Thesis Defense

MACHINE LEARNING DEPARTMENT

Gradient Descent for Non-convex Problems in Modern Machine Learning

Abstract:

Machine learning has become an important toolset for artificial intelligence and data science across many fields. A modern machine learning methods can be often reduced to a mathematical optimization problem. Among algorithms to solve the optimization problem, gradient descent and its variants like stochastic gradient descent and momentum methods are the most popular ones. The optimization problem induced from classical machine learning methods is often a convex and smooth one, for which gradient descent is guaranteed to solve it efficiently. On the other hand, modern machine learning methods, like deep neural networks, often require solving a non-smooth and non-convex problem. Theoretically, non-convex mathematical optimization problems cannot be solved efficiently. However, in practice, gradient descent and its variants can find a global optimum efficiently. These competing facts show that often there are special structures in the optimization problems that can make gradient descent succeed in practice. This talk presents technical contributions to fill the gap between theory and practice on the gradient descent algorithm. The outline of the thesis is as follows.

In the first part, we consider applying gradient descent to minimize the empirical risk of a neural network. We will show if a multi-layer neural network with smooth activation function is sufficiently wide, then randomly initialized gradient descent can efficiently find a global minimum of the empirical risk. We will also show the same result for the two-layer neural network with Rectified Linear Unit (ReLU) activation function. It is quite surprising that although the objective function of multi-layer neural networks is non-convex, gradient descent can still find their global minimum.

In the second part, we study conditions under which gradient descent fails. We will show gradient descent can take exponential time to optimize a smooth function with the strict saddle point property for which the noise-injected gradient can optimize in polynomial time.

Speaker:

Simon Du

Thesis Committee:

Barnabas Poczos (Co-Chair)
Aarti Singh (Co-Chair)
Ruslan Salakhutidnov
Michael I. Jordan (UC, Berkeley)

Apr. 23, 2019
4:00pm
GHC 4405

Link to draft document:
https://www.dropbox.com/s/bmzh40xf6faj0id/du_dissertations.pdf?dl=0