Bias/Variance Trade Off

ML Model

In ML, for some response *Y* and predictors X_1, X_2, …, X_p , there is a relationship of the form:

$$Y = f(X) + \epsilon$$

In prediction, we are trying to estimate the response:

$$\hat{Y} = \hat{f}(X)$$

Where $\hat{f}$ is the model we are trying to fit.

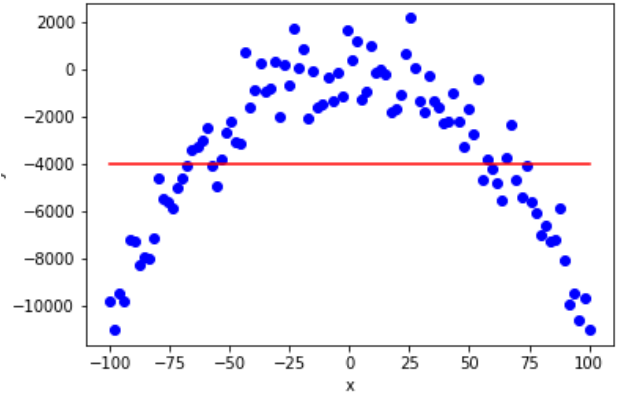
Variance & Bias

Variance is the amount $\hat{f}$ would change if we used a different training set to fit the model.

Bias is the error that is produced when trying to approximate a real-life problem using a simpler model.

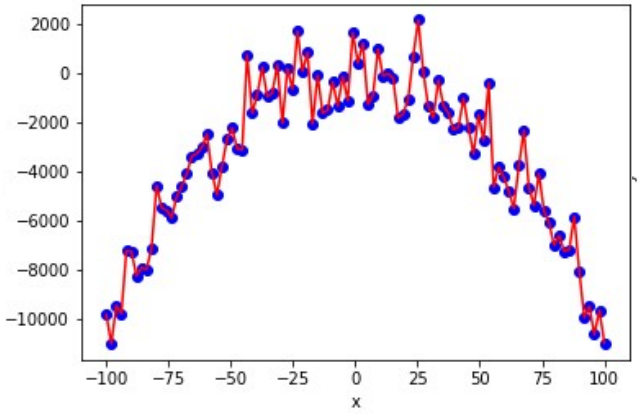
Underfitting vs Overfitting

Low Variance, High Bias



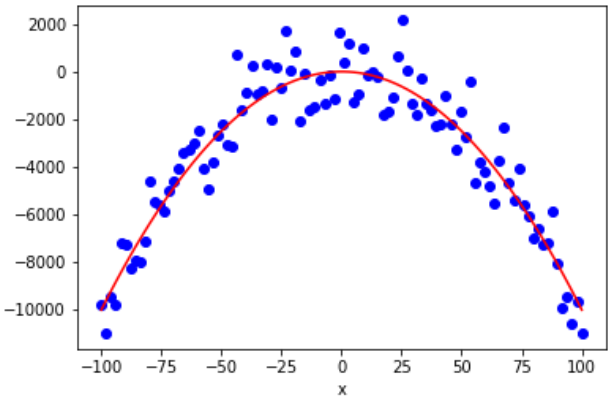
Underfitting

High Variance, Low Bias



Overfitting

Low Variance, Low Bias



Underfitting vs Overfitting

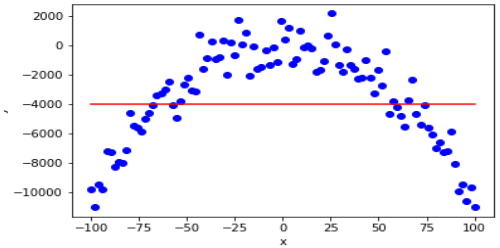
Underfitting

- Model performs poorly on training data AND poorly on testing data
- Model is unable to capture relationship between input data and target data (x and y) values
- Model is not detailed enough - i.e. a straight line to model something that is exponential

Overfitting

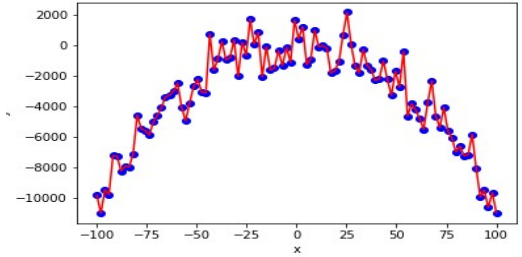
- Model performs well on training data but NOT on testing data
- "Memorizing the training data"
- Unable to generalize

Low Variance, High Bias



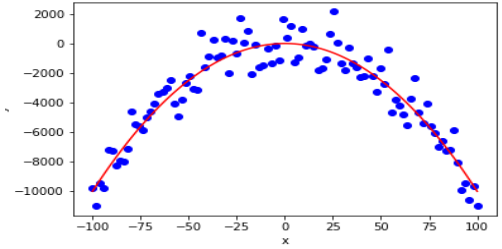
Underfitting

High Variance, Low Bias



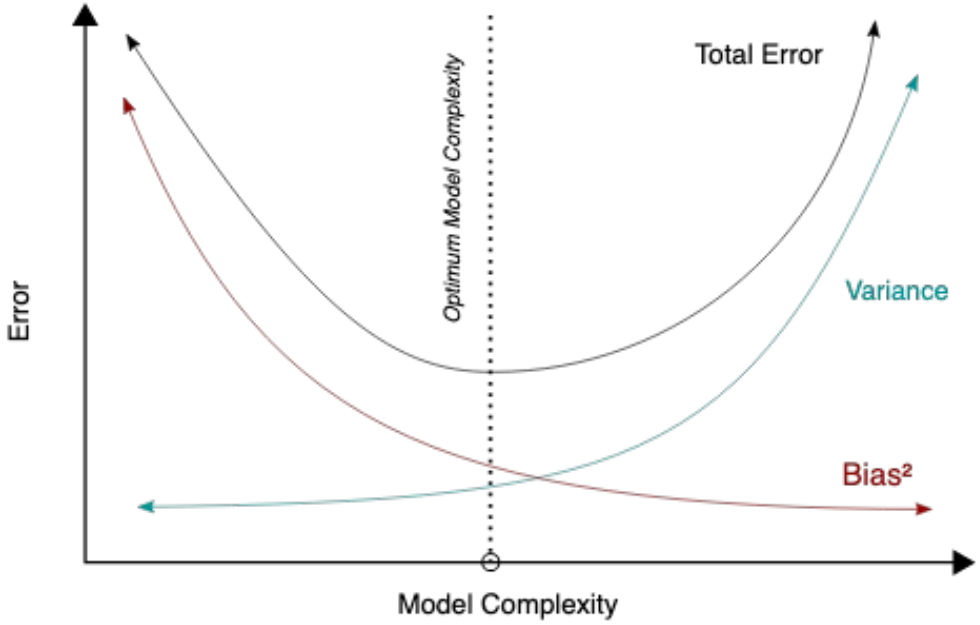
Overfitting

Low Variance, Low Bias



What Does This Mean For Us?

- Generally, more flexible methods will have more variance and less bias.
- However, at some point, the reduction in bias will slow down and the change in variance will increase rapidly.
- Therefore, **the goal in machine learning is to find a method for which both the variance and the bias are low**.



Source: Wikimedia Commons

How Do We Address This in Practice?

It is generally not possible to explicitly compute the bias and variance for a method. However, it is important to keep these concepts in mind.

Flexible methods that reduce bias are not always the best solutions as they can have high variance (and they may not perform well when tested under a new data set).