# High-Dimensional Robust Mean Estimation via Outlier-Sparsity Minimization

Aditya Deshmukh*
Department of Electrical and Computer Engineering, and Coordinated Science Laboratory
UIUC
ad11@illinois.edu

Jing Liu*
Department of Electrical and Computer Engineering, and Coordinated Science Laboratory
UIUC
jil292@illinois.edu

Venugopal V. Veeravalli
Department of Electrical and Computer Engineering, and Coordinated Science Laboratory
UIUC
vvv@illinois.edu

2021 55th Asilomar Conference on Signals, Systems, and Computers | 978-1-6654-5828-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEEECONF53345.2021.9723212

***Abstract*—We study the robust mean estimation problem in high dimensions, where less than half of the datapoints can be arbitrarily corrupted. Motivated by compressive sensing, we formulate the robust mean estimation problem as the minimization of the $\ell_0$-'norm' of an *outlier indicator vector*, under a second moment constraint on the datapoints. We further relax the $\ell_0$-'norm' to the $\ell_p$-norm ($0 < p \leq 1$) in the objective and prove that the global minima for each of these objectives are order-optimal for the robust mean estimation problem. Then we propose a computationally tractable iterative $\ell_p$-minimization and hard thresholding algorithm based on the proposed optimization problems. Empirical studies demonstrate that the proposed algorithm outperforms state-of-the-art robust mean estimation methods.**

## I. Introduction

Robust mean estimation in high dimensions has received considerable interest recently, and has found applications in areas such as data analysis (e.g., spectral data in astronomy [1]), outlier detection [2], [3], [4] and distributed machine learning [5], [6], [7]. Classical robust mean estimation methods such as coordinate-wise median and geometric median have error bounds that scale with the dimension of the data [8], which results in poor performance in the high dimensional regime. A notable exception is Tukey's Median [9] that has an error bound that is independent of the dimension, when the fraction of outliers is less than a threshold [10], [11]. However, the computational complexity of Tukey's Median algorithm is exponential in the dimension.

A number of recent papers have proposed polynomial-time algorithms that have dimension independent error bounds under certain distributional assumptions (e.g., bounded covariance or concentration properties). For a recent comprehensive survey on robust mean estimation, we refer the interested readers to [12]. One of the first such algorithms is Iterative Filtering [13], [14], [15], in which one finds the top eigenvector of the sample covariance matrix and removes (or down-weights) the points with large projection scores on that eigenvector, and then repeat this procedure on the rest of points until the top eigenvalue is small. However, as discussed in [4], the drawback of this approach is that it only looks at one direction/eigenvector at a time, and the outliers may not exhibit unusual bias in only one direction or lie in a single cluster.

There are interesting connections between existing methods for robust mean estimation and those used in compressive sensing. The Iterative Filtering algorithm has similarities to greedy Matching Pursuit type compressive sensing algorithm [16]. In the latter algorithm, one finds a single column of sensing matrix $\boldsymbol{A}$ that has largest correlation with the measurements $\boldsymbol{b}$, removes that column and its contribution from $\boldsymbol{b}$, and repeats this procedure on the remaining columns of $\boldsymbol{A}$. Dong et al. [4] proposed a new scoring criterion for finding outliers, in which one looks at multiple directions associated with large eigenvalues of the sample covariance matrix in every iteration of the algorithm. Interestingly, this approach is conceptually similar to Iterative Thresholding techniques in compressive sensing (e.g., Iterative Hard Thresholding [17] or Hard Thresholding Pursuit [18]), in which one simultaneously finds multiple columns of matrix $\boldsymbol{A}$ that are more likely contribute to $\boldsymbol{b}$. Although this type of approach is also greedy, it is more accurate than the Matching Pursuit technique in practice.

A common assumption in robust mean estimation problem is that the fraction of the corrupted datapoints is small. In this paper, we explicitly use this information through the introduction of an *outlier indicator vector* whose $\ell_0$-'norm' we minimize under a second moment constraint on the datapoints. This new formulation not only enables us to leverage advanced compressive sensing techniques to solve the robust mean estimation problem, but also makes it possible for our algorithm to *not* require prior knowledge of the fraction of outliers.

