Noise, Overfitting, and Bias vs Variance

Hamid Dehghani
School of Computer Science
Birmingham
September 2020

Slides adapted from Iain Styles, School of Computer Science

UNIVERSITY OF BIRMINGHAM

Intended Learning Outcome

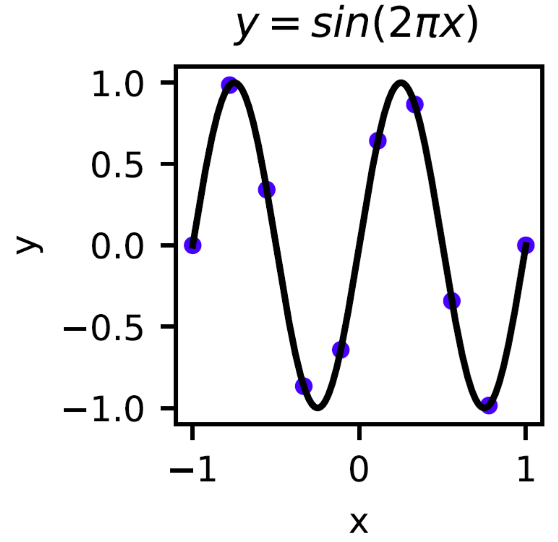
- Understand the effect of noise on machine learning problems
- Understand and explain the concepts of over and underfitting
- Be able to explain these concepts using the idea of bias-variance decomposition

Model choice

- Remember
  - We want to model the relationship between x and y as a mathematical function
    - $f(w, x)$
  - If we know model, we can use that
    - $y(x) = \sin(2\pi x)$;
  - But often we do not!
    - We will use a ‘representation’ that describes the data well, which we called a basis

Visual example

- Keeping to linear models will help us explore the details
- $y(x) = \sin(2\pi x)$;
  - So $\mathrm{f}(\mathbf{w}, \mathrm{x}) \; = w_0 \sin(2\pi x)$; would be a trivial choice
- Assume we have no knowledge about model
  - $\mathrm{f}(\mathbf{w}, \mathrm{x}) \; = \sum_{i=0}^{M-1} w_i x_i$



