Outlier-Robust Sparse Estimation via Non-Convex Optimization

Yu Cheng
Brown University
Providence, RI 02912
yu_cheng@brown.edu

Ilias Diakonikolas
University of Wisconsin-Madison
Madison, WI 53706
ilias@cs.wisc.edu

Rong Ge
Duke University
Durham, NC 27708
rongge@cs.duke.edu

Shivam Gupta
University of Texas at Austin
Austin, TX 78712
shivamgupta@utexas.edu

Daniel M. Kane
University of California, San Diego
La Jolla, CA 92093
dakane@cs.ucsd.edu

Mahdi Soltanolkotabi
University of Southern California
Los Angeles, CA 90089
soltanol@usc.edu

Abstract

We explore the connection between outlier-robust high-dimensional statistics and non-convex optimization in the presence of sparsity constraints, with a focus on the fundamental tasks of robust sparse mean estimation and robust sparse PCA. We develop novel and simple optimization formulations for these problems such that *any* approximate stationary point of the associated optimization problem yields a near-optimal solution for the underlying robust estimation task. As a corollary, we obtain that any first-order method that efficiently converges to stationarity yields an efficient algorithm for these tasks.[1] The obtained algorithms are simple, practical, and succeed under broader distributional assumptions compared to prior work.

1 Introduction

In several modern machine learning (ML) applications, such as ML security [BNJT10, BNL12, SKL17, DKK$^+$19a] and exploratory analysis of real datasets, e.g., in population genetics [RPW$^+$02, PLJD10, LAT$^+$08, DKK$^+$17], typical datasets contain a non-trivial fraction of arbitrary (or even adversarial) outliers. Robust statistics [HRRS86, HR09] is the subfield of statistics aiming to design estimators that are tolerant to a *constant fraction* of outliers, independent of the dimensionality of the data. Early work in this field, see, e.g., [Tuk60, Hub64, Tuk75] developed sample-efficient robust estimators for various basic tasks, alas with runtime exponential in the dimension.

During the past five years, a line of work in computer science, starting with [DKK$^+$16, LRV16], has developed the first *computationally efficient* robust high-dimensional estimators for a range of tasks. This progress has led to a revival of robust statistics from an algorithmic perspective (see, e.g., [DK19, DKK$^+$21] for surveys on the topic). In this work, we focus on high-dimensional estimation tasks in the presence of sparsity constraints. To rigorously study these problems, we need to formally define the model of data corruption. Throughout this work, we work with the following standard contamination model.

[1] An implementation of our algorithms is available at https://github.com/guptashvm/Sparse-GD.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Definition 1.1 (Strong Contamination Model, see [DKK$^+$16])**.** *Given a parameter $0 < \epsilon < 1/2$ and a distribution family $\mathcal{D}$ on $\mathbb{R}^d$, the* adversary *operates as follows: The algorithm specifies a number of samples n, and n samples are drawn from some unknown $D \in \mathcal{D}$. The adversary is allowed to inspect the samples, remove up to ϵn of them and replace them with arbitrary points. This modified set of n points is then given as input to the algorithm. We say that a set of samples is* ϵ-corrupted *if it is generated by the above process.*

High-dimensional robust statistics is algorithmically challenging because the natural optimization formulations of such tasks are typically non-convex. The recent line of work on algorithmic robust statistics has led to a range of sophisticated algorithms. In some cases, such algorithms require solving large convex relaxations, rendering them computationally prohibitive for large-scale problems. In other cases, they involve a number of hyper-parameters that may require careful tuning. Motivated by these shortcomings of known algorithms, recent work [CDGS20, ZJS20] established an intriguing connection between high-dimensional robust estimation and non-convex optimization. The high-level idea is quite simple: Even though typical robust statistics tasks lead to non-convex formulations, it may still be possible to leverage the underlying structure to show that standard first-order methods provably and efficiently reach near-optimal solutions. Indeed, [CDGS20, ZJS20] were able to prove such statements for robust mean estimation under natural distributional assumptions. Specifically, these works established that any (approximate) stationary point of a well-studied non-convex formulation for robust mean estimation yields a near-optimal solution for the underlying robust estimation task.

In this work, we continue this line of work with a focus on *sparse* estimation tasks. Leveraging sparsity in high-dimensional datasets is a fundamental problem of significant practical importance. Various formalizations of this problem have been investigated in statistics and machine learning for at least the past two decades (see, e.g., [HTW15] for a textbook on the topic). We focus on *robust sparse mean estimation* and *robust sparse PCA*. Sparse mean estimation is arguably one of the most fundamental sparse estimation tasks and is closely related to the Gaussian sequence model [Tsy08, Joh17]. The task of sparse PCA in the spiked covariance model, initiated in [Joh01], has been extensively investigated (see Chapter 8 of [HTW15] and references therein).

In the context of robust sparse mean estimation, we are given an ϵ-corrupted set of samples from a distribution with unknown mean $\mu \in \mathbb{R}^d$ where μ is k-sparse, and we want to compute a vector $\widehat{\mu}$ close to μ. In the context of robust sparse PCA (in the spiked covariance model), we are given an ϵ-corrupted set of samples from a distribution with covariance matrix $I + \rho v v^T$, where $v \in \mathbb{R}^d$ is k-sparse and the goal is to approximate v. It is worth noting that for both problems, we have access to much fewer samples compared to the non-sparse case (roughly $O(k^2 \log d)$ instead of $\Omega(d)$). Consequently, the design and analysis of optimization formulations for robust sparse estimation requires new ideas and techniques that significantly deviate from the standard (non-sparse) case.

1.1 Our Results and Contributions

We show that standard first-order methods lead to robust and efficient algorithms for sparse mean estimation and sparse PCA. Our main contribution is to propose novel (non-convex) formulations for these robust estimation tasks, and to show that *approximate stationarity suffices for near-optimality*. We establish landscape results showing that *any* approximate stationary point of our objective function yields a near-optimal solution for the underlying robust estimation task. Consequently, gradient descent (or any other methods converging to stationarity) can solve these problems.

Our results provide new insights and techniques in designing and analyzing (non-convex) optimization formulations of robust estimation tasks. Our formulations and structural results immediately lead to simple and practical algorithms for robust sparse estimation. Importantly, the gradient of our objectives can be computed efficiently via a small number of basic matrix operations. In addition to their simplicity and practicality, our methods provably succeed under more general distributional assumptions compared to prior work.

For robust sparse mean estimation and robust sparse PCA, our landscape results require deterministic conditions on the original set of good samples. We refer to these conditions as *stability conditions* (Definitions 2.1 and 2.2, formally defined in Section 2). At a high level, they state that the first and second moments of a set of samples are stable when *any* ϵ-fraction of the samples are removed. These stability conditions hold with high probability for a set of clean samples drawn from natural families of distributions (e.g., subgaussian).

For robust sparse mean estimation, we establish the following result.

Theorem 1.2 (Robust Sparse Mean Estimation)**.** *Let $0 < \epsilon < \epsilon_0$ for some universal constant ϵ_0 and let $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (k, ϵ, δ)-stable (per Definition 2.1) w.r.t. a distribution with unknown k-sparse mean $\mu \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$.* [2] *There is an algorithm that on inputs S, k, ϵ, and δ, runs in polynomial time and returns a k-sparse vector $\widehat{\mu} \in \mathbb{R}^d$ such that $\|\widehat{\mu} - \mu\|_2 \leq O(\delta)$.*

We emphasize that a key novelty of Theorem 1.2 is that the underlying algorithm is a *first-order method* applied to a *novel non-convex formulation* of the problem. The major advantage of our algorithm over prior work [BDLS17, DKK$^+$19b] is its simplicity, practicality, and the fact that it seamlessly applies to a wider class of distributions on the clean data.

As we will discuss in Section 3, when the ground-truth distribution D is subgaussian with unknown k-sparse mean $\mu \in \mathbb{R}^d$ and identity covariance, a set of $n = \widetilde{\Omega}(k^2 \log d/\epsilon^2)$ samples drawn from D is (k, ϵ, δ)-stable (Definition 2.1) with high probability for $\delta = O(\epsilon\sqrt{\log(1/\epsilon)})$. It follows as an immediate corollary of Theorem 1.2 that, given an ϵ-corrupted set of samples, we can compute a vector $\widehat{\mu}$ that is $O(\delta) = O(\epsilon\sqrt{\log(1/\epsilon)})$ close to the true mean μ. This sample complexity matches the known computational-statistical lower bounds [DKS17, BB20]. More generally, one can relax the concentration assumption on the clean data and obtain qualitatively similar error guarantees.

Next we state our main result for robust sparse PCA.

Theorem 1.3 (Robust Sparse PCA)**.** *Let $0 < \rho \leq 1$ and $0 < \epsilon < \epsilon_0$ for some universal constant ϵ_0. Let $G^\star$ be a set of n samples that is (k, ϵ, δ)-stable (as in Definition 2.2) w.r.t. a centered distribution with covariance $\Sigma = I + \rho v v^\top$, for an unknown k-sparse unit vector $v \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$. There is an algorithm that on inputs S, k, and ϵ, runs in polynomial time and returns a unit vector $u \in \mathbb{R}^d$ such that $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\delta/\rho})$.*

Interestingly, our algorithm for robust sparse PCA is a first-order method applied to a simple *convex* formulation of the problem. We view the existence of a convex formulation as an intriguing fact that, surprisingly, was not observed in prior work.

As we will discuss in Section 4, when the ground-truth distribution D is centered subgaussian with covariance $\Sigma = I + \rho v v^\top$, for an unknown k-sparse unit vector $v \in \mathbb{R}^d$, a set of $n = \widetilde{\Omega}(k^2 \log d/\epsilon^2)$ samples drawn from D is (k, ϵ, δ)-stable (Definition 2.2) with high probability for $\delta = O(\epsilon \log(1/\epsilon))$. Therefore, our algorithm outputs a vector that is $O(\sqrt{\epsilon \log(1/\epsilon)/\rho})$ close to the true direction v. The sample complexity in this case nearly matches the computational-statistical lower bound of $\Omega(k^2 \log d/\epsilon^2)$ [BR13] which holds even without corruptions. While the error guarantee of our algorithm is slightly worse compared to prior work [BDLS17, DKK$^+$19b] for Gaussian data (we get $O(\sqrt{\delta/\rho})$ rather than $O(\delta/\rho)$), we note that our algorithm works for a broader family of distributions.

Prior Work on Robust Sparse Estimation. We provide a detailed summary of prior work for comparison. [BDLS17] obtained the first sample-efficient and polynomial-time algorithms for robust sparse mean estimation and robust sparse PCA. These algorithms succeed for Gaussian inliers and inherently use the ellipsoid method. The separation oracle required for the ellipsoid algorithm turns out to be another convex program — corresponding to an SDP to solve sparse PCA. As a consequence, the running time of these algorithms, while polynomially bounded, is impractically high. [LLC19] proposed an algorithm for robust sparse mean estimation via iterative trimmed hard thresholding, which can only tolerate a *sub-constant* fraction of corruptions. [DKK$^+$19b] gave iterative spectral robust algorithms for sparse mean estimation and sparse PCA. These algorithms are still quite complex and are only shown to succeed under Gaussian inliers.

1.2 Overview of Our Approach

In this section, we give an overview of our approach for robust sparse mean estimation. At a very high level, we assign a nonnegative weight to each data point and try to find a good set of $(1-\epsilon)n$ samples. The constraint on the weight vector is that it represents at least a (fractional) set of $(1-\epsilon)$-portion of the input dataset. Formally, given n datapoints $(X_i)_{i=1}^n$, the goal is to find a weight vector $w \in \mathbb{R}^n$

[2] For two sets of samples S and T, we say S is an ϵ-corrupted version of T if $|S| = |T|$ and $|S \setminus T| \leq \epsilon|S|$.

such that $\mu_w = \sum_i w_i X_i$ is close to the true mean μ. The constraint on w is that it belongs to

$$\Delta_{n,\epsilon} = \left\{ w \in \mathbb{R}^n : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \tfrac{1}{(1-\epsilon)n} \ \forall i \right\} ,$$

which is the convex hull of all uniform distributions over subsets $S \subseteq [n]$ of size $|S| = (1-\epsilon)n$.

Let $\Sigma_w = \sum_i w_i (X_i - \mu_w)(X_i - \mu_w)^\top$ denote the weighted empirical covariance matrix. It is well-known that if one can find $w \in \Delta_{n,\epsilon}$ that minimizes the weighted empirical variance $v^\top \Sigma_w v$ for all k-sparse unit vectors v, then μ_w must be close to μ. Unfortunately, it is NP-Hard to find the sparse direction v with the largest variance. To get around this issue, [BDLS17] considered the following convex relaxation, minimizing the variance for convex combinations of sparse directions:

$$\min_w \max_{\operatorname{tr}(A)=1, \sum_{ij}|A_{ij}| \le k, A \succeq 0} (A \bullet \Sigma_w) \ . \tag{1}$$

Given w, the optimal A can be found using semidefinite programming (SDP). [ZJS20] observed that any stationary point w of (1) gives a good solution for robust sparse mean estimation. However, solving (1) requires convex programming to compute the gradient in each iteration. As explained in the proceeding discussion, our approach circumvents this shortcoming, leading to a formulation for which each gradient can be computed *using only basic matrix operations*.

In this work, we propose and analyze the following optimization formulation:

$$\min_w \ f(w) = \|\Sigma_w - I\|_{F,k,k} \quad \text{subject to } \ w \in \Delta_{n,\epsilon} \ ,$$

where $\|A\|_{F,k,k}$ is the Frobenius norm of the k^2 entries of A with largest magnitude, with the additional constraint that these k^2 entries are chosen from k rows with k entries in each row.

We prove that any stationary point of $f(w)$ yields a good solution for robust sparse mean estimation. Here we provide a brief overview of our proof (see Section 3 for more details). Given a weight vector w, we show that if w is not a good solution, then moving toward $w^\star$ (the weight vector corresponding to the uniform distribution on the clean input samples) will decrease the objective value. Formally, we will show that, for any $0 < \eta < 1$,

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta \Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top \ .$$

We can then take $\|\cdot\|_{F,k,k}$ norm on both sides (after subtracting I) and show that the third term can be essentially ignored. If the third term were not there, we would have

$$\begin{aligned} f((1-\eta)w + \eta w^\star) &= \left\|\Sigma_{(1-\eta)w+\eta w^\star} - I\right\|_{F,k,k} \\ &\le (1-\eta)\|\Sigma_w - I\|_{F,k,k} + \eta \|\Sigma_{w^\star} - I\|_{F,k,k} = (1-\eta) f(w) + \eta f(w^\star) \ . \end{aligned}$$

Therefore, if w is a bad solution with $f(w)$ much larger than $f(w^\star)$, then w cannot be a stationary point because f decreases when we move from w to $(1-\eta)w + \eta w^\star$.

