Noise and Its Impact On Overfitting

Anahita Zarei, Ph.D.

Overview

- Role of Noise in Overfitting
- Reading: 4.1 Learning From Data

Motivation Question:

- You’re given 15 sample points that were generated from a **noiseless** 50th order polynomial. (i.e. all sample points fall exactly on the 50th order target function)
- Which one of the following models will you try to fit to these points:
  A) 2nd order
  B) 10th order
  C) 50th order

Overfitting - Background

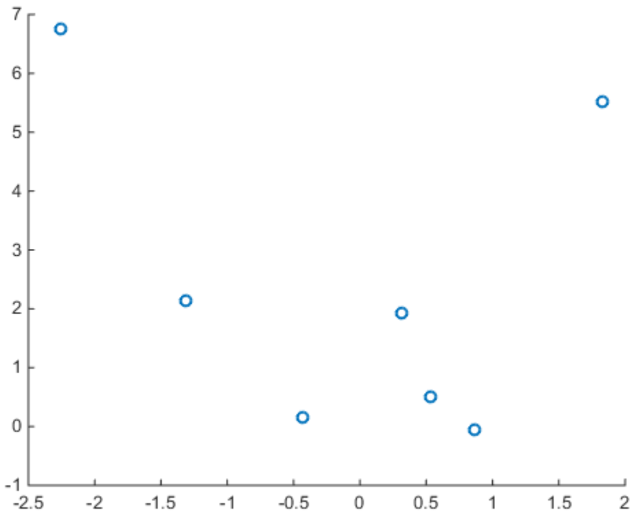
- Paraskavedekatriaphobia
- Superstitions are examples of the human ability to overfit:
  - Unfortunate events are memorable.
  - Given a few such memorable events, people try to find an explanation.
  - Will there be more unfortunate events on Friday the 13th's than any other day in the future?
- Overfitting is the phenomenon where fitting the observed facts no longer indicates a better out-of-sample performance.
- E.g. when a model uses its additional degrees of freedom to fit noise in the data, resulting in a final hypothesis that is inferior.

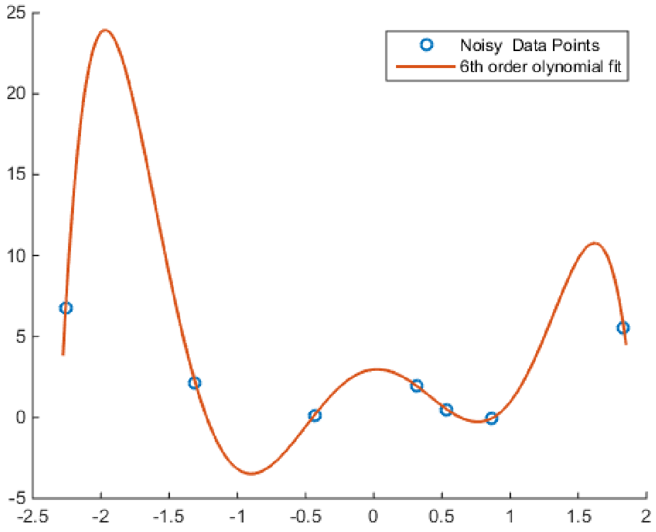
When does overfitting occur?

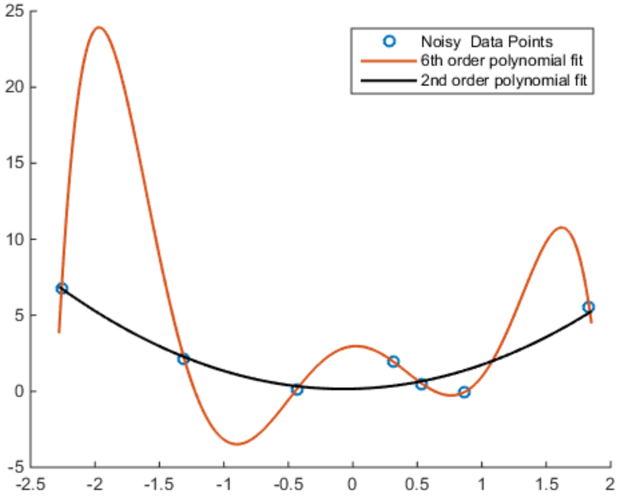
- Overfitting occurs when you **fit the data more than is justified.**
- Overfitting can happen in any type of models (linear/logistic regression, multiple layer neural network, etc.)
- The main case of overfitting is when a hypothesis with **lower in-sample** (training) error results in **higher out of sample** (test) error.
- The underlying reason for overfitting is *noise*.

