Linear Models for Regression(Linear Basis Function Models)

- Background of Linear Regression
- The Regression Problem
- Linear Function Model
- Constructing the Basis Function
- Introducing a non-linear function

Linear Models of Regression

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \varphi_j(x)$$

- Regression is of the technique under Supervised Learning. The other is Classification.
- The objective of the regression model is to determine the value of one or more of a target variable t, given the value of a D-dimensional vector, x of input variables.
- In other words, you need to find the function that relates the input and the output.
- This can be done using Linear Models.
- One of such modes is the polynomial curve fitting which gives a function that is a linear function of a particular parameter.
- A better model is the *Linear Basis Function*.

Linear Basis Function

- Given a set of input dataset of N samples $\{x_n\}$, where $n = 1, \ldots, N$, as well as the corresponding target values $\{t_n\}$, the goal is to deduce the value of t for new value of x.
- The set of input data set together with the corresponding target values t is known as the training data set.
- On way to handle this is by constructing a function $y(x)$ that maps x to t such that:

$$y(x) = t \quad \text{for a new input value of } x.$$

- Then we can examine this model by finding the probability that the results are correct.
- This means that we need to examine the probability of t given x

$$p(t|x)$$

Constructing the Linear Basis Function

- The basic linear model for regression is a model that involves a linear combination of the input variables:

$$y(w,x) = w_o + w_1x_1 + w_2x_2 + \dots + w_Dx_D$$

where $x = (x_1, x_2, \dots, x_D)T$

- This is what is generally known as *linear regression*.

- The key attribute of this function is that it is a linear function of the parameters $w_0, w_1,\ldots, w_D$.
- It is also a linear function of the input variable x.
- Being a linear function of the input variable x, limits the usefulness of the function.
- This is because most of the observations that may be encountered does not necessarily follow a linear relationship.
- To solve this problem consider modifying to model to be a combination of fixed non-linear functions of the input variable.

Non-linear function

- If we assume that the non-linear function of the input variable is $\varphi(x)$, then we can re-write the original function as :

$$y(x,w) = w0 + w_1\varphi(x_1) + w_2\varphi(x_2) + \ldots + w_D\varphi(x_D)$$

- Summing it up, we will have:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x)$$

where $\varphi(x)$ are known as *basis functions.*

The total number of parameters in this function will be M, therefore the summation of terms is from ***j = 1 to M.***

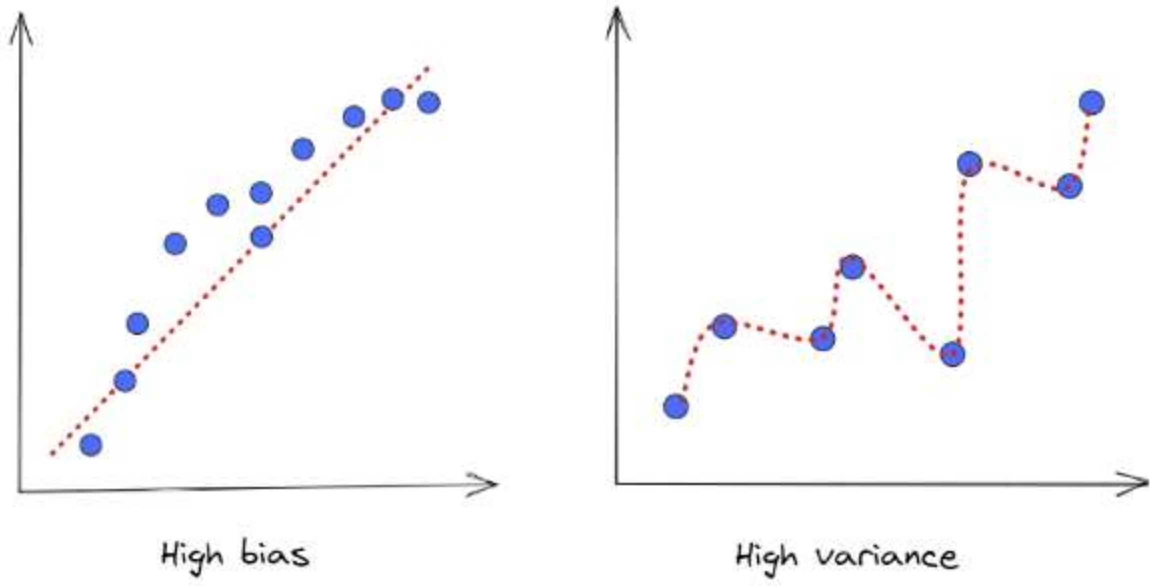
The parameter w_0 is known as the bias parameter which allows for a fixed offset in the data.