Remark 1.4. The technical overview for robust sparse PCA follows a similar high-level approach, but is somewhat more technical. It is deferred to Section 4.

Roadmap. In Section 2, we introduce basic notations and the deterministic stability conditions that we require on the good samples. We present our algorithms and analysis for robust sparse mean estimation in Section 3 and robust sparse PCA in Section 4. In Section 5, we evaluate our algorithm on synthetic datasets and show that it achieves good statistical accuracy under various noise models.

2 Preliminaries and Background

Notation. For a positive integer n, let $[n] = \{1, \dots, n\}$. For a vector v, we use $\|v\|_0$, $\|v\|_1$, $\|v\|_2$, and $\|v\|_\infty$ for the number of non-zeros, the ℓ_1, ℓ_2, and ℓ_∞ norm of v respectively. Let I be the identity matrix. For a matrix A, we use $\|A\|_2$, $\|A\|_F$, $\operatorname{tr}(A)$ for the spectral norm, Frobenius norm, and trace of A respectively. For two vectors x, y, let $x^\top y$ denote their inner product. For two matrices A, B, we use $A \bullet B = \operatorname{tr}(A^\top B)$ for their entrywise inner product. A matrix A is said to be positive semidefinite (PSD) if $x^\top A x \ge 0$ for all x. We write $A \preceq B$ iff $(B - A)$ is PSD.

For a vector $w \in \mathbb{R}^n$, let $\operatorname{diag}(w) \in \mathbb{R}^{n \times n}$ denote a diagonal matrix with w on the diagonal. For a matrix $A \in \mathbb{R}^{n \times n}$, let $\operatorname{diag}(A) \in \mathbb{R}^n$ denote a column vector with the diagonal of A. For a vector $v \in \mathbb{R}^d$ and a set $S \subseteq [d]$, we write $v_S \in \mathbb{R}^d$ for a vector that is equal to v on S and zero everywhere else. Similarly, for a matrix $A \in \mathbb{R}^{d \times d}$ and a set $S \subseteq ([d] \times [d])$, we write A_S for a matrix that is equal to A on S and zero everywhere else.

For a vector v, we define $\|v\|_{2,k} = \max_{|S|=k} \|v_S\|_2$ to be the maximum ℓ_2-norm of any k entries of v. For a matrix A, we define $\|A\|_{F,k^2}$ to be the maximum Frobenius norm of any k^2 entries of A. Moreover, we define $\|A\|_{F,k,k}$ to be the maximum Frobenius norm of any k^2 entries with the extra requirement that these entries must be chosen from k rows with k entries in each row. Formally,

$$\|A\|_{F,k^2} = \max_{|Q|=k^2} \|A_Q\|_F \quad \text{and} \quad \|A\|_{F,k,k}^2 = \max_{|S|=k} \sum_{i \in S} \|A_i\|_{2,k}^2 \text{ where } A_i \text{ is } i\text{-th row of } A \ . \quad (2)$$

Sample Reweighting Framework. We use n for the number of samples, d for the dimension, and ϵ for the fraction of corrupted samples. For sparse estimation, we use k for the sparsity of the ground-truth parameters. We use $G^\star$ for the original set of n good samples. We use $S = G \cup B$ for the input samples after the adversary replaced ϵ-fraction of $G^\star$, where $G \subset G^\star$ is the set of remaining good samples and B is the set of bad samples (outliers) added by the adversary. Note that $|G| = (1-\epsilon)n$ and $|B| = \epsilon n$.

Given n samples $X_1, \ldots, X_n$, we write $X \in \mathbb{R}^{d \times n}$ as the sample matrix where the i-th column is X_i. For a weight vector $w \in \mathbb{R}^n$, we use $\mu_w = Xw = \sum_i w_i X_i$ for the weighted empirical mean and $\Sigma_w = X \operatorname{diag}(w) X - \mu_w \mu_w^\top = \sum_i w_i (X_i - \mu_w)(X_i - \mu_w)^\top$ for the weighted empirical covariance. Let $\Delta_{n,\epsilon}$ be the convex hull of all uniform distributions over subsets $S \subseteq [n]$ of size $|S| = (1-\epsilon)n$: $\Delta_{n,\epsilon} = \{w \in \mathbb{R}^n : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)n} \ \forall i\}$, In other words, every $w \in \Delta_{n,\epsilon}$ corresponds to a fractional set of $(1-\epsilon)n$ samples. We use $w^\star$ to denote the uniform distribution on G (the remaining good samples in S).

Deterministic Stability Conditions. For robust sparse mean estimation and robust sparse PCA, we require the following conditions respectively.

Definition 2.1 (Stability Conditions for Sparse Mean)**.** *A set of n samples $G^\star = (X_i)_{i=1}^n$ is said to be (k, ϵ, δ)-stable (w.r.t. a distribution with mean μ) iff for any weight vector $w \in \Delta_{n,2\epsilon}$, we have $\|\mu_w - \mu\|_{2,k} \le \delta$ and $\|\Sigma_w - I\|_{F,k,k} \le \delta^2/\epsilon$, where μ_w and Σ_w are the weighted empirical mean and covariance matrix respectively, and the $\|\cdot\|_{F,k,k}$ norm is defined in Equation* (2)*.*

Definition 2.2 (Stability Conditions for Sparse PCA)**.** *A set of n samples $G^\star = (X_i)_{i=1}^n$ is (k, ϵ, δ)-stable (w.r.t. a centered distribution with covariance $I + \rho v v^\top$) iff for any weight vector $w \in \Delta_{n,2\epsilon}$, $\left\|M_w - (I + \rho v v^\top)\right\|_{F,2k^2} \le \delta$, where $M_w = \sum_i w_i X_i X_i^\top$ and the $\|\cdot\|_{F,2k^2}$ norm is defined in Equation* (2)*.*

First-Order Stationary Points. We give a formal definition of the notion of (approximate) first-order stationary point that we use in this paper.

Definition 2.3 (Approximate Stationary Points)**.** *Fix a convex set $\mathcal{K}$ and a differentiable function f. For $\gamma \ge 0$, we say that $x \in \mathcal{K}$ is a γ-stationary point of f iff the following condition holds: For any unit vector u where $x + \alpha u \in \mathcal{K}$ for some $\alpha > 0$, we have $u^\top \nabla f(x) \ge -\gamma$.*

We note that the objective functions studied in this paper are not everywhere differentiable. This is because, taking the $\|\cdot\|_{F,k,k}$ norm as an example, there can be ties in choosing the largest k^2 entries. When the function f is not differentiable, we use ∇f informally to denote an element of the sub-differential. We will show in Appendix C that, while f is not differentiable, it does have a nonempty subdifferential, as it can be written as the pointwise maximum of differentiable functions.

3 Robust Sparse Mean Estimation

In this section, we present our non-convex approach for robust sparse mean estimation. We will optimize the following objective, where $\|\cdot\|_{F,k,k}$ is defined in Equation (2):

$$\min_w \ f(w) = \|\Sigma_w - I\|_{F,k,k} \quad \text{subject to} \ \ w \in \Delta_{n,\epsilon} \ . \quad (3)$$

We will show that the objective function (3) has no bad stationary points (Theorem 3.1). In other words, *every* first-order stationary point of f yields a good solution for robust sparse mean estimation.

Our algorithm is stated in Algorithm 1. As a consequence of our landscape result (Theorem 3.1), we know that Algorithm 1 works *no matter how* we find a stationary point of f (because any stationary point works), so we intentionally did not specify how to find such a point. As a simple illustration, we show that (projected) gradient descent can be used to minimize f. The convergence analysis and iteration complexity are provided in Appendix C.

Algorithm 1: Robust sparse mean estimation.

```
Input: $k > 0$, $0 < \epsilon < \epsilon_0$, and an $\epsilon$-corrupted set of samples $(X_i)_{i=1}^n$ drawn from a distribution with $k$-sparse mean $\mu$. [3]
Output: a vector $\widehat{\mu}$ that is close to $\mu$.
Find a first-order stationary point $w \in \Delta_{n,\epsilon}$ of the objective $\min_w f(w) = \|\Sigma_w - I\|_{F,k,k}$.
Return $\widehat{\mu} = (\mu_w)_Q$ where $Q$ is a set of $k$ entries of $\mu_w$ with largest magnitude.
```

Formally, we first prove that Algorithm 1 can output a vector $\widehat{\mu} \in \mathbb{R}^d$ that is close to μ in $\|\cdot\|_{2,k}$ norm, as long as the good samples satisfies the stability condition in Definition 2.1.

Theorem 3.1. *Fix $k > 0$, $0 < \epsilon < \epsilon_0$, and $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (k, ϵ, δ)-stable (as in Definition 2.1) w.r.t. a distribution with unknown k-sparse mean $\mu \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$. Let $f(w) = \|\Sigma_w - I\|_{F,k,k}$. Let $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2})$. Then, for any $w \in \Delta_{n,\epsilon}$ that is a γ-stationary point of $f(w)$, we have $\|\mu_w - \mu\|_{2,k} = O(\delta)$.*

Once we have a vector μ_w that is $O(\delta)$-close to μ in $\|\cdot\|_{2,k}$ norm, we can guarantee that a truncated version of μ_w (the output $\widehat{\mu}$ of Algorithm 1) is $O(\delta)$-close to μ in the ℓ_2-norm:

Lemma 3.2. *Fix two vectors x, y with $\|x\|_0 \le k$ and $\|x - y\|_{2,k} \le \delta$. Let z be a vector that keeps the k entries of y with largest absolute values and sets the rest to 0. We have $\|x - z\|_2 \le \sqrt{5}\delta$.*

Theorem 1.2 follows immediately from Theorem 3.1 and Lemma 3.2.

We can apply Theorem 1.2 to get an end-to-end result for subgaussian distributions. We show that the required stability conditions are satisfied with a small number of samples.

Lemma 3.3. *Fix $k > 0$ and $0 < \epsilon < \epsilon_0$. Let $G^\star$ be a set of n samples that are drawn i.i.d. from a subgaussian distribution with mean μ and covariance I. If $n = \Omega(k^2 \log d/\epsilon^2)$, then with probability at least $1 - \exp(-\Omega(k^2 \log d))$, $G^\star$ is (k, ϵ, δ)-stable (as in Definition 2.1) for $\delta = O(\epsilon \log(1/\epsilon))$.*

Combining Theorem 1.2 and Lemma 3.3, we know that given an ϵ-corrupted set of $O(k^2 \log d/\epsilon^2)$ samples drawn from a subgaussian distribution with k-sparse mean μ, the output of Algorithm 1 is $O(\epsilon\sqrt{\log(1/\epsilon)})$-close to μ in ℓ_2-norm.

In the rest of this section, we will prove Theorem 3.1. Omitted proofs in this section are in Appendix A.

We start with some intuition on why we choose our objective function (3). We would like to design $f(w) = g(\Sigma_w - I)$ to satisfy the following properties:

1. $g(\Sigma_w - I)$ is an upper bound on $v^\top(\Sigma_w - I)v$ for all k-sparse unit vectors $v \in \mathbb{R}^d$. This way, a small objective value implies that $\|\mu_w - \mu\|_{2,k}$ is small.
2. $g(\Sigma_{w^\star} - I)$ is small for $w^\star$ (the uniform distribution on G). This guarantees that a good w exists.
3. Triangle inequality on g. This allows us to upper bound the objective value when we move w toward $w^\star$ by the sum of $g(\cdot)$ of each term on the right-hand side:
$$\Sigma_{(1-\eta)w+\eta w^\star} - I = (1-\eta)(\Sigma_w - I) + \eta(\Sigma_{w^\star} - I) + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top .$$
4. $g(uu^\top)$ is close to $g(vv^\top)$ where v keeps only the k largest entries of u. We want to approximate μ in $\|\cdot\|_{2,k}$ norm, so intuitively $g(\Sigma_w - I)$ should depend only on the largest k entries of $(\mu_w - \mu)$.

[3] Without loss of generality we can assume that ϵ is given to the algorithm. This is because we can run a binary search to determine ϵ: if our guess of ϵ is too small, then the algorithm will output a w whose objective value $f(w)$ is much larger than it should be.

Our choice of $f(w) = g(\Sigma_w - I) = \|\Sigma - I\|_{F,k,k}$ is motivated by (and satisfies) all these properties.

Lemma 3.4. *Fix $A \in \mathbb{R}^{d\times d}$. We have $\left|v^\top A v\right| \le \|A\|_{F,k,k}$ for any k-sparse unit vector $v \in \mathbb{R}^d$.*

Lemma 3.5. *For any vector $v \in \mathbb{R}^d$, $\left\|vv^\top\right\|_{F,k,k} = \|v\|_{2,k}^2$.*

We now continue to present key technical lemmas for proving our main structural result (Theorem 3.1). Lemma 3.6 gives the weighted empirical covariance for a convex combination of two weight vectors.

