The Annals of Statistics
2020, Vol. 48, No. 2, 1193–1213
https://doi.org/10.1214/19-AOS1843

MEAN ESTIMATION WITH SUB-GAUSSIAN RATES IN POLYNOMIAL TIME

BY SAMUEL B. HOPKINS[1]

[1] *Department of Electrical Engineering and Computer Science, University of California, Berkeley,* hopkins@berkeley.edu

We study polynomial time algorithms for estimating the mean of a heavy-tailed multivariate random vector. We assume only that the random vector X has finite mean and covariance. In this setting, the radius of confidence intervals achieved by the empirical mean are large compared to the case that X is Gaussian or sub-Gaussian.

We offer the first polynomial time algorithm to estimate the mean with sub-Gaussian-size confidence intervals under such mild assumptions. Our algorithm is based on a new semidefinite programming relaxation of a high-dimensional median. Previous estimators which assumed only existence of finitely many moments of X either sacrifice sub-Gaussian performance or are only known to be computable via brute-force search procedures requiring time exponential in the dimension.

1. Introduction. This paper studies estimation of the mean of a heavy-tailed multivariate random vector from independent samples. In particular, we address the question: *Are statistically optimal confidence intervals for heavy-tailed multivariate mean estimation achievable by polynomial-time computable estimators?* Our main result answers this question affirmatively, up to some explicit constants.

Estimating the mean of a distribution from independent samples is among the oldest problems in statistics. From the *asymptotic* viewpoint (i.e., when the number of samples n tends to infinity) it is well understood. If $X_1, \dots, X_n$ are n independent copies of a random variable X on $\mathbb{R}^d$, the empirical mean $\overline{\mu}_n = \frac{1}{n}\sum_{i\le n} X_i$ converges in probability to the mean $\mu = \mathbb{E}X$. If X has finite variance, the limiting distribution of $\overline{\mu}_n$ is Gaussian.

Aiming for finer-grained (finite-sample) guarantees, this paper takes a *nonasymptotic* view. For every $\delta > 0$ and $n \in \mathbb{N}$ we ask for an estimator $\hat{\mu}_{n,\delta}$ which comes with a tail bound of the form

$$\mathop{\mathbb{P}}_{X_1,\dots,X_n}\{\|\hat{\mu}_{n,\delta}(X_1,\dots,X_n) - \mu\| > r_\delta\} \le \delta$$

for as small a radius r_δ (which may depend on n and the distribution of X) as possible. That is, we are interested in estimators with the smallest-possible confidence intervals.

When X is Gaussian or sub-Gaussian, strong nonasymptotic guarantees are available on confidence intervals of the sample mean $\overline{\mu}_n$. Applying Gaussian concentration, if X has covariance Σ, then in the Gaussian setting,

$$\mathbb{P}\left\{\|\overline{\mu}_n(X_1,\dots,X_n) - \mu\| > \sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{2\|\Sigma\|\log(1/\delta)}{n}}\right\} \le \delta, \tag{1.1}$$

where $\|\Sigma\| = \lambda_{\max}(\Sigma)$ is the operator norm/maximum eigenvalue of Σ.

However, if one tries to replace the assumption that X is Gaussian with something weaker, equation (1.1) breaks down for the sample mean $\overline{\mu}_n$. For instance, consider a much weaker

Received October 2018; revised February 2019.
MSC2010 subject classifications. Primary 62H12; secondary 68W.
Key words and phrases. Multivariate estimation, heavy tails, confidence intervals, sub-Gaussian rates, semidefinite programming, sum of squares method.

assumption: X has finite covariance Σ. Then the best possible tail inequality for the sample mean becomes

$$\mathbb{P}\left\{\|\overline{\mu}_n(X_1,\dots,X_n)-\mu\| > \sqrt{\frac{\operatorname{Tr}\Sigma}{\delta n}}\right\} \le \delta \tag{1.2}$$

(see, e.g., [14], Section 6). By comparison with equation (1.1), the tail bound equation (1.2) has degraded in two ways: first, the $\log(1/\delta)$ term has become $1/\delta$, and second, that term multiplies $\operatorname{Tr}\Sigma$ rather than $\|\Sigma\|$; note that $\operatorname{Tr}\Sigma$ may be as large as $d\|\Sigma\|$, as in the case of isotropically distributed data.

This paper focuses on finding estimators $\hat{\mu}$ which can match (1.1) under milder assumptions than sub-Gaussianity, such as the existence of finitely many moments. Weak assumptions like this allow for the presence of *heavy tails*. A d-dimensional random vector X is heavy-tailed if for some unit $u \in \mathbb{R}^d$, the tail of $\langle X, u\rangle$ outgrows any exponential distribution; that is, for all $s > 0$ one has $\lim_{t\to\infty} e^{ts}\mathbb{P}\{\langle X-\mu, u\rangle > t\} = \infty$.

There are many situations in which one may wish to avoid a Gaussian or sub-Gaussian assumption. One may simply wish to be conservative, or there may reason to believe a Gaussian assumption is unjustified—heavy-tailed and high-dimensional data are not unusual. Many distibutions in big-data settings have heavy tails: for example, *power law* distributions consistently emerge from statistics of large networks (the internet graph, social network graphs, etc.) [21, 37]. And no matter how nice the underlying distribution, corruptions and noise in collected data often result in an empirical distribution with many outliers [51]. As a result, such X may have only a few finite moments; that is, $\mathbb{E}X^p$ may not exist for large-enough $p \in \mathbb{N}$.

This suggests the question of whether an estimator with a guarantee matching equation (1.1) (up to universal constants) exists under only the assumption that X has finite mean and covariance. (These assumptions are necessary to obtain the $1/\sqrt{n}$ rate in both equations (1.1) and (1.2).) One may show this is impossible if a single estimator is desired to satisfy an inequality like equation (1.1) [19].

Quite remarkably, the story changes if the estimator may additionally depend on the desired confidence level $1-\delta$. Indeed, by now in the classical case $d = 1$, many such δ-*dependent* estimators are known which achieve equation (1.1) up to explicit constants for $\delta \ge 2^{-O(n)}$, even when X has only finite mean and variance [14, 19]. Since the δ-dependence is a necessary concession to achieve concentration like equation (1.1) with only two finite moments, for this paper our estimators are all allowed to depend on δ: it is an interesting future direction to explore what fraction of the theory may be reproduced without the δ-dependence [19, 42]. The lower bound $\delta \ge 2^{-O(n)}$ is also information-theoretically necessary [19].

The high-dimensional case is much more difficult, and has been resolved only recently: the culmination of a series of works [28, 36, 38, 41] is the following theorem of Lugosi and Mendelson, who gave the family of estimators matching equation (1.1) (up to constants) for any d under only the assumption of finite second moments. (In fact, their result also holds in the infinite-dimensional Banach space setting.)

THEOREM 1.1 (Lugosi–Mendelson estimator, [38]). *There is a universal constant C such that for every n, d, and $\delta \ge 2^{-n/C}$ there is an estimator $\hat{\mu}_{\delta,n} : \mathbb{R}^{dn} \to \mathbb{R}^d$ such that for every random variable X on $\mathbb{R}^d$ with finite mean and covariance,*

$$\mathbb{P}\left\{\|\hat{\mu}_{n,\delta}(X_1,\dots,X_n)-\mu\| > C\left(\sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{n}}\right)\right\} \le \delta,$$

where $X_1,\dots,X_n$ are i.i.d. copies of X and $\mu = \mathbb{E}X$ and $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^\top$.

In high-dimensional estimation, especially with large data sets, it is important to study estimators with guarantees both on statistical accuracy and algorithmic tractability. Indeed, there is growing evidence that some basic high-dimensional estimation tasks which appear possible from a purely information-theoretic perspective altogether lack computationally efficient algorithms. There are many examples of such *information-computation gaps*, including the problem of finding sparse principal components of high-dimensional data sets (the *sparse PCA problem*) and optimal detection of hidden communities in random graphs with latent community structure (the *k-community stochastic block model*) [5, 9, 18, 25, 27, 40].

From this perspective, a major question left open by Theorem 1.1 is whether there exists an estimator matching Theorem 1.1 but which is efficiently computable. In this paper, *efficiently computable* means computable by an algorithm running in time $(nd\log(1/\delta))^{O(1)}$—that is, polynomial in both the number of samples and the ambient dimension, as well as the number of bits needed to describe the input $\delta > 0$. Indeed, the *median-of-means* estimator used by Lugosi and Mendelson lacks any obvious algorithm running in time less than $\exp(cd)$, for some fixed $c > 0$, which is the time required for brute-force search over every direction in a d-dimensional ε-net. More worringly, the key idea of Lugosi and Mendelson is a combinatorial notion of a multivariate median, which appears to place the problem dangerously near those high-dimensional combinatorial statistics problems which lack efficient algorithms altogether.

The main result of this paper shows that there is a family of estimators matching Theorem 1.1 and computable by polynomial-time algorithms.

THEOREM 1.2 (Main theorem). *There are universal constants C_0, C_1, C_2 such that for every $n, d \in \mathbb{N}$ and $\delta > 2^{-n/C_2}$ there is an algorithm which runs in time $O(nd) + (d\log(1/\delta))^{C_0}$ such that for every random variable X on $\mathbb{R}^d$, given i.i.d. copies $X_1, \dots, X_n$ of X the algorithm outputs a vector $\hat{\mu}_\delta(X_1, \dots, X_n)$ such that*

$$\mathbb{P}\left\{\|\mu - \hat{\mu}_\delta\| > C_1\left(\sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{n}}\right)\right\} \le \delta,$$

where $\mathbb{E}X = \mu$ and $\mathbb{E}(X-\mu)(X-\mu)^\top = \Sigma$.

On constants and running times. No effort has been made to optimize the constants C_0, C_1, C_2. By careful analysis they may certainly be made less than 1000, but we expect substantial improvements beyond this are possible.

Because of the large polynomial running time, we regard Theorem 1.2 as mainly *a (constructive) proof of the existence of a polynomial-time algorithm*: of course, we do not suggest anyone attempt to run an $(nd)^{1000}$-time algorithm in practice! Polynomial-time algorithms are qualitatively different from exponential-time brute-force searches, however, and very often the insights from a slow polynomial-time algorithm can be leveraged to design a fast one, while the same cannot be said of a brute-force search procedure. Thus, when addressing challenging algorithmic questions in high-dimensional statistics, the first question is whether there is a polynomial-time algorithm at all: Theorem 1.2 answers this affirmatively.

