Lecture 18: Some open problems in robust statistics (and friends)

December 1, 2025

1 Optimal sample complexity for resilient cores

We'll start off with an easy-to-state, self-contained problem which is to complete the picture for robust mean estimation for bounded second moments. The missing piece is this issue that the best known sample complexity for this problem is off from the non-robust rate by a logarithmic factor, which in the grand scheme of things probably isn't the biggest deal, but makes the landscape unfortunately messy.

More formally, the question is the following:

Problem 1.1. *Let D be a distribution over $\mathbb{R}^d$ with mean 0 and covariance $\Sigma \preceq I$, and let $\varepsilon > 0$. Given a set of $n = O(d/\varepsilon)$ samples $X_1, \ldots, X_n$ from D, is it true that with high probability, there exists a set $S \subseteq [n]$ of size $|S| \geq (1-\varepsilon)n$ satisfying that*

$$\frac{1}{|S|} \sum_{i \in S} X_i X_i^\top \preceq C \cdot I \ ,$$

for some universal constant C?

On the homework, you were asked to show that this holds given $O(\frac{d \log d}{\varepsilon})$ samples, but unfortunately, for that approach it feels like the logarithmic factor is unavoidable. Ultimately, this logarithmic factor comes from the application of matrix Chernoff, where such a factor is inevitable.

The downstream application of this for robust mean estimation is that the best known efficient algorithms for robust mean estimation with bounded second moments need to pay this additional logarithmic factor in the sample complexity, when we know that this is not necessary non-robustly. Resolving this would finally allow us to obtain minimax optimal rates for this basic problem, with an efficient algorithm.

2 Efficient private sparse mean estimation

Recall from the lectures on private estimation there is this nice reduction from a robust estimation algorithm to a pure DP estimation algorithm, via the exponential mechanism applied as follows. Suppose we are trying to estimate some parameter $\theta \in \mathbb{R}^d$. Given a dataset X, a robust estimation algorithm $\mathcal{A}$ which, for all corruption thresholds η, achieved error $\|\mathcal{A}(S) - \theta^*\|_2 \leq \alpha(\eta)$, and some parameter α_0, and for some ground truth θ^*, we define the score function

$$s(\theta) = \min_{X'} \{ \|X' - X\|_0 : \|\mathcal{A}(X') - \theta\|_2 \leq \alpha_0 \} \ .$$

The exponential mechanism with this score function then is guaranteed to output a θ so that $\|\theta - \theta^*\|_2 \leq O(\alpha_0)$, so long as

$$n \geq \max_{\eta \in [\eta_0, 1]} \frac{d \cdot \log \frac{2\alpha(\eta)}{\alpha(\eta_0)} + \log 1/\beta + O(\log \eta n)}{\eta \varepsilon} \ .$$

Our main focus in this problem will be on this first term in the sum in the numerator on the RHS. Recall that this term is really a logarithm of a ratio of volumes, that is, it is of the form

$$d \cdot \frac{2\alpha(\eta)}{\alpha(\eta_0)} = \log\left(\frac{V(2\alpha(\eta))}{V(\alpha_0)}\right) ,$$

where $V(r)$ denotes the volume of a d-dimensional ℓ_2 ball with radius r.

The problem is that this ratio fundamentally grows exponentially with d, resulting in a sample complexity which will be fundamentally dimension-dependent. But there are many tasks for which we believe that we can obtain sublinear rates in d, and indeed, we often can, with suboptimal rates, for efficient algorithms. Despite this, the powerful tools of the exponential mechanism seem to fail us in this setting, because of these volume arguments.

A simple but representative example of this is the setting of *sparse mean estimation*. Here, we are given n samples from $\mathcal{N}(\mu, I)$, for some μ which is k-sparse, and the goal is to learn μ to error α, given as few samples as possible. One can easily demonstrate that non-privately, the number of samples we need is $n = O\left(\frac{k \log d}{\alpha^2}\right)$. A major question is if a similar rate can be achieved privately:

Problem 2.1. *Is there an ε-DP algorithm that runs in polynomial time that achieves sample complexity*

$$n = O\left(\frac{k \log d}{\alpha^2} + \frac{k \log d}{\alpha\varepsilon}\right) ,$$

and which solves robust sparse mean estimation to error α, with high probability?

Again, here the big problem is that our most powerful tool (the exponential mechanism) cannot witness the fact that there is sparsity structure here. In particular, there does not appear to be any reasonable score function that can be implemented in polynomial time, which can witness this. There is actually some evidence that this is inevitable, at least in some settings [1]. This relates to the next class of questions:

3 Computational statistical gaps in robust statistics

There is a growing literature in the field of statistical learning exploring this phenomenon where some statistical problems appear to require more samples than necessary information-theoretically, if we insist on algorithms which run in polynomial time. This phenomenon is usually known by the somewhat unwieldy moniker of *computational-statistical tradeoffs* in statistical learning.