We consider the setting in which the distribution of the datapoints before corruption has bounded covariance, as is commonly assumed in many recent works (e.g., [14], [4], [19], [20]). In particular, in [19], the authors propose to minimize the spectral norm of the weighted sample covariance matrix and use the knowledge of the outlier fraction $\alpha$ to constrain the weights. Along these lines, in two recent works [21], [22] it is shown that any approximate stationary point of the objective

*Equal contribution.

This research was supported by the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 (IOBT CRA), through the University of Illinois.

in [19] gives a near-optimal solution. In contrast, our objective is designed to minimize the sparsity of an *outlier indicator vector*, and we show that *any* sparse enough solution is nearly optimal.

There is another line of related work on mean estimation of heavy tailed distributions. See, e.g. the recent survey article [23] and the references therein. Also, the connection between robust mean estimation and heavy-tailed mean estimation is discussed in [24].

Our contributions are as follows:

- At a fundamental level, a contribution of this paper is the formulation of the robust mean estimation problem as minimizing the $\ell_0$-'norm' of the proposed *outlier indicator vector*, under a second moment constraint on the datapoints. In addition, order-optimal estimation error guarantees and optimal breakdown point are shown for this objective. We relax the $\ell_0$ objective to $\ell_p (0 < p \leq 1)$ as in compressive sensing, and establish corresponding order-optimal estimation error guarantees.
- Motivated by the proposed $\ell_0$ and $\ell_p$ objectives and their theoretical justifications, we propose a computationally tractable *iterative* $\ell_p (0 < p \leq 1)$ minimization and hard thresholding algorithm. Empirical studies show that the proposed algorithm significantly outperforms state-of-the-art methods in robust mean estimation in high dimensions.

## II. Proposed optimization problems

We begin by defining what we mean by a corrupted sample of datapoints.

**Definition 1.** ($\alpha$-corrupted sample [4]) Let $P$ be a distribution on $\mathbb{R}^d$ with unknown mean $\boldsymbol{\mu}$, and let $\tilde{\boldsymbol{y}}_1, ..., \tilde{\boldsymbol{y}}_n$ be independent and identically distributed (i.i.d.) drawn from $P$. These datapoints are then modified by an adversary who can inspect all the datapoints, remove $\alpha n$ of them, and replace them with arbitrary vectors in $\mathbb{R}^d$. We then obtain an $\alpha$-corrupted sample, denoted as $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$.

Our primary goal is to robustly estimate the true population mean, given an $\alpha$-corrupted sample. A key insight exploited in previous works on the problem is that it suffices to find a large subset of the $\alpha$-corrupted sample, whose sample covariance matrix has bounded spectral norm. In order for such a subset to exist and for the mean of this large subset to be close to the true mean, we need some form of concentration of the datapoints (before corruption) around the mean of their distribution. A constrained second moment condition is sufficient to guarantee this, and this assumption is also used in previous works.

Based on this motivation, we propose an $\ell_0$-minimization problem to find the largest subset, whose sample covariance matrix is close to the covariance matrix of the underlying distribution. Let the datapoints before corruption be generated from a distribution whose covariance matrix is bounded: $\Sigma \preceq \sigma^2 I$. We first introduce an *outlier indicator vector* $\boldsymbol{h}$: for the $i$-th datapoint, $h_i$ indicates that whether it is an outlier ($h_i = 1$) or not ($h_i = 0$). Given an $\alpha$-corrupted sample of size $n$, we propose the following optimization problem:

$$\min_{\boldsymbol{h},\boldsymbol{x}} \|\boldsymbol{h}\|_0 \quad s.t.\ h_i \in \{0,1\}, \forall i, \qquad \left\| \sum_{i=1}^{n} (1-h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top \right\|_2 \leq c_1^2 \sigma^2 n. \tag{1}$$

We further relax the problem to the following:

$$\min_{\boldsymbol{h},\boldsymbol{x}} \|\boldsymbol{h}\|_0 \quad s.t.\ 0 \leq h_i \leq 1, \forall i, \qquad \left\| \sum_{i=1}^{n} (1-h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top \right\|_2 \leq c_1^2 \sigma^2 n. \tag{2}$$

Note that any globally optimal solution of (1) is also globally optimal solution of (2). We show in Theorem 1, that any sparse enough feasible solution including the global optimum of (2) achieves order-optimality.