Indeed, Theorem 1.2 and the algorithm behind it have already inspired further investigation into the (rather distinct) question of just how fast an algorithm is possible. After the present work was initially circulated, Cherapanamjeri, Flammarion and Bartlett combined the ideas in our Section 2 with a nonconvex gradient descent procedure to obtain an algorithm with the statistical same guarantees as Theorem 1.2 but with running time $O(n^{3.5} + n^2 d) \cdot (\log nd)^{O(1)}$ [15]. It is more than plausible that further developments will lead to a truly practical algorithm (with running time, say, $nd \cdot \log(nd)^{O(1)}$—note that input vectors consist of nd real numbers, so this running time would correspond to reading the data $\log(nd)^{O(1)}$ times).

Semidefinite programming, proofs to algorithms and the sum of squares method. Our algorithm is based on semidefinite programming (SDP). It is not an attempt to directly compute the estimator proposed by Lugosi and Mendelson. Instead, inspired by that estimator, we introduce MEDIAN-SDP, a new semidefinite programming approach to computation of a high-dimensional median. We hope that the ideas behind it will find further uses in algorithms for high-dimensional statistics.

Our SDP arises from the *sum of squares (SoS)* method, which is a powerful and flexible approach to SDP design and analysis. Rather than design an SDP from scratch and invent a new analysis, guided by the SoS method we construct an SDP whose variables and constraints allow for the *proof* of Lugosi and Mendelson's Theorem 1.1 to translate directly to an *analysis* of the SDP, proving our Theorem 1.2. (More prosaically: Lugosi and Mendelson's proof inspires the construction of a family of dual solutions to our SDP, which then we use to argue that it recovers a good estimate for the mean.)

This technique, which turns sufficiently simple proofs of identifiability like the proof of Theorem 1.1 into algorithms as in Theorem 1.2, has recently been employed in algorithm design for several computationally challenging statistics problems. For instance, recent works offer the best available polynomial-time guarantees for parameter estimation of high-dimensional mixture models and for estimation in Huber's contamination model [26, 29, 31, 32]. SoS has also been key to progress in computationally challenging tensor problems with statistical applications, such as tensor decomposition (a key primitive for moment-method algorithms in high dimensions) and tensor completion [6, 39, 48]. For further discussion see the survey [49]. We expect many further basic statistical problems for which efficient algorithms are presently unknown to be successfully attackable with the SoS method.

Organization. In the remainder of this introduction we discuss the *median of means* estimation paradigm which underlies both Lugosi and Mendelson's estimator (Theorem 1.1) and our own (Theorem 1.2) and briefly introduce the SoS method, as well as offer some comparisons of the SDP used in this paper to some common SDPs employed in statistics. Before turning to technical material, in Section 1.3 we give a brief overview of our estimator.

In Section 2, we describe an algorithm for a twist on the mean estimation problem, called the *certification* problem. The main lemma analyzes an SDP whose solutions capture information about quantiles of a set of high-dimensional vectors. It is the key tool in the design of our algorithm to estimate the mean. This section requires no background on SoS.

Then, in Section 3 we give some formal definitions and standard theorems about SoS. In Section 4 we prove our main theorem from technical lemmas, whose proofs can be found in the Supplementary Material [24].

1.1. *The median of means paradigm.*

The *median of means* is an approach to mean estimation for heavy-tailed distributions which combines the reduction in variance offered by averaging independent samples (thus achieving $1/\sqrt{n}$ convergence rates) with the outlier-robustness of the median (thus achieving $\sqrt{\log(1/\delta)}$ tail behavior) [2, 30, 44]. Consider the $d = 1$ case first. Suppose $X_1, \dots, X_n$ are i.i.d. copies of a real-valued random variable X with mean $\mu \in \mathbb{R}$ and variance σ^2. Let $k = \Theta(\log 1/\delta)$ be an integer, and for $i \le k$ let Z_i be the average of samples $X_{i \cdot n/k}$ to $X_{(i+1)\cdot n/k}$.[1] Then it is an exercise to show that the median (or indeed any fixed quantile) of the Z_i's satisfies

$$\mathbb{P}\left\{\left|\mathrm{median}(Z_1, \dots, Z_k) - \mu\right| > C\sigma\sqrt{\frac{\log(1/\delta)}{n}}\right\} \le \delta$$

[1] Throughout the paper we will assume that n is divisible by $C\log(1/\delta)$ for an appropriate constant C. One may achieve this from general n, k and $\delta \ge 2^{-O(n)}$ by throwing out samples to reach the nearest multiple of $C\log(1/\delta)$; the effect on the error rates is only a constant.

for some universal constant C (given the correct choice of k). There are estimators achieving this $\sqrt{\log(1/\delta)}$ rate using ideas other than the median of means in the case $d = 1$ [14, 19], but we focus here on median of means since it is the only approach known to prove a theorem like Theorem 1.1 in the high dimensional case.

Correctly extending this median of means idea to higher dimensions d is not simple. Suppose that X is d-dimensional, with mean μ and covariance Σ. Replacing $X_1, \dots, X_n \in \mathbb{R}^d$ with grouped averages $Z_1, \dots, Z_k \in \mathbb{R}^d$ remains possible, but the sticking point is to choose an appropriate notion of median or quantile in d dimensions.

A first attempt would be to use as a median of $Z_1, \dots, Z_k$ any point in $\mathbb{R}^d$ which has at most some distance r to at least ck of $Z_1, \dots, Z_k$ for some $c > 1/2$. Let us call such a point a *simple r-median*. It is straightforward to prove, by the same ideas as in the $d = 1$ case, that

$$\mathbb{P}\left\{ \|\mu - Z_i\| > C\sqrt{\frac{\operatorname{Tr}\Sigma \log(1/\delta)}{n}} \text{ for at least } ck \text{ vectors } Z_i \right\} \le \delta$$

for some universal constant $C = C(c)$. It follows that with probability at least $1 - \delta$ the mean μ is a simple r-median for $r = C\sqrt{\operatorname{Tr}\Sigma \log(1/\delta)/n}$. When $c > 1/2$, any two simple r-medians must each have distance at most r to some Z_i, so by the triangle inequality,

$$\mathbb{P}\left\{ \|\text{simple } 2r\text{-median}(Z_1, \dots, Z_k) - \mu\| > 2C\sqrt{\frac{\operatorname{Tr}\Sigma \log(1/\delta)}{n}} \right\} \le \delta, \tag{1.3}$$

where simple $2r$-median$(Z_1, \dots, Z_k)$ is any simple $2r$-median of $Z_1, \dots, Z_k$. At the cost of replacing $2r$ by $4r$, a simple r-median can be found easily in polynomial time (in fact in quadratic time) because if there is any simple $2r$-median of $Z_1, \dots, Z_k$ then by triangle inequality some Z_i must be a simple $4r$-median.

In prior work, Minsker shows that the geometric median of $Z_1, \dots, Z_k$ achieves the same guarantee equation (1.3) as the simple median (perhaps with a different universal constant C) [41]. Geometric median is computable in nearly linear time (i.e., time $dk \cdot (\log dk)^{O(1)}$) [16].

The guarantee equation (1.3) represents the smallest confidence intervals previously known to be achievable by polynomial-time computable mean estimators under the assumption that X has finite mean and covariance. This tail bound is an intermediate between the $\sqrt{\operatorname{Tr}\Sigma/\delta n}$-style tail bound achieved by the empirical mean equation (1.2) and the Gaussian-style guarantee of Lugosi and Mendelson from Theorem 1.1. It fails to match Theorem 1.1 because the $\log(1/\delta)$ term multiplies $\operatorname{Tr}\Sigma$ rather than $\|\Sigma\|$—this introduces an unnecessary dimension-dependence. That is, if X has covariance identity, then informally speaking the rate of tail decay has a dimension-dependent factor when it should be dimension-independent: it decays as $\exp(-ct^2/d)$ rather than $\exp(-ct^2)$ (where c is some fixed constant).[2] This is not a failure of the analysis: if the approach is to draw a ball around the population mean μ which contains at least a constant fraction of $Z_1, \dots, Z_k$ with probability $1 - \delta$, the ball must have radius of order $\sqrt{\operatorname{Tr}\Sigma \log(1/\delta)/n}$, which grows with the dimension of X.

To prove Theorem 1.1, Lugosi and Mendelson introduce a new notion of high-dimensional median, which arises from what they call a *median of means tournament*. This tournament median of $Z_1, \dots, Z_k$ is

$$\arg\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} \|x - y\| \quad \text{such that } \|Z_i - x\| \ge \|Z_i - y\| \text{ for at least } \frac{k}{2} \ Z_i\text{'s}. \tag{1.4}$$

Rephrased, the tournament median is the point $x \in \mathbb{R}^d$ minimizing the number r such that for every unit $u \in \mathbb{R}^d$, the projection $\langle x, u \rangle$ is at distance at most r from a median of the projections $\{\langle Z_i, u \rangle\}$.[3]

[2] Of course, formally we are talking about one estimator $\hat{\mu}_\delta$ for every δ, so it is not correct to speak of tail decay with respect to δ.

[3] Thanks to Jerry Li for pointing out this reinterpretation of the tournament median to me.

In fact, Lugosi and Mendelson's arguments apply to any x which *r-central* in the following sense: for every unit u, there are at least $0.51k$ vectors among $Z_1, \dots, Z_k$ such that $|\langle Z_i, u\rangle - \langle x, u\rangle| \leq r$. Their proof shows that an estimator which outputs any r-central point will achieve the guarantee in Theorem 1.1. This interpretation shows that their estimator is related to a weak notion of Tukey median: a Tukey median (at least in the typical case that it has constant Tukey depth) should be between a 49th and 51st percentile in every direction u, while an r-central point has distance at most r to such a percentile in every direction u [53]. Thus our result Theorem 1.2 adds to several in the literature which demonstrate that although the Tukey median of vectors $v_1, \dots, v_k \in \mathbb{R}^d$ is NP-hard to compute if $v_1, \dots, v_k$ are chosen adversarially, under reasonable assumptions (in this case that $Z_1, \dots, Z_k$ are i.i.d. from a distribution with bounded covariance) one may find some kind of approximate Tukey median in polynomial time [8, 20, 34].

The heart of the proof of Theorem 1.1 shows that with probability at least $1 - \delta$, the mean μ is r-central for $r = C(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{\|\Sigma\| \log(1/\delta)/n})$. The difficulty in *computing* the tournament median—or finding some r-central point—comes from the fact that in each direction u it may be different collection of $0.51k$ vectors which satisfy $|\langle Z_i, u\rangle - \langle x, u\rangle| \leq r$. Thus even if an algorithm is given $Z_1, \dots, Z_k$ *and* μ, to efficiently check that μ is a tournament median or is r-central seems naively to require brute-force search over $\exp(cd)$ directions in $\mathbb{R}^d$, for some fixed $c > 0$. The heart of our algorithm is a semidefinite program which (with high probability) can efficiently *certify* that μ is r-central: this algorithm is described in Section 2.

