Algorithms for Heavy-Tailed Statistics: Regression, Covariance Estimation, and Beyond

Yeshwanth Cherapanamjeri
U.C. Berkeley
USA
yeshwanth@berkeley.edu

Samuel B. Hopkins
U.C. Berkeley
USA
hopkins@berkeley.edu

Tarun Kathuria
U.C. Berkeley
USA
tarunkathuria@berkeley.edu

Prasad Raghavendra
U.C. Berkeley
USA
raghahvendra@berkeley.edu

Nilesh Tripuraneni
U.C. Berkeley
USA
nilesh_tripuraneni@berkeley.edu

ABSTRACT

We study polynomial-time algorithms for linear regression and covariance estimation in the absence of strong (Gaussian) assumptions on the underlying distributions of samples, making assumptions instead about only finitely-many moments. We focus on how many samples are required to perform estimation and regression with high accuracy and exponentially-good success probability in the face of heavy-tailed data.

For covariance estimation, linear regression, and several other problems in high-dimensional statistics, estimators have recently been constructed whose sample complexities and rates of statistical error match what is possible when the underlying distribution is Gaussian, but known algorithms for these estimators require exponential time. We narrow the gap between the Gaussian and heavy-tailed settings for polynomial-time estimators with: (a) a polynomial-time estimator which takes n samples from a d-dimensional random vector X with covariance Σ and produces $\hat{\Sigma}$ such that in spectral norm $\|\hat{\Sigma} - \Sigma\|_2 \leq \tilde{O}(d^{3/4}/\sqrt{n})$ w.p. $1 - 2^{-d}$ where the information-theoretically optimal error bound is $\tilde{O}(\sqrt{d/n})$, while previous approaches to polynomial-time algorithms were stuck at $\tilde{O}(d/\sqrt{n})$ and (b) a polynomial-time algorithm which takes n samples (X_i, Y_i) where $Y_i = \langle u, X_i \rangle + \varepsilon_i$ where both X and ε have a constant number of bounded moments and produces $\hat{u}$ such that the loss $\|u - \hat{u}\|^2 \leq O(d/n)$ w.p. $1 - 2^{-d}$ for any $n \geq d^{3/2}\,\mathrm{poly}\log(d)$. This (information-theoretically optimal) error is achieved by inefficient algorithms for any $n \gg d$, while previous approaches to polynomial-time algorithms suffer loss $\Omega(d^2/n)$ and require $n \gg d^2$.

Our algorithms make crucial use of degree-8 sum-of-squares semidefinite programs. Both apply to any X which has constantly-many *certifiably hypercontractive moments*. We offer preliminary evidence that improving on these rates of error in polynomial time is not possible in the *median of means* framework our algorithms employ. Our work introduces new techniques to high-probability estimation, and suggests numerous new algorithmic questions in the following vein: *when is it computationally feasible to do statistics in high dimensions with Gaussian-style errors when data is far from Gaussian*?

CCS CONCEPTS

• **Theory of computation → Sample complexity and generalization bounds**; *Rounding techniques*; • **Mathematics of computing → Multivariate statistics**.

KEYWORDS

Heavy-Tailed Estimation, Sum-of-squares, Algorithms

ACM Reference Format:
Yeshwanth Cherapanamjeri, Samuel B. Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. 2020. Algorithms for Heavy-Tailed Statistics: Regression, Covariance Estimation, and Beyond. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC '20), June 22–26, 2020, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3357713.3384329

1 INTRODUCTION

Much work in theoretical computer science on algorithms for high-dimensional learning and statistics focuses on the dependence of rates of error (in estimation, regression, PAC learning, etc.) on the number of samples n given to a learning/regression/estimation algorithm and the dimension/number of features d of those samples[1]. *In statistics it is also of fundamental importance to understand the dependence on the* ***level of confidence*** $1 - \delta$ *– predictions and estimates made from samples are most useful if they come with small confidence intervals.* Classical estimators for elementary estimation and regression problems often have error rates $r(n, d, \delta)$ with far-from-optimal dependence on δ unless strong assumptions are made on the underlying distribution of samples. In this work, we study algorithms for high-dimensional statistics without strong (sub-Gaussian) assumptions, focusing on achieving *small errors with high probability in polynomial time.*

[1]For formal theorem statements and detailed proofs of results we refer the reader to our full version in arXiv 1912.11071.

STOC '20, June 22–26, 2020, Chicago, IL, USA

ACM ISBN 978-1-4503-6979-4/20/06…$15.00
https://doi.org/10.1145/3357713.3384329

Consider a prototypical estimation problem: the goal is to take independent samples $X_1, \dots, X_n \sim\sim p_\theta$, where p_θ is a member of a family of d-dimensional probability distributions indexed by parameters θ and find $\hat{\theta}$ such that $\|\theta - \hat{\theta}\| \le r$ with probability $1 - \delta$ for some norm $\|\cdot\|$ and some *rate* $r(n, d, \delta)$. If we make only a weak assumption on p_θ – e.g. that it has a small number of finite moments – then the rates $r(n, d, \delta)$ achieved by classical approaches are typically exponentially-far from optimal with respect to δ (i.e. $r(n, d, \delta)$ scales like $1/\text{poly}(\delta)$ rather than $\log(1/\delta)$).

Since at least the 1980s it has been known that in low-dimensional settings (e.g. $d = 1$) there are estimators for basic problems like estimating the mean which achieve rates $r(n, \delta)$ whose dependence on δ under such weak assumptions is comparable to that of classical estimators (the empirical mean) under (sub)-Gaussian assumptions (up to constants). For instance, the *median of means* estimator of the mean achieves the same $r(n, \delta)$ as the empirical mean does in the Gaussian setting but assuming only that p_θ has finite variance [AMS99a, JVV86, NY83b]. This immediately proved useful in streaming algorithms [AMS99a].

Achieving similar guarantees for large dimensions d is much more challenging, even without asking for computationally-efficient algorithms. A series of exciting developments in the last decade in statistics, however, constructs estimators with $r(n, d, \delta)$ matching the rates achievable in the Gaussian case by classical approaches but with much weaker assumptions. Such estimators are now known for high dimensional mean estimation, covariance estimation, (sparse) linear regression, and more [LM19b]. Unlike their one-dimensional counterparts and classical approaches, however, *naive algorithms to compute this new generation of optimal estimators take time exponential in n, d, or both.* This suggests a key question applying to a wide range of estimation, regression, and learning problems:

> *Are there efficiently computable estimators achieving optimal $r(n, d, \delta)$ under weak assumptions (like finitely-many bounded moments) on underlying data?*

Recent work in algorithms shows that such optimal and computationally efficient estimators do exist for the problem of estimating the mean of a random vector X under only the assumption that X has finite covariance [Hop18a, CFB19]. The resulting algorithms, however, are heavily tailored to estimating the mean in ℓ_2; although they introduce useful techniques, it is unclear whether they suggest any broader answers to the above.

In this work we tackle covariance estimation and linear regression with these goals in mind. We contribute new algorithms for both problems whose error rates $r(n, d, \delta)$ improve by $\text{poly}(d, \log(1/\delta))$ factors on the previous best polynomial-time algorithms when the underlying data is drawn from a distribution with only finitely-many bounded moments. Unlike the situation in mean estimation, however, our estimators do not achieve information-theoretically optimal error rates. We offer evidence (by constructing certain moment-matching distributions) that no efficient algorithm using the median-of-means approach we use here can significantly improve on rates achieved by our algorithms. This suggests the possibility that the computational landscape for covariance estimation and regression is more complicated than for mean estimation: in particular, it could be that these problems suffer from a novel kind of tradeoff between computational efficiency and error rate *in the small δ regime.* (By contrast in the regime $\delta = \Omega(1)$ classical estimators typically have $r(n, d, \delta)$ which is information-theoretically optimal with respect to n, d and are also efficiently computable.) Whether there is indeed such a tradeoff is a fascinating open question.

