Robust Mean Estimation in High Dimensions: An Outlier-Fraction Agnostic and Efficient Algorithm

Aditya Deshmukh*, *Student Member, IEEE*, Jing Liu*, *Member, IEEE*, Venugopal V. Veeravalli, *Fellow, IEEE*

Abstract

The problem of robust mean estimation in high dimensions is studied, in which a certain fraction (less than half) of the datapoints can be arbitrarily corrupted. Motivated by compressive sensing, the robust mean estimation problem is formulated as the minimization of the ℓ_0-'norm' of an *outlier indicator vector*, under a second moment constraint on the datapoints. The ℓ_0-'norm' is then relaxed to the ℓ_p-norm $(0 < p \le 1)$ in the objective, and it is shown that the global minima for each of these objectives are order-optimal and have optimal breakdown point for the robust mean estimation problem. Furthermore, a computationally tractable iterative ℓ_p-minimization and hard thresholding algorithm is proposed that outputs an order-optimal robust estimate of the population mean. The proposed algorithm (with breakdown point ≈ 0.3) does not require prior knowledge of the fraction of outliers, in contrast with most existing algorithms, and for $p = 1$ it has near-linear time complexity. Both synthetic and real data experiments demonstrate that the proposed algorithm outperforms state-of-the-art robust mean estimation methods.

Index Terms

Robust estimation, High-dimensional statistics, Global outlier pursuit, Linear time complexity algorithm

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.

*Equal contribution.

A. Deshmukh and V.V. Veeravalli are with the ECE Department and Coordinated Science Lab, University of Illinois at Urbana-Champaign, Illinois, USA. Email: ad11,vvv@illinois.edu

J. Liu is with the Mitsubishi Electric Research Laboratories, Massachusetts, USA. Email: jiliu@merl.com

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. Figure 1 illustrates an example for which Iterative Filtering might have poor empirical performance. In this figure, the inlier datapoints in blue are randomly generated from the standard Gaussian distribution in (high) dimension d, and therefore their ℓ_2-distances to the origin are roughly $\sqrt{d}$ (see, e.g., Theorem 3.1 of [16]). There are two clusters of outliers in red, and their ℓ_2-distances to the origin are also roughly $\sqrt{d}$. If there is only one cluster of outliers, Iterative Filtering can effectively identify them; however, in this example, this method may remove many inlier points and perform suboptimally.
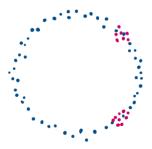

Fig. 1: Illustration of two clusters of outliers (red points). The inlier points (blue) are drawn from standard Gaussian distribution in high dimension d. Both the outliers and inliers are at roughly $\sqrt{d}$ distance from the origin.

There are interesting connections between existing methods for robust mean estimation and those used in compressive sensing. The Iterative Filtering algorithm has similarities to the greedy Matching Pursuit compressive sensing algorithm [17]. 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 criteria 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 multi-directional approach is conceptually similar to Iterative Thresholding techniques in compressive sensing (e.g., Iterative Hard Thresholding [18] or Hard Thresholding Pursuit [19]), in which one simultaneously finds multiple columns of matrix $\boldsymbol{A}$ that are more likely contribute to $\boldsymbol{b}$. Although iterative thresholding techniques are also greedy, they are more accurate than the Matching Pursuit technique in practice [20], [21].

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 ℓ_0-'norm' we minimize under a second moment constraint on the datapoints. This is partially motivated by

compressive sensing and shares the same principle of 'fitting the majority of the data' that is common in robust statistics. This new formulation not only enables us to leverage advanced compressive sensing techniques to solve the robust mean estimation problem, but also allow us to design algorithms that do not require prior knowledge of the fraction of outliers. There are some works in sparse recovery (see, e.g. [22], [23]), in which ℓ_0/ℓ_p minimization is used to remove outliers in data. In these works, a linear model $y = Ax + e$ is considered, wherein y denotes the measurements, the matrix A is known, and the unknown sparse vector e models the potential outlier corruption on each datapoint. Consequently, the analyses in the works on sparse recovery methods heavily rely on the assumption that the underlying model is linear (e.g., some works exploit the range-space/null-space properties of the matrix A). On the other hand, in robust mean estimation, a general observation model (not necessarily linear) is considered. In light of this, the analyses in the works on sparse recovery cannot be transferred in an obvious way to the robust mean estimation problem.

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], [24], [25]). In particular, in [24], the authors propose to minimize the spectral norm of the weighted sample covariance matrix and use the knowledge of the outlier-fraction ϵ to constrain the weights. Along this line, two very recent works [26], [27] show that any approximate stationary point of the objective in [24] 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.

Contributions:

- At a fundamental level, a contribution of this paper is the formulation of the robust mean estimation problem as minimizing the ℓ_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 ($\epsilon < 1/2$) are shown for this objective. We relax the ℓ_0 objective to $\ell_p(0 < p \leq 1)$ as in compressive sensing, and establish corresponding order-optimal estimation error guarantees. The guarantees are order-optimal with respect to the number of datapoints(n), dimension of the data (d), and the fraction of corrupted datapoints(ϵ). Henceforth we use the term 'order-optimal' in this sense.

- Motivated by the proposed ℓ_0 and ℓ_p objectives and their theoretical justifications, we propose a computationally tractable *iterative* $\ell_p(0 < p \leq 1)$ minimization and hard thresholding algorithm, and establish the order optimality of the algorithm. Empirical studies show that the proposed algorithm significantly outperforms state-of-the-art methods in robust mean estimation.

- The proposed algorithm (with maximal breakdown point of $1 - 1/\sqrt{2}$) does not require the knowledge of the fraction of outliers (in contrast to most existing algorithms). For $p = 1$, the algorithm has near-linear time complexity.

II. PROPOSED OPTIMIZATION PROBLEMS

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

Definition 1. (*ϵ-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 ϵn of them, and replace them with arbitrary vectors in $\mathbb{R}^d$. We then obtain an ϵ-corrupted sample, denoted as $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$.

Throughout the rest of the paper, we adhere to the notation given above: we represent a datapoint before corruption as $\tilde{\boldsymbol{y}}_i$, and after corruption as $\boldsymbol{y}_i$. Given a set of datapoints $\{\boldsymbol{x}_i, i = 1, \ldots, n\}$, we term the following as *sample covariance matrix around $\boldsymbol{z}$*:

$$\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{z})(\boldsymbol{x}_i - \boldsymbol{z})^{\top}. \tag{1}$$

There are other types of contamination one can consider, e.g., Huber's ϵ-contamination model [28]. The contamination model described in Definition 1 is the strongest in the sense that the adversary is not oblivious to the original datapoints, and can replace any subset of ϵn datapoints with any vectors in $\mathbb{R}^d$. We refer the reader to [12] for a more detailed discussion on contamination models.

Our primary goal is to robustly estimate the true population mean, given an ϵ-corrupted sample. We assume that the underlying distribution has bounded second moment. A powerful and useful key insight that was exploited in previous work on the problem is that if the outliers in an ϵ-corrupted sample (of large size) shift the average of datapoints before corruption by $\Omega(\xi)$ in a direction ν, then the variance of the projected sample along ν increases by $\Omega(\xi^2/\epsilon)$. Thus, intuitively, it suffices to find a large subset of the ϵ-corrupted sample, whose sample covariance matrix is close to the covariance matrix of the underlying distribution. 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 such an assumption is also used in previous works. In the following, we provide a brief high-level explanation (details can be found in the Appendix). Suppose we are given a sufficiently large sample of datapoints of size n, generated from a distribution with mean $\boldsymbol{\mu}$ and spectral norm of the covariance matrix bounded by σ^2. Then, with high probability, there exists a large subset of the sample with spectral norm of the sample covariance matrix around $\boldsymbol{\mu}$ bounded by $O(\sigma^2)$. Hence, after corruption, with high probability there still exists a sufficiently large subset, say $\boldsymbol{G}^*$, of the resulting ϵ-corrupted sample, of size $(1 - \epsilon')n$ (where $\epsilon' \to \epsilon$ as $n \to \infty$), such that the spectral norm of the sample covariance matrix around $\boldsymbol{\mu}$ is bounded by $O(\sigma^2)$. Utilizing this, the concentration of the sample before corruption around $\boldsymbol{\mu}$, and a fundamental result [27, Lemma C.2] about closeness of population mean and conditional mean, it can be shown that the distance between $\boldsymbol{\mu}$ and the sample average of $\boldsymbol{G}^*$ is $O(\sigma\sqrt{\epsilon'})$.

Based on this motivation, we propose an ℓ_0-minimization problem to find the largest subset, whose sample covariance matrix exhibits bounded spectral norm. 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 ϵ-corrupted sample of size n, we propose the following optimization problem, for which the solution in $\boldsymbol{x}$ should yield a robust estimate of the mean:

$$\min_{\boldsymbol{h}, \boldsymbol{x}} \|\boldsymbol{h}\|_0 \quad s.t. \ h_i \in \{0, 1\}, \forall i, \tag{2}$$

$$\lambda_{\max}\left(\sum_{i=1}^{n}(1 - h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^{\top}\right) \leq c_1^2\sigma^2 n,$$

where c_1 is a constant that controls the inflation of the constraint with respect to the bound (σ^2) on the spectral norm of the covariance matrix of the underlying distribution.

We further relax the problem to the following:

$$\min_{\boldsymbol{h},\boldsymbol{x}} \|\boldsymbol{h}\|_0 \quad s.t.\ 0 \le h_i \le 1, \forall i, \tag{3}$$

$$\lambda_{\max}\left(\sum_{i=1}^{n}(1-h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \le c_1^2\sigma^2 n.$$

Note that any globally optimal solution of (2) is also globally optimal solution of (3). To see this, let $\tilde{\boldsymbol{h}}$ be a global optimum of (3). Let $\boldsymbol{h}'$ be the vector obtained after setting the non-zero values of $\tilde{\boldsymbol{h}}$ to 1. Note that $\boldsymbol{h}'$ has the same ℓ_0-norm as $\tilde{\boldsymbol{h}}$, and is also a feasible point of (2). Since the constraint set of (3) is larger than (2), the optimum value of (2) must be greater than or equal to the optimum value of (2). This implies that $\boldsymbol{h}'$ is a global optimum of (2). Hence, the claim holds. We show in Theorem 1, that any sparse enough feasible pair including the global optimum of (3) achieves order-optimality in terms of the error in estimating the mean.

However, minimizing the above ℓ_0 objective is not computationally tractable. Motivated by compressive sensing, we further propose to relax the ℓ_0-'norm' to the ℓ_p-norm $(0 < p \le 1)$, which leads to the following optimization problem:

$$\min_{\boldsymbol{h},\boldsymbol{x}} \|\boldsymbol{h}\|_p \quad s.t.\ 0 \le h_i \le 1, \forall i, \tag{4}$$

$$\lambda_{\max}\left(\sum_{i=1}^{n}(1-h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \le c_1^2\sigma^2 n.$$

We show in Theorem 2, that even in this case any 'good' feasible pair including the global optimum is order-optimal in terms of the error in estimating the mean.

In the approaches taken in prior works (see, e.g., [27]), the robust mean estimation problem is the following feasibility problem:

$$\text{Find } h \text{ s.t. } h_i \in [0,1], \sum_{i=1}^{n} h_i \le \epsilon n, \lambda_{\max}(\sum_{i=1}^{n}(1-h_i)(y_i - x)(y_i - x)^\top) \le C\sigma^2 n. \tag{5}$$

Most works (see, e.g., [4], [24]) consider the following problem (or its variant), which is obtained by changing the feasibility problem into the following optimization problem:

$$\min \lambda_{\max}(\sum_{i=1}^{n}(1-h_i)(y_i - x)(y_i - x)^\top) \text{ s.t. } h_i \in [0,1], \sum_{i=1}^{n} h_i \le \epsilon n \tag{6}$$

where x is either fixed or is the weighted average of y_i's with weights as $1 - h_i$. Landscape results related to the optimization problem (6) were obtained in [24] and [27]. Our formulation (4), for the special case of $p = 1$, corresponds to minimizing the feasibility condition related to the sum of "weights" in (5). We provide landscape results for the optimization problem given in (4) (Theorems 1 and 2). An advantage of our formulation, which we will exploit in Algorithm 1, is that it does not require knowledge of the fraction of outliers ϵ.

We now provide theoretical guarantees for the estimator which is given by the solution of the optimization problem (3). We show that given an ϵ-corrupted sample of sufficiently large size, then with high probability, the ℓ_2-norm of

the estimator's error is $O\left(\sigma\sqrt{\frac{\epsilon+\frac{d\log d}{n}}{1-2\left(\epsilon+\frac{d\log d}{n}\right)}}\right)$. We formalize this in the following theorem. It is well known that

an information-theoretic lower bound on the ℓ_2-norm of any estimator's error $\|\hat{x} - \boldsymbol{\mu}\|_2$ is $\Omega\left(\sigma\sqrt{\frac{\epsilon}{1-2\epsilon}}\right)$ (see [27]).