What is Bias-Variance Decomposition?

- The bias is defined as the difference between the ML model's prediction of the values and the correct value.
- Biasing causes a substantial inaccuracy in both training and testing data.
- To prevent the problem of under fitting, it is advised that an algorithm be low biased at all times.

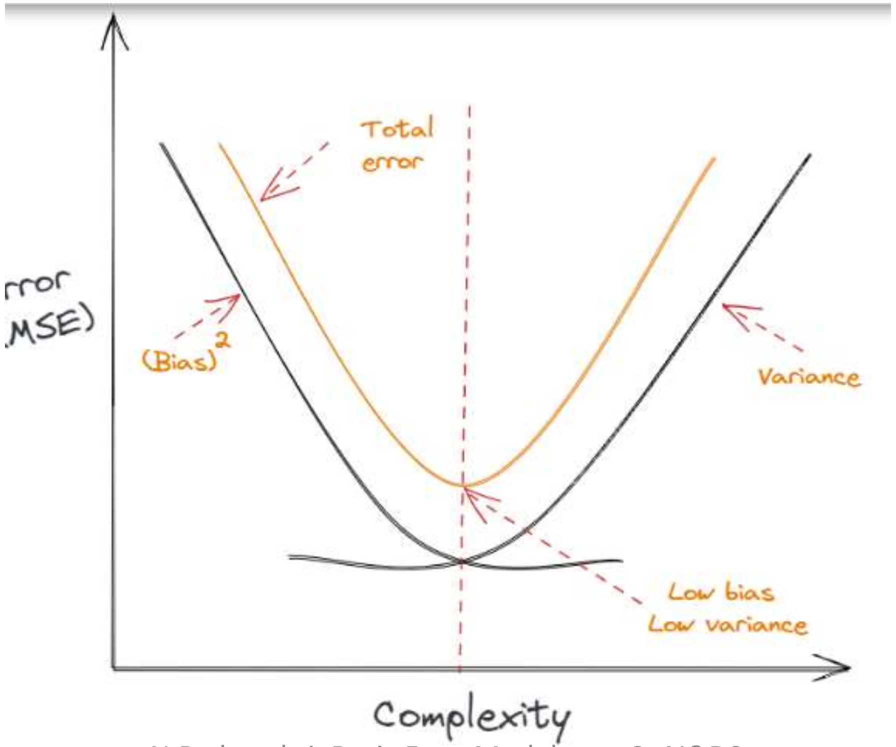
- The data predicted with high bias is in a straight-line format, which does not fit the data in the data set adequately.
- Under fitting of data is a term used to describe this type of fitting. This occurs when the theory is overly simplistic or linear in form.
- The variance of the model is the variability of model prediction for a particular data point, which tells us about the dispersion of the data.
- The model with high variance has a very complicated fit to the training data and so is unable to fit correctly on new data.

- As a result, while such models perform well on training data, they have large error rates on test data.
- When a model has a large variance, this is referred to as Overfitting of Data. Variability should be reduced to a minimum while training a data model.



- Bias and variance are negatively related, therefore it is essentially difficult to have an ML model with both a low bias and a low variance.
- When we alter the ML method to better match a specific data set, it results in reduced bias but increases variance.
- In this manner, the model will fit the data set while increasing the likelihood of incorrect predictions.

- when developing a low variance model with a bigger bias.
- The model will not fully fit the data set, even though it will lower the probability of erroneous predictions.
- As a result, there is a delicate balance between biases and variance.



When to use bias-variance decompositior.

- **Low Bias:** Tends to suggest fewer implications about the target function's shape.
- **High-Bias:** Suggests additional assumptions about the target function's shape.
- **Low Variance:** Suggests minor changes to the target function estimate when the training dataset changes.
- **High Variance:** Suggests that changes to the training dataset cause considerable variations in the target function estimate.

