Stanford University
Department of Statistics

Departmental Seminar

4:30pm, Tuesday, November 6, 2018
Sloan Mathematics Center Room 380C

Refreshments served at 4pm in Sequoia Lounge.

Speaker: Jason Lee, *USC*

Title: Foundations of Deep Learning: SGD, Overparametrization, and Generalization

Abstract:

We provide new results on the effectiveness of SGD and overparametrization in deep learning.

a) SGD: We show that SGD converges to stationary points for general nonsmooth, nonconvex functions, and that stochastic subgradients can be efficiently computed via Automatic Differentiation. For smooth functions, we show that gradient descent, coordinate descent, ADMM, and many other algorithms, avoid saddle points and converge to local minimizers. For a large family of problems including matrix completion and shallow ReLU networks, this guarantees that gradient descent converges to a global minimum.

b) Overparametrization: For a k hidden-node shallow network with quadratic activation and n training data points, we show as long as $k \geq \sqrt{(2n)}$, over-parametrization enables local search algorithms to find a *globally* optimal solution. Further, despite the fact that the number of parameters may exceed the sample size, we show that with weight decay the solution also generalizes well.

   For general neural networks, we establish a margin-based theory. The minimizer of the cross-entropy loss with weak regularization is a max-margin predictor, and enjoys stronger generalization guarantees as the amount of overparametrization increases.

c) Next, we analyze the implicit regularization effects of various optimization algorithms on overparametrized networks. In particular we prove that for least squares with mirror descent, the algorithm converges to the closest solution in terms of the Bregman divergence. For linearly separable classification problems, we prove that the steepest descent with respect to a norm solves SVM with respect to the same norm. For overparameterized non-convex problems such as matrix sensing or neural net with quadratic activation, we prove that gradient descent converges to the minimum nuclear norm solution, which allows for both meaningful optimization and generalization guarantees.