Eliminating Sharp Minima from SGD with Truncated Heavy-Tailed Noise

Dr. Chang-Han Rhee
Assistant Professor
Industrial Engineering and Management Sciences
Northwestern University

Abstract: The empirical success of deep learning is often attributed to SGD's mysterious ability to avoid sharp local minima in the loss landscape, as sharp minima are known to lead to poor generalization. Recently, empirical evidence of heavy-tailed gradient noise was reported in many deep learning tasks, and it was argued that SGD can escape sharp local minima under the presence of such heavy-tailed gradient noise, providing a partial explanation to the mystery. This talk analyzes a popular variant of SGD where gradients are truncated above a fixed threshold. We show that it achieves a stronger notion of avoiding sharp minima: it can effectively eliminate sharp local minima entirely from its training trajectory. Further, we rigorously characterize the first exit times from local minima and prove that under some structural conditions, the dynamics of heavy-tailed truncated SGD with small learning rates closely resemble those of a continuous-time Markov chain that never visits any sharp minima. Real data experiments on deep neural networks confirm our theoretical prediction that SGD with truncated heavy-tailed gradient noise finds flatter local minima and achieves better generalization.
This talk is based on the joint work with Xingyu Wang and Sewoong Oh.

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 in the Stochastics Group at Centrum Wiskunde & Informatica and Industrial & Systems Engineering and Biomedical Engineering at Georgia Tech. He received his Ph.D. in Computational and Mathematical Engineering from Stanford University. His research interests include stochastic simulation, applied probability, experimental design, and machine learning. He is a recipient of the 2022 NSF CAREER Award, 2016 INFORMS Simulation Society Outstanding Publication Award,2012 Winter Simulation Conference Best Student Paper Award (MS/OR focused), and a finalist of 2013 INFORMS George Nicholson Student Paper Competition.