Thus, the estimator is order-optimal in terms of the error as $\alpha \to 0$ and $n \to \infty$.

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*
$\delta \in (0, 1/4)$ *and* $c_1 > 1$ *be fixed. Let* $c_1' = c_1^2 \min\{c_1^2 \log c_1^2 + 1 - c_1^2, 1\}$, $n > \frac{2e}{c_1'\delta^2}d\log(d/\delta)$ *and* $\alpha = \frac{ed\log(d/\delta)}{n\delta^2 c_1'}$.
Let $\epsilon \in (0, 1/2 - \alpha)$ *and* $\epsilon' = \epsilon + \alpha$. *Given an ϵ-fraction corrupted set of n datapoints from P, let*

$$\mathcal{S} = \left\{ (\boldsymbol{h}, \boldsymbol{x}) : \|\boldsymbol{h}\|_0 < (1-\epsilon')n; \boldsymbol{x} = \frac{\sum_{\{i:h_i=0\}} \boldsymbol{y}_i}{|\{i : h_i = 0\}|} \right\}. \tag{7}$$

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 (3) such that $(\hat{\boldsymbol{h}}, \hat{\boldsymbol{x}}) \in \mathcal{S}$ satisfies

$$\|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 \leq \left(\sqrt{\frac{c_1^2\sigma^2}{1-\epsilon'}} + \sqrt{\frac{c_1^2\sigma^2}{1-\frac{\|\hat{\boldsymbol{h}}\|_0}{n}}}\right) \sqrt{\frac{\max\left\{\epsilon', \frac{\|\hat{\boldsymbol{h}}\|_0}{n}\right\}}{1-\epsilon'-\frac{\|\hat{\boldsymbol{h}}\|_0}{n}}} + \sqrt{\frac{c_1^2\sigma^2}{1-\alpha} \cdot \frac{\epsilon}{1-\epsilon}} + \sigma\sqrt{\alpha\delta}\left(1 + 2\sqrt{\frac{c_1'}{e\log(\frac{d}{\delta})}}\right). \tag{8}$$

2) A global optimum $(\boldsymbol{h}^{\text{opt}}, \boldsymbol{x}^{\text{opt}})$ of (3) lies in $\mathcal{S}$ with $\|\boldsymbol{h}^{\text{opt}}\|_0 \leq \epsilon'n$.

The proof is deferred to the Appendix. A high-level sketch of the proof of Theorems 1 is as follows. We use the idea in [27, Lemma 2.2] stated in Lemma 2. Informally, if two probability distributions on a set of datapoints are close in total variation distance, then the weighted means of the distributions are close. Consider the uniform distribution on the set $\{\boldsymbol{y}_i : \hat{h}_i = 0\}$ (say P_1). Note that the estimator $\hat{\boldsymbol{x}}$ in Theorem 1 is the mean of P_1. We show that the total variation distance between P_1 and the uniform distribution (say P') on the set of inlier datapoints (that are within a distance of $\sigma\sqrt{\frac{d}{\alpha\delta}} = \sigma\sqrt{\frac{n\delta c_1'}{e\log(d/\delta)}}$ from $\boldsymbol{\mu}$), is small. Therefore one can show that the distance between $\hat{\boldsymbol{x}}$ and the mean of P' is $O\left(\sigma\sqrt{\frac{\max\left\{\epsilon', \frac{\|\hat{\boldsymbol{h}}\|_0}{n}\right\}}{1-\epsilon'-\frac{\|\hat{\boldsymbol{h}}\|_0}{n}}}\right)$. Using Lemma 2, we show that the distance between the mean of P' and $\boldsymbol{\mu}$ is $O(\sigma\sqrt{\epsilon'})$. Using triangle inequality, it follows that the distance between $\hat{\boldsymbol{x}}$ and $\boldsymbol{\mu}$ is $O\left(\sigma\sqrt{\frac{\max\left\{\epsilon', \frac{\|\hat{\boldsymbol{h}}\|_0}{n}\right\}}{1-\epsilon'-\frac{\|\hat{\boldsymbol{h}}\|_0}{n}}}\right)$.

Remark 1. *Theorem 1 shows that, as long as we find a feasible point $\hat{\boldsymbol{h}}$ that is sparse enough, i.e., $\|\hat{\boldsymbol{h}}\|_0 \leq (\epsilon+\alpha)n$, the average of the estimated inliers $\frac{\sum_{\{i:\hat{h}_i=0\}} \boldsymbol{y}_i}{|\{i:\hat{h}_i=0\}|}$ is close to the true mean in the optimal sense. It is not necessary to reach the global optimum of the objective (3).*

We now provide a similar order-optimal error guarantee for the solution of the optimization problem in (4).

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*
$\delta \in (0, 1/4)$, $c_1 > 1$ *and* $p \in (0, 1]$ *be fixed. Let* $c_1' = c_1^2 \min\{c_1^2 \log c_1^2 + 1 - c_1^2, 1\}$, $n > \frac{2e}{c_1'\delta^2}d\log(d/\delta)$ *and*
$\alpha = \frac{ed\log(d/\delta)}{n\delta^2 c_1'}$. *Let* $\epsilon \in (0, 1/2 - \alpha)$ *and* $\epsilon' = \epsilon + \alpha$. *Given an ϵ-fraction corrupted set of n datapoints from P, let*

$$\mathcal{S}' = \left\{ (\boldsymbol{h}, \boldsymbol{x}) : \|\boldsymbol{h}\|_p^p < (1-\epsilon')n; \ \boldsymbol{x} = \frac{\sum_{i=1}^{n}(1-h_i)\boldsymbol{y}_i}{\sum_{i=1}^{n}(1-h_i)} \right\}. \tag{9}$$

Then the following holds with probability at least $1 - 4\delta$:

1) *Any feasible pair $(\hat{\boldsymbol{h}}, \hat{\boldsymbol{x}})$ of (4) such that $(\hat{\boldsymbol{h}}, \hat{\boldsymbol{x}}) \in \mathcal{S}'$ satisfies*

$$
\|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 \leq \left(\sqrt{\frac{c_1^2 \sigma^2}{1 - \epsilon'}} + \sqrt{\frac{c_1^2 \sigma^2}{1 - \frac{\|\hat{\boldsymbol{h}}\|_p^p}{n}}} \right) \sqrt{\frac{\max\left\{ \epsilon', \frac{\|\hat{\boldsymbol{h}}\|_p^p}{n} \right\}}{1 - \epsilon' - \frac{\|\hat{\boldsymbol{h}}\|_p^p}{n}}} + \sqrt{\frac{c_1^2 \sigma^2}{1 - \alpha} \cdot \frac{\epsilon}{1 - \epsilon}}
$$
$$
+ \sigma \sqrt{\alpha \delta} \left(1 + 2\sqrt{\frac{c_1'}{e \log(\frac{d}{\delta})}} \right). \tag{10}
$$

2) *A global optimum $(\boldsymbol{h}^{\mathrm{opt}}, \boldsymbol{x}^{\mathrm{opt}})$ of (4) lies in $\mathcal{S}'$ with $\|\boldsymbol{h}^{\mathrm{opt}}\|_p^p \leq \epsilon' n$.*

The proof is deferred to the Appendix. The high-level idea is similar to that of the proof of Theorem 1. We consider the distribution on the α-corrupted samples with (normalized) probability weights $1 - h_i$ (say P_2). Note that the estimator $\hat{\boldsymbol{x}}$ in Theorem 2 is the mean of P_2. We show that the total variation distance between P_2 and the uniform distribution (say P') on the set of inlier datapoints (that are within a distance of $\sigma \sqrt{\frac{d}{\alpha \delta}} = \sigma \sqrt{\frac{n \delta c_1'}{e \log(d/\delta)}}$ from $\boldsymbol{\mu}$), is small. Therefore one can show that the distance between $\hat{\boldsymbol{x}}$ and the mean of P' is $O\left(\sigma \sqrt{\frac{\max\left\{ \epsilon', \frac{\|\hat{\boldsymbol{h}}\|_p^p}{n} \right\}}{1 - \epsilon' - \frac{\|\hat{\boldsymbol{h}}\|_p^p}{n}}} \right)$. Using Lemma 2, we show that the distance between the mean of P' and $\boldsymbol{\mu}$ is $O(\sigma \sqrt{\epsilon'})$. Using triangle inequality, it follows that the distance between $\hat{\boldsymbol{x}}$ and $\boldsymbol{\mu}$ is $O\left(\sigma \sqrt{\frac{\max\left\{ \epsilon', \frac{\|\hat{\boldsymbol{h}}\|_p^p}{n} \right\}}{1 - \epsilon' - \frac{\|\hat{\boldsymbol{h}}\|_p^p}{n}}} \right)$.

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

Remark 3. *From Lemma 6 in the Appendix, we know that given any feasible pair of (4) with $\|\hat{\boldsymbol{h}}\|_p \leq (\epsilon' n)^{1/p}$, we have that $\left(\hat{\boldsymbol{h}}, \frac{\sum_{i=1}^{n} (1 - \hat{h}_i) \boldsymbol{y}_i}{\sum_{i=1}^{n} (1 - \hat{h}_i)} \right)$ is also a feasible pair, and therefore it lies in the set $\mathcal{S}'$ defined in (9). Theorem 2 further shows that this weighted average of the datapoints $\frac{\sum_{i=1}^{n} (1 - \hat{h}_i) \boldsymbol{y}_i}{\sum_{i=1}^{n} (1 - \hat{h}_i)}$ is close to the true mean. Again, we note that it is not necessary to reach the global optimum of the objective (4); we only need to find a feasible point $\boldsymbol{h}$ of (4) whose ℓ_p-norm is small enough.*

III. ALGORITHM

A. ℓ_p minimization and thresholding

Motivated by the ℓ_p objective and its theoretical guarantee, we propose an iterative ℓ_p minimization algorithm. The algorithm, which is detailed in Algorithm 1, alternates between updating the outlier indicator vector $\boldsymbol{h}$ via

---

Algorithm 1 Robust Mean Estimation via ℓ_p Minimization and Thresholding

---

Inputs:

1) An ϵ-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: $\check{\epsilon}$

3) Upper bound on spectral norm of Σ: σ^2.

4) Threshold: $0 < \tau \leq 1$ such that $f(\tau) > \check{\epsilon}$, where $f(\tau)$ is defined in (13), if such a τ exists.

5) Set $c_1 > 1$.

6) Set $0 < p \leq 1$ in ℓ_p.

Initialize:

1) $\boldsymbol{x}^{(0)}$ as the coordinate-wise median of $\{\boldsymbol{y}_i\}_{i=1}^n$.

2) $c_2^{(0)} = 3\sqrt{d} + 2c_1$.

3) Iteration number $t = 0$.

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 (11).

Step 2: Given $\boldsymbol{h}^{(t)}$, update $\boldsymbol{x}$:

$\boldsymbol{x}^{(t+1)} = \frac{\sum_{i=1}^n (1-h_i^{(t)})\mathbb{1}\{h_i^{(t)} \leq \tau\}\boldsymbol{y}_i}{\sum_{i=1}^n (1-h_i^{(t)})\mathbb{1}\{h_i^{(t)} \leq \tau\}}$.

$c_2^{(t+1)} = \gamma(\check{\epsilon})c_2^{(t)} + \beta(\check{\epsilon})$,

where γ and β are defined in (14) and (15)

$t = t + 1$.

While: $t < T = 1 + \frac{\log c_2^{(0)}}{\log |\gamma(\check{\epsilon})|}$ and $c_2^{(t)} < c_2^{(t-1)}$

Output: $\boldsymbol{x}^{(T)}$

---

minimizing its ℓ_p-norm and updating the estimated mean $\boldsymbol{x}$. To describe Algorithm 1, let $\mathcal{H}$ be the set defined by

$$\mathcal{H}(\boldsymbol{x}, c_2) := \arg\min_{\boldsymbol{h}} \|\boldsymbol{h}\|_p \tag{11}$$

$$s.t. \quad 0 \leq h_i \leq 1, \forall i,$$

$$\lambda_{\max}\left(\sum_{i=1}^n (1-h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \leq (c_1^2 + c_2^2)\sigma^2 n.$$

When updating the estimated mean $\boldsymbol{x}$ in Step 2 of Algorithm 1, we add an option to threshold the h_i by τ, 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 ℓ_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. The breakdown point of Algorithm 1 depends on the threshold τ and is given by $f(\tau)$ (see (13)). The maximal breakdown point corresponds to no thresholding, i.e., $f(1) = 1 - 1/\sqrt{2}$. Algorithm 1 requires an upper bound $\check{\epsilon}$ on the true fraction of outliers. This upper bound can be set arbitrarily close to (but less than) the breakdown point.

With this intuitive updating rule in Step 2, Algorithm 1 has following order-optimal guarantee.