1.2. *Semidefinite programming and the SoS method in statistics.*

One of the main tools in our algorithm is semidefinite programming, and in particular the sum of squares method. Recall that a semidefinite program (SDP) is a convex optimization problem of the following form:

$$\min_X \langle X, C\rangle \quad \text{such that } \langle A_1, X\rangle \geq 0, \dots, \langle A_m, X\rangle \geq 0 \text{ and } X \succeq 0, \tag{1.5}$$

where X ranges over symmetric $n \times n$ real matrices and $\langle M, N\rangle = \operatorname{Tr} MN^\top$. Subject to mild conditions on C and $A_1, \dots, A_m$, semidefinite programs are solvable to arbitrary accuracy in polynomial time [11].

Semidefinite programming as a tool for algorithm design has by now seen numerous uses across both theoretical computer science and statistics. Familiar SDPs in statistics include the nuclear-norm minimization SDP, used for matrix sensing and matrix completion [12, 13], the Goemans–Williamson cut SDP, variants of which are used for community detection in sparse graphs [1, 23, 43], SDPs for finding sparse principal components [4, 17, 33], SDPs used for high-dimensional change-point detection [55], SDPs used for optimal experiment design [54] and more.

While much work has focused on detailed analyses of a small number of canonical semidefinite programs—the nuclear-norm SDP, the Goemans–Williamson SDP, etc.—the SoS method offers a rich variety of semidefinite programs suited to many purposes [35, 45, 47, 52]. For every *polynomial optimization problem with semialgebraic constraints*, SoS offers a *hierarchy* of SDP relaxations. That is, for every collection of multivariate polynomials $p, q_1, \dots, q_m \in \mathbb{R}[x_1, \dots, x_n]$ and every even $r \geq \max(\deg p, \deg q_1, \dots, \deg q_m)$, SoS offers a relaxation of the problem

$$\min p(x) \quad \text{such that } q_1(x) \geq 0, \dots, q_m(x) \geq 0.$$

As r increases, the relaxations become stronger, more closely approximating the true optimum value of the optimization problem, but the complexity of the relaxations also increases. Typically, the rth relaxation is solvable in time $(nm)^{O(r)}$. In many applications, such as when

$q_1, \dots, q_m$ include the constraints $x_i^2 - x \geq 0$, $x_i^2 - x \leq 0$ which imply $x \in \{0,1\}^n$, when $r = n$ the SoS SDP exactly captures the optimum of the underlying polynomial optimization problem. However, the resulting SDP has at least 2^n variables, so is not generally solvable in polynomial time. This paper focuses on SoS SDPs with $r = O(1)$ (in fact $r = 8$), leading to polynomial-time algorithms.

SoS carries at least two advantages relevant to this paper over more classical approaches to semidefinite programming. First is the flexibility which comes from the possibility of beginning with any set of polynomials $p, q_1, \dots, q_m$; we choose polynomials which capture the idea of r-centrality. Second is ease of analysis: SoS SDPs in statistical settings are amenable to an analysis strategy which converts proofs of statistical identifiability into analysis of an SDP-based algorithm by phrasing the identifiability proof as a dual solution to the SDP. This style of analysis is feasible in our case because the SoS SDP has enough constraints that many properties of r-centrality carry over to the relaxed version: it is not clear whether a more elementary SDP would share this property.

1.3. *Algorithm overview.* Recall where we left off in Section 1.1. Having taken samples $X_1, \dots, X_n$ from a distribution with mean μ and covariance Σ and averaged groups of n/k of them to form vectors $Z_1, \dots, Z_k$, the goal is to find a median of $Z_1, \dots, Z_k$. As we discussed, the appropriate notion of a median is any point $x \in \mathbb{R}^d$ which is r-central for $r = O(\sqrt{\operatorname{Tr}\Sigma/n} + \sqrt{\|\Sigma\| \log(1/\delta)/n})$, meaning that for every 1-dimensional projection $\langle x, u\rangle$, $\langle Z_1, u\rangle$, ..., $\langle Z_k, u\rangle$, the point $\langle x, u\rangle$ has distance at most r to a 0.51-quantile of of $\{\langle Z_i, u\rangle\}$

Let us change the problem temporarily with a thought experiment: imagine being given $Z_1, \dots, Z_k$ *and* the population mean μ and being asked to verify (or in computer science jargon, *certify*) that indeed μ is r-central. Even for this apparently simpler task there is no obvious polynomial-time algorithm: a brute-force inspection of $\{\langle Z_i, u\rangle\}$ for, say, all u in an ε-net of the unit ball in $\mathbb{R}^d$ will require time $(1/\varepsilon)^d$.

Our first technical contribution is to show that with high probability over $Z_1, \dots, Z_k$ there is a *short certificate*, or *witness*, to the fact that the population mean μ has distance at most r to a median in every direction. This certificate takes the form of a dual solution to a semidefinite relaxation of the following combinatorial optimization problem: given $Z_1, \dots, Z_k, \mu$ and $r > 0$, maximize over all directions u the number of $i \in [k]$ such that $\langle Z_i - \mu, u\rangle \geq r$. Solving this SDP gives an algorithm for the certification problem: we show that with probability at least $1 - \delta$ the maximum value is at most $k/3$ for the choice of r above. We note that this SDP and its analysis do not rely on the SoS technology, so all of Section 2 can be read without this background.

Returning to the problem of estimating μ given $Z_1, \dots, Z_k$, the task is made simpler by the existence of the certificate that μ is r-central. In particular, it gives a concrete object which our estimation algorithm can search for: we know it will suffice to find any point in the set:

$$\begin{aligned}&\textsc{certifiable-centers}(Z_1, \dots, Z_k)\\ &\quad = \{(x, M) : x \in \mathbb{R}^d, M \in \mathbb{R}^{(d+k+1)\times(d+k+1)} \text{ certifies } x \text{ is } r\text{-central}\},\end{aligned}$$

which is nonempty because it in particular contains (μ, M_μ), where M_μ is the aforementioned SDP dual solution. (It is not yet obvious why $(d+k+1) \times (d+k+1)$ is the appropriate dimension for M; we will see this in the next section.) Our second technical contribution is an algorithm which we call MEDIAN-SDP, based on the SoS method, which takes $Z_1, \dots, Z_k$ and finds $x' \in \mathbb{R}^d$ such that $\|x - x'\| = O(r)$ for every $x \in$ CERTIFIABLE-CENTERS.

The algorithm is based on an SDP relaxation of the set CERTIFIABLE-CENTERS, this time based on the SoS method. The relaxation is designed to accommodate the following kind of analysis: we turn the following simple argument about r-central points into a dual solution to the SDP (in the SoS context this object is called an *SoS proof*), then use the latter to show that the SDP finds a good estimator x.

The argument which we must turn into an SoS proof is the following: if x, x' are r-central then consider in particular the direction $v = (x - x')/\|x - x'\|$. There exists some Z_i such that $\langle x, v\rangle \le r + \langle Z_i, v\rangle$ and $\langle x', -v\rangle \le r + \langle Z_i, -v\rangle$. Adding the inequalities gives $\langle x - x', v\rangle = \|x - x'\| \le 2r$. When we make this argument into an SoS proof, it will imply (roughly speaking) not just when x is r-central but also when x is in our relaxation of the set CERTIFIABLE-CENTERS.

This strategy will rely crucially on both the existence of the certificate μ (needed to turn the above argument into an SoS proof) and the SoS strategy for designing SDPs (to accommodate the complexity of the resulting dual solution). For more discussion, see Section 2 of the Supplementary Material [24].

2. Certifying centrality. In this section we describe and analyze one of the key components of our algorithm: a semidefinite program to certify the main property of the population mean our algorithm exploits—(r, p)-centrality.

DEFINITION 2.1 (Centrality). Let $Z_1, \dots, Z_k \in \mathbb{R}^d$, $r > 0$, and $p \in [0, 1]$. We say that $x \in \mathbb{R}^d$ is (r, p)-central (with respect to $Z_1, \dots, Z_k$) if for every unit $u \in \mathbb{R}^d$ there are at most pk vectors $Z_1, \dots, Z_k$ such that $\langle Z_i - x, u\rangle \ge r$.

At the heart of Lugosi and Mendelson's mean estimator is the following remarkable lemma, characterizing centrality of the population mean.

LEMMA 2.2 ([38], rephrased). *Let Z be a d-dimensional random vector with mean $\mu = \mathbb{E}Z$ and covariance Σ. Let $Z_1, \dots, Z_k$ be i.i.d. copies of Z. With probability at least $1 - 2^{-\Omega(k)}$, the population mean μ is $(r, 1/3)$-central with respect to $Z_1, \dots, Z_k$, for $r = O(\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|})$.*[4]

The main difficulty in proving Lemma 2.2 (and our later algorithmic versions of it) is to simultaneously obtain the tight quantitative bound $r = O(\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|})$ and the high probability $1 - 2^{-\Omega(k)}$. Without both, one does not get an estimator matching Theorem 1.1.

Suppose, as in the median of means paradigm, Z is taken as the empirical average of n/k i.i.d. copies $X_1, \dots, X_{n/k}$ of another random vector X having covariance Σ'. Then $\Sigma = \frac{k}{n}\Sigma'$. One may see that if $k = \Theta(\log(1/\delta))$ the mean μ is $(r, 1/3)$-central for $r = O(\sqrt{\operatorname{Tr}\Sigma'/n} + \sqrt{\|\Sigma'\|\log(1/\delta)/n})$ with probability at least $1 - \delta$. Any two $(r, 1/3)$ central points x, y also have $\|x - y\| \le 2r$ (see Section 1.3), and thus it follows that to obtain the guarantees of Theorem 1.1, given $Z_1, \dots, Z_k$ one only needs to output any $(r, 1/3)$-central point.

2.1. *Certification and the failure of empirical moments.* A natural avenue to designing an efficient algorithm matching Theorem 1.1 is to try to compute an $(r, 1/3)$-central point given $Z_1, \dots, Z_k$. A first roadblock is that there is not an obvious efficient algorithm for the following apparently simpler problem: given $x \in \mathbb{R}^d$, decide whether x is an $(r, 1/3)$-central point—brute-force search over 2^d one-dimensional projections must be avoided. In this section we give an efficient algorithm for a slight twist of this problem, which we call the *certification* problem.

[4] We write $f(n) = O(g(n))$ if there is a constant C such that for all large-enough n one has $f(n) \le Cg(n)$. Similarly, we write $f = \Omega(g(n))$ if there is c such that $f(n) \ge cg(n)$ for large-enough n. We write $f = \Theta(g(n))$ if both $f = O(g(n))$ and $f = \Omega(g(n))$.

