The Unreasonable Effectiveness of Gradient Descent

John Lafferty

Abstract

In Eugene Wigner's classic 1960 essay "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," the uncanny success of mathematics in modeling the physical world is likened to a person who has a large collection of keys, and who opens many doors in succession by always finding the right key on the first or second try. Gradient descent is acting like a master key across many machine learning applications, so it is important to better understand its success, as well as its limitations. In this talk we present work that sheds some light on the effectiveness of gradient descent in machine learning.

First, we present results on the optimization landscape of certain machine learning problems related to low rank approximation, semidefinite programming, and matrix completion, where the optimality of gradient descent can be well understood. Next, we consider the problem of characterizing the complexity of minimizing a specific convex function. This is tricky to formalize,as traditional complexity analysis is expressed in terms of the worst case over a large class of instances. We extend the classical minimax analysis of stochastic convex optimization by introducing a localized form of complexity for individual functions. In this setting gradient descent can be shown to be locally adaptive to the geometry of the function. Finally, we present work that shows how modified gradient descent algorithms can effectively handle non-convex optimization for deep neural networks, by "surfing" over the evolving optimization surface as learning is carried out.

Along the way, we highlight work of others that characterizes the landscape of learning problems for which gradient descent is effective, and that extends the usefulness of gradient descent to other statistical inference problems.