However, the above $\ell_0$ objective is not computationally tractable. Motivated by compressive sensing, we further propose to relax the $\ell_0$-'norm' to the $\ell_p$-norm ($0 < p \leq 1$), which leads to the following optimization problem:

$$\min_{\boldsymbol{h},\boldsymbol{x}} \|\boldsymbol{h}\|_p \quad s.t. 0 \leq h_i \leq 1, \forall i, \qquad \left\| \sum_{i=1}^{n} (1-h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top \right\|_2 \leq c_1^2 \sigma^2 n. \tag{3}$$

We show in Theorem 2 that even in this case, any 'good' feasible solution including the global optimum is order-optimal.

We now provide theoretical guarantees for the estimators which are given by the solutions of the optimization problems (2) and (3). We show that given an $\alpha$-corrupted sample of size $\Omega\left(\frac{d \log d}{\epsilon}\right)$, with high probability, the $\ell_2$-norm of both estimators' error is bounded by $O\left(\sigma \sqrt{\frac{\alpha+\epsilon}{1-2(\alpha+\epsilon)}}\right)$. We formalize this in the following theorems. The parameter $\epsilon$ is independent of the fraction of outliers $\alpha$ and controls the tradeoff between the error bound and the number of datapoints required. It is well known that an information-theoretic lower bound on the $\ell_2$-norm of any estimator's error $\|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_2$ is $\Omega\left(\sigma \sqrt{\frac{\alpha}{1-2\alpha}}\right)$. By setting $\epsilon = O(\alpha)$, we see that the estimators are information-theoretically order-optimal.

**Theorem 1.** *Let $P$ be a distribution on $\mathbb{R}^d$ with unknown mean $\boldsymbol{\mu}$ and unknown covariance matrix $\Sigma \preceq \sigma^2 I$. Let $0 < \epsilon < 1/2$, $0 < \delta < 1/4$ and $c_1 > 1$ be fixed. Let $0 < \alpha < 1/2 - \epsilon$. Given an $\alpha$-fraction corrupted set of $n \geq \frac{ed}{\epsilon \delta^2 c_1'} \log\left(\frac{d}{\delta}\right)$ datapoints from $P$, let*

$$\mathcal{S} = \left\{ (\boldsymbol{h}, \boldsymbol{x}) : \|\boldsymbol{h}\|_0 \leq \alpha' n; \boldsymbol{x} = \frac{\sum_{\{i:h_i=0\}} \boldsymbol{y}_i}{|\{i : h_i = 0\}|} \right\}, \tag{4}$$

*where $c_1' = c_1^2 \min\{c_1^2 \log c_1^2 + 1 - c_1^2, 1\}$, $\alpha' = \alpha + \epsilon$.*

*Then the following holds with probability at least* $1-4\delta$*:*
*1) Any feasible pair* $(\hat{\boldsymbol{h}},\hat{\boldsymbol{x}})$ *for the optimization problem* (2) *such that* $(\hat{\boldsymbol{h}},\hat{\boldsymbol{x}})\in\mathcal{S}$ *satisfies*

$$\|\hat{\boldsymbol{x}}-\boldsymbol{\mu}\|_2 \le \sqrt{\frac{c_1^2\sigma^2}{1-\epsilon}.\frac{\alpha}{1-\alpha}} + c_3\sigma + \left(\sqrt{\frac{c_1^2\sigma^2}{1-\alpha'}}+\sqrt{\frac{c_1^2\sigma^2}{1-\frac{\|\hat{\boldsymbol{h}}\|_0}{n}}}\right)\sqrt{\frac{\alpha'}{1-\alpha'-\frac{\|\hat{\boldsymbol{h}}\|_0}{n}}}, \quad (5)$$

*where* $c_3=\sqrt{\epsilon\delta}\Big(1+2\sqrt{\frac{c_1'}{e\log(d/\delta)}}\Big)$.
*2) A global optimum of* (2) *lies in* $\mathcal{S}$.

We now provide a similar order-optimal error guarantee for the solution of the optimization problem in (3).

