A Fast Spectral Algorithm for Mean Estimation with Sub-Gaussian Rates

Zhixian Lei ZHIXIANLEI@G.HARVARD.EDU
Harvard University

Kyle Luh KYLELUH@GMAIL.COM
Harvard University

Prayaag Venkat PVENKAT@G.HARVARD.EDU
Harvard University

Fred Zhang Z0@BERKELEY.EDU
UC Berkeley

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study the algorithmic problem of estimating the mean of a heavy-tailed random vector in $\mathbb{R}^d$, given n i.i.d. samples. The goal is to design an efficient estimator that attains the optimal sub-gaussian error bound, only assuming that the random vector has bounded mean and covariance. Polynomial-time solutions to this problem are known but have high runtime due to their use of semi-definite programming (SDP). Moreover, conceptually, it remains open whether convex relaxation is truly necessary for this problem.

In this work, we show that it is possible to go beyond SDP and achieve better computational efficiency. In particular, we provide a *spectral* algorithm that achieves the optimal statistical performance and runs in time $\widetilde{O}\left(n^2 d\right)$, improving upon the previous fastest runtime $\widetilde{O}\left(n^{3.5} + n^2 d\right)$ by Cherapanamjeri *et al.* (COLT '19). Our algorithm is spectral in that it only requires (approximate) eigenvector computations, which can be implemented very efficiently by, for example, power iteration or the Lanczos method.

At the core of our algorithm is a novel connection between the furthest hyperplane problem introduced by Karnin *et al.* (COLT '12) and a structural lemma on heavy-tailed distributions by Lugosi and Mendelson (Ann. Stat. '19). This allows us to iteratively reduce the estimation error at a geometric rate using only the information derived from the top singular vector of the data matrix, leading to a significantly faster running time.

Keywords: High-dimensional statistics, mean estimation, robust statistics, spectral algorithm

1. Introduction

Estimating the mean of a multivariate distribution from samples is among the most fundamental statistical problems. Surprisingly, it was only recently that a line of works in the statistics literature culminated in an estimator achieving the optimal statistical error under minimal assumptions (Lugosi and Mendelson (2019b)). However, from an algorithmic point of view, computation of this estimator appears to be intractable. On the other hand, fast estimators, such as the empirical average, tend to achieve sub-optimal statistical performance. The following question remains open:

Can we provide simple, fast algorithm that computes a statistically optimal mean estimator in high dimensions, under minimal assumptions?

In this paper, we make progress towards this goal, under the classic setting where only finite mean and covariance are assumed. Formally, our problem is defined as follows. Given n i.i.d. copies $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of a random vector $\boldsymbol{X} \in \mathbb{R}^d$ with bounded mean $\boldsymbol{\mu} = \mathbb{E}\,\boldsymbol{X}$ and covariance $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T$, compute an estimate $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ of the mean $\boldsymbol{\mu}$. Our goal is to show that for any failure probability $\delta \in (0, 1]$,

$$\Pr\left(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| > r_\delta\right) \le \delta,$$

for as small a radius r_δ as possible. Moreover, we would like to compute $\widehat{\boldsymbol{\mu}}$ efficiently. The naïve estimator is simply the empirical mean $\overline{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$. It is well known that among all estimators, the empirical mean minimizes mean squared error. However, if we instead use the size of the deviations to quantify the quality of the estimator, the empirical mean is only optimal for sub-gaussian random variables (Catoni (2012)). When $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we have with probability at least $1 - \delta$,

$$\|\overline{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \le \sqrt{\frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{n}} + \sqrt{\frac{2\|\boldsymbol{\Sigma}\|\log(1/\delta)}{n}} \tag{1.1}$$

An estimator that achieves above is said to have *sub-gaussian performance* or *sub-gaussian rate*.

In practical settings, assuming that the samples obey a Gaussian distribution may be unrealistic. In an effort to design robust estimators, it is natural to study the mean estimation problem under very weak assumptions on the data. A recent line of works (Catoni (2012); Minsker (2015); Devroye et al. (2016); Joly et al. (2017); Lugosi and Mendelson (2019a,b); Lerasle et al. (2019)) study the mean estimation problem when the samples obey a heavy-tailed distribution.

For heavy-tailed distributions the performance of the empirical mean is abysmal. If we only assume that $\boldsymbol{X}$ has finite mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then by Chebyshev's inequality, the empirical mean only achieves error of order $\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma})/\delta n}$, which is worse than the sub-gaussian rate in two ways. First, its dependence on $\frac{1}{\delta}$ is exponentially worse. Second, the $\mathrm{Tr}(\boldsymbol{\Sigma})$ term, which may grow with the dimension d, is multiplied the dimension-independent term $\sqrt{1/\delta n}$, whereas in the Gaussian case, the two are separate.

Median-of-means paradigm Surprisingly, recent work has shown that it is possible to improve on the performance of the empirical mean using the *median-of-means* approach. For $d = 1$, the following construction, originally due to Nemirovsky and Yudin (1983); Jerrum et al. (1986); Alon et al. (1999), achieves sub-gaussian performance:

(i) First, bucket the data into $k = \lceil 10\log(1/\delta)\rceil$ disjoint groups and compute their means $Z_1, Z_2, \cdots, Z_k$.

(ii) Then, output the median $\widehat{\mu}$ of $Z_1, Z_2, \cdots, Z_k$.

