

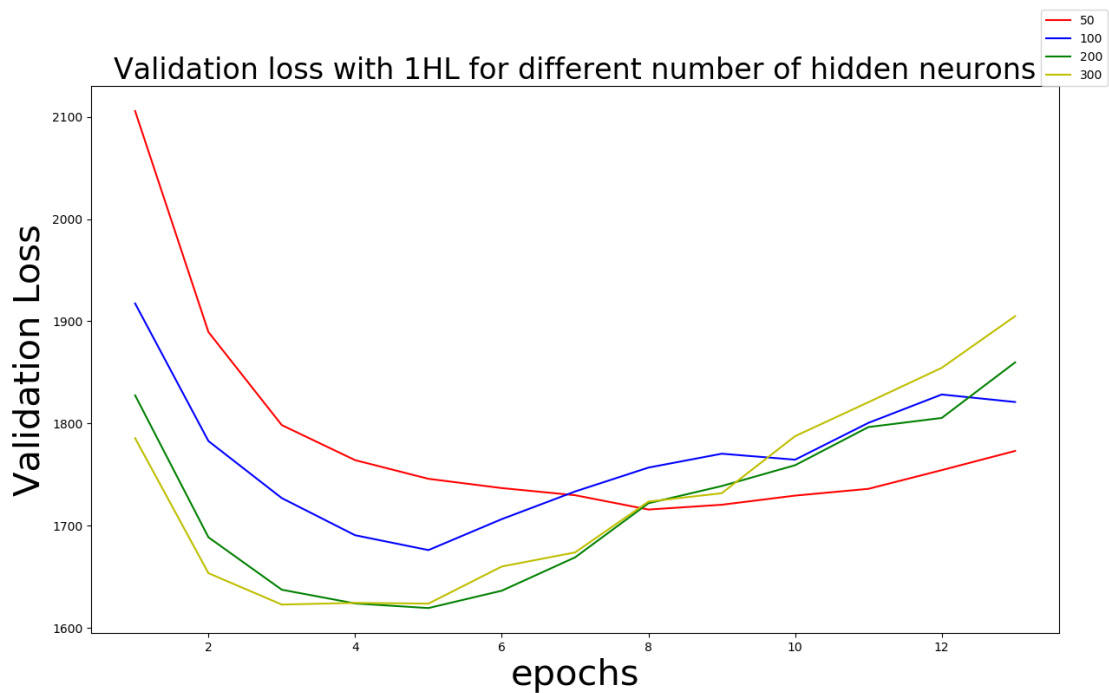
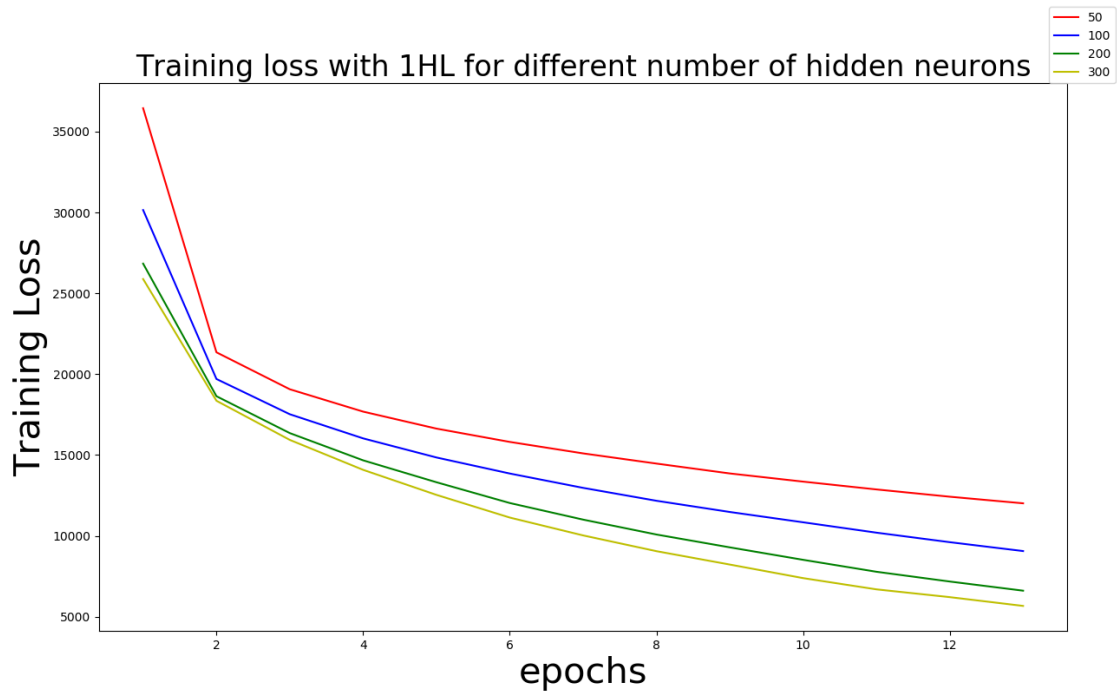
Deep Learning

