Programme du 50^{e} : Nouveaux développements en probabilités libres et applications
« Atelier sur les probabilités libres : la perspective appliquée »
25 au 29 mars 2019

50th anniversary program : New Developments in Free Probability and Applications
"Workshop on Free Probability : the applied perspective"
March 25-29, 2019

A new theory for sketching in linear regression

Edgar Dobriban *

dobriban@wharton.upenn.edu

Large datasets create opportunities as well as analytic challenges. A recent development is to use random projection or sketching methods for dimension reduction in statistics and machine learning. In this work, we study the statistical performance of sketching algorithms for linear regression. Suppose we randomly project the data matrix and the outcome using a random sketching matrix reducing the sample size, and do linear regression on the resulting data. How much do we lose compared to the original linear regression? The existing theory does not give a precise enough answer, and this has been a bottleneck for using random projections in practice.

In this paper, we introduce a new mathematical approach to the problem, relying on very recent results from asymptotic random matrix theory and free probability theory. This is a perfect fit, as the sketching matrices are random in practice. We allow the dimension and sample sizes to have an arbitrary ratio. We study the most popular sketching methods in a unified framework, including random projection methods (Gaussian and iid projections, uniform orthogonal projections, subsampled randomized Hadamard transforms), as well as sampling methods (including uniform, leverage-based, and greedy sampling). We find precise and simple expressions for the accuracy loss of these methods. These go beyond classical Johnson-Lindenstrauss type results, because they are exact, instead of being bounds up to constants. Our theoretical formulas are surprisingly accurate in extensive simulations and on two empirical datasets.

*Department of Statistics, University of Pennsylvania, 3730 Walnut Street, Suite 400, Philadelphia, PA 19104, USA