A long line of work has followed this paradigm and generalized it to higher dimensions (Catoni (2012); Devroye et al. (2016); Joly et al. (2017); Lugosi and Mendelson (2019a,b)). The key challenge is to correctly define a notion of median for a collection of points in $\mathbb{R}^d$. Minsker (2015) considered $\widehat{\boldsymbol{\mu}}_{GM}$ defined to be the *geometric median* of the bucket means $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k$. For some constant c_{GM}, with probability at least $1 - \delta$, it satisfies

$$\|\widehat{\boldsymbol{\mu}}_{GM} - \boldsymbol{\mu}\| \le c_{GM}\sqrt{\frac{\mathrm{Tr}\,\boldsymbol{\Sigma}\cdot\log(1/\delta)}{n}}. \tag{1.2}$$

This achieves the correct dependence on δ, but the dimension dependent and independent terms are still not separated. Following this work, Lugosi and Mendelson (2019a) described a tournament-based estimator, which finally achieved the optimal sub-gaussian radius. The idea is to consider *all* 1*-dimensional projections* of the bucket means and try to find an estimate that is close to the median of the means of all projections. This construction has been further simplified by Hopkins (2018). Formally, it was shown that the following estimator achieves the optimal, sub-gaussian error:

$$\widehat{\boldsymbol{\mu}}_{LM} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^d} \max_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \left|\text{median}\left\{\langle \boldsymbol{Z}_i, \boldsymbol{u}\rangle\right\}_{i=1}^{k} - \langle \boldsymbol{x}, \boldsymbol{u}\rangle\right|. \tag{1.3}$$

Clearly, searching over all directions in $\mathbb{S}^{d-1}$ requires exponential time. The key question, therefore, is whether one can achieve both computational and statistical efficiency simutaneously.

Computational considerations A priori, it is unclear that the Lugosi-Mendelson estimator can be computed in polynomial time as a direct approach involves solving an intractable optimization problem. Moreover, the Lugosi-Mendelson analysis seems to suggest that estimation in the heavy-tailed model is conceptually harder than under (adversarial) corruptions. In the latter, each sample can be classified as either an inlier or an outlier. In the heavy-tailed setting, Lugosi-Mendelson shows that there is a majority of the bucket means that cluster around the true mean along any projection. However, a given sample may be an inlier by being close to the mean when projected onto one direction, but an outlier when projected onto another. In other words, the *set* of inliers may change from one direction to another.

Surprisingly, a recent line of works have established the polynomial-time computability of Lugosi-Mendelson estimator. Hopkins (2018) formulates $\widehat{\boldsymbol{\mu}}_{LM}$ as the solution of a low-degree polynomial optimization problem and showed that using the Sum-of-Squares SDP hierarchy to relax this problem yields a sub-gaussian estimator. While the run-time of this algorithm is polynomial, it involves solving a large SDP. Soon after, Cherapanamjeri et al. (2019a) provided an iterative method in which each iteration involves solving a smaller, explicit SDP, leading to a run-time of $\widetilde{O}\left(n^{3.5} + n^2 d\right)$[1]. Even more recently, a concurrent and independent work by Lecué and Depersin (2019) gave an estimator with sub-gaussian performance that can be computed in time $\widetilde{O}(n^2 d)$. The construction is inspired by a near-linear time algorithm for robust mean estimation under adversarial corruptions due to Cheng et al. (2019). The algorithm requires solving (covering) SDPs.

We note, however, that a common technique in these algorithms is SDP, which tends to be impractical for large sample sizes and in high dimensions. In contrast, our algorithm only requires approximate eigenvector computations. For a problem as fundamental as mean estimation, it is desirable to obtain simple and ideally practical solutions. A key conceptual message of our work is that SDP is indeed unnecessary and can be replaced by simple spectral techniques.

Our result In this work, we demonstrate for the first time that mean estimation with sub-gaussian rates can be achieved efficiently *without* SDP. The runtime of the algorithm matches the independent work of Lecué and Depersin (2019). In addition, our algorithm enjoys robustness against (additive) corruptions, where the number of adversarial points is a small fraction of k.

It is known that there exists an information-theoretic requirement for achieving such rates—that is, $\delta \geq 2^{-O(n)}$ (Devroye et al. (2016)). Under this assumption, we give an efficient spectral algorithm.

1. Throughout we use $\widetilde{O}(\cdot)$ to hide polylogarithmic factors (in n, d and $\log(1/\delta)$).

Theorem 1 *Let $\delta \geq Ae^{-n}$ for a constant A and $k = \lceil 3600 \log(1/\delta) \rceil$. Given n points $\mathcal{G} \cup \mathcal{B}$, where $\mathcal{G}$ are i.i.d. samples from a distribution over $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and $\mathcal{B}$ a set of arbitrary points with $|\mathcal{B}| \leq k/200$, there is an efficient algorithm that outputs an estimate $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^d$ such that with probability at least $1 - \delta$,*

$$\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\| \leq C \left(\sqrt{\frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{n}} + \sqrt{\frac{\|\boldsymbol{\Sigma}\| \log(1/\delta)}{n}} \right),$$

for a constant C. Furthermore, the algorithm runs in time $O\left(nd + k^2 d \ \mathrm{polylog}(k, d)\right)$.

The algorithm is iterative. Each iteration only requires an (approximate) eigenvector computation, which can be implemented in nearly linear time by power iteration or the Lanczos algorithm. We believe that our algorithm can be fairly practical.

Other related work Recently, Prasad et al. (2019) established a formal connection between the Huber contamination model and the heavy-tailed model we study in this paper. They leverage this connection to use an existing $\widetilde{O}(nd^2)$-time mean estimation algorithm of Diakonikolas et al. (2019) to design estimators for the heavy-tailed model. Under moment assumptions, their estimator achieves performance better than geometric median (1.2), yet worse than sub-gaussian.

In addition, algorithmic robust statistics has gained much attention in the theoretical computer science community in recent years. A large body of works have studied the mean estimation problem with *adversarially* corrupted samples, with the focus on providing efficient algorithms (Diakonikolas et al. (2019); Lai et al. (2016); Cheng et al. (2019); Dong et al. (2019)). For a more complete survey, see Diakonikolas and Kane (2019)

Going beyond mean estimation, there has been a recent spate of works on other statistical problems under heavy-tailed distributions. We refer the readers to Lugosi and Mendelson (2019c) for a survey.