Why Weak Assumptions? We study polynomial-time algorithms for high-dimensional statistics under *weak assumptions* on underlying data. Both linear regression and covariance estimation boast well-studied and computationally-efficient algorithms which achieve statistically optimal rates $r(n, d, \delta)$ with respect to both n and δ under (sub)-Gaussian assumptions on X (and ε): ordinary least squares regression and the empirical covariance, respectively. These estimators are among the oldest in statistics: Gauss and Legendre both studied the least-squares estimator for linear regression around 1800 [Wik19a] and study of the empirical covariance dates at least to Pearson's invention of principal component analysis [Pea01].

However, data cannot assumed to be Gaussian in every situation. In this paper we only assume boundedness conditions on a small number of moments of a random vector X (generally 8th moments). Under such assumptions, the error rates of the empirical covariance and ordinary least squares grow polynomially in $1/\delta$, while optimal error rates are logarithmic in $1/\delta$. Beyond allowing us to address basic questions about which error rates are achievable in polynomial time, working under weak assumptions makes our algorithms potentially useful in a variety of settings where classical estimators break down.

First, our algorithms are useful in statistical settings involving heavy-tailed data – data drawn from distributions with only a finite number of bounded moments. Large networks, for instance, are well known to generate heavy-tailed data, often following a power law distribution. Other common heavy-tailed distributions in statistics include the *Student's t* distribution, and the Log-Normal distribution – the latter describes a number of real-world phenomena, such as the distribution of English sentence lengths, the distribution of elements in the Earth's crust, the distribution of species' abundances, and more [Wik19b]. Even when data are not known to follow a particular heavy-tailed distribution, the conservative statistician may wish to avoid a Gaussian assumption if also lacking good reason to believe that the underlying population is Gaussian-distributed.

Second, it is often convenient to use algorithmic primitives for basic tasks like covariance estimation and regression as parts of more sophisticated algorithms. Algorithms for the complicated high-dimensional statistics problems often studied in theoretical machine learning can have many moving parts. In such situations, the samples $X_1, \dots, X_n$ may themselves be the output of a complex random process or another "upstream" algorithm. This can make it it difficult or impossible to guarantee that $X_1, \dots, X_n$ satisfy sub-Gaussian concentration properties, but it can be much easier to establish that the outputs of such upstream algorithms satisfy the kind of weak finite-moment bounds required by our algorithms. Indeed, one of the first uses of the *median of means* technique we employ here (for estimation of frequency moments in a streaming setting) was for exactly this purpose [AMS99b].

1.1 Results

"Nice" distributions. Since the main goal of our work is to achieve Gaussian-style error bounds while avoiding Gaussian assumptions in high-dimensional parameter estimation, before we lay out our results, we must describe the class of distributions to which they apply. Obtaining Gaussian-style error rates does require some assumptions on the underlying random variables, for information-theoretic reasons – typically the existence of 2nd moments is a minimal requirement [C+12]. (For covariance estimation this becomes 4th moments of a random vector X, which are the 2nd moments of the random matrix $XX^\top$.)

In this paper we make an assumption called *certifiable hypercontractivity:* we assume that *as a polynomial in variables* $u = (u_1, \dots, u_d)$, there exists a universal constant L such that,

$$L^2 \cdot (\mathbb{E}\langle X, u\rangle^2)^4 - \mathbb{E}\langle X, u\rangle^8 .$$

is a sum of squares of polynomials in u.[2] This in particular implies the more standard 8th moment bound $\mathbb{E}\langle X, u\rangle^2 \le O(\mathbb{E}\langle X, u\rangle^8)^{1/4}$. We often call $(2, 8)$ certifiably-hypercontractive distributions *nice*. We emphasize that niceness is an "infinite-sample" assumption: it concerns population moments $\mathbb{E}\, X^{\otimes 8}$.

Certifiable hypercontractivity holds for numerous interesting heavy-tailed distributions for which previous polynomial-time algorithms could not have achieved Gaussian-style error guarantees. For instance, any product of univariate distributions with bounded 8-th moments, and any linear transformation thereof (in particular for multivariate t-distributions) is certifiably hypercontractive. In fact, the certifiable hypercontractivity assumption has been shown to hold for any distribution whose 8-th moments match those of some strongly log-concave distribution [KSS18] (even if, say, 9th moments do not exist). The certifiable hypercontractivity assumption also underlies recent results in on polynomial-time high-dimensional clustering of mixture models and several robust parameter estimation problems [KSS18, HL18, KKM18].

1.1.1 Covariance Estimation. Covariance estimation is the following simple problem. Given samples $X_1, \dots, X_n$ from a d-dimensional random vector with covariance Σ, find $\hat{\Sigma}$ with the smallest possible spectral norm error $\|\hat{\Sigma} - \Sigma\|_2$. For simplicity, let us focus for now on the setting that $\operatorname{Tr}\Sigma \le O(d)$ and $\|\Sigma\|_2 \le O(1)$ and $\delta = 2^{-d}$. (Our main theorem for covariance estimation handles the case of general Σ and δ.) We also assume throughout that $\mathbb{E}\, X = 0$; otherwise one may replace X with $(X - X')/\sqrt{2}$ for pairs of independent samples X, X' without affecting the covariance and losing only a factor of 2 in the sample complexity.

Consider the Gaussian setting $X \sim \mathcal{N}(0, \Sigma)$. In this case, classical results offer the following type of concentration bound for the empirical covariance $\overline{\Sigma} = \frac{1}{n}\sum_{i \le n} X_i X_i^\top$ of n independent samples: for a universal constant C,

$$\mathbb{P}\left(\left\|\overline{\Sigma} - \Sigma\right\|_2 \ge C\left(\sqrt{\frac{d}{n}} + t\right)\right) \le \exp(-t^2 n) . \tag{1.1}$$

(This bound becomes meaningful only when $n \ge d$.) Note that by Eq. (1.1), $\left\|\overline{\Sigma} - \Sigma\right\|_2 \le O(\sqrt{d/n})$ with probability $1 - 2^{-d}$.

[2] Our algorithms also work if instead the inequality $\mathbb{E}\langle X, u\rangle^8 \le (\mathbb{E}\langle X, u\rangle^2)^4$ has an SoS proof of higher degree, at a commensurate cost in running time to allow for higher-degree SoS relaxations.

Recent work by Mendelson and Zhivotovskiy [MZ18], building on earlier works by Lugosi and Mendelson [LM18a] shows that there is an estimator $\hat{\Sigma}$ for the covariance Σ which matches this error guarantee under only the assumption that X has hypercontractive 4-th moments. (In all the following informal theorem statements we assume $\operatorname{Tr}\Sigma \le O(d)$, $\|\Sigma\|_2 \le O(1)$.)