**Theorem 2.** *Let* $P$ *be a distribution on* $\mathbb{R}^d$ *with unknown mean* $\boldsymbol{\mu}$ *and unknown covariance matrix* $\Sigma\preceq\sigma^2 I$*. Let* $0<p\le 1$*,* $0<\epsilon<1/2$*,* $0<\delta<1/4$ *and* $c_1>1$ *be fixed. Let* $0<\alpha<1/2-\epsilon$*. Given an* $\alpha$*-fraction corrupted set of* $n\ge\frac{ed}{\epsilon\delta^2 c_1'}\log\big(\frac{d}{\delta}\big)$ *datapoints from* $P$*, let*

$$\mathcal{S}'=\left\{(\boldsymbol{h},\boldsymbol{x}):\|\boldsymbol{h}\|_p^p\le\alpha' n;\ \ \boldsymbol{x}=\frac{\sum_{i=1}^n(1-h_i)\boldsymbol{y}_i}{\sum_{i=1}^n(1-h_i)}\right\},$$

*where* $c_1'=c_1^2\min\big\{c_1^2\log c_1^2+1-c_1^2,1\big\}$*,* $\alpha'=\alpha+\epsilon$.
*Then the following holds with probability at least* $1-4\delta$*:*

1) *Any feasible solution* $(\hat{\boldsymbol{h}},\hat{\boldsymbol{x}})$ *of* (3) *such that* $(\hat{\boldsymbol{h}},\hat{\boldsymbol{x}})\in\mathcal{S}'$ *satisfies*

$$\|\hat{\boldsymbol{x}}-\boldsymbol{\mu}\|_2 \le \sqrt{\frac{c_1^2\sigma^2}{1-\epsilon}.\frac{\alpha}{1-\alpha}} + c_3\sigma + \left(\sqrt{\frac{c_1^2\sigma^2}{1-\alpha'}}+\sqrt{\frac{c_1^2\sigma^2}{1-\frac{\|\hat{\boldsymbol{h}}\|_p^p}{n}}}\right)\sqrt{\frac{\alpha'}{1-\alpha'-\frac{\|\hat{\boldsymbol{h}}\|_p^p}{n}}}, \quad (6)$$

*where* $c_3=\sqrt{\epsilon\delta}\Big(1+2\sqrt{\frac{c_1'}{e\log(d/\delta)}}\Big)$.
2) *A global optimum of* (3) *lies in* $\mathcal{S}'$.

**Remark 1.** Theorems 1 and 2 show that, as long as we find a feasible solution $\hat{\boldsymbol{h}}$ whose norm is small enough, e.g., $\|\hat{\boldsymbol{h}}\|_0\le\alpha' n$, the corresponding $\hat{\boldsymbol{x}}$ is close to the true mean. It is not necessary to reach the global optimum of the objectives (2) and (3).

**Remark 2.** The breakdown point of the estimators in Theorem 1 and 2 is nearly the maximal possible $1/2$ (as $\epsilon\to 0$ and $n\to\infty$), that is, the estimator can tolerate any corruption level $\alpha<1/2$, assuming the number of samples $n$ satisfies the lower bound.

A high-level sketch of the proofs of Theorems 1 and 2 is as follows. We use the idea in [22, Lemma 2.2]. Informally, if two probability distributions on a set a datapoints are close in total variation distance, then the weighted means of the distribution are close. For Theorem 1, we consider the uniform distribution on the set $\{\boldsymbol{y}_i:\hat{h}_i=0\}$ (say $P_1$). For Theorem 2, we consider the distribution on the $\alpha$-corrupted samples with (normalized) probability weights $1-h_i$ (say $P_2$). Note that the estimator $\hat{\boldsymbol{x}}$ in Theorem 1 is the mean of $P_1$, and similarly, the estimator in Theorem 2 is the mean of $P_2$. We show that for both $i=1,2$, the total variation distance between $P_i$ and the uniform distribution (say $P_3$) on the set of inlier datapoints (that are within a distance of $\sigma\sqrt{\frac{d}{\epsilon\delta}}$ from $\boldsymbol{\mu}$), is small. Therefore one can show that the distance between $\hat{\boldsymbol{x}}$ and the mean of $P_3$ is $O\left(\sigma\sqrt{\frac{\alpha+\epsilon}{1-2(\alpha+\epsilon)}}\right)$. Using the same result in [22, Lemma 2.2], we show that the distance between the mean of $P_3$ and $\boldsymbol{\mu}$ is $O(\sigma\sqrt{\alpha+\epsilon})$. Using triangle inequality, we show that the distance between $\hat{\boldsymbol{x}}$ and $\boldsymbol{\mu}$ is $O\left(\sigma\sqrt{\frac{\alpha+\epsilon}{1-2(\alpha+\epsilon)}}\right)$.