Programming Assignment # 1

Author:
Pawandeep Singh CS17S027
Sidharth Aggarwal CS17S012

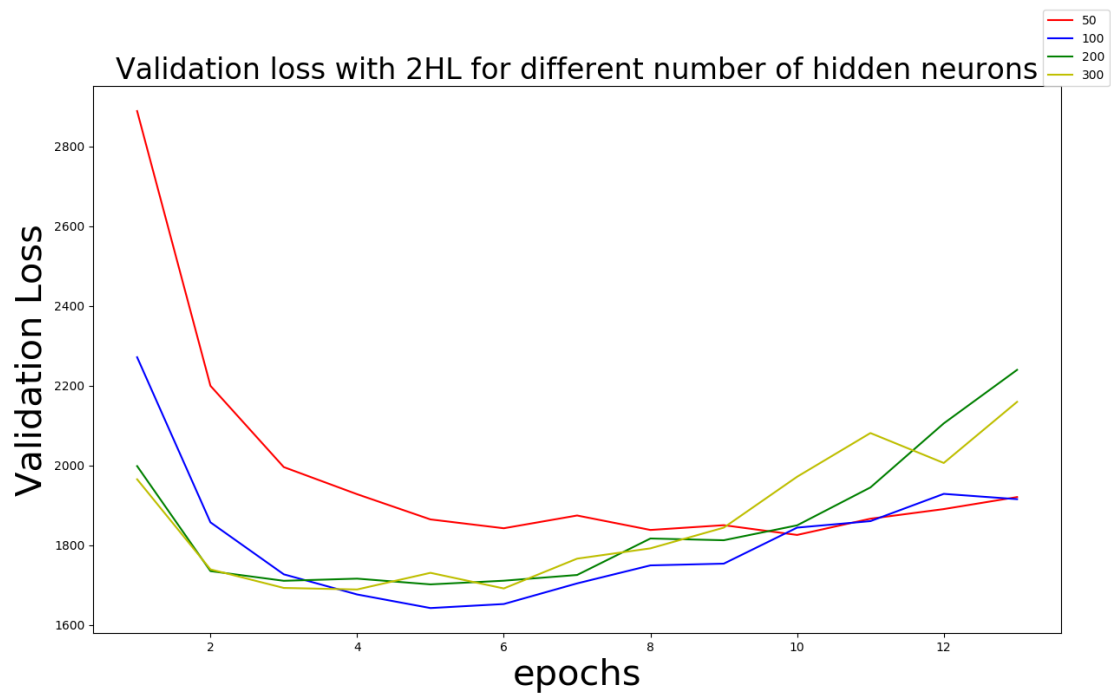
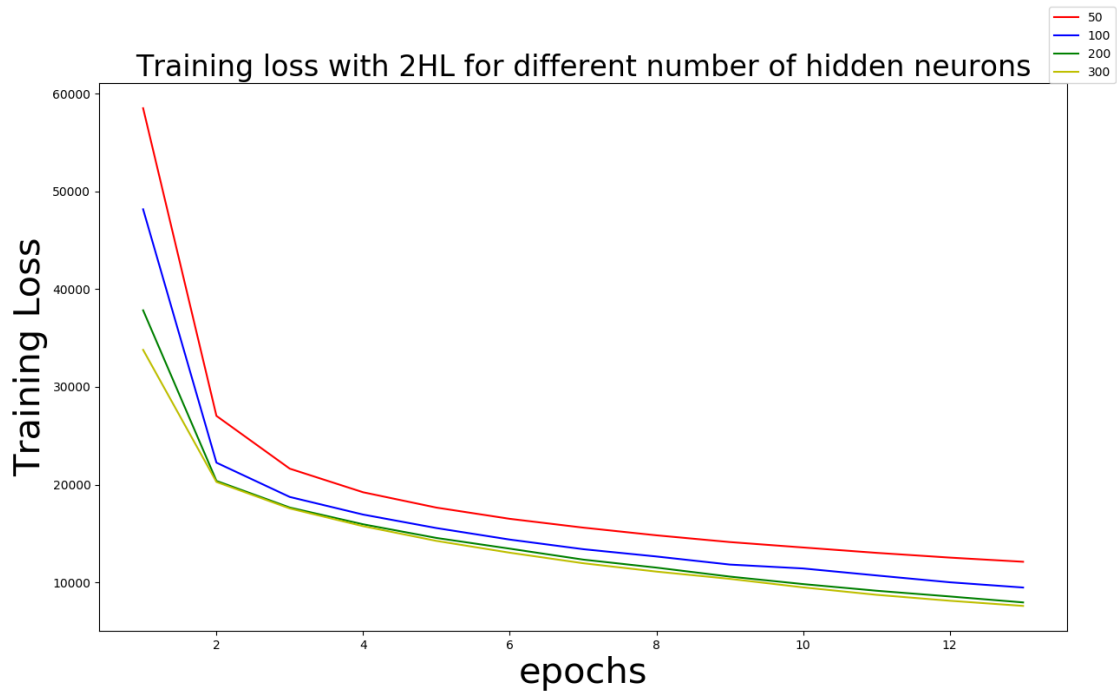
Professor:
Mitesh Khapra