UNIVERSITY OF BIRMINGHAM

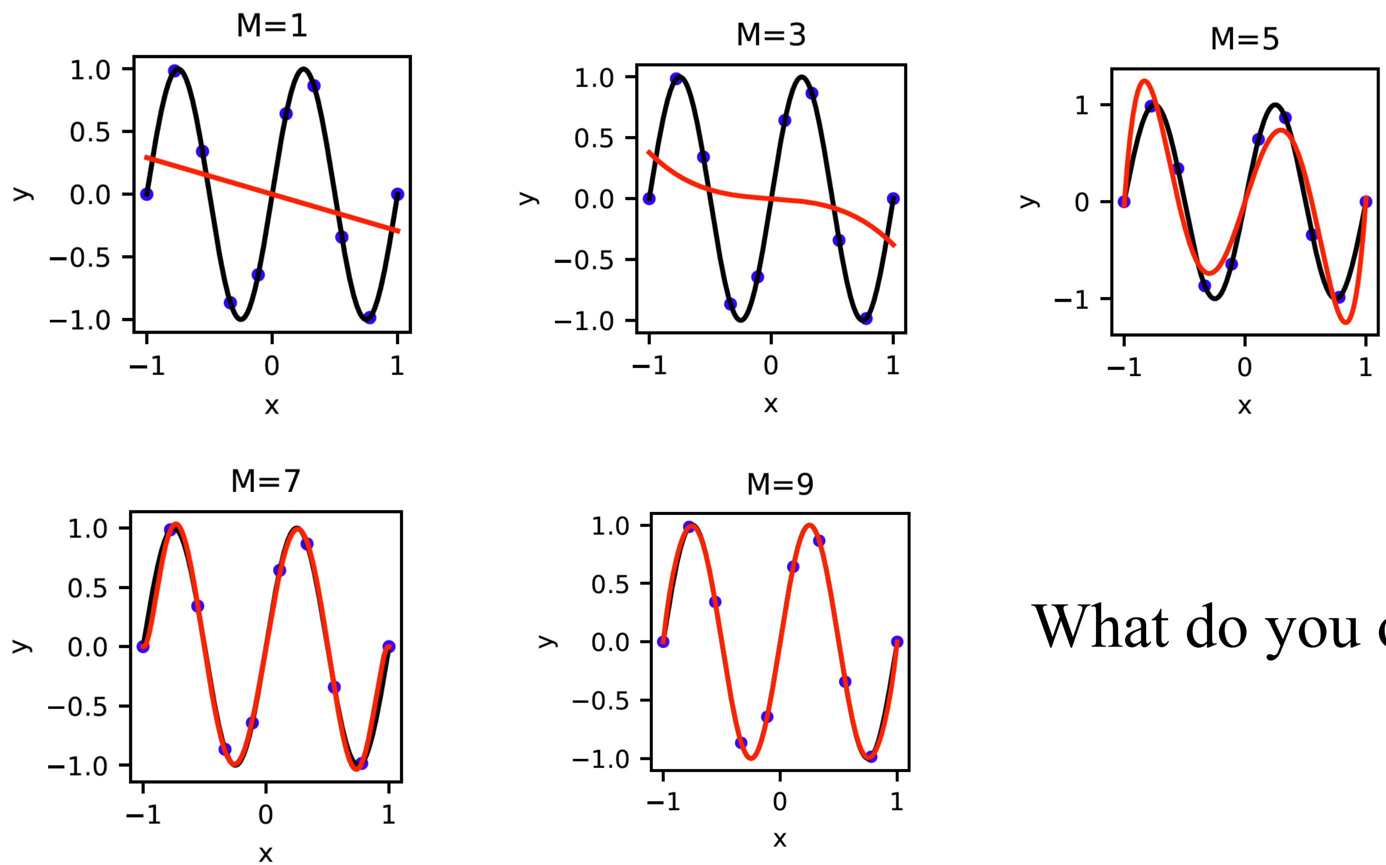
Visual example

- $y(x) = \sin(2\pi x) = \sum_{i=0}^{M-1} w_i x_i$
- Expand y(x) using Maclaurin series:

$$\sin(ax) = ax - \frac{a^3x^3}{3!} + \frac{a^5x^5}{5!} - \frac{a^7x^7}{7!} + \cdots$$

  - $w \approx (0, 6.28, -41.34, 0, 81.61, \ldots)$
- So if we start generating data using this approximation by increasing order, and compare to actual model, with no noise!

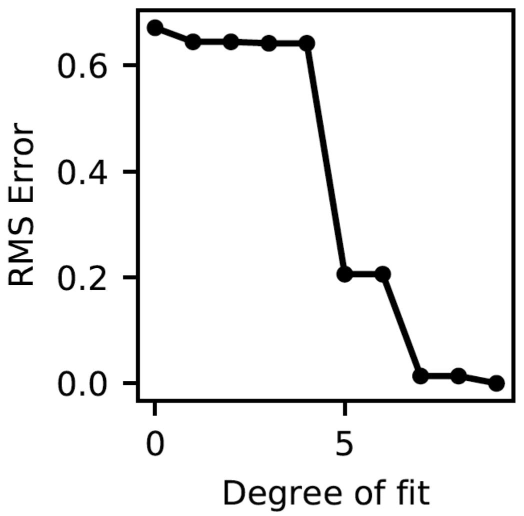
Polynomial fit of $y(x) = \sin(2\pi x)$;



What do you observe?

How good is the fit?