Theorem 1.1 ([MZ18]). *There is an estimator $\hat{\Sigma} = \hat{\Sigma}(X_1, \dots, X_n)$ which given n independent samples from a random variable X with covariance Σ and which is $(2, 4)$-hypercontractive has the guarantee*

$$\left\|\hat{\Sigma} - \Sigma\right\|_2 \le O\left(\sqrt{\frac{d \log d}{n}}\right) \text{ with probability at least } 1 - 2^{-d} .$$

Up to logarithmic factors, this rate of error is information-theoretically optimal, but no algorithm is known which achieves this guarantee in polynomial time. Prior to this work, the strongest result known for polynomial-time algorithms was weaker by a $\operatorname{poly}(d)$ factor:

Theorem 1.2 ([MW18]). *Under the same hypotheses as Theorem 1.1 there is a polynomial-time algorithm which finds $\hat{\Sigma}$ such that $\|\hat{\Sigma} - \Sigma\|_2 \le O(d/\sqrt{n})$ with probability at least $1 - 2^{-d}$.*

Our main result for covariance estimation in the setting $\operatorname{Tr}\Sigma \approx d, \|\Sigma\| \approx 1$ is the following.

Theorem 1.3 (Main Theorem on Covariance Estimation, Informal). *There is an algorithm with running time $\operatorname{poly}(n, d)$ which when given n i.i.d. samples $X_1, \dots, X_n$ from a nice random vector X in d dimensions returns an estimate $\hat{\Sigma}$ of the covariance Σ of X such that*

$$\left\|\hat{\Sigma} - \Sigma\right\|_2 \le \tilde{O}\left(\frac{d^{3/4}}{\sqrt{n}}\right) \text{ with probability at least } 1 - 2^{-d} .$$

Here $\tilde{O}(\cdot)$ hides logarithmic factors in the dimension d.

The general statement of our main theorem obtains an error rate which avoids explicit dependence on the ambient dimension d (except for logarithmic factors); instead, it depends only on the "effective" rank $\operatorname{sr}(\Sigma) = \frac{\operatorname{Tr}(\Sigma)}{\|\Sigma\|_2} \le d$ and the operator norm $\|\Sigma\|_2$. Thus if X lies in or near a low-dimensional subspace, our algorithm exploits this additional structure to estimate Σ with fewer samples.

Finally, we note that our algorithm assumes access to a small number of additional parameters: bounds on $\operatorname{Tr}\Sigma$, $\|\Sigma\|_2$, and (as with all the algorithms described in this paper beyond empirical averages) in the case of general confidence levels $1 - \delta$ it depends on the value of δ. The latter dependence may be intrinsic: it is not information-theoretically possible to obtain Gaussian-style error rates in the heavy-tailed setting with estimators which do not depend on δ with minimal moment assumptions [C+12]. We expect that techniques similar to those of [MZ18] can avoid the dependence on $\operatorname{Tr}\Sigma$, $\|\Sigma\|_2$ by estimating them from samples.

The improvement from $d/\sqrt{n}$ to $d^{3/4}/\sqrt{n}$ moves the algorithmic state of the art for covariance estimation closer to information-theoretic optimality. Of course the possibility of an information-theoretically optimal covariance estimation algorithm is tantalizing, but just as interesting from a complexity viewpoint is the possibility that $d^{3/4}/\sqrt{n}$ cannot be improved upon in polynomial time. In Section 1.1.4 we discuss evidence in this direction.

1.1.2 Linear Regression. We study the following classical linear regression problem. Let $f^* : \mathbb{R}^d \to \mathbb{R}$ be a linear function – that is $f^*(x) = \langle f^*, x\rangle$ for some vector $f^* \in \mathbb{R}^d$. Let X be a d-dimensional mean-zero random vector, and let ε be an $\mathbb{R}$-valued random variable with $\mathbb{E}\,\varepsilon = 0$. To avoid a preponderance of parameters, in this paper we focus on the case that $\mathbb{E}\,XX^\top = \mathrm{Id}$ and $\mathbb{E}\,\varepsilon^2 = 1$.[3]

The goal is to take n independent samples of the form (X_i, Y_i), where $Y_i = f^*(X_i) + \varepsilon_i$, and find a linear function $\hat{f}$ such that $\|f^* - \hat{f}\|$ is as small as possible. Here the norm $\|f^* - \hat{f}\|$ is the 2-norm induced by X; that is, $(\mathbb{E}(f^*(X) - \hat{f}(X))^2)^{1/2}$. However, since we assume $\mathbb{E}\,XX^\top = \mathrm{Id}$, this is identical to the Euclidean norm of $f^* - \hat{f}$ considered as a vector of coefficients.

In most respects the situation for linear regression is similar to that for covariance estimation. The classical algorithm is empirical risk minimization, also known in this setting as ordinary least squares regression (OLS). The algorithm is simple: given $(X_1, Y_1), \ldots, (X_n, Y_n)$, output $\hat{f}$ which minimizes the empirical loss $\mathbb{E}_{i\sim[n]}(f(X_i) - Y_i)^2$. This minimization problem is convex, so $\hat{f}$ can be obtained in polynomial time; it also admits a closed-form linear-algebraic solution.

Analogously to the empirical covariance in the previous section, when X and ε are Gaussian, OLS achieves small error with high probability. Concretely, one has the following:[4]

$$\|\hat{f}_{\mathrm{OLS}} - f^*\|^2 \le O\left(\frac{d}{n}\right) \text{ with probability } 1 - 2^{-d} \text{ so long as } n \gg d\,.$$

We focus for now on the setting of regression with confidence $1 - 2^{-d}$: this regime provides a useful litmus test because it is the highest probability for which the $O(d/n)$ guarantee holds for OLS. When X or ε has only a finite number of bounded moments, the error bound on $\|\hat{f}_{\mathrm{OLS}} - f^*\|$ degrades badly, becoming $\frac{\exp(O(d))}{n}$ for confidence $1 - 2^{-d}$.

Recent work by Lugosi and Mendelson [LM16] shows that a guarantee matching that of OLS in the Gaussian setting is possible without Gaussian assumptions. Concretely we have the following:

Theorem 1.4 ([LM16], Informal). *There exists an (exponential-time) estimator $\hat{f}$ which given n independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y = f^*(X) + \varepsilon$, $\mathbb{E}\,XX^\top = \mathrm{Id}$, X is $(2,4)$-hypercontractive, and $\mathbb{E}\,\varepsilon^2 = 1$, has*[5]

$$\left\|\hat{f} - f\right\|^2 \le O\left(\frac{d}{n}\right) \text{ with probability } 1 - 2^{-d} \text{ so long as } n \gg d\,.$$

Once again, the state of the art for polynomial-time algorithms is somewhat worse (though still far better than OLS). Until this paper, the polynomial-time algorithm with smallest error guarantees in the $1 - 2^{-d}$ probability regime were achieved by an algorithm of [HS16b].

[3] It is trivial to show that our results also work if $\mathbb{E}\,\varepsilon^2 = \sigma^2$, with appropriate dependence of the error rates on σ. We also believe that our techniques will be useful in designing algorithms which achieve small error $\mathbb{E}(\hat{f}(X) - f^*(X))^2$ when $\mathbb{E}\,XX^\top = \Sigma$ for general Σ, but we defer this challenge to future work. If X is not mean zero then it can be replaced with $X - X'$ for pairs of samples X, X', so this assumption is without loss of generality.

[4] It is traditional here to state bounds on $\|\hat{f} - f\|^2$ rather than $\|\hat{f} - f\|$; note that the bound $O(d/n)$ represents the so-called *fast rate* for regression – in this paper we are exclusively concerned with fast rates, rather than the *slow rate* $O(\sqrt{d/n})$.)

[5] The results of [LM16] apply to a wide variety of convex function classes rather than just linear regression; we state here the special case for linear regression.