Simple Example

- Consider a simple one dimensional regression problem with **7 data** points. What degree polynomial will you consider to minimize the training error?
- A **6-th order** polynomial fit will result in zero in-sample error, however the test error will not be promising.
- How about a 2nd order fit?
- Note that the underlying target function was indeed a 2nd order, though data points were a bit noisy.

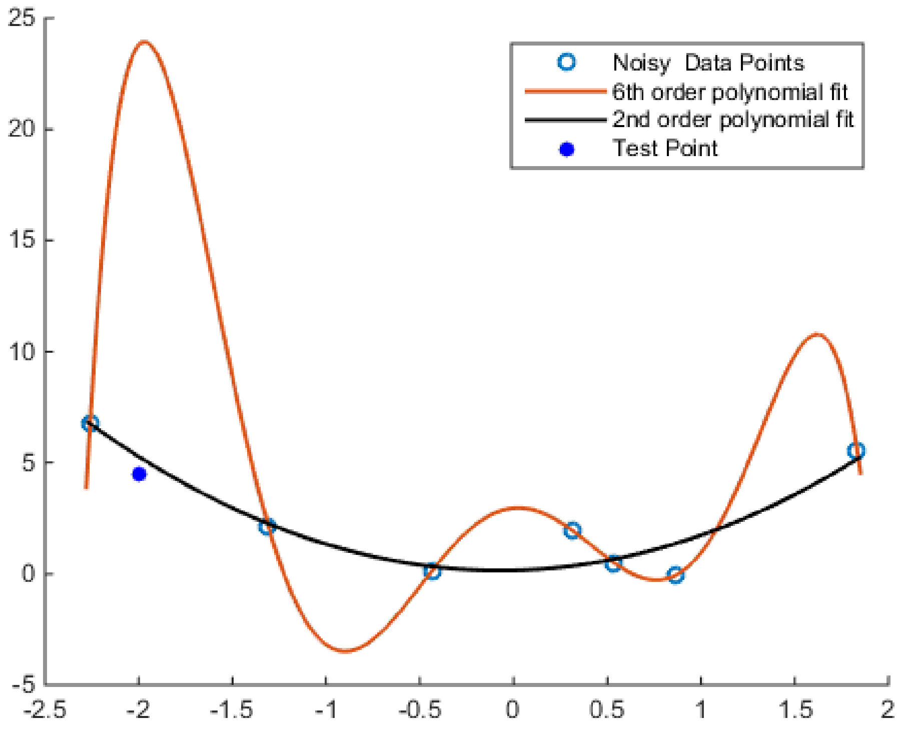






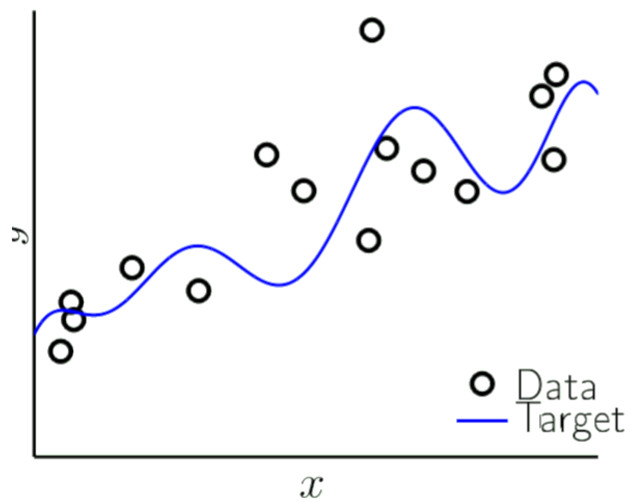
Why is overfitting harmful?