Technical overview Our main algorithm builds upon the iterative approach of Cherapanamjeri et al. (2019a). For simplicity, assume there is no adversarial point. At a high level, for each iteration t, the algorithm will maintain a current guess $\boldsymbol{x}_t$ of the true mean. To update, Cherapanamjeri *et al.* study the inner maximization of $\widehat{\boldsymbol{\mu}}_{LM}$ (1.3) with $\boldsymbol{x} = \boldsymbol{x}_t$. They showed that under Lugosi-Mendelson structral condition, the problem is essentially equivalent of following program, which we call $\mathcal{M}(\boldsymbol{x}_t, \boldsymbol{Z})$:

$$\begin{aligned} \max \quad & \theta \\ \text{subject to} \quad & b_i \langle \boldsymbol{Z}_i - \boldsymbol{x}_t, \boldsymbol{u} \rangle \geq b_i \theta \text{ for } i = 1, \ldots, k \\ & \sum_{i=1}^{k} b_i \geq 0.95k \\ & \boldsymbol{b} \in \{0,1\}^k, \boldsymbol{u} \in \mathbb{S}^{d-1}. \end{aligned}$$

It can be shown that an optimal solution $\boldsymbol{u} \in \mathbb{S}^{d-1}$ will align with the unit vector in the direction of $\boldsymbol{\mu} - \boldsymbol{x}_t$, and θ approximates $\|\boldsymbol{\mu} - \boldsymbol{x}_t\|$. Hence, one can perform the update $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t + \gamma\theta\boldsymbol{u}$, for some appropriate constant γ, to geometrically decrease the distance of $\boldsymbol{x}_t$ to $\boldsymbol{\mu}$.

In this work, we start by drawing a connection between the above program and the furthest hyperplane problem (FHP) of Karnin et al. (2012). This allows us to avoid the SDP approach in

Cherapanamjeri et al. (2019a). The problem can be formulated as the following:

$$\begin{aligned} \max \quad & \theta && \text{(FHP)} \\ \text{subject to} \quad & |\langle \boldsymbol{Z}_i - \boldsymbol{x}_t, \boldsymbol{u}\rangle| \geq \theta \text{ for } i = 1, \ldots, k && (1.4) \\ & \boldsymbol{u} \in \mathbb{S}^{d-1}. \end{aligned}$$

In the original formulation due to Karnin *et al.*, the goal is to find a maximum margin linear classifier for a collection of points, where the margin is *two-sided*. Notice that any feasible solution to $\mathcal{M}(\boldsymbol{x}_t, \boldsymbol{Z})$ satisfies at least $0,95k$ constraints of FHP as well. For an arbitrary dataset, the two-sided margin requirement indeed provides a relaxation. One technical observation of this work is that it is *not* a significant one, for the random data we care about—if a major fraction of the constraint (1.4) are satisfied, then most constraints of $\mathcal{M}(\boldsymbol{x}_t, \boldsymbol{Z})$ are satisfied as well.

Unfortunately, the algorithm of Karnin *et al.* cannot directly apply, as it only works under a strong promise that there exists a feasible solution that satisfies *all* of the constraints (1.4). In our setting, there may not be such a feasible solution; we can only guarantee that there exists a unit vector (namely, the one in the direction of $\boldsymbol{\mu} - \boldsymbol{x}_t$) that satisfies most of constraints with large margin.

Our main contribution is to provide an algorithm that works even under this weak promise. We now briefly review the algorithm of Karnin *et al.*, show why it fails for our purpose, and explain how we address the issues that arise. Suppose that there exists a unit vector $\boldsymbol{u}^*$ and θ^* which are feasible for the FHP problem. Then, averaging the constraints tells us that

$$\frac{1}{k}\sum_{i=1}^{k} \langle \boldsymbol{Z}_i, \boldsymbol{u}^*\rangle^2 \geq \theta^{*2}.$$

Hence, if we define $\boldsymbol{u}$ to be the top right singular vector of the matrix $\boldsymbol{A}$ whose rows are $\boldsymbol{Z}_i$, then

$$\|\boldsymbol{A}\boldsymbol{u}\|_2^2 = \sum_{i=1}^{k} \langle \boldsymbol{Z}_i, \boldsymbol{u}\rangle^2 \geq \sum_{i=1}^{k} \langle \boldsymbol{Z}_i, \boldsymbol{u}^*\rangle^2 \geq k\theta^{*2},$$

so $\boldsymbol{u}$ satisfies the constraints in (FHP) *on average*. However, the distribution of the quantities $\langle \boldsymbol{Z}_i, \boldsymbol{u}\rangle^2$ may be extremely skewed, so that $\boldsymbol{u}$ only satisfies a few of the constraints with large margin. If this happens, however, we can downweight those constraints which are satisfied by $\boldsymbol{u}$ with large slack to encourage it to satisfy more constraints. This reweighting procedure is repeated several times, and at the end we use a simple rounding scheme to yield a single output vector with the desired properties from all the repetitions. In particular, this weighting scheme is essentially the same as the classic *multiplicative weights update* (MWU) method (Arora et al. (2012)) for regret minimization, as we show in Appendix I.

If we are only guaranteed that $\boldsymbol{u}^*$ satisfies most, but not all, of the constraints, then the inequality $\sum_{i=1}^{k} \langle \boldsymbol{Z}_i, \boldsymbol{u}^*\rangle^2 \geq k\theta^{*2}$ may no longer hold when the points $\boldsymbol{Z}_i$ get re-weighted and the algorithm of Karnin *et al.* cannot be guaranteed to converge. To illustrate this point, consider the following extreme case. Suppose that after the first iteration, the algorithm finds the vector $\boldsymbol{u}^*$ as the top right singular vector of $\boldsymbol{A}$. In the re-weighting procedure, the constraints i for which $\langle \boldsymbol{Z}_i, \boldsymbol{u}^*\rangle^2 \geq \theta^{*2}$ may be down-weighted significantly, whereas the remaining constraints may be unaffected. This may result in most of the weight being concentrated on the constraints i where $\langle \boldsymbol{Z}_i, \boldsymbol{u}^*\rangle^2 \ll \theta^{*2}$. In the second iteration, we have no guarantee of the behavior of the top singular vector of the re-weighted matrix because all the weight is concentrated on a small set consisting of these "bad" constraints.

