Abstract:
We study the problem of identifying correlations in multivariate data, under information constraints: Either on the amount of memory that can be used by the algorithm, or the amount of communication when the data is distributed across several machines. We prove a tight trade-off between the memory/communication complexity and the sample complexity, implying (for example) that to detect pairwise correlations with optimal sample complexity, the number of required memory/communication bits is at least quadratic in the dimension. Our results substantially improve those of Shamir (2014), which studied a similar question in a much more restricted setting. To the best of our knowledge, these are the first provable sample/memory/communication trade-offs for a practical estimation problem, using standard distributions, and in the natural regime where the memory/communication budget is larger than the size of a single data point. To derive our theorems, we prove a new information-theoretic result, which may be relevant for studying other information-constrained learning problems. Joint work with Ohad Shamir