Theorem 1.5 ([HS16b], Informal). *There is a polynomial-time algorithm which computes an estimator $\hat{f}$ which given n i.i.d. samples (X_i, Y_i) where X is $(2, 4+\delta)$-hypercontractive for some $\delta > 0$ and $Y = f^*(X) + \varepsilon$ for some linear function f^* for a random variable ε with $\mathbb{E}\,\varepsilon = 0$ and $\mathbb{E}\,\varepsilon^2 = 1$ achieves*

$$\left\|\hat{f} - f\right\|^2 \le O\left(\frac{d^2}{n}\right) \text{ with probability } 1 - 2^{-d} \text{ so long as } n \gg d^2\,.$$

Note that the error guarantees of Theorem 1.5 are weaker than what is information-theoretically possible (Theorem 1.4) in two key ways: first of all, the error scales with d^2 rather than with d, and second, the error rate does not kick in until $n \gg d^2$. Our main theorem on regression completely fixes the first problem and partially fixes on the second (but does not reach information-theoretic optimality), for nice X.

Theorem 1.6 (Main Theorem on Linear Regression, Informal). *There is an algorithm with running time $\mathrm{poly}(n, d)$ with the following guarantees. Suppose X is nice, ε is a univariate random variable with $\mathbb{E}\,\varepsilon^2 = 1$ and $\mathbb{E}\,\varepsilon = 0$, and f^* is a linear function. Given n i.i.d. samples (X_i, Y_i) of the form $Y_i = f^*(X_i) + \varepsilon_i$, the algorithm finds a linear function $\hat{f}$ such that*

$$\left\|\hat{f} - f^*\right\|^2 \le O\left(\frac{d}{n}\right) \text{ with probability } 1 - 2^{-d}$$
$$\text{so long as } n \gg d^{3/2} \cdot (\log d)^{O(1)}\,.$$

Our main result (and all the prior work) gracefully tolerates confidence levels other than $1 - 2^{-d}$.

1.1.3 Faster Algorithms for Mean Estimation in General Norms. Our final algorithmic result concerns the problem of estimating the mean of a random vector X on $\mathbb{R}^d$ with respect to an arbitrary norm $\|\cdot\|$. Our starting point is the following theorem of Lugosi and Mendelson which constructs an estimator of the mean with respect to any norm $\|\cdot\|$ on $\mathbb{R}^d$. In such a general setting the question of information-theoretic optimality is somewhat murky. Nonetheless, for many natural norms (ℓ_2 and spectral norm, for instance) one may see that the guarantees of their estimator match those of the empirical mean in the Gaussian setting. We refer the reader to [LM18a] for further interpretation of the guarantees of the following theorem.

Theorem 1.7 ([LM18a], Informal, Id-Covariance case). *For every $n, d \in \mathbb{N}$ and $\delta > 2^{-n}$ and norm $\|\cdot\|$ on $\mathbb{R}^d$ there is an estimator with the following guarantee. Given n i.i.d. samples $X_1, \ldots, X_n$ of a random vector X with mean μ and covariance Id, it finds $\hat{\mu}$ such that*

$$\|\mu - \hat{\mu}\| \le \frac{1}{\sqrt{n}} \cdot O\left(\mathbb{E}\left\|\sum_{i\le n}\sigma_i X_i\right\| + R\sqrt{\log(1/\delta)}\right)$$
with probability at least $1 - \delta$

where $\sigma_1, \ldots, \sigma_n \sim \{\pm 1\}$ are independent signs and $R = \sup_{\|x\|_*=1}\|x\|_2$ is the norm-equivalence constant between the dual norm $\|\cdot\|_*$ and ℓ_2. Note that the first term is essentially the expected

error achieved by the empirical mean for the norm $\|\cdot\|$, and in particular is independent of δ, while the second term determines the decay of the bound as δ becomes small.[6]

The naive algorithm to compute the estimator $\hat{\mu}$ from Theorem 1.7 requires brute-force search for a point in a non-convex set in d dimensions, taking $\exp(\Omega(d))$ time. We slightly modify the estimator from Theorem 1.7 and show that subject to a mild computational assumption on the norm $\|\cdot\|$ it can be computed by an algorithm whose running time is exponential only in $\log(1/\delta)$ rather than in d.

Theorem 1.8 (Informal, Id-Covariance Case). *With the same setting and guarantees as Theorem 1.7, under the additional assumption that there is a polynomial-time separation oracle for the dual ball of $\|\cdot\|$, there is an algorithm to compute $\hat{\mu}$ in time $\mathrm{poly}(n, d, 1/\delta)$.*

1.1.4 Roadblock to Improved Error Rates: Single-Spike Block Mixtures. Our main results on covariance estimation and linear regression (Theorems 1.3 and 1.6) push the state of the art in terms of error rates achievable for heavy-tailed statistics in polynomial time, but they do not achieve information-theoretic optimality. Our covariance estimation algorithm in the setting of $\mathrm{Tr}\,\Sigma \le O(d)$, $\|\Sigma\|_2 \le O(1)$ achieves error $\|\hat{\Sigma} - \Sigma\|_2 \le \tilde{O}(d^{3/4}/\sqrt{n})$, while in exponential time it is possible to achieve $\tilde{O}(\sqrt{d/n})$. (Similarly, our linear regression algorithm requires $n \gg d^{3/2}$ rather than $n \gg d$.)

It is a fascinating open problem to understand whether these gaps can be closed. We offer here some evidence that this is unlikely to be possible with techniques in the present paper. We focus on covariance estimation – the relation to linear regression is more subtle. The key subroutine in our covariance estimation algorithm is an algorithm for the following problem:

Problem 1.9 (Find High-Variance Direction). *Given $\Sigma_1, \dots, \Sigma_d \in \mathbb{R}^{d\times d}$, with $\Sigma_i \succeq 0$, find a unit vector $x \in \mathbb{R}^d$ such that $\langle x, \Sigma_i x\rangle \ge r$ for at least $d/4$ matrices Σ_i, or certify that none exists.*

In fact, Problem 1.9 must be solved when $\Sigma_1, \dots, \Sigma_d$ are empirical covariance matrices by any algorithm performing covariance estimation using the *median-of-means* framework, which is the dominant approach in constructing high-dimensional estimators with optimal $r(n, d, \delta)$ (even ignoring running time considerations). It will have to wait until Section 1.2 to see in more detail why an algorithm solving Problem 1.9 is useful for covariance estimation. For now, let us note that our subroutine solves Problem 1.9 when Σ_i is the empirical covariance of n/d samples from the heavy-tailed distribution whose covariance we are estimating, and the Σ_i's are all independent. Problem 1.9 gets easier as r gets larger, but it turns out that the value of r for which we can solve it translates directly to the error rate of our covariance estimation algorithm. *Summarizing: in the case of estimating the covariance Σ of a random variable X with $\mathrm{Tr}\,\Sigma \approx d$, $\|\Sigma\|_2 \approx 1$, our key subroutine solves Problem 1.9 with Σ_i being the empirical covariance of n/d of the samples $X_1, \dots, X_n$ and $r \le \tilde{O}(d^{3/4}/\sqrt{n})$.*

Improving the error rates of our algorithm (or any other median-of-means-based algorithm) would thus seem to require solving Problem 1.9 with smaller r. To investigate whether this may be possible in polynomial time, we consider an easier variant, which we call the *single-spike block mixtures* problem. It is easier in two respects: it is a decision problem rather than a search problem, and the underlying random variable X is distributed in a known, Gaussian fashion. (Note that it appears no longer relevant that we were initially interested in heavy-tailed random vectors – we believe computational hardness for Problem 1.9 appears even when Σ_i's are empirical covariances formed from Gaussian samples.)

