STAD80: Final Project

Proposal

Shawn Santhoshgeorge (1006094673) Anaqi Amir Razif (1005813880)

19 March 2023

The conventional wisdom in machine learning is that larger models and more data is better as it should better approximate the true distribution of the data thus allowing it to perform well on future data. Thus creating overparametrized models, were the number of the parameters (d) well exceeds the number of samples (n) like in Neural Networks. These models tend to exhibit a “double-descent” phenomenon were, as the number of samples increase the performance is non-monotonic. The paper we intent on reviewing “More Data Can Hurt for Linear Regression: Sample-wise Double Descent” (Nakkiran, 2019) looks deeper into this phenomenon and try to understand why this may occur. This papers looks at the Ordinary Least Squares Solution with a constant parameter size (d) but with varying sample size (n) with a particular focus on when $n \approx d$ since this the regime were the noise starts to really affect the model due to poor conditioning of the data matrix.

There seems to be a few specific areas this paper explores regarding the issue which are Excess Risk and Bias-Variance Trade off, Conditioning of the Data Matrix which looks at why the data matrix is well-conditioned at $n << d$ but is not well-conditioned at $n \approx d$ by looking at points of criticality, the change in variance when increasing sample size. We intend on analyzing the results from these section to get a better understanding of the phenomenon. We intend on testing the conclusions by creating a experiment using R based on the way the problem was setup in the paper to recreate the “Double Descent” phenomenon and look into if adding a regularization term like in Ridge Regression will help to minimize this from happening.