CS492 CLT Proposal - Benign Overfitting without Linearity

Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data

Team 4 Jaeryeong Kim, Yujun Kim

1 Motivation for the Selection of the Topic

Classical principle in statistical learning is not to make model of prediction to complicated. Due to the bias-variance trade off, more complex your model gets, more variance it produces, reducing the generalization power of the model. However, recently there has been discovery of phenomenon so called "Benign overfitting". With this phenomenon, the interpolating model also generalize as well. There has been a series of works to theoretically explain this phenomenon.

As to use this good phenomenon of good generalization, we need understand on which setting this happens. For the kernel regression problem, the phenomenon occurs depending on the type of kernel and the number of features we use. There are several research explaining benign overfitting for linear or kernel regression.

2 Work in This Paper

In contrast, this paper deals with finite width one hidden layer neural networks trained for logistic loss to solve a classification problem. Although it has a limitation of learning only the weight matrix connecting the input layer and the hidden layer, it allows the analysis of benign overfitting for nonlinear model. The nonlinearity arises from the activation function. The main theorem is stated under interpolating setting where the dimension of input is large compared to the number of samples. It states that with reasonable additional assumptions on the distribution of samples, minimization of empirical risk through gradient descent leads to reduction of population risk as well.

3 Further Investigation Direction

- It would be interesting to further investigate on which practical distribution satisfies the assumption used in the theorem, and empirically check whether benign overfitting happens. We may implement the generator to produce samples ourselves, and train the model with gradient descent algorithm to reproduce the experimental result. For kernel regression problems, plotting the prediction value visually demonstrates the generalization. For a classification problem, we can empirically check a train and test accuracy, and visualize the decision boundary of the classifier to demonstrate the generalization from benign overfitting.
- In addition, finding a real world dataset for a classification where its trained model exhibits benign overfitting may give hints on the relaxation of assumption in the theorem.
- Theoretically, this research still has quiet strong assumption in the sense of dimension of input(p) compared to the number of data point(n) satisfies $p \gg n$. Understanding sufficient parameters to fully memorize as $mp > n$ where m is the width of the network, there is still a regime of interpolation that lacks theoretical guarantees.
- We may theoretically extend the benign overfitting conditions with neural networks trained with different optimization algorithms other than GD.

4 References

1. S. Frei, N. S. Chatterji, and P. L. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data, *Annual Conference on Learning Theory 2022*

2. A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. Preprint, arXiv:2009.14286, 2020. *Journal of Machine Learning and Research (JMLR), 23(123):1-76, 2023*

3. S. Frei, G. Vardi, P. Bartlett, and N. Srebro. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. *The Thirty Sixth Annual Conference on Learning Theory (PMLR), (pp. 3173-3228), 2023*