Definition 1.10 (Single-Spike Block Mixtures). Let $d, m \in \mathbb{N}$ and $1 > \lambda > 0$. In the *single-spike block mixtures testing problem* the goal is to distinguish, given vectors $y_1, \dots, y_{md} \in \mathbb{R}^d$, between the following two cases:

NULL: $y_1, \dots, y_{md} \sim \mathcal{N}(0, \mathrm{Id})$ i.i.d.

PLANTED: First $x \sim \{\pm 1/\sqrt{d}\}^d$ and $s_1, \dots, s_d \sim \{\pm 1\}$. Then, $y_1, \dots, y_m \sim \mathcal{N}(0, \mathrm{Id} + s_1\lambda xx^\top)$ and $y_{m+1}, \dots, y_{2m} \sim \mathcal{N}(0, +s_2\lambda xx^\top)$, and so forth. That is, each *block* of vectors $y_{im}, \dots, y_{(i+1)m-1}$ has either slightly larger variance in the x direction (if $s_i = 1$) or slightly lesser variance (if $s_i = -1$) than they would in the null case.

It turns out that so long as $\lambda \gg 1/\sqrt{m} = \sqrt{d/n}$ (where $n = md$) it is possible to distinguish NULL from PLANTED in exponential time. (This is closely related to the fact that heavy-tailed mean estimation can be solved with error rate $\tilde{O}(\sqrt{d/n})$.) But what about polynomial time? A consequence of our main subroutine is the following theorem:

Theorem 1.11 (Informal). *If $\lambda \ge (d^{3/4}/\sqrt{n})\,\mathrm{poly}\log(d, m)$ then there is a polynomial-time algorithm which distinguishes NULL from PLANTED with high probability.*

We make the following conjecture regarding optimality of this algorithm.

Conjecture 1.12. *If $\lambda \le d^{3/4-\Omega(1)}/\sqrt{n}$ then no polynomial time algorithm solves the single-spike block mixture problem.*

In support of Conjecture 1.12, we prove a lower bound against a certain class of restricted algorithms, called *low degree tests*. A degree-D test is a function $f : \mathbb{R}^{d\times md} \to \mathbb{R}$ such that as a polynomial $\deg f \le D$ and $\mathbb{E}_{Y=y_1,\dots,y_{md}\sim\text{NULL}} f(Y) = 0$. We say the test is *successful* if $\mathbb{E}_{Y\sim\text{PLANTED}} f(Y)/(\mathbb{E}_{Y\sim\text{NULL}} f(Y)^2)^{1/2} \to \infty$ as $d, m \to \infty$.

While such low degree tests (for D relatively small – say at most $(md)^{o(1)}$) would seem to be a quite restrictive model compared to the class of all polynomial time algorithms, it turns out that the existence of a successful low degree test solving a hypothesis testing problem is a remarkably accurate predictor for the existence of any polynomial time algorithm. For instance, successful low degree tests (of logarithmic degree) appear exactly at the predicted *computational thresholds* for the planted clique problem (clique size $\Omega(\sqrt{n})$), the random 3-SAT problem ((number of variables)$^{3/2}$) clauses), the k-community stochastic block model (the *Kesten-Stigum threshold*), the sparse PCA problem (the k^2 sample threshold) and beyond. Lower bounds on low degree tests are technically distinct from but conceptually similar to statistical query lower bounds. They are also closely related to the *pseudocalibration* technique for proving

[6] In [LM18a] this theorem is stated with an extra term in the error guarantee (which is typically dominated by the first term); we provide a simplified proof which also shows that the additional term is unnecessary.

lower bounds against SoS algorithms. For further discussion, see [Hop18b, KWB19].

We rule out the existence of successful low degree tests for $D = (md)^{o(1)}$ when $\lambda \le d^{3/4-\Omega(1)}/\sqrt{n}$. Obtaining an impossibility result for such large D is relatively strong: in this low-degree test model the typical proxy for polynomial time is D of degree logarithmic in the input size (in this case md^2).

Theorem 1.13 (Informal). *If $\lambda \le d^{3/4-\Omega(1)}/\sqrt{n}$ then there is no successful degree $(md)^{o(1)}$ test for the single-spike block mixtures problem.*

1.2 Techniques

For purposes of this technical overview, we focus on covariance estimation. Our algorithm for linear regression employs broadly similar ideas.

The Median of Means Framework. Let us first explain the basic median-of-means trick in one dimension. Consider the problem of estimating the mean $\mu \in \mathbb{R}$ of a one-dimensional random variable X from independent samples, and suppose $\mathbb{E}(X-\mu)^2 \le 1$, but make no further assumptions on X. In this setting, the empirical mean $\overline{\mu} = \sum_{i=1}^n X_i$ of n independent samples has $\mathbb{P}(|\overline{\mu} - \mu| > t) \le 1/t^2 n$ by Chebyshev's inequality, and no tighter bound is possible. By contrast, if X were Gaussian, we would have the exponentially-better bound $\mathbb{P}(|\overline{\mu} - \mu| > t) \le \exp(-t^2 n/2)$.

The simplest median-of-means trick offers a family of estimators $\hat{\mu}_\delta$ for each $\delta \ge 2^{-0.01n}$ such that $\mathbb{P}(|\hat{\mu}_\delta - \mu| > 100\sqrt{\log(1/\delta)/n}) \le \delta$. First we place $X_1, \dots, X_n$ into $\Theta(\log(1/\delta))$ equal-size buckets. In each bucket $i \le \Theta(\log(1/\delta))$ we let Z_i be the average of the samples in bucket i. Then we let $\overline{\mu}_\delta$ be the median of $Z_1, \dots, Z_{\Theta(\log(1/\delta))}$.

The analysis is a straightforward use of Chebyshev's inequality to show that each Z_i has $|Z_i - \mu| \le O(\sqrt{\log(1/\delta)/n})$ with probability at least 0.9, followed by a binomial tail bound ensuring that with probability at least $1-\delta$ at least a 0.7 fraction of the Z_i's satisfy this inequality. Then the key step: if more than half of $Z_1, \dots, Z_k$ have distance at most r to μ, then so does their median.

Medians in High Dimensions. Extending this idea to high dimensional settings requires surmounting several hurdles. The first one is to design an appropriate high-dimensional notion of median. In the last few years, however, the techniques to do this have become relatively well understood in statistics [LM19b]. For example, the key notion in recent heavy-tailed estimators of the mean of a random vector in d dimensions with respect to Euclidean distance is the following: for a set of points $Z_1, \dots, Z_k \in \mathbb{R}^d$ and $r > 0$, $x \in \mathbb{R}^d$ is an r-median if for every unit direction u we have $|\langle Z_i, u\rangle - \langle x, u\rangle| \le r$ for at least a 0.51-fraction of $Z_1, \dots, Z_k$. It turns out that using the median of means trick with this notion of median leads to an information-theoretically optimal estimator of the mean in d dimensions assuming only that the underlying random vector has finite covariance.

