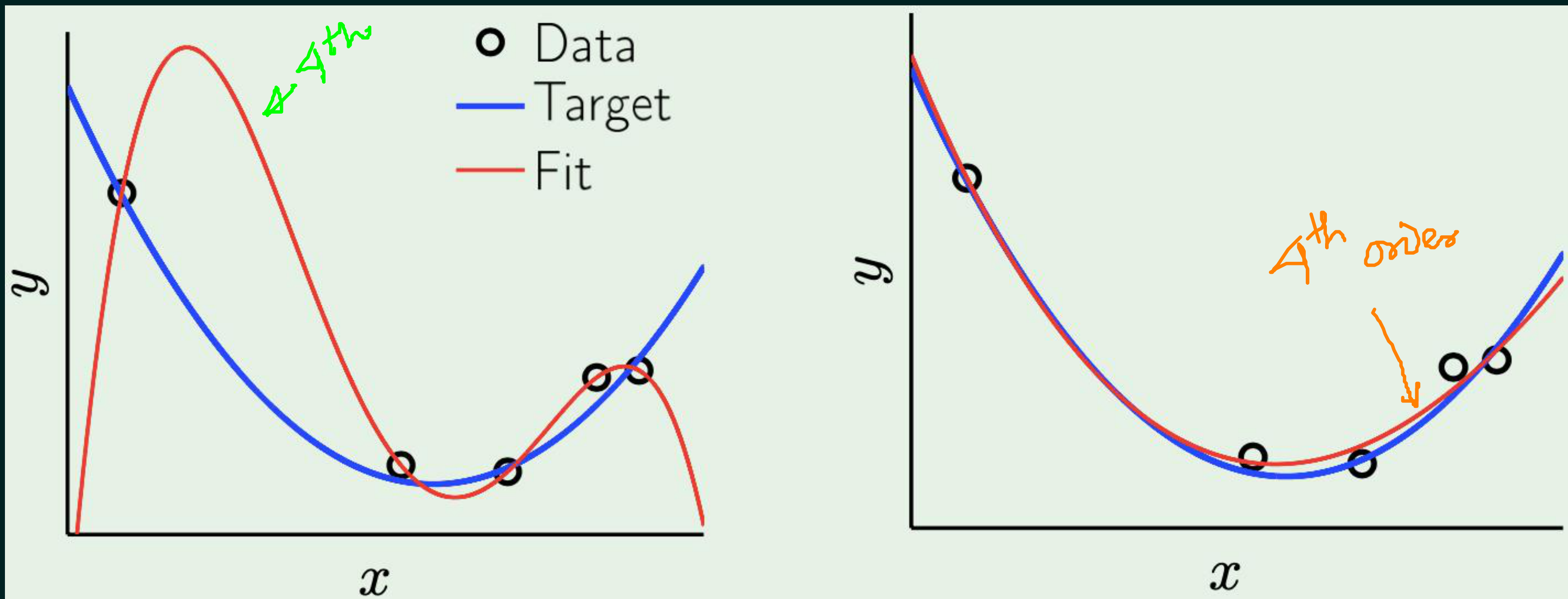
$$\sum_{i=0}^{Q} a_i x^i$$

$\delta \rightarrow (a_i \pm \delta)$

unrestrained fit

(Regularization)

restricted fit



① Regularization

Put brake

② Validation

Check bottomline