Benign underfitting of Stochastic Gradient Descent, and some other peculiarities

Abstract:

Why do gradient methods, such as Stochastic Gradient Descent (SGD), generalize well even when used to train predictors with far more parameters than training samples, and without any explicit form of capacity control (regularization)? In this talk, I will revisit this elusive question in the context of (stochastic) convex optimization and discuss some surprising theoretical results and phenomena that challenge the current conventional wisdom in machine learning. Based on joint works with Idan Amir, Amit Attia, Roi Livni, Yishay Mansour, Uri Sherman and Matan Schlizerman.