I M B M istanbul matematiksel bilimler merkezi
istanbul center for mathematical sciences

Towards Building a Heavy-Tailed Theory of Stochastic Gradient Descent for Deep Neural Networks: Part 2

Umut Simsekli

INRIA - ENS Paris

Abstract

In this talk, I will focus on the ’tail behavior’ in Stochastic Gradient Descent in deep learning. I will first empirically illustrate that heavy tails arise in the gradient noise (i.e., the difference between the stochastic gradient and the true gradient). Accordingly I will propose to model the gradient noise as a heavy-tailed α-stable random vector, and accordingly propose to analyze SGD as a discretization of a stochastic differential equation (SDE) driven by a stable process. As opposed to classical SDEs that are driven by a Brownian motion, SDEs driven by stable processes can incur ‘jumps’, which force the SDE (and its discretization) transition from ’narrow minima’ to ’wider minima’, as proven by existing metastability theory and the extensions that we proved recently. These results open up a different perspective and shed more light on the view that SGD ’prefers’ wide minima. In the second part of the talk, I will focus on the generalization properties of such heavy-tailed SDEs and show that the generalization error can be controlled by the Hausdorff dimension of the trajectories of the SDE, which is closely linked to the tail behavior of the driving process. Our results imply that heavier-tailed processes should achieve better generalization; hence, the tail-index of the process can be used as a notion of ”capacity metric”. Finally, if time permits, I will talk about the ’originating cause’ of such heavy-tailed behavior and present theoretical results which show that heavy-tails can even emerge in very sterile settings such as linear regression with iid Gaussian data.

Date : Wednesday, March 9, 2022

Time: 14:00

Place: Boğaziçi University, South Campus