Theorem 3. *Let P be a distribution on $\mathbb{R}^d$ with unknown mean $\boldsymbol{\mu}$ and unknown covariance matrix $\Sigma \preceq \sigma^2 I$. Let $\delta \in (0, 1/5)$, $c_1 > 1$ and $p \in (0, 1]$ be fixed. Let $\tau \in (0, 1]$, $c'_1 = c_1^2 \min\{c_1^2 \log c_1^2 + 1 - c_1^2, 1\}$, $n > \max\left\{90, \frac{e}{c'_1 \delta^2 f(\tau)} d\right\} \log(d/\delta)$, $\alpha = \frac{ed \log(d/\delta)}{n \delta^2 c'_1}$. Let $\epsilon \geq 0$ be such that $0 < \epsilon' \coloneqq \epsilon + \alpha \leq \check{\epsilon} < f(\tau)$. Given an ϵ-fraction corrupted set of n datapoints from P, with probability at least $1 - 5\delta$, all the iterates of Algorithm 1 (for $t \geq 1$) satisfy*

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{\mu}\|_2 \leq \sigma \left[\gamma(\epsilon') \left(c_2^{(0)} \gamma(\check{\epsilon})^{t-1} + \frac{1 - \gamma(\check{\epsilon})^{t-1}}{1 - \gamma(\check{\epsilon})} \beta(\check{\epsilon}) \right) + \beta(\epsilon') \right] + c_1 \sigma \sqrt{\frac{\epsilon}{(1-\alpha)(1-\epsilon)}} \\ + \sigma \sqrt{\alpha \delta} \left(1 + 2 \sqrt{\frac{c'_1}{e \log(d/\delta)}} \right) \tag{12}$$

where $c_2^{(0)}$ is given in Algorithm 1, and

$$f(\tau) = \frac{3\tau + \tau^2 - \sqrt{\tau^4 + 2\tau^3 + 5\tau^2}}{2(1+\tau)} \tag{13}$$

$$\gamma(\epsilon) = \sqrt{\frac{\epsilon/\tau}{(1 - \epsilon/\tau)(1 - \epsilon - \epsilon/\tau)}} \tag{14}$$

$$\beta(\epsilon) = c_1 \left((1 - \epsilon/\tau)^{-1/2} + (1 - \epsilon)^{-1/2} \right) \sqrt{\frac{\epsilon/\tau}{1 - \epsilon - \epsilon/\tau}}. \tag{15}$$

The output of Algorithm 1 at the end of $T = 1 + \frac{\log c_2^{(0)}}{|\log \gamma(\check{\epsilon})|} = O\left(\frac{\log d}{|\log \check{\epsilon}|} \right)$ (when $c_2^{(0)} \geq \frac{\beta(\check{\epsilon})}{1 - \gamma(\check{\epsilon})}$) or $T = 1$ (when $c_2^{(0)} < \frac{\beta(\check{\epsilon})}{1 - \gamma(\check{\epsilon})}$) iterations is order-optimal:

$$\|\boldsymbol{x}^{(T)} - \boldsymbol{\mu}\|_2 \leq \sigma \left[\gamma(\epsilon') \left(1 + \frac{\beta(\check{\epsilon})}{1 - \gamma(\check{\epsilon})} \right) + \beta(\epsilon') \right] + c_1 \sigma \sqrt{\frac{\epsilon}{(1-\alpha)(1-\epsilon)}} + \sigma \sqrt{\alpha \delta} \left(1 + 2 \sqrt{\frac{c'_1}{e \log(d/\delta)}} \right) \tag{16}$$

$$= O(\sigma \sqrt{\epsilon'}).$$

The proof is deferred to the Appendix, but we briefly discuss the design of the algorithm and the high-level approach. Let $\bar{\boldsymbol{x}}^*$ be the average of the set of inlier datapoints that are within a distance of $\sigma \sqrt{\frac{d}{\alpha \delta}}$ from $\boldsymbol{\mu}$. We use induction to show that $\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^*\| \leq c_2^{(t)} \sigma$. We show in the Appendix that the coordinate-wise median satisfies $\|\boldsymbol{x}^{(0)} - \boldsymbol{\mu}\|_2 \leq c_2^{(0)} \sigma$ with high probability. Firstly, observe that in Step 1 of Algorithm 1, the constraint on the spectral norm of the weighted covariance matrix around $\boldsymbol{x}^{(t)}$ is $\left(c_1^2 + (c_2^{(t)})^2 \right) \sigma^2 n$ instead of $c_1^2 \sigma^2 n$ as in (4). This ensures that with high probability that the optimization problem in Step 1 has a feasible point, and that the optimum solution satisfies $\|\boldsymbol{h}^{(t)}\|_p \leq (\epsilon' n)^{1/p}$. Secondly, we exploit the boundedness of $\|\boldsymbol{h}^{(t)}\|_p$ and the fact that the spectral norm of the weighted covariance matrix around $\boldsymbol{x}^{(t)}$ is bounded (similar to the idea used in Theorem 2), along with some concentration bounds to show that in each iteration the iterate $\boldsymbol{x}^{(t+1)}$ in Step 2 moves closer to μ than $\boldsymbol{x}^{(t)}$. Specifically, we show that $\|\boldsymbol{x}^{(t+1)} - \boldsymbol{\mu}\|_2 \leq \gamma \|\boldsymbol{x}^{(t)} - \boldsymbol{\mu}\|_2 + \beta \sigma \leq (\gamma c_2^{(t)} + \beta)\sigma = c_2^{(t+1)} \sigma$, where $\gamma < 1$. From the proof we can see that it is not necessary to reach the global optimum in Step 1, we only need to find a feasible point whose ℓ_p-norm is small enough.

Remark 4. *The results of Theorems 1, 2 and 3 can be easily extended to establish the estimators' closeness to the average of the datapoints before corruption, $\tilde{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{y}}_i$, using the fact that $\tilde{\boldsymbol{\mu}}$ is close to $\boldsymbol{\mu}$, which is shown in the Appendix (see (32)). We obtain the following extension to the above theorems with the same probability guarantees:*

$$\|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{\mu}}\|_2 \le \|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 + \sigma \sqrt{\frac{d}{n\delta}}. \tag{17}$$

Moreover, it can be also shown that the estimators are close to the average of inliers, that are at most a distance of $\sigma \sqrt{\frac{d}{\alpha\delta}} = \sigma \sqrt{\frac{n\delta c_1'}{e \log(d/\delta)}}$ from $\boldsymbol{\mu}$.

Remark 5. *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 $\|\boldsymbol{x}^{(0)} - \boldsymbol{\mu}\|_2 \le c_2^{(0)}\sigma$ with high probability.*

An important aspect of the proposed algorithm is that it does not require the true fraction of outliers ϵ and is still order-optimal. To the best of our knowledge no other algorithm for our corruption model has this property.

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 [29] with $w_i \triangleq 1 - h_i$, and e_i being the i-th standard basis vector in $\mathbb{R}^n$. The details can be found in the Appendix.

$$\max_{\boldsymbol{w}} \ \mathbf{1}^\top \boldsymbol{w} \quad s.t. \ w_i \ge 0, \forall i \tag{18}$$

$$\sum_{i=1}^{n} w_i \begin{bmatrix} e_i e_i^\top & \\ & (\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top \end{bmatrix} \preceq \begin{bmatrix} I_{n\times n} & \\ & cn\sigma^2 I_{d\times d} \end{bmatrix}$$

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 ℓ_2 [30], [31] or ℓ_1 techniques [32] from compressive sensing.[1] And it is well-known in compressive sensing that such iterative re-weighted approaches often performs better than ℓ_1 [32], [30].

C. Complexity analysis

Theorem 3 guarantees that the total number of iterations of Algorithm 1 required to achieve optimality is upper bounded by $O(\frac{\log d}{\log|\bar{\epsilon}|})$. In each iteration, the computational complexity of Step 2 is $O(nd)$. It follows easily from the proof of Theorem 3, that it suffices to solve the SDP in step 1 of Algorithm 1 (with $p = 1$) to a *constant* precision. As a result, the error is affected by a constant and thus remains order-optimal and the time complexity is $\tilde{O}(nd)$ parallelizable work using positive SDP solvers [33] (the notation $\tilde{O}(m)$ hides the poly-log factors: $\tilde{O}(m) = O(m.\text{polylog}(m))$). A comparison of our theoretical results with those in state-of-the-art works is given in Table I.

---

[1] We observe that iterative re-weighted ℓ_2 achieves better empirical performance.

If we use ℓ_p with $0 < p < 1$ in Step 1, we iteratively construct and minimize a *tight* upper bound on the ℓ_p objective via iterative re-weighted ℓ_2 [30], [31] or iterative re-weighted ℓ_1 techniques [32][2]. Minimizing the resulting weighted ℓ_1 objective can be also solved very efficiently to a *constant* precision by formulating it as a Packing SDP (see Appendix) with computational complexity of $\tilde{O}(nd)$ [33]. If we use iterative re-weighted ℓ_2, minimizing the resulting weighted ℓ_2 objective is a SDP constrained least squares problem, whose computational complexity is in general polynomial in both d and n. We will explore more efficient solutions for this objective in future work.

Algorithm	Time complexity	Error guarantee	Breakdown point	Requires ϵ
Tukey median [11]	NP-hard	$O(\sigma\sqrt{\epsilon})$	$\frac{1}{d+1}$	No
IF [14]	$\tilde{O}(nd^2)$	$O(\sigma\sqrt{\epsilon})$	NA	Yes
GF [27]	$\tilde{O}(n^2 d)$	$O\left(\sigma\frac{\sqrt{\epsilon}}{1-2\epsilon}\right)$	$\frac{1}{2}$	Yes
CDG [24]	$\tilde{O}\left(\frac{nd}{\epsilon^6}\right)$	$O(\sigma\sqrt{\epsilon})$	$\frac{1}{3}$	Yes
QUE [4]	$\tilde{O}(nd)$	$O(\sigma\sqrt{\epsilon})$	NA	Yes
Proposed optimization problems ($\ell_p, p \in [0,1]$)	NA	$O\left(\sigma\sqrt{\frac{\epsilon}{1-2\epsilon}}\right)$	$\frac{1}{2}$	No
Proposed algorithm ($p = 1$)	$\tilde{O}(nd)$	$O(\sigma\sqrt{\epsilon})$	$1 - \frac{1}{\sqrt{2}} \approx 0.3$	No

TABLE I: Theoretical comparison

IV. EMPIRICAL STUDIES

In this section, we present empirical results on the performance of Algorithm 1 and compare with the following state-of-the-art high dimension robust mean estimation methods: Iterative Filtering (IF) [14], Generalized Filtering (GF) [27, Algorithm 2], the method proposed in [8] (denoted as LRV), the method for bounded covariance distributions in [24] (denoted as CDG), and Quantum Entropy Scoring (QUE) [4], which scores the outliers based on multiple directions. We briefly discuss the implementation details of these algorithms. We implemented the QUE method by utilizing the code provided in [4]. In [24], the authors provide a way to implement the CDG method approximately; we provide results for an exact implementation of the CDG method. Since the number of datapoints considered in the following simulations is less than the minimum requirement, a reasonable approach to compare the performance of the algorithms is to tune the hyper-parameters of all algorithms to get the best possible error. For example, the hyper-parameter c_4 that appears in CDG method [24], is set to be 1.05, which produced the smallest empirical error. The value of σ provided to the algorithms is not the theoretical value, but the empirical one (precisely, the spectral norm of the sample covariance matrix of G^* (see Section II)). For evaluation purposes, we report the recovery error, which we define it as the ℓ_2 distance of the estimated mean to the oracle solution, i.e., the average of the uncorrupted datapoints after corruption.

A. Synthetic data

We consider two experimental settings. For the first setting we follow [4]. The dimension of the data is d, and the number of datapoints is n. The inlier datapoints are generated i.i.d. according to the standard Gaussian distribution

[2]We run fewer than 10 re-weighted iterations in our implementation.

with zero mean. Randomly (uniformly) chosen ϵ fraction of the datapoints are replaced by outliers. For the outliers, half of them are set to be $(\sqrt{d/2}, \sqrt{d/2}, 0, ..., 0)^\top$, and the other half are set as $(\sqrt{d/2}, -\sqrt{d/2}, 0, ..., 0)^\top$, so that their ℓ_2 distances to the population mean $(0, ..., 0)^\top$ are all $\sqrt{d}$, similar to that of the inlier points. These are two clusters of outliers, and their ℓ_2 distances to the true mean $\boldsymbol{x}$ are similar to that of the inlier points. In Algorithm 1, we set the threshold $\tau = 0.6$, $c_1 = 1.1$, and we initialize $c_2^{(0)}$ as the ℓ_2 error of the coordinate-wise Median relative to the true mean. We implemented the IF method for sub-Gaussian parameters [14, Theorem 3.1]. We vary the total fraction ϵ of the outliers and report the average recovery error of each method over 10 trials in Table II with $d = 100, n = 1000$. The proposed ℓ_1 and $\ell_{0.5}$ methods show significant improvements over the competing methods, and the $\ell_{0.5}$ method performs the best.