1 Plots for 1 hidden layer



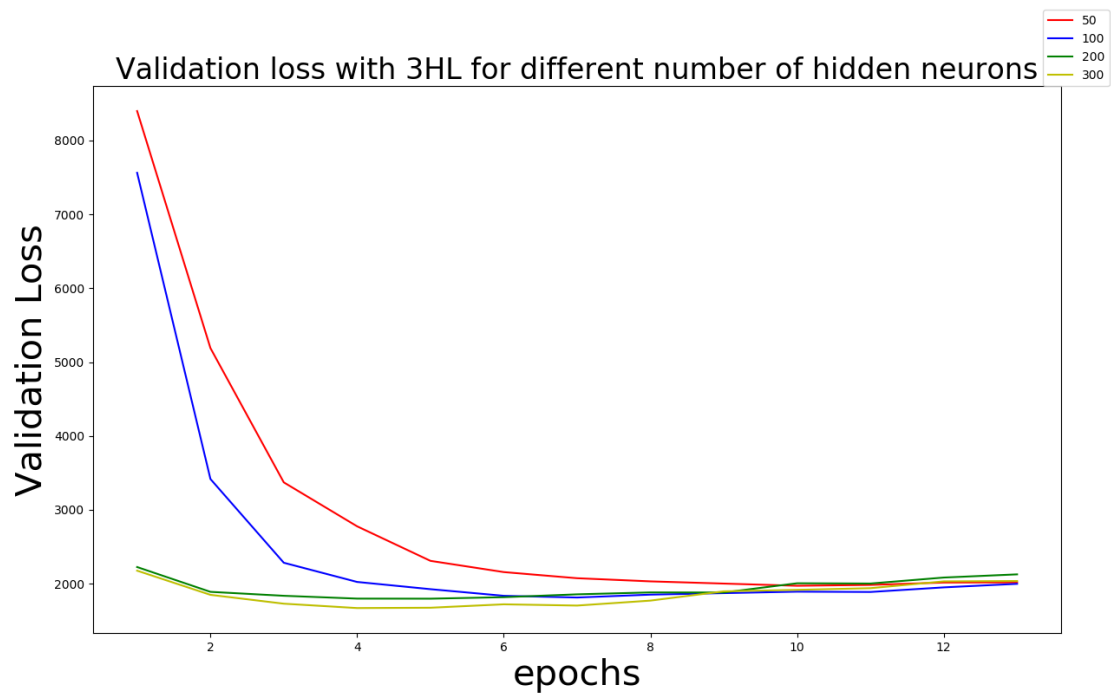
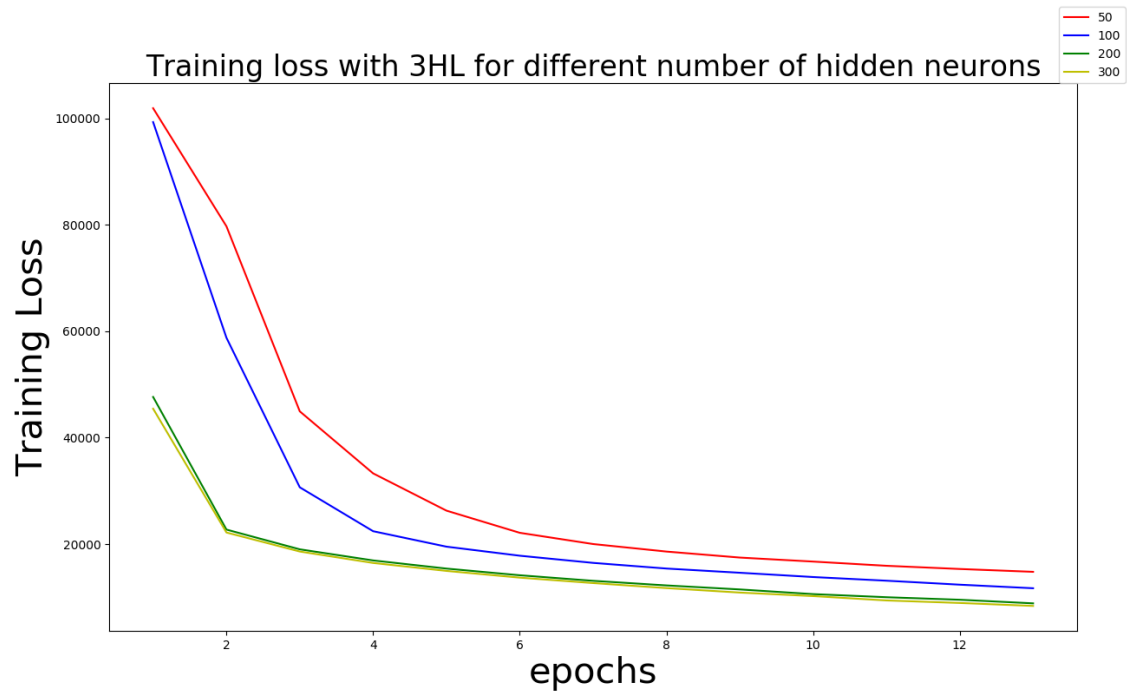
Observations and inferences: Training loss decreases with increasing number of hidden layers as the model becomes more complex. Validation loss is high initially as the model is not trained yet. It decreases with increasing epoch count but it increases afterwards denoting over fitting. Best model for 1HL among these is with 200 neurons per layer.

