ISTA 421/521 – Homework 3

Due: Sunday, October 15, 8pm

24 pts total for Undergrads, 30 pts total for Grads

Kargi Chauhan

STUDENT NAME

Undergraduate / Graduate

Instructions

In this assignment you are required to modify/write 4 scripts in python. Details of what this will involve are specified in exercises 1, 2, 5 and 6, below.

Included in the homework 3 release are following files in `code`:

- `poisson.py` - This script will be used in Exercise 1.
- `random_generator.py` - This script implements a class, `RNG`, that will simulate a random number generator, but where the random numbers are read from file. This is an attempt to make it so that all random number draws will be uniform across platforms. This will be used in Exercise 2.
- `approx_expected_value.py` - This script includes a demonstration of how to approximate an expected value through sampling. You will modify this code and submit your solution for Exercise 2.
- `predictive_variance.py` - This script contains code relevant to exercises 5 and 6. You will need to fill in two functions (as described in Exercise 5) as well as `exercise_6`
- `gauss_surf.py` - This is provided for fun – it is not required for any exercise here. It generates a 2d multivariate Gaussian and plots it as both a contour and surface plot. This provides an example of how to code making use of numpy array-computation; in particular, look at how the "mesh" of points in 2-d surface is computed simultaneously on lines 84-95 (see comments).
- `w_variation_demo.py` - This script is also provided for fun and is not required for the assignment. (It also provides more example python code!) This implements the simulated experiment demonstrating the theoretical and empirical bias in the estimate, $\widehat{\sigma^2}$, of the model variance, σ^2, as a function of the sample size used for estimation.

All exercises except exercise 1 require that you provide some "written" answer (in some cases also figures), so you will also submit a .pdf of your written answers. You can use LaTeX or any other system (including handwritten; plots, of course, must be program-generated) as long as the final version is in PDF.

As with the previous homework, the final submitted PDF written answers must be named `hw3-answers.pdf`.

NOTE: Exercises 3 and 7 are required for Graduate students only; Undergraduates may complete them for extra credit equal to the point value.

As in previous homework, pytest "unit tests" are provided to help guide your progress.

NOTE: For the unit test for Exercise 5c,
`code/test_ex5c_cov_w.py::test_ex5c_plot_functions_sampling_from_covw`
it is expected that you will get the following warning (line number could vary):

```
hw3_solution/code/predictive_variance.py:344:
RuntimeWarning: covariance is not symmetric positive-semidefinite.
```

You may work with others in the course on the homework. However, if you do, you. **must** list he names of everyone you worked with, along with which problems you collaborated. Your final submissions of code and

written answers **MUST ALL BE IN YOUR OWN CODE FORMULATION AND WORDS**; you cannot submit copies of the same work – doing so will be considered cheating.

(FCML refers to the course text: Rogers and Girolami (2016), *A First Course in Machine Learning, Second Edition*. For general notes on using LaTeX to typeset math, see: `http://en.wikibooks.org/wiki/LaTeX/Mathematics`)

1. [3 points] Adapted from **Exercise 2.3** of FCML:

   Let Y be a random variable that can take any non-negative integer value. The likelihood of these outcomes is given by the Poisson pmf (probability mass function):

   $$P(y) = \frac{\lambda^y}{y!} e^{-\lambda} \tag{1}$$

   By using the fact that for a discrete random variable the pmf gives the probabilities of the individual events occurring and the probabilities are additive, fill in the two functions in `poisson.py` as follows:

   (a) In `calculate_posson_pmf_a` compute the probability that $Y \geq 2$ and $Y \leq 6$ for $\lambda = 3$, i.e., $P(2 \leq Y \leq 6)$, and assign that to the return value `probability`.

   (b) In `calculate_posson_pmf_b`, using the fact that one outcome has to happen, compute the `probability` that $Y < 2$ or $Y > 6$ (again, for $\lambda = 3$).

   You are only allowed to use the python `math` package; no other packages are allowed.

   There is no need for a written answer to this exercise, only your code submission.

2. [4 points] Adapted from **Exercise 2.4** of FCML:

   Let X be a random variable with uniform density, $p(x) = \mathcal{U}(a, b)$.

   Work out analytically $\mathbf{E}_{p(x)} \left\{35 + 3x - 3x^2 + 0.2x^3 + 0.01x^4\right\}$ for $a = -1$, $b = 9$ (show the steps).

   The script `approx_expected_value.py` includes a function that demonstrates how you can use uniform random samples to approximate an expectation, as described in Section 2.5.1 of FCML. The script estimates the expectation of the function y^2 when $Y \sim \mathcal{U}(0, 1)$ (that is, Y is uniformly distributed between 0 and 1). This script generates a plot of how the estimation improves as larger samples are considered, up to 10,000 samples. NOTE: the script uses `random_generator.py`, which emulates a uniform random number generator interface but uses a fixed sequence of random numbers (`data/rand_uniform_10000.txt`); this ensures that correct solutions should be exact across platforms.