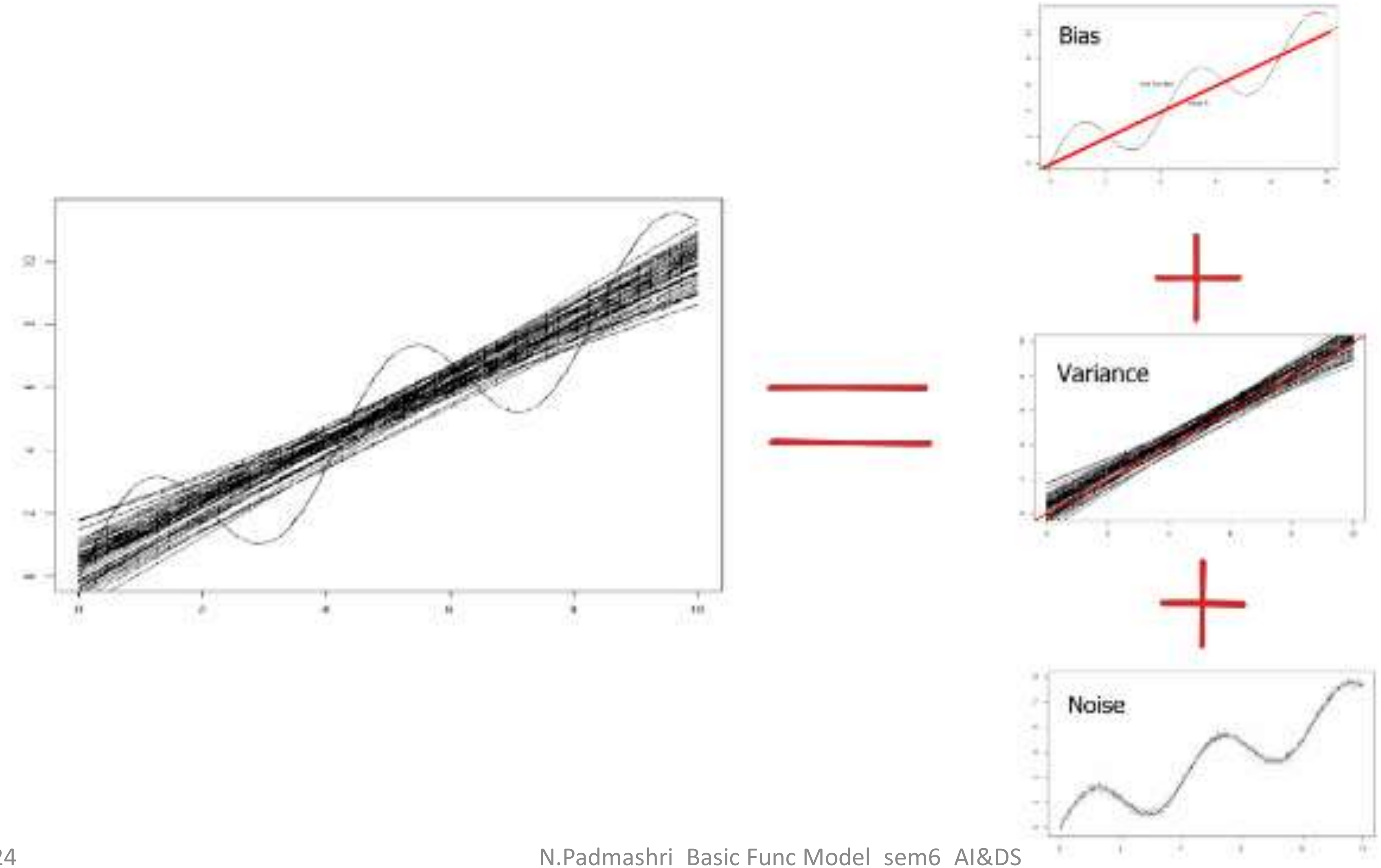
When to use bias-variance decomposition

- Theoretically, a model should have low bias and low variance but this is impossible to achieve.
- So, an optimal bias and variance are acceptable.
- Linear models have low variance but high bias and non-linear models have low bias but high variance.

How does this work?

- The total error of a machine learning algorithm has three components: bias, variance and noise.
- So decomposition is the process of derivation of total error in this case we are taking Mean Squared Error (MSE).
- Total error = Bias2 + Variance + Noise

EXAMPLE



- In this example, we’re attempting to match a sine wave with lines, which are obviously not realistic.
- On the left, we produced 50 distinct lines.
- The red line in the top right corner represents the anticipated hypothesis which is an average of infinitely many possibilities.
- The black curve depicts test locations along with the true function.
- Because lines do not match sine waves well, we notice that most test points have a substantial bias.
- Here the bias is the squared difference between the black and red curves.

- Some of the test locations, exhibit a slight bias, where the sine wave crosses the red line.
- The variance in the middle represents the predicted squared difference between a random black line and the red line.
- The irreducible error is the predicted squared difference between a random test point and the sine wave.