TABLE II: Recovery error of each method under different fraction ϵ of the outlier points ($d = 100, n = 1000$)

ϵ	IF	GF	QUE	LRV	CDG	ℓ_1	$\ell_{0.5}$
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 III. We can see that the performance of all methods get better when the number of datapoints is increased. Again, our proposed methods consistently perform better than the other methods.

TABLE III: Recovery error of each method w.r.t. different number of samples ($d = 100, \epsilon = 0.2$)

n	IF	GF	QUE	LRV	CDG	ℓ_1	$\ell_{0.5}$
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**

The second experimental setting is as follows. The dimension of the data is d, and the number of datapoints is n. The inlier datapoints are generated i.i.d. such that each coordinate follows the Pareto distribution with scale parameter as 1 and shape parameter as 2.5. This implies that each coordinate has bounded second moment but no higher moments. All outliers are set to be the same vector $\boldsymbol{v}$ which is chosen as follows. Let g be the average of the ℓ_2 norms of the datapoints. The vector $\boldsymbol{v}$ is set as $(2 + \sqrt{g/d}, 2 + \sqrt{g/d}, \ldots, 2 + \sqrt{g/d})^\top$. Randomly (uniformly) chosen ϵ fraction of the datapoints are replaced by outliers. The LRV method is applicable only to cases where distributions have bounded fourth moment, and hence is not applicable to this setting. The CDG method was not implemented due to the high computational complexity of its implementation. We implemented the IF method for bounded second moment parameters [14, Theorem 3.2]. We implemented Algorithm 1 with $p = 1$, $\tau = 1$(no thresholding during the iterations), $c_1 = 1$ and $c_2^{(0)} = 3\sqrt{d} + 2c_1$. However, we threshold the last iterate $h^{(T)}$ with

threshold 0.6. We vary the total fraction ϵ of the outliers and report the average recovery error of each method over 100 trials in Table IV with $d = 1000, n = 100000$. The proposed ℓ_1 method show significant improvement over the competing methods. We also tested the performance of each method for two different numbers of datapoints. The dimension of the data is fixed to be 1000. The fraction of the corrupted points is fixed to be 20%. We consider the number of datapoints to be 10000 and 100000, and report the average recovery error for each method over 100 trials in Table V. Again, the proposed method consistently perform better than the other methods.

TABLE IV: Recovery error of each method under different fraction ϵ of the outlier points ($d = 1000, n = 100000$)

ϵ	IF	GF	QUE	ℓ_1
10%	0.0164	0.0550	0.1096	**0.0161**
20%	0.0845	0.0726	0.2489	**0.0190**

TABLE V: Recovery error of each method w.r.t. different number of samples ($d = 1000, \epsilon = 0.2$)

n	IF	GF	QUE	ℓ_1
10000	0.3027	0.0780	0.3991	**0.0257**
100000	0.0845	0.0726	0.2489	**0.0190**

B. Corrupted image dataset

Here we use a dataset of real face images to test the effectiveness of the robust mean estimation methods. The average face of particular regions or certain groups of people is useful for many social and psychological studies [34]. Here we use 100 frontal human face images from the Brazilian face database[3] as inliers. For the outliers, we choose 15 face images of cats and dogs from the CIFAR10 [35] database. In order to be able to run the CDG method [24], we scale the size of images to 18×15 pixels, so the dimension of each datapoint is 270. The oracle solution is the average of the 100 human faces. Table VI reports the recovery error, which is the ℓ_2 distance of the estimated mean to the oracle solution, for each method. The proposed methods achieve smaller recovery error than the state-of-the-art methods. The sample inlier and outlier images as well as the estimated mean for each method can be found in the Appendix.

TABLE VI: Recovery error of the mean face by each method

Sample average	IF	LRV	CDG	ℓ_1	$\ell_{0.5}$
141	63	83	81	**38**	**46**

[3]https://fei.edu.br/ cet/facedatabase.html

V. Conclusion

We formulated the robust mean estimation problems as the minimization of the ℓ_0-'norm' of the introduced outlier indicator vector, under a second moment constraint on the datapoints. We further relaxed the ℓ_0 objective to an ℓ_p $(0 < p \leq 1)$ objective, and theoretically justified the new objective. The proposed ℓ_0 and ℓ_p optimization problems do not need to know ϵ, and still achieve information-theoretically order-optimal error bounds with optimal breakdown points. Then we proposed a computationally tractable iterative $\ell_p(0 < p \leq 1)$ minimization and hard thresholding algorithm, which significantly outperforms state-of-the-art robust mean estimation methods, and is order-optimal. In the empirical studies, we observed strong numerical evidence that using the ℓ_p $(0 < p \leq 1)$ norm in the optimization leads to sparse solutions; theoretically justifying this phenomenon is also of interest. It is worth noting that almost all previous polynomial-time methods (with dimension-independent error bound) need to know ϵ, while our Algorithm 1 does not require to know ϵ. It has a maximal breakdown point of $1 - 1/\sqrt{2}$, and has near-linear time complexity for $p = 1$.

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*. New York: Wiley, 2004.

[3] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. New York: Wiley, 2019.

[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*, vol. 2, Vancouver, August 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. Adams, "High-dimensional probability lecture notes," 2020, available online.

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

[18] 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.

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

[20] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier analysis and Applications*, vol. 14, no. 5, pp. 629–654, 2008.

[21] J.-L. Bouchot, S. Foucart, and P. Hitczenko, "Hard thresholding pursuit algorithms: Number of iterations," *Applied and Computational Harmonic Analysis*, vol. 41, no. 2, pp. 412–435, 2016, sparse Representations with Applications in Imaging Science, Data Analysis, and Beyond, Part II. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1063520316000130

[22] M. Wang, W. Xu, and A. Tang, "The limits of error correction with lp decoding," in *2010 IEEE International Symposium on Information Theory*, 2010, pp. 749–753.

[23] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[24] 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*, San Diego, CA, USA, 2019, pp. 2755–2771.

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

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

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

[28] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 03 1964.

[29] 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*, Berlin, June 2005, pp. 152–166.

[30] 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.

[31] 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.

[32] 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.

[33] Z. Allen-Zhu, Y. T. Lee, and L. Orecchia, "Using optimization to obtain a width-independent, parallel, simpler, and faster positive sdp solver," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, Arlington, VA, Jan 2016.

[34] A. C. Little, B. C. Jones, and L. M. DeBruine, "Facial attractiveness: evolutionary based research," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, no. 1571, pp. 1638–1659, 2011.

[35] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," Master's thesis, University of Toronto, 2009.

VI. Appendix

A. Technical preliminaries

We introduce the following parameters that control the minimum number of datapoints needed, error and confidence level. Let $\delta > 0$, $c_1 > 1$ and $c_1' = [c_1^2 \min\{c_1^2 \log c_1^2 + 1 - c_1^2, 1\}]$. Let $n > \frac{ed \log(d/\delta)}{\delta^2 c_1'}$ and $\alpha = \frac{ed \log(d/\delta)}{n \delta^2 c_1'}$. Let $\boldsymbol{S} = \{\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n\}$ be a set of n datapoints drawn from a distribution P with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma \preceq \sigma^2 I$. We now define $\boldsymbol{G}$ as the set of datapoints which are less than $\sigma \sqrt{\frac{d}{\alpha \delta}} = \sigma \sqrt{\frac{n \delta c_1'}{e \log(d/\delta)}}$ distance away from $\boldsymbol{\mu}$:

$$\boldsymbol{I} = \left\{ i : \|\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}\|_2 \leq \sigma \sqrt{\frac{d}{\alpha \delta}} \right\} \tag{19}$$

$$\boldsymbol{G} = \{\tilde{\boldsymbol{y}}_i : i \in \boldsymbol{I}\}. \tag{20}$$

It follows from Lemma 4 that for the event

$$\mathcal{E}_1 = \{|\boldsymbol{I}| \geq n - \alpha n\}, \tag{21}$$

$$\mathsf{P}(\mathcal{E}_1) \geq 1 - \delta. \tag{22}$$

Let $\mathcal{E}_2$ be the event:

$$\mathcal{E}_2 = \left\{ \lambda_{\max} \left(\sum_{i \in \boldsymbol{I}} (\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})^\top \right) \leq c_1^2 \sigma^2 n \right\}. \tag{23}$$

It follows from Lemma 5 that

$$\mathsf{P}(\mathcal{E}_2) \geq 1 - \delta. \tag{24}$$

Thus, we have that

$$\mathsf{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 2\delta. \tag{25}$$

For analysis purposes, we consider the *far away* uncorrupted datapoints $\boldsymbol{S} \setminus \boldsymbol{G}$ as outliers also.

Let $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ be an ϵ-corrupted version of the set $\boldsymbol{S}$. Let $\boldsymbol{h}^*$ be such that $h_i^* = 1$ for the outliers (both far away uncorrupted datapoints and corrupted datapoints), and $h_i^* = 0$ for the rest of uncorrupted datapoints, i.e.,

$$h_i^* = \begin{cases} 1, & \text{if } \boldsymbol{y}_i \neq \tilde{\boldsymbol{y}}_i \quad \text{or} \quad \tilde{\boldsymbol{y}}_i \in \boldsymbol{S} \setminus \boldsymbol{G} \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

Let the set of *inliers* be given by $\boldsymbol{G}^*$:

$$\boldsymbol{I}^* = \{i : h_i^* = 0\} \tag{27}$$

$$\boldsymbol{G}^* = \{\boldsymbol{y}_i : i \in \boldsymbol{I}^*\} = \{\tilde{\boldsymbol{y}}_i : i \in \boldsymbol{I}^*\} \tag{28}$$

Note that $\boldsymbol{I}^* \subseteq \boldsymbol{I}$ and $\boldsymbol{G}^* \subseteq \boldsymbol{G}$. Since $(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})^\top$ is positive semi-definite (PSD), we must have

$$\lambda_{\max} \left(\sum_{i=1}^n (1 - h_i^*)(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^\top \right) \leq \lambda_{\max} \left(\sum_{i \in \boldsymbol{I}} (\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})^\top \right).$$

This implies that

$$\left\{\lambda_{\max}\left(\sum_{i=1}^{n}(1-h_i^*)(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^\top\right) \leq c_1^2\sigma^2 n\right\} \supseteq \mathcal{E}_2. \tag{29}$$

Then, we have:

$$\mathsf{P}\left\{\lambda_{\max}\left(\sum_{i=1}^{n}(1-h_i^*)(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^\top\right) \leq c_1^2\sigma^2 n\right\} \geq \mathsf{P}(\mathcal{E}_2) \geq 1-\delta. \tag{30}$$

Our intended solution is to have $h_i = 0$ for the inlier points and $h_i = 1$ for the outlier points.

Let $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{x}}^*$ be the averages of datapoints in $\boldsymbol{G}$ and $\boldsymbol{G}^*$ respectively. Applying Lemma C.2 from [27], we have

$$\|\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}^*\|_2 \leq \sqrt{\frac{c_1^2\sigma^2}{1-\alpha}\cdot\frac{\epsilon}{1-\epsilon}}. \tag{31}$$

We now introduce some more events (c.f. [14, Lemma A.18]):

$$\mathcal{E}_3 = \left\{\left\|\frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i-\boldsymbol{\mu})\right\|_2 \leq \sigma\sqrt{\frac{d}{n\delta}}\right\} \tag{32}$$

$$\mathcal{E}_4 = \left\{\left\|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{z}_i-\mathsf{E}[\boldsymbol{z}_1])\right\|_2 \leq \sigma\sqrt{\frac{d}{n\delta}}\right\}, \tag{33}$$

where $\boldsymbol{z}_i = (\tilde{\boldsymbol{y}}_i-\boldsymbol{\mu})\mathbb{1}\left\{\|\tilde{\boldsymbol{y}}_i-\boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}}\right\}$. From Lemma 4, we get that

$$\mathsf{P}(\mathcal{E}_3) \geq 1-\delta, \text{ and } \mathsf{P}(\mathcal{E}_4) \geq 1-\delta. \tag{34}$$

Let $\mathcal{E}$ be the event given by

$$\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4. \tag{35}$$

Let $\Delta_{n,\xi}$ be the set of probability vectors given by:

$$\Delta_{n,\xi} = \left\{\boldsymbol{w} \in \mathbb{R}^n : 0 \leq w_i \leq \frac{1}{1-\xi}, \sum_{i=1}^{n} w_i = 1\right\}. \tag{36}$$

Let $\mathrm{TV}(.,.)$ denote the total variation distance between probability measures.

