COMS 474
Homework 2

Haadi Majeed

Spring 2022

Problem 1

[26 points total]
For the following, you will look at how the underlying $f(X)$ and the number of samples play a role in the occurrence and extent of both under-fitting and over-fitting. You will use the Jupyter notebook posted on canvas along with these directions. You should only need to modify two variables in the last code block, setting *f_type* to 'linear' or 'sinusoidal' and setting *n_samples* to 15 or 1000.

Recall that there are many factors that affect how well a fitted model will perform on future data, including the class of models we are using for fitting (here polynomials), the underlying trend $f(X)$, the noise (here additive Gaussian), and the number of samples.

Additional Notes

- The results are random, since the data set itself is random.
- You are encouraged to re-run the code for each setting a few times to gain some insight into the variability of the results. We will explore this issue in more depth later on.
- The Y axis for the MSE figure does not have fixed limits. Pay attention to the range of the MSE values as you compare plots.

(a) Using a sinusoidal $f(X)$ and noise standard deviation $\sigma = 5$, plot the estimated models and the MSE curves for

(i) $n = 15$ samples

(ii) $n = 1000$ samples

Using those four plots, comment on whether under-fitting and/or over-fitting occur and the extent to which they do for each of i. and ii. Then discuss how the number of samples affects the result (eg similarities and differences between i. and ii.)

(b) Repeat (a) with a linear $f(X)$.

(c) Now discuss similarities and differences between the results of (a) and (b), given that the main difference was $f(X)$ (which for (a) was not

only non-linear in X but also not in the model class we were fitting with, while for (b) it was linear and in the model class).