- Root-mean-square (RMS) error
  - $R = \sqrt{\frac{1}{N}\Sigma_i \; r_i}$
- Modelled (approximated) function converges with little change at M-7



UNIVERSITY OF BIRMINGHAM

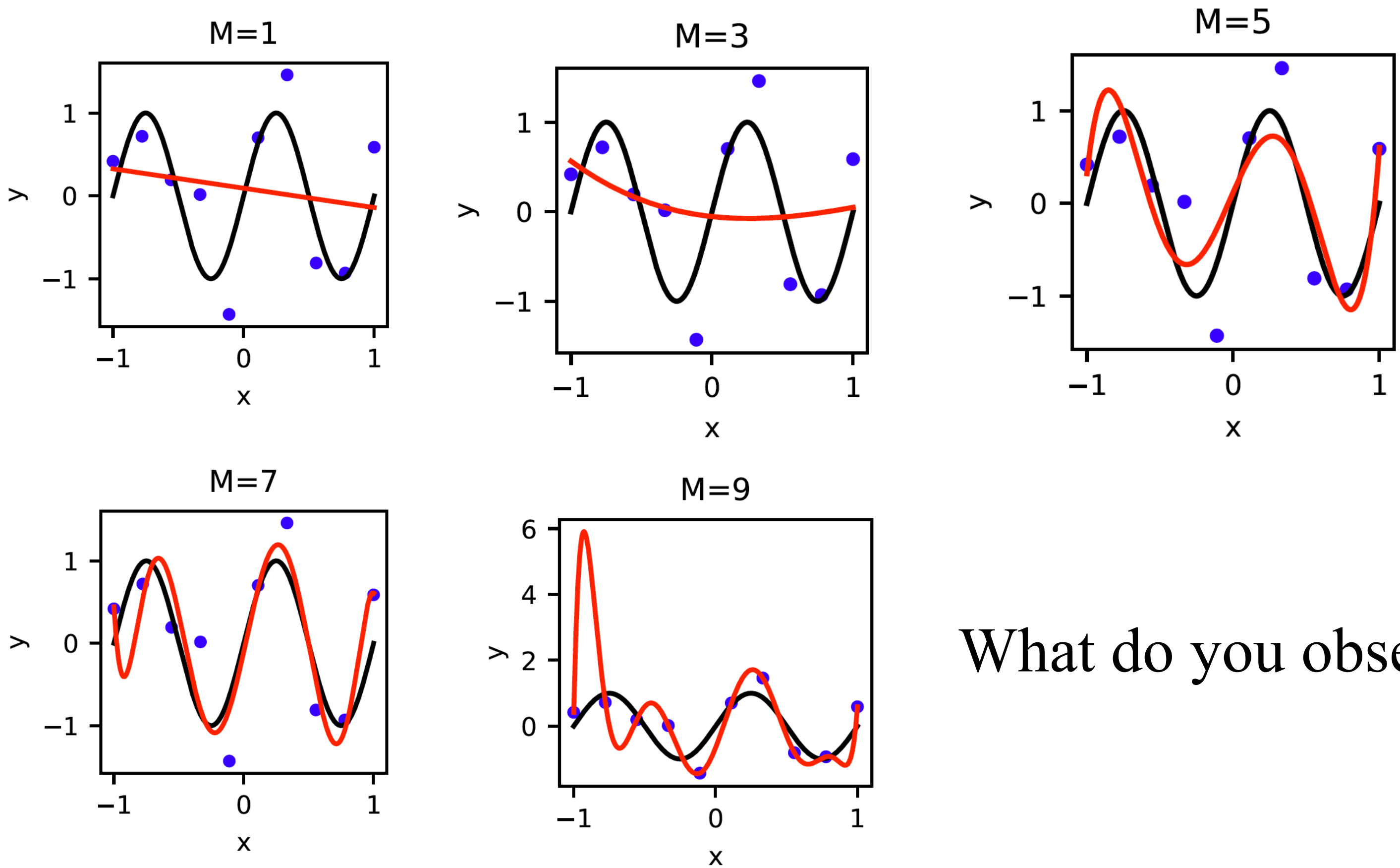
Coefficients of the fitted polynomial

M	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
0	0.00									
1	0.00	-0.29								
2	0.00	-0.29	-0.00							
3	0.00	-0.07	-0.00	-0.31						
4	0.00	-0.07	0.00	-0.31	-0.00					
5	0.00	3.85	-0.00	-16.51	0.00	12.69				
6	0.00	3.85	-0.00	-16.51	0.00	12.69	-0.00			