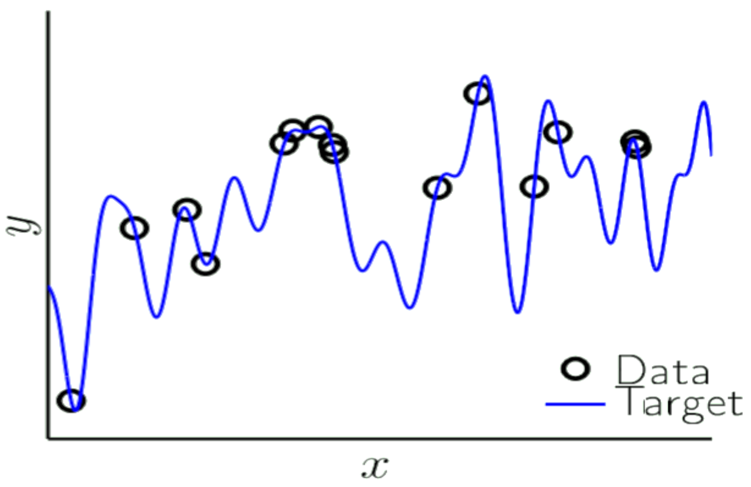
- When the model fits the in-sample noise, it **extrapolates** for the out of sample data.
- This means that it's actually **taking you away** from the correct solution.
- It thinks of an imaginary pattern and this will end-up hurting the performance.



A Case Study:
Understanding the Role of Noise

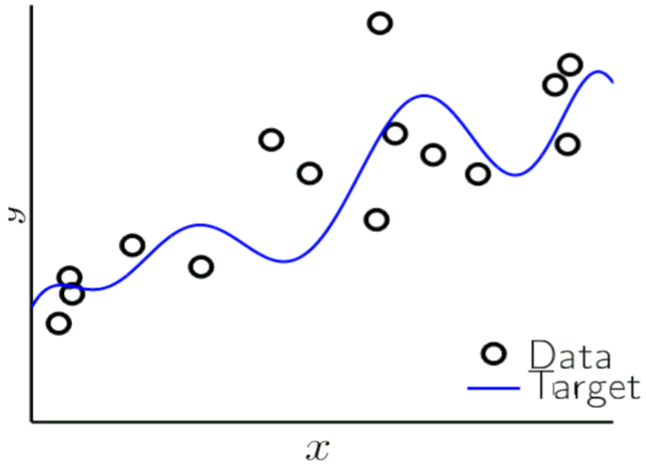
- We use overfitting with polynomial regression to gain a better understanding of when overfitting occurs.
- Given 2 cases, each with 15 data points, provide the *best* fit.
  - Case 1: Simple target (10th order), but noisy data
  - Case 2: Complex target (50th order), but noiseless

