Spencer Frei: Implicit regularization and benign overfitting for neural networks in high dimensions

In this talk we go over some of our recent work on understanding optimization and generalization in neural networks trained by gradient descent. We focus on the setting of two-layer neural networks that are trained by the logistic loss for a binary classification task. In the first part of the talk, we show that when the training data is sufficiently high-dimensional, gradient descent has an implicit bias towards low-rank solutions: although the weight matrix is full rank at (random) initialization, gradient descent rapidly reduces the rank to something that is independent of the number of rows and columns of the weight matrix. We show how the choice of the initialization variance has a significant effect on this low-rank implicit bias of gradient descent. In the second part of the talk, we consider a particular distributional setting where we show that the trained network exhibits 'benign overfitting': the network perfectly fits noisy training data and simultaneously generalizes near-optimally. Our results hold when the training data is sufficiently high-dimensional, and we discuss experiments which suggest this assumption is essential for overfitting to be 'benign'. Based on joint work with Peter Bartlett, Niladri Chatterji, Wei Hu, Nati Srebro, and Gal Vardi.