For covariance estimation the appropriate notion of median was first defined in [LM18a] and fully analyzed in [MZ18]. We will call M an r-median for matrices $Z_1, \dots, Z_k$ if for all unit $x \in \mathbb{R}^d$ it holds that $|\langle Z_i, xx^\top\rangle - \langle M, xx^\top\rangle| \le r$ for at least a 0.51-fraction of $Z_1, \dots, Z_k$. Then (ignoring some technical details regarding truncation of large samples) one may design a nearly information-theoretically optimal covariance estimator for random vectors X with bounded 4th moments as follows. Given samples $X_1, \dots, X_n$, as before, place them in $\approx \log(1/\delta)$ buckets. Let Σ_i be the empirical covariance in bucket i, and output an r-median of $\Sigma_1, \dots, \Sigma_{\Theta(\log(1/\delta))}$ for the least r for which such an r-median exists.

How to Compute a Median in High Dimensions. The next hurdle is computational: naive algorithms to compute the medians described above would seem to require exponential time in n or d. Hopkins [Hop18a] uses the sum of squares method to compute the relevant median for mean estimation in ℓ_2. Our main technical contribution for covariance estimation is an algorithm to compute the relevant median *for values of r somewhat larger (hence making finding the median easier) than information-theoretically optimal (but exponential time) algorithms would do.* We stress that our algorithm only outputs a valid r-median when the $X_1, \dots, X_n$ are sampled i.i.d. from a nice distribution.

The key difficulty in computing a median is knowing when we have found one. We first aim to solve a simpler *certification* problem. Suppose given $\Sigma_1, \dots, \Sigma_k$ which are the empirical covariances of independent bucketed copies $X_1, \dots, X_n$ of a random vector X with covariance Σ, *and suppose also given* Σ. How can we *certify*, for as small a value of r as possible, that Σ is an r-median of $\Sigma_1, \dots, \Sigma_k$? That is, we aim to find a certificate that for all unit directions u we have $|\langle \Sigma_i, uu^\top\rangle - \langle \Sigma, uu^\top\rangle| \le r$ for at least a 0.51-fraction of $\Sigma_1, \dots, \Sigma_k$. To leverage the power of the median-of-means trick to obtain estimators whose error is small with high probability, we need to successfully find such a certificate with high probability, $1 - 2^{-k}$. (This need for a high-probability guarantee will play the same role in the algorithmic and high-dimensional context as the simple binomial concentration bound does in the one-dimensional median-of-means estimator.)

To certify that Σ is an r-median for $\Sigma_1, \dots, \Sigma_k$ we start by setting up an optimization problem in variables $b_1, \dots, b_k \in \{0,1\}^k$ and $u \in \mathbb{R}^d$ with $\|u\|^2 = 1$.

$$\max \sum_{i \le k} b_i \text{ s.t. } b_i \langle \Sigma_i - \Sigma, uu^\top\rangle \ge b_i r, \|u\|^2 = 1, b_i^2 = b_i\,. \qquad (1.2)$$

Notice that a feasible solution of value $0.52k$ to the above problem corresponds to a subset of $0.52k$ of $\Sigma_1, \dots, \Sigma_k$ and a unit direction u such that for all Σ_i in the subset, $|\langle \Sigma_i, uu^\top\rangle - \langle \Sigma, uu^\top\rangle| \ge r$. *Ruling out* such solutions (i.e. placing an upper bound on the value of the optimization problem) would thus certify that Σ is an r-median (ignoring some small technical issues about the sign of $\langle \Sigma_i - \Sigma, uu^\top\rangle$).

We will pass to an efficiently-computable convex relaxation of the optimization problem above. In particular, we use the degree-8 Sum of Squares (SoS) semidefinite programming relaxation of Eq. (1.2). Sum of Squares semidefinite programs are convex relaxations of polynomial optimization problems – they have seen extensive recent use in algorithm design for high-dimensional statistics. (See e.g. [RSS18a, Hop18b].) Roughly speaking, to show that SoS SDPs can efficiently certify a bound on the optimum of the above optimization problem, we need to prove such an upper bound *using only arguments involving low-degree polynomials in u, b_i*. Now

we sketch that proof, which is the technical heart of our algorithm for covariance estimation.

First, we show by applying a bounded-differences concentration inequality to the value of the SoS SDP that the optimum value of the relaxation of Eq. (1.2) concentrates around its expectation with high probability $(1 - 2^{-k})$. (This bounded-differences step appears in the non-algorithmic context in [LM18b] and in the algorithmic context in [Hop18a].) Then we bound the expected value of the above problem via

$$\sum_{i \le k} b_i \le \frac{1}{r} \sum_{i \le k} b_i \langle \Sigma_i - \Sigma, uu^\top \rangle \le \frac{1}{r} \cdot \sqrt{k} \cdot \left(\sum_{i \le k} \langle \Sigma_i - \Sigma, uu^\top \rangle^2 \right)^{1/2},$$

where we have used Cauchy-Schwarz.

The polynomial on the right-hand side is a degree-4 polynomial in u with random coefficients; the goal is to upper bound its expected maximum on the unit sphere (via an argument which applies also to the SoS relaxation, which rules out standard approaches using ε-nets). In fact, since we need the bound $0.51k$ on the $\sum_{i \le k} b_i$, we will eventually take r large enough to compensate for whatever is our bound on $\sum_{i \le k} \langle \Sigma_i - \Sigma, uu^\top \rangle^2$. We want to keep r small, so we want the tighest bound possible.

Note that $\sum_{i \le k} \langle \Sigma_i - \Sigma, uu^\top \rangle^2$ is a sum of i.i.d. random polynomials. A standard approach to analyze the performance of SoS for such random polynomials is to first "unfold" the polynomial to a matrix (in this case $\sum_{i \le k} (\Sigma_i - \Sigma)^{\otimes 2}$) and then use matrix concentration inequalities to analyze the maximum eigenvalue of this random matrix. Such eigenvalue bounds will also apply to the SoS relaxation we work with in the end.

We use a similar approach, with a key technical twist: in previous applications of this idea, it was usually necessary to have an explicit expression for $\mathbb{E} M$, where M is the random matrix analogous to $(\Sigma_i - \Sigma)^{\otimes 2}$, and typically also for its inverse, in order to correctly "precondition" the random matrix before analyzing its top eigenvalue. Such an explicit representation would be easily accessible if the underlying data X were Gaussian or had independent coordinates, for example, which was the case in previous applications of SoS to random degree-4 polynomials. We do not have this luxury, since we only make the niceness assumption on the underlying random vector X.

Nonetheless, we are able to carry out the preconditioning strategy (which removes spurious large eigenvalues of $(\Sigma_i - \Sigma)^{\otimes 2}$) for any nice random variable X. Along the way we prove a new (albeit simple) SoS Bernstein inequality which may be of independent use (and in particular allows for simplifed proofs of some previous applications of SoS to random degree-4 polynomials – e.g. that of [BBH+12a]). See the full version of our paper for the SoS Bernstein inequality and our application to the random polynomial $\sum_{i \le k} \langle \Sigma_i - \Sigma, uu^\top \rangle^2$.

Certification to Search. Using similar techniques as [CFB19] developed for the case of ℓ_2 mean estimation, we turn our certification into an algorithm to *find* an r-median. Suppose that instead of knowing the true covariance Σ as above, in its place we have some guess $M \in \mathbb{R}^{d \times d}$. If the certification algorithm certifies that M is a median, then we can output M as our estimator for Σ. If not, we show that by rounding the above SoS relaxation we can instead update M to make it closer to Σ – we can replace it with $M + \Delta$ such that $\|M + \Delta - \Sigma\| \ll \|M - \Sigma\|$.