To address this scenario, our key technical idea is to project the weights onto the set of *smooth distributions* after each update. Informally, the notion of smooth distribution enforces that no point can take too much probability mass—say, more than $4/k$. This prevents the weights from ever being concentrated on too small a subset and allows us to guarantee that $\sum_{i=1}^{k} \langle \boldsymbol{Z}_i, \boldsymbol{u}^* \rangle^2 \geq k\theta^{*2}$ still holds approximately. Moreover, the appropriate notion of projection here is that of a Bregman projection. Leveraging our earlier MWU interpretation of the algorithm (Section I), we apply a classic regret bound for MWU under Bregman projection (Arora et al. (2012)), and this yields the same guarantee of the original algorithm. Finally, we remark that the projection can be computed quickly. Combining all these ideas together, we manage to bypass the barrier of having bad points, under the much weaker assumption on $\boldsymbol{u}^*$.

Organization The remainder of this article is organized as follows. In Section 2, we set up the notations and specify assumptions on the data. In Section 3, we explain the high level approach based on an iterative descent procedure from Cherapanamjeri et al. (2019a). The procedure requires us to approximately maximize a (non-convex) objective, and we discuss its properties in Section 4. Section 5 contains the main technical innovations of this work, where we design and analyze a faster algorithm for the aforementioned optimization problem.

2. Preliminaries and Assumptions

In the following, we use $r_\delta = \sqrt{\mathrm{Tr}(\boldsymbol{\Sigma})/n} + \sqrt{\|\boldsymbol{\Sigma}\| \log(1/\delta)/n}$ to denote the optimal, sub-gaussian error rate and $k = \lceil 3200 \log(8/\delta) \rceil$. The input data $\{\boldsymbol{X}_i\}_{i=1}^n$ consist of $\mathcal{G}$, a set of i.i.d. points, and $\mathcal{B}$, a set of adversarial points, with $|B| \leq k/200$. Our algorithm preprocesses the data $\boldsymbol{X}_i$ into the bucket means $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_{2k} \in \mathbb{R}^d$.[2] Let $\mathcal{B}_j$ be the set of $\boldsymbol{X}_i$ in bucket j. We say that a bucket mean $\boldsymbol{Z}_j$ is *contaminated* if B_j contains an adversarial $X_i \in B$ and *uncontaminated* otherwise. Note that the number of contaminated bucket means is at most $k/200$.

Our argument is built on the Lugosi-Mendelson condition. It states that under any one-dimensional projection, most of the (uncontaminated) bucket means are close to the true mean, by an additive factor of $O(r_\delta)$. Throughout, we pessimistically assume all contaminated bucket means do not satisfy this property (under any projection) and condition on the following event.

Assumption 1 (Lugosi-Mendelson condition) *Under the setting above, for all unit $\boldsymbol{v}$, we have*

$$|\{i : \langle \boldsymbol{v}, \boldsymbol{Z}_i \rangle - \langle \boldsymbol{v}, \boldsymbol{\mu} \rangle \geq 600 r_\delta\}| \leq 0.05k.$$

Lemma 2 (Lugosi and Mendelson (2019b)) *Assumption 1 holds with probability at least $1 - \delta/8$.*

3. Descent Procedure

At a high level, our algorithm builds upon the iterative descent paradigm of Cherapanamjeri et al. (2019a). It maintains a sequence of estimates and updates via distance and gradient estimate.

Definition 3 (distance estimate) *We say that d_t is a distance estimate (with respect to $\boldsymbol{x}_t$) if*

(i) when $\|\boldsymbol{\mu} - \boldsymbol{x}_t\| \leq 14000 r_\delta$, we have $d_t \leq 28000 r_\delta$; and

2. We assume δ is such that $k \leq n/2$; as we mentioned in the introduction, this is information-theoretically necessary, up to a constant (Devroye et al. (2016)).

(ii) when $\|\boldsymbol{\mu} - \boldsymbol{x}_t\| > 14000 r_\delta$*, we have*

$$\frac{1}{21}\|\boldsymbol{\mu} - \boldsymbol{x}_t\| \le d_t \le 2\|\boldsymbol{\mu} - \boldsymbol{x}_t\| \tag{3.1}$$

Definition 4 (gradient estimate) *We say that* $\boldsymbol{g}_t$ *is a gradient estimate (with respect to* $\boldsymbol{x}_t$*) if*

$$\left\langle \boldsymbol{g}_t, \frac{\boldsymbol{\mu} - \boldsymbol{x}_t}{\|\boldsymbol{\mu} - \boldsymbol{x}_t\|} \right\rangle \ge \frac{1}{200} \tag{3.2}$$

whenever $\|\boldsymbol{\mu} - \boldsymbol{x}_t\| > 14000 r_\delta$.

1. **Input:** Buckets means $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k \in \mathbb{R}^d$, initial estimate $\boldsymbol{x}_0$, iteration count T_{des}, and step size η.
2. For $t = 1, \ldots, T_{\mathsf{des}}$:
   (a) Compute $d_t = \textsc{DistEst}(\boldsymbol{Z}', \boldsymbol{x}_t)$.
   (b) Compute $\boldsymbol{g}_t = \textsc{GradEst}(\boldsymbol{Z}', \boldsymbol{x}_t)$.
   (c) Update $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta d_t \boldsymbol{g}_t$.
3. **Output:** $\boldsymbol{x}_{t^*}$, where $t^* = \arg\min_t d_t$.

Algorithm 3.1: Main algorithm—DESCENT

Suppose we initialize the estimate with coordinate-wise median-of-means which achieves an error rate $\sqrt{\|\boldsymbol{\Sigma}\| kd/n}$ (Lemma 16). The following lemma states that if DISTEST and GRADEST provide distance and gradient estimate, then the algorithm DESCENT succeeds in logarithmic iterations. The lemma has essentially appeared in Cherapanamjeri et al. (2019a), albeit with a general initialization and a different set of constants. We give a proof in Appendix C for completeness.