   Fill in the function `exercise_2` in `approx_expected_value.py` to compute a sample-based approximation to the expectation of the function $35 + 3x - 3x^2 + 0.2x^3 + 0.01x^4$ when $X \sim \mathcal{U}(-1, 9)$ and observe how the approximation improves with the number of samples drawn. In your written answer, include the generated plot `ex2_fn_approx.png` showing the evolution of the approximation, relative to the true value, over $10,000$ samples. Include a description of the trend you see in the plot.

   Solution.

3. [2 points; **Required only for Graduates**] Adapted from **Exercise 2.6** of FCMA:

   Assume that $p(\mathbf{w})$ is the Gaussian pdf for a D-dimensional vector $\mathbf{w}$ given in

   $$p(\mathbf{w}) = \frac{1}{(2\pi)^{D/2} |\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \mu)^\top \mathbf{\Sigma}^{-1} (\mathbf{w} - \mu)\right\}. \tag{2}$$

Solution 2 :

$$E_{p(x)} = \{35 + 3x - 3x^2 + 0.2x^3 + 0.01x^4\}$$

$$P(x) = U(a, b)$$

$$a = -1$$
$$b = 9$$

$$P(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{else} \end{cases}$$

$$E_p(x) = \int_a^b f(x)\, p(x)\, dx$$

$$= \frac{1}{9-(-1)} \int_{-1}^{9} \left(35x + \frac{3}{2}x^2 - x^3 + \frac{0.2x^4}{4} + \frac{0.01x^5}{5}\right) fx$$

$$= \frac{1}{10}\left(35(9) + \frac{3}{2}(81) - 729 + \frac{0.2}{4}(6561) + \frac{0.01}{5}(59049) - \left(35(4) + \frac{3}{2} + 1 + \frac{0.2}{4} - \frac{0.01}{5}\right)\right)$$

$$= \frac{1}{10}(315 + 121.5 - 729 + 328.05 + 118.09 + 35 - 1.5 - 1 - 0.05 + 0.02)$$

$$= \frac{1}{10}(186.11)$$

$$= \underline{\underline{18.61}}$$

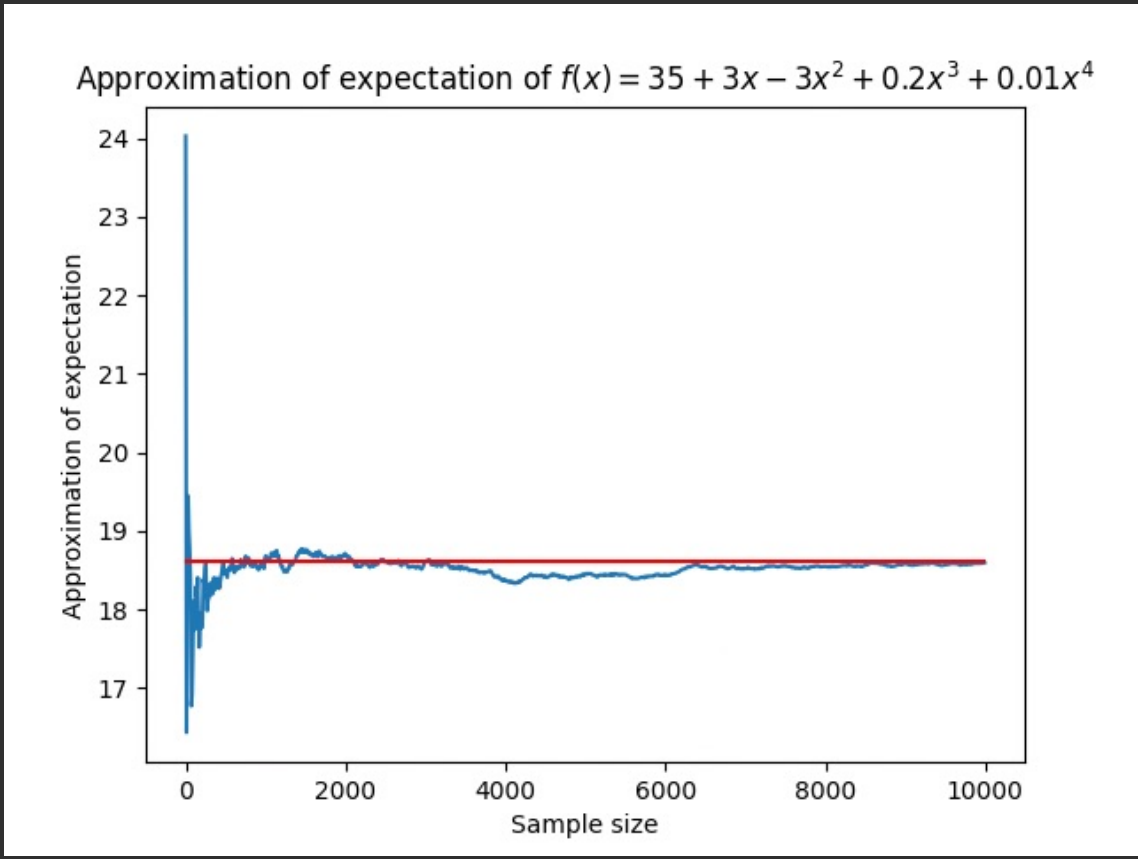


Figure 2

It shows that estimated value it closer to expected value when the sample size increases. Also, this is the reason why it is better to have larger size of sample or data to evaluate in order to get better estimated results.

