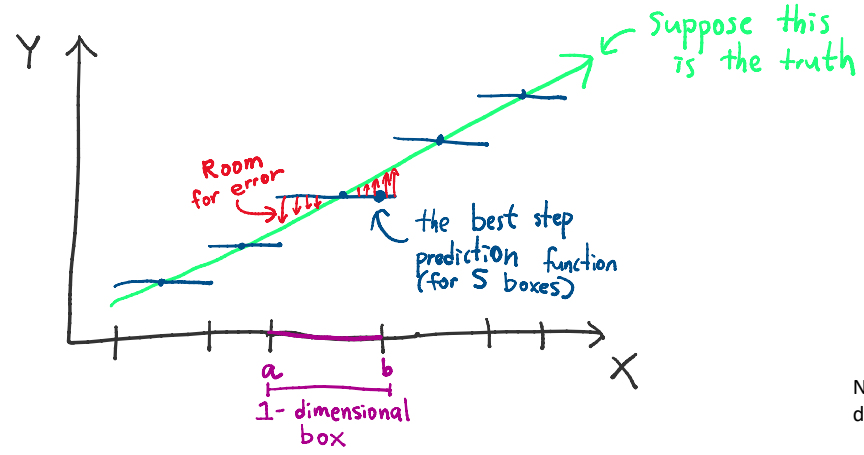
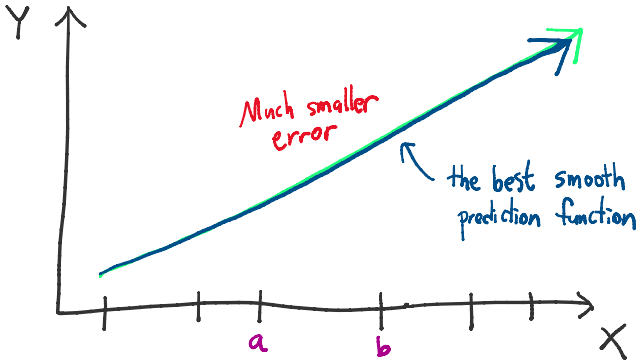
RF underperforms when?

Saturday, May 1, 2021 7:45 PM

A random forest works by partitioning the p-dimensional predictor space into p-dimensional boxes. Any p-dimensional observation which falls into this box is assigned a single value no matter where in the box it is. This is like a p-dimensional step function. But if the underlying relationship between the predictors and the response is truly a smooth, p dimensional surface, then a prediction function that is a step function will always perform worse than a prediction function which is a smooth function. We can see this easily in the p=1 dimensional case as I illustrate here:





Note that the intervals here (such as (a,b) shown) are really 1-dimesional boxes.

In the case with one predictor (p=1), the prediction surface is a line which approximates a smooth truth better than a step function would. This is because within any given step function interval, predictions will map more closely to the truth in a smooth prediction function than for a step prediction function. For our project we likely have a higher dimensional analog of this phenomenon.

The step function could be improved by increasing the number of boxes, but it will never outperform the a smooth prediction function, such as what is achieved in regularization. So, it is possible that the underlying relationship between the predictors and the response in our project is very smooth, in which case regularization methods would outperform the random forest.