November 16, 2018

STUDENT SEMINAR

UW-Madison | Department of Statistics | www.stat.wisc.edu

Title

A PDE Perspective of Stochastic Gradient Descent in Deep Learning

Speaker

Yuhua Zhu

(Ph.D. candidate in Mathematics, UW-Madison)

Time & Place

Friday, Nov 16, 3pm, SMI **133**

Snacks @2:45pm, SMI **133**

THE UNIVERSITY of WISCONSIN MADISON

Abstract

Stochastic gradient descent (SGD) is almost ubiquitously used for training nonconvex optimization tasks including deep neural networks. Recently, a hypothesis that *large batch SGD tends to converge to sharp minimizers* has received increasing attention by machine learning researchers. We theoretically justify this hypothesis by providing new properties of SGD in both finite-time and asymptotic regime, with the tools from Partial Differential Equations (PDE). In particular, we give an explicit escaping time of SGD from a local minimum in the finite-time regime and prove that SGD tends to converge to flatter minima in the asymptotic regime (although may take exponential time to converge) regardless of the batch size. We also find that SGD with a larger learning rate to batch size ratio tends to converge to a flat minimum faster, however, its generalization performance could be worse than the SGD with a smaller learning rate to batch size ratio. We include experiments to corroborate these theoretical findings.