Observe that in Theorems 1 and 2, $\epsilon$ controls the error tolerance level. Also, the lower bound on the required number of datapoints is $\Omega\left(\frac{d\log d}{\epsilon}\right)$, which is independent of the corruption level $\alpha$. Previous works (see, e.g., [13], [14], [19]) do not consider a tolerance level, and in these works the lower bound on the required number of datapoints is inversely proportional to the corruption level $\alpha$, which blows up as $\alpha\to 0$. Moreover, $\alpha$ is typically unknown in practice. Specifying $\epsilon$ to control the estimator's error helps us remove the dependence of the number of datapoints required on the fraction of corruption $\alpha$. Note that we can recover the order-optimal results in the form as given in the previous works by setting $\epsilon=O(\alpha)$ in Theorems 1 and 2,

## III. Computationally Tractable Algorithm

### A. $\ell_p$ minimization and thresholding

Motivated by the $\ell_p$ objective and its theoretical guarantees, we propose an iterative $\ell_p$ minimization algorithm. The algorithm alternates between updating the outlier indicator vector $\boldsymbol{h}$ via minimizing its $\ell_p$-norm and updating the estimated mean $\boldsymbol{x}$ (see Algorithm 1).

When updating the estimated mean $\boldsymbol{x}$ in Step 2 of Algorithm 1, we add an option to threshold the $h_i$ by $\tau$, so one can use the weighted average of the estimated 'reliable' datapoints (i.e., those for which $h_i\approx 0$) to estimate $\boldsymbol{x}$. This is motivated by the analysis of the original $\ell_0$ objective in Theorem 1, where the average of the estimated 'reliable' datapoints $\frac{\sum_{\{i:\hat{h}_i=0\}}\boldsymbol{y}_i}{|\{i:\hat{h}_i=0\}|}$ is close to the true mean as long as the outlier indicator vector $\hat{\boldsymbol{h}}$ is sparse enough. We now define some notations used in the description of Algorithm 1. Let

$$f(\tau)=\frac{3\tau+\tau^2-\sqrt{\tau^4+2\tau^3+5\tau^2}}{2(1+\tau)} \quad (7)$$

$$\gamma(\alpha)=\sqrt{\frac{\alpha/\tau}{(1-\alpha/\tau)(1-\alpha-\alpha/\tau)}} \quad (8)$$

$$\beta(\alpha)=\left(\frac{c_1}{\sqrt{1-\alpha/\tau}}+\frac{c_1}{\sqrt{1-\alpha}}\right)\sqrt{\frac{\alpha/\tau}{1-\alpha-\alpha/\tau}}. \quad (9)$$

