The implicit bias phenomenon in deep learning

Holger Rauhut

Abstract

Deep neural networks are usually trained by minimizing a non-convex loss functional via (stochastic) gradient descent methods. Unfortunately, the convergence properties are not very well- understood. Moreover, a puzzling empirical observation is that learning neural networks with a number of parameters exceeding the number of training examples often leads to zero loss, i.e., the network exactly interpolates the data. Nevertheless, it generalizes very well to unseen data, which is in stark contrast to intuition from classical statistics which would predict a scenario of overfitting. A current working hypothesis is that the chosen optimization algorithm has a significant influence on the selection of the learned network. In fact, in this overparameterized context there are many global minimizers so that the optimization method induces an implicit bias on the computed solution. It seems that gradient descent methods and their stochastic variants favor networks of low complexity (in a suitable sense to be understood), and, hence, appear to be very well suited for large classes of real data. Initial attempts in understanding the implicit bias phenomen considers the simplified setting of linear networks, i.e., (deep) factorizations of matrices. This has revealed a surprising relation to the field of sparse and low rank recovery (compressive sensing) in the sense that gradient descent favors sparse diagonal or low rank matrices in certain situations. Moreover, initial results on learning two-layer ReLU networks show that sparse ReLU-expansions may be favored by gradient flow. Despite such initial theoretical results on simplified scenarios, the understanding of the implicit bias phenomenon in deep learning is widely open.

Based on joint works with El Mehdi Achour, Hung-Hsu Chou, Cristina Cipriani, Johannes Maly, Maria Matveev, Ulrich Terstiege, Rachel Ward