Generalization and Optimization of Interpolating Models

Niladri S. Chatterji, Philip M. Long, Peter L. Bartlett

The recent success of neural network models has shone light on a rather surprising statistical phenomenon: statistical models that perfectly fit noisy data can generalize well to unseen test data. Understanding this phenomenon of benign overfitting has attracted intense theoretical and empirical study.

To understand this phenomenon we studied three different problems.

In the first problem, we studied the training of finite-width two-layer smoothed ReLU networks for binary classification using the logistic loss [2]. We showed that gradient descent drives the training loss to zero if the initial loss is small enough. When the data satisfies certain cluster and separation conditions and the network is wide enough, we showed that one step of gradient descent reduces the loss sufficiently that the first result applies. In contrast, all past analyses of fixed-width networks that we knew of did not guarantee that the training loss goes to zero.

Next we extended these results to deep networks [3]. Specifically, we established conditions under which gradient descent applied to fixed-width deep networks drives the logistic loss to zero, and proved bounds on the rate of convergence. Our analysis applies for smoothed approximations to the ReLU, such as Swish and the Huberized ReLU, proposed in previous applied work. We provided two sufficient conditions for convergence. The first is simply a bound on the loss at initialization. The second is a data separation condition used in prior analyses.

Finally, in the last problem we study generalization [1]. We considered interpolating two-layer linear neural networks trained with gradient flow on the squared loss and derived bounds on the excess risk when the covariates satisfy sub-Gaussianity and anti-concentration properties, and the noise is independent and sub-Gaussian. By leveraging recent results that characterize the implicit bias of this estimator, our bounds emphasize the role of both the quality of the initialization as well as the properties of the data covariance matrix in achieving low excess risk.

References

[1] Niladri S Chatterji, Philip M Long, and Peter L Bartlett. "The Interplay Between Implicit Bias and Benign Overfitting in Two-Layer Linear Networks". In: *arXiv preprint arXiv:2108.11489* (2021).

[2] Niladri S Chatterji, Philip M Long, and Peter L Bartlett. "When Does Gradient Descent with Logistic Loss Find Interpolating Two-Layer Networks?" In: *Journal of Machine Learning Research* 22.159 (2021), pp. 1–48.

[3] Niladri S Chatterji, Philip M Long, and Peter L Bartlett. "When does gradient descent with logistic loss interpolate using deep networks with smoothed ReLU activations?" In: *Proceedings of the 34th Conference on Learning Theory*. 2021.