2 Plots for 2 hidden layer



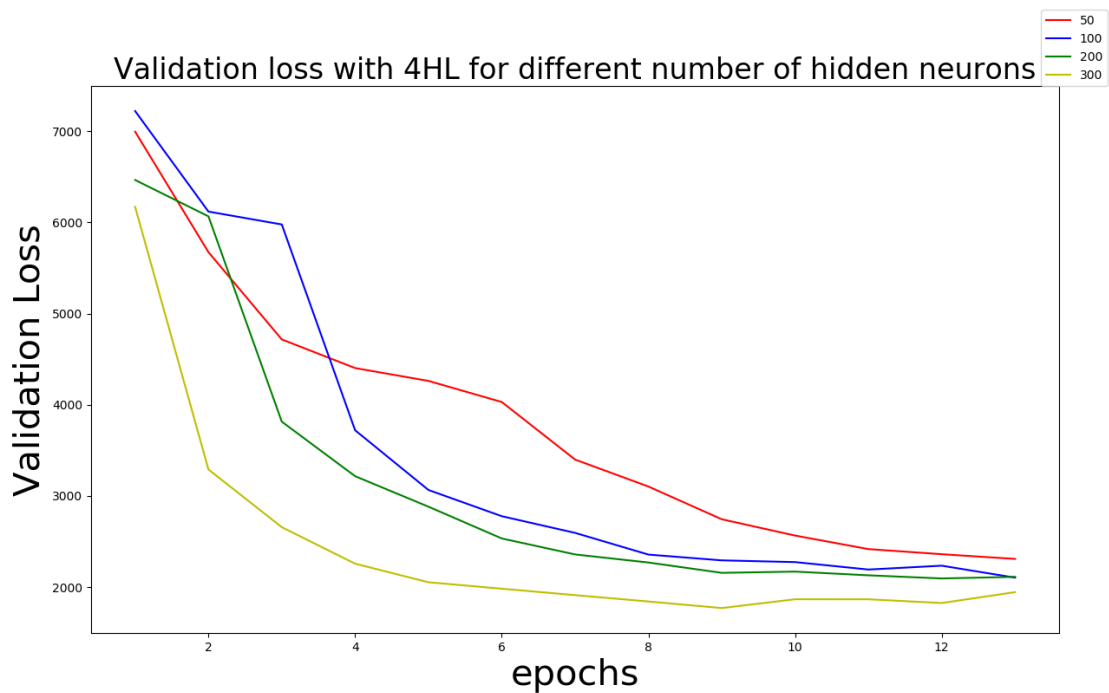
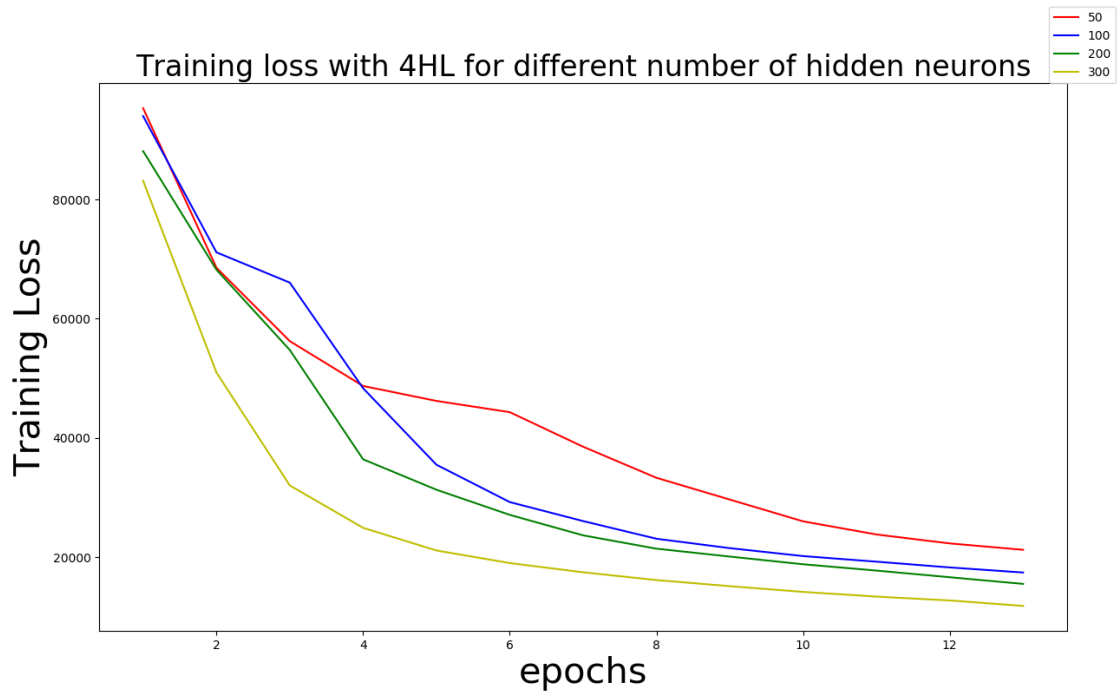
Observations and inferences: Training loss decreases with increasing number of hidden layers as the model becomes more complex. Validation loss is high initially as the model is not trained yet. It decreases with increasing epoch count but it increases afterwards denoting over fitting. Best model for 2HL among these is with 100 neurons per layer.