Lemma 5 (convergence rate; see Cherapanamjeri et al. (2019a)) *Assume that for all* $t \le T_{\mathsf{des}}$*,* d_t *is a distance estimate and* $\boldsymbol{g}_t$ *is a gradient estimate (with respect to* $\boldsymbol{x}_t$*). Suppose* $\|\boldsymbol{\mu} - \boldsymbol{x}_0\| \le O\left(\sqrt{\|\boldsymbol{\Sigma}\| kd/n}\right)$*. Then the output of Algorithm 3.1* DESCENT *instantiated with* $T_{\mathsf{des}} = \Theta(\log d)$ *and* $\eta = 1/8000$ *satisfies* $\|\boldsymbol{x}_{t^*} - \boldsymbol{\mu}\| \le O(r_\delta)$.

4. Inner Maximization and its Two-Sided Relaxation

Cherapanamjeri et al. (2019a) obtains gradient and distance estimates by solving the *inner maximization* problem of the Lugosi-Mendelson estimator, denoted by $\mathcal{M}(\boldsymbol{x}, \boldsymbol{Z})$:

$$\begin{aligned} \max \quad & \theta \\ \text{subject to} \quad & b_i \langle \boldsymbol{Z}_i - \boldsymbol{x}, \boldsymbol{w} \rangle \ge b_i \theta \text{ for } i = 1, \ldots, k \\ & \sum_{i=1}^k b_i \ge 0.95k \end{aligned}$$

$$\boldsymbol{b} \in \{0,1\}^k, \boldsymbol{w} \in \mathbb{S}^{d-1}.$$

We also denote its feasibility version for a fixed θ by $\mathcal{M}(\theta, \boldsymbol{x}, \boldsymbol{Z})$. Note that the constraint of $\mathcal{M}(\boldsymbol{x}, \boldsymbol{Z})$ dictates that 0.95 fraction of the data must lie on *one* side of the hyperplane $\boldsymbol{w}$ with a margin θ. As discussed in the introduction, we relax it by allowing a two-sided margin: $\mathcal{M}_2(\boldsymbol{x}, \boldsymbol{Z})$.

$$\begin{aligned} \max \quad & \theta \\ \text{subject to} \quad & b_i |\langle \boldsymbol{Z}_i - \boldsymbol{x}, \boldsymbol{w}\rangle| \geq b_i\theta \text{ for } i = 1, \ldots, k \\ & \sum_{i=1}^{k} b_i \geq 0.95k \\ & \boldsymbol{b} \in \{0,1\}^k, \boldsymbol{w} \in \mathbb{S}^{D-1}. \end{aligned}$$

One technical observation here is that under the Lugosi-Mendelson condition, this relaxation is insignificant. Indeed, approximately solving the problem suffices for gradient and distance estimates.

Lemma 6 *Let θ^* be the optimal value of $\mathcal{M}(\boldsymbol{x}, \boldsymbol{Z})$ and $\boldsymbol{w}$ be a unit vector such that for at least $k/8$ of the $\boldsymbol{Z}_i$, we have $|\langle \boldsymbol{w}, \boldsymbol{Z}_i - \boldsymbol{x}\rangle| \geq \theta$, where $\theta = 0.1\theta^*$. We have that (i) θ is a distance estimate and (ii) either $\boldsymbol{w}$ or $-\boldsymbol{w}$ is a gradient estimate.*

We give a proof in Appendix D. The intuition here is simple. If $\|\boldsymbol{x} - \boldsymbol{\mu}\| \ll r_\delta$, then the Lugosi-Mendelson condition ensures at most $0.05k$ points are far from $\boldsymbol{x}$ by $O(r_\delta)$ (under any projection), so $\theta = O(r_\delta)$. On the other hand, if $\|\boldsymbol{x} - \boldsymbol{\mu}\| \gg r_\delta$, along the gradient direction, a majority of data lie only on *one* side of the hyperplane, the side that contains the true mean, so the two-sided constraint does not make a difference.

5. Approximating the Inner Maximization

We now give an algorithm that efficiently computes a approximate solution to the relaxation of the inner maximization. This will provide gradient and distance estimates for each iteration of the main DESCENT algorithm (Algorithm 3.1).

The run-time of the algorithm is proportional $1/\theta^2$. For technical reasons, we need to ensure that $\|\boldsymbol{Z}_i - \boldsymbol{x}\| \leq 1$ for all i. However, naïvely scaling all the data would decrease θ, thereby blowing up the running time. Hence, as a preprocessing step, we prune out a small fraction of points $\boldsymbol{Z}_i - \boldsymbol{x}$ with large norm before scaling.

5.1. Pruning and scaling

The preprocessing step (Algorithm 5.1) will be executed *only once* in the algorithm. After the pruning step and an appropriate scaling, we may assume the following structures on the data.

Assumption 2 *Given a current estimate $\boldsymbol{x}$, the pruned dataset $\boldsymbol{Z} \in \mathbb{R}^{k' \times d}$ of size k', let $\boldsymbol{Z}'_i = \frac{1}{B}(\boldsymbol{Z}_i - \boldsymbol{x})$, where $B = \max_i \|\boldsymbol{Z}'_i - \boldsymbol{x}\|$. We assume (i) $\|\boldsymbol{Z}'_i\| \leq 1$; (ii) $k' \geq 0.9k$; and (iii) there exists $\theta = \Omega(1/\sqrt{d})$ and a unit vector $\boldsymbol{w}$ such that for at least $0.8k$ points $|\langle \boldsymbol{Z}'_i, \boldsymbol{w}\rangle| \geq \theta$.*

We analyze the subroutine and prove the following lemma in Appendix E.

1. **Input:** Dataset $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_k \in \mathbb{R}^d$, initial estimate $\boldsymbol{x}_0$
2. Compute the distances $d_i = \|\boldsymbol{Z}_i - \boldsymbol{x}_0\|$.
3. Sort the points by d_i in decreasing order.
4. Remove the top $1/10$ fraction of them. Let $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_{k'}$ be the remaining data.
5. **Output:** $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_{k'}$

Algorithm 5.1: PRUNE

Lemma 7 *With probability at least $1 - \delta/8$, Assumption 2 holds for any $\boldsymbol{x}$ such that $\|\boldsymbol{x} - \boldsymbol{\mu}\| \leq O\left(\sqrt{\|\boldsymbol{\Sigma}\| kd/n}\right)$ and $\|\boldsymbol{x} - \boldsymbol{\mu}\| \geq \Omega(r_\delta)$.*

In the remainder of the section, given a current estimate $\boldsymbol{x}$, we work with the pruned and scaled data, centered at $\boldsymbol{x}$, which we call $\boldsymbol{Z}' \in \mathbb{R}^{k' \times d}$.