Suppose we use a diagonal covariance matrix with different elements on the diagonal, i.e.,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D^2 \end{bmatrix}$$

Does this assume independence of the D elements of $\mathbf{w}$? If so, show how by expanding the vector notation of Eqn. 2 and re-arranging. You will need to be aware that the determinant of a matrix that only has entries on the diagonal is the product of the diagonal values and that the inverse of the same matrix is constructed by simply inverting each element on the diagonal. (Hint, a product of exponentials can be expressed as an exponential of a sum. Also, just a reminder that $\exp\{x\}$ is e^x.)

Solution.

4. [4 points] Adapted from **Exercise 2.9** of FCML:

   Assume that a dataset of N binary values, $x_1, ..., x_N$, was sampled from a Bernoulli distribution, and each sample x_i is independent of any other sample. Explain why this is *not* a Binomial distribution. Derive the maximum likelihood estimate for the Bernoulli parameter of this distribution.

   Solution.

5. [8 points] Adapted from **Exercise 2.12** of FCML:

   Familiarize yourself with the provided script `predictive_variance.py`. It is mostly implemented, but you will have to fill in the details for two functions:

   - `calculate_prediction_variance`, which calculates the *variance* for a prediction at x_{new} given the design matrix, $\mathbf{X}$, the estimated parameters, $\hat{\mathbf{w}}$, and target responses, $\mathbf{t}$.
   - `calculate_cov_w`, which calculates the estimated covariance of $\hat{\mathbf{w}}$ given the design matrix, $\mathbf{X}$, the estimated parameters, $\hat{\mathbf{w}}$, and target responses, $\mathbf{t}$.

   Once implemented, then you can run the script.

   When you run the script, it will generate a dataset based on a function (implemented in `true_function`) and then remove all values for which $2.5 \leq x \leq 4.5$. Three groups of plots will be generated:

   (a) First is a plot of the data (this will be generated by Part 5a of the script).

   (b) Next, the script will plot the error bar plots for predictions of values for model orders 1, 3, 5 and 9 (this will be generated by Part 5b).

   (c) Finally, in Part 5c, the script samples model parameters $\hat{\mathbf{w}}$ from the covariance $\text{cov}(\hat{\mathbf{w}})$ and plots the resulting functions (again, for model orders 1, 3, 5 and 9).

   In total, you will plot 9 figures. You must include the plots in your written submission and do the following: Include a caption for each figure that qualitatively describes what the figure shows; contrast the figures within group (b) with each other. Do the same for group (c). Also, clearly explain what effect removing the points $2.5 \leq x \leq 4.5$ has done in contrast to if they were left in.

   (NOTE: the script `predictive_variance` also contains elements for Exercise 6 (namely, the function `exercise_6`), but you can ignore those components while completing this exercise.)

   Solution.

6. [5 points]

Solution 3: cite: FML / lecture slides

$$P(w) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)\right\}$$

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sigma_D^2 \end{bmatrix}$$

$$= \text{diag}.(\sigma_1^2, \sigma_2^2, \sigma_3^2 \cdots \sigma_D^2)$$

$$= \sigma_1^2$$

$$\Sigma^{-1} = \text{diag}\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \frac{1}{\sigma_3^2} \cdots \frac{1}{\sigma_D^2}\right) = \frac{1}{\sigma_i^2}$$

$$P(w) = \frac{1}{(2\pi)^{D/2}(\sigma_i^2)^{1/2}}$$

$$\exp\left(-\frac{1}{2}(w-\mu)^T \frac{1}{\sigma_i^2}(w-\mu)\right)$$

$$= \frac{1}{(2\pi)^{D/2}(\sigma_i^2)^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{D}\left(\frac{w_i - \mu_i}{\sigma_i^2}\right)^2\right)$$

Replacing $(\sigma_i^2)^{1/2}$ with $(\sigma^2)^{D/2}$

$$P(w) = \frac{1}{(2\pi)^{D/2}(\sigma_i^2)^{1/2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{w_1 - \mu_1}{\sigma}\right)^2 + \left(\frac{w_2 - \mu_2}{\sigma_2}\right)^2 \cdots\right)$$

However, it shows that w element has its own mean μ and own variance σ^2

Solution 4 : Cite: From FML Pg 55-56

A Binomial distribution is turned as the probability of success by observing certain number of attempts.
A Bernoulli gives the probability of one attempt with having dataset of binary values which doesn't confirms whether it is Binomial because it can be just random numbers.