7	-0.00	6.00	0.00	-35.84	-0.00	54.04	0.00	-24.20		
8	-0.00	6.00	0.00	-35.84	-0.00	54.04	0.00	-24.20	-0.00	
9	-0.00	6.28	0.00	-41.12	-0.00	78.61	0.00	-63.77	-0.00	20.00
True	0	6.28	0	-41.34	0	81.61	0	-76.7	0	42.1

- Coefficients are not quite correct
  - Effect of limited sample domain (Maclaurin series is over [-∞ ∞]
- Low order terms match well
- M = 9 has zero error
  - Exactly fits all data point
- A strong hint as to what can go wrong

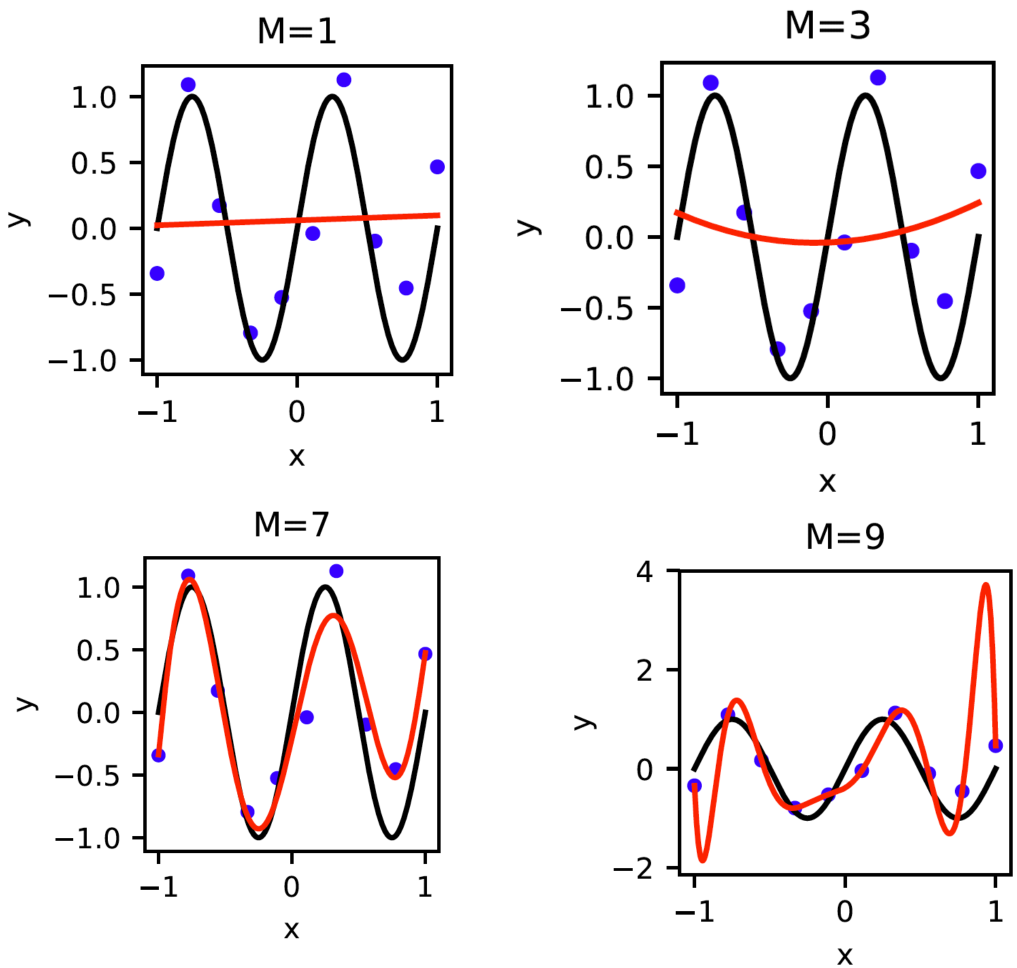
UNIVERSITY OF BIRMINGHAM

Polynomial fit of $y(x) = \sin(2\pi x) + \varepsilon$;