1.3 Related Work

Robust Statistics. The questions we address here are distinct from those addressed by a recent flurry of algorithmic work in *robust* statistics [DKK+16, LRV16] (see also [Li18, Ste18] for further references). In the latter setting, one studies statistics when the list of samples $X_1, \ldots, X_n$ contains a small constant fraction ε of adversarially-chosen outliers, and the primary focus is on achieving statistical error nearly as small as would be achieved by the classical estimators when $\varepsilon = 0$. By contrast, *our goal is to beat the error rate of the classical estimators when Gaussianity is violated.* One consequence is that we give estimators which come with small confidence intervals even for error probabilities as low as 2^{-d}; this high-probability regime is not addressed by the adversarial corruptions model.[7]

Median of Means. In heavy-tailed (constantly-many moments exist) settings, estimators based on empirical averages typically have poor statistical performance, because they are sensitive to large outliers. Our work falls in a long line which develop the *median of means* technique for high-probability estimators in the face of heavy tails. The median of means framework was first developed to estimate univariate heavy-tailed random variables [NY83a, JVV86, AMS99a]. Recent extensions to the multivariate case typically have two flavors: they are polynomial-time computable (e.g. [HS16a, LO11, Min15]) but statistically suboptimal, or statistically optimal ([LM19a, LM18a, LM16]) but apparently require exponential computation time. The first major exceptions to this rule came in 2018, starting with a polynomial-time statistically-optimal algorithm for mean estimation in ℓ_2 [Hop18a]. Because of reliance on high-degree sum of squares semidefinite programs, this algorithm has an enormous polynomial running time. The subsequent work [CFB19] brought the running time much closer to practicality by replacing some of the sum of squares tools with a gradient-descent style algorithm. ([LLVZ19, LD19] brought the running times down even further.) The present work builds substantially on ideas from both these papers.

Covariance Estimation. There is a long and rich literature on the problem of covariance estimation (see [FLL16] for an expository review). However, strong high-confidence guarantees for many such estimators rely on the assumption that the samples are drawn from a sub-Gaussian distribution. The problem of robustly estimating covariance only assuming boundedness of low-order moments on the underlying distribution has also received attention; however many rigorous theoretical results in this vein are either asymptotic (i.e. concern only the $n \to \infty$ limit for fixed dimensions d) and/or often impose strong parametric assumptions on the underlying distribution (i.e. requiring elliptical symmetry). See [T+87, FLL16] for example, for a coverage of several such results.

[7] One recent work, [LD19], shows that while the adversarial robustness model and the ones we consider here are incomparable, under some circumstances the same algorithm can give information-theoretically optimal estimates in both models. This work, however, does not address covariance estimation or linear regression – it is an interesting direction to understand to what extent algorithms for covariance estimation and linear regression can perform well across different models.

The state-of-the-art results for the problem we consider here have been recently achieved in the works of [MW18] and [MZ18]. These results have come in two flavors, paralleling recent work in the problem of heavy-tailed mean estimation: [MW18][Corollary 4.1] provides computationally-efficient but information-theoretically suboptimal estimators while [MZ18][Theorem 1.9] provides statistically-optimal estimators that require exponential time (in n, d) to compute.

Linear Regression. Like covariance estimation, linear regression is an old and well-studied topic and a thorough survey is out of the scope of this paper. Regression in the heavy-tailed and high-dimensional setting has been studied via the median-of-means framework in [LM16, HS16b, LM17]. There are also efficient outlier-robust algorithms for linear regression which use techniques besides median-of-means estimation – for instance, the iterative methods of [SBRJ19] – but none are yet known to achieve information-theoretically optimal error. In particular we are not aware of any which improve on the guarantees of [HS16b] in our setting, while our algorithms offer $\mathrm{poly}(d)$ improvements on the error rates of [HS16b].

Sum of Squares Algorithms for High-Dimensional Statistics. There has been a significant amount of recent work using the sum of squares (SoS) semidefinite programming hierarchy to design computationally efficient algorithms for unsupervised learning problems (see [RSS18b] for a survey). By now, SoS algorithms are the only ones known which gives state-of-the-art statistical performance among polynomial-time algorithms for a wide range of problems: dictionary learning, tensor decomposition, high-dimensional clustering, robust parameter estimation and regression, and more [BKS15, HL18, KSS18, MSS16, KKM18, BM16].

We note that one of our techniques for exploiting 8-th moments is inspired by a certain approach to using the Cauchy-Schwarz inequality in SoS proofs for bounding degree-3 random polynomials by degree-4 random polynomials. This technique is in turn inspired by refutation algorithms for random constraint satisfaction problems [FO07], and has been used in the design of SoS algorithms for several learning problems [GM15, HSS15, BM16]. We also note that the certify-or-gradient paradigm used by our algorithms, where gradients are furnished by solving SDPs, has previously appeared in robust and heavy-tailed mean estimation [CDG19, CFB19]; these works do not combine this technique with SoS SDPs of degree greater than 2.

Our algorithms using the SoS hierarchy run in polynomial time, but because of their reliance on solving large semindefinite programs they are impractical. However, numerous slow-but-polynomial-time SoS algorithms for high-dimensional statistics have led to algorithms with practical nearly-linear running times [SS17, DHL19, HSSS16, LD19, HSS19, CFB19]. We therefore hope that additional investigation can lead to SoS-inspired and practical algorithms with improved guarantees for heavy-tailed covariance estimation and regression.

Certifiable Hypercontractivity. Our algorithms for covariance estimation and linear regression assume the underlying random vector X is $(2, 8)$ certifiably hypercontractive. The certifiable hypercontractivity assumption was introduced in [KSS18, HL18] where it was used in designing algorithms for robust estimation and mixture model clustering. It has been used in the context of regression by [KKM18]. Previous work using certifiable hypercontractivity assumptions (for example in clustering mixture models) typically assumed the presence of a $\mathrm{poly}(d)$-factor more samples than information-theoretically necessary in order to ensure the convergence of empirical moments to these population averages. Since we are interested in fine-grained questions about the number of samples required to achieve certain rates of statistical error, a major portion of the technical work in our paper is to show that SoS algorithms can exploit structure in the population moments even with relatively few samples. [HL19, BBH+12b] investigate computational hardness questions surrounding certifiable hypercontractivity.

Recent work by [LM19c] designs an (inefficient) mean estimator not based on the median-of-means framework; instead the estimator is constructing by trimming the mean estimate. It is an interesting and open question to see whether generalizations of this trimmed estimator can be adapted to the problem of covariance estimation considered herein, and be made computable in poly-time.

ACKNOWLEDGEMENTS

We thank Tselil Schramm for helpful remarks as this manuscript was being prepared.

REFERENCES

[AMS99a] Noga Alon, Yossi Matias, and Mario Szegedy, *The space complexity of approximating the frequency moments*, Journal of Computer and system sciences **58** (1999), no. 1, 137–147.

[AMS99b] Noga Alon, Yossi Matias, and Mario Szegedy, *The space complexity of approximating the frequency moments*, J. Comput. Syst. Sci. **58** (1999), no. 1, 137–147.

[BBH+12a] Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou, *Hypercontractivity, sum-of-squares proofs, and their applications*, STOC, 2012, pp. 307–326.

[BBH+12b] Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou, *Hypercontractivity, sum-of-squares proofs, and their applications*, STOC, ACM, 2012, pp. 307–326.

[BKS15] Boaz Barak, Jonathan A. Kelner, and David Steurer, *Dictionary learning and tensor decomposition via the sum-of-squares method*, STOC, ACM, 2015, pp. 143–151.