Bernoulli distribution :

$$P(X_n = x \mid r) = r^x (1-r)^{1-x} \quad \text{where } x \text{ can be } 0 \text{ or } 1$$

$$L(P, X) = \prod_{n=1}^{N} r^{x_n} (1-r)^{1-x_n}$$

$$\log L = \sum_{n=1}^{N} x_n \log r + (1-x_n) \log (1-r)$$

$$\frac{\partial \log}{\partial r} = \sum_{n=1}^{N} \left(\frac{x_n}{r} - \frac{1-x_n}{1-r} \right) = 0$$

$$\sum_{n=1}^{N} \left(\frac{x_n}{r} \right) = \sum_{n=1}^{N} \left(\frac{1-x_n}{1-r} \right)$$

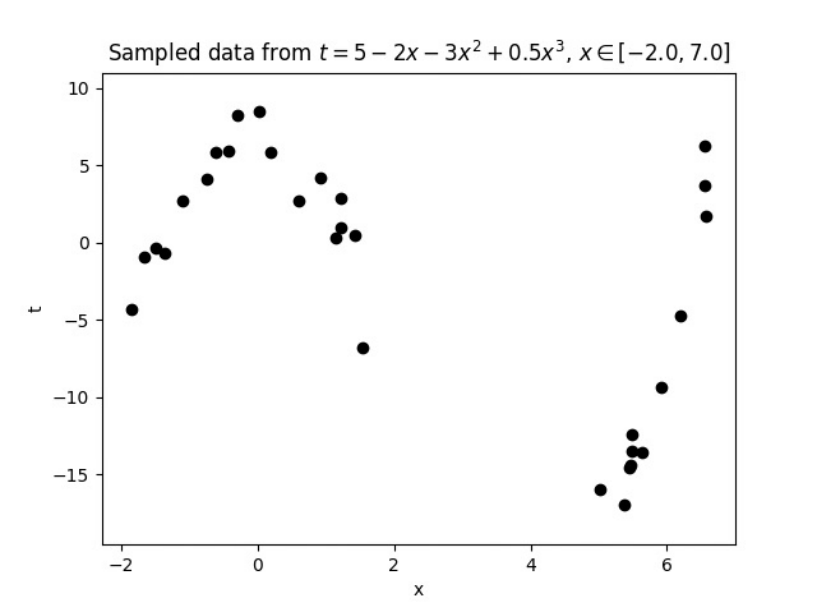
$$\sum_{n=1}^{N} r_n - r \sum_{n=1}^{N} x_n = rN - r \sum_{n=1}^{N} x_n$$

$$\boxed{r = \frac{1}{N} \sum_{n=1}^{N} x_n}$$

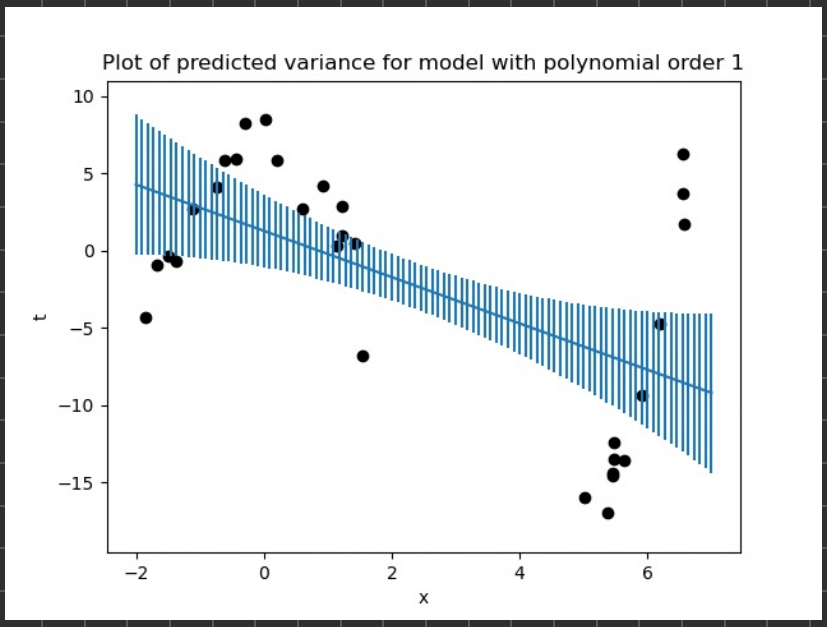
Solution 5:

When Removing points $2.5 \leq x \leq 4.5$ from the data given it shows the part Variance is larger as seen in figure 5. If they are in left then variance is less in the range and predicted value will get closer to original value.