What do you observe?

Polynomial fit of $y(x) = \sin(2\pi x) + \varepsilon_1$;



What do you observe?

Noise corrupts

- Low-order fits similar in most cases
- High-order fits very differently
- Noise in the data leads to noise in the estimated model
  - Robust models cannot model the data very well
- How can we understand this?

Bias-Variance Decomposition

- Underlying data generating function $h(x)$
- Data $y = h(x) + \epsilon$
- Estimated model $f(x)$

What is the expected value of the least-squares loss?

$$\mathbb{E}[\mathcal{L}] = \mathbb{E}[(y - f)^2] \tag{1}$$

UNIVERSITY OF BIRMINGHAM

Bias-Variance Decomposition

We first expand the square

$$\begin{aligned} \mathbb{E}[\mathcal{L}] &= \mathbb{E}[(y-f)^2] & (2) \\ &= \mathbb{E}[y^2] + \mathbb{E}[f^2] - 2\mathbb{E}[yf] & (3) \end{aligned}$$

The variance of a random variable is:

$$\mathrm{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \tag{4}$$

and for independent variables X and Y

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \tag{5}$$

This allows us to rewrite the loss as

$$\mathbb{E}[\mathcal{L}] = \mathrm{var}[y] + (\mathbb{E}[y])^2 + \mathrm{var}[f] + (\mathbb{E}[f])^2 - 2\mathbb{E}[y]\mathbb{E}[f] \tag{6}$$

UNIVERSITY OF BIRMINGHAM

Bias-Variance Decomposition

- Recall $y = h(x) + \epsilon$
- Noise distribution: $\mathbb{E}[\epsilon] = 0$ and $\mathrm{var}[\epsilon] = \sigma^2$
- So $\mathbb{E}[y] = h$ and $\mathrm{var}[y] = \sigma^2$.

The expected loss becomes

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}] &= \sigma^2 + h^2 + \mathrm{var}[f] + (\mathbb{E}[f])^2 - 2h\mathbb{E}[f] && (7) \\
&= \sigma^2 + \mathrm{var}[f] + h^2 + (\mathbb{E}[f])^2 - 2h\mathbb{E}[f] && (8) \\
&= \sigma^2 + \underbrace{\mathrm{var}[f]}_{\text{variance}} + \underbrace{(h - \mathbb{E}[f])^2}_{\text{bias}} && (9)
\end{aligned}
$$

UNIVERSITY OF BIRMINGHAM

What does it all mean?

- How can we interpret this result?
- Only contribution from data y is its variance σ^2.
- All dependency on the *specific sample*, y, of the data has been absorbed into the other terms.
- The variance of f is a consequence of the variance in the data
  - No noise $\rightarrow$ always learn the same model
  - Noisy samples $\rightarrow$ different models.
  - var f is sensitivity of learned model to the choice of data.

What does it all mean?

- $h(x) - \mathbb{E}[f(x)]$ is the ability of the estimated model to accurately represent the true model
- It is the *bias* of the estimate.
- Fitting $f(x) = mx * c$ to $h(x) = \sin(2\pi x)$ has a high bias: cannot represent the data
- But it has a low variance: insensitive to particular data choice
- Loss minimisation requires simultaneous minimisation of both bias and variance
  - Models that both fit *and* generalise well
  - Nearly always in conflict
- A fundamental limitation of machine learning.