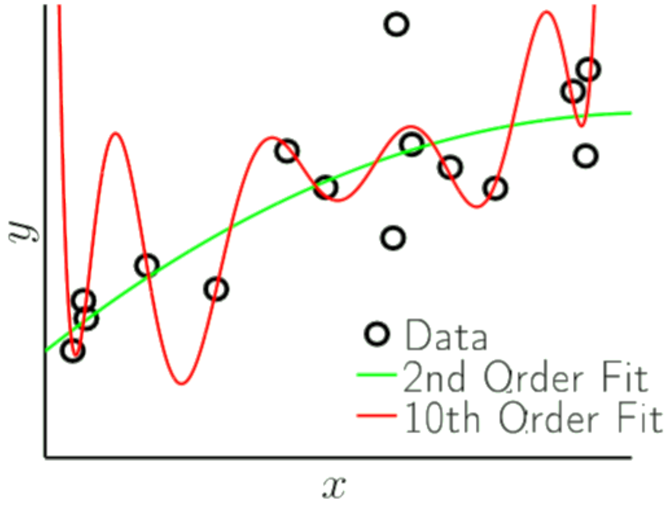

10th-order target + noise


50th-order target

Case Study: 2nd and 10th Order Fits

10th-order target + noise

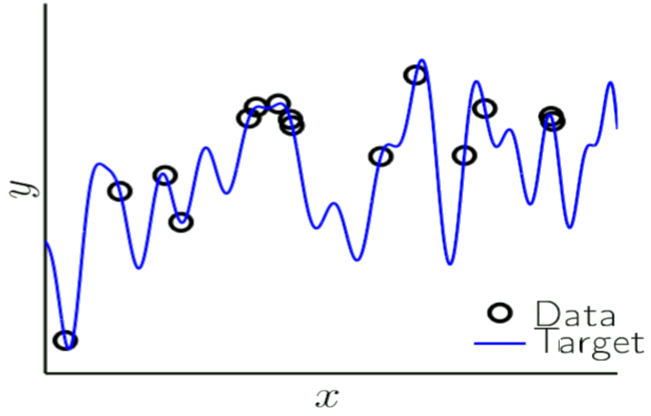


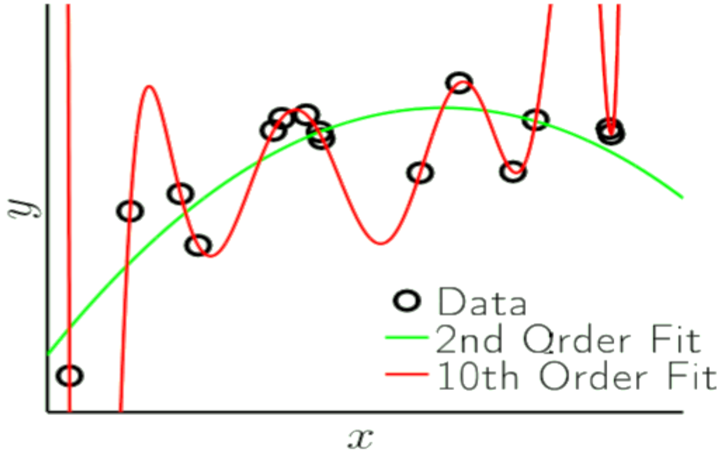


Noisy low-order target

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

50th-order target



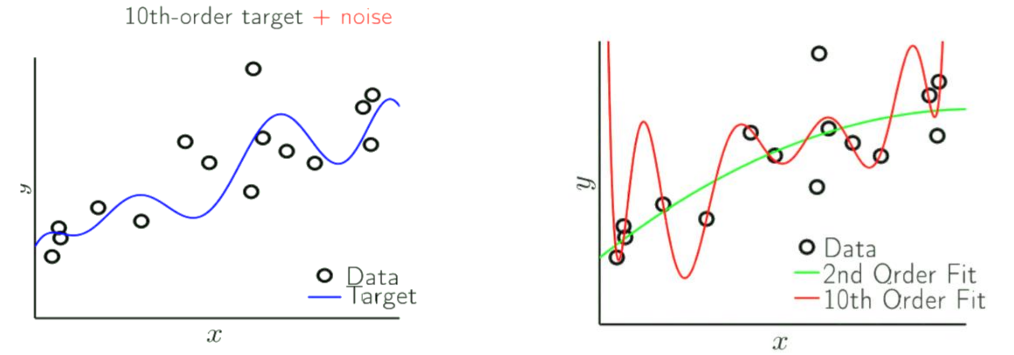


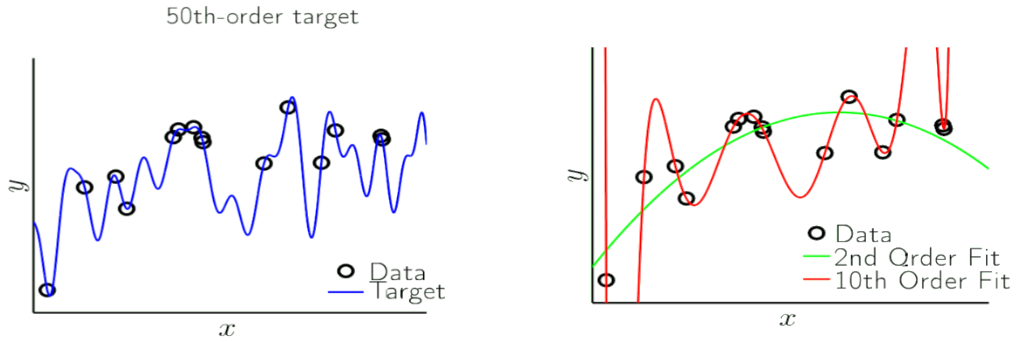
Noiseless high-order target

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

Analysis of Case Study

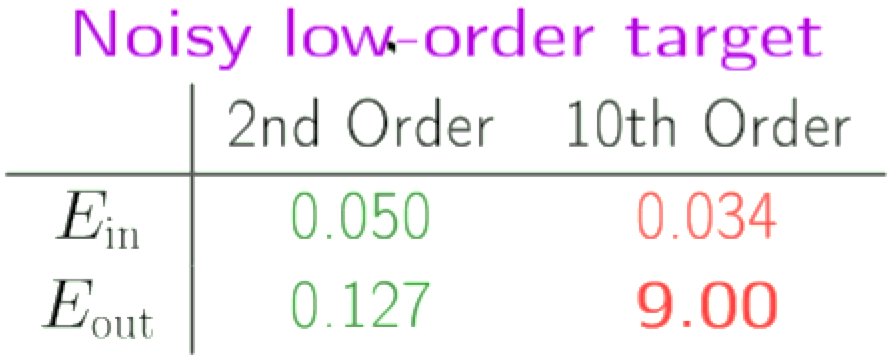
- In both cases, the 10th order polynomials heavily overfit the data and don’t resemble the true function.
- The 2nd order fits do not capture the full nature of the true function either, but they capture its general trend, resulting in smaller E_{out}.
- In both cases, the 10th order has a lower E_{in} and higher E_{out}, indicating a case of overfitting.