We will aim at proving the following lemma, under Assumption 2.

Lemma 8 (key lemma) *Assume Assumption 2. Let $\delta \in (0, 1)$ and $T_{\mathsf{des}} = \Theta(\log d)$. Suppose that there exists $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$ which satisfies $|\langle \boldsymbol{Z}'_i, \boldsymbol{w}^* \rangle| \geq \theta^*$ for $0.8k$ points in $\{\boldsymbol{Z}'_i\}$. Then there is an algorithm* APPROXBREGMAN *which, with probability at least $1 - \delta/4T_{\mathsf{des}}$, outputs $\boldsymbol{w} \in \mathbb{S}^{d-1}$ such that for at least 0.45 fraction of the points $\boldsymbol{Z}'_i$, it holds that $|\langle \boldsymbol{Z}'_i, \boldsymbol{w} \rangle| \geq 0.1\theta^*$.*

Further, APPROXBREGMAN *runs in time $\widetilde{O}\left(k^2 d\right)$.*

5.2. Approximation via Bregman Projection

In this section, we give the main algorithm for approximating $\mathcal{M}_2$. Suppose (by binary search) that we know the optimal margin θ in Lemma 8. The goal is to find a unit vector $\boldsymbol{w}$ such that a constant fraction of $\boldsymbol{Z}'_i$ has margin $|\langle \boldsymbol{Z}'_i, \boldsymbol{w} \rangle| \geq \theta$. The intuition is that we can start by computing the top singular vector of $\boldsymbol{Z}'$. Then the margin would be large on average: certain points may overly satisfy the margin demand while other may under-satisfy it. Hence, we would downweight those data poitns that achieve large margin and compute the top singular vector of the weighted matrix again.

However, it may stop making progress if it puts too much weight on the points that do not satisfy the margin bound. In this section, we show how to prevent this scenario from occurring. The key idea is that at every iteration, we "smooth" the weight vector τ_t so that we can guarantee progress is being made. We will formulate our algorithm in the well-studied regret-minimization framework and appeal to existing machinery Arora et al. (2012) to derive the desired approximation guarantees.

First, we define what type of distribution we would like τ_t to be.

Definition 9 (Smooth distributions) *The set of* smooth distributions *on $[k']$ is defined to be*

$$\mathcal{K} = \left\{ p \in \Delta_{k'} : p(i) \leq \frac{4}{k'} \text{ for every } i \in [k'] \right\},$$

where $\Delta_{k'}$ is the set of probability distributions on $[k']$,

$$\Delta_{k'} = \left\{ p : [k'] \to [0,1] : \sum_{i \in [k']} p(i) = 1 \right\}.$$

In the course of the algorithm, after updating τ_t as in the previous section, it may no longer be smooth. Hence, we will replace it by the closest smooth weight vector (under KL divergence). The following fact confirms that finding this closest smooth weight vector can be done quickly.

Fact 10 (Barak et al. (2009)) *For any $p \in \Delta_k$ with support size at least $k'/2$, computing*

$$\Pi_{\mathcal{K}}(p) = \arg\min_{q \in \mathcal{K}} KL(p||q)$$

can be done in $\tilde{O}(k')$ time, where $KL(\cdot||\cdot)$ denotes the Kullback-Leibler divergence.

Remark 11 *In our algorithm, we will only compute Bregman projections of distributions of support size at least $k'/2$. This is because neither our reweighting procedure nor the actual projection algorithm of Barak et al. (2009) sets any coordinates to 0 and the initial weight is uniform.*

1. **Input:** Buckets means $\boldsymbol{Z}' \in \mathbb{R}^{k' \times d}$, margin θ, iteration count $T \in \mathbb{N}$
2. Initialize weights: $\boldsymbol{\tau}_1 = \frac{1}{k'}(1, \ldots, 1) \in \mathbb{R}^{k'}$.
3. For $t = 1, \ldots, T$, repeat:
   (a) Let $\boldsymbol{A}_t$ be the $k' \times d$ matrix whose ith row is $\sqrt{\boldsymbol{\tau}_t(i)}(\boldsymbol{Z}'_i)$ and $\boldsymbol{w}_t$ be its approximate top right singular vector .
   (b) Set $\sigma_t(i) = |\langle \boldsymbol{Z}'_i, \boldsymbol{w}_t \rangle|$.
   (c) Reweight: If $\|\boldsymbol{A}_t \boldsymbol{w}_t\|_2^2 \geq \frac{\theta^2}{10}$, then $\boldsymbol{\tau}_{t+1}(i) = \boldsymbol{\tau}_t(i)\left(1 - \boldsymbol{\sigma}_t(i)^2/2\right)$ for $i \in [k']$. Otherwise, do not change the weights.
   (d) Normalize: Let $Z = \sum_{i \in [k']} \boldsymbol{\tau}_{t+1}(i)$ and redefine $\boldsymbol{\tau}_{t+1} \leftarrow \frac{1}{Z}\boldsymbol{\tau}_{t+1}$.
   (e) Compute the Bregman projection: $\boldsymbol{\tau}_{t+1} \leftarrow \Pi_{\mathcal{K}}(\boldsymbol{\tau}_{t+1})$.
4. **Output**: $\boldsymbol{w} \leftarrow$ROUND$\left(\boldsymbol{Z}', \{\boldsymbol{w}_j\}_{t=1}^T, \theta\right)$ (or report FAIL if ROUND fails).

