Title: Near Optimal Heteroscedastic Regression with Symbiotic Learning

Speaker: Praneeth Netrapalli, Research Scientist at Google Research India

Area: DS

Date: 05.09.2023, Venue: M21 @ 3.30PM

Abstract:

We consider the classical problem of heteroscedastic linear regression where, given n independent and identically distributed samples (x_i,y_i) drawn from the model $y_i=w^T x_i + \eps_i·(f^T x_i)$ where $x_i ~ N(0,I)$ and $\epsilon_i ~ N(0,1)$, our aim is to estimate the regressor `w` without prior knowledge of the noise parameter `f`. In addition to classical applications of such models in statistics (Jobson and Fuller, 1980), econometrics (Harvey, 1976), time series analysis (Engle, 1982) etc., it is also particularly relevant in machine learning problems where data is collected from multiple sources of varying (but apriori unknown) quality, e.g., in the training of large models (Devlin et al., 2019) on web-scale data. In this work, we develop an algorithm called SymbLearn (short for Symbiotic Learning) which estimates `w` in squared norm up to an error of $\tilde O(||f||^2(1/n+(d/n)^2))$, and prove that this rate is minimax optimal modulo logarithmic factors. This represents a substantial improvement upon the previous best known upper bound of \Otilde(||f||^2 d/n). Our algorithm is essentially an alternating minimization procedure which comprises of two key subroutines: (1) An adaptation of the classical weighted least squares heuristic to estimate `w` (dating back to at least (Davidian and Carroll 1987)), for which our work presents the first non-asymptotic guarantee; (2) a novel non-convex pseudo gradient descent procedure for estimating `f`, which draws inspiration from the phase retrieval literature. As corollaries of our analysis, we obtain fast non-asymptotic rates for two important problems, linear regression with multiplicative noise, and phase retrieval with multiplicative noise, both of which could be of independent interest. Beyond this, the proof of our lower bound, which involves a novel adaptation of Le Cam's two point method for handling infinite mutual information quantities (which prevents a direct application of standard techniques such as Fano's method), could also be of broader interest for establishing lower bounds for other heteroscedastic or heavy tailed statistical problems.

Speaker Profile:

Praneeth Netrapalli is a research scientist at Google Research India, Bengaluru. He is also an adjunct professor at IIT Bombay, TIFR, Mumbai and a faculty associate of ICTS, Bengaluru. Prior to this, he was a researcher at Microsoft Research. He obtained MS and PhD in ECE from UT Austin, and B-Tech in EE from IIT Bombay. He is a co-recipient of IEEE Signal Processing Society Best Paper Award 2019, Indian National Science Academy (INSA) Medal for Young Scientists 2021 and is an associate of Indian Academy of Sciences (IASc) 2019-2022. His current research interest is to make training and inference of large language models more efficient.

Webpage Link: https://praneethnetrapalli.org/