PROBLEM 2.3 (Certification). Given $Z_1, \dots, Z_k, x \in \mathbb{R}^d$ and $r > 0$ and $p \in [0,1]$, a certification algorithm may output YES or DO NOT KNOW. If the output is YES, then x must be (r,p)-central with respect to $Z_1, \dots, Z_k$. If the output is DO NOT KNOW, then x may or may not be (r,p)-central.

Our goal is to design a certification algorithm with parameters matching Lemma 2.2. That is, we would like a certification algorithm which outputs YES with probability at least $1 - 2^{-\Omega(k)}$ over $Z_1, \dots, Z_k$ when given $x = \mu$ and $r = O(\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|})$ and p a small constant. This is an easier task than deciding (r,p)-centrality exactly, since we care only about those configurations of $Z_1, \dots, Z_k$ which may arise as i.i.d. copies of a random vector Z with covariance Σ, and even when μ is (r,p)-central we allow the algorithm to output DO NOT KNOW, so long as this does not happen too often. We prove the following theorem, which we view as an algorithmic version of Lemma 2.2.

THEOREM 2.4. *There is an algorithm for the certification problem with running time $(kd)^{O(1)}$ and the guarantee that if $Z_1, \dots, Z_k$ are i.i.d. copies of a random variable Z with mean μ and covariance Σ then the algorithm outputs* YES *with probability at least* $1 - 2^{-\Omega(k)}$ *given $p = 1/100$*[5] *and $r = O(\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|})$.*

Our algorithm for the certification problem will be based on semidefinite programming. While our final algorithm to estimate μ (Theorem 1.2) will not directly employ this certification algorithm as a subroutine, the semidefinite program we analyze for the latter is at the heart of the former.

On the failure of empirical moments. Before we describe our certification algorithm and prove Theorem 2.4, we offer some intuition as to why a powerful tool such as semidefinite programming is necessary, by assessing simpler potential approaches to certification. A natural approach would involve the maximum eigenvalue $\lambda = \|\overline{\Sigma}\|$ of the empirical covariance $\overline{\Sigma} = \frac{1}{k}\sum_{i=1}^k (Z_i - \mu)(Z_i - \mu)^\top$. If a unit vector u has $\langle Z_i - \mu, u\rangle \geq r$ for more than $k/3$ vectors Z_i (thus violating $(r, 1/3)$-centrality), then $\frac{1}{k}\sum \langle Z_i - \mu, u\rangle^2 \geq r^2/3$. Thus the maximum eigenvalue λ (which is of course computable in polynomial time) would certify that μ is $(O(\sqrt{\lambda}), 1/3)$-central.

Unfortunately, because of our weak assumptions on Z—again, we only assume the second moment Σ exists—the maximum eigenvalue of the empirical covariance is poorly concentrated: for instance, with probability about 2^{-k} some vector Z_i may have norm as large as $\sqrt{\operatorname{Tr}\Sigma} \cdot 2^k$, resulting in $\sqrt{\lambda} \geq \sqrt{\operatorname{Tr}\Sigma} \cdot 2^{k/2}$. (Indeed, even the typical value of $\sqrt{\lambda}$ could be much larger than $\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|}$.) Straightforward approaches to address this—for example, discarding a constant fraction of the samples $Z_1, \dots, Z_k$ of largest norm, or replacing the second moment $\frac{1}{k}\sum_{i=1}^k \langle Z_i - \mu, u\rangle^2$ with the first moment $\frac{1}{k}\sum_{i=1}^k |\langle Z_i - \mu, u\rangle|$—offer some quantiative improvement over the empirical covariance, but still do not match the $\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|}$ bound with probability $1 - 2^{-\Omega(k)}$ which we are aiming for. Our semidefinite programming-based algorithm for certification can be viewed as a more sophisticated approach to improve the outlier-robustness of the maximum eigenvalue of the empirical covariance.

2.2. *The centrality SDP.* We turn to our certification algorithm and the proof of Theorem 2.4. To start, we design a convex relaxation of the following (nonconvex) optimization problem, which captures centrality: given $Z_1, \dots, Z_k, x$ and $r \geq 0$, find the minimum p such that x is (r,p)-central. Or, rephrased, find the *maximum* over directions u of the number of Z_i such that $\langle Z_i - x, u\rangle \geq r$. The latter we capture as the following *quadratic program*.

[5] The constant $1/100$ differs from the $1/3$ in Lemma 2.2 only for technical convenience later in this paper.

FACT 2.5. The minimum p such that x is (r, p)-central with respect to $Z_1, \dots, Z_k \in \mathbb{R}^d$ is given by the optimum of the following quadratic program in variables $b_1, \dots, b_k$ and $u_1, \dots, u_d$:

$$\begin{aligned} &\max_{u,b} \frac{1}{k}\sum_{i=1}^{k} b_i \quad \text{such that} \\ &b_1, \dots, b_k \in \{0, 1\}, \\ &\|u\|^2 \le 1, \\ &b_i\langle Z_i - x, u\rangle \ge b_i r \quad \text{for } i = 1, \dots, k. \end{aligned} \tag{2.1}$$

We relax the quadratic program (2.1) to a semidefinite program in standard fashion.

DEFINITION 2.6 (Centrality SDP). Given $Z_1, \dots, Z_k, x \in \mathbb{R}^d$ and $r \ge 0$, we define a semidefinite program over $(d + k + 1) \times (d + k + 1)$ positive semidefinite matrices with the following block structure:

$$Y(B, W, U, b, u) = \begin{pmatrix} 1 & b^\top & u^\top \\ b & B & W \\ u & W^\top & U \end{pmatrix},$$

where $B \in \mathbb{R}^{k\times k}$, $U \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^k$, $u \in \mathbb{R}^d$. As usual, the intended solutions of the SDP are rank-one matrices $(1, b, u)(1, b, u)^\top$ where $(b, u) \in \mathbb{R}^{d+k}$ is a solution to (2.1). The SDP is

$$\begin{aligned} &\max_{Y(B,W,U,b,u)} \frac{1}{k}\sum_{i=1}^{k} b_i \quad \text{such that} \\ &B_{ii} \le 1 \quad \text{for } i = 1, \dots, k, \\ &\operatorname{Tr} U \le 1, \\ &\langle Z_i - x, W_i\rangle \ge r \cdot b_i \quad \text{for } i = 1, \dots, k, \\ &Y(B, W, U, b, u) \succeq 0. \end{aligned}$$

Here W_i is the ith row of the $k \times d$ matrix W. It stands in for the vector $b_i \cdot u$ in (2.1).[6]

The centrality SDP is a relaxation of centrality proper: there is no *a priori* reason to believe that it faithfully captures the quadratic program (2.1). For instance, it could be that for most $Z_1, \dots, Z_k$ the r-centrality SDP value is 1, even though Lemma 2.2 says that with high probability the value of (2.1) is at most 1/3 in the median of means setting (for appropriate choice of r).

Remarkably, the opposite is true: at least in our median of means setting, the centrality SDP is a good approximation to the quadratic program it relaxes.[7] This is captured by the following key technical lemma, from which Theorem 2.4 follows immediately (because the SDP can be solved in polynomial time [11]).

[6] We remark that a more traditional SDP relaxation might only involve the large $(d + k) \times (d + k)$ block of Y, replacing b_i with B_{ii} in all constraints. However, the extra row and column $(1, b, u)$ will be of some technical use later in this paper; it is possible with some technical modifications to other proofs they could be removed.

[7] Here we do not mean approximation in the sense the word is used in *approximation algorithms*, since we are studying only the behavior of the SDP for $Z_1, \dots, Z_k$ being a collection of random vectors, and we prove only high probability guarantees, rather than probability-1 guarantees.

DEFINITION 2.7 (Certifiable Centrality). Let $Z_1, \dots, Z_k \in \mathbb{R}^d$, $r > 0$ and $p \in [0, 1]$. We say that $x \in \mathbb{R}^d$ is *certifiably* (r, p)-central (with respect to $Z_1, \dots, Z_k$) if the value of the centrality SDP with parameters $Z_1, \dots, Z_k, x, r$ is at most p.

LEMMA 2.8. *Let Z be a d-dimensional random vector with mean $\mu = \mathbb{E}Z$ and covariance Σ. Let $Z_1, \dots, Z_k$ be i.i.d. copies of Z. With probability at least $1 - 2^{-\Omega(k)}$, μ is certifiably $(O(\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|}), 1/100)$-central.*

Since the centrality SDP can be solved in polynomial time, Lemma 2.8 comprises an analysis of the following algorithm for the certification problem: given $Z_1, \dots, Z_k, x$, solve the centrality SDP, and output YES if the optimum value is at most $1/100$ (otherwise output DO NOT KNOW).

In the rest of this section we prove Lemma 2.8. The proof follows a similar strategy to that used by Lugosi and Mendelson to prove Lemma 2.2. We find it surprising that this is possible, given that Lugosi and Mendelson's argument only needs to address the quadratic program (2.1) (almost equivalently, it would only address rank-one solutions to the centrality SDP), while we need to argue about all relaxed solutions.

We will be able to establish, however, that the properties of (2.1) used by (an adaptation of) Lugosi and Mendelson's proof also hold for the centrality SDP. In particular, we will use a bounded-differences property of the centrality SDP to establish concentration. While bounded-differences arguments are standard, using bounded differences to show exponential concentration of the optimum value of a convex program appears to be novel.

2.3. *Proof of Lemma* 2.8. We need to assemble a few tools for the proof of Lemma 2.8. The first concern the 2 → 1 norm of a matrix—in particular, we will be interested in the matrix M with rows $Z_1, \dots, Z_k$.

For our purposes, the 2 → 1 norm of M serves as a moderately outlier-robust modification of the spectral norm (a.k.a. 2 → 2 norm) of the empirical covariance of $Z_1, \dots, Z_k$. This robustness is achieved by replacing an ℓ_2 norm with an ℓ_1 norm. We say "moderately" outlier robust because under our 2nd moment assumption on $Z_1, \dots, Z_k$ we will only be able to establish bounds *in expectation* on the 2 → 1 norm of M, rather than high-probability bounds.