Algorithm 5.2: Approximate inner maximization via Bregman projection—APPROXBREGMAN

Since Algorithm 5.2 is the MWU method with Bregman projections onto the set $\mathcal{K}$, we will apply the following regret guarantee.[3]

3. To be more precise, the iterations t in which $\|\boldsymbol{A}_t \boldsymbol{w}_t\|_2^2 \geq \frac{\theta^2}{10}$ behave according to the MWU method. Whenever $\|\boldsymbol{A}_t \boldsymbol{w}_t\|_2^2 < \frac{\theta^2}{10}$, the algorithm does not update the weights, which has no effect on the other iterations.

Theorem 12 (Theorem 2.4 of Arora et al. (2012)) *Suppose that for $\sigma_t^2(i) \in [0,1]$ for all $i \in [k']$ and $t \in [T]$. Then after T iterations of Algorithm 5.2, for any $p \in \mathcal{K}$, it holds that:*

$$\sum_{t=1}^{T} \left\langle \boldsymbol{\tau}_t, \boldsymbol{\sigma}_t^2 \right\rangle \leq \frac{3}{2} \sum_{t=1}^{T} \left\langle \boldsymbol{p}, \boldsymbol{\sigma}_t^2 \right\rangle + 2KL(\boldsymbol{p}||\boldsymbol{\tau}_1).$$

Finally, we comment that we cannot naïvely apply the power method for the singular vector computation. The power method has failure probability of $1/10$, whereas our algorithm should fail with probability at most $\delta = O(\exp(-k))$ that is exponentially low. However, we note that the algorithm computes the top singular vectors of a sequence of matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_T$. Observe that as long as $T = \Omega(\log(1/\delta)) = \Omega(k)$, with probability at least $1 - \delta/8$, the power method will succeed for $0.9T$ of the matrices. We will show that this many successes suffice to guarantee correctness of our algorithm.

We first prove the following lemma, a requirement for the rounding algorithm to succeed.

Lemma 13 (regret analysis) *After $T = O\left(\max\left(\frac{\log k'}{\theta^2}, \log(T_{\mathsf{des}}/\delta)\right)\right)$ iterations of Algorithm 5.2, for all but a $1/4$ fraction of $i \in [k']$:*

$$\sum_{t=1}^{T} \left\langle \boldsymbol{Z}_i', \boldsymbol{w}_t \right\rangle^2 \geq 100 \log k'.$$

Proof Let $S = \{i \in [k'] : |\left\langle \boldsymbol{Z}_i', \boldsymbol{w}^* \right\rangle| \geq \theta\}$ be the set of constraints satisfied by the unit vector $\boldsymbol{w}^*$ whose existence is guaranteed in the hypothesis of Lemma 8. By assumption, we have that $|S| \geq 0.8k'$. We simply calculate each of the terms in Theorem 12.

First, let $\mathcal{I} = \{t \in [T] : \boldsymbol{w}_t$ is a $1/2$-approximate top singular vector of $\boldsymbol{A}_t\}$. Then we have for any $t \in \mathcal{I}$:

$$\begin{aligned}
\left\langle \boldsymbol{\tau}_t, \boldsymbol{\sigma}_t^2 \right\rangle &= \sum_{i=1}^{k'} \boldsymbol{\tau}_t(i) \left\langle \boldsymbol{Z}_i', \boldsymbol{w}_t \right\rangle^2 && \text{(by definition)} \\
&\geq \frac{1}{2} \sum_{i=1}^{k'} \boldsymbol{\tau}_t(i) \left\langle \boldsymbol{Z}_i', \boldsymbol{w}^* \right\rangle^2 && \text{(because } \boldsymbol{w}_t \text{ is an approximate top eigenvector)} \\
&\geq \frac{1}{2} \sum_{i \in S} \boldsymbol{\tau}_t(i) \left\langle \boldsymbol{Z}_i', \boldsymbol{w}^* \right\rangle^2 \\
&\geq \frac{1}{2} \sum_{i \in S} \boldsymbol{\tau}_t(i) \theta^2 && \text{(by definition of } S) \\
&\geq \frac{1}{2} \cdot \frac{1}{5} \theta^2 = \frac{\theta^2}{10} && \text{(because } |S| \geq 0.8k' \text{ and } \boldsymbol{\tau}_t \in \mathcal{K}).
\end{aligned}$$

Summing this inequality over $t \in [T]$, we have that

$$\sum_{t=1}^{T} \left\langle \boldsymbol{\tau}_t, \boldsymbol{\sigma}_t^2 \right\rangle \geq \sum_{t \in \mathcal{I}} \left\langle \boldsymbol{\tau}_t, \boldsymbol{\sigma}_t^2 \right\rangle \geq \frac{|\mathcal{I}|}{10} \theta^2.$$

By Chernoff-Hoeffding bound combined with the guarantee of power iteration (Fact 17), as long as $T = \Omega(\log(T_{\mathsf{des}}/\delta))$, then with probability at least $1 - \frac{\delta}{8T_{\mathsf{des}}}$, for at least $\frac{4}{5}T$ iterations, it will be the case that $\boldsymbol{w}_t$ is an approximate top singular vector. In other words, $|\mathcal{I}| \geq \frac{4}{5}T$, so that we have:

$$\sum_{t=1}^{T} \langle \boldsymbol{\tau}_t, \boldsymbol{\sigma}_t^2 \rangle \geq \frac{2T}{25}\theta^2.$$

Next, note that if we choose $\boldsymbol{p} = \boldsymbol{e}_i$, then

$$\sum_{t=1}^{T} \langle \boldsymbol{p}, \boldsymbol{\sigma}_t^2 \rangle = \sum_{t=1}^{T} \langle \boldsymbol{Z}_i', \boldsymbol{w}_t \rangle^2 .$$

