Exercises Chapter 2

Exercise 1

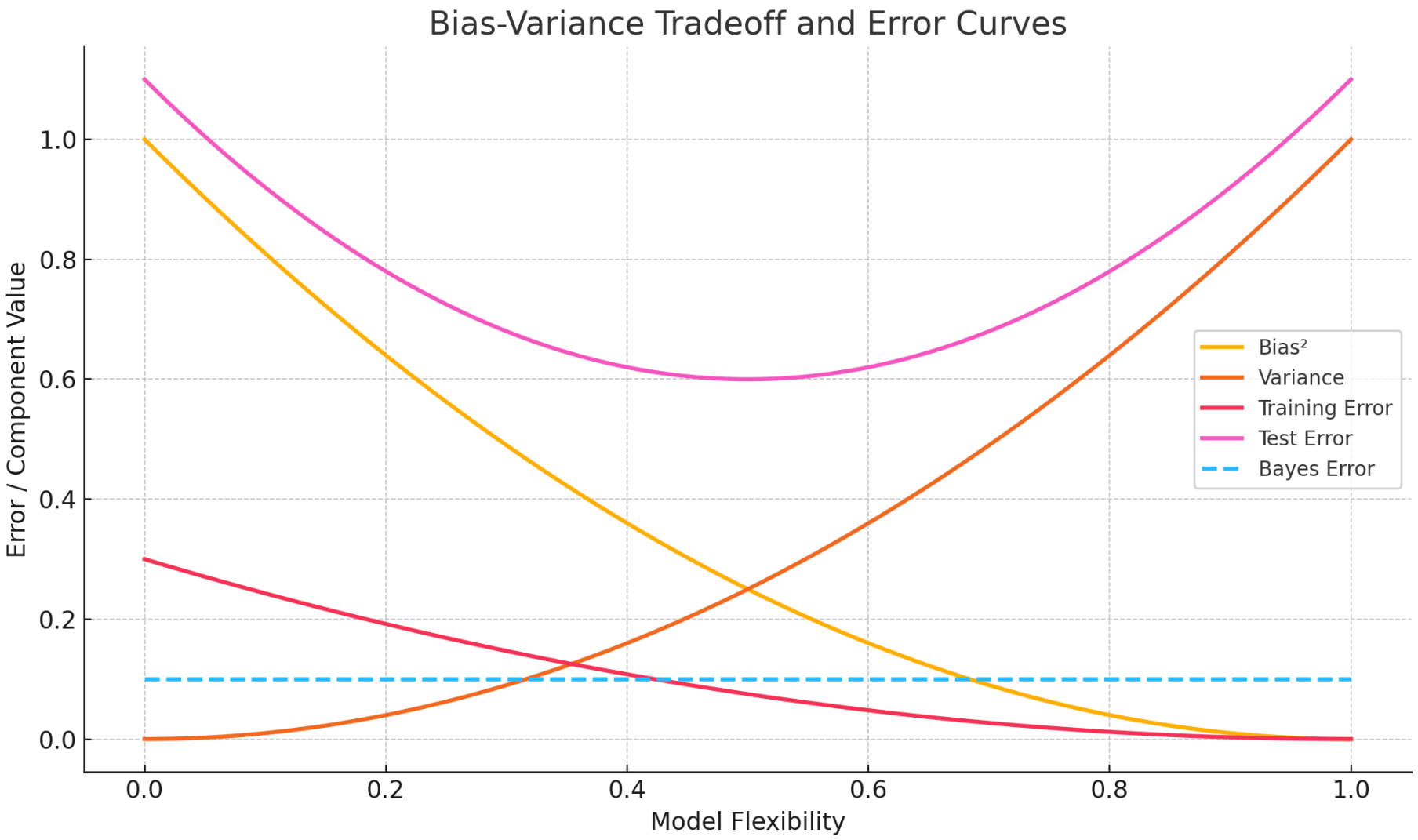
- Sample size n is extremely large with small Predictors P
  - Flexible methods will perform better cause they can capture complex patterns without overfitting , the high n provide enough data to estimate complex relationships
- Sample size n is small while the predictors P is extremely large
  - Flexible methods will perform worse cause they have high variance and can easily over fit, Linear regression is more stable in high dimensional settings
- The relationship between the predictors and the response is highly non-linear
  - inflexible models like linear regression assume a simple relationship between the predictions and the response
- The variance of the error terms $\sigma^2 = Var(\varepsilon)$ is extremely high
  - Flexible is worse cause they amplifies the variance unlike the more Linear approaches

Exercise 2

- We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
  - This is a Regression Problem $\rightarrow$ Salary is a quantitative
  - We interested in Inference $\rightarrow$ Relationship (How all these factors they affect) CEO salary
  - $n = 500$ $p = 3$
- We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
  - This is a Classification Problem $\rightarrow$ Two Classes Success or Failure and they are qualitative value
  - We interested in Prediction -> either success or failure
  - $n = 20$, $p = 13$
- We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market
  - This is a Regression Problem $\rightarrow$ % is a quantitative value
  - We interested in Prediction
  - $n = 53$, $p = 3 \rightarrow n$ is the number of weeks in a year

Exercise 3

- This Exercise challenged my understanding of Variance and the curve related to it and also clarified some concepts



- Explaining The curves :
  - **Bias** : Error introduce by approximating real world complex problems into a simple Model so increasing the model flexibility initially reduce the Bias till a point where increasing flexibility the decreases slows down and level off
  - **Variance :** How much estimate results change with different **training datasets,** less Flexible models have very low variance, Flexible Models becomes **sensitive to small fluctuations** in training data , causing high variance
  - Training error : Its the error rate the model make on a Training Data set flexible methods have a very low rate or none when they overfit,
  - **Test error** : The error rate on new unseen data, typically follows a **U-shape curve** which means the test error will decrease as the **flexibility increase** (due to lower bias), till a point where it start overfitting (The increasing variance dominates)
  - **The Bayes Error** : Irreducible error its the **minimum possible error** due to noise in the data no model can do better than this , its a horizontal line on the plot because it **doesn't depend on the model flexibility**

Exercise 4

Obs	X_1	X_2	X_3	Y	Distance
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	$\sqrt{10}$
4	0	1	2	Green	$\sqrt{5}$
5	-1	0	1	Green	$\sqrt{2}$
6	1	1	1	Red	$\sqrt{3}$

- Predicting Y when $X_1 = X_2 = X_3 = 0$

1. Compute the Euclidean distance between each observation and the test points X_i Using K-Nearest Neighbors
   - Distance = $\sqrt{(x_1 - 0)^2 + (x_2 - 0)^2 + (x_3 - 0)^2} = \sqrt{(x_1)^2 + (x_2)^2 + (x_3)^2}$
   - The Y prediction when $K = 1$ is Green , cause the closes point to the $(0, 0, 0)$ is $\sqrt{2}$ observation 5
   - The Y prediction when $K = 3$ is Red , cause the closes Three points to the $(0, 0, 0)$, are $\sqrt{2}$ green ,two red $\sqrt{3}$ and 2

- If the Bayes decision boundary is highly non-linear than the best value for K is to be small , Higher K would smooth over those curves