DEFINITION 2.9. Let $A \in \mathbb{R}^{n \times m}$ be a matrix with rows $A_1, \dots, A_n$. The 2-to-1 norm of A is defined as

$$\|A\|_{2\to 1} = \max_{\|u\|=1} \|Au\|_1 = \max_{\|u\|=1, \sigma \in \{\pm 1\}^n} \sum_{i \le n} \sigma_i \langle A_i, u \rangle.$$

Computing the 2 → 1-norm of a matrix A exactly is computationally intractable [10]. Nonetheless, we will profitably use a convex program—again, an SDP – whose optimal values can be related to the 2 → 1 norm. Eventually we will relate the centrality SDP to the following slightly different SDP. It is one of a well-studied family of SDPs for $p \to q$-norm problems, the most famous of which is the $\infty \to 1$-norm SDP appearing in Grothendieck's inequality and used to approximate the cut norm of a matrix [3].

DEFINITION 2.10. For $n, m \in \mathbb{N}$, let $\mathcal{S}_{n,m}^{2\to 1}$ be the following subset of $\mathbb{R}^{(n+m)\times(n+m)}$, treated as the set of block matrices

$$X(S, R, U) = \begin{pmatrix} S & R \\ R^\top & U \end{pmatrix}$$

with $S \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{m\times m}$:

$$\mathcal{S}_{n,m}^{2\to 1} = \{X(S,R,U) : S_{ii} = 1 \text{ for } i = 1,\dots,n, \operatorname{Tr} U \leq 1, \text{ and } X \succeq 0\}.$$

Here we think of S as a relaxation of rank-one matrices $\sigma\sigma^\top$, where $\sigma \in \{\pm 1\}$ is as in Definition 2.9, and U as a relaxation of $uu^\top$ where u is a unit vector as in Definition 2.9.

The following theorem is due to Nesterov. It will allow us to control the optimum value of an SDP relaxation of the $2 \to 1$ norm in terms of the $2 \to 1$ norm itself. It follows fairly easily from the observation that $\|A\|_{2\to 1}^2 = \max_{\sigma\in\{\pm 1\}^n} \sigma^\top A^\top A\sigma$ and the fact (also due to Nesterov) that semidefinite programming yields a $\frac{2}{\pi}$-approximation algorithm for the maximization of a positive semidefinite quadratic form over $\{\pm 1\}^n$ (see, e.g., [56], Section 6.3 for a simple proof).

THEOREM 2.11 ([46]). *There is a constant $K_{2\to 1} = \sqrt{\pi/2} < 2$ such that for every $n \times m$ matrix A, one has the following inequality*:

$$\max_{X(S,R,U)\in\mathcal{S}_{n,m}^{2\to 1}} \langle R, A\rangle \leq K_{2\to 1}\|A\|_{2\to 1}.$$

The following lemma affords control over $\mathbb{E}\|M\|_{2\to 1}$, where M has rows $Z_1,\dots,Z_k$. The proof uses standard tools from empirical process theory; a similar argument appears in [38]. We provide the proof in Section 1 of the Supplementary Material [24].

LEMMA 2.12. *Let Z be an $\mathbb{R}^d$-valued random variable with mean $\mathbb{E}Z = 0$ and covariance $\mathbb{E}ZZ^\top = \Sigma$. Let $Z_1,\dots,Z_k$ be i.i.d. copies of Z, and let $M \in \mathbb{R}^{k\times d}$ be the matrix whose rows are $Z_1,\dots,Z_k$. Then*

$$\mathbb{E}\|M\|_{2\to 1} \leq 2\sqrt{k \operatorname{Tr}\Sigma} + k\sqrt{\|\Sigma\|},$$

where $\|\Sigma\|$ denotes the operator norm, or maximum eigenvalue, of Σ.

Finally, the last lemma on the way to Lemma 2.8 shows that the centrality SDP satisfies a bounded differences property: this is crucial to establishing the high-probability bound in Lemma 2.8. The proof is in [24], Section 1.

LEMMA 2.13. *Let $r \geq 0$ and $x \in \mathbb{R}^d$. Let $Z_1,\dots,Z_k \in \mathbb{R}^d$, $i \in [k]$ and $Z_i' \in \mathbb{R}^d$. Let $\operatorname{SDP}(Z_1,\dots,Z_k,x,r)$ be the optimum value of the centrality SDP with parameters $Z_1,\dots,Z_k,x,r$. Then*

$$\left|\operatorname{SDP}(Z_1,\dots,Z_k,x,r) - \operatorname{SDP}(Z_1,\dots,Z_{i-1},Z_i',Z_{i+1},\dots,Z_k,x,r)\right| \leq \frac{1}{k}.$$

Now we are ready to prove Lemma 2.8.

PROOF OF LEMMA 2.8. The proof has an expectation step and a concentration step. Let $\operatorname{SDP}(Z_1,\dots,Z_k,\mu,r)$ be the optimum value of the centrality SDP. Since $Z_1,\dots,Z_k$ are independent, by the bounded differences inequality together with Lemma 2.13,

$$\mathbb{P}\big(\operatorname{SDP}(Z_1,\dots,Z_k,\mu,r) - \mathbb{E}\operatorname{SDP}(Z_1,\dots,Z_k,\mu,r) > 1/200\big) < 2^{-\Omega(k)}.$$

Thus, it will suffice to show that $\mathbb{E}\operatorname{SDP}(Z_1,\dots,Z_k,\mu,r) \leq 1/200$ for some $r = O(\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|})$.

By definition of the centrality SDP, using the constraints $\langle Z_i - x, W_i\rangle \geq r \cdot b_i$, we have

$$\mathbb{E} \max_{B,W,U,b,u} \frac{1}{k}\sum_{i\leq k} b_i \leq \frac{1}{kr}\mathbb{E} \max_{B,W,U,b,u} \sum_{i\leq k}\langle W_i, Z_i - \mu\rangle.$$

Let $\mathcal{S}'$ be the set $\mathcal{S}_{k,d}^{2\to 1}$ with the modified constraint $S_{ii} \leq 1$ rather than $S_{ii} = 1$. Then we have $\mathcal{S}' \supseteq \{Y(B, W, U)\}$ where the latter is the set of feasible solutions to the centrality SDP (restricted to the large $(d+k)\times(d+k)$ block), and hence

$$\frac{1}{kr}\mathbb{E} \max_{B,W,U,b,u} \sum_{i\leq k}\langle W_i, Z_i - \mu\rangle \leq \frac{1}{kr}\mathbb{E} \max_{X(S,R,U)\in\mathcal{S}'} \sum_{i\leq k}\langle R_i, Z_i - \mu\rangle,$$

where R has rows $R_1, \dots, R_k$, since the left-hand side maximizes over a larger set of PSD matrices.

We would like to replace $\mathcal{S}'$ with $\mathcal{S}_{k,d}^{2\to 1}$. For this we need to argue that the constraints $S_{ii} = 1$ are satisfied by the optimal $X(S, R, U)$. First of all, note that the maximum on the right-hand side is obtained at $X(S, R, U)$ where $\langle R_i, Z_i - \mu\rangle \geq 0$, otherwise we may replace X with $\frac{1}{2}X + \frac{1}{2}(-E_{ii})X(-E_{ii})$ and remain inside $\mathcal{S}'$ while only increasing $\langle R_i, Z_i - \mu\rangle$—here E_{ii} is the matrix with exactly one nonzero entry, at the (i, i)th position, with value 1.

Hence also the maximum is obtained at $X(S, R, U)$ with $S_{ii} = 1$, otherwise we may rescale the ith row and column by $1/\sqrt{S_{ii}}$ and remain in $\mathcal{S}'$ while only increasing $\langle R_i, Z_i - \mu\rangle$ (here we used that $\langle R_i, Z_i - \mu\rangle \geq 0$, so $\langle R_i, Z_i - \mu\rangle/\sqrt{S_{ii}} \geq \langle R_i, Z_i - \mu\rangle$). Ultimately, we can conclude that

$$\frac{1}{kr}\mathbb{E} \max_{X(S,R,U)\in\mathcal{S}'} \sum_{i\leq k}\langle R_i, Z_i - \mu\rangle = \frac{1}{kr}\mathbb{E} \max_{X(S,R,U)\in\mathcal{S}_{k,d}^{2\to 1}} \sum_{i\leq k}\langle R_i, Z_i - \mu\rangle.$$

The right-hand side is exactly the 2 → 1-norm SDP relaxation from Definition 2.10. So if M is the matrix with rows $Z_i - \mu$, we get

$$\begin{aligned}\mathbb{E} SDP(Z_1, \dots, Z_k, \mu, r) &\leq \frac{K_{2\to 1}}{kr} \cdot \mathbb{E}\|M\|_{2\to 1}\\ &\leq \frac{K_{2\to 1}}{r} \cdot \bigl(2\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|}\bigr),\end{aligned}$$

where we have used Theorem 2.11 and $K_{2\to 1}$ is the constant from that theorem. By choosing $r = 1000(\sqrt{\operatorname{Tr}\Sigma/k} + \sqrt{\|\Sigma\|})$ the lemma follows. □

3. SoS preliminaries. Now that we have established certifiable centrality of the mean, we can turn back to our main goal: design an algorithm to estimate the mean μ in order to prove Theorem 1.2. While in Section 2 we employed a traditional style of semidefinite program (arising as a relaxation of a quadratic program), to prove Theorem 1.2 we will need a larger semidefinite program (i.e., having more variables and constraints). The sum of squares method offers a principled way to exploit the addition of extra variables and constraints to semidefinite programs.

Treating SoS-style semidefinite programs with the traditional language and notation of semidefinite programming is often cumbersome. Recent work in theoretical computer science has pioneered an alternative point of view, involving *pseudoexpectations*, which correspond to SDP primal solutions, and *SoS proofs*, which correspond to SDP dual solutions. Analyzing a complex semidefinite program can often be reduced to the construction of an appropriate dual solution. The pseudoexpectation/SoS proof point of view is designed to make this construction possible in a modular fashion, building a complicated dual solutions out of many simpler ones.

In this section we get set up to use the SoS approach for our main algorithm. We review the preliminaries we need and refer the reader to other resources for a full exposition—see, for example, [7].

DEFINITION 3.1 (SoS Polynomials). Let $x = x_1, \dots x_n$ be some indeterminates, and let $p \in \mathbb{R}[x]$. We say that p is SoS if it is expressible as $p = \sum_{i=1}^m q_i(x)^2$ for some other polynomials q_i. We write $p \succeq 0$, and if $p - q \succeq 0$ we write $p \succeq q$.