Because $\boldsymbol{\tau}_1$ is uniform, the relative entropy term in Theorem 12 is at most $\log k'$. Let's pretend for a moment that $\boldsymbol{e}_i \in \mathcal{K}$ (it is not). Then after plugging in the above calculations to Theorem 12 and rearranging, we have that for every $i \in [k']$

$$\sum_{t=1}^{T} \langle \boldsymbol{Z}_i', \boldsymbol{w}_t \rangle^2 \geq \frac{2T}{25}\theta^2 - 2\log k' \geq 100 \log k',$$

by setting $T \geq \frac{10^5 \log k'}{\theta^2}$. This gives the bound claimed in the statement of the lemma, but it remains to fix the invalid assumption that $\boldsymbol{e}_i \in \mathcal{K}$. To do so, we will construct, for most $i \in [k']$, another distribution $\boldsymbol{p}' \in \mathcal{K}$ such that

$$\sum_{t=1}^{T} \langle \boldsymbol{e}_i, \boldsymbol{\sigma}_t^2 \rangle \geq \sum_{t=1}^{T} \langle \boldsymbol{p}', \boldsymbol{\sigma}_t^2 \rangle .$$

Combining this with $\sum_{t=1}^{T} \langle \boldsymbol{p}', \boldsymbol{\sigma}_t^2 \rangle \geq 100 \log k'$ gives the desired lower bound, for most i. Write $\boldsymbol{\alpha} = \sum_{t=1}^{T} \boldsymbol{\sigma}_t^2$, and without loss of generality assume that

$$\boldsymbol{\alpha}_1 \geq \boldsymbol{\alpha}_2 \geq \ldots \geq \boldsymbol{\alpha}_{k'}.$$

For $i = 1, \ldots, 4k'/5$, take $\boldsymbol{p}'$ to be uniform on those $j \in [k']$ such that $\boldsymbol{\alpha}_i \geq \boldsymbol{\alpha}_j$ (there are at least $k'/5$ such i). By construction, we have that $\langle \boldsymbol{\alpha}, \boldsymbol{e}_i \rangle \geq \langle \boldsymbol{\alpha}, \boldsymbol{p}' \rangle$. Finally, observe that $\boldsymbol{p}' \in \mathcal{K}$ because $\boldsymbol{p}'$ is uniform on a set of size at least $k'/5$. ■

Observe that the APPROXBREGMAN produces a sequence of vectors by the end. Karnin et al. (2012) provides a rounding algorithm that combines them into one with the desired margin bound. We describe the algorithm and prove the following lemma in Appendix F.

Lemma 14 *The algorithm* ROUND *(Algorithm F.1) outputs $\boldsymbol{w}$ that satisfies $|\langle \boldsymbol{Z}_i', \boldsymbol{w} \rangle| \geq 0.1\theta$ for $0.45k$ of the points, with probability at least $1 - \delta/4T_{\mathsf{des}}$.*

Finally, we can prove the key lemma using APPROXBREGMAN (Appendix G) and put everything together in Appendix H.

Acknowledgments

The authors would like to thank Boaz Barak and Jelani Nelson for helpful conversations. In particular, we would like to thank Boaz for directing us to the paper (Barak et al. (2009)).

Zhixian Lei and Prayaag Venkat have been supported by NSF awards CCF 1565264 and CNS 1618026, and the Simons Foundation. Kyle Luh has been partially supported by NSF postdoctoral fellowship DMS-1702533. Prayaag Venkat has also been supported by an NSF Graduate Fellowship under grant DGE1745303.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1, part 2):137–147, 1999. ISSN 0022-0000. doi: 10.1006/jcss.1997.1545.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012. doi: 10.4086/toc.2012.v008a006.

Boaz Barak, Moritz Hardt, and Satyen Kale. The uniform hardcore lemma via approximate Bregman projections. In *ACM-SIAM Symposium on Discrete Algorithms (SODA '09)*, pages 1193–1200, 2009. URL http://dl.acm.org/citation.cfm?id=1496770.1496899.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2019.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012. ISSN 0246-0203. doi: 10.1214/11-AIHP454.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *ACM-SIAM Symposium on Discrete Algorithms (SODA '19)*, pages 2755–2771, 2019. doi: 10.1137/1.9781611975482.171.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory (COLT '19)*, pages 786–806, 2019a. URL https://arxiv.org/abs/1902.01998.

Yeshwanth Cherapanamjeri, Samuel B. Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. *arXiv preprint arXiv:1912.11071*, 2019b.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-gaussian mean estimators. *Annals of Statistics*, 44(6):2695–2725, 2016. doi: 10.1214/16-AOS1440.

Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019. URL https://arxiv.org/abs/1911.05911.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019. doi: 10.1137/17M1126680.

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. URL https://arxiv.org/abs/1906.11366.

Samuel B Hopkins. Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018. URL https://arxiv.org/abs/1809.07425.

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986. ISSN 0304-3975. doi: 10.1016/0304-3975(86)90174-X.

Emilien Joly, Gábor Lugosi, and Roberto Imbuzeiro Oliveira. On the estimation of the mean of a random vector. *Electronic Journal of Statistics*, 11(1):440–451, 2017. doi: 10.1214/17-EJS1228.

Zohar Karnin, Edo Liberty, Shachar Lovett, Roy Schwartz, and Omri Weinstein. Unsupervised SVMs: On the complexity of the furthest hyperplane problem. In *Conference on Learning Theory (COLT '12)*, pages 1–17, 2012. URL http://proceedings.mlr.press/v23/karnin12.html.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Symposium on Foundations of Computer Science (FOCS '16)*, pages 665–674, 2016. doi: 10.1109/FOCS.2016.76.

Guillaume Lecué and Jules Depersin. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019. URL https://arxiv.org/abs/1906.03058.

Matthieu Lerasle, Zoltan Szabo, Timothée Mathieu, and Guillaume Lecue. Monk outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning*, pages 3782–3793, 2019.

Gábor Lugosi and Shahar Mendelson. Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, 2019a. ISSN 1432-2064. doi: 10.1007/s00440-019-00906-4.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794, 2019b. ISSN 0090-5364. doi: 10.1214/17-AOS1639.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, Aug 2019c. doi: 10.1007/s10208-019-09427-x.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015. doi: 10.3150/14-BEJ645.

A. S. Nemirovsky and D. B. and Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. ISBN 0-471-10345-4. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Masashi Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the institute of Statistical Mathematics*, 10(1):29–35, 1959.

Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019. URL https://arxiv.org/abs/1907.00927.