For instance, a classic example of a problem where we believe such a phenomena exists is the *planted clique* problem. Here, we are given a graph G, and the goal is to distinguish between the following two cases:

- **Null:** $G \sim G(n, 1/2)$, that is, G is an Erdős-Renyi graph.
- **Planted:** G is an Erdős-Renyi graph, but with an additional clique of size k added.

It is well-known that information-theoretically, the largest clique in an Erdős-Renyi graph of size in has size $k = (1 + o(1)) \cdot 2 \log n$ with overwhelming probability. However, the only known polynomial-time algorithm for this problem requires $k = \Omega(\sqrt{n})$ to succeed with any non-trivial probability, despite quite a lot of work attempting to improve this. This has given rise to the following popular conjecture:

Conjecture 3.1 (Planted clique conjecture)**.** No polynomial time algorithm can solve the planted clique problem, unless $k = \Omega(\sqrt{n})$.

While we are not particularly close to proving such a lower bound, since it is strictly harder than the $\mathsf{P} = \mathsf{NP}$ problem, there is a lot of conditional evidence that this, and related problems, have statistical-computational gaps. Typically, this is in the form of lower bounds against known algorithmic techniques, such as lower bounds against SoS algorithms, statistical query (SQ) algorithms, smooth algorithms, or low-degree algorithms.

Of particular interest to us is the fact that such lower bounds and separations are also conjectured for robust statistics. The peculiar thing is that sometimes, it appears that these lower bounds arise even when the non-robust problem does not itself exhibit such tradeoffs. A classic example of this is the setting of *robust sparse mean estimation*, where it seems like robustly, to learn the k-sparse mean of a Gaussian, one requires $\Omega(k^2 \log d)$ samples, even though, as discussed above, non-robustly, $O(k \log d)$ samples suffice. In fact, these lower bounds form the basis of some of the conditional evidence for the hardness of private sparse mean estimation. An interesting open direction is to provide more evidence, or find more instances, where robustness seems to result in these gaps.

Another, somewhat orthogonal, direction, is to demonstrate hardness for robust mean estimation more directly. Here, unlike many other statistical estimation problems, it seems like there is hope that one can do this directly from standard complexity theoretic assumptions, as opposed to these more hand-wavy lower bounds. In particular, we have the following question:

Problem 3.1. *Let $k \geq 4$. Can one show worst-case hardness for the following problem: given an ε-corrupted set of samples from a distribution D with mean μ so that $\mathbb{E}_{X \sim D}[\langle X - \mu, v \rangle^k] \leq 1$ for all $\|v\|_2 = 1$, output a $\widehat{\mu}$ so that with high probability, $\|\widehat{\mu} - \mu\|_2 \leq o(\varepsilon^{1/2})$.*

Note that here, information-theoretically, the right rate is $O(\varepsilon^{1-1/k})$, but all known efficient algorithms get stuck at this rate, at least for k constant. The fundamental issue here is this issue alluded to before in the lectures on SoS about the difficulty of certifying these injective norms of tensors. This is deeply related to the *small-set expansion* hypothesis:

Conjecture 3.2 (Small-set expansion hypothesis (SSE) [2])**.** For every $\varepsilon > 0$ there exists a $\delta > 0$ so that, given a graph G, it is NP-hard to distinguish between the following two instances:

- **Null:** Every set of vertices S of G of size δn has expansion $\Phi(S) \geq 1 - \varepsilon$, or
- **Planted:** There exists a set of vertices S of size δn has expansion $\Phi(S) \leq \varepsilon$.

Here $\Phi(S)$ is the probability that a random walk initialized in S, and which chooses a uniformly random edge, leaves the set S.

This conjecture is closely related both to the more famous *unique games conjecture*, as well as to problems related to certifying injective norms. In fact, [3] demonstrated that certifying injective norms is SSE-hard (or at least, can be reduced to some very related hypothesis). There has been some partial progress on this: prior work of [4] demonstrated that a strengthening of this conjecture does in fact imply hardness for the problem. However, there is also a bit of a recent wrinkle: work of [5] implies that if k is sufficiently large, (as of writing, it seems to require that k is polynomial in d), then this problem actually is easy.

4 Robust statistics and quantum learning

Come to the lecture to hear my 30 minute spiel about quantum learning :)

References

[1] Kristian Georgiev and Samuel Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in neural information processing systems*, 35:6829–6842, 2022.

[2] Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 755–764, 2010.

[3] Boaz Barak, Fernando GSL Brandao, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 307–326. ACM, 2012.

[4] Samuel B Hopkins and Jerry Li. How hard is robust mean estimation? *arXiv preprint arXiv:1903.07870*, 2019.

[5] Ilias Diakonikolas, Samuel B Hopkins, Ankit Pensia, and Stefan Tiegel. Sos certifiability of subgaussian distributions and its algorithmic applications. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1689–1700, 2025.