Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers

Zeyuan Allen-Zhu (MS Research AI), Yuanzhi Li (Stanford), Yingyu Liang (UW-Madison)

Deep learning

- Great empirical success vs. largely open theoretical questions

- Key theoretical questions:
- What functions can neural networks provably learn?
- Why do overparameterized networks (found by those training algorithms) generalize?

What functions can neural networks provably learn?

- We prove an important concept class that contains three-layer (resp. two-layer) neural networks equipped with smooth activations can be efficiently learned by three-layer (resp. two-layer) ReLU neural networks via SGD or its variants.

- First to show using hidden layers to learn two/three-layer NNs with non-trivial activation functions
- Three-layer result proved with sophisticated non-convex interactions across layers

Why do overparameterized networks generalize?

- Our explanation: with larger overparameterization, one can hope to learn better target functions with possibly larger size, more complex activations, smaller risk of the target function, and to a smaller error

Proof ideas

- **(Approximation)** Good networks with small risks are plentiful: with high probability over random initialization, there exists a good network in the close neighborhood of any point on the SGD training trajectory
- **(Optimization)** The optimization in overparameterized neural networks has benign properties: essentially along the training trajectory, there is no second-order critical points for learning three-layer networks, and no first-order critical points for two-layer
- **(Generalization)** In the learned networks, information is also evenly distributed among neurons. This structure allows a new generalization bound that is (almost) independent of the number of neurons