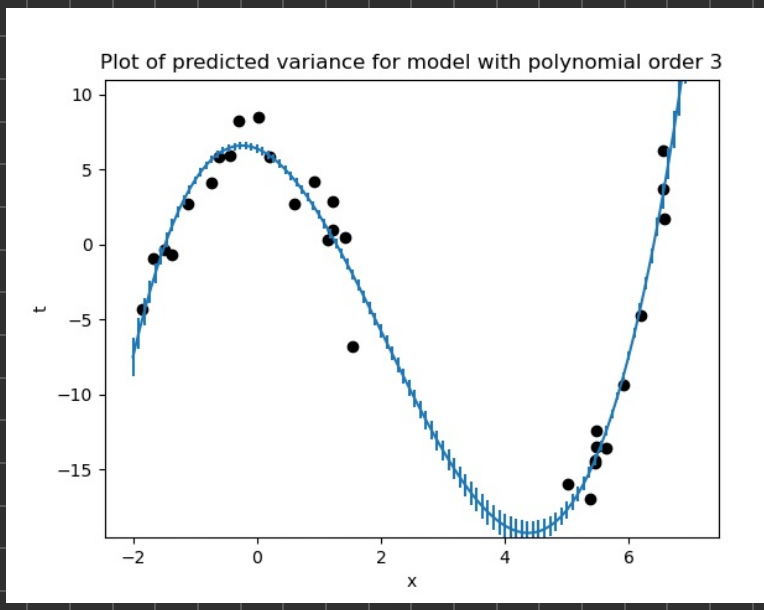
In figure 6 all the models have same results in graphs but they are different in terms of range.



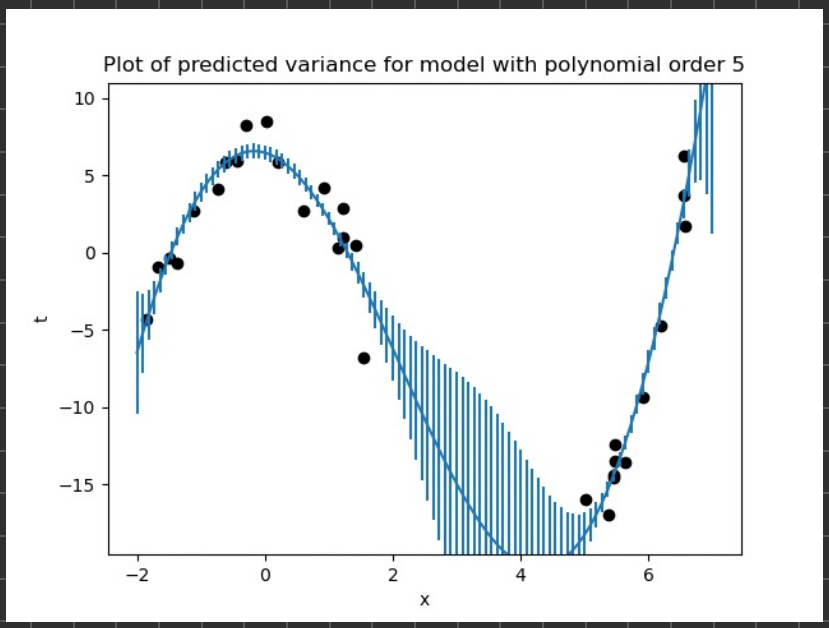
sample data



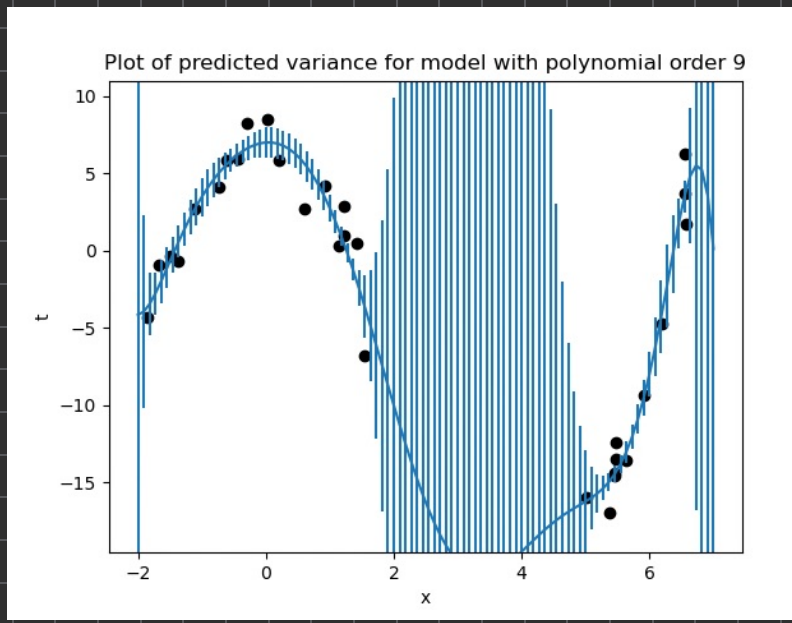
(a)



(b)

