UNIVERSITY AT ALBANY
State University of New York

College of Arts and Sciences
Department of Mathematics and Statistics

Colloquium

Chang-Han Rhee
Northwestern University

Heavy-Tailed Large Deviations and Sharp Characterization of Global Dynamics of SGDs in Deep Learning

Friday, March 8, 2024
3:00 p.m. in Massry 221

Abstract. While the typical behaviors of stochastic systems are often deceptively oblivious to the tail distributions of the underlying uncertainties, the ways rare events arise are vastly different depending on whether the underlying tail distributions are light-tailed or heavy-tailed. Roughly speaking, in light-tailed settings, a system-wide rare event arises because everything goes wrong a little bit as if the entire system has conspired up to provoke the rare event (conspiracy principle), whereas, in heavy-tailed settings, a system-wide rare event arises because a small number of components fail catastrophically (catastrophe principle). In the first part of this talk, I will introduce the recent developments in the theory of large deviations for heavy-tailed stochastic processes at the sample path level and rigorously characterize the catastrophe principle for such processes.

The empirical success of deep learning is often attributed to the mysterious ability of stochastic gradient descents (SGDs) to avoid sharp local minima in the loss landscape, as sharp minima are believed to lead to poor generalization. To unravel this mystery and potentially further enhance such capability of SGDs, it is imperative to go beyond the traditional local convergence analysis and obtain a comprehensive understanding of SGDs' global dynamics within complex non-convex loss landscapes. In the second part of this talk, I will characterize the global dynamics of SGDs building on the heavy-tailed large deviations and local stability framework developed in the first part. This leads to the heavy-tailed counterparts of the classical Freidlin-Wentzell and Eyring-Kramers theories. Moreover, we reveal a fascinating phenomenon in deep learning: by injecting and then truncating heavy-tailed noises during the training phase, SGD can almost completely avoid sharp minima and hence achieve better generalization performance for the test data.

This talk is based on the joint work with Mihail Bazhba, Jose Blanchet, Bohan Chen, Sewoong Oh, Zhe Su, Xingyu Wang, and Bert Zwart.

Bio

Chang-Han Rhee is an Assistant Professor in Industrial Engineering and Management Sciences at Northwestern University. Before joining Northwestern University, he was a postdoctoral researcher at Centrum Wiskunde & Informatica and Georgia Tech. He received his Ph.D. from Stanford University. His research interests include applied probability, stochastic simulation, experimental design, and the theoretical foundation of machine learning. His research has been recognized with the 2016 INFORMS Simulation Society Outstanding Publication Award, the 2012 Winter Simulation Conference Best Student Paper Award, the 2023 INFORMS George Nicholson Student Paper Competition (2nd place), and the 2013 INFORMS George Nicholson Student Paper Competition (finalist). Since 2022, his research has been supported by the NSF CAREER Award.