Lemma 3.6. *Fix n samples $X_1, \dots, X_n \in \mathbb{R}^d$. Let $\overline{w}, \widehat{w} \in \mathbb{R}^n$ be two non-negative weight vectors with $\|\overline{w}\|_1 = \|\widehat{w}\|_1 = 1$. For any $\alpha, \beta \ge 0$ with $\alpha + \beta = 1$, letting $w = \alpha\overline{w} + \beta\widehat{w}$, we have*

$$\Sigma_w = \alpha\Sigma_{\overline{w}} + \beta\Sigma_{\widehat{w}} + \alpha\beta(\mu_{\overline{w}} - \mu_{\widehat{w}})(\mu_{\overline{w}} - \mu_{\widehat{w}})^\top \ .$$

Proof. Because $w = \alpha\overline{w} + \beta\widehat{w}$ and μ_w is linear in w, we have $\mu_w = \alpha\mu_{\overline{w}} + \beta\mu_{\widehat{w}}$. The lemma follows from the following calculations:

$$\begin{aligned}\Sigma_w = \sum_i w_i X_i X_i^\top - \mu_w\mu_w^\top &= \sum_i \alpha\overline{w}_i X_i X_i^\top - \alpha\mu_{\overline{w}}\mu_{\overline{w}}^\top + \sum_i \beta\widehat{w}_i X_i X_i^\top - \beta\mu_{\widehat{w}}\mu_{\widehat{w}}^\top \\ &\qquad + \alpha\mu_{\overline{w}}\mu_{\overline{w}}^\top + \beta\mu_{\widehat{w}}\mu_{\widehat{w}}^\top - (\alpha\mu_{\overline{w}} + \beta\mu_{\widehat{w}})(\alpha\mu_{\overline{w}} + \beta\mu_{\widehat{w}})^\top \\ &= \alpha\Sigma_{\overline{w}} + \beta\Sigma_{\widehat{w}} + \alpha\beta(\mu_{\overline{w}} - \mu_{\widehat{w}})(\mu_{\overline{w}} - \mu_{\widehat{w}})^\top \ .\end{aligned}$$

The last step uses $\alpha - \alpha^2 = \beta - \beta^2 = \alpha\beta$ as $\alpha + \beta = 1$. □

Let $w^\star$ denote the uniform distribution on G, i.e., $w_i^\star = \frac{1}{(1-\epsilon)n}$ if $i \in G$ and $w_i^\star = 0$ otherwise. By Lemma 3.6 for any w, if we move toward $w^\star$, we have

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta\Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top \ .$$

We will show that we can essentially ignore the last rank-one term using Lemma 3.7.

Lemma 3.7. *Let $G^\star$ be a (k, ϵ, δ)-stable set of samples with respect to the ground-truth distribution with $0 < \epsilon \le \delta$. Let S be an ϵ-corrupted version of $G^\star$. Then, we have*

$$\left\|(\mu_w - \mu_{w_\star})(\mu_w - \mu_{w_\star})^\top\right\|_{F,k,k} \le 4\epsilon\left(\|\Sigma_w - I\|_{F,k,k} + O(\delta^2/\epsilon)\right) \ .$$

We are now ready to prove our main result (Theorem 3.1).

Proof of Theorem 3.1. Fix any weight vector $w \in \Delta_{n,\epsilon}$. We will show that if w is a bad solution, then $f(w)$ decreases if w moves toward $w^\star$, so w cannot be a stationary point.

Let c_1 be the constant in $O(\cdot)$ in Lemma 3.7. By Lemma 3.7, if $\|\mu_w - \mu\|_{2,k} \ge c_2\delta$ for a sufficiently large constant c_2, then $\|\Sigma_w - I\|_{F,k,k} \ge (\frac{c_2^2}{4} - c_1)\frac{\delta^2}{\epsilon} = \Omega(\frac{\delta^2}{\epsilon})$.

By Lemma 3.6, $\Sigma_{(1-\eta)w+\eta w^\star} - I = (1-\eta)(\Sigma_w - I) + \eta(\Sigma_{w^\star} - I) + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top$.

Using the triangle inequality for $\|\cdot\|_{F,k,k}$, we have

$$\begin{aligned}\left\|\Sigma_{(1-\eta)w+\eta w^\star} - I\right\|_{F,k,k} &\le (1-\eta)\|\Sigma_w - I\|_{F,k,k} \\ &\qquad + \eta\|\Sigma_{w^\star} - I\|_{F,k,k} + \eta(1-\eta)\left\|(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top\right\|_{F,k,k} \ .\end{aligned}$$

We know that $\|\Sigma_{w^\star} - I\|_{F,k,k} \le \frac{\delta^2}{\epsilon}$ by the stability condition in Definition 2.1. By Lemma 3.7 and $\|\Sigma_w - I\|_{F,k,k} = \Omega(\delta^2/\epsilon)$, we can show that for all $0 < \eta < 1$,

$$\begin{aligned}&f((1-\eta)w + \eta w^\star) = \left\|\Sigma_{(1-\eta)w+\eta w^\star} - I\right\|_{F,k,k} \\ &\le (1-\eta)\|\Sigma_w - I\|_{F,k,k} + \tfrac{\eta\delta^2}{\epsilon} + 4\epsilon\eta\left(\|\Sigma_w - I\|_{F,k,k} + O(\tfrac{\delta^2}{\epsilon})\right) \\ &\le (1-\eta+4\epsilon\eta)\|\Sigma_w - I\|_{F,k,k} + (4c_1+1)\tfrac{\eta\delta^2}{\epsilon} \\ &\le (1-\tfrac{\eta}{2})\|\Sigma_w - I\|_{F,k,k} = (1-\tfrac{\eta}{2})f(w) \ .\end{aligned} \tag{4}$$

The last step uses $(\frac{1}{2} - 4\epsilon) \left\|\Sigma_w - I\right\|_2 \geq (4c_1 + 1)\frac{\delta^2}{\epsilon}$ which holds if $\epsilon \leq 1/10$ and $c_2^2 \geq 164\, c_1 + 40$.

It follows immediately that w cannot be a stationary point. Let $u = \frac{w^\star - w}{\left\|w^\star - w\right\|_2}$ and $h = \eta \left\|w^\star - w\right\|_2$. We have $w + hu = (1-\eta)w + \eta w^\star \in \Delta_{n,\epsilon}$ because $\Delta_{n,\epsilon}$ is convex. Since $\left\|w^\star - w\right\|_2 = O(\sqrt{\epsilon/n})$,

$$u^\top \nabla f(w) = \lim_{h \to 0} \frac{f(w+hu) - f(w)}{h} \leq \lim_{\eta \to 0} \frac{-(\eta/2) f(w)}{\eta \left\|w^\star - w\right\|_2} \leq -\frac{\Omega(\delta^2/\epsilon)}{\left\|w^\star - w\right\|_2} \leq -\Omega(n^{1/2}\delta^2\epsilon^{-3/2}) \; .$$

By Definition 2.3, we know w cannot be a γ-stationary point of f for some $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2})$. $\square$

4 Robust Sparse PCA

We consider a spiked covariance model for sparse PCA. In this model, there is a direction $v \in \mathbb{R}^d$ with at most k nonzero entries. The good samples are drawn from a ground-truth distribution with covariance $\Sigma = I + \rho v v^\top$, where $\rho > 0$ is a parameter that intuitively measures the strength of the signal. We consider the more interesting case when $\rho \leq 1$ (if ρ is larger the problem becomes easier).

To solve the sparse PCA problem, we consider the following optimization problem, where $M_w = \sum_i w_i X_i X_i^\top$ and $\left\|A\right\|_{F,2k^2} = \max_{|Q| = 2k^2} \left\|A_Q\right\|_F$:

$$\min_{w} \; f(w) = \left\|M_w - I\right\|_{F,2k^2} \;\text{ subject to }\; w \in \Delta_{n,\epsilon}. \tag{5}$$

The objective function minimizes the Frobenius norm of the largest $2k^2$ entries of a reweighted second-moment matrix M_w. Note that $f(w)$ is actually convex in w, because the matrix M_w is linear in w and the $\left\|\cdot\right\|_{F,2k^2}$ norm is convex.

Let R be the support of $vv^\top$. Intuitively, the k^2 entries in R could be large due to spiked covariance. By minimizing the norm of the largest $2k^2$ entries, we hope to make the entries outside of R very small. Our algorithm is given in Algorithm 2.

Algorithm 2: Robust sparse PCA.

```
Input: $k > 0$, $0 < \epsilon < \epsilon_0$, and an $\epsilon$-corrupted set of samples $(X_i)_{i=1}^n$ drawn from a distribution
with covariance $I + \rho v v^\top$ for a $k$-sparse unit vector $v$.
Output: a vector $u$ that is close to $v$.
  Find a first-order stationary point $w \in \Delta_{n,\epsilon}$ of the objective $\min_w f(w) = \left\|M_w - I\right\|_{F,2k^2}$.
  Let $A = M_w - I$. Let $Q$ be the $k^2$ entries of $A$ with largest magnitude.
  Return $u =$ the top eigenvector of $(A_Q + A_Q^\top)$.
```

Theorem 4.1. *Let $0 < \rho \leq 1$, $0 < \epsilon < \epsilon_0$, and $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (k, ϵ, δ)-stable (as in Definition 2.2) w.r.t. a centered distribution with covariance $I + \rho v v^\top$ for an unknown k-sparse unit vector $v \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$. Algorithm 2 outputs a vector u such that $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\delta/\rho})$.*

Theorem 1.3 is an immediate corollary of Theorem 4.1.

We can apply Theorem 4.1 to get an end-to-end result for subgaussian distributions. Algorithm 2 requires the stability conditions (Definition 2.2) of the original good samples $G^\star$. We show that these conditions are satisfied with a small number of samples.

Lemma 4.2. *Let $0 < \rho \leq 1$ and $0 < \epsilon < \epsilon_0$. Let D be a centered subgaussian distribution with covariance $I + \rho v v^\top$ for a k-sparse unit vector $v \in \mathbb{R}^d$. Let $G^\star$ be a set of $n = \Omega(k^2 \log d / \delta^2)$ samples drawn from D. Then then with probability at least $1 - \exp(-\Omega(k^2 \log d))$, $G^\star$ is (k, ϵ, δ)-stable (as in Definition 2.2) w.r.t. D for $\delta = O(\epsilon \log(1/\epsilon))$.*

Combining Theorem 4.1 and Lemma 4.2, given as input an ϵ-corrupted set of $n = \widetilde{\Omega}(k^2 \log d / \epsilon^2)$ samples drawn from a centered subgaussian distribution with covariance $I + \rho v v^\top$, Algorithm 2 returns a vector u with $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\epsilon \log(1/\epsilon)/\rho})$.

We defer the proofs of Lemma 4.2 and Theorem 4.1 to Appendix B and give an overview of the proof of Theorem 4.1.

Proof Sketch of Theorem 4.1. We can use the stability conditions to upper bound the optimal objective value: note that for $w^\star$ (uniform distribution on the remaining good samples), we must have $\left\|M_{w^\star} - (I + \rho v v^\top)\right\|_{F,2k^2} \le \delta$ by the stability conditions, therefore $\|M_{w^\star} - I\|_{F,2k^2} \le \left\|M_{w^\star} - (I + \rho v v^\top)\right\|_{F,2k^2} + \left\|\rho v v^\top\right\|_{F,2k^2} \le \rho + \delta$. Because the objective function $f(w)$ is convex, any stationary point w must be globally optimal and satisfies $f(w) \le \rho + \delta$.

Fix a stationary point w and let $A = M_w - I$. Let R be the support of $vv^\top$ and let Q be the set of k^2 largest entries of A. The stability conditions implies for any w, the projection in the v direction must be large (formally $v^\top A v \ge \rho - \delta$). Because the objective function measures the norm of the largest $2k^2$ entries of A and it is not much larger than the norm of the largest k^2 entries, we can argue that A_R and A_Q are close, so $v^\top A_Q v \ge \rho - O(\delta)$.

Now A_Q is a matrix with Frobenius norm at most $\rho + \delta$ while $v^\top A_Q v \ge \rho - O(\delta)$. Together these imply that the norm of A_Q in direction orthogonal to $vv^\top$ is at most $O(\sqrt{\rho\delta})$, and then by standard matrix perturbation bounds we know the top eigenvector of $(A_Q + A_Q^\top)$ is $O(\sqrt{\delta/\rho})$ close to v.

5 Experiments

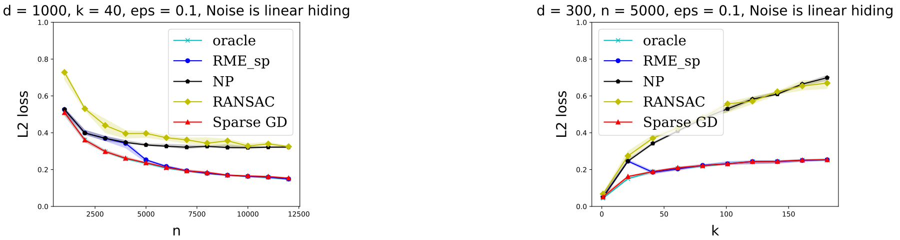
We perform an experimental evaluation of our robust sparse mean estimation algorithm on synthetic datasets with a focus on statistical accuracy (ℓ_2-distance between the output and the true sparse mean). We evaluate our algorithm (Sparse Gradient Descent, `Sparse GD`) on different noise models, and compare it to the following previous algorithms:

- `oracle`, which is told exactly which samples are inliers, and outputs their empirical mean,
- the robust sparse mean estimation algorithm `RME_sp` from [DKK+19b],
- `NP` (Naive Pruning), which removes samples far from the median and output the mean of the rest,
- `RANSAC`, which randomly selects half of the points and computes their mean. One solution is preferred to another if it has more points in a ball of radius $O(\sqrt{d})$ around it.

For algorithms that output non-sparse vectors, we take the largest k entries before measuring the ℓ_2 distance to the true mean. We evaluate the algorithms on various noise models:

- **Linear-hiding noise.** The inliers are drawn from $\mathcal{N}(0, I)$. Let S be a size k set. Then, half the outliers are drawn from $\mathcal{N}(1_S, I)$ and the other half are drawn from $\mathcal{N}(0, 2I - I_S)$.
- **Tail-flipping noise.** This noise model picks a k-sparse direction v and replaces the ϵ fraction of points farthest in the $-v$ direction with points in the $+v$ direction.
- **Constant-bias noise.** This model adds a constant to every coordinate of the outlier points. In Figure 3, we add 2 to every coordinate of every outlier point.

We ran our experiments on a computer with a 1.6 GHz Intel Core i5 processor and 8 GB RAM. We built on the codebase of [DKK+19b] [4] and implemented our new algorithm for the experiments. For each pair of algorithm and noise model, we repeat the same experiment 10 times and plot the median value of the measurements. We shade the interquartile region around the reported points in the figure as confidence intervals.

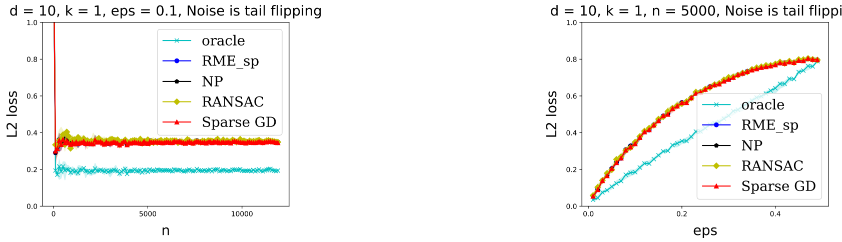


(a) Fix the sparsity k and change the number of samples n.