**Algorithm 1** Robust Mean Estimation via $\ell_p$ Minimization and Thresholding

**Inputs:**
1) An $\alpha$-corrupted set of datapoints $\{\boldsymbol{y}_i\}_{i=1}^n \in \mathbb{R}^d$ generated by a distribution whose covariance matrix satisfies $\Sigma \preceq \sigma^2 I$.
2) Upper bound on corruption level: $\check{\alpha}$
3) Upper bound on spectral norm of $\Sigma$: $\sigma^2$.
4) Threshold: $0 < \tau \leq 1$ such that $f(\tau) > \check{\alpha}$, where $f(\tau)$ is defined in (7).
5) Set $c_1 > 1$.
6) Set $0 < p \leq 1$ in $\ell_p$.
**Initialize:**
1) $\boldsymbol{x}^{(0)}$ = coordinate-wise median of $\{\boldsymbol{y}_i\}_{i=1}^n$.
2) $c_2^{(0)} = 3\sqrt{d} + 2c_1$.
3) Iteration number $t = 0$.
**while** $t < T = 1 + \frac{\log c_2^{(0)}}{\log|\gamma(\check{\alpha})|}$ and $c_2^{(0)} \geq \frac{\beta(\check{\alpha})}{1-\gamma(\check{\alpha})}$ **do**
*Step 1*: Given $\boldsymbol{x}^{(t)}$, update $\boldsymbol{h}$:
$\boldsymbol{h}^{(t)} \in \mathcal{H}(\boldsymbol{x}^{(t)}, c_2^{(t)})$, where $\mathcal{H}$ is defined in (10).
*Step 2*: Given $\boldsymbol{h}^{(t)}$, update $\boldsymbol{x}$:
$\boldsymbol{x}^{(t+1)} = \frac{\sum_{i=1}^n (1-h_i^{(t)}) 1\{h_i^{(t)} \leq \tau\} \boldsymbol{y}_i}{\sum_{i=1}^n (1-h_i^{(t)}) 1\{h_i^{(t)} \leq \tau\}}$.
$c_2^{(t+1)} = \sigma(\gamma(\check{\alpha}) c_2^{(t)} + \beta(\check{\alpha}))$,
where $\gamma$ and $\beta$ are defined in (8) and (9)
$t = t + 1$.
**end while**
**Output:** $\boldsymbol{x}^{(T)}$

Let $\mathcal{H}$ be the set defined by

$$\begin{aligned} &\mathcal{H}(\boldsymbol{x}, c_2) \coloneqq \arg\min_{\boldsymbol{h}} \|\boldsymbol{h}\|_p \qquad (10)\\ &s.t. \quad 0 \leq h_i \leq 1, \forall i, \\ &\left\| \sum_{i=1}^n (1-h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top \right\|_2 \leq (c_1^2 + c_2^2)\sigma^2 n. \end{aligned}$$

Theoretical guarantees for Algorithm 1 are out of scope of this work. We plan to establish such guarantees in future work.

**Remark 3.** The initialization $c_2^{(0)} = 3\sqrt{d} + 2c_1$ can be replaced by a smaller value as long as it is possible to guarantee $\|\boldsymbol{x}^{(0)} - \boldsymbol{\mu}\| \leq c_2^{(0)}\sigma$ with high probability.

### *B. Solving Step 1 of Algorithm 1*

When we set $p = 1$ in the objective $\|\boldsymbol{h}\|_p$ in Step 1 of Algorithm 1, the resulting problem is convex, and can be reformulated as the following packing SDP [25] with $w_i \triangleq 1 - h_i$, $A_i = (\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top$ and $e_i$ being the $i$-th standard basis vector in $\mathbb{R}^n$:

$$\begin{aligned} &\max_{\boldsymbol{w}} \mathbf{1}^\top \boldsymbol{w} \qquad (11)\\ &s.t. \quad w_i \geq 0, \forall i \\ &\sum_{i=1}^n w_i \begin{bmatrix} e_i e_i^\top & \\ & A_i \end{bmatrix} \preceq \begin{bmatrix} I_{n\times n} & \\ & cn\sigma^2 I_{d\times d} \end{bmatrix}. \end{aligned}$$

When $0 < p < 1$, the equivalent objective function $\|\boldsymbol{h}\|_p^p = \sum_i h_i^p$ is concave, *not* convex. So it may be difficult to find its global minimum. Nevertheless, we can iteratively construct and minimize a *tight* upper bound on this objective function via iterative re-weighted $\ell_2$ [26], [27] or $\ell_1$ techniques [28] from compressive sensing.[1] And it is well-known in compressive sensing that such iterative re-weighted approaches often performs better than $\ell_1$ [28], [26].

