On the Implicit Bias of Gradient Descent with Linear Teachers

Abstract:

Artificial neural networks have recently revolutionized the field of machine learning. However, we still do not have sufficient theoretical understanding of how such models can be successfully learned. Two specific questions in this context are: how can neural nets be learned despite the non-convexity of the learning problem, and how can they generalize well despite often having more parameters than training data. I will describe our recent work showing that gradient-descent optimization indeed leads to "simpler" models, where simplicity is captured by lower weight norm and in some cases clustering of weight vectors. We demonstrate this for several teacher and student architectures, including learning linear teachers with ReLU networks, learning boolean functions and learning convolutional pattern detection architectures. I will also discuss results on fine-tuning of large models (e.g., BERT), and what affects generalization in these cases.