TABLE VII: Description of variables

Variable	Description
$\boldsymbol{\mu}$	Mean (expected value) of population distribution
$\tilde{\boldsymbol{\mu}}$	Average of all datapoints before corruption
$\boldsymbol{G}$	Set of datapoints within $\sigma\sqrt{\frac{d}{\alpha\delta}} = \sigma\sqrt{\frac{n\delta c_1'}{e\log(d/\delta)}}$ of $\boldsymbol{\mu}$ before corruption
$\boldsymbol{G}^*$	Maximal subset of $\boldsymbol{G}$ which is uncorrupted by adversary
$\bar{\boldsymbol{x}}$	Average of vectors in $\boldsymbol{G}$, the set of datapoints within $\sigma\sqrt{\frac{d}{\alpha\delta}} = \sigma\sqrt{\frac{n\delta c_1'}{e\log(d/\delta)}}$ of $\boldsymbol{\mu}$
$\bar{\boldsymbol{x}}^*$	Average of vectors in $\boldsymbol{G}^*$, the set of inliers within $\boldsymbol{G}$

B. Technical Lemmas

Lemma 1 (Lemma 2.2 [27]). *For a finite set of datapoints $\{\boldsymbol{y}_i\}_{i=1}^n$, let $\boldsymbol{x_w} = \sum\limits_{i \in [n]} w_i \boldsymbol{y}_i$ and $\Sigma_{\boldsymbol{w}} = \sum\limits_{i \in [n]} w_i(\boldsymbol{y}_i - \boldsymbol{x_w})(\boldsymbol{y}_i - \boldsymbol{x_w})^\top$ be the weighted average and weighted covariance with respect to a probability weight vector $\boldsymbol{w}$. Let $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ be two probability weight vectors such that $\mathrm{TV}(\boldsymbol{w}_1, \boldsymbol{w}_2) \leq \zeta$. Then,*

$$\|\boldsymbol{x}_{\boldsymbol{w}_1} - \boldsymbol{x}_{\boldsymbol{w}_2}\|_2 \leq \left(\sqrt{\lambda_{\max}(\Sigma_{\boldsymbol{w}_1})} + \sqrt{\lambda_{\max}(\Sigma_{\boldsymbol{w}_2})} \right) \sqrt{\frac{\zeta}{1 - \zeta}} \tag{37}$$

Lemma 2 (Lemma 2.3 [27]). *Let $\boldsymbol{w}_1 \in \Delta_{n,\epsilon_1}$ and $\boldsymbol{w}_2 \in \Delta_{n,\epsilon_2}$. Then*

$$\mathrm{TV}(\boldsymbol{w}_1, \boldsymbol{w}_2) \leq \frac{\max\{\epsilon_1, \epsilon_2\}}{1 - \min\{\epsilon_1, \epsilon_2\}}. \tag{38}$$

Lemma 3. *Let P be a distribution on $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma \preceq \sigma^2 I$. Let $\epsilon \leq 1/3$. Given an ϵ-fraction corrupted set of n datapoints from P, the coordinate-wise median of the corrupted set, $\hat{\boldsymbol{x}}$, satisfies with probability at least $1 - d\exp(-n/90)$ that*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 \leq 3\sigma\sqrt{d}. \tag{39}$$

Proof. We first show that with high probability the error in each dimension is bounded by 3σ. Fix a coordinate, and let $\tilde{y}_i$, y_i, μ and $\hat{x}$ be the component of $\tilde{\boldsymbol{y}}_i$, $\boldsymbol{y}_i$, $\boldsymbol{\mu}$ and $\hat{\boldsymbol{x}}$ respectively in that coordinate. By Markov's inequality, we have

$$\mathsf{P}(|\tilde{y}_i - \mu_i| \geq 3\sigma) \leq 1/9. \tag{40}$$

Let $b_i = \mathbb{1}\{|\tilde{y}_i - \mu_i| \geq 3\sigma\}$. By Chernoff's inequality, we obtain

$$\mathsf{P}\left(\sum_{i=1}^n b_i \geq n/6 \right) \leq \exp\left(-\frac{(0.5)^2 n}{9(2 + 0.5)} \right) = \exp(-n/90). \tag{41}$$

Thus with high probability more than five-sixth of the datapoints satisfy $|\tilde{y}_i - \mu_i| \leq 3\sigma$, which implies that even if $\epsilon \leq 1/3$ fraction of datapoints are corrupted, we would have

$$|\hat{x} - \mu| \leq 3\sigma. \tag{42}$$

Applying union bound, we get that with probability at least $1 - d\exp(-n/90)$, the error in each dimension is bounded by 3σ and hence $\|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 \leq 3\sigma\sqrt{d}$ holds.

$\square$

Lemma 4. *Let $0 < \delta \leq 1$. Let $\mathcal{E}_1$, $\mathcal{E}_3$ and $\mathcal{E}_4$ be the events as described in* (21), (32) *and* (33). *Then,*

$$\mathsf{P}(\mathcal{E}_1) \geq 1 - \delta, \ \mathsf{P}(\mathcal{E}_3) \geq 1 - \delta, \ and \ \mathsf{P}(\mathcal{E}_4) \geq 1 - \delta,$$

Proof. By Markov's inequality we have

$$P(|\boldsymbol{G}^c| > \alpha n) \le \frac{\mathsf{E}[|\boldsymbol{G}^c|]}{\alpha n} \tag{43}$$

$$= \frac{\mathsf{E}\left[\sum\limits_{i=1}^{n} 1\left\{\|\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}}\right\}\right]}{\alpha n} \tag{44}$$

$$= \frac{P\left(\|\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}}\right)}{\alpha}. \tag{45}$$

Applying Markov's inequality again, we have

$$P\left(\|\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}}\right) \le \frac{\alpha\delta\,\mathsf{E}\left[\|\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu}\|_2^2\right]}{\sigma^2 d} \tag{46}$$

$$= \frac{\alpha\delta\mathrm{Tr}(\mathsf{E}[(\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu})(\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu})^\top])}{\sigma^2 d} \tag{47}$$

$$\le \frac{\alpha\delta\sigma^2 d}{\sigma^2 d} \tag{48}$$

$$= \alpha\delta. \tag{49}$$

Thus, we get

$$P(|\boldsymbol{G}^c| > \alpha n) \le \delta \tag{50}$$

$$P(|\boldsymbol{G}| \ge (1-\alpha)n) \ge 1 - \delta. \tag{51}$$

This proves the result for $\mathcal{E}_1$. Applying Markov's inequality again, we obtain

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})\right\|_2 \le \sigma\sqrt{\frac{d}{n\delta}}\right) \le \frac{\mathsf{E}\left[\left\|\frac{1}{n}\sum\limits_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})\right\|_2^2\right]}{\frac{\sigma^2 d}{n\delta}} \tag{52}$$

$$= \frac{n\delta}{\sigma^2 d}\sum_{k=1}^{d}\mathsf{E}\left[(\tilde{\mu}_k - \mu_k)^2\right] \tag{53}$$

$$\le \frac{n\delta}{\sigma^2 d}\cdot\frac{d\sigma^2}{n} \tag{54}$$

$$= \delta. \tag{55}$$

This proves the result for $\mathcal{E}_3$. By similar reasoning, the result for $\mathcal{E}_4$ follows. $\square$

Lemma 5. *Let* $0 < \delta \le 1$, $c_1 > 1$, $c_1' = [c_1^2\min\{c_1^2\log c_1^2 + 1 - c_1^2, 1\}]$, $n > \frac{ed\log(d/\delta)}{\delta^2 c_1'}$ *and* $\alpha = \frac{ed\log(d/\delta)}{n\delta^2 c_1'}$. *Let* $\mathcal{E}_2$ *be the event described in* (23). *Then*

$$P(\mathcal{E}_2) \ge 1 - \delta.$$

Proof. We adopt the approach in [14, Lemma A.18 (iv)]. Lemma A.19 from [14] states that the following: Let $\{X_i\}_{i=1}^n$ be $d \times d$ positive semi-definite random matrices such that $\lambda_{\max}(X_i) \le L$ almost surely for all i. Let $S = \sum\limits_{i=1}^{n} X_i$ and $M = \lambda_{\max}(\mathsf{E}[S])$. Then, for any $\theta > 0$,

$$\mathsf{E}[\lambda_{\max}(S)] \le (e^\theta - 1)M/\theta + L\log(d)/\theta, \tag{56}$$

and for any $\eta > 0$,

$$P(\lambda_{\max}(S) \geq (1+\eta)M) \leq d\left(\frac{e^\eta}{(1+\eta)^{1+\eta}}\right)^{M/L}. \tag{57}$$

We apply this result by assigning $X_i = (\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})^\top \mathbb{1}\left\{\|\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}\|_2 \leq \sigma\sqrt{\frac{d}{\alpha\delta}}\right\}$. Note that $\lambda_{\max}(X_i) \leq L = \frac{\sigma^2 d}{\alpha\delta}$ for all $i \in [n]$, and $M \leq n\lambda_{\max}(\mathsf{E}[X_1]) \leq n\sigma^2$. We consider two mutually exclusive cases:

1) Suppose that $M < e^{-1}\delta c_1^2 \sigma^2 n$. Applying (56) with $\theta = 1$, we obtain

$$\mathsf{E}[\lambda_{\max}(S)] \leq (e-1)M + L\log d. \tag{58}$$

Applying Markov's inequality, we obtain

$$P(\lambda_{\max}(S) \geq c_1^2 \sigma^2 n) \leq \frac{\mathsf{E}[\lambda_{\max}(S)]}{c_1^2 \sigma^2 n} \tag{59}$$

$$\leq \frac{(e-1)\delta c_1^2 \sigma^2 n}{e c_1^2 \sigma^2 n} + \frac{\sigma^2 d \log d}{\alpha\delta c_1^2 \sigma^2 n} \tag{60}$$

$$\leq \frac{(e-1)\delta}{e} + \frac{\delta}{e} \tag{61}$$

$$= \delta. \tag{62}$$

The inequality in (60) follows from the assumption that $M < e^{-1}\delta c_1^2 \sigma^2 n$ and the inequality in (61) follows from the fact that $\alpha = \frac{ed\log(d/\delta)}{n\delta^2 c_1'}$ and $c_1' \leq c_1^2$.

2) Suppose that $M \geq e^{-1}\delta c_1^2 \sigma^2 n$. Applying (57) with $\eta = c_1^2 - 1$, we obtain

$$P(\lambda_{\max}(S) \geq c_1^2 \sigma^2 n) \leq P(\lambda_{\max}(S) \geq c_1^2 M) \tag{63}$$

$$\leq d\left(\frac{e^{c_1^2 - 1}}{(c_1^2)^{c_1^2}}\right)^{\frac{\delta c_1^2 \sigma^2 n}{e} \cdot \frac{\alpha\delta}{\sigma^2 d}} \tag{64}$$

$$\leq \delta. \tag{65}$$

The inequality in (63) follows from the fact that $M \leq n\sigma^2$, the inequality in (65) follows from the fact that $e^\alpha < (1+\alpha)^{1+\alpha}$ for any $\alpha > 0$, and the fact that $\alpha = \frac{ed\log(d/\delta)}{n\delta^2 c_1'}$ and $c_1' = [c_1^2 \min\{c_1^2 \log c_1^2 + 1 - c_1^2, 1\}]$. $\qquad\square$

Lemma 6. *Given a set of points $\boldsymbol{y}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, then for any $\boldsymbol{w} \in \mathbb{R}^n$ we have*

$$\boldsymbol{x}_w \triangleq \frac{\sum_{i=1}^n w_i \boldsymbol{y}_i}{\|\boldsymbol{w}\|_1} \in \arg\min_{\boldsymbol{x}} \lambda_{\max}\left(\sum_{i=1}^n w_i(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \tag{66}$$

Proof. We have

$$\min_{\boldsymbol{x}} \lambda_{\max}\left(\sum_{i=1}^n w_i(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\nu}:\|\boldsymbol{\nu}\|_2=1} \sum_{i=1}^n w_i\langle \boldsymbol{y}_i - \boldsymbol{x}, \boldsymbol{\nu}\rangle^2 \tag{67}$$

$$\geq \max_{\boldsymbol{\nu}:\|\boldsymbol{\nu}\|_2=1} \min_{\boldsymbol{x}} \sum_{i=1}^n w_i\langle \boldsymbol{y}_i - \boldsymbol{x}, \boldsymbol{\nu}\rangle^2 \tag{68}$$

$$= \max_{\boldsymbol{\nu}:\|\boldsymbol{\nu}\|_2=1} \sum_{i=1}^n w_i\langle \boldsymbol{y}_i - \boldsymbol{x}_w, \boldsymbol{\nu}\rangle^2 \tag{69}$$

$$= \lambda_{\max}\left(\sum_{i=1}^n w_i(\boldsymbol{y}_i - \boldsymbol{x}_w)(\boldsymbol{y}_i - \boldsymbol{x}_w)^\top\right). \tag{70}$$

The equality (69) follows from the fact that the minimum in the RHS of (68) is attained at $\boldsymbol{x}_w = \frac{\sum_{i=1}^n w_i \boldsymbol{y}_i}{\|\boldsymbol{w}\|_1}$. Consequently, (66) holds.