[BM16] Boaz Barak and Ankur Moitra, *Noisy tensor completion via the sum-of-squares hierarchy*, COLT, JMLR Workshop and Conference Proceedings, vol. 49, JMLR.org, 2016, pp. 417–445.

[C+12] Olivier Catoni et al., *Challenging the empirical mean and empirical variance: a deviation study*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, vol. 48, Institut Henri Poincaré, 2012, pp. 1148–1185.

[CDG19] Yu Cheng, Ilias Diakonikolas, and Rong Ge, *High-dimensional robust mean estimation in nearly-linear time*, Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2019, pp. 2755–2771.

[CFB19] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett, *Fast mean estimation with sub-gaussian rates*, arXiv preprint arXiv:1902.01998 (2019).

[DHL19] Yihe Dong, Samuel B Hopkins, and Jerry Li, *Quantum entropy scoring for fast robust mean estimation and improved outlier detection*, arXiv preprint arXiv:1906.11366 (2019).

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart, *Robust estimators in high dimensions without the computational intractability*, FOCS, IEEE Computer Society, 2016, pp. 655–664.

[FLL16] Jianqing Fan, Yuan Liao, and Han Liu, *An overview of the estimation of large covariance and precision matrices*, 2016.

[FO07] Uriel Feige and Eran Ofek, *Easily refutable subformulas of large random 3cnf formulas*, Theory of Computing **3** (2007), no. 1, 25–43.

[GM15] Rong Ge and Tengyu Ma, *Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms*, arXiv preprint arXiv:1504.05287 (2015).

[HL18] Samuel B Hopkins and Jerry Li, *Mixture models, robustness, and sum of squares proofs*, Proceedings of the 50th Annual ACM SIGACT Symposium

on Theory of Computing, ACM, 2018, pp. 1021–1034.

[HL19] ———, *How hard is robust mean estimation?*, arXiv preprint arXiv:1903.07870 (2019).

[Hop18a] Samuel B Hopkins, *Sub-gaussian mean estimation in polynomial time*, arXiv preprint arXiv:1809.07425 (2018).

[Hop18b] Samuel Brink Klevit Hopkins, *Statistical inference and the sum of squares method.*

[HS16a] D. Hsu and S. Sabato, *Loss minimization and parameter estimation with heavy tails*, J. Mach. Learn. Res. **17** (2016).

[HS16b] Daniel Hsu and Sivan Sabato, *Loss minimization and parameter estimation with heavy tails*, The Journal of Machine Learning Research **17** (2016), no. 1, 543–582.

[HSS15] Samuel B. Hopkins, Jonathan Shi, and David Steurer, *Tensor principal component analysis via sum-of-square proofs*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 956–1006.

[HSS19] Samuel B Hopkins, Tselil Schramm, and Jonathan Shi, *A robust spectral algorithm for overcomplete tensor decomposition*, Conference on Learning Theory, 2019, pp. 1683–1722.

[HSSS16] Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer, *Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors*, STOC, ACM, 2016, pp. 178–191.

[JVV86] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani, *Random generation of combinatorial structures from a uniform distribution*, Theoretical Computer Science **43** (1986), 169–188.

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka, *Efficient algorithms for outlier-robust regression*, arXiv preprint arXiv:1803.03241 (2018).

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer, *Robust moment estimation and improved clustering via sum of squares*, Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, ACM, 2018, pp. 1035–1046.

[KWB19] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira, *Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio*, arXiv preprint arXiv:1907.11636 (2019).

[LD19] Guillaume Lecué and Jules Depersin, *Robust subgaussian estimation of a mean vector in nearly linear time*, arXiv preprint arXiv:1906.03058 (2019).

[Li18] Jerry Zheng Li, *Principled approaches to robust machine learning and beyond*, Ph.D. thesis, Massachusetts Institute of Technology, 2018.

[LLVZ19] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang, *A fast spectral algorithm for mean estimation with sub-gaussian rates*, arXiv preprint arXiv:1908.04468 (2019).

[LM16] Gabor Lugosi and Shahar Mendelson, *Risk minimization by median-of-means tournaments*, arXiv preprint arXiv:1608.00757 (2016).

[LM17] Gábor Lugosi and Shahar Mendelson, *Regularization, sparse recovery, and median-of-means tournaments*, arXiv preprint arXiv:1701.04112 (2017).

[LM18a] ———, *Near-optimal mean estimators with respect to general norms*, arXiv preprint arXiv:1806.06233 (2018).

[LM18b] ———, *Sub-gaussian estimators of the mean of a random vector*, Annals of Statistics (2018).

[LM19a] G. Lugosi and S. Mendelson, *Sub-Gaussian estimators of the mean of a random vector*, Ann. Statist. **47** (2019), no. 2, 783–794.

[LM19b] Gabor Lugosi and Shahar Mendelson, *Mean estimation and regression under heavy-tailed distributions–a survey*, arXiv preprint arXiv:1906.04280 (2019).

[LM19c] ———, *Robust multivariate mean estimation: the optimality of trimmed mean*, arXiv preprint arXiv:1907.11391 (2019).

[LO11] Matthieu Lerasle and Roberto I Oliveira, *Robust empirical mean estimators*, arXiv preprint arXiv:1112.3914 (2011).

[LRV16] Kevin A. Lai, Anup B. Rao, and Santosh Vempala, *Agnostic estimation of mean and covariance*, FOCS, IEEE Computer Society, 2016, pp. 665–674.

[Min15] S. Minsker, *Geometric median and robust estimation in Banach spaces*, Bernoulli **21** (2015), no. 4, 2308–2335.

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer, *Polynomial-time tensor decompositions with sum-of-squares*, FOCS, IEEE Computer Society, 2016, pp. 438–446.

[MW18] Stanislav Minsker and Xiaohan Wei, *Robust modifications of u-statistics and applications to covariance estimation problems*, arXiv preprint arXiv:1801.05565 (2018).

[MZ18] Shahar Mendelson and Nikita Zhivotovskiy, *Robust covariance estimation under l_4 − l_2 norm equivalence*, arXiv preprint arXiv:1809.10462 (2018).

[NY83a] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons, 1983.

[NY83b] Arkadii Semenovich Nemirovsky and David Borisovich Yudin, *Problem complexity and method efficiency in optimization.*

[Pea01] Karl Pearson, *Liii. on lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2** (1901), no. 11, 559–572.

[RSS18a] Prasad Raghavendra, Tselil Schramm, and David Steurer, *High-dimensional estimation via sum-of-squares proofs*, arXiv preprint arXiv:1807.11419 (2018).

[RSS18b] ———, *High-dimensional estimation via sum-of-squares proofs*, arXiv preprint arXiv:1807.11419 (2018).

[SBRJ19] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain, *Adaptive hard thresholding for near-optimal consistent robust regression*, arXiv preprint arXiv:1903.08192 (2019).

[SS17] Tselil Schramm and David Steurer, *Fast and robust tensor decomposition with applications to dictionary learning*, Proceedings of Machine Learning Research vol **65** (2017), 1–34.

[Ste18] Jacob Steinhardt, *Robust learning: Information theory and algorithms*, Ph.D. thesis, Stanford University, 2018.

[T+87] David E Tyler et al., *A distribution-free m-estimator of multivariate scatter*, The annals of Statistics **15** (1987), no. 1, 234–251.

[Wik19a] Wikipedia contributors, *Least squares— Wikipedia, the free encyclopedia*, 2019, [Online; accessed 24-July-2019].

[Wik19b] ———, *Log-normal distribution*, 2019, [Online; accessed 22-July-2019].