Over-parameterized nonlinear learning: Gradient descent follows the shortest path?

Mahdi Soltanolkotabi, University of Southern California

Many modern learning tasks including deep neural networks involve fitting highly nonlinear models to data which are trained in an over-parameterized regime where the parameters of the model exceed the size of the training dataset. A major challenge is that training these models correspond to extremely high-dimensional and nonconvex optimization problems and it is not clear how to provably solve them to global optimality. Furthermore, due to over-parameterization, the training loss may have infinitely many global minima and it is critical to understand the properties of the solutions found by first-order optimization schemes such as (stochastic) gradient descent starting from different initial estimates. In this talk I will show that (stochastic) gradient methods have a few intriguing properties: 1) the iterates converge at a geometric rate to a global optima despite the nonconvex nature of the landscape, (2) among all global optima of the loss the iterates converge to one with a near minimal distance to the initial estimate, (3) the iterates take a near direct route from the initial estimate to this global optima.
I will demonstrate the utility of our general theory for a variety of problem domains spanning low-rank matrix recovery to neural network training. (This talk is based on joint work with Samet Oymak)