$\square$

Lemma 7. *Let $n > \frac{ed\log(d/\delta)}{\delta^2 c_1'}$ and $\alpha = \frac{ed\log(d/\delta)}{n\delta^2 c_1'}$. Suppose $\|\boldsymbol{x} - \bar{\boldsymbol{x}}^*\|_2 \leq c_2\sigma$, where $\bar{\boldsymbol{x}}^*$ is the average of datapoints in $\boldsymbol{G}^*$, defined in (28). Then on event $\mathcal{E}_2$ defined in (23), $\boldsymbol{h}^*$ satisfies*

$$\lambda_{\max}\left(\sum_{i=1}^n (1 - h_i^*)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \leq (c_1^2 + c_2^2)\sigma^2 n. \tag{71}$$

Proof. Let $\boldsymbol{I}$ and $\boldsymbol{I}^*$ be the sets defined in (19) and (27). We have

$$\lambda_{\max}\left(\sum_{i=1}^n (1 - h_i^*)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \tag{72}$$

$$= \lambda_{\max}\left(\sum_{i\in\boldsymbol{I}^*} (\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \tag{73}$$

$$= \lambda_{\max}\left(\sum_{i\in\boldsymbol{I}^*} (\boldsymbol{y}_i - \bar{\boldsymbol{x}}^* + \bar{\boldsymbol{x}}^* - \boldsymbol{x})(\boldsymbol{y}_i - \bar{\boldsymbol{x}}^* + \bar{\boldsymbol{x}}^* - \boldsymbol{x})^\top\right) \tag{74}$$

$$\leq \lambda_{\max}\left(\sum_{i\in\boldsymbol{I}^*} (\boldsymbol{y}_i - \bar{\boldsymbol{x}}^*)(\boldsymbol{y}_i - \bar{\boldsymbol{x}}^*)^\top\right) + \lambda_{\max}\left(\sum_{i\in\boldsymbol{I}^*} (\boldsymbol{x} - \bar{\boldsymbol{x}}^*)(\boldsymbol{x} - \bar{\boldsymbol{x}}^*)^\top\right) \tag{75}$$

$$+ 2\lambda_{\max}\left(\sum_{i\in\boldsymbol{I}^*} (\boldsymbol{y}_i - \bar{\boldsymbol{x}}^*)(\bar{\boldsymbol{x}}^* - \boldsymbol{x})^\top\right) \tag{76}$$

$$= \lambda_{\max}\left(\sum_{i\in\boldsymbol{I}^*} (\boldsymbol{y}_i - \bar{\boldsymbol{x}}^*)(\boldsymbol{y}_i - \bar{\boldsymbol{x}}^*)^\top\right) + |\boldsymbol{I}^*|\|\boldsymbol{x} - \bar{\boldsymbol{x}}^*\|^2 + 0 \tag{77}$$

$$\leq c_1^2\sigma^2 n + c_2^2\sigma^2 n. \tag{78}$$

The last inequality follows from the definition of $\mathcal{E}_2$ in (23) and Lemma 6. $\square$

Lemma 8. *Let $n > \frac{ed\log(d/\delta)}{\delta^2 c_1'}$ and $\alpha = \frac{ed\log(d/\delta)}{n\delta^2 c_1'}$. Let $\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n$ be i.i.d. datapoints drawn from a distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma \preccurlyeq \sigma^2 I$. Let $\boldsymbol{G}$ be the set defined in (20). Let $\bar{\boldsymbol{x}}$ be the average of datapoints in $\boldsymbol{G}$. Then the following holds on the event $\mathcal{E}_1 \cap \mathcal{E}_3 \cap \mathcal{E}_4$, where the events are defined in (21), (32) and (33):*

$$\|\bar{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 \leq \sigma\sqrt{\alpha\delta}\left(1 + 2\sqrt{\frac{c_1'}{e\log(d/\delta)}}\right). \tag{79}$$

Proof. Note that

$$\left\| \frac{|\boldsymbol{G}|}{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \right\|_2 \tag{80}$$

$$= \left\| \frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}) - \frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})\mathbb{1}\left\{ \|\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}} \right\} \right\|_2 \tag{81}$$

$$\leq \left\| \frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}) \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i \right\|_2 \tag{82}$$

$$\leq \left\| \frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}) \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{z}_i - E[\boldsymbol{z}_1]) \right\|_2 + \|E[\boldsymbol{z}_1]\|_2, \tag{83}$$

where $\boldsymbol{z}_i = (\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu})\mathbb{1}\left\{ \|\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}} \right\}$.

The last term is upper bounded as follows,

$$\|E[\boldsymbol{z}_1]\|_2 = \left\| E\left[(\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu})\mathbb{1}\left\{ \|\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}} \right\} \right] \right\|_2 \tag{84}$$

$$= \max_{\|v\|_2=1} v^\top E\left[(\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu})\mathbb{1}\left\{ \|\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}} \right\} \right] \tag{85}$$

$$= \max_{\|v\|_2=1} E\left[v^\top(\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu})\mathbb{1}\left\{ \|\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}} \right\} \right] \tag{86}$$

$$\overset{(a)}{\leq} \max_{\|v\|_2=1} \sqrt{ E[v^\top(\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu})]^2 P\left(\|\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}} \right) } \tag{87}$$

$$= \sqrt{ \lambda_{\max}(\Sigma) P\left(\|\tilde{\boldsymbol{y}}_1 - \boldsymbol{\mu}\|_2 > \sigma\sqrt{\frac{d}{\alpha\delta}} \right) } \tag{88}$$

$$\overset{(b)}{\leq} \sqrt{\sigma^2\alpha\delta} \tag{89}$$

$$= \sigma\sqrt{\alpha\delta}. \tag{90}$$

The inequality (a) follows from Cauchy-Schwarz inequality, and (b) follows from Markov's inequality.

From (83), (22), (34), and (90), we get that on the event $\mathcal{E}_1 \cap \mathcal{E}_3 \cap \mathcal{E}_4$,

$$\|\bar{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 \leq \sigma\sqrt{\alpha\delta}\left(1 + 2\sqrt{\frac{c_1'}{e\log(d/\delta)}} \right). \tag{91}$$

$\square$

Lemma 9. *Let $0 < \tau \leq 1$. Suppose $\boldsymbol{h} \in \mathbb{R}^n$ such that $\forall i$, $0 \leq h_i \leq 1$, and $\|\boldsymbol{h}\|_1 \leq \epsilon n$ for some $\epsilon \in [0, 1)$. Then*

$$\sum_{i=1}^{n}(1 - h_i)\mathbb{1}\{h_i \leq \tau\} \geq \left(1 - \frac{\epsilon}{\tau} \right)n. \tag{92}$$

Proof. We first show that $\sum_{i=1}^{n} 1\{h_i > \tau\} \leq \frac{\epsilon n}{\tau}$. Observe that

$$\epsilon n \geq \sum_{i=1}^{n} h_i = \sum_{i=1}^{n} h_i 1\{h_i \leq \tau\} + \sum_{i=1}^{n} h_i 1\{h_i > \tau\} \tag{93}$$

$$\geq \tau \sum_{i=1}^{n} 1\{h_i > \tau\}. \tag{94}$$

Hence, we have

$$\sum_{i=1}^{n} 1\{h_i > \tau\} \leq \frac{\epsilon n}{\tau}. \tag{95}$$

Consequently, we obtain

$$\sum_{i=1}^{n}(1 - h_i)1\{h_i \leq \tau\} = \sum_{i=1}^{n}(1 - h_i) - \sum_{i=1}^{n}(1 - h_i)1\{h_i > \tau\} \tag{96}$$

$$\geq \sum_{i=1}^{n}(1 - h_i) - (1 - \tau)\sum_{i=1}^{n} 1\{h_i > \tau\} \tag{97}$$

$$\geq (1 - \epsilon)n - \frac{(1 - \tau)\epsilon n}{\tau} \tag{98}$$

$$= \left(1 - \frac{\epsilon}{\tau}\right)n. \tag{99}$$

$\square$

C. Proof of Theorem 1

Proof. Let $(\hat{h}, \hat{x})$ be a feasible pair for (3) lying in $\mathcal{S}$. Note that we get a corresponding feasible pair lying in $\mathcal{S}$ by only setting non-zero $\hat{h}_i$ to be 1. With slight abuse of notation, let $(\hat{h}, \hat{x})$ be this feasible pair.

Let $\epsilon' \triangleq \alpha + \epsilon$. Let $\hat{w} = \frac{1 - \hat{h}}{n - \|\hat{h}\|_0}$ and $\beta = \|\hat{h}\|_0/n$. Note that $\hat{w} \in \Delta_{n,\beta}$. Consider h^* as defined in (26). Let $\bar{x}^*$ be the average of datapoints in the set G^* defined in (28) and let $w^* = \frac{1 - h^*}{n - \|h^*\|_0}$. Observe that on event $\mathcal{E}_1$, $w^* \in \Delta_{n,\epsilon'}$. From Lemma 2, we obtain

$$\mathrm{TV}(\hat{w}, w^*) \leq \frac{\max(\beta, \epsilon')}{1 - \min(\beta, \epsilon')}. \tag{100}$$

As a consequence of Lemma 6, on event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\lambda_{\max}(\Sigma_{w^*}) \leq \lambda_{\max}\left(\frac{1}{|\mathbf{I}^*|}\sum_{i=1}^{n}(1 - h_i^*)(y_i - \mu)(y_i - \mu)^\top\right) \leq \frac{c_1^2\sigma^2 n}{|\mathbf{I}^*|} \leq \frac{c_1^2\sigma^2}{1 - \epsilon'} \tag{101}$$

$$\lambda_{\max}(\Sigma_{\hat{w}}) \leq \lambda_{\max}\left(\frac{1}{n - \|\hat{h}\|_0}\sum_{i=1}^{n}(1 - \hat{h}_i)(y_i - \hat{x})(y_i - \hat{x})^\top\right) \leq \frac{c_1^2\sigma^2 n}{n - \|\hat{h}\|_0} = \frac{c_1^2\sigma^2}{1 - \beta}. \tag{102}$$

Consider the case $\beta \leq \epsilon' < 1/2$. This implies $\mathrm{TV}(\hat{w}, w^*) \leq \frac{\epsilon'}{1 - \beta} < 1$. From Lemma 1, on event $\mathcal{E}_1 \cap \mathcal{E}_2$, we get

$$\|\hat{x} - \bar{x}^*\|_2 \leq \left(\sqrt{\frac{c_1^2\sigma^2}{1 - \epsilon'}} + \sqrt{\frac{c_1^2\sigma^2}{1 - \beta}}\right)\sqrt{\frac{\epsilon'}{1 - \epsilon' - \beta}} \tag{103}$$

$$\leq \frac{2c_1\sigma\sqrt{\epsilon'}}{\sqrt{(1 - \epsilon')(1 - 2\epsilon')}}. \tag{104}$$

Consider the case $\epsilon' \le \beta < 1 - \epsilon'$. This implies $\mathrm{TV}(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \le \frac{\beta}{1-\epsilon'} < 1$. From Lemma 1, on event $\mathcal{E}_1 \cap \mathcal{E}_2$, we get

$$\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}^*\|_2 \le \left(\sqrt{\frac{c_1^2 \sigma^2}{1-\epsilon'}} + \sqrt{\frac{c_1^2 \sigma^2}{1-\beta}} \right) \sqrt{\frac{\beta}{1-\epsilon'-\beta}}. \tag{105}$$

Consequently, on the event $\mathcal{E}$ defined in (35), using Lemma 8, (31) and applying triangle inequality, we obtain that with probability at least $1 - 4\delta$

$$\|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 \le \|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}^*\|_2 + \sqrt{\frac{c_1^2 \sigma^2}{1-\alpha} \cdot \frac{\epsilon}{1-\epsilon}} + \sigma \sqrt{\alpha \delta} \left(1 + 2\sqrt{\frac{c_1'}{e \log(d/\delta)}} \right).$$

It follows from (29) that on the event $\mathcal{E}_2$, $\mathcal{E}_2$, $(\boldsymbol{h}^*, \bar{\boldsymbol{x}}^*)$ is feasible. We also have that $\mathcal{E}_1 = \{|\boldsymbol{I}| \ge (1-\alpha)n\} \subseteq \{n - \|\boldsymbol{h}^*\|_0 \ge (1-\epsilon')n\} = \{\|\boldsymbol{h}^*\|_0 \le \epsilon'n\}$. Note that for any globally optimal solution of (3), by setting all its non-zero h_i to be 1, we can always get corresponding feasible and globally optimal $(\boldsymbol{h}^{\mathrm{opt}}, \boldsymbol{x}^{\mathrm{opt}})$ with $h_i^{\mathrm{opt}} \in \{0,1\}$ and $\boldsymbol{x}^{\mathrm{opt}} = \frac{\sum_{\{i : h_i^{\mathrm{opt}}=0\}} \boldsymbol{y}_i}{|\{i : h_i^{\mathrm{opt}}=0\}|}$ (i.e., $\boldsymbol{x}^{\mathrm{opt}}$ is the average of the $\boldsymbol{y}_i$'s corresponding to $h_i^{\mathrm{opt}} = 0$), and the objective value remains unchanged. Since $(\boldsymbol{h}^{\mathrm{opt}}, \boldsymbol{x}^{\mathrm{opt}})$ is globally optimal, and $(\boldsymbol{h}^*, \boldsymbol{\mu})$ is feasible, we have $\|\boldsymbol{h}^{\mathrm{opt}}\|_0 \le \|\boldsymbol{h}^*\|_0 \le \epsilon'n$. Hence, $(\boldsymbol{h}^{\mathrm{opt}}, \boldsymbol{x}^{\mathrm{opt}}) \in \mathcal{S}'$ with $\|\boldsymbol{h}^{\mathrm{opt}}\|_0 \le \epsilon'n$.