(b) Fix n and change the sparsity k.

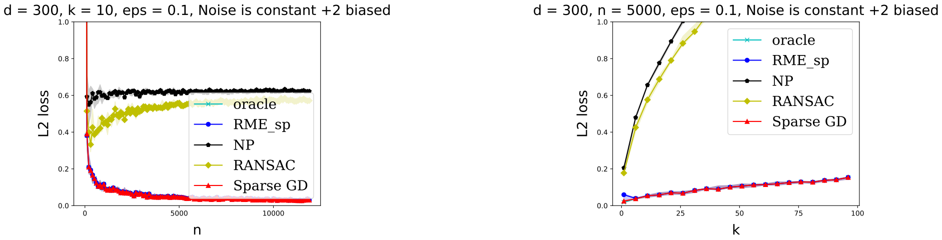
Figure 1: The performance of various algorithms under linear-hiding noise. Notably, when the number of samples n or the sparsity k is small, our algorithm `Sparse GD` outperforms `RME_sp`.

[4] Available at: https://github.com/sushrutk/robust_sparse_mean_estimation, MIT license



(a) Fix the sparsity k and change the number of samples n. (b) Fix n, k and change the fraction of corruption ϵ.

Figure 2: The performance of various algorithms in the tail-flipping noise model.



(a) Fix the sparsity k and change the number of samples n.

(b) Fix n and change the sparsity k.

Figure 3: The performance of various algorithms in the constant-bias noise model.

Our experimental results are summarized in Figures 1, 2 and 3. For the linear-hiding and constant-bias noise models, we run two experiments: 1) fix the sparsity k and change the number of samples n, and 2) fix n and change k. For the tail-flipping noise model, we run two experiments: 1) fix the sparsity k and change the number of samples n, and 2) fix k and n and change the fraction of corruption ϵ.

In terms of statistical accuracy, our algorithm (`Sparse GD`), outperforms the filter-based `RME_sp` algorithm [DKK+19b] in the linear-hiding noise model when the number of samples or the sparsity is small, as shown in Figure 1. Our algorithm matches the performance of `RME_sp` under the tail flipping and constant-bias noise models, as shown in Figures 2 and 3.

Matching our theoretical results, our `Sparse GD` algorithm has accuracy $O(\epsilon\sqrt{\log(1/\epsilon)})$ and is within a constant factor of the $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ worst-case performance of `oracle`. In contrast, the naive algorithms `NP` and `RANSAC` both incur error that scales as $\epsilon\sqrt{k}$. The tail-flipping noise (Figure 2) illustrates that $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ error can occur no matter which algorithm is used (including `oracle`), because ϵ-fraction of the original good samples was removed.

Acknowledgments and Disclosure of Funding

Yu Cheng is supported in part by NSF Award CCF-2307106. Ilias Diakonikolas is supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), NSF Award AiTF-2006206, a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. Rong Ge is supported by NSF Award CCF-1704656, NSF Award CCF-1845171 (CAREER), NSF Award CCF-1934964, a Sloan Research Fellowship, and a Google Faculty Research Award. Shivam Gupta is supported by NSF Award CCF-2008868, NSF Award AiTF-2006206, and the NSF AI Institute for Foundations of Machine Learning (IFML). Some of this work was done while Shivam Gupta was visiting the University of Wisconsin-Madison. Daniel M. Kane is supported by NSF Award CCF-1652862 (CAREER), NSF Award CCF-2107547, a Sloan Research Fellowship, and a grant from CasperLabs. Mahdi Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship, NSF Award CCF-1846369 (CAREER), NSF Award CCF-1813877, DARPA Learning with Less Labels (LwLL) and Fast Network Interface Cards (FastNICs) grants, and Faculty Research Awards from Google and Amazon.

References

[BB20] M. Brennan and G. Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 648–847. PMLR, 2020.

[BDLS17] S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proc. 30th Annual Conference on Learning Theory*, pages 169–212, 2017.

[BNJT10] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.

[BNL12] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012.

[BR13] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT 2013 - The 26th Annual Conference on Learning Theory*, pages 1046–1066, 2013.

[Bub14] S. Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[CDGS20] Y. Cheng, I. Diakonikolas, R. Ge, and M. Soltanolkotabi. High-dimensional robust mean estimation via gradient descent. In *Proc. 37th International Conference on Machine Learning (ICML)*, pages 1768–1778, 2020.

[DD18] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

[DK19] I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019.

[DKK$^+$16] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 655–664, 2016.

[DKK$^+$17] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. In *Proc. 34th International Conference on Machine Learning (ICML)*, pages 999–1008, 2017.

[DKK$^+$19a] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart. SEVER: A robust meta-algorithm for stochastic optimization. In *Proc. 36th International Conference on Machine Learning (ICML)*, pages 1596–1606, 2019.

[DKK$^+$19b] I. Diakonikolas, S. Karmalkar, D. Kane, E. Price, and A. Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *Advances in Neural Information Processing Systems 32, NeurIPS 2019*, pages 10688–10699, 2019.

[DKK$^+$21] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustness meets algorithms. *Commun. ACM*, 64(5):107–115, 2021.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *Proc. 58th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017.

[HLZ20] S. B. Hopkins, J. Li, and F. Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. In *Advances in Neural Information Processing Systems 33: NeurIPS 2020*, 2020.

[HR09] P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley New York, 2009.

[HRRS86] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics. The approach based on influence functions*. Wiley New York, 1986.

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[Hub64] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[Joh01] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[Joh17] I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. Available at http://statweb.stanford.edu/~imj/GE_08_09_17.pdf, 2017.

[LAT$^+$08] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104, 2008.

[LLC19] L. Liu, T. Li, and C. Caramanis. High dimensional robust estimation of sparse models via trimmed hard thresholding. *CoRR*, abs/1901.08237, 2019.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, 2016.

[lTB01] F. De la Torre and M. J. Black. Robust principal component analysis for computer vision. In *Proc. 8th International Conference On Computer Vision (ICCV)*, pages 362–369, 2001.

[PLJD10] P. Paschou, J. Lewis, A. Javed, and P. Drineas. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of Medical Genetics*, 47:835–847, 2010.

[RPW$^+$02] N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 298:2381–2385, 2002.

[SKL17] J. Steinhardt, P. W. Koh, and P. S. Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems 30*, pages 3520–3532, 2017.

[Tsy08] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

[Tuk60] J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

[Tuk75] J. W. Tukey. Mathematics and picturing of data. In *Proceedings of ICM*, volume 6, pages 523–531, 1975.

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[ZJS20] B. Zhu, J. Jiao, and J. Steinhardt. Robust estimation via generalized quasi-gradients. *CoRR*, abs/2005.14073, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Omitted proofs are in the supplemental material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Omitted Proofs in Section 3

In this section, we provide the proofs omitted from Section 3.

We start with the key technical lemmas we used for our structural result for robust sparse mean estimation. The sample complexity and the stability conditions for sparse mean will be proved in Appendix A.1.

We will restate the lemmas before proving them.

Lemma 3.2. *Fix two vectors x, y with $\|x\|_0 \le k$ and $\|x - y\|_{2,k} \le \delta$. Let z be a vector that keeps the k entries of y with largest absolute values and sets the rest to 0. We have $\|x - z\|_2 \le \sqrt{5}\delta$.*

Proof. Without loss of generality we assume $\|x\|_0 = k$. We partition the coordinates into three disjoint sets based on the sparsity of x and z. Let $A = \{i : x_i \neq 0 \text{ and } z_i = 0\}$, $B = \{i : x_i = 0 \text{ and } z_i \neq 0\}$, and $C = \{i : x_i \neq 0 \text{ and } z_i \neq 0\}$.

We know that $|y_i| \le |y_j|$ for all $i \in A$ and $j \in B$ because the k largest entries of $|y|$ are in $B \cup C$. Since $|A \cup C| = |B \cup C| = k$, we have $|A| = |B|$ and therefore $\|y_A\|_2 \le \|y_B\|_2$.

By the definition of z, we have $z_i = 0$ for $i \in A$ and $z_i = y_i$ for all $i \in B \cup C$. We have

$$\begin{aligned}
\|x - z\|_2^2 &= \|x_A\|_2^2 + \|y_B\|_2^2 + \|x_C - y_C\|_2^2 && (y_A = 0 \text{ and } x_B = 0)\\
&\le 2(\|x_A - y_A\|_2^2 + \|y_A\|_2^2) + \|y_B\|_2^2 + \|x_C - y_C\|_2^2 && \text{(triangle inequality)}\\
&\le 2\,\|x_A - y_A\|_2^2 + 3\,\|y_B\|_2^2 + \|x_C - y_C\|_2^2 && (\|y_A\|_2 \le \|y_B\|_2)\\
&\le 2\,\|x_{A\cup B\cup C} - y_{A\cup B\cup C}\|_2^2 + \|y_B\|_2^2 && (A, B, C \text{ are disjoint})\\
&\le 5\,\|x - y\|_{2,k}^2 \ .
\end{aligned}$$

The last inequality uses $|A \cup B \cup C| \le 2k$ and $\|x-y\|_{2,2k}^2 \le 2\,\|x-y\|_{2,k}^2$. □

Lemma 3.4. *Fix $A \in \mathbb{R}^{d\times d}$. We have $\left|v^\top A v\right| \le \|A\|_{F,k,k}$ for any k-sparse unit vector $v \in \mathbb{R}^d$.*

Proof. Without loss of generality, we assume $\|v\|_0 = k$.

Let $R \subseteq ([d] \times [d])$ be the support of $vv^\top$. We have

$$v^\top A v \le \|A_R\|_2 \le \|A_R\|_F \le \|A\|_{F,k,k}$$

The last inequality is because $\|A\|_{F,k,k}$ chooses a set of k^2 entries to maximize the ℓ_2-norm of these entries, subject to choosing these entries from k rows with k entries on each row, and R is a feasible way to choose such k^2 entries. □

Lemma 3.5. *For any $v \in \mathbb{R}^d$, $\left\|vv^\top\right\|_{F,k,k} = \|v\|_{2,k}^2$.*

Proof. Without loss of generality, we can assume v is non-negative because the norms on both sides are independent of signs. Moreover, we can assume w.l.o.g. that the coordinates of v are sorted from large to small, i.e., $v_1 \ge v_2 \ge \ldots \ge v_d \ge 0$.

The rows are multiples of each other, so the k rows with largest (ℓ_2, k)-norms are the first k rows, and the (ℓ_2, k)-norm of each row is just the ℓ_2-norm of the first k entries. Therefore, we have $\left\|vv^\top\right\|_{F,k,k}^2 = \sum_{i=1}^k \sum_{j=1}^k (v_i v_j)^2 = \left(\sum_{i=1}^k v_i^2\right)^2 = \|v\|_{2,k}^4$ as claimed. □

Lemma 3.7. *Let $G^\star$ be a (k, ϵ, δ)-stable set of samples with respect to the ground-truth distribution with $0 < \epsilon \le \delta$. Let S be an ϵ-corrupted version of $G^\star$. Then, we have*

$$\left\|(\mu_w - \mu_{w_\star})(\mu_w - \mu_{w_\star})^\top\right\|_{F,k,k} \le 4\epsilon\left(\|\Sigma_w - I\|_{F,k,k} + O(\tfrac{\delta^2}{\epsilon})\right) .$$

Proof. Recall that $S = G \cup B$ where G is the set of (remaining) good samples and B is the set of corrupted samples. Let $\alpha = \|w_G\|_1$ and $\beta = \|w_B\|_1$. Let $\overline{w} = w_G/\alpha$ and $\widehat{w} = w_B/\beta$ denote the normalized version of w_G and w_B.

We can write $w = \alpha\overline{w} + \beta\widehat{w}$, by Lemma 3.6, we know that

$$\Sigma_w = \alpha\Sigma_{\overline{w}} + \beta\Sigma_{\widehat{w}} + \alpha\beta(\mu_{\overline{w}} - \mu_{\widehat{w}})(\mu_{\overline{w}} - \mu_{\widehat{w}})^\top . \tag{6}$$

Since $\beta \le \|w\|_\infty |B| \le \frac{\epsilon}{1-\epsilon}$, we have $\|\overline{w}\|_\infty = \frac{\|w_G\|_\infty}{\alpha} \le \frac{1}{(1-2\epsilon)n}$. Because $G^\star$ is (k, ϵ, δ)-stable and $\overline{w} \in \Delta_{n,2\epsilon}$ can be viewed as weights on $G^\star$, by the stability condition in Definition 2.1,

$$\|\Sigma_{\overline{w}} - I\|_{F,k,k} \le \frac{\delta^2}{\epsilon} . \tag{7}$$

Using Lemma 3.4, Equations (6) and (7), and that $\Sigma_{\widehat{w}} \succeq 0$, for any k-sparse unit vector $v \in \mathbb{R}^d$,

$$\begin{aligned}\|\Sigma_w - I\|_{F,k,k} \ge v^\top(\Sigma_w - I)v &= \alpha v^\top \Sigma_{\overline{w}} v + \beta v^\top \Sigma_{\widehat{w}} v + \alpha\beta\left((\mu_{\overline{w}} - \mu_{\widehat{w}})^\top v\right)^2 - 1 \\ &\ge \alpha\left(1 + v^\top(\Sigma_{\overline{w}} - I)v\right) + \alpha\beta\left((\mu_{\overline{w}} - \mu_{\widehat{w}})^\top v\right)^2 - 1 \\ &\ge \alpha\left(1 - \tfrac{\delta^2}{\epsilon}\right) - 1 + \alpha\beta\left((\mu_{\overline{w}} - \mu_{\widehat{w}})^\top v\right)^2 .\end{aligned} \tag{8}$$

We know $\alpha(1 - \delta^2/\epsilon)$ is close to 1, so we are essentially left with only the last term on the right-hand side. We will relate this term to $\|\mu_w - \mu_{w_\star}\|_{2,k}^2$, which is what appears in the lemma statement.