3 Plots for 3 hidden layer



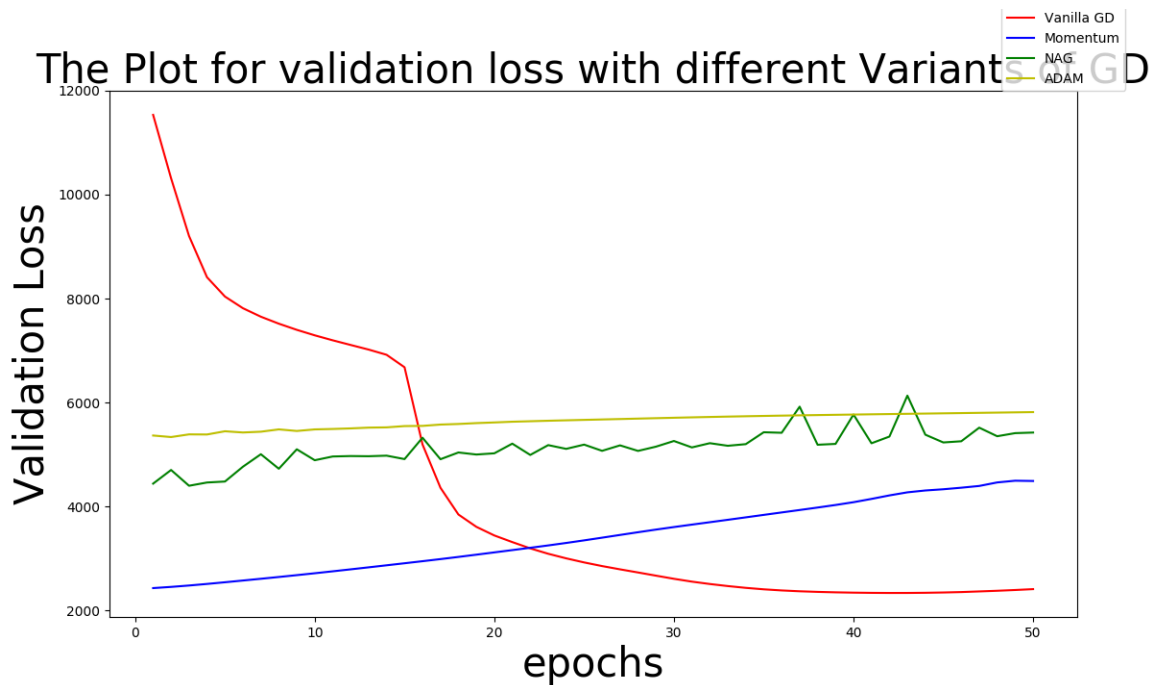
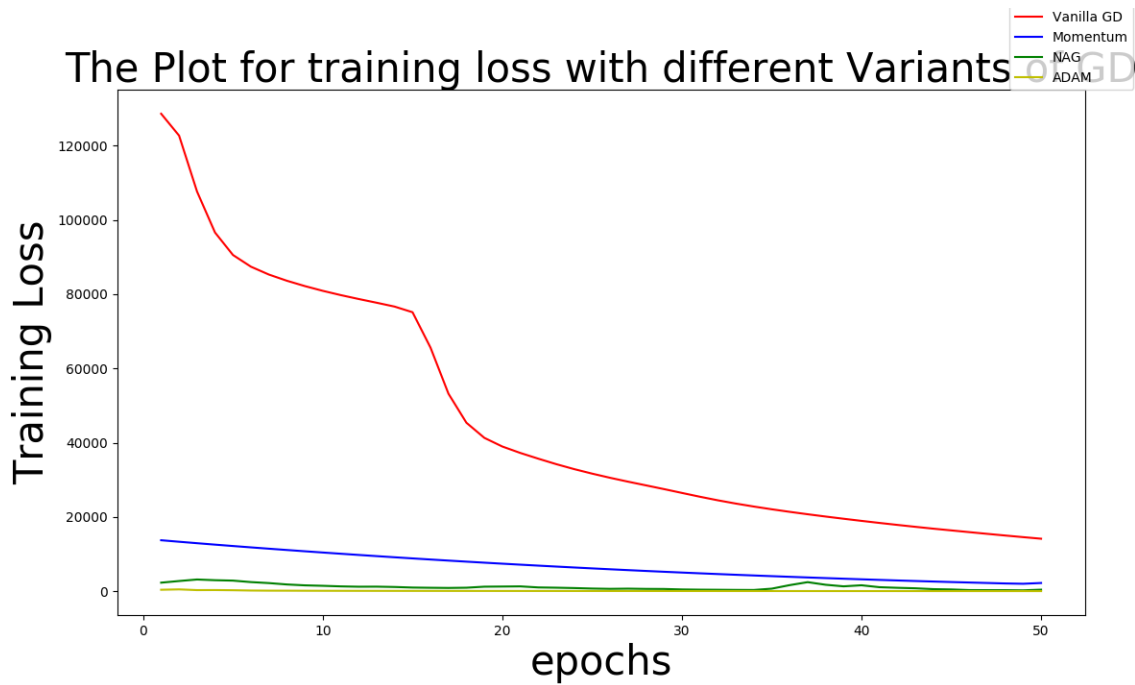
Observations and inferences: Training loss decreases with increasing number of hidden layers as the model becomes more complex. Validation loss is high initially as the model is not trained yet. It decreases with increasing epoch count but it increases afterwards denoting over fitting. Best model for 3HL among these is with 300 neurons per layer.