$\square$

D. Proof of Theorem 2

Proof. Let $(\hat{\boldsymbol{h}}, \hat{\boldsymbol{x}}) \in \mathcal{S}'$ be a feasible pair for (4) with some $0 < p \le 1$. We have

$$\|\hat{\boldsymbol{h}}\|_p \le ((1-\epsilon')n)^{1/p}. \tag{106}$$

Since $0 \le \hat{h}_i \le 1$ for all i, we have

$$\left[\sum_{i=1}^n \hat{h}_i \right]^{1/p} \le \left[\sum_{i=1}^n \hat{h}_i^p \right]^{1/p} \le ((1-\epsilon')n)^{1/p}. \tag{107}$$

This implies the following

$$\|\hat{\boldsymbol{h}}\|_1 \le \|\hat{\boldsymbol{h}}\|_p^p \le (1-\epsilon')n \tag{108}$$

$$\|\mathbf{1} - \hat{\boldsymbol{h}}\|_1 \ge n - \|\hat{\boldsymbol{h}}\|_p^p \ge \epsilon'n. \tag{109}$$

Let $\hat{\boldsymbol{w}} = \frac{\mathbf{1} - \hat{\boldsymbol{h}}}{\|\mathbf{1} - \hat{\boldsymbol{h}}\|_1}$ and $\beta = \|\hat{\boldsymbol{h}}\|_p^p / n$. Note that $\hat{\boldsymbol{w}} \in \Delta_{n,\beta}$. Consider $\boldsymbol{h}^*$ as defined in (26). Let $\bar{\boldsymbol{x}}^*$ be the average of datapoints in the set $\boldsymbol{G}^*$ defined in (28) and let $\boldsymbol{w}^* = \frac{\mathbf{1} - \boldsymbol{h}^*}{n - \|\boldsymbol{h}^*\|_0}$. Observe that on event $\mathcal{E}_1$, $\boldsymbol{w}^* \in \Delta_{n,\epsilon'}$. As a consequence of Lemma 6, on event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\lambda_{\max}(\Sigma_{\boldsymbol{w}^*}) \le \lambda_{\max} \left(\frac{1}{|\boldsymbol{I}^*|} \sum_{i=1}^n (1 - h_i^*)(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^\top \right) \le \frac{c_1^2 \sigma^2 n}{|\boldsymbol{I}^*|} \le \frac{c_1^2 \sigma^2}{1-\epsilon'} \tag{110}$$

$$\lambda_{\max}(\Sigma_{\hat{\boldsymbol{w}}}) \le \lambda_{\max} \left(\frac{1}{n - \|\boldsymbol{h}\|_1} \sum_{i=1}^n (1 - \hat{h}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{x}})(\boldsymbol{y}_i - \hat{\boldsymbol{x}})^\top \right) \le \frac{c_1^2 \sigma^2 n}{n - \|\boldsymbol{h}\|_p^p} = \frac{c_1^2 \sigma^2}{1-\beta}. \tag{111}$$

From Lemma 2, we obtain

$$\mathrm{TV}(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \le \frac{\max(\beta, \epsilon')}{1 - \min(\beta, \epsilon')}. \tag{112}$$

Consider the case $\beta \leq \epsilon' < 1/2$. This implies $\mathrm{TV}(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \leq \frac{\epsilon'}{1-\beta} < 1$. From Lemma 1, on event $\mathcal{E}_1 \cap \mathcal{E}_2$, we get

$$\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}^*\|_2 \leq \left(\sqrt{\frac{c_1^2 \sigma^2}{1-\epsilon'}} + \sqrt{\frac{c_1^2 \sigma^2}{1-\beta}} \right) \sqrt{\frac{\epsilon'}{1-\epsilon'-\beta}} \tag{113}$$

$$\leq \frac{2c_1 \sigma \sqrt{\epsilon'}}{\sqrt{(1-\epsilon')(1-2\epsilon')}}. \tag{114}$$

Consider the case $\epsilon' \leq \beta < 1 - \epsilon'$. This implies $\mathrm{TV}(\hat{\boldsymbol{w}}, \boldsymbol{w}^*) \leq \frac{\beta}{1-\epsilon'} < 1$. From Lemma 1, on event $\mathcal{E}_1 \cap \mathcal{E}_2$, we get

$$\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}^*\|_2 \leq \left(\sqrt{\frac{c_1^2 \sigma^2}{1-\epsilon'}} + \sqrt{\frac{c_1^2 \sigma^2}{1-\beta}} \right) \sqrt{\frac{\beta}{1-\epsilon'-\beta}}. \tag{115}$$

Consequently, on the event $\mathcal{E}$ defined in (35), using Lemma 8, (31) and applying triangle inequality, we obtain that with probability at least $1 - 4\delta$

$$\|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_2 \leq \|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}^*\|_2 + \sqrt{\frac{c_1^2 \sigma^2}{1-\alpha} \cdot \frac{\epsilon}{1-\epsilon}} + \sigma \sqrt{\alpha \delta} \left(1 + 2\sqrt{\frac{c_1'}{e \log(d/\delta)}} \right).$$

Let $(\boldsymbol{h}^{\mathrm{opt}}, \boldsymbol{x}^{\mathrm{opt}})$ be an optimal solution to (4). From Lemma 6 we have that $\left(\boldsymbol{h}^{\mathrm{opt}}, \dfrac{\sum_{i=1}^{n} (1-h_i^{\mathrm{opt}}) \boldsymbol{y}_i}{\sum_{i=1}^{n} (1-h_i^{\mathrm{opt}})} \right)$ is also an optimal solution. Note that on the event $\mathcal{E}$, we have that $(\boldsymbol{h}^*, \boldsymbol{\mu})$ is a feasible pair for (4). Hence,

$$\|\boldsymbol{h}^{\mathrm{opt}}\|_p \leq \|\boldsymbol{h}^*\|_p \leq (\epsilon' n)^{1/p}. \tag{116}$$

This implies

$$\left(\boldsymbol{h}^{\mathrm{opt}}, \frac{\sum_{i=1}^{n} (1 - h_i^{\mathrm{opt}}) \boldsymbol{y}_i}{\sum_{i=1}^{n} (1 - h_i^{\mathrm{opt}})} \right) \in \mathcal{S}'. \tag{117}$$

$\square$

E. Proof of Theorem 3

Proof. We prove the result by the method of induction.

Let $\boldsymbol{x}^{(0)}$ be the coordinate-wise median of the corrupted sample. It is easy to check that under the conditions stated in Theorem 3, it follows that $c_3 \leq c_1$ and $\epsilon \leq (1-\alpha)(1-\epsilon)$. Note that if $n \geq 90 \log\left(\frac{d}{\delta}\right)$, then by Lemma 3, Lemma 8, (31) and triangle inequality, we have that the following holds with probability at least $1 - \delta$:

$$\|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}^*\|_2 = \|\boldsymbol{x}^{(0)} - \boldsymbol{\mu} + \boldsymbol{\mu} - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}^*\|_2 \tag{118}$$

$$\leq \|\boldsymbol{x}^{(0)} - \boldsymbol{\mu}\|_2 + \|\boldsymbol{\mu} - \bar{\boldsymbol{x}}\|_2 + \|\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}^*\|_2 \tag{119}$$

$$\leq 3\sigma\sqrt{d} + \sigma c_3 + \sigma \sqrt{\frac{c_1^2 \sigma^2}{1-\alpha} \cdot \frac{\epsilon}{1-\epsilon}} \leq \sigma(3\sqrt{d} + 2c_1) = \sigma c_2^{(0)}. \tag{120}$$

Let $\mathcal{E}'$ be the event

$$\|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}^*\|_2 \leq \sigma c_2^{(0)}. \tag{121}$$

All the following statements hold on the event $\mathcal{E} \cup \mathcal{E}'$, where $\mathcal{E}$ is defined in (35). Also note that $\mathsf{P}(\mathcal{E} \cup \mathcal{E}') \geq 1 - 5\delta$, when $n \geq \max\left\{90, \frac{2e}{c_1'\delta^2}d\right\}\log\left(\frac{d}{\delta}\right)$.

Suppose $\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^*\|_2 \leq c_2^{(t)}\sigma$ and $\|\boldsymbol{h}^{(t-1)}\|_p^p \leq \epsilon'n$. Let $\boldsymbol{h}^{(t)}$ be an optimal solution to

$$\min_{\boldsymbol{h}} \|\boldsymbol{h}\|_p \tag{122}$$

$$\text{s.t. } \lambda_{\max}\left(\sum_{i=1}^{n}(1 - h_i)(\boldsymbol{y}_i - \boldsymbol{x}^{(t)})(\boldsymbol{y}_i - \boldsymbol{x}^{(t)})^\top\right) \leq \left(c_1^2 + (c_2^{(t)})^2\right)\sigma^2 n \tag{123}$$

$$0 \leq h_i \leq 1, \ \forall i \in [n]. \tag{124}$$

From Lemma 7, we have that $\boldsymbol{h}^*$ is a feasible point for the above optimization problem. Hence,

$$\|\boldsymbol{h}^{(t)}\|_p \leq \|\boldsymbol{h}^*\|_p \leq (\epsilon'n)^{1/p}. \tag{125}$$

Since $0 \leq h_i^{(t)} \leq 1$ for all i, we have

$$\left[\sum_{i=1}^{n} h_i^{(t)}\right]^{1/p} \leq \left[\sum_{i=1}^{n}\left(h_i^{(t)}\right)^p\right]^{1/p} \leq (\epsilon'n)^{1/p}. \tag{126}$$

This implies

$$\|\boldsymbol{h}^{(t)}\|_1 \leq \epsilon'n. \tag{127}$$

Let $\boldsymbol{w}$ be such that

$$w_i = \frac{(1 - h_i^{(t)})\mathbf{1}\{h_i^{(t)} \leq \tau\}}{\sum\limits_{i=1}^{n}(1 - h_i^{(t)})\mathbf{1}\{h_i^{(t)} \leq \tau\}}. \tag{128}$$

By Lemma 9, we have that $\boldsymbol{w} \in \Delta_{n, \frac{\epsilon'}{\tau}}$. Now we follow the proof of Theorem 2. Let $\boldsymbol{x}^{(t+1)} = \sum\limits_{i=1}^{n} w_i \boldsymbol{y}_i$. Observe that $\boldsymbol{w}^* \in \Delta_{n,\epsilon'}$. As a consequence of Lemma 6, we have

$$\lambda_{\max}(\Sigma_{\boldsymbol{w}^*}) \leq \lambda_{\max}\left(\frac{1}{|\boldsymbol{I}^*|}\sum_{i=1}^{n}(1 - h_i^*)(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^\top\right) \leq \frac{c_1^2\sigma^2 n}{|\boldsymbol{I}^*|} \leq \frac{c_1^2\sigma^2}{1 - \epsilon'}, \tag{129}$$

$$\lambda_{\max}(\Sigma_{\boldsymbol{w}}) = \lambda_{\max}\left(\frac{1}{\sum\limits_{i=1}^{n}(1 - h_i^{(t)})\mathbf{1}\{h_i^{(t)} \leq \tau\}}\sum_{i=1}^{n}(1 - h_i^{(t)})\mathbf{1}\{h_i^{(t)} \leq \tau\}(\boldsymbol{y}_i - \boldsymbol{x}^{(t+1)})(\boldsymbol{y}_i - \boldsymbol{x}^{(t+1)})^\top\right) \tag{130}$$

$$\leq \lambda_{\max}\left(\frac{1}{\sum\limits_{i=1}^{n}(1 - h_i^{(t)})\mathbf{1}\{h_i^{(t)} \leq \tau\}}\sum_{i=1}^{n}(1 - h_i^{(t)})\mathbf{1}\{h_i^{(t)} \leq \tau\}(\boldsymbol{y}_i - \boldsymbol{x}^{(t)})(\boldsymbol{y}_i - \boldsymbol{x}^{(t)})^\top\right) \tag{131}$$

$$\leq \lambda_{\max}\left(\frac{1}{(1 - \epsilon'/\tau)n}\sum_{i=1}^{n}(1 - h_i^{(t)})(\boldsymbol{y}_i - \boldsymbol{x}^{(t)})(\boldsymbol{y}_i - \boldsymbol{x}^{(t)})^\top\right) \tag{132}$$

$$\leq \frac{\left(c_1^2 + (c_2^{(t)})^2\right)\sigma^2}{1 - \epsilon'/\tau}. \tag{133}$$