Recall that $\alpha + \beta = 1$ and $w = \alpha\overline{w} + \beta\widehat{w}$, and thus

$$\beta(\mu_{\widehat{w}} - \mu_{\overline{w}}) = \beta\mu_{\widehat{w}} + \alpha\mu_{\overline{w}} - \mu_{\overline{w}} = \mu_w - \mu_{\overline{w}} = (\mu_w - \mu_{w^\star}) + (\mu_{w^\star} - \mu_{\overline{w}}) . \tag{9}$$

Since $\overline{w}, w^\star \in \Delta_{n,2\epsilon}$ put weights only on G, it follows from the stability conditions (Definition 2.1)

$$\left|(\mu_{w^\star} - \mu_{\overline{w}})^\top v\right| \le \|\mu_{w^\star} - \mu_{\overline{w}}\|_2 \le \|\mu_{w^\star} - \mu\|_2 + \|\mu - \mu_{\overline{w}}\|_2 \le 2\delta \ . \tag{10}$$

Let $u \in \mathbb{R}^d$ be a vector that keeps the k entries of $(\mu_w - \mu_{w^\star})$ with largest magnitude. We choose $v = \frac{u}{\|u\|_2}$. Notice that $(\mu_w - \mu_{w^\star})^\top v = \|\mu_w - \mu_{w^\star}\|_{2,k}$. From Equations (9) and (10), we have

$$\begin{aligned}\left(\beta \cdot (\mu_{\widehat{w}} - \mu_{\overline{w}})^\top v\right)^2 &= \left((\mu_w - \mu_{w^\star})^\top v + (\mu_{w^\star} - \mu_{\overline{w}})^\top v\right)^2 \\ &\ge \frac{\left((\mu_w - \mu_{w_\star})^\top v\right)^2}{2} - \left((\mu_{\overline{w}} - \mu_{w^\star})^\top v\right)^2 \ge \frac{\|\mu_w - \mu_{w_\star}\|_{2,k}^2}{2} - 4\delta^2 \ .\end{aligned} \tag{11}$$

The first inequality in Equation (11) uses the fact that $(x+y)^2 \ge \frac{x^2}{2} - y^2$ for any $x, y \in \mathbb{R}$.

Putting Equations (8) and (11) together for our choice of v, we have

$$\begin{aligned}\|\Sigma_w - I\|_{F,k,k} &\ge \alpha\left(1 - \tfrac{\delta^2}{\epsilon}\right) - 1 + \frac{\alpha}{\beta}\left(\beta \cdot (\mu_{\overline{w}} - \mu_{\widehat{w}})^\top v\right)^2 \\ &\ge \frac{1-2\epsilon}{1-\epsilon}\left(1 - \tfrac{\delta^2}{\epsilon}\right) - 1 + \frac{1-2\epsilon}{\epsilon}\left(\frac{\|\mu_w - \mu_{w_\star}\|_{2,k}^2}{2} - 4\delta^2\right) \\ &\ge \frac{1}{4\epsilon}\|\mu_w - \mu_{w_\star}\|_{2,k}^2 - O\left(\tfrac{\delta^2}{\epsilon}\right) \ .\end{aligned}$$

The lemma follows since $\|\mu_w - \mu_{w_\star}\|_{2,k}^2 = \left\|(\mu_w - \mu_{w_\star})(\mu_w - \mu_{w_\star})^\top\right\|_{F,k,k}$ by Lemma 3.5. □

A.1 Stability Conditions for Robust Sparse Mean

In this section, we prove Lemma 3.3, which states that stability conditions (Definition 2.1) needed for our robust sparse mean algorithm is satisfied with a small number of samples (Lemma 3.3).

A version of the $\|\cdot\|_{2,k}$ part of of Lemma 3.3 was known in prior works (e.g., [BDLS17]). In this paper, we define the stability conditions with weights and we include the proof for the $\|\cdot\|_{2,k}$ part to be self-contained.

Lemma 3.3. *Fix $k > 0$ and $0 < \epsilon < \epsilon_0$. Let $G^\star$ be a set of n samples that is generated according to a subgaussian distribution with mean μ and covariance I, if $n = \Omega(k^2 \log d/\delta^2)$, then with probability at least $1 - \exp(-\Omega(k^2 \log d))$, $G^\star$ is (k, ϵ, δ)-stable (as in Definition 2.1) for $\delta = O(\epsilon \log(1/\epsilon))$.*

Proof. Recall that the stability conditions in Definition 2.1 state that for all $w \in \Delta_{n,2\epsilon}$,

$$\|\mu_w - \mu\|_{2,k} \le \delta \quad \text{and} \quad \|\Sigma_w - I\|_{F,k,k} \le \delta^2/\epsilon,$$

Without loss of generality, we assume that the samples $G^\star = (X_i)_{i=1}^n$ are drawn from a subgaussian distribution with mean $\mu = 0$ and identity covariance matrix. This is because shifting the samples by μ does not change the lemma statement.

For ease of presentation, we will upper bound the norms with $O(\delta)$ and $O(\delta^2/\epsilon)$ and prove the lemma for $\Delta_{n,\epsilon}$ instead of $\Delta_{n,2\epsilon}$. This is sufficient because the constants in $O(\cdot)$ and the constants due to changing ϵ to 2ϵ can be put into $\delta = O(\epsilon \log(1/\epsilon))$.

(i) First we prove $\|\mu_w\|_{2,k} \le O(\delta)$ with probability at least $1 - \exp(k \log d - \Omega(n\delta^2))$.

Due to convexity of $\|\cdot\|_{2,k}$, it is sufficient to upper bound $\|\mu_w\|_{2,k}$ for all vertices of $\Delta_{n,\epsilon}$. In other words, we need to show that

$$\left\|\frac{1}{(1-\epsilon)n}\sum_{i \in G^\star \setminus L} X_i\right\|_{2,k} = O(\delta) \ \text{ for every } L \subseteq G^\star \text{ with } |L| = \epsilon n.$$

Ignoring the constants and by the triangle inequality, it suffices to show

$$\left\|\frac{1}{n}\sum_{i\in G^\star} X_i\right\|_{2,k} = O(\delta) \quad \text{and} \quad \left\|\frac{1}{n}\sum_{i\in L} X_i\right\|_{2,k} = O(\delta) \text{ for all } |L| = \epsilon n \ . \tag{12}$$

By the definition of $\|\cdot\|_{2,k}$, we need to show that for all k-sparse unit vector v,

$$v^\top\left(\frac{1}{n}\sum_{i\in G^\star} X_i\right) = O(\delta) \quad \text{and} \quad v^\top\left(\frac{1}{n}\sum_{i\in L} X_i\right) = O(\delta) \text{ for all } |L| = \epsilon n \ . \tag{13}$$

We first prove Equation (13) for a fixed v and then a take union bound over a net of k-sparse vectors to prove Equation (12).

Fix a unit vector $v \in \mathbb{R}^d$.

(i.a) For $G^\star$, by the definition of subgaussian distributions and the Chernoff bound,

$$\Pr\left[\left(\frac{1}{n}\sum_{i=1}^{n} v^\top X_i\right) > \delta\right] \le \exp(-\Omega(n\delta^2)) \ .$$

(i.b) For L, in the worst case, L contains the samples with the largest $(v^\top X_i)$. We will show that very few $(v^\top X_i)$ can be large. Let

$$h_r(z) = \begin{cases} 0 & z \le r \ , \\ z & z > r \ . \end{cases}$$

We have that, for every L with $|L| = \epsilon n$,

$$\frac{1}{n}\sum_{i\in L} v^\top X_i \le \frac{1}{n}\sum_{i\in L} r + \frac{1}{n}\sum_{i\in L} h_r(v^\top X_i) \le \epsilon r + \frac{1}{n}\sum_{i=1}^{n} h_r(v^\top X_i) \ .$$

We set $r = 2\sqrt{\ln(1/\epsilon)} > 1$. The first term is $\epsilon r = O(\epsilon\sqrt{\log(1/\epsilon)}) = O(\delta)$, so we can focus on the second term. Note that $h_r(v^\top X)$ is not bounded, but we can bound it using Chernoff-bound like arguments.

$$\begin{aligned} \mathbb{E}_X\left[\exp\left(h_r(v^\top X)/4\right)\right] &= \Pr[v^\top X \le r] + \int_r^\infty \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)\exp(x/4)\ dx \\ &\le 1 + \int_r^\infty \frac{1}{\sqrt{2\pi}}\exp(-x^2/4)\ dx \\ &\le 1 + \exp(-r^2/4) \le \exp(\epsilon) \ . \end{aligned}$$

Because the X_i's are independent, we have

$$\mathbb{E}\left[\exp\left(\frac{1}{4}\sum_{i=1}^{n} h_r(v^\top X_i)\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n}\exp\left(h_r(v^\top X_i)\right)\right] \le \exp(\epsilon n) \ .$$

By Markov's inequality, $\frac{1}{4}\sum_{i=1}^{n} h_r(v^\top X_i) > 2\delta n$ with probability at most $\exp((\epsilon - 2\delta)n) \le \exp(-n\delta)$.

Therefore, Equation (13) hold for a fixed $v \in \mathbb{R}^d$ with probability at least $1 - \exp(-\Omega(n\delta^2))$.

We conclude Part *(i)* via a union bound over a net of k-sparse vectors. Fix a sparsity pattern $R \subseteq [d]$ with $|R| = k$. There exists a net $\mathcal{C}_R$ of $2^{O(k)}$ unit vectors such that for any $y \in \mathbb{R}^d$, there exists a vector $v \in \mathcal{C}_R$ such that $v^\top y \ge (1/2)\|y\|_2$. Consequently, for any $y \in \mathbb{R}^d$,

$$\|y\|_{2,k} = \max_{|R|=k}\|y_R\|_2 \le 2\max_R \max_{v\in\mathcal{C}_R} v^\top y \ .$$

Taking a union bound over $\binom{d}{k}$ sparsity patterns R and every $v \in \mathcal{C}_R$, we know that Equation (12) holds with probability at least $1 - \exp(O(k\log d) - \Omega(n\delta^2))$, which then immediately implies $\|\mu_w\|_{2,k} \le O(\delta)$.

(ii) Next we prove $\|\Sigma_w - I\|_{F,k,k} \le O(\delta^2/\epsilon)$. The proof structure is similar to Part *(i)*. The main difference is that we will need concentration inequality and tail bounds for the second moment, and for $\|\cdot\|_{F,k,k}$, we will consider a fixed matrix U with $\|U\|_F = 1$ (rather than a unit vector) and then union bound over all k^2-sparse matrices.

We first argue that it is sufficient to prove

$$\|M_w - I\|_{F,k,k} = O(\delta^2/\epsilon) \text{ where } M_w = \sum_w w_i X_i X_i^\top \ .$$

This is because

$$\begin{aligned}\|\Sigma_w - I\|_{F,k,k} &= \left\|M_w - \mu_w\mu_w^\top - I\right\|_{F,k,k} \\ &\le \|M_w - I\|_{F,k,k} + \left\|\mu_w\mu_w^\top\right\|_{F,k,k} \\ &= \|M_w - I\|_{F,k,k} + \|\mu_w\|_{2,k}^2 && \text{(Lemma 3.5)} \\ &\le \|M_w - I\|_{F,k,k} + O(\delta^2) && (\|\mu_w\|_{2,k} \le O(\delta) \text{ from Part } \textit{(i)})\end{aligned}$$

Moreover, since $\|\cdot\|_{F,k,k}$ is convex and $(M_w - I)$ is linear in w, it is sufficient to consider all w that is a vertex of $\Delta_{n,\epsilon}$. In other words, we need to show for every $|L| = \epsilon n$,

$$\left\|\frac{1}{(1-\epsilon)n}\sum_{i\in G^\star\setminus L} X_i X_i^\top - I\right\|_{F,k,k} = O(\delta^2/\epsilon) \ .$$

Notice that

$$\frac{1}{(1-\epsilon)n}\sum_{i\in G^\star\setminus L} X_i X_i^\top - I = \frac{1}{1-\epsilon}\left(\frac{1}{n}\sum_{i\in G^\star}\left(X_i X_i^\top - I\right) - \frac{1}{n}\sum_{i\in L}\left(X_i X_i^\top - I\right)\right) \ .$$

Ignoring the constants and by the triangle inequality, it suffices to show

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i\in G^\star}\left(X_i X_i^\top - I\right)\right\|_{F,k,k} &= O(\delta^2/\epsilon) \text{ and} \\ \left\|\frac{1}{n}\sum_{i\in L}\left(X_i X_i^\top - I\right)\right\|_{F,k,k} &= O(\delta^2/\epsilon) \text{ for all } |L| = \epsilon n \ .\end{aligned} \tag{14}$$

Because $\|A\|_{F,k,k} \le \|A\|_{F,k^2} = \max_{\|U\|_0 \le k^2, \|U\|_F = 1}(U \bullet A)$, it is sufficient to show that for all k^2-sparse matrix $U \in \mathbb{R}^{d\times d}$ with $\|U\|_F = 1$,

$$\begin{aligned}U \bullet \left(\frac{1}{n}\sum_{i\in G^\star}\left(X_i X_i^\top - I\right)\right) &= O(\delta^2/\epsilon) \text{ and} \\ U \bullet \left(\frac{1}{n}\sum_{i\in L}\left(X_i X_i^\top - I\right)\right) &= O(\delta^2/\epsilon) \text{ for all } |L| = \epsilon n \ .\end{aligned} \tag{15}$$

We first prove Equation (15) for a fixed U and then a take union bound over a net of k^2-sparse matrices to prove Equation (14).

Fix a matrix $U \in \mathbb{R}^{d\times d}$ with $\|U\|_F = 1$.

(ii.a) Recall that $G^\star = (X_i)_{i=1}^n$ are drawn independently from a centered subgaussian distribution with covariance $\Sigma = I$. Note that $\mathbb{E}_X\left[U \bullet (XX^\top - I)\right] = 0$. By the Hanson-Wright inequality, for any $\|U\|_F = 1$ and $0 < t \le 1$, we have

$$\Pr\left[U \bullet \left(\frac{1}{n}\sum_{i\in G^\star}\left(X_i X_i^\top - I\right)\right) > t\right] \le \exp(-\Omega(nt^2)) \ .$$

Therefore, $U \bullet \left(\frac{1}{n}\sum_{i\in G^\star}\left(X_i X_i^\top - I\right)\right) = O(\delta^2/\epsilon)$ holds with probability at least $1 - \exp(-\Omega(n\delta^4/\epsilon^2)) \ge 1 - \exp(-\Omega(n\delta^2))$.

