Abstract:

Noise plays a central role in many machine learning algorithms; one well-studied example is stochastic gradient descent (SGD). Despite its simplicity, i.e., being a noisy first-order optimization method, SGD empirically outperforms gradient descent (GD) and second-order methods. In the first part of the talk, I will present Gaussian Stochastic Gates (STGs), a differentiable non-convex relaxation of the L0 norm highly effective for feature selection. Under a linear sparse regression model, I will show that STGs can recover the informative features successfully. Then, I will show how a randomly aggregated least squares procedure can improve the probability of exact recovery. In the second part, I will revisit the noise model of SGD and present strong evidence that it is well characterized by a symmetric $S\alpha S$ Lévy distribution. This allows us to use Lévy-driven stochastic differential equation (SDE) to analyze different properties of deep nets near local minima. (based and join works with Yutaro Tamada, Yuval Kluger, Sahand Negahban, Stefan Steinerberger, and Barak Battash).