From Lemma 2, we obtain

$$\text{TV}(\boldsymbol{w}, \boldsymbol{w}^*) \leq \frac{\max(\epsilon'/\tau, \epsilon')}{1 - \min(\epsilon'/\tau, \epsilon')} = \frac{\epsilon'/\tau}{1 - \epsilon'}. \tag{134}$$

From Lemma 1, we get

$$\|\boldsymbol{x}^{(t+1)} - \bar{\boldsymbol{x}}^*\| \leq (\sqrt{\lambda_{\max}(\Sigma_{\boldsymbol{w}})} + \sqrt{\lambda_{\max}(\Sigma_{\boldsymbol{w}^*})})\sqrt{\frac{\text{TV}(\boldsymbol{w}, \boldsymbol{w}^*)}{1 - \text{TV}(\boldsymbol{w}, \boldsymbol{w}^*)}} \tag{135}$$

$$\leq \left(\sqrt{\frac{\left(c_1^2 + (c_2^{(t)})^2\right)\sigma^2}{1 - \frac{\epsilon'}{\tau}}} + \sqrt{\frac{c_1^2\sigma^2}{1 - \epsilon'}}\right)\sqrt{\frac{\epsilon'/\tau}{1 - \epsilon' - \epsilon'/\tau}} \tag{136}$$

$$\leq \left(\frac{(c_1 + c_2^{(t)})\sigma}{\sqrt{1 - \epsilon'/\tau}} + \frac{c_1\sigma}{\sqrt{1 - \epsilon'}}\right)\sqrt{\frac{\epsilon'/\tau}{1 - \epsilon' - \epsilon'/\tau}} \tag{137}$$

$$= \sigma(\gamma(\epsilon')c_2^{(t)} + \beta(\epsilon')) \tag{138}$$

$$\leq \sigma(\gamma(\check{\epsilon})c_2^{(t)} + \beta(\check{\epsilon})) \tag{139}$$

$$= \sigma c_2^{(t+1)}. \tag{140}$$

We established that $\|\boldsymbol{x}^{(t+1)} - \bar{\boldsymbol{x}}^*\|_2 \leq \sigma c_2^{(t+1)}$ and $\|\boldsymbol{h}^{(t)}\|_p^p \leq \epsilon'n$. Hence, by the principle of mathematical induction, the result follows. It is easy to check that $\gamma(\check{\epsilon}) < 1$ holds if and only if $\check{\epsilon} < f(\tau)$. Furthermore, $\check{\epsilon} < f(\tau)$ implies $\check{\epsilon} < \tau$. Thus, we have that

$$\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^*\|_2 \leq \sigma\left(\gamma(\epsilon')c_2^{(t-1)} + \beta(\epsilon')\right) \tag{141}$$

$$= \sigma\left[\gamma(\epsilon')\left(c_2^{(0)}\gamma(\check{\epsilon})^{t-1} + \frac{1 - \gamma(\check{\epsilon})^{t-1}}{1 - \gamma(\check{\epsilon})}\beta(\check{\epsilon})\right) + \beta(\epsilon')\right]. \tag{142}$$

Consequently, using Lemma 8, (31) and applying triangle inequality, we obtain that with probability at least $1 - 5\delta$

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{\mu}\|_2 \leq \sigma\left[\gamma(\epsilon')\left(c_2^{(0)}\gamma(\check{\epsilon})^{t-1} + \frac{1 - \gamma(\check{\epsilon})^{t-1}}{1 - \gamma(\check{\epsilon})}\beta(\check{\epsilon})\right) + \beta(\epsilon')\right] + c_1\sigma\sqrt{\frac{\epsilon}{(1 - \alpha)(1 - \epsilon)}} + c_3\sigma.$$

It is easy to see that for $T = 1 + \frac{\log c_2^{(0)}}{|\log \gamma(\check{\epsilon})|}$, we have

$$\|\boldsymbol{x}^{(T)} - \boldsymbol{\mu}\|_2 \leq \sigma\left[\gamma(\epsilon')\left(1 + \frac{\beta(\check{\epsilon})}{1 - \gamma(\check{\epsilon})}\right) + \beta(\epsilon')\right] + c_1\sigma\sqrt{\frac{\epsilon}{(1 - \alpha)(1 - \epsilon)}} + c_3\sigma = O(\sigma\sqrt{\epsilon'}). \tag{143}$$

$\square$

F. Solving ℓ_1 objective via Packing SDP

$$\min_{\boldsymbol{h}}\|\boldsymbol{h}\|_1 \tag{144}$$

$$s.t.\ 0 \leq h_i \leq 1, \forall i,$$

$$\lambda_{\max}\left(\sum_{i=1}^{n}(1 - h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \leq cn\sigma^2.$$

Define the vector $\boldsymbol{w}$ with $w_i \triangleq 1 - h_i$. Since $0 \leq h_i \leq 1$, we have $0 \leq w_i \leq 1$. Further, $\|\boldsymbol{h}\|_1 = \sum_{i=1}^{n} h_i = \sum_{i=1}^{n}(1 - w_i) = n - \sum_{i=1}^{n} w_i = n - \mathbf{1}^{\top}\boldsymbol{w}$. Therefore, solving (144) is equivalent to solving the following:

$$\max_{\boldsymbol{w}} \ \mathbf{1}^{\top}\boldsymbol{w} \tag{145}$$

$$s.t. \ 0 \leq w_i \leq 1, \forall i,$$

$$\lambda_{\max}\left(\sum_{i=1}^{n} w_i(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^{\top}\right) \leq cn\sigma^2.$$

Then, we rewrite the constraints $0 \leq w_i \leq 1, \forall i$ as $0 \leq w_i$, and $\sum w_i e_i e_i^{\top} \preceq I_{n \times n}$, where e_i is the i-th standard basis vector in $\mathbb{R}^n$. This establishes the equivalence between (145) and (18).

G. Minimizing ℓ_p via iterative re-weighted ℓ_2

Consider ℓ_p $(0 < p < 1)$ in Step 1 of Algorithm 1. We have the following equivalent objective:

$$\min_{\boldsymbol{h}} \|\boldsymbol{h}\|_p^p \tag{146}$$

$$s.t. \ 0 \leq h_i \leq 1, \forall i,$$

$$\lambda_{\max}\left(\sum_{i=1}^{n}(1 - h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^{\top}\right) \leq c\sigma^2 n.$$

Note that $\|\boldsymbol{h}\|_p^p = \sum_{i=1}^{n} h_i^p = \sum_{i=1}^{n}(h_i^2)^{\frac{p}{2}}$. Consider that we employ the iterative re-weighted ℓ_2 technique [30], [31]. Then at $(k+1)$-th inner iteration, we construct a tight upper bound on $\|\boldsymbol{h}\|_p^p$ at $\boldsymbol{h}^{(k)^2}$ as

$$\sum_{i=1}^{n}\left[\left(h_i^{(k)^2}\right)^{\frac{p}{2}} + \frac{p}{2}\left(h_i^{(k)^2}\right)^{\frac{p}{2}-1}\left(h_i^2 - h_i^{(k)^2}\right)\right]. \tag{147}$$

We minimize this upper bound:

$$\min_{\boldsymbol{h}} \sum_{i=1}^{n}\left(h_i^{(k)^2}\right)^{\frac{p}{2}-1} h_i^2 \tag{148}$$

$$s.t. \ 0 \leq h_i \leq 1, \forall i,$$

$$\lambda_{\max}\left(\sum_{i=1}^{n}(1 - h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^{\top}\right) \leq cn\sigma^2,$$

Define $u_i = \left(h_i^{(k)}\right)^{\frac{p}{2}-1}$, the objective in (148) becomes $\sum_{i=1}^{n} u_i^2 h_i^2$. Define the vector $\boldsymbol{w}$ with $w_i \triangleq 1 - h_i$. Since $0 \leq h_i \leq 1$, we have $0 \leq w_i \leq 1$. Further, $\sum_{i=1}^{n} u_i^2 h_i^2 = \sum_{i=1}^{n} u_i^2(1 - w_i)^2 = \sum_{i=1}^{n}(u_i - u_i w_i)^2$. So, solving (148) is equivalent to solving the following:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n}(u_i - u_i w_i)^2 \tag{149}$$

$$s.t. \ 0 \leq w_i \leq 1, \forall i,$$

$$\lambda_{\max}(\sum_{i=1}^{n} w_i(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^{\top}) \leq cn\sigma^2.$$

Further, define the vector $\boldsymbol{z}$ with $z_i \triangleq u_i w_i$. Then solving (149) is equivalent to solving the following:

$$\min_{\boldsymbol{z}} \|\boldsymbol{u} - \boldsymbol{z}\|_2^2 \tag{150}$$

$$s.t.\ 0 \le z_i \le u_i, \forall i,$$

$$\lambda_{\max}\left(\sum_{i=1}^n z_i[(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top/u_i]\right) \le cn\sigma^2.$$

Then, we rewrite the constraints $0 \le z_i \le u_i, \forall i$ as $0 \le z_i$, and $\sum_{i=1}^n z_i e_i e_i^\top \preceq \operatorname{diag}(\boldsymbol{u})$, where e_i is the i-th standard basis vector in $\mathbb{R}^n$. Finally, we can turn (150) into the following least squares problem with semidefinite cone constraints:

$$\min_{\boldsymbol{z}} \|\boldsymbol{u} - \boldsymbol{z}\|_2^2 \tag{151}$$

$$s.t.\ \ z_i \ge 0, \forall i,$$

$$\sum_{i=1}^n z_i \begin{bmatrix} e_i e_i^\top & \\ & (\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top/u_i \end{bmatrix} \preceq \begin{bmatrix} \operatorname{diag}(\boldsymbol{u}) & \\ & cn\sigma^2 I_{d\times d} \end{bmatrix}.$$

H. Solving weighted ℓ_1 objective via Packing SDP

Consider ℓ_p ($0 < p < 1$) in Step 1 of Algorithm 1 (see objective (146)). If we employ iterative re-weighted ℓ_1 approach [32], [30], we need to solve the following problem:

$$\min_{\boldsymbol{h}} \sum_{i=1}^n u_i h_i \tag{152}$$

$$s.t.\ 0 \le h_i \le 1, \forall i,$$

$$\lambda_{\max}\left(\sum_{i=1}^n (1 - h_i)(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \le cn\sigma^2,$$

where u_i is the weight on corresponding h_i. Define the vector $\boldsymbol{w}$ with $w_i \triangleq 1 - h_i$. Since $0 \le h_i \le 1$, we have $0 \le w_i \le 1$. Further, $\sum_{i=1}^n u_i h_i = \sum_{i=1}^n u_i(1 - w_i) = \sum_{i=1}^n u_i - \sum_{i=1}^n u_i w_i$. So, solving (152) is equivalent to solving the following:

$$\max_{\boldsymbol{w}} \boldsymbol{u}^\top \boldsymbol{w} \tag{153}$$

$$s.t.\ 0 \le w_i \le 1, \forall i,$$

$$\lambda_{\max}\left(\sum_{i=1}^n w_i(\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top\right) \le cn\sigma^2.$$

Then, we rewrite the constraints $0 \le w_i \le 1, \forall i$ as $0 \le w_i$, and $\sum w_i e_i e_i^\top \preceq I_{n\times n}$, where e_i is the i-th standard basis vector in $\mathbb{R}^n$. Finally, we can turn (153) into the following Packing SDP:

$$\max_{\boldsymbol{w}} \boldsymbol{u}^\top \boldsymbol{w} \tag{154}$$

$$s.t.\ w_i \ge 0, \forall i,$$

$$\sum_{i=1}^n w_i \begin{bmatrix} e_i e_i^\top & \\ & (\boldsymbol{y}_i - \boldsymbol{x})(\boldsymbol{y}_i - \boldsymbol{x})^\top \end{bmatrix} \preceq \begin{bmatrix} I_{n\times n} & \\ & cn\sigma^2 I_{d\times d} \end{bmatrix}.$$

I. Corrupted image dataset

We use real face images to test the effectiveness of the robust mean estimation methods. The average face of particular regions or certain groups of people is useful for many social and psychological studies [34]. Here we use 100 frontal human face images from Brazilian face database[4] as inliers. For the outliers, we choose 15 face images of cat and dog from CIFAR10 [35]. In order to run the CDG method [24], we scale the size of images to 18×15 pixels, so the dimension of each datapoint is 270. Fig. 2 and Fig. 3 show the sample inlier and outlier images. Fig. 4 shows the oracle solution (the average of the 100 inlier human faces) and the estimated mean by each method, as well as their ℓ_2 distances to the oracle solution. The proposed ℓ_1 and ℓ_p methods achieve smaller recovery error than the state-of-the-art methods. The estimated mean faces by the proposed methods also look visually similar to the oracle solution, which illustrates the efficacy of the proposed ℓ_1 and ℓ_p methods.

Fig. 2: Sample inlier human face images.

Fig. 3: Sample outlier cat and dog face images from CIFAR10.

Fig. 4: Reconstructed mean face and its recovery error by each method.

---

[4]https://fei.edu.br/ cet/facedatabase.html