Population Recovery and Partial Identification

Avi Wigderson* Amir Yehudayoff†

Abstract

We study several problems in which an *unknown* distribution over an *unknown* population of vectors needs to be recovered from partial or noisy samples, each of which nearly completely erases or obliterates the original vector. Such problems naturally arise in a variety of contexts in learning, clustering, statistics, computational biology, data mining and database privacy, where loss and error may be introduced by nature, inaccurate measurements, or on purpose. We give fairly efficient algorithms to recover the data under fairly general assumptions.

Underlying our algorithms is a new structure we call a *partial identification (PID) graph* for an arbitrary finite set of vectors over any alphabet. This graph captures the extent to which certain subsets of coordinates in each vector distinguish it from other vectors. PID graphs yield strategies for dimension reductions and re-assembly of statistical information.

The quality of our algorithms (sequential and parallel runtime, as well as numerical stability) critically depends on three parameters of PID graphs: width, depth and cost. The combinatorial heart of this work is showing that *every* set of vectors posses a PID graph in which all three parameters are small (we prove some limitations on their trade-offs as well). We further give an efficient algorithm to find such near-optimal PID graphs for any set of vectors.

Our efficient PID graphs imply general algorithms for these recovery problems, even when loss or noise are just below the information-theoretic limit! In the learning/clustering context this gives a new algorithm for learning mixtures of binomial distributions (with known marginals) whose running time depends only quasi-polynomially on the number of clusters. We discuss implications to privacy and coding as well.