- Keep in mind that the algorithm has only the data points to work with, has no knowledge of the true function.





Does Incorporating Information About the Target Function Improve Results?

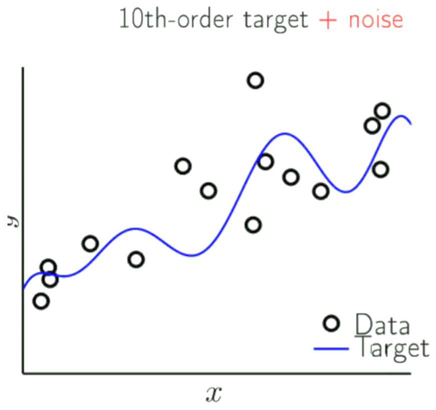
- One of the surprising facts in the case study is that the **simpler** model **knowingly** gives up the ability to implement the true target function, and yet has a **better** performance (i.e. smaller E_{out})
- The conclusion is even if we DO know the order of the target and naively incorporate this knowledge in creating the model, the performance is inferior compared to the stable 2^{nd} order model.
- Fit a model that matches your data resources not target function complexity.

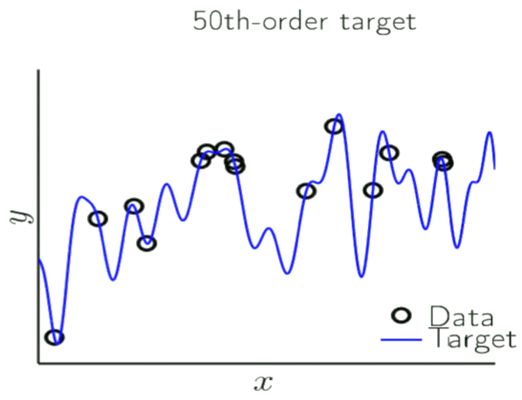


Noisy low-order target

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

Does the Simple Model Always Win?