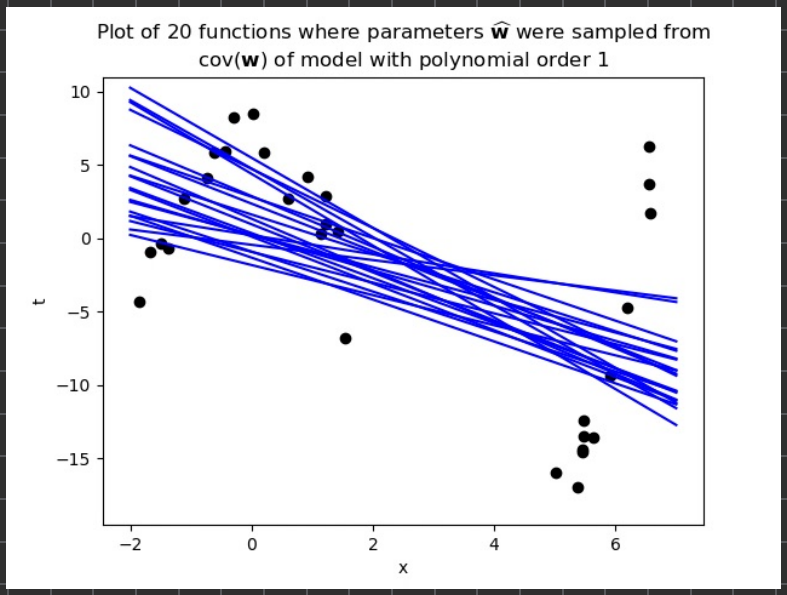


(c)

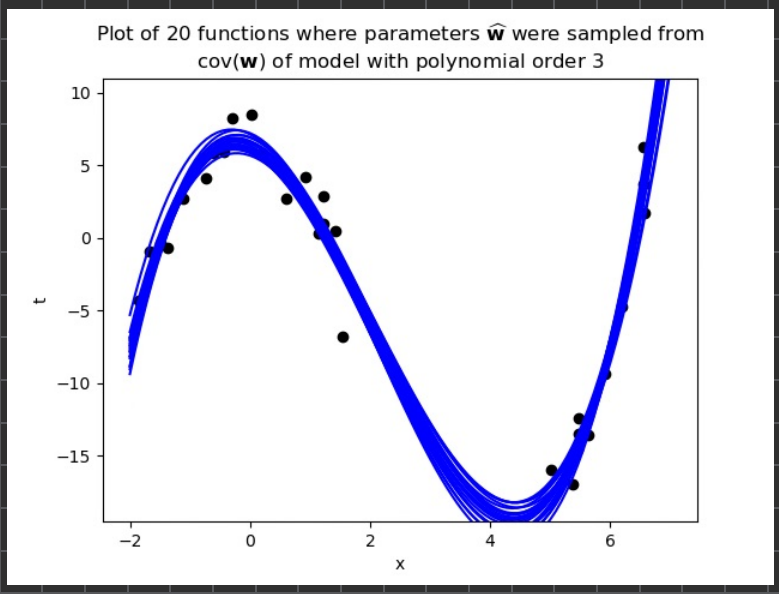


(d)

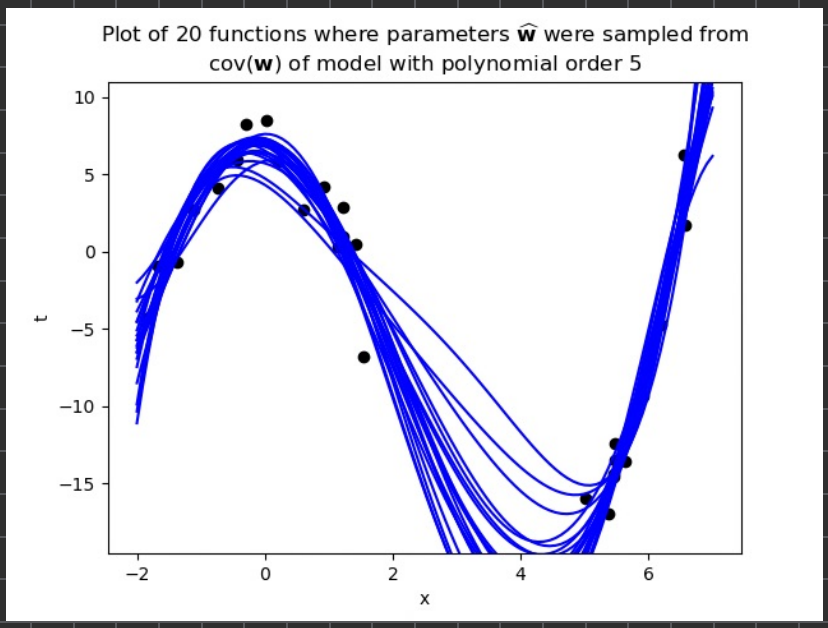
Predicted Variance for model in order 1, 3, 5, 9



