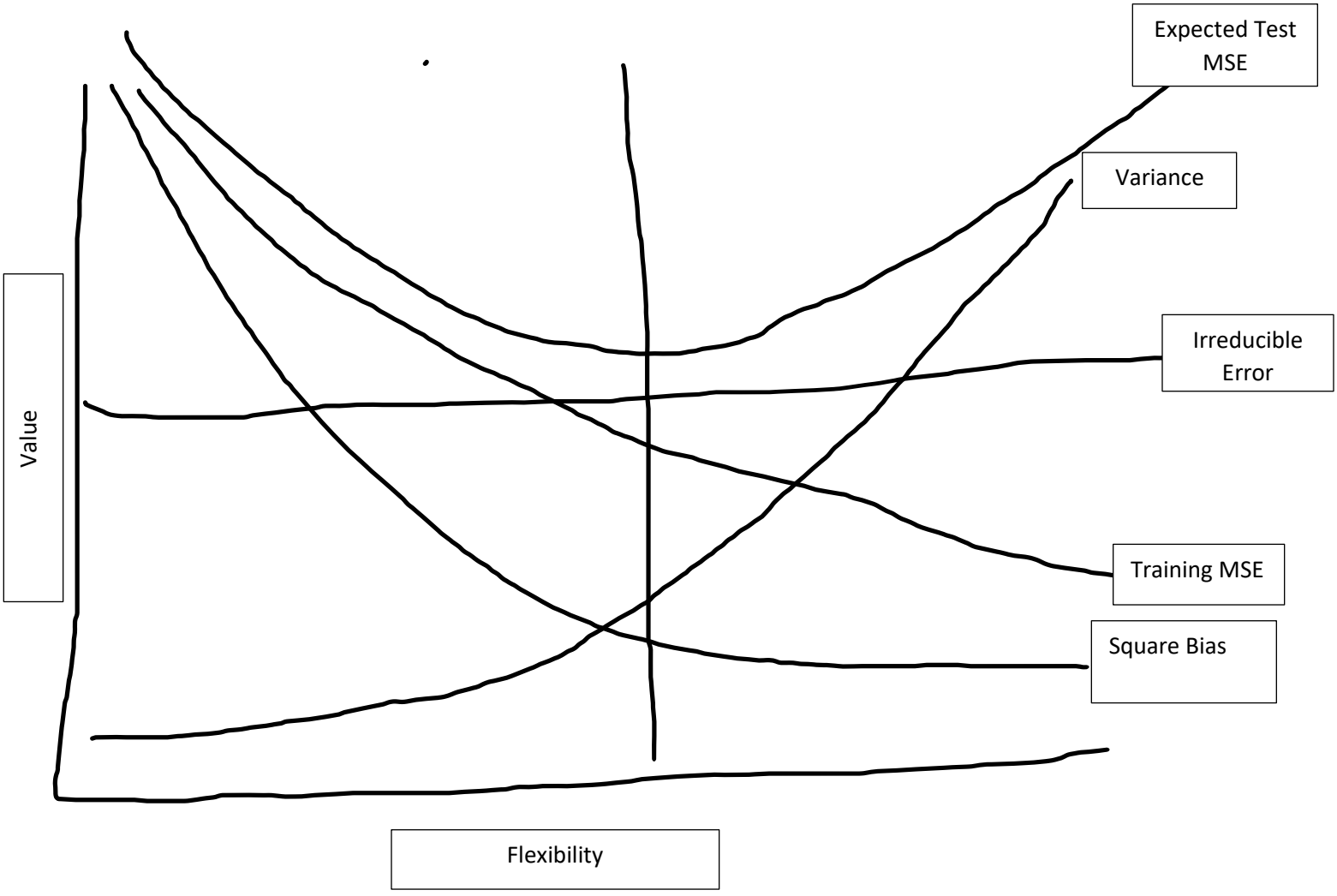
Keven Lin

2/1/2022

DS 301: HW 1

Problem 1: Bias-Variance Decomposition

a. On a single plot, provide a sketch of typical curves for (squared) bias, variance, expected test MSE, training MSE, and the irreducible error as we go from less flexible statistical learning methods towards more flexible methods. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be 5 curves. Make sure to label each one.



b. Define in plain language (so that a non-data scientist can understand) what the quantities expected test MSE, training MSE, bias, variance and irreducible error mean.

Training MSE → Measure of the average square errors of the model compared to the training data set.

Test MSE → Measure of the average square errors of the model compared to the test data set.

Bias → The natural bias/error that occurs by incorrect assumption by the model.

Variance → The variability of the model predication base on different portions of training data that is fed to it.

Irreducible Error Mean → it is the errors that we cannot remove with our model.

c. Explain why each of the five curves has the shape displayed in part (a).

Training MSE will less values as flexibility increase as more flexible the data is, the more the model will fit the data and produce less values specific values that we are looking for.

The Test MSE is a U shape because too much flexibility and too little flexibility will produce results that overfitting and underfitting (thus the curve producing the wings on the U because those are bias data that are forced out of the model. Where we want to be on the test MSE is near the bottom of the U shape.

Our Square Bias curves is the inverse of the variance curve as the smaller the variance, the more bias the model as. The reason for that is if we are really forcing our model to produce a specific output without letting the model the value naturally, then the model could be overfit and skewed.

The variance curve scales with the flexibility that we allow with our model. So, say we have a very strict model (value of 2), what is the variance of 2? 2 is a constant (very strict model) so the variance is 0. However, if we have a very flexible model then our variance would be very large.

Irreducible Error is a straight line as this is a constant number of errors that cannot be removed by our model.

d. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression?

The advantage of a very flexible model is that it may produce a better fit for non-linear models, and it has the possibility of reducing the bias.

However, a very flexible model can produce a lot of noise, produce extra estimation of parameters, and increase the variance.

Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

When the data has a non-linear character (like voltage usage dataset) it is best to use a more flexible model.

Strict model should be used on model that is linear.