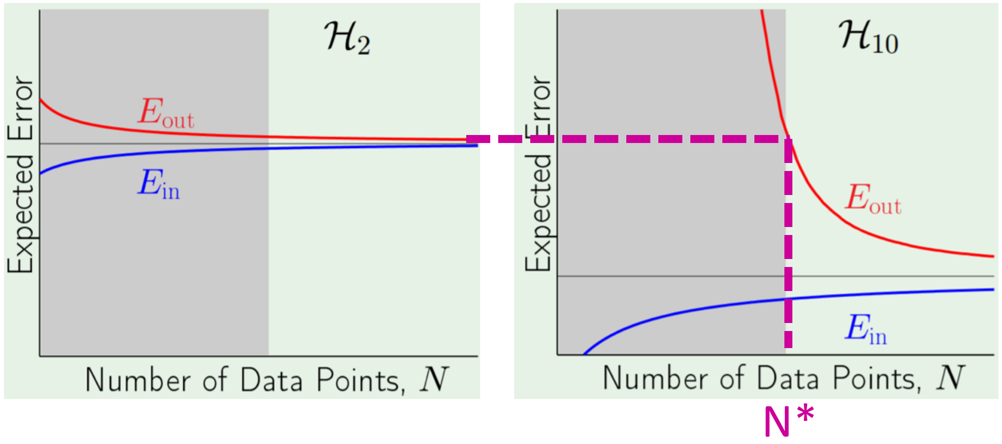




In the case of simple target function, if the data were noiseless, the 10th order fit will be able to create the exact target function from 15 data points. i.e. complex hypothesis will win.

In the case of (noiseless) complex target function, once the data resources increase beyond N*, the complex model will win.

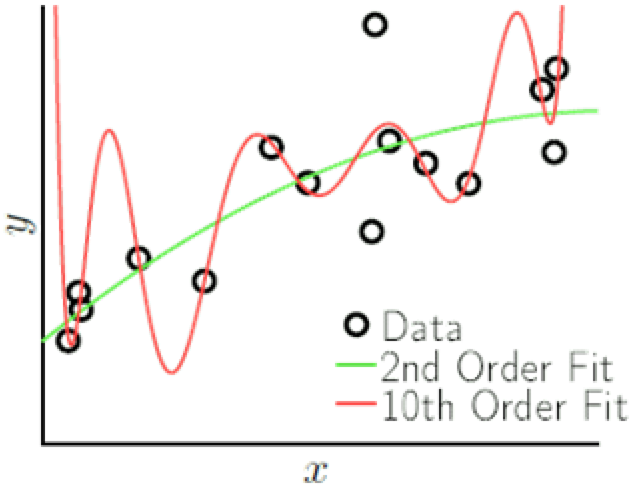
The grey area corresponds to overfitting. For those N, complex model has lower E$_{in}$, but higher E$_{out}$ compare to simple model.



The Overfit Measure

We fit the data set $(x_1, y_1), \cdots, (x_N, y_N)$ using our two models:

$\mathcal{H}_2$: 2nd-order polynomials $\mathcal{H}_{10}$: 10th-order polynomials



Compare out-of-sample errors of

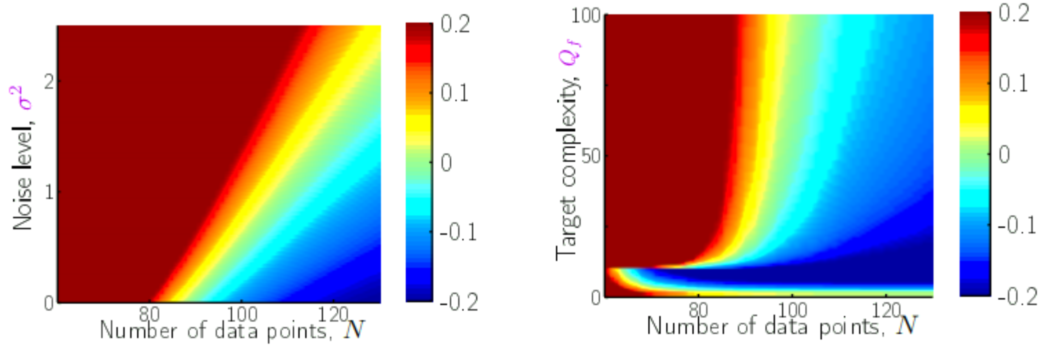
$g_2 \in \mathcal{H}_2$ and $g_{10} \in \mathcal{H}_{10}$

overfit measure: $E_{\text{out}}(g_{10}) - E_{\text{out}}(g_2)$

- If overfit measure is positive, then there is overfit.
- If it’s negative, it means the complex model is better and therefore there’s no overfit.
- If it’s zero, it mean both models are the same.