4 Plots for 4 hidden layer



Observations and inferences: Training loss decreases with increasing number of hidden layers as the model becomes more complex. Validation loss is high initially as the model is not trained yet. It decreases with increasing epoch count but it increases afterwards denoting over fitting. Best model for 4HL among these is with 300 neurons per layer.

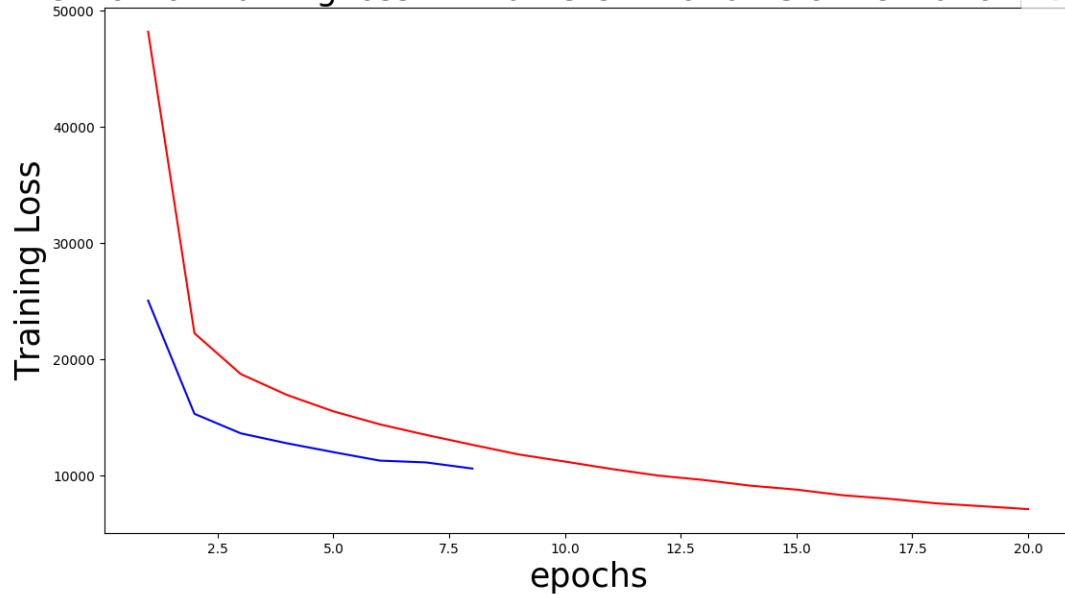
5 Variants of Gradient Descent



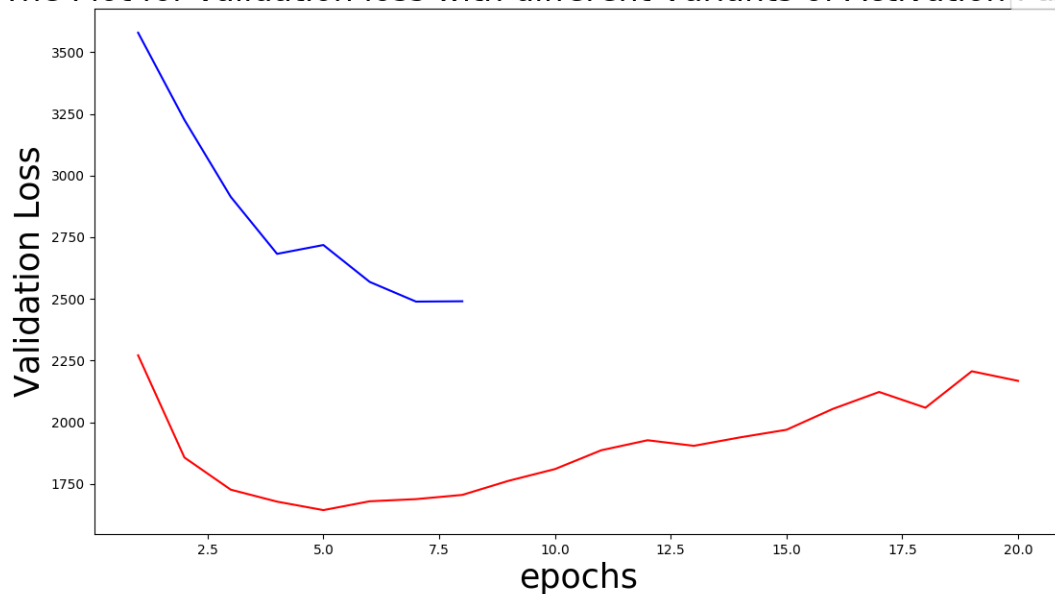
Observations and Inferences: During training Vanilla GD is outperformed by all the other GD variants. For NAG training loss increases few times as it may have overshoot the minimum.