(ii.b) For L, we will show that very few $U \bullet (X_i X_i^\top - I)$ can be large. Recall that

$$h_r(z) = \begin{cases} 0 & z \le r \ , \\ z & z > r \ . \end{cases}$$

For every L with $|L| = \epsilon n$, we have

$$\begin{aligned} \frac{1}{n}\sum_{i \in L} U \bullet (X_i X_i^\top - I) &\le \frac{1}{n}\sum_{i \in L} r + \frac{1}{n}\sum_{i \in L} h_r(U \bullet (X_i X_i^\top - I)) \\ &\le \epsilon r + \frac{1}{n}\sum_{i=1}^{n} h_r(U \bullet (X_i X_i^\top - I)) \ . \end{aligned} \tag{16}$$

We set $r = \delta^2/\epsilon^2$.
The first term is $\epsilon r = O(\delta^2/\epsilon)$, so we focus on the second term. For the second term, let c be a sufficiently small constant and consider $\mathbb{E}\big[\exp(c \cdot \sum_{i=1}^n h_r(U \bullet (X_i X_i^\top - I))\big]$.
Notice that by hypercontractivity, $U \bullet (XX^\top - I)$ has exponential tails (see, e.g., [Ver18]). If δ is a sufficiently large multiple of $\epsilon\sqrt{\ln(1/\epsilon)}$, then r will be a sufficiently large multiple of $\ln(1/\epsilon)$, and it will be the case that $h_r(U \bullet (XX^\top - I)) = 0$ except with probability at most ϵ^3. Observe that $\exp(c \cdot h_r(U \bullet (XX^\top - I))) > z$ iff $(U \bullet (XX^\top - I)) > \max(r, \ln(z)/c)$, which by the exponential tails (when c is small enough) happens with probability at most $\min(\epsilon^3, 1/z^3)$.
We are now ready to bound the second term in Equation (16).

$$\begin{aligned} &\mathbb{E}\big[\exp(c \cdot h_r(U \bullet (XX^\top - I)))\big] \\ &= 1 + \int_1^\infty \Pr\big[\exp\big(c \cdot h_r(U \bullet (XX^\top - I))\big) > z\big]\, dz \\ &\le 1 + \int_r^{1/\epsilon} \epsilon^3 \, dz + \int_{1/\epsilon}^\infty 1/z^3 \, dz \\ &\le 1 + O(\epsilon^2) \le \exp(O(\epsilon^2)) \ . \end{aligned}$$

Because the X_i's are independent, we have

$$\mathbb{E}\left[\exp\left(c \cdot \sum_{i=1}^n h_r(U \bullet (X_i X_i^\top - I))\right)\right] = \exp(O(n\epsilon^2)) \ .$$

By Markov's inequality, $c \cdot \sum_{i=1}^n h_r(U \bullet (XX^\top - I)) > 2(\delta^2/\epsilon)n$ with probability at most $\exp(n(O(\epsilon^2) - 2\delta^2/\epsilon)) \le \exp(-n\delta^2/\epsilon) \le \exp(-\Omega(n\delta^2))$.

Therefore, Equation (15) hold for a fixed $U \in \mathbb{R}^{d \times d}$ with probability at least $1 - \exp(-\Omega(n\delta^2))$.

There is a set $\mathcal{C}$ of k^2-sparse matrices with unit Frobenius norm with size $|\mathcal{C}| = d^{O(k^2)}$, such that for any $Y \in \mathbb{R}^{d \times d}$, there exists a matrix $U \in \mathcal{C}$ such that $U \bullet Y \ge (1/2)\|Y\|_{F,k^2}$. Taking a union bound over $\mathcal{C}$, we know that Equation (14) holds with probability at least $1 - \exp(O(k^2 \log d) - \Omega(n\delta^2))$, which then immediately implies $\|\Sigma_w - I\|_{F,k,k} \le O(\delta)$.

Taking a union bound over Part *(i)* and *(ii)*, we have that $G^\star$ is (k, ϵ, δ)-stable with probability at least $1 - \exp(O(k^2 \log d) - \Omega(n\delta^2))$. Therefore, when $n = \Omega(k^2 \log d/\delta^2)$, $G^\star$ is (k, ϵ, δ)-stable with probability at least $1 - \exp(-\Omega(k^2 \log d))$. □

B Omitted Proofs in Section 4

In this section, we provide our main structural result for robust sparse PCA, which states that Algorithm 2 works as long as the original good samples satisfy the stability conditions in Definition 2.2. The sample complexity and the stability conditions for sparse PCA will be proved in Appendix B.1.

Theorem 4.1. *Let $0 < \rho \le 1$, $0 < \epsilon < \epsilon_0$, and $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (k, ϵ, δ)-stable (as in Definition 2.2) w.r.t. a centered distribution with covariance $I + \rho vv^\top$ for an* *unknown k-sparse unit vector $v \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$. Then the output u of Algorithm 2 satisfies that $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\delta/\rho})$.*

Recall that our objective function for sparse PCA is

$$\min_w \; f(w) = \|M_w - I\|_{F,2k^2} \quad \text{subject to } \; w \in \Delta_{n,\epsilon} \; ,$$

where $M_w = \sum_i w_i X_i X_i^\top$ and $\|A\|_{F,2k^2} = \max_{Q \subseteq ([d]\times[d]), |Q|=2k^2} \|A_Q\|_F$.

We will in fact prove a stronger statement that any approximately optimal solution w suffices for robust sparse PCA. Formally, we show that the minimum objective value $\min_w f(w) \le \rho + \delta$, and given any $w \in \Delta_{n,\epsilon}$ with $f(w) \le \rho + O(\delta)$, Algorithm 2 can achieve the guarantee stated in Theorem 4.1.

Throughout this section, we fix an approximately optimal solution $w \in \Delta_{n,\epsilon}$ with $f(w) \le \rho + O(\delta)$. Let $A = M_w - I$. Recall that R is the support of $vv^\top$ and Q is the largest k^2 entries of A.

Proof of Theorem 4.1. We assume without loss of generality that $\rho = \Omega(\delta)$ for some sufficiently large constant. Otherwise, the theorem holds vacuously because $\left\|uu^\top - vv^\top\right\|_F \le 2 \le O(\sqrt{\delta/\rho})$.

Let $w^\star$ be the uniform distribution over the remaining good samples G. By the stability conditions in Definition 2.2, we can upper bound the objective value at $w^\star$.

$$f(w^\star) = \|M_{w^\star} - I\|_{F,2k^2} \le \left\|M_{w^\star} - (I + \rho vv^\top)\right\|_{F,2k^2} + \left\|\rho vv^\top\right\|_{F,2k^2} \le \delta + \rho \; .$$

We will show that given such any $w \in \Delta_{n,\epsilon}$ with $f(w) \le \rho + O(\delta)$, the output u of Algorithm 2 satisfies that $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\delta/\rho})$.

Recall that u is the top eigenvector of $\overline{A_Q} = (A_Q + A_Q^\top)/2$. At a high level, we will show that $\overline{A_Q}$ is close to $\rho vv^\top$ and then use matrix perturbation theorem to show their top eigenvectors are close.

We will show in Lemma B.1 that $v^\top A_Q v \ge \rho - O(\delta)$. Consequently, we can write $A_Q = \lambda vv^\top + B$, where $v^\top B v = 0$ and $\lambda \ge \rho - O(\delta)$. Because $v^\top B v = 0$ and $\|A_Q\|_F = f(w) \le \rho + O(\delta)$,

$$\|B\|_F^2 = \|A_Q\|_F^2 - \lambda^2 \le (\rho + O(\delta))^2 - (\rho - O(\delta))^2 = O(\rho\delta + \delta^2) = O(\rho\delta) \; .$$

Let $\overline{B} = (B + B^\top)/2$. We have $v^\top \overline{B} v = 0$, $\left\|\overline{B}\right\|_F \le \|B\|_F = O(\sqrt{\rho\delta})$, and

$$\overline{A_Q} = \lambda vv^\top + \overline{B} \; .$$

Notice that u is the top eigenvector of $\overline{A_Q}$ and v is the top eigenvector of $\rho vv^\top$. By the matrix perturbation theorem (e.g., Davis-Kahan), we have

$$\left\|uu^\top - vv^\top\right\|_2 = O\left(\frac{\|\overline{B}\|_2}{\lambda - \lambda_2}\right) \le O\left(\frac{\sqrt{\rho\delta}}{\rho}\right) = O(\sqrt{\delta/\rho}) \; ,$$

where λ_2 is the second largest eigenvalue of $\overline{A_Q}$. The eigengap $\lambda - \lambda_2 = \Omega(\rho)$, because the top eigenvalue of $\overline{A_Q}$, which is at least λ, is close to its Frobenius norm. More specifically, we can have say $\lambda \ge \rho - O(\delta) \ge \frac{8}{9}\rho$, and $\lambda_2 \le \frac{2}{3}\rho$ due to $\lambda^2 + \lambda_2^2 \le \left\|\overline{A_Q}\right\|_F^2 \le \|A_Q\|_F^2 \le (\rho + O(\delta))^2 \le (\frac{10}{9}\rho)^2$.

We conclude the proof by noticing that

$$\left\|uu^\top - vv^\top\right\|_F \le \sqrt{2}\left\|uu^\top - vv^\top\right\|_2 = O(\sqrt{\delta/\rho}) \; . \qquad \square$$

We used the following lemma in the proof of Theorem 4.1, which intuitively states that A_Q is close to $\rho vv^\top$ when measured by $vv^\top$.

Lemma B.1. *Consider the same setting as in Theorem 4.1. We have $v^\top A_Q v \ge \rho - O(\delta)$.*

Proof. Let $A = \lambda vv^\top + B$ with $v^\top B v = 0$. Recall that $w_G = \sum_{i \in G} w_i \ge \frac{1-2\epsilon}{1-\epsilon}$.

By the stability conditions in Definition 2.2, we have

$$\begin{aligned}
\lambda = v^\top A v = v^\top \left(\sum_{i=1}^{n} w_i X_i X_i^\top - I \right) v \\
&\geq v^\top \left(\sum_{i \in G} w_i X_i X_i^\top - I \right) v \\
&= v^\top \left(\sum_{i \in G} w_i \left(X_i X_i^\top - I - \rho v v^\top \right) \right) v - (1 - w_G) + w_G \rho \\
&\geq w_G \rho - \left\| \sum_{i \in G} w_i (X_i X_i^\top - I - \rho v v^\top) \right\|_{F,k^2} - (1 - w_G) \\
&\geq \frac{1-2\epsilon}{1-\epsilon} \rho - \frac{1-\epsilon}{1-2\epsilon} O(\delta) - \frac{\epsilon}{1-\epsilon} \geq \rho - O(\delta) \ .
\end{aligned}$$

The last inequality uses that $\epsilon < \delta$ and $\rho \leq 1$. Note that this also implies

$$\|A_R\|_F \geq v^\top A_R v = v^\top A_R v \geq \rho - O(\delta) \ .$$

At a high level, we want to show that $v^\top A_Q v$ is close to $v^\top A_R v$. Because

$$v^\top A_Q v = v^\top A_R v - v^\top A_{R \setminus Q} v \ ,$$

we will focus on the quadratic form $v^\top A_{R \setminus Q} v = \mathrm{vec}(A_{R \setminus Q})^\top \mathrm{vec}(v v^\top)$, where $\mathrm{vec}(\cdot)$ is the vectorization of a matrix. We will upper bound this term by $\left\| \mathrm{vec}(A_{R \setminus Q}) \right\|_\infty \left\| \mathrm{vec}((v v^\top)_{R \setminus Q}) \right\|_1$.

(i) We first prove that every entry in $A_{R \setminus Q}$ has magnitude at most $O(\sqrt{\rho\delta}/k)$.

In particular, we will show that the smallest entry in Q has magnitude $O(\sqrt{\rho\delta}/k)$.

Let $\overline{R} = ([d] \times [d]) \setminus R$. We have $\|B_{\overline{R}}\|_{F,k^2} = O(\sqrt{\rho\delta})$, otherwise $f(w) = \|A\|_{F,2k^2}^2 \geq \|A_R\|_F^2 + \|B_{\overline{R}}\|_{F,k^2}^2$ would be larger than $\rho + O(\delta)$.

Observe that the $2k^2$-th largest entry of A (i.e., the smallest entry of A in Q) is upper bounded by the k^2-th largest entry of $A_{\overline{R}} = B_{\overline{R}}$, so its magnitude is at most $\frac{\|B_{\overline{R}}\|_{F,k^2}}{k} = O(\sqrt{\rho\delta}/k)$.

(ii) Next we show that the average magnitude of $(v v^\top)_{R \setminus Q}$ is small.

Let $t = |R \setminus Q| > 0$. (If $t = 0$, then $Q = R$ and the lemma follows from previous calculations.) Let

$$r = \frac{\sum_{(i,j) \in R \setminus Q} |v_i v_j|}{t}$$

be the average magnitude of entries in $(v v^\top)_{R \setminus Q}$. We will show that $r = O(\sqrt{\delta/(\rho t)})$.

Notice that $\left\| (\lambda v v^\top)_{R \setminus Q} \right\|_F = \Omega(\rho r \sqrt{t})$ (because $\lambda = \Omega(\rho)$ and the Frobenius norm is minimized when all entries are equal) and $\left\| B_{R \setminus Q} \right\|_F \leq \|B_R\|_F = O(\sqrt{\rho\delta})$ (otherwise $f(w) \geq \|A_R\|_F^2 = \lambda^2 + \|B_R\|_F^2$ would be larger than $\rho + O(\delta)$).

By the triangle inequality,

$$\left\| A_{R \setminus Q} \right\|_F = \left\| (\lambda v v^\top + B)_{R \setminus Q} \right\|_F \geq \Omega(\rho r \sqrt{t}) - O(\sqrt{\rho\delta}) \ .$$

On the other hand, by Part *(i)*, we know that every entry of $A_{R \setminus Q}$ is small,

$$\left\| A_{R \setminus Q} \right\|_F \leq O(\sqrt{\rho\delta}/k \cdot \sqrt{t})$$

Putting the above two inequalities together and solving for r, we get

$$r \leq \frac{\sqrt{\delta/\rho}}{k} + \frac{\sqrt{\delta/\rho}}{\sqrt{t}} = O\left(\sqrt{\frac{\delta/\rho}{t}} \right) \ .$$

The last step uses $t \leq |R| = k^2$.