DEFINITION 3.2 (SoS Proof). Let $\mathcal{A} = \{p_1(x) \geq 0, \dots, p_m(x) \geq 0\}$ be a set of polynomial inequalities. We sometimes include polynomial equations $p_i(x) = 0$, by which we mean that $\mathcal{A}$ contains both $p_i(x) \geq 0$ and $-p_i(x) \geq 0$. We say that $\mathcal{A}$ SoS-proves that $q(x) \geq 0$ if there are SoS polynomials $q_S(x)$ for every $S \subseteq [m]$ such that

$$q(x) = \sum_{S \subseteq [m]} q_S(x) \prod_{i \in S} p_i(x).$$

The polynomials $q_S(x)$ form an *SoS proof* that $q(x) \geq 0$ for every x such that $p_i(x) \geq 0$. If $\deg q_S(x) \cdot \prod_{i \in S} p_i(x) \leq d$ for every S, then we say that the proof has *degree d*, and write

$$\mathcal{A} \vdash_d q(x) \geq 0.$$

SoS proofs obey many natural inference rules, which we will freely use in this paper—see, for example, [7].

Critically, the set of SoS proofs of $q(x) \geq 0$ using axioms $\mathcal{A}$ form a convex set (in fact, a semidefinite program). Their convex duals are called *pseudodistributions* or *pseudoexpectations* (we use the terms interchangeably).

DEFINITION 3.3 (Pseudoexpectation). A degree-d *pseudoexpectation* in variables $x = x_1, \dots, x_n$ is a linear operator $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$, where $\mathbb{R}[x]_{\leq d}$ are the polynomials in x with real coefficients and degree at most d. A pseudoexpectation is:

1. Normalized: $\tilde{\mathbb{E}} 1 = 1$, where $1 \in \mathbb{R}[x]_{\leq d}$ on the left side is the constant polynomial.
2. Nonnegative: $\tilde{\mathbb{E}} p(x)^2 \geq 0$ for every p of degree at most $d/2$.

DEFINITION 3.4 (Satisfying constraints). A pseudoexpectation of degree d *satisfies* a polynomial equation $p(x) = 0$ if for every $q(x)$ such that $p(x)q(x)$ has degree at most d it holds that $\tilde{\mathbb{E}} p(x)q(x) = 0$. The pseudodistribution satisfies an inequality $p(x) \geq 0$ if for every $q(x)^2$ such that $\deg q(x)^2 p(x) \leq d$ it holds that $\tilde{\mathbb{E}} p(x)q(x)^2 \geq 0$.

EXAMPLE 3.5. To demystify pseudoexpectations slightly, consider the classic semidefinite relaxation of the set $\{\pm 1\}^n$ to the set $\{X \in \mathbb{R}^{n \times n} : X \succeq 0, X_{ii} = 1\}$. (This is exactly the set of PSD matrices employed in the SDP-based MAX-CUT algorithm of Goemans and Williamson [22].)

Each such X defines a degree-2 pseudoexpectation, by setting $\tilde{\mathbb{E}} x_i x_j = X_{ij}$ for $1 \leq i \leq n$, $\tilde{\mathbb{E}} x_i = 0$, and finally $\tilde{\mathbb{E}} 1 = 1$. Since $X \succeq 0$, it also follows that for every polynomial $p \in \mathbb{R}[x_1, \dots, x_n]_{\leq 2}$, one has $\tilde{\mathbb{E}} p(x)^2 = p_1^\top X p_1 + \hat{p}(\varnothing)^2 \geq 0$, where p_1 is the vector of coefficients of the homogeneous linear part of p and $\hat{p}(\varnothing)$ is the constant term in p. Last, since $\tilde{\mathbb{E}} x_i^2 = X_{ii} = 1$, the pseudoexpectation satisfies $x_i^2 - 1 = 0$ for each i; these equations exactly characterize $\{\pm 1\}^n$ as a variety in $\mathbb{R}^n$.[8]

[8] In this case, $\tilde{\mathbb{E}}$ is defined by a few more parameters than X—namely the values $\tilde{\mathbb{E}} x_i$, which we set to zero. For most algorithms involving degree-2 pseudoexpectations the main focus is on the n^2 variables $\tilde{\mathbb{E}} x_i x_j$, so this is not too surprising. However, as we will see in the algorithm in Section 2 of the Supplementary Material [24], pseudoexpectations of degree higher than 2 can contain useful information about polynomials of various degrees.

As in this simple example, it is always possible to write an explicit semidefinite program whose solutions are pseudoexpectations satisfying some chosen set of polynomial inequalities. However, as the degrees and complexity of the of polynomials grow, these SDPs become notationally unwieldy. In this regard, the pseudoexpectation approach carries significant advantages.

The most elementary fact relating pseudodistributions and SoS proofs is the following:

FACT 3.6. Suppose $\mathcal{A} \vdash_d p(x) \geq 0$. Then any degree-$d$ pseudodistribution $\tilde{\mathbb{E}}$ which satisfies $\mathcal{A}$ also has $\tilde{\mathbb{E}} p(x) \geq 0$.

We will make use of the following theorem, which can be proved via semidefinite programming.

THEOREM 3.7 (Adapted from [7]). *For every $d \in \mathbb{N}$ there exists an $(mn)^{O(d)}$-time algorithm which given a set of m n-variate polynomial inequalities $\mathcal{A}$ which*:

- *has coefficients with bit complexity at most $(mn)^{O(d)}$,*
- *contains a constraint of the form $\|x\|^2 \leq M$ for a positive constant M and*
- *is satisfied by some $x \in \mathbb{R}^n$*

finds a degree-d pseudodistribution which satisfies $\mathcal{A}$ up to an additive error of $2^{-(mn)^d}$ in each inequality.

In general, the additive $2^{-(mn)^d}$ errors will not bother us, because the magnitudes of coefficients in the SoS proofs we construct will be bounded by $\text{poly}(n, m)$. See [7, 50] for more discussion of such numerical considerations.

We will use the following simple fact about pseudodistributions.

FACT 3.8. Let $\tilde{\mathbb{E}}$ be a pseudodistribution of degree 2 in variables $x_1, \dots, x_n$ and let $\mu \in \mathbb{R}^n$. Then $\|\tilde{\mathbb{E}} x - \mu\|^2 \leq \tilde{\mathbb{E}} \|x - \mu\|^2$.

PROOF. Follows from $\tilde{\mathbb{E}}(x_i - \mu_i)^2 \geq (\tilde{\mathbb{E}} x_i - \mu_i)^2$ for every $i \leq n$, which follows from the more general fact $\tilde{\mathbb{E}} p(x)^2 \geq (\tilde{\mathbb{E}} p(x))^2$ for every degree-1 polynomial p. The latter follows by $\tilde{\mathbb{E}}(p(x) - \tilde{\mathbb{E}} p(x))^2 \geq 0$. □

4. Main algorithm and analysis.

Our main lemma for this section gives an algorithm which recovers a central point given vectors $Z_1, \dots, Z_k$, provided that a *certifiably* central point exists (and some minor additional regularity conditions on $Z_1, \dots, Z_k$ are met).

LEMMA 4.1. *For every $d, k \in \mathbb{N}$ and $C, r > 0$ there is an algorithm* MEDIAN-SDP *which runs in time $(dk \log C)^{O(1)}$ and has the following guarantees. Let $Z_1, \dots, Z_k \in \mathbb{R}^d$. Suppose that $\mu \in \mathbb{R}^d$ is certifiably $(r, 1/100)$-central with respect to $Z_1, \dots, Z_k$. And, suppose that at most $k/100$ of the vectors $Z_1, \dots, Z_k$ have $\|Z_i - \mu\| > Cr$. Then given $Z_1, \dots, Z_k$,* MEDIAN-SDP *returns a point $\hat{\mu}$ with $\|\mu - \hat{\mu}\| = O(r)$.*

Together Lemmas 2.8 and 4.1 suffice to prove Theorem 1.2, with the small modification that the algorithm is given access to r, C in addition to the samples $X_1, \dots, X_n$. We discuss in Section 5 of the Supplementary Material [24] how to use standard ideas to avoid this dependence.

PROOF OF THEOREM 1.2. Let $k = c\log(1/\delta)$ for a big-enough constant c. Given samples $X_1, \dots, X_n$, for $i \leq k$ let Z_i be the average of samples $X_{i\cdot(n/k)}, \dots, X_{(i+1)\cdot n/k-1}$ (throwing out samples as necessary so that n is divisible by k). Then $Z_1, \dots, Z_k$ are i.i.d. copies of a random variable Z with $\mathbb{E}Z = \mu$ and $\mathbb{E}(Z-\mu)(Z-\mu)^\top = \frac{k}{n}\Sigma$. By Lemma 2.8, μ is certifiably $(r, 1/100)$-central with respect to $Z_1, \dots, Z_k$ for $r = O(\sqrt{\operatorname{Tr}\Sigma/n} + \sqrt{\|\Sigma\|k/n})$ with probability at least $1 - \exp(-\Omega(k))$. We can choose c so that this probability is at least $1-\delta$ and $\sqrt{\|\Sigma\|k/n} = O(\sqrt{\|\Sigma\|\log(1/\delta)/n})$.

Furthermore, by Chebyshev's inequality and a binomial tail bound, with probability at least $1 - \exp(-\Omega(k))$ we have that $\|Z_i - \mu\| \leq O(\sqrt{\operatorname{Tr} k\Sigma/n}) \leq O(kr)$ for all but $k/100$ vectors Z_i. Hence, except with probability $2^{-\Omega(k)}$, calling MEDIAN-SDP with $C = O(k)$ yields a vector x with $\|\mu - x\| \leq O(r)$. □

In the remainder of this section we prove Lemma 4.1 from technical lemmas which are proved in the Supplementary Material [24]. We will make use of the SoS method, which will require some setup and technical arguments, so we describe the main idea first. Given $Z_1, \dots, Z_k$, we will define a system of polynomial equations $\mathcal{A}$ whose feasible solutions are the certifiably $(r, 1/10)$-central points. (For technical convenience actually $\mathcal{A}$ has feasible solutions which are the certifiably $(r, 1/10)$-central points satisfying an additional mild regularity condition, as we discuss below.) Our main algorithm will find a pseudodistribution which satisfies $\mathcal{A}$ and extract from it an estimator $\hat{\mu} \in \mathbb{R}^d$.

To argue about $\|\hat{\mu} - \mu\|$, we will construct SoS proofs (using $\mathcal{A}$ as axioms) of several inequalilties concerning certifiable $(r, 1/10)$-central points. Together these inequalities will capture the fact that any two $(r, 1/10)$-central points x, y have $\|x - y\| \leq 2r$; we will use the SoS proofs of these inequalities as duals to the set of pseudodistributions satisfying $\mathcal{A}$, ultimately showing that $\|\hat{\mu} - \mu\| = O(r)$.

Before we can construct $\mathcal{A}$, we need to observe a consequence of SDP duality—certifiable centrality of μ implies the existence of a witness to its centrality. (Here it may help to recall the set CERTIFIABLE-CENTERS from Section 1.) Our construction of $\mathcal{A}$ will exploit these witnesses.