6 Sigmoid Vs Tanh

The Plot for training loss with different Variants of Activation Function

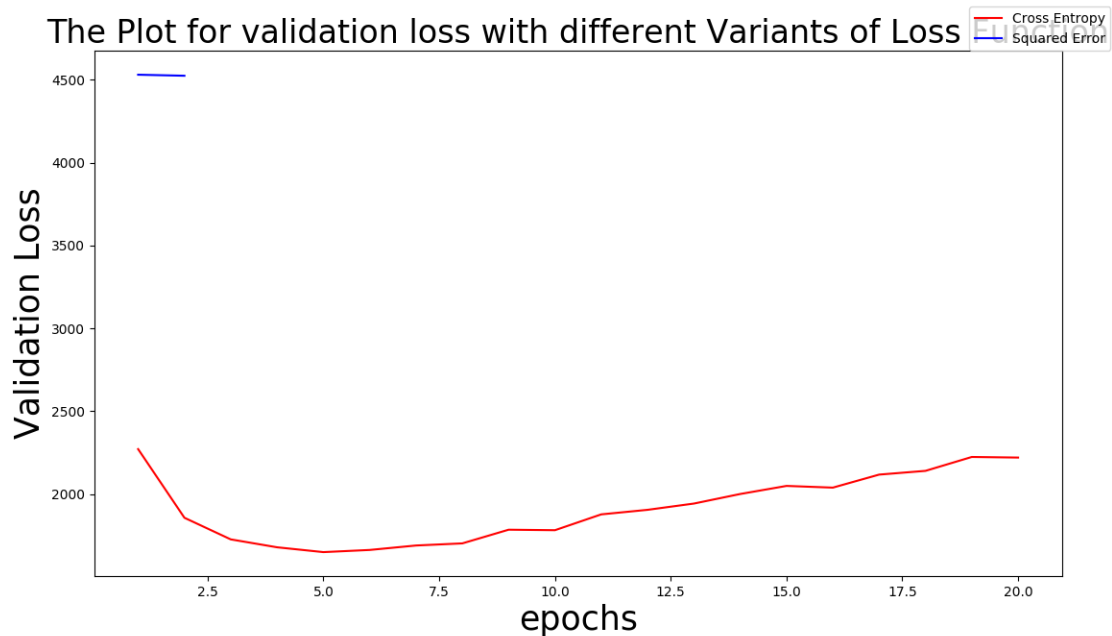
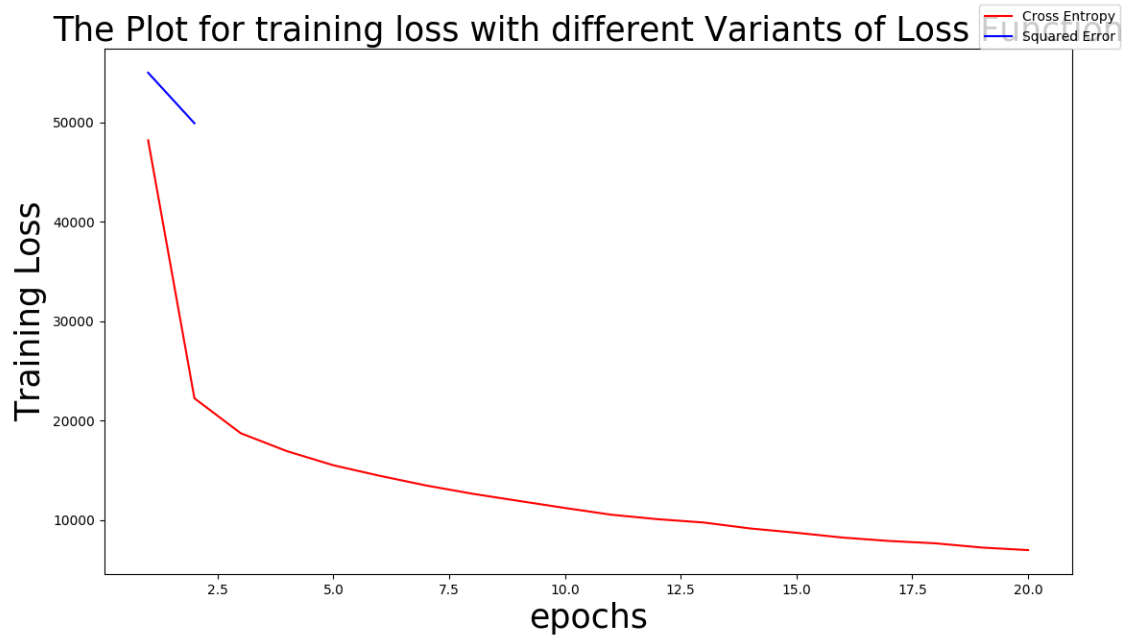


