MANUELA GIROTTI, Saint Mary's University

Neural Networks Efficiently Learn Low-Dimensional Representations with SGD

We study the problem of training a two-layer neural network (NN) of arbitrary width using stochastic gradient descent (SGD) where the input $\boldsymbol{x} \in \mathbb{R}^d$ is Gaussian and the target $y \in \mathbb{R}$ follows a multiple-index model, i.e., $y = g(\langle \boldsymbol{u_1}, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{u_k}, \boldsymbol{x} \rangle)$ with a noisy link function g. We prove that the first-layer weights of the NN converge to the k-dimensional *principal subspace* spanned by the vectors $\boldsymbol{u_1}, \ldots, \boldsymbol{u_k}$ of the true model, when online SGD with weight decay is used for training. This phenomenon has several important consequences when $k \ll d$. First, by employing uniform convergence on this smaller subspace, we establish a generalization error bound of $O(\sqrt{kd/T})$ after T iterations of SGD, which is independent of the width of the NN. We further demonstrate that, SGD-trained ReLU NNs can learn a single-index target of the form $y = f(\langle \boldsymbol{u}, \boldsymbol{x} \rangle) + \epsilon$ by recovering the principal direction, with a sample complexity linear in d (up to log factors), where f is a monotonic function with at most polynomial growth, and ϵ is the noise. This is in contrast to the known $d^{\Omega(p)}$ sample requirement to learn any degree p polynomial in the kernel regime, and it shows that NNs trained with SGD can outperform the neural tangent kernel at initialization. Finally, we also provide compressibility guarantees for NNs using the approximate low-rank structure produced by SGD.

This is a joint work with Alireza Mousavi-Hosseini (UofT, Vector), Sejun Park (Korea Univeristy), Ioannis Mitliagkas (UdeM, Mila), and Murat A. Erdogdu (UofT, Vector).