## IV. Empirical Studies

In this section, we compare performance of Algorithm 1 with the following state-of-the-art high dimensional robust mean estimation methods: iterative filtering algorithm [14] (denoted as DKK), method proposed in [22] (denoted as ZJS), method proposed in [8] (denoted as LRV), method in [19] (denoted as CDG), and Quantum Entropy Scoring (QUE) [4]. We fix $p = 0.5$ for the proposed $\ell_p$ method. In Algorithm 1, we set the threshold $\tau = 0.6$, the constant $c_1 = 1.1$, the upper bound on corruption level $\check{\alpha} = \frac{161}{160}\alpha$ and we initialize $c_2^{(0)}$ as the $\ell_2$ error of the Coordinate-wise Median relative to the true mean. We carefully tune the parameters in the compared methods. For evaluation, we define the recovery error as the $\ell_2$ distance of the estimated mean to the oracle solution, i.e., the average of the datapoints before corruption.

We use a similar experiment setting as in [4]. The dimension of the data is $d$, and the number of datapoints is $n$. There are two clusters of outliers, and their $\ell_2$ distances to the true mean $\boldsymbol{x}$ are similar to that of the inlier points. The inlier datapoints are randomly generated from the standard Gaussian distribution with zero mean. For the outliers, half of them are set to be $[\sqrt{d/2}, \sqrt{d/2}, 0, ..., 0]$, and the other half are set as $[\sqrt{d/2}, -\sqrt{d/2}, 0, ..., 0]$, so that their $\ell_2$ distances to the true mean $[0, ..., 0]$ are all $\sqrt{d}$, similar to that of the inlier points. We vary the total fraction $\alpha$ of the outliers and report the average recovery error of each method over 10 trials in Table I with $d = 100, n = 1000$. The proposed $\ell_1$ and $\ell_p$ methods show significant improvements over the competing methods, and the $\ell_p$ method performs the best.

TABLE I
Recovery error of each method under different fraction $\alpha$ of the outlier points ($d = 100, n = 1000$)

| $\alpha$ | DKK | ZJS | QUE | LRV | CDG | $\ell_1$ | $\ell_p$ |
|---|---|---|---|---|---|---|---|
| 10% | 0.124 | 0.098 | 0.429 | 0.367 | 0.064 | **0.013** | **0.006** |
| 20% | 0.131 | 0.115 | 0.492 | 0.659 | 0.084 | **0.013** | **0.007** |

We also tested the performance of each method for different numbers of datapoints. The dimension of the data is fixed to be 100. The fraction of the corrupted points is fixed to be 20%. We vary the number of datapoints from 100 to 1000, and report the average recovery error for each method over 50 trials in Table II. We can see that the performance of all methods get better when the number of datapoints is increased. Again, the proposed methods consistently perform better than the other methods.

[1] We observe that iterative re-weighted $\ell_2$ achieves better empirical performance.

TABLE II
RECOVERY ERROR OF EACH METHOD W.R.T. DIFFERENT NUMBER OF SAMPLES ($d = 100, \alpha = 0.2$)

| $n$ | DKK | ZJS | QUE | LRV | CDG | $\ell_1$ | $\ell_p$ |
|---|---|---|---|---|---|---|---|
| 100 | 0.493 | 0.293 | 1.547 | 1.423 | 0.316 | **0.060** | **0.033** |
| 200 | 0.313 | 0.239 | 1.038 | 1.084 | 0.198 | **0.036** | **0.021** |
| 500 | 0.186 | 0.170 | 0.680 | 0.794 | 0.148 | **0.021** | **0.012** |
| 1000 | 0.131 | 0.115 | 0.492 | 0.659 | 0.084 | **0.013** | **0.007** |

## V. CONCLUSION

We formulated the robust mean estimation problem as the minimization of the $\ell_0$-'norm' of the introduced outlier indicator vector, under a second moment constraint on the datapoints. We further relaxed the $\ell_0$ objective to $\ell_p$ $(0 < p \leq 1)$ and theoretically justified the new objective. We also proposed a computationally tractable iterative $\ell_p (0 < p \leq 1)$ minimization and hard-thresholding algorithm, Algorithm 1, which significantly outperforms state-of-the-art robust mean estimation methods. In empirical studies, we observed strong numerical evidence that $\ell_p$ $(0 < p \leq 1)$ leads to sparse solutions; theoretically justifying this phenomenon is of interest and is an avenue for future research. The proposed $\ell_0$ and $\ell_p$ optimization problems do not require knowledge of $\alpha$, and their solutions are order optimal in terms of error. While we do not currently have computationally tractable algorithms to solve the $\ell_0$ and $\ell_p$ optimization problems, we conjecture that it should be possible to design computationally tractable algorithms that do not require prior knowledge of $\alpha$ and are order optimal in terms of error. The proposed Algorithm 1 is one such algorithm, and establishing theoretical guarantees for this algorithm is a direction of research that we are currently pursuing.

