COM S 474/574 Spring 2022

Introduction to Machine Learning

Homework 3

1Directions:

- **Due: Thursday February 17, 2022 at 9pm.** Late submissions will be accepted for 24 hours after that time without penalty, and then for another 24 hours with a 15% penalty. (the enforcement is strict, beginning at 9:01pm, except for extreme situations; having a poor wifi connection or minor computer problems is not sufficient for the penalty to be waived.)
- Upload the homework to Canvas as a single pdf file.
- If the graders cannot easily read your submission (writing is illegible, image is too dark, or if the contrast is too low) then you might receive a zero or only partialcredit.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

2Problems

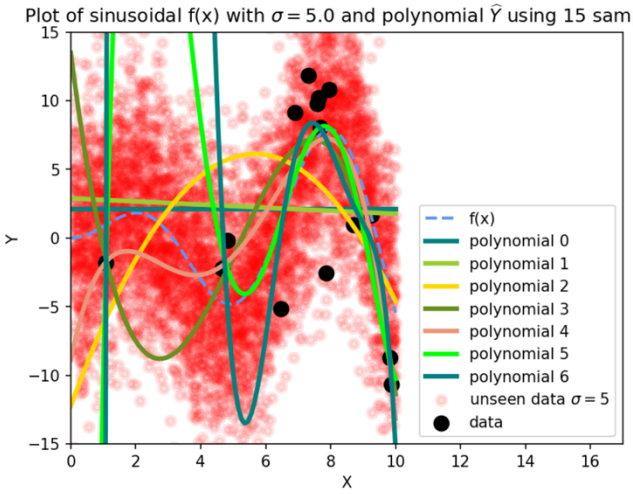
Problem 1. [26 points total]

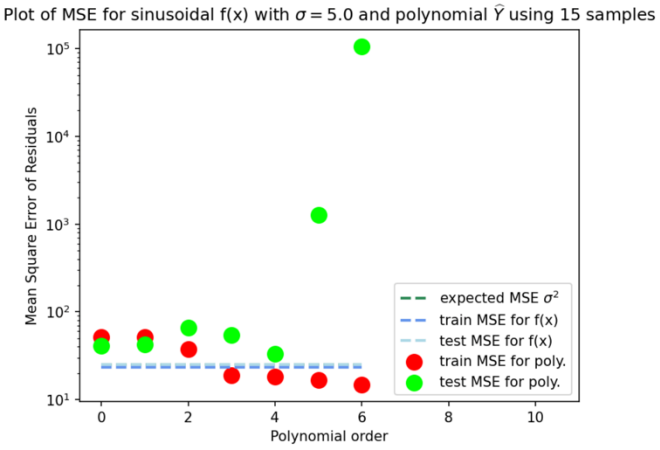
For the following, you will look at how the underlying $f(X)$ and the number of samples play a role in the occurrence and extent of both under-fitting and over-fitting. You will use the Jupyter notebook posted on canvas along with these directions. You should only need to modify two variables in the last code block, setting ftype 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.

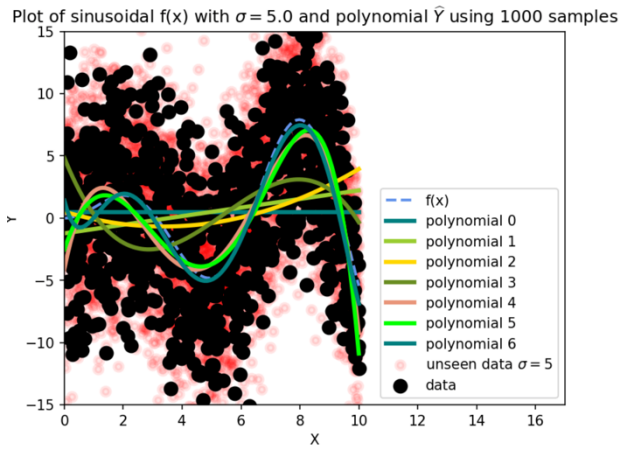
(a). Using a sinusoidal $f(X)$ and noise standard deviation $\sigma = 5$, plot the estimated models and the MSE curves for

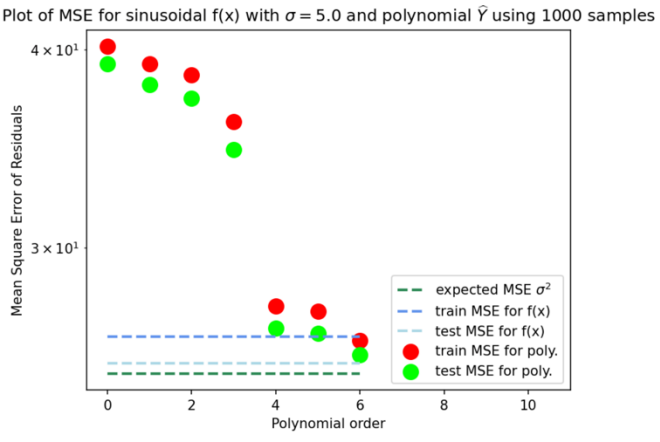
i. n = 15 samples and





ii. n = 1000 samples.