Role of Noise in Overfitting

- The colors map to the level of overfitting, the redder the worse the overfitting.
- The left figure shows that there is less overfitting when the noise level drops or when the number of data points N increases.
- The right figure shows that target function complexity affects overfitting in a similar way to stochastic noise.

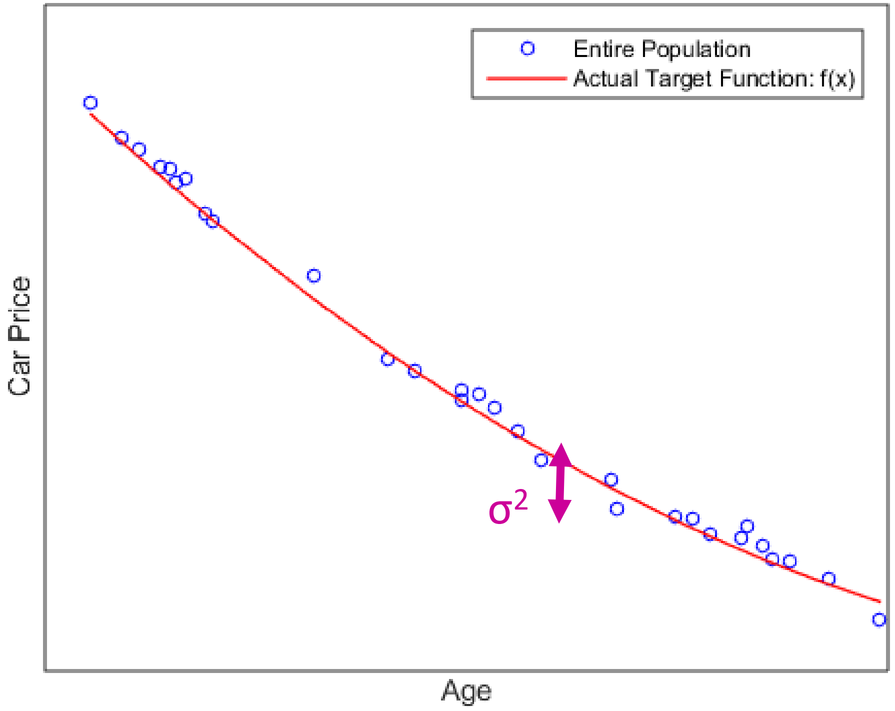


Stochastic noise Deterministic noise

number of data points	↑	Overfitting	↓
stochastic noise	↑	Overfitting	↑
deterministic noise	↑	Overfitting	↑