## VI. ACKNOWLEDGEMENTS

The authors would like to thank Akshayaa Magesh for fruitful discussions.

## REFERENCES

[1] R. A. Maronna and R. H. Zamar, "Robust estimates of location and dispersion for high-dimensional datasets," *Technometrics*, vol. 44, no. 4, pp. 307–317, 2002.

[2] P. J. Huber, *Robust statistics*. Springer, 2011.

[3] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. Wiley, 2018.

[4] Y. Dong, S. Hopkins, and J. Li, "Quantum entropy scoring for fast robust mean estimation and improved outlier detection," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 6067–6077.

[5] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," in *Proc. ACM Measurement and Analysis of Computing Systems*, vol. 1, no. 2. ACM New York, NY, USA, 2017, pp. 1–25.

[6] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*, 2018, pp. 5650–5659.

[7] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7711–7717, 2013.

[8] K. A. Lai, A. B. Rao, and S. Vempala, "Agnostic estimation of mean and covariance," in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 2016, pp. 665–674.

[9] J. W. Tukey, "Mathematics and the picturing of data," in *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, vol. 2, 1975, pp. 523–531.

[10] D. L. Donoho, M. Gasko *et al.*, "Breakdown properties of location estimates based on halfspace depth and projected outlyingness," *The Annals of Statistics*, vol. 20, no. 4, pp. 1803–1827, 1992.

[11] B. Zhu, J. Jiao, and J. Steinhardt, "When does the tukey median work?" *arXiv preprint arXiv:2001.07805*, 2020.

[12] I. Diakonikolas and D. M. Kane, "Recent advances in algorithmic high-dimensional robust statistics," *arXiv preprint arXiv:1911.05911*, 2019.

[13] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high dimensions without the computational intractability," in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 2016, pp. 655–664.

[14] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Being robust (in high dimensions) can be practical," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 999–1008.

[15] J. Steinhardt, "Robust learning: Information theory and algorithms," Ph.D. dissertation, Stanford University, 2018.

[16] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, pp. 3397–3415, 1993.

[17] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[18] S. Foucart, "Hard thresholding pursuit: an algorithm for compressive sensing," *SIAM Journal on Numerical Analysis*, vol. 49, no. 6, pp. 2543–2563, 2011.

[19] Y. Cheng, I. Diakonikolas, and R. Ge, "High-dimensional robust mean estimation in nearly-linear time," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '19. USA: Society for Industrial and Applied Mathematics, 2019, p. 2755–2771.

[20] J. Steinhardt, M. Charikar, and G. Valiant, "Resilience: A criterion for learning in the presence of arbitrary outliers," *arXiv preprint arXiv:1703.04940*, 2017.

[21] Y. Cheng, I. Diakonikolas, R. Ge, and M. Soltanolkotabi, "High-dimensional robust mean estimation via gradient descent," *arXiv preprint arXiv:2005.01378*, 2020.

[22] B. Zhu, J. Jiao, and J. Steinhardt, "Robust estimation via generalized quasi-gradients," *arXiv preprint arXiv:2005.14073*, 2020.

[23] G. Lugosi and S. Mendelson, "Mean estimation and regression under heavy-tailed distributions–a survey," *arXiv preprint arXiv:1906.04280*, 2019.

[24] S. B. Hopkins, J. Li, and F. Zhang, "Robust and heavy-tailed mean estimation made simple, via regret minimization," in *Advances in Neural Information Processing Systems 33*, 2020.

[25] G. Iyengar, D. J. Phillips, and C. Stein, "Approximation algorithms for semidefinite packing problems with applications to maxcut and graph coloring," in *International Conference on Integer Programming and Combinatorial Optimization*. Springer, 2005, pp. 152–166.

[26] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 3869–3872.

[27] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[28] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.