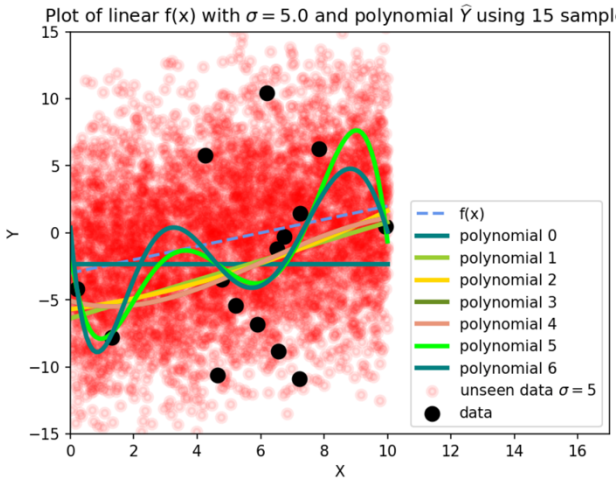


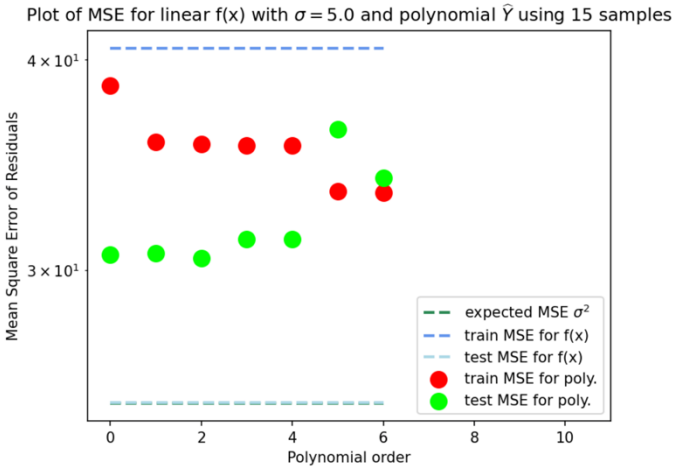


Using those four plots, comment on whether under-fitting and/or over-fitting occur and the extent to which they do for each of i. and ii. Then discuss how the number of samples affects the result (eg similarities and differences between i. and ii.).

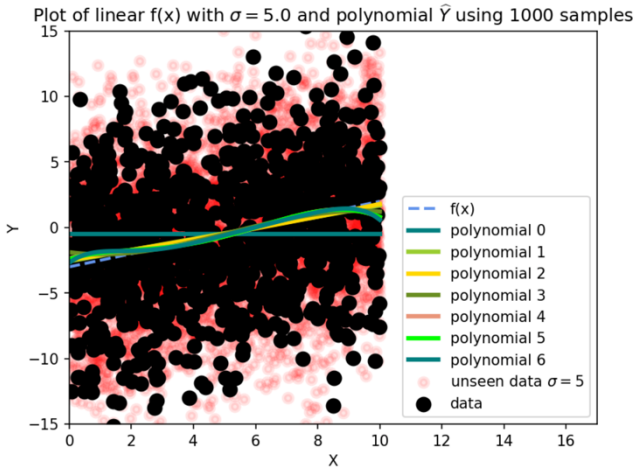
(b).Repeat (a). with a linear $f(X)$.

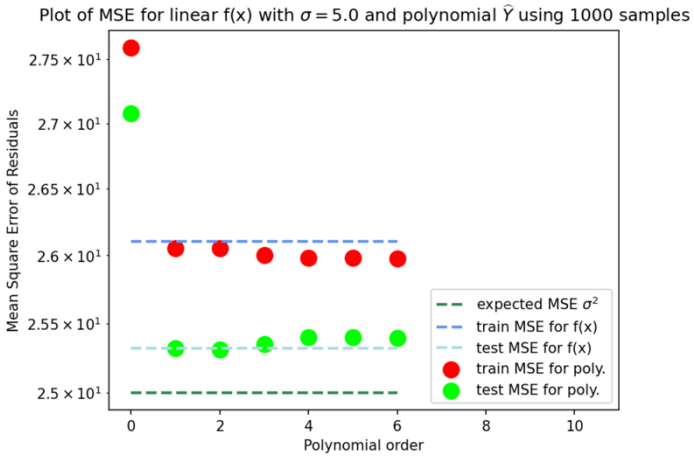
i. n = 15 samples and





ii. n = 1000 samples





(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).

The similarity between (a) and (b) is that as the sample size increase, MSE decreases. Also, as the polynomial order increases it tends to over-fit.
The difference between (a) and (b) is that the increment in the polynomial order does not effect on (b) but for (a) it makes a big difference in MSE. That is to say, it is easier to overfit in (a). Also, it is h to say as the polynomial order of the actual function, it gets harder to fit.

Formatting: Use a word processor for this assignment. Arrange the figures and corresponding discussions in a readable layout

- The plots (a). should each be on a single page and arranged neatly. Likewise for the plots for (b).
- the figures should not be too small to read or too low resolution that they are fuzzy. Two figures side by side of equal size, together spanning a column should be a good balance.

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.