Types of Noise: Stochastic

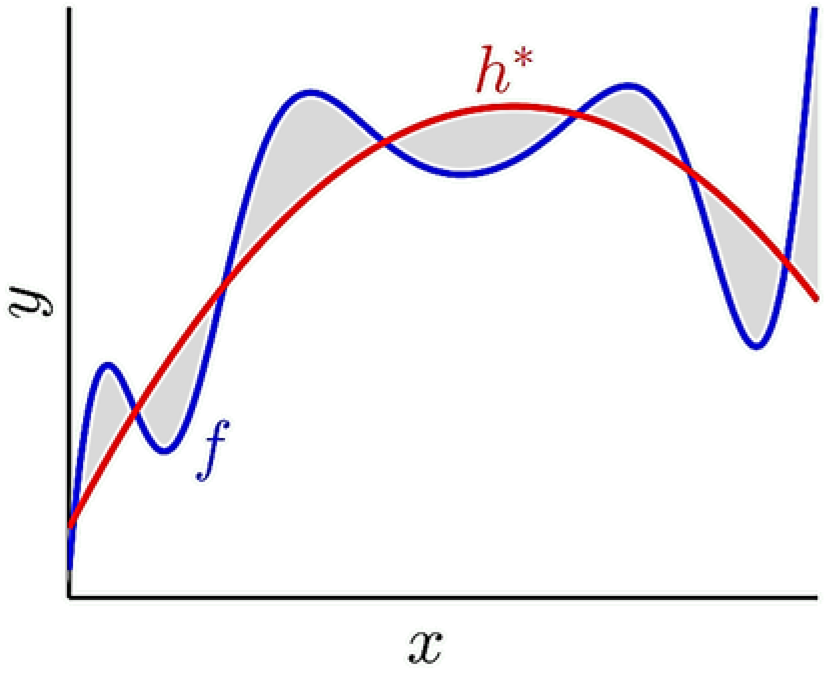
- **Stochastic noise** is the irreducible inherent noise in the data that has nothing to do with the fit of our model.
- You can not reduce this noise by choosing a better hypothesis.
- For example one can assume that there is an underlying true relationship between the price of a car and its age (f(x)).
- However, there are other contributing factors that can’t be captured in f(x), like how a customer feels about a particular car or a salesperson when walking into a dealership and purchasing a car. i.e. there is not a perfect description between x and y.
- The behavior that can’t be captured in f(x) is the inherent random noise, represented by ε.
- We assume that noise has expected value of 0 and at any x (e.g. Age) has a spread of σ^2.



y=f(x)+ε(x)

Types of Noise: Deterministic

- For a given learning model, there is a best approximation to the target function.
- The part of the target function outside this best fit acts like noise in the data.
- We call it deterministic noise to differentiate fit from the stochastic noise.
- **Deterministic Noise:** The inability of the model to capture part of the target function due to limited data.



Mathematical Decomposition of Error

- We now modify our decomposition expression to show that there are three different sources of error: Stochastic Noise, Deterministic Noise, and Variance.
- $Error\ at\ x = MSE\ at\ x = E_D\left[\left(g^D(x) - f(x)\right)^2\right] = var(x) + bias^2(x)$ (if there’s no random noise).
- $Error\ at\ x = E_{D,y}[(g^D(x) - y)^2] = E_{D,y}[(g^D(x) - f(x) + f(x) - y)^2]$=
- $E_{D,y}\left[\left(g^D(x) - f(x)\right)^2\right] + 2E_{D,y}\left[\left(g^D(x) - f(x)\right)(f(x) - y)\right] + E_{D,y}[(f(x) - y)^2] =$
- $E_D\left[\left(g^D(x) - f(x)\right)^2\right] + 2E_{D,y}\left[\left(g^D(x) - f(x)\right)\varepsilon\right] + E_y[\varepsilon^2] =$
- $E_D\left[\left(g^D(x) - f(x)\right)^2\right] + 2E_D\left[\left(g^D(x) - f(x)\right)\right]\cancel{E_y}[\varepsilon] + \sigma^2(x) = var(x) + bias^2(x) + \sigma^2(x)$
- $E_X[var(x) + bias(x)^2 + \sigma^2(x)] = var + bias^2 + \sigma^2$

Error Decomposition and Noise

- Recall that $\mathrm{error} = \sigma^2 + bias^2 + var$
- The first two terms reflect the direct impact of the stochastic and deterministic noise:
    - The variance of stochastic noise is σ^2
    - The bias is directly related to deterministic noise in that it captures the model's inability to approximate target function.
- The variance term is indirectly impacted by both types of noise capturing a model's susceptibility to being led astray by the noise.
- $error = stochastic\ noise + deterministic\ noise + var$

Motivation Question:

- You're given 15 sample points that were generated from a **noiseless** 50th order polynomial. (i.e. all sample points fall exactly on the 50th target function)
- Which one of the following models will you try to fit to these points:
  - A) 2nd order
  - B) 10th order
  - C) 50th order

References

Learning From Data by Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin