National Taiwan University **Introduction to Machine Learning and Deep Learning**
Department of Civil Engineering Instructor: C.-S. CHEN

Homework 2 Report
Essay and Programming, Due 21:00, Wednesday, October 2, 2024

Student ID: R12521521
Name: 吳竣名

---

1. Consider a five-bit representation 0_1_1 **with the following ground truth. Please list all the possible target functions.**

```
00101 | ○
00101 | ●
00111 | ○
00111 | ●
01101 | ○
01101 | ●
01111 | ○
01111 | ●

Total number of hypotheses: 8 = 2^2 * 2
```

2. Report your error rate using 5 training examples and 12 training examples.

5 training examples	12 training examples
error rate = 0.12	error rate = 0.0267

3.
(a) Calculate and report the expected out-of-sample error and its bias and variance components.

error	Bias	variance
0.2193	0.2044	0.0149

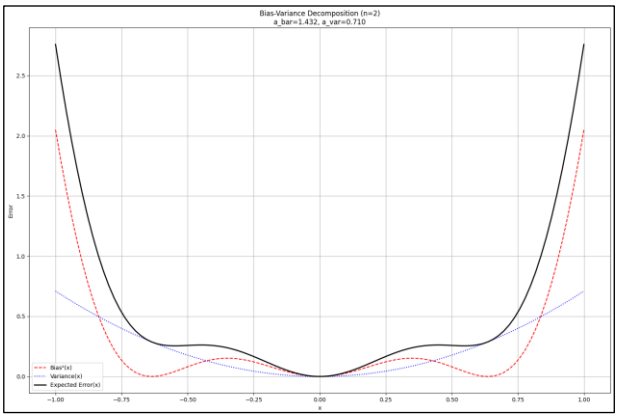
(b) Write a short essay to rationalize and comment on your findings.

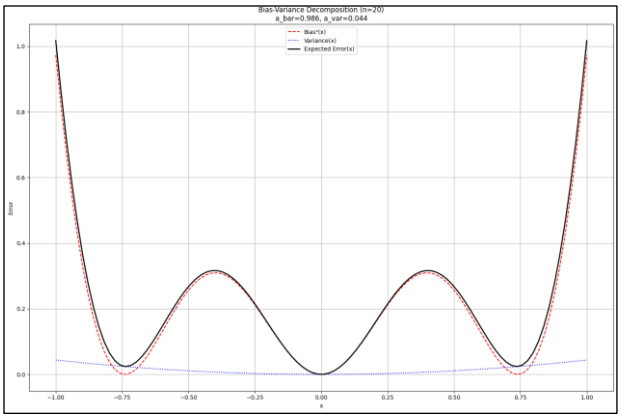
We can see the results of training data D using 20 data points and only 2 data points:

1. Bias:
    - 2-point data: The bias is larger, especially noticeable at both ends of the x-axis.
    - 20-point data: The bias is significantly reduced, with a smoother and more consistent curve.
2. Variance:
    - 2-point data: The variance is larger.
    - 20-point data: The variance is greatly reduced, maintaining a lower level across almost the entire x-axis range.
3. Expected Error:
    - 2-point data: The error is larger, showing peaks at both ends of the x-axis.
    - 20-point data: The error is significantly reduced, with a smoother curve and only slight increases at the ends of the x-axis.

Based on these observations, we can :have some conclusion:

- Effect of data quantity on bias: Increasing data points from 2 to 20 actually increased the model's bias. This might suggest that more data points reveal the underlying complexity of the data, making it hard for a simple model to fully capture the true data distribution.
- Effect of data quantity on variance: Increasing data points significantly reduced variance. This meets our expectations, as more data usually decreases the model's sensitivity to individual data points.
- Bias-variance trade-off: This example well demonstrates the bias-variance trade-off. Increasing data quantity reduced variance but increased bias at the same time. This shows that simply increasing data quantity may not improve all aspects of performance simultaneously.
- Consideration of model complexity: The results from 20 data points might suggest our model is too simple. Although variance decreased, the increase in bias indicates the model may need higher complexity to better fit the data.





4.

(a) Calculate and report the expected out-of-sample error and its bias and variance components.

Bias	variance
Bias: 0.896	**Variance: 0.049**

(b) Write a short essay to rationalize and comment on your findings.

We can see the result in the picture, it compare the 2 point & 20 point to training data.

1. Bias:

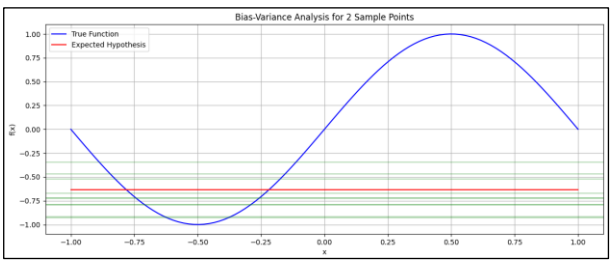
- Both the 2-point and 20-point data show large bias. This is because there is a big gap between our expected hypothesis function and the actual target function. As a result, the models based on both data sizes have high bias.

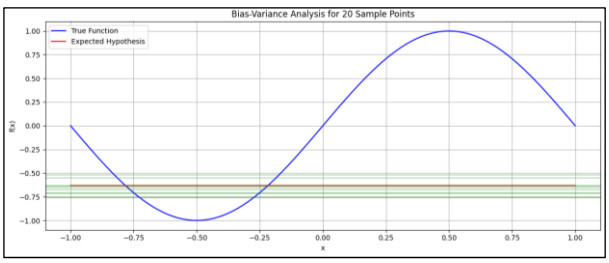
2. Variance:

- 2-point data: The variance is large, which can be clearly seen from the width of the green area in the graph.
- 20-point data: The variance is significantly reduced, with the green area becoming noticeably narrower.

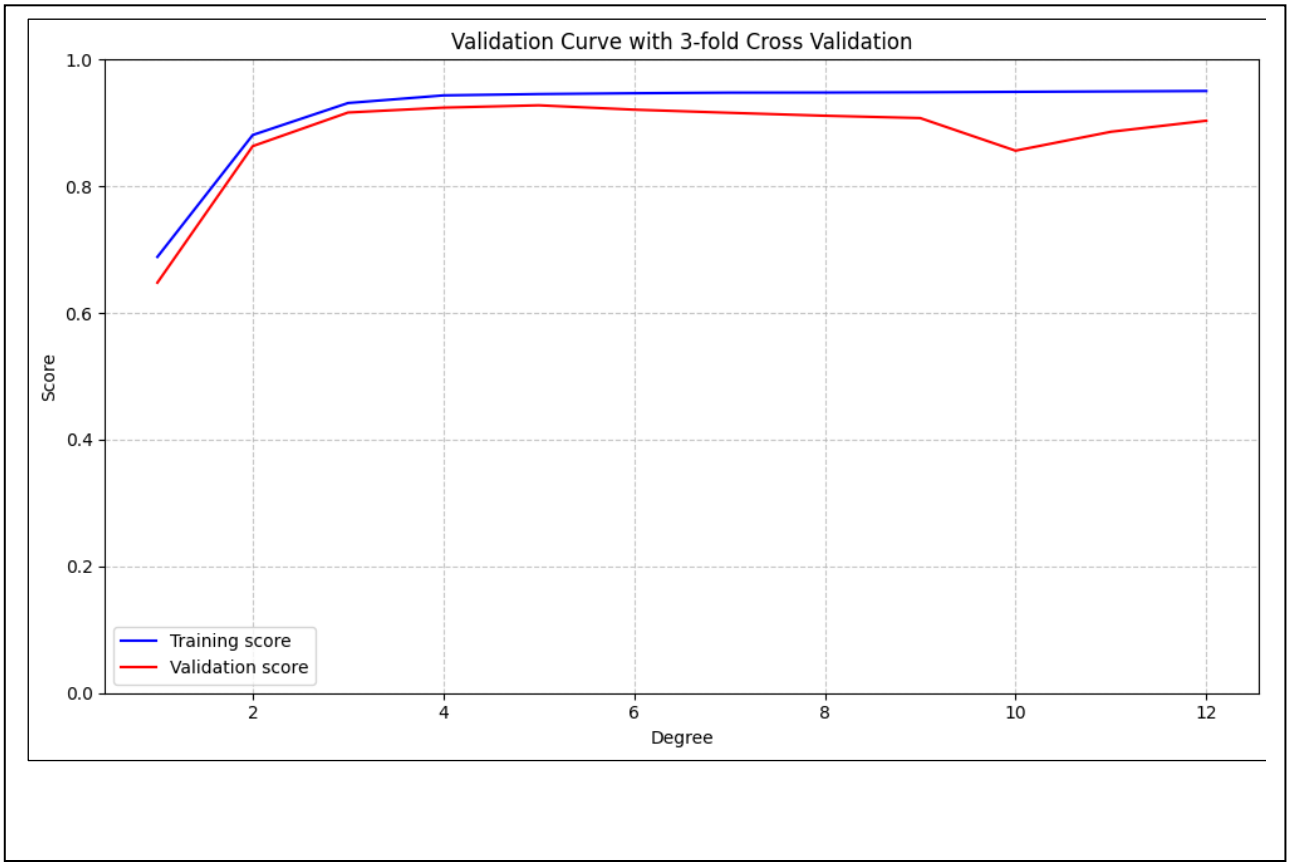
Based on our observation, Here are some conclusion in this case.

- Limitation of the hypothesis set: As you mentioned, our hypothesis set only includes constant functions h(x) = b. This type of function cannot fit the sine function f(x) = sin(πx) well.
- Model's expressive power: A constant function can only represent a horizontal line, while a sine function is a wavy curve. No matter how we choose the value of constant b, we cannot accurately represent the shape of the sine function.
- Bias-variance trade-off: In this case, even with more data points, the model's bias remains high. This is because the model's limitations (constant function) prevent it from capturing the true distribution of the data (sine function).
- Impact of data quantity: Increasing data points from 2 to 20 mainly affects the variance, with relatively little impact on bias. This is because more data points help us estimate the best constant value more accurately, but they cannot change the basic form of the model.





5. Please plot the mean score of training and validation curves using a polynomial regression model with 3-fold cross validation



Submission Format

Please convert `HW2_report_template.docx` to `HW2_report.pdf`.