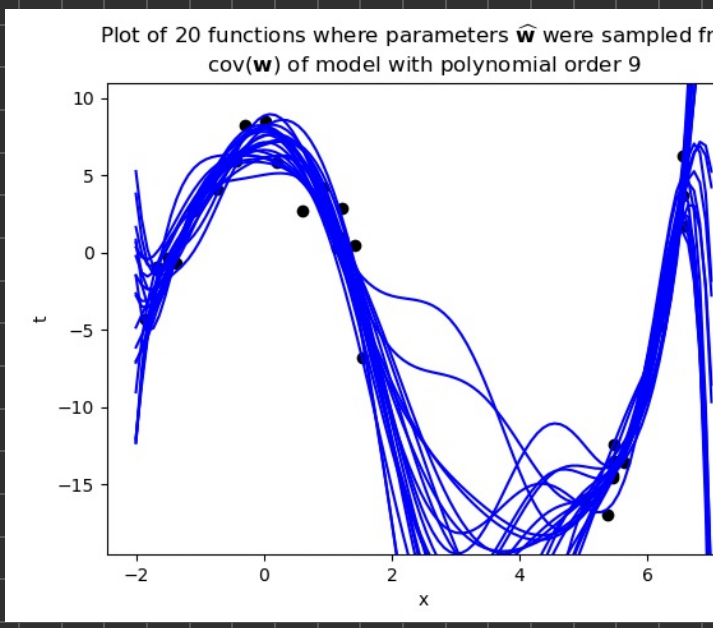
(a)



(b)



(c)



(d)

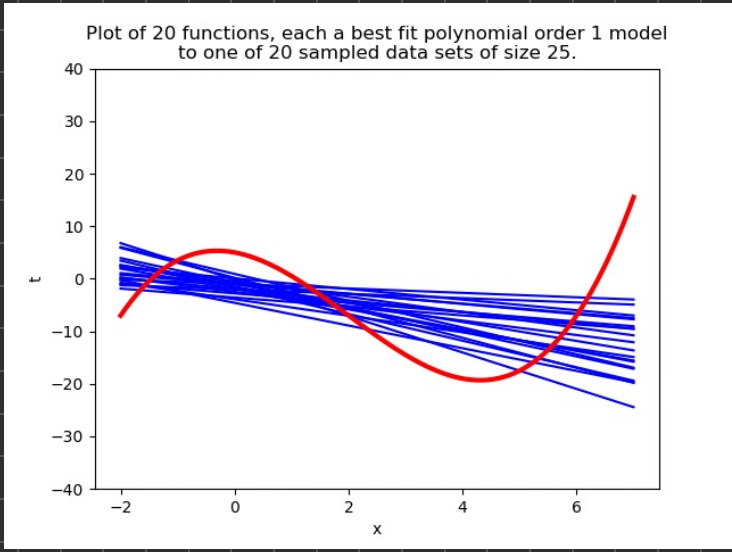
Sample models with parameter from coveriance

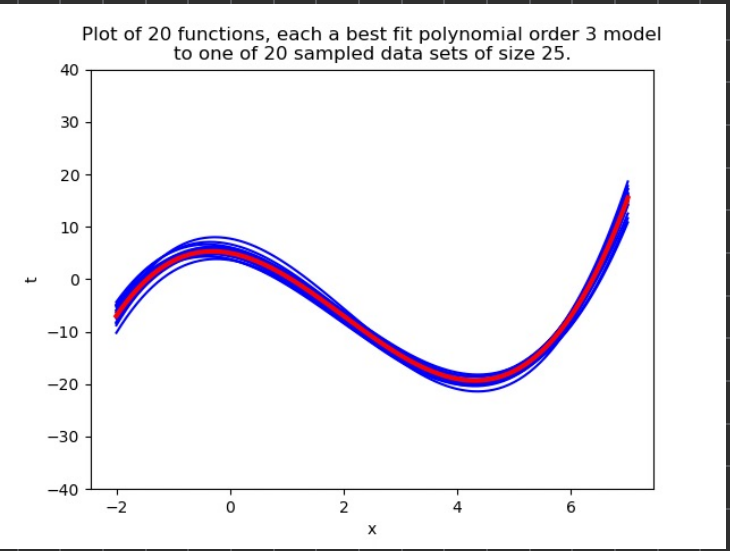
Summary :

