Interactive Fingerprinting Codes and the Hardness of Preventing False Discovery

Thomas Steinke & Jonathan Ullman
Harvard & Northeastern
Motivation: False Discovery
Problem: Data Over-Use

• “Ideally,” data is used once and discarded. In practice, data is reused.

• Data analysis may depend on previous analysis of data.

• e.g. Freedman’s paradox: Use same dataset for variable selection and model fitting.
Model: Statistical Queries

Oracle(S)

Sample S of size n

Interactive Analyst

Unknown Distribution D

Query q_i

Answer a_i

Query q_k

Answer a_k
Key Question [DFHPRR15]

• Q: How many samples \( n \) do I need to answer \( k \) adaptive statistical queries?

• Alternatively: Suppose I have a SQ learning algorithm that makes \( k \) statistical queries. If I simulate this as a PAC learning algorithm (in the obvious way) how many samples \( n \) does it need?
Our Results

• Q: How many samples \( n \) do I need to answer \( k \) adaptive statistical queries?

• For non-adaptive queries: \( n = \Theta(\log k) \) suffices

• **Main Result:** \( n = \Omega(\sqrt{k}) \) *

• Previously: \( n = \tilde{\Omega}(k^{1/3}) \) [HU14]

• Almost Tight: \( n = \tilde{O}(\sqrt{k}) \) [DFHPRR15,BSSU15,NS15]

• *Assuming either dimension \( O(n^2) \) or a computationally bounded oracle, super-logarithmic dimension, and OWFs.
Intuition

• If the analyst knows the sample, she can easily over-fit it.

• Answering lots of queries reveals enough information to identify the sample, which means the analyst may over-fit.

• Key tool: Interactive Fingerprinting Codes. [FT01,T03,LDRSdW13,this]