LEMMA 4.2. *Let $Z_1, \dots, Z_k, x \in \mathbb{R}^d$ and suppose x is certifiably (r, p)-central with respect to $Z_1, \dots, Z_k$. Then there are nonnegative numbers $\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k, \gamma$ and a degree-2 SoS polynomial $\sigma \in \mathbb{R}[b_1, \dots, b_k, v_1, \dots, v_d]_{\leq 2}$ such that the following polynomial identity holds in variables $b_1, \dots, b_k, v_1, \dots, v_d$:*

$$pk - \sum_{i=1}^k b_i = \sum_{i=1}^k \alpha_i b_i(\langle Z_i - x, v\rangle - r) + \sum_{i=1}^k \beta_i(1 - b_i^2) + \gamma(1 - \|v\|^2) + \sigma(b, v). \tag{4.1}$$

The proof is a direct application of SDP duality—see, for example, [11]. (The polynomial identity is obtained by evaluating the quadratic form of an optimal dual solution to the centrality SDP at the vector of indeterminates $(1, b, u)$.) The numbers α, β, γ and SoS polynomial σ are an SoS proof that x is (r, p)-central: they witness

$$\bigcup_{i \leq k}\{b_i^2 \leq 1, \|v\|^2 \leq 1, b_i\langle Z_i - x, v\rangle - b_i r \geq 0\} \vdash_2 \sum_{i=1}^k b_i \leq pk.$$

Indeed one may check that if v is any unit vector and b is the $0/1$ indicator for those $i \in [k]$ such that $\langle Z_i - x, v\rangle \geq r$, then the right-hand side of equation (4.1) is nonnegative when evaluated at b, v. Hence the left-hand side must be as well, which means that $\sum_{i\in k} b_i \leq pk$.

The last step before constructing the polynomial system $\mathcal{A}$ is to observe a consequence of the regularity condition from Lemma 4.1 that $\|Z_i - \mu\| \leq Cr$ for at least $99k/100$ Z_i's. Namely, it affords some control over the magnitudes of the numbers $\alpha_1, \ldots, \alpha_k, \gamma$ from Lemma 4.2, ensuring that the witness $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k, \gamma$ has a certain well-conditioned-ness property. We will capture the well-conditioned-ness property in $\mathcal{A}$ and make use of it in our SoS proofs. The proof of the following lemma involves elementary manipulations on equations like equation (4.1); we defer it to the Supplementary Material [24].

LEMMA 4.3. *Let $Z_1, \ldots, Z_k, x \in \mathbb{R}^d$ and suppose x is (r, p)-central with respect to $Z_1, \ldots, Z_k$. Suppose also that $\|Z_i - x\| \leq Cr$ for all but qk vectors Z_i, where $C \geq 1$. Then there are nonnegative numbers $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k, \gamma$ and a degree-2 SoS polynomial $\sigma \in \mathbb{R}[b_1, \ldots, b_k, v_1, \ldots, v_d]_{\leq 2}$ such that the following polynomial identity holds in variables $b_1, \ldots, b_k, v_1, \ldots, v_d$:*

$$(4.2)\qquad \begin{aligned}(p+q+1/20)k - \sum_{i=1}^{k} b_i = \sum_{i=1}^{k} \alpha_i b_i\big(\langle Z_i - x, v\rangle - r\big) + \sum_{i=1}^{k} \beta_i\big(1 - b_i^2\big) \\ + \gamma\big(1 - \|v\|^2\big) + \sigma(b, v).\end{aligned}$$

Furthermore, γ is in the finite set $\{0, 1/100, 2/100, \ldots, k\}$, and $\alpha_1, \ldots, \alpha_k$ are in the set $\{0\} \cup [1/100Cr, 4k/r]$.

Now we are able to construct our main polynomial system $\mathcal{A}$, whose solutions correspond to $x, \alpha, \beta, \gamma, \sigma$ such that $\alpha, \beta, \gamma, \sigma$ form a witness that x is a certifiably $(r, 1/10)$-central. For technical convenience, we take γ to be a *parameter* of this system rather than one of its indeterminates. Part of our algorithm will involve a brute-force search for a good choice of γ—by Lemma 4.3 there will only be $O(k)$ possibilities to search over.

DEFINITION 4.4 (The polynomial system $\mathcal{A}(Z_1, \ldots, Z_k, r, C, c, \gamma)$). For vectors $Z_1, \ldots, Z_k \in \mathbb{R}^d$, $r > 0$, and $c, C > 0$ we define a system of equations in the following variables:

$$\begin{gathered}\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k, \sigma_{ij} \quad \text{for } i, j \in [d+k+1], \\ x_1, \ldots, x_d, \quad \text{and} \quad a_{i,t} \quad \text{for } i \in [k] \text{ and } t \in [\log C/c + 1].\end{gathered}$$

Let $\mathcal{A}_{\text{sos}}$ be the set of linear equations among $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k, \sigma_{ij}, x$ which ensure that the polynomial identity

$$\begin{aligned}\frac{k}{10} - \sum_{i=1}^{k} b_i = \sum_{i \in S} \alpha_i b_i\big(\langle Z_i - \mu, v\rangle - r\big) + \sum_{i=1}^{k} \beta_i\big(1 - b_i^2\big) \\ + \gamma\big(1 - \|v\|^2\big) + \sum_{i \in [d+k+1]} \big\langle \sigma_i, (1, b, v)\big\rangle^2\end{aligned}$$

holds in variables $b_1, \ldots, b_k, v_1, \ldots, v_d$, where σ_i is the vector with jth entry σ_{ij} and $(1, b, v)$ is the $(d+k+1)$-dimensional concatenation $1, b_1, \ldots, b_k, v_1, \ldots, v_d$. We often abuse notation and write $\sigma(b, v)$ for the expression $\sum_{i \in [d+k+1]} \langle \sigma_i, (1, b, v)\rangle^2$. Let $\mathcal{A}_{\text{nonneg}}$ be the inequalities

$$\alpha_i \geq 0 \quad \text{for } i \in [k] \quad \text{and} \quad \beta_i \geq 0 \quad \text{for } i \in [k].$$

Let $\mathcal{A}_{\text{a}}$ be the equations and inequalities

$$\begin{gathered}a_{i,t}^2 = a_{i,t} \quad \text{for } t \in [\log C/c + 1], \\ a_{i,t} \cdot 2^{t-1} \cdot c \leq a_{i,t} \cdot \alpha_i \quad \text{for } t \in [1, \log C/c + 1],\end{gathered}$$

Algorithm 1 MEDIAN-SDP

```
Given: $Z_1, \dots, Z_k \in \mathbb{R}^d, r, C > 0$
1. For each $\gamma \in \{0, 1/100, 2/100, \dots, k\}$, try to find a degree-8 pseudodistribution satisfying $\mathcal{A}(Z_1, \dots, Z_k, r, 1/100Cr, 4k/r, \gamma)$. If none exists for any $\gamma$, output REJECT. Otherwise, let $\tilde{\mathbb{E}}$ be the pseudodistribution obtained for any $\gamma$ for which one exists.
2. Output $\tilde{\mathbb{E}}x$.
```

$$\begin{aligned} a_{i,t} \cdot \alpha_i &\le a_{i,t} \cdot 2^t \cdot c \quad \text{for } t \in [1, \log C/c + 1], \\ a_{i,0} \cdot \alpha_i &= 0, \\ \sum_{t \le \log C/c+1} a_{i,t} &= 1 \quad \text{for all } i \le k, \\ a_{i,t} a_{i,t'} &= 0 \quad \text{for all } i \le k \text{ and } t \ne t'. \end{aligned}$$

The inequalities $\mathcal{A}_a$ ensure that $a_{i,t} \in \{0, 1\}$ and $a_{i,t} = 1$ if and only if $\alpha \in [2^{t-1}c, 2^t c]$ (or $\alpha_i = 0$ in the case of $a_{i,0}$). We will use the variables $a_{i,t}$ to approximate some functions of α_i which are not polynomials. For instance, if α, a satisfy $\mathcal{A}_a$ and $\alpha_i > 0$ then $\sum_{1 \le t \le \log C/c+1} a_{i,t}/(c \cdot 2^t) \in [1/2\alpha_i, 1/\alpha_i]$.

Finally, let $\mathcal{A} = \mathcal{A}_{\text{sos}} \cup \mathcal{A}_{\text{nonneg}} \cup \mathcal{A}_{\text{a}}$.

Now we can describe the algorithm MEDIAN-SDP (Algorithm 1) and its main analysis.

LEMMA 4.5 (Main lemma for MEDIAN-SDP). *Let $Z_1, \dots, Z_k \in \mathbb{R}^d$. Let μ be certifiably $(r, 1/10)$-central. Then for every c, C, γ, any degree-8 pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\mathcal{A}$ has $\tilde{\mathbb{E}}\|x - \mu\|^2 = O(r^2)$.*

We will prove Lemma 4.5 in [24], Section 2. We wrap up this section by proving Lemma 4.1 from Lemmas 4.2, 4.3 and 4.5.

PROOF OF LEMMA 4.1. Since at most $k/100$ of of $Z_1, \dots, Z_k$ have $\|Z_i - \mu\| > Cr$, and because μ is $(r, 1/100)$-certifiable, together Lemmas 4.2 and 4.3 show that there exist nonnegative $\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k, \gamma$ and a degree-2 SoS polynomial σ such that

$$\begin{aligned} 0.07k - \sum_{i \le k} b_i = \sum_{i=1}^k \alpha_i b_i \big(\langle Z_i - x, v\rangle - r\big) + \sum_{i=1}^k \beta_i \big(1 - b_i^2\big) \\ + \gamma\big(1 - \|v\|^2\big) + \sigma(b, v) \end{aligned}$$

holds as a polynomial identity in b, v. Furthermore, $\alpha_i \in \{0\} \cup [1/100Cr, 4k/r]$ and $\gamma \in \{1/100, 2/100, \dots, k\}$. So, $\mathcal{A}(Z_1, \dots, Z_k, r, 1/100Cr, 4k/r, \gamma)$ is feasible. Thus, MEDIAN-SDP with parameters r, C eventually finds a pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\mathcal{A}$ for some γ'. So by Lemma 4.5 we have $\tilde{\mathbb{E}}\|x - \mu\|^2 \le O(r^2)$. Then the main conclusion of Lemma 4.1 follows by Fact 3.8.