Finally, we can upper bound $v^\top A_{R\setminus Q} v$ by

$$\begin{aligned} v^\top A_{R\setminus Q} v &= \mathrm{vec}(A_{R\setminus Q})^\top \mathrm{vec}((vv^\top)_{R\setminus Q}) \\ &\le \left\|\mathrm{vec}(A_{R\setminus Q})\right\|_\infty \left\|\mathrm{vec}((vv^\top)_{R\setminus Q})\right\|_1 \\ &\le O\left(\frac{\sqrt{\rho\delta}}{k}\right) \cdot (rt) \le O\left(\frac{\sqrt{\rho\delta}}{k} \cdot \sqrt{\frac{\delta/\rho}{t}} \cdot t\right) = O(\delta)\ . \end{aligned}$$

The lemma follows immediately because

$$v^\top A_Q v = v^\top A_R v - v^\top A_{R\setminus Q} v \ge \rho - O(\delta) - O(\delta) = \rho - O(\delta)\ . \qquad \square$$

B.1 Stability Conditions for Robust Sparse PCA

Lemma 4.2. *Let $0 < \rho \le 1$ and $0 < \epsilon < \epsilon_0$. Let D be a centered subgaussian distribution with covariance $I + \rho v v^\top$ for a k-sparse unit vector $v \in \mathbb{R}^d$. Let $G^\star$ be a set of $n = \Omega(k^2 \log d / \delta^2)$ samples drawn from D. Then then with probability at least $1 - \exp(-\Omega(k^2 \log d))$, $G^\star$ is (k, ϵ, δ)-stable (as in Definition 2.2) w.r.t. D for $\delta = O(\epsilon \log(1/\epsilon))$.*

The proof of Lemma 4.2 is almost identical to Part *(ii)* of Lemma 3.3. We give a proof sketch highlighting the differences.

Notice that the PCA stability conditions are only on the second moment, and the δ in the PCA stability conditions plays the role of the "δ^2/ϵ" in the second-moment stability conditions for sparse mean.

Similar to the proof Lemma 4.2, it is sufficient to upper bound the norm with $O(\delta)$ and for all vertices of $\Delta_{n,\epsilon}$. Or equivalently,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i \in G^\star} \left(X_i X_i^\top - I - \rho v v^\top \right) \right\|_{F, 2k^2} &= O(\delta) \ \text{ and} \\ \left\| \frac{1}{n} \sum_{i \in L} \left(X_i X_i^\top - I - \rho v v^\top \right) \right\|_{F, 2k^2} &= O(\delta) \ \text{ for all } \ |L| = \epsilon n\ . \end{aligned}$$

Fix some $U \in \mathbb{R}^{d \times d}$ with $\|U\|_F = 1$.

Notice that since $0 < \rho \le 1$, the Hanson-Wright inequality continues to hold when the covariance matrix is $\Sigma = I + \rho v v^\top \preceq 2I$.

$$\Pr\left[U \bullet \left(\frac{1}{n} \sum_{i \in G^\star} \left(X_i X_i^\top - (I + \rho v v^\top) \right) \right) > \delta \right] \le \exp(-\Omega(n\delta^2))\ .$$

By hypercontractivity that $U \bullet (XX^\top - (I + \rho v v^\top))$ has exponential tails. Consequently, we can show that with probability at least $1 - \exp(-\Omega(n\delta))$,

$$U \bullet \left(\frac{1}{n} \sum_{i \in L} (X_i X_i^\top - (I + \rho v v^\top)) \right) \le O(\delta)$$

for all $|L| = \epsilon n$. Therefore, the desired conditions hold for a fixed U with probability at least $1 - \exp(-\Omega(n\delta^2))$.

We can then take a union bound over a net $|\mathcal{C}|$ of $2k^2$-sparse matrices U of size $|\mathcal{C}| = d^{O(k^2)}$ to show that, when $n = \Omega(k^2 \log d / \delta^2)$,

$$\left\| \sum_{i \in G^\star \setminus L} \frac{1}{(1-\epsilon)n} X_i X_i^\top - (I + \rho v v^\top) \right\|_{F, 2k^2} \le O(\delta) \quad \text{for all } |L| = \epsilon n\ ,$$

with probability at least $1 - \exp(\Omega(-k^2 \log d))$.

C Algorithmic Results: Finding Stationary Points

In this section, we present our algorithmic results for robust sparse mean estimation and robust sparse PCA. We show that one can find an approximate stationary point that suffices for the underlying robust estimation problem in a polynomial number of iterations.

When the true distribution is subgaussian, we prove that projected gradient descent can compute a good stationary point in $\widetilde{O}(d^4/\epsilon^2)$ iterations for robust sparse mean estimation, and in $\widetilde{O}(nd^2/\epsilon)$ iterations for robust sparse PCA.

We note that our iteration complexity is fairly loose and we did not make an effort to optimize the polynomial dependence. [5]

Theorem C.1. *Fix $k > 0$ and $0 < \epsilon < \epsilon_0$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted set of $n = \Omega(k^2 \log d/\epsilon^2)$ samples drawn from a subgaussian distribution with unknown mean $\mu \in \mathbb{R}^d$ and covariance I. Consider the optimization problem $\min_{w \in \Delta_{n,\epsilon}} f(w)$ where $f(w) = \|\Sigma_w - I\|_{F,k,k}$. After $\widetilde{O}(d^4/\epsilon^2)$ iterations, projected subgradient descent can output $w \in \Delta_{n,\epsilon}$ such that, with high probability, $\|\mu_w - \mu\|_{2,k} = O(\epsilon\sqrt{\log(1/\epsilon)})$.*

Theorem C.2. *Let $0 < \rho \leq 1$ and $0 < \epsilon < \epsilon_0$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted set of $n = \Omega(k^2 \log d/\epsilon^2)$ samples drawn from a centered subgaussian distribution with covariance $I + \rho vv^\top$ for an unknown k-sparse unit vector $v \in \mathbb{R}^d$. Consider the optimization problem $\min_{w \in \Delta_{n,\epsilon}} f(w)$ where $f(w) = \|M_w - I\|_{F,2k^2}$. After $\widetilde{O}(d^2/\epsilon)$ iterations, projected subgradient descent can output $w \in \Delta_{n,\epsilon}$ such that, with high probability, Algorithm 2 can obtain $u \in \mathbb{R}^d$ from w such that $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\epsilon \log(1/\epsilon)/\rho})$.*

C.1 Algorithmic Results: Robust Sparse Mean Estimation

Recall the objective function $f(w) = \|\Sigma_w - I\|_{F,k,k}$ for robust sparse mean estimation.

Note that $f(w)$ may not be differentiable. To circumvent this, we view $\min_w f(w)$ as a minimax optimization problem:

$$\min_{w \in \Delta_{n,\epsilon}} f(w) = \min_{w \in \Delta_{n,\epsilon}} \max_{Y \in \mathcal{Y}} F(w, Y) \quad \text{where} \quad F(w, Y) = (\Sigma_w - I) \bullet Y ,$$

and $\mathcal{Y} = \{Y \in \mathbb{R}^{d \times d} : \|Y\|_F = 1$ and Y is non-zero in at most k rows and k entries in each row$\}$.

We use projected subgradient descent (PGD) to minimize $f(w) = \max_Y F(w, Y)$ (a formal description of PGD is given in Lemma C.4). In each iteration, we first compute a matrix Y that maximizes $F(w, Y)$ for the current w: Let Q denote the set of k^2 entries that maximizes $\|(\Sigma_w - I)_Q\|_F$, with the constraint that Q contains entries from k rows with k entries in each row (breaking ties arbitrarily).

$$f(w) = \|\Sigma_w - I\|_{F,k,k} = (\Sigma - I) \bullet Y \quad \text{where} \quad Y = \frac{(\Sigma_w - I)_Q}{\|(\Sigma_w - I)_Q\|_F} .$$

We then run (one iteration of) PGD to update w using the subgradient $\nabla_w F(w, Y)$:

$$w \leftarrow \mathcal{P}_{\Delta_{n,\epsilon}}(w - \eta \nabla_w F(w, Y))$$
$$\text{where} \;\; \nabla_w F(w, Y) = \operatorname{diag}(X^\top Y X) - X^\top (Y + Y^\top) X w ,$$

η is the step size of PGD that we will decide later, and $\mathcal{P}_{\mathcal{K}}(\cdot)$ is the ℓ_2 projection operator onto $\mathcal{K}$.

Because f may not be differentiable, we cannot use the notion of stationarity in Definition 2.3. Instead, to prove Theorem C.1, we show that after we run PGD for a sufficient number of iterations, a different kind of approximate stationarity holds. For this notion of approximate stationarity, we need to work with a smoothed variant of the objective function known as the Moreau envelope.

Definition C.3 (Moreau Envelope)**.** *For any function f and closed convex set $\mathcal{K}$, its associated Moreau envelope $f_\beta(w)$ is defined as*

$$f_\beta(w) := \min_{\widetilde{w} \in \mathcal{K}} f(\widetilde{w}) + \beta \|w - \widetilde{w}\|_2^2 .$$

[5] We believe the iteration complexity of robust sparse mean estimation can be improved if we run mirror descent to minimize f (similar to the way [HLZ20] improved the iteration complexity of [CDGS20] for the non-sparse case).

The Moreau envelope can be thought of as a form of convolution between the original function f and a quadratic, so as to smoothen the landscape. In particular, when $f(w)$ takes the form of a maximization problem $f(w) = \max_Y F(w, Y)$ with F a mapping that is β-smooth in the w parameter, the Moreau envelope is also β-smooth. Therefore, the approximate stationarity of the Moreau envelop can be directly defined through its gradient.

To continue, we state a result from [CDGS20] to prove this form of approximate stationarity holds. We omit the proof of Lemma C.4, which generalizes the analysis in recent work (e.g., [DD18]) that provides convergence guarantees for weakly convex optimization problems.

Lemma C.4 (Lemma 4.2 of [CDGS20]). *Let $\mathcal{K}$ be a closed convex set. Let $F(w, Y)$ be a function which is L-Lipschitz and β-smooth with respect to w. Consider the following optimization problem $\min_{w \in \mathcal{K}} f(w)$ where $f(w) = \max_{Y \in \mathcal{Y}} F(w, Y)$.*

Starting from any initial point $w_0 \in \mathcal{K}$, we run iterative updates of the form:

$$Y_\tau = \arg\max_{Y \in \mathcal{Y}} F(w_\tau, Y)$$
$$w_{\tau+1} = \mathcal{P}_{\mathcal{K}}(w_\tau - \eta \nabla_w F(w_\tau, Y_\tau))$$

for T iterations with step size $\eta = \frac{\xi}{\sqrt{T}}$. Then, we have

$$\min_{0 \le \tau < T} \|\nabla f_\beta(w_\tau)\|_2^2 \le \frac{2}{\sqrt{T}} \left(\frac{f_\beta(w_0) - \min_w f(w)}{\xi} + \xi \beta L^2 \right)$$

where $f_\beta(w)$ is the Moreau envelope of f as in Definition C.3.

In our setting, we have $f(w) = \max_{Y \in \mathcal{Y}} F(w, Y)$ with $F(w, Y) = Y \bullet (\Sigma_w - I)$. We will show in Lemma C.5 that $F(w, Y)$ obeys the required Lipschitz and smoothness properties.

We are now ready to prove Theorem C.1.

Proof of Theorem C.1. Note that when $n = \Omega(k^2 \log d / \epsilon^2)$, the original good samples $G^\star$ is (k, ϵ, δ)-stable for $\delta = O(\epsilon \sqrt{\log(1/\epsilon)})$ with probability at least $1 - \exp(-\Omega(k^2 \log d))$.

In addition, we can assume without loss of generality that $\|X_i\|_2 \le O(\sqrt{d \log d})$ for all $i \in S$. We can throw away samples in S that are $\Omega(\sqrt{d \log d})$-far from the empirical median, since with high probability, all good samples are $O(\sqrt{d \log d})$-close to the empirical median. Then we shift all samples by the empirical median, which does not affect the final error guarantee $\|\mu_w - \mu\|_{2,k}$.

We will show in Lemma C.5 that $F(w, Y)$ is L-Lipschitz and β-smoothness with $L = \widetilde{O}(\sqrt{n}d)$ and $\beta = \widetilde{O}(nd)$. In addition, we have $B := f_\beta(w_0) - \min_w f(w) \le f_\beta(w_0) \le f(w_0) \le \widetilde{O}(d)$, and $\gamma = O(n^{1/2} \delta^2 \epsilon^{-3/2})$ from Theorem 3.1.

Therefore, we can apply Lemma C.4 with $\mathcal{K} = \Delta_{n,\epsilon}$ to obtain that after $T \ge O(B \beta L^2 / \gamma^4) = \widetilde{O}(d^4/\epsilon^2)$ iterations, we have $\|\nabla f_\beta(w_\tau)\|_2 \le \gamma$.

The condition $\|\nabla f_\beta(w)\|_2 \le \gamma$ implies that there exists a vector $\widehat{w}$ such that

$$\|\widehat{w} - w\|_2 = \frac{\gamma}{2\beta} \quad \text{and} \quad \min_{g \in \partial f(\widehat{w}) + \partial \mathcal{I}_{\mathcal{K}}(\widehat{w})} \|g\|_2 \le \gamma \; .$$

We first show that $\widehat{w}$ is a good solution. We note that a similar argument was used in [CDGS20] for working with Moreau envelope of the spectral norm.