The Plot for validation loss with different Variants of Activation Function



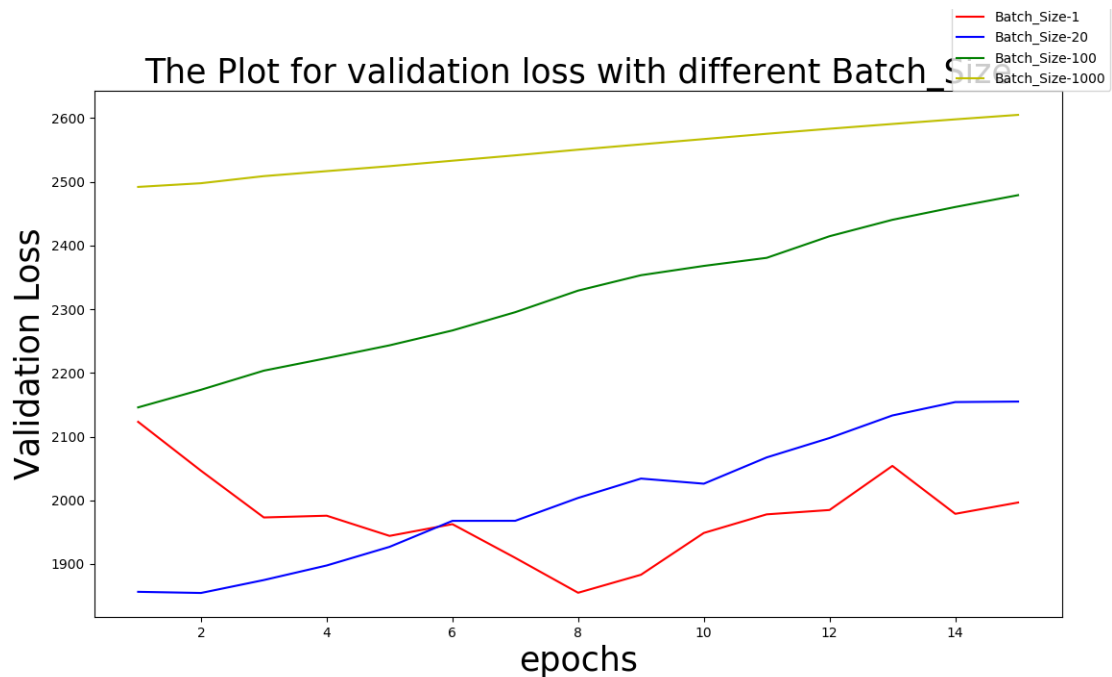
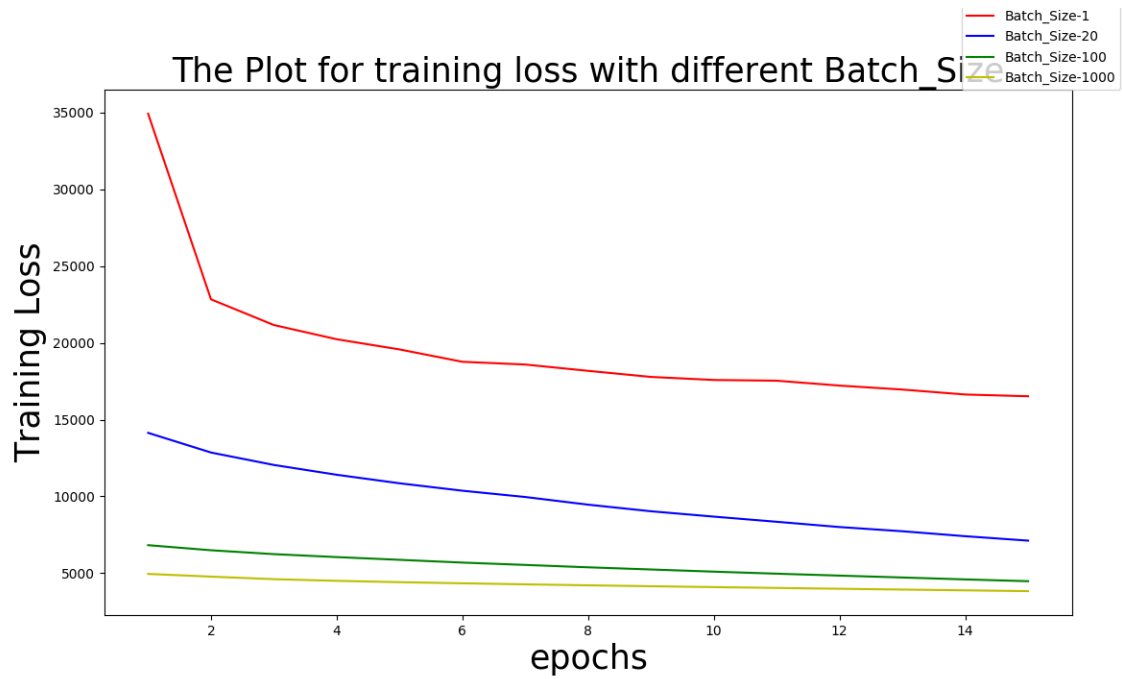
Observations and Inferences: Training loss for Tanh is lesser than Sigmoid function but for validation Loss Sigmoid is better than Tanh. We have not plotted Tanh after few epoch values for both training and validation loss as it's value was NaN.

7 Cross Entropy Vs Squared Loss



Observations and Inferences: Here it is difficult to judge which loss function is better since we don't have Loss corresponding to Squared Loss after few epoch iterations since it came out to be NAN.

8 Different batch Size



Observations and Inferences: During training as the batch size increases gradient increases and hence loss decreases at higher rate but 1 epoch time for bigger batch size is more.