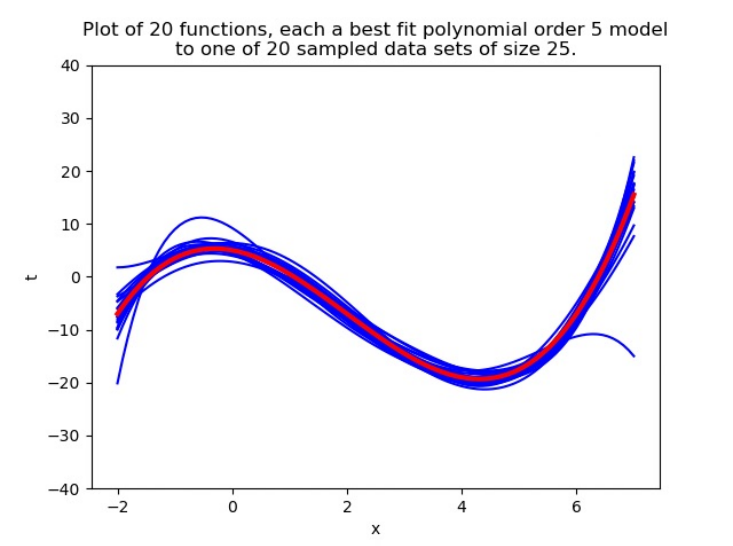
The graphs show the predicted variance including all the functions of polynomial models of different orders. The Figure shows how as the polynomial order increases from 1 to 9 the model's complexity is also seen to increase. Whereas, the plots with more curve are seen to fit in data.

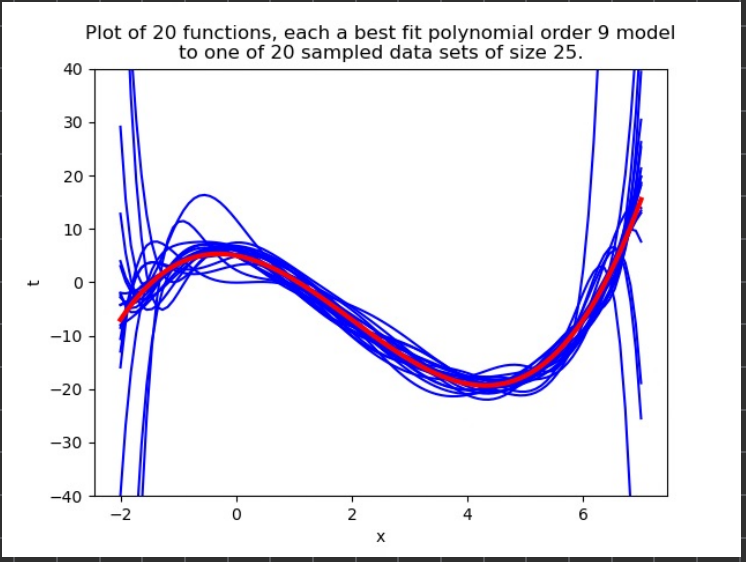
Graphs displays the predicted variance of the model in the increases in the polynomial order. The shade part in the plot shows the uncertainity at every point. The polynomial order increase that the given model takes details from the data but the cost of higher variance in regions having some data point.

Solution 6:









Sample model parameters with Variance

Summary :

The plot shows the visualization of the best-fit polynomials starting 1, 3, 5 & 9 with 20 sampled data with 25 sets of size. As the polynomial order increases the model takes details from the data.

(a) graph shows the linear fit showing the tend in data given data points whereas in the (d) graph the models are seen the increasing extensively capturing the fluctuation which shows the overfitting. Therefore, the model starts from simple to complex showing the bias & Variance in Regression.

The code for this exercise is also found in `predictive_variance.py`, in the function `exercise_6`. In this exercise, you will fill in the missing pieces as indicated. Once done, running `exercise_6` demonstrates how model bias impacts variance, similar to the demonstration in Lecture 9. Once implemented, the script will generate four plots: a separate plot for each of the model polynomial orders 1, 3, 5 and 9. In your written answer, first describe what the code is doing, in your own words. Then, include the four plots with descriptive captions, and describe what happens to the variance in the functions in the plots as the model order is changed.

Solution.

7. [4 points; **Required only for Graduates**] Adapted from **Exercise 2.13** of FCML:

   Derive the Fisher Information Matrix for the parameter of a Bernoulli distribution.

   Solution.

Solution 7:

cite : FML / Lecture slides

Bernoulli's distribution

$$P\left(\frac{x}{\theta}\right) = \theta^{x}(1-\theta)^{1-x}$$

Fisher information

$$F = -E\left[\frac{d^2 L(\theta)}{d\theta^2}\right]$$

Log likelihood $f^n = x \log(\theta) + (1-x)\log(1-\theta) = L(\theta)$

$$\frac{d(\theta)}{d(\theta)} = \frac{x}{\theta} - \frac{(1-x)}{(1-\theta)}$$

$$\frac{d^2 L(\theta)}{d^2(\theta)} = \frac{-x}{\theta^2} - \frac{(1-x)}{(1-\theta)^2}$$

$$E(x) = \theta$$

$$E(1-x) = 1-\theta$$

$$E(\theta) = \left(-\frac{E(x)}{\theta^2} - \frac{E(1-x)}{(1-\theta)^2}\right)$$

$$E(\theta) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1-\theta+\theta}{\theta-\theta^2}$$

$$= \frac{1}{\theta(1-\theta)}$$