The running time bound follows by observation that $\mathcal{A}$ has $(dk \log C)^{O(1)}$ variables and inequalities with this choice of parameters, then application of Theorem 3.7. □

5. Conclusion. We have described the first polynomial-time algorithm capable of estimating the mean of a distribution with confidence intervals asymptotically matching those of the empirical mean in the Gaussian setting, under only the assumption that the distribution has finite mean and covariance. Previous estimators with matching rates under such weak assumptions required exponential computation time. Our algorithm uses semidefinite programming, and in particular the SoS method. The SDP we employ is sufficiently powerful that Lugosi and Mendelson's analysis of their tournament-based estimator can be transformed to an analysis of the SoS SDP.

Our algorithm runs in polynomial time, but it is not close to practical for any substantially high-dimensional data set. Work building on the present paper has already reduced the running time to $O(n^{3.5}+n^2 d)\cdot(\log nd)^{O(1)}$ [15]. It remains an interesting direction for future study whether there is a *practical* algorithm whose empirical performance improves on that of fast, practical algorithms (like geometric median) which achieve a $\sqrt{\operatorname{Tr}\Sigma\log(1/\delta)/n}$-style confidence interval.

Acknowledgments. Thanks to Siva Balakrishnan and Stas Minsker for bringing the mean estimation problem to my attention. I am most grateful to Peter Bartlett, Tarun Kathuria, Pravesh Kothari, Jerry Li, Gabor Lugosi, Prasad Raghavendra and Jacob Steinhardt for helpful conversations as this manuscript was being prepared, and to anonymous reviewers for many suggestions in improving its presentation and correcting errors. Finally, an earlier version of this manuscript contained a serious technical error: I am greatly indebted to Yeshwanth Cherapanamjeri for pointing it out to me (and explaining it to me several times), and to Prasad Raghavendra for several suggestions in correcting it.

This work was supported by a Miller Fellowship at UC Berkeley. This project was also supported by NSF Award No. 1408673.

SUPPLEMENTARY MATERIAL

Supplement to "Mean estimation with sub-Gaussian rates in polynomial time" (DOI: 10.1214/19-AOS1843SUPP; .pdf). We provide deferred proofs of technical results.

REFERENCES

[1] ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

[2] ALON, N., MATIAS, Y. and SZEGEDY, M. (1999). The space complexity of approximating the frequency moments. *J. Comput. System Sci.* **58** 137–147. MR1688610 https://doi.org/10.1006/jcss.1997.1545

[3] ALON, N. and NAOR, A. (2006). Approximating the cut-norm via Grothendieck's inequality. *SIAM J. Comput.* **35** 787–803. MR2203567 https://doi.org/10.1137/S0097539704441629

[4] AMINI, A. A. and WAINWRIGHT, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *ISIT* 2454–2458. IEEE.

[5] BANKS, J., KLEINBERG, R. and MOORE, C. (2017). The Lovász theta function for random regular graphs and community detection in the hard regime. In *Approximation*, *Randomization*, *and Combinatorial Optimization*. *Algorithms and Techniques*. *LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **81** Art. No. 28, 22. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3695595

[6] BARAK, B. and MOITRA, A. (2016). Noisy tensor completion via the sum-of-squares hierarchy, COLT. *J. Mach. Learn. Res. Workshop Conf. Proc.* **49** 417–445.

[7] BARAK, B. and STEURER, D. (2017). The sos algorithm over general domains. In *Lecture Notes*: *Proofs*, *Beliefs and Algorithms Through the Lens of Sum of Squares*. https://www.sumofsquares.org/public/lec-definitions-general.html.

[8] BERNHOLT, T. (2006). Robust estimators are hard to compute. Technical Report, Univ. Dortmund.

[9] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection, COLT. *J. Mach. Learn. Res. Workshop Conf. Proc.* **30** 1046–1066.

[10] BHATTIPROLU, V., GHOSH, M., GURUSWAMI, V., LEE, E. and TULSIANI, M. (2018). Inapproximability of matrix $p\to q$ norms. In *Electronic Colloquium on Computational Complexity* (*ECCC*) **25** 37.

[11] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

[12] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

[13] CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

[14] CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

[15] CHERAPANAMJERI, Y., FLAMMARION, N. and BARTLETT, P. L. (2019). Fast mean estimation with sub-Gaussian rates. Available at arXiv:1902.01998.

[16] COHEN, M. B., LEE, Y. T., MILLER, G., PACHOCKI, J. and SIDFORD, A. (2016). Geometric median in nearly linear time. In *STOC'*16—*Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 9–21. ACM, New York. MR3536551

[17] D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49** 434–448. MR2353806 https://doi.org/10.1137/050645506

[18] DECELLE, A., KRZAKALA, F., MOORE, C. and ZDEBOROVÁ, L. (2011). Inference and phase transitions in the detection of modules in sparse networks. *Phys. Rev. Lett.* **107** 065701. https://doi.org/10.1103/PhysRevLett.107.065701

[19] DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440

[20] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2016). Robust estimators in high dimensions without the computational intractability. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 655–664. IEEE Computer Soc., Los Alamitos, CA. MR3631028

[21] FALOUTSOS, M., FALOUTSOS, P. and FALOUTSOS, C. (1999). On power-law relationships of the Internet topology. In *ACM SIGCOMM Computer Communication Review* **29** 251–262. ACM, New York.

[22] GOEMANS, M. X. and WILLIAMSON, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.* **42** 1115–1145. MR1412228 https://doi.org/10.1145/227683.227684

[23] GUÉDON, O. and VERSHYNIN, R. (2016). Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields* **165** 1025–1049. MR3520025 https://doi.org/10.1007/s00440-015-0659-z

[24] HOPKINS, S. B. (2020). Supplement to "Mean estimation with sub-Gaussian rates in polynomial time." https://doi.org/10.1214/19-AOS1843SUPP.

[25] HOPKINS, S. B., KOTHARI, P. K., POTECHIN, A., RAGHAVENDRA, P., SCHRAMM, T. and STEURER, D. (2017). The power of sum-of-squares for detecting hidden structures. In 58*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2017 720–731. IEEE Computer Soc., Los Alamitos, CA. MR3734275

[26] HOPKINS, S. B. and LI, J. (2018). Mixture models, robustness, and sum of squares proofs. In *STOC'*18—*Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 1021–1034. ACM, New York. MR3826314

[27] HOPKINS, S. B. and STEURER, D. (2017). Efficient Bayesian estimation from few samples: Community detection and related problems. In 58*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2017 379–390. IEEE Computer Soc., Los Alamitos, CA. MR3734245

[28] HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40. MR3491112

[29] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

[30] JERRUM, M. R., VALIANT, L. G. and VAZIRANI, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** 169–188. MR0855970 https://doi.org/10.1016/0304-3975(86)90174-X

[31] KLIVANS, A., KOTHARI, P. K. and MEKA, R. (2018). Efficient algorithms for outlier-robust regression. arXiv preprint, arXiv:1803.03241.

[32] KOTHARI, P. K., STEINHARDT, J. and STEURER, D. (2018). Robust moment estimation and improved clustering via sum of squares. In *STOC'*18—*Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 1035–1046. ACM, New York. MR3826315

[33] KRAUTHGAMER, R., NADLER, B. and VILENCHIK, D. (2015). Do semidefinite relaxations solve sparse PCA up to the information limit? *Ann. Statist.* **43** 1300–1322. MR3346704 https://doi.org/10.1214/15-AOS1310

[34] LAI, K. A., RAO, A. B. and VEMPALA, S. (2016). Agnostic estimation of mean and covariance. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 665–674. IEEE Computer Soc., Los Alamitos, CA. MR3631029

[35] LASSERRE, J. B. (2000/01). Global optimization with polynomials and the problem of moments. *SIAM J. Optim.* **11** 796–817. MR1814045 https://doi.org/10.1137/S1052623400366802

[36] LERASLE, M. and OLIVEIRA, R. I. (2011). Robust empirical mean estimators. arXiv preprint, arXiv:1112.3914.

[37] LESKOVEC, J., KLEINBERG, J. and FALOUTSOS, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* 177–187. ACM New York.

[38] LUGOSI, G. and MENDELSON, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794. MR3909950 https://doi.org/10.1214/17-AOS1639

[39] MA, T., SHI, J. and STEURER, D. (2016). Polynomial-time tensor decompositions with sum-of-squares. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 438–446. IEEE Computer Soc., Los Alamitos, CA. MR3631006

[40] MA, T. and WIGDERSON, A. (2015). Sum-of-squares lower bounds for sparse PCA. *NIPS* 1612–1620.

[41] MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468 https://doi.org/10.3150/14-BEJ645

[42] MINSKER, S. (2018). Uniform bounds for robust mean estimators. arXiv preprint, arXiv:1812.03523.

[43] MONTANARI, A. and SEN, S. (2016). Semidefinite programs on sparse random graphs and their application to community detection. In *STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing* 814–827. ACM, New York. MR3536616

[44] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience Publication*. Wiley, New York. Translated from the Russian and with a preface by E. R. Dawson. MR0702836

[45] NESTEROV, Y. (2000). Squared functional systems and optimization problems. In *High Performance Optimization. Appl. Optim.* **33** 405–440. Kluwer Academic, Dordrecht. MR1748764 https://doi.org/10.1007/978-1-4757-3216-0_17

[46] NESTEROV, YU. (1998). Semidefinite relaxation and nonconvex quadratic optimization. *Optim. Methods Softw.* **9** 141–160. MR1618100 https://doi.org/10.1080/10556789808805690

[47] PARRILO, P. A. (2000). Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

[48] POTECHIN, A. and STEURER, D. (2017). Exact tensor completion with sum-of-squares. *Proceedings of Machine Learning Research* **65** 1–54.

[49] RAGHAVENDRA, P., SCHRAMM, T. and STEURER, D. (2018). High-dimensional estimation via sum-of-squares proofs. Available at arXiv:1807.11419.

[50] RAGHAVENDRA, P. and WEITZ, B. (2017). On the bit complexity of sum-of-squares proofs. In *44th International Colloquium on Automata, Languages, and Programming. LIPIcs. Leibniz Int. Proc. Inform.* **80** Art. No. 80, 13. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3685820

[51] RAHM, E. and DO HAI, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* **23** 3–13.

[52] SHOR, N. Z. (1987). An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics* **23** 695–700.

[53] TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics* 448–485. Stanford Univ. Press, Stanford, CA. MR0120720

[54] VANDENBERGHE, L., BOYD, S. and WU, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Appl.* **19** 499–533. MR1614078 https://doi.org/10.1137/S0895479896303430

[55] WANG, T. and SAMWORTH, R. J. (2018). High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 57–83. MR3744712 https://doi.org/10.1111/rssb.12243

[56] WILLIAMSON, D. P. and SHMOYS, D. B. (2011). *The Design of Approximation Algorithms*. Cambridge Univ. Press, Cambridge. MR2798112 https://doi.org/10.1017/CBO9780511921735