It is well-known that the subdifferential of the support function is the normal cone, which is in turn the polar of the tangent cone. That is,

$$\partial \mathcal{I}_{\mathcal{K}}(\widehat{w}) = \mathcal{N}_{\mathcal{K}}(\widehat{w}) = (\mathcal{C}_{\mathcal{K}}(\widehat{w}))^\circ \; .$$

Thus, there exists a vector $g = \nu + v$ with $\|g\|_2 \le \gamma$ such that $\nu \in \partial f(\widehat{w})$ and $v \in (\mathcal{C}_{\mathcal{K}}(\widehat{w}))^\circ$. Now consider any unit vector $u \in \mathcal{C}_{\mathcal{K}}(\widehat{w})$:

$$-\gamma \le u^\top g = u^\top \nu + u^\top v \le u^\top \nu \; ,$$

where the last step follows from the definition of the polar set. In other words, there exists a vector $\nu \in \partial f(\widehat{w})$ such that

$$-\nu^\top u \le \gamma \quad \text{for all unit vectors } u \in \mathcal{C}_{\mathcal{K}}(\widehat{w}) \; . \tag{17}$$

This is the notion of stationarity in Definition 2.3 which is used in Theorem 3.1. Because $G^\star$ is (k, ϵ, δ)-stable, by Theorem 3.1, we must have $\|\mu_{\widehat{w}} - \mu\|_2 \le O(\delta)$.

We conclude the proof by noticing that w is very close to $\widehat{w}$, so if $\widehat{w}$ is a good solution, then w must also be a good solution:

$$\begin{aligned} \|\mu_w - \mu\|_2 &\le \|\mu_w - \mu_{\widehat{w}}\|_2 + \|\mu_{\widehat{w}} - \mu\|_2 \\ &\le \|X\|_2 \|w - \widehat{w}\|_2 + O(\delta) = O(\delta) \; . \end{aligned}$$

The last step uses $\|X\|_2 \|\widehat{w} - w\|_2 = \sqrt{n} \max_i \|X_i\|_2 \cdot O(\gamma/\beta) = \widetilde{O}(\delta^2 d^{-1/2} \epsilon^{-3/2}) = O(\delta)$. □

Lemma C.5 upper bounds the Lipschitzness and smoothness parameters of the function $F(w, Y)$ with respect to w.

Lemma C.5. *Fix a set of samples $(X_i)_{i=1}^n$ with $\max_i \|X_i\|_2 = \widetilde{O}(\sqrt{d})$. Fix some $Y \in \mathbb{R}^{d\times d}$ with $\|Y\|_F = 1$. The function $F(w, Y) = (\Sigma_w - I) \bullet Y$ defined over $w \in \Delta_{n,\epsilon}$ is L-Lipschitz and β-smooth with respect to w for $L = \widetilde{O}(\sqrt{n}d)$ and $\beta = \widetilde{O}(nd)$.*

Proof. Recall that $X \in \mathbb{R}^{d\times n}$ is the sample matrix whose i-th column is X_i.

Fix any $w \in \Delta_{n,\epsilon}$.

Recall that

$$\nabla_w F(w, Y) = \operatorname{diag}(X^\top Y X) - X^\top (Y + Y^\top) X w \; .$$

Therefore,

$$\begin{aligned} \|\nabla_w F(w, Y)\|_2 &\le \left\|\operatorname{diag}(X^\top Y X)\right\|_2 + \left\|X^\top (Y + Y^\top) X w\right\|_2 \\ &\le \sqrt{n} \max_i X_i^\top Y X_i + 2 \|X\|_2^2 \|Y\|_2 \|w\|_2 \\ &\le \sqrt{n} \max_i \|X_i\|_2^2 \|Y\|_2 + 2n(\max_i \|X_i\|_2^2) \|Y\|_2 \|w\|_2 \\ &\le \widetilde{O}(\sqrt{n}d) \; . \end{aligned}$$

The last inequality uses the fact that $\max_i \|X_i\|_2 = \widetilde{O}(\sqrt{d})$, $\|Y\|_2 \le \|Y\|_F \le 1$, and $\|w\|_2 \le \sqrt{n} \|w\|_\infty = O(1/\sqrt{n})$.

For the smoothness parameter, note that

$$\nabla_w^2 F(w, Y) = -X^\top (Y + Y^\top) X \; .$$

Thus,

$$\left\|\nabla_w^2 F(w, Y)\right\|_2 \le 2 \|X\|_2^2 \|Y\|_2 \le n(\max_i \|X_i\|_2^2) = \widetilde{O}(nd) \; .$$

This concludes that $L = \widetilde{O}(\sqrt{n}d)$ and $\beta = \widetilde{O}(nd)$. □

C.2 Algorithmic Results: Robust Sparse PCA

Recall that for robust sparse PCA, our objective function is $f(w) = \|M_w - I\|_{F,2k^2}$ where $M_w = \sum_i w_i X_i X_i^\top$.

Note that $f(w)$ is convex in w. Therefore, we can obtain an upper bound on the number of iterations from well-known results on the convergence of projected subgradient descent for L-Lipschitz convex functions (see, e.g., [Bub14]).

Proof of Theorem C.2. Similar to the proof of Theorem C.1, we assume without loss of generality that $G^\star$ is (k, ϵ, δ)-stable for $\delta = O(\epsilon \log(1/\epsilon))$ and $\max_{i\in S} \|X_i\|_2 = \widetilde{O}(\sqrt{d})$, which happens with probability at least $1 - \exp(-\Omega(k^2 \log d))$.

We will show in Lemma C.6 that $f(w)$ is L-Lipschitz for $L = \widetilde{O}(\sqrt{n}d)$. In the proof of Theorem 4.1 in Appendix B, we know that there exists $w \in \Delta_{n,\epsilon}$ with $f(w) \le \rho + \delta$, and moreover, given the stability of $G^\star$, it is sufficient to find a weight vector w with $f(w) \le \rho + O(\delta)$ to obtain the claimed error guarantee $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\epsilon \log(1/\epsilon)/\rho})$.

Therefore, it is sufficient to compute a solution w with $f(w) - \min_w f(w) \le O(\delta)$. It is well-known that (e.g., Theorem 3.2 in [Bub14]), the number of iterations required to compute such w is $T \ge O(R^2 L^2/\delta^2)$, where R is the radius of the feasible region $\Delta_{n,\epsilon}$, and L is the Lipschitz parameter of f. Plugging in $R = O(\sqrt{\epsilon/n})$ and $L = \widetilde{O}(\sqrt{n}d)$, the iteration complexity is $T \ge \widetilde{O}(\frac{\epsilon}{n} \cdot nd^2 \cdot \delta^{-2}) = \widetilde{O}(d^2/\epsilon)$. □

Lemma C.6. *Fix a set of samples $(X_i)_{i=1}^n$ with $\max_i \|X_i\|_2 = \widetilde{O}(\sqrt{d})$. Let $M_w = \sum_{i=1}^n w_i X_i X_i^\top$. The function $f(w) = \|M_w - I\|_{F,2k^2}$ defined over $w \in \Delta_{n,\epsilon}$ is L-Lipschitz for $L = \widetilde{O}(\sqrt{n}d)$.*

Proof. By the definition of the $\|\cdot\|_{F,2k^2}$ norm, we can define

$$f(w) = \max_{Y \in \mathcal{Y}} F(w, Y) \quad \text{where} \quad F(w, Y) = (M_w - I) \bullet Y$$

and $\mathcal{Y} = \{Y \in \mathbb{R}^{d \times d} : \|Y\|_F = 1 \text{ and } Y \text{ has at most } 2k^2 \text{ non-zeros}\}$.

Because the maximum of L-Lipschitz functions is still L-Lipschitz, it is sufficient to upper bound the Lipschitz parameter of $F(w, Y)$ for fixed Y.

We have

$$\nabla_w F(w, Y) = \operatorname{diag}(X^\top Y X) \ .$$

Therefore,

$$\|\nabla_w F(w, Y)\|_2 \le \sqrt{n} \max_i X_i^\top Y X_i \ \le \sqrt{n} \max_i \|X_i\|_2^2 \|Y\|_2 \le \widetilde{O}(\sqrt{n}d) \ .$$

This concludes that $L = \widetilde{O}(\sqrt{n}d)$. □

D Simpler Analysis for Robust Mean Estimation via Gradient Descent

Our analysis in Section 3 can be applied almost directly to general (non-sparse) robust mean estimation. We present the corresponding structural results and proofs in this section. As for the objective function, we will instead use the spectral norm $f(w) = \|\Sigma_w - I\|_2$. We can obtain the main result of [CDGS20] (that natural non-convex formulations of robust mean estimation has no bad stationary points) and greatly simplify their analysis.

The main advantages of our analysis include: (1) our structural result holds under broader distributional assumptions (e.g., subgaussian and bounded covariance distributions), (2) our analysis is shorter and conceptually simpler, and (3) we can show that any γ-approximate stationary point w suffices for a larger value of γ, and consequently, a good w can be found faster.

Theorem D.1. *Fix $0 < \epsilon < \epsilon_0$ and $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (ϵ, δ)-stable (as in Definition D.2) with respect to a d-dimensional ground-truth distribution with unknown mean μ. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$. Let $f(w) = \|\Sigma_w - I\|_2$. Let $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2})$. Then, for any $w \in \Delta_{n,\epsilon}$ that is a γ-stationary point of $f(w)$, we have $\|\mu_w - \mu\|_2 = O(\delta)$.*

When the ground-truth distribution is a spherical Gaussian $\mathcal{N}(\mu, I)$ as in [CDGS20], our sample complexity $n = \widetilde{\Omega}(d/\epsilon^2)$ and error guarantee $\delta = O(\epsilon\sqrt{\log(1/\epsilon})$ both match those in prior works (which are optimal up to logarithmic factors). Moreover, our result (Theorem D.1) states that any γ-stationary point suffices for $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2}) = O(n^{1/2}\epsilon^{1/2}\log(1/\epsilon))$, which is $\sqrt{\epsilon n}$ times larger than the $\gamma = O(\log(1/\epsilon))$ in previous work [CDGS20].

Overview. The high-level idea of our proof is identical to that in Section 3: Given a weight vector w, we show that if w is not a good solution, then moving toward $w^\star$ (the uniform distribution on the good input samples) will decrease the objective value. The changes in the proofs are mostly in using different norms and not having to handle sparsity. Formally, by Lemma 3.6, we get

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta\Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top \ .$$

We can then take spectral norm on both sides (after subtracting the identity matrix) and show that the third term can be essentially ignored. Because w is a bad solution and $w^\star$ is a good solution, $\|\Sigma_w - I\|_2$ must be much larger than $\|\Sigma_{w^\star} - I\|_2$, so the objective function must decrease when we move from w to $(1-\eta)w + \eta w^\star$.

We start with the stability conditions we need for (non-sparse) robust mean estimation.

Deterministic Conditions. For (non-sparse) robust mean estimation, we require the following deterministic conditions (Definition D.2) on the original set of good samples $G^\star$. We refer to these conditions as *stability conditions* because, at a high level, they state that the first and second moments of the good samples are stable when a small fraction of the samples are removed.

Definition D.2 (Stability Conditions)**.** *A set of n samples $G^\star = (X_i)_{i=1}^n$ is said to be (ϵ, δ)-stable (with respect to a distribution with mean μ) iff for any weight vector $w \in \Delta_{n,2\epsilon}$, we have*

$$\|\mu_w - \mu\|_2 \le \delta \quad \text{and} \quad \|\Sigma_w - I\|_2 \le \delta^2/\epsilon \; .$$

Key Lemma. We use Lemma D.3 to relate the spectral norm of the rank-one matrix $(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top$ to that of $(\Sigma_w - I)$, showing that we can essentially ignore this term.

Lemma D.3. *Let $G^\star$ be an (ϵ, δ)-stable set of n samples with $0 < \epsilon \le \delta$. Let S be an ϵ-corrupted version of $G^\star$. Then, for any $w \in \Delta_{n,\epsilon}$, we have*

$$\|\mu_w - \mu_{w_\star}\|_2^2 \le 4\epsilon \left(\|\Sigma_w - I\|_2 + O(\tfrac{\delta^2}{\epsilon}) \right) \; .$$

Proof. Recall that $S = G \cup B$ where G is the set of (remaining) good samples and B is the set of corrupted samples. Let $\alpha = \|w_G\|_1$ and $\beta = \|w_B\|_1$. Let $\overline{w} = w_G/\alpha$ and $\widehat{w} = w_B/\beta$ denote the normalized version of w_G and w_B.

We can write $w = \alpha\overline{w} + \beta\widehat{w}$, by Lemma 3.6, we know that

$$\Sigma_w = \alpha\Sigma_{\overline{w}} + \beta\Sigma_{\widehat{w}} + \alpha\beta(\mu_{\overline{w}} - \mu_{\widehat{w}})(\mu_{\overline{w}} - \mu_{\widehat{w}})^\top \; . \tag{18}$$

Since $\beta \le \|w\|_\infty \cdot |B| \le \frac{\epsilon}{1-\epsilon}$, we have $\|\overline{w}\|_\infty = \frac{\|w_G\|_\infty}{\alpha} \le \frac{1}{(1-\epsilon)n} \cdot \frac{1}{1-\beta} \le \frac{1}{(1-2\epsilon)n}$. Because $G^\star$ is (ϵ, δ)-stable and we can view $\overline{w} \in \Delta_{n,2\epsilon}$ as a weight vector on $G^\star$, by the stability conditions in Definition D.2,

$$\|\Sigma_{\overline{w}} - I\|_2 \le \tfrac{\delta^2}{\epsilon} \; . \tag{19}$$

Using Equations (18) and (19) and that $\Sigma_{\widehat{w}} \succeq 0$, for any unit vector $v \in \mathbb{R}^d$,

$$\begin{aligned} \|\Sigma_w - I\|_2 \ge v^\top(\Sigma_w - I)v &= \alpha v^\top \Sigma_{\overline{w}} v + \beta v^\top \Sigma_{\widehat{w}} v + \alpha\beta\left((\mu_{\overline{w}} - \mu_{\widehat{w}})^\top v\right)^2 - 1 \\ &\ge \alpha\left(1 - \tfrac{\delta^2}{\epsilon}\right) - 1 + \alpha\beta\left((\mu_{\overline{w}} - \mu_{\widehat{w}})^\top v\right)^2 \; . \end{aligned} \tag{20}$$

We know $\alpha(1 - \delta^2/\epsilon)$ is close to 1, so we are essentially left with only the last term on the right-hand side. We will relate this term to $\|\mu_w - \mu_{w_\star}\|_2^2$, which is what appears in the lemma statement.

Recall that $\alpha + \beta = 1$ and $w = \alpha\overline{w} + \beta\widehat{w}$, and thus

$$\beta(\mu_{\widehat{w}} - \mu_{\overline{w}}) = \beta\mu_{\widehat{w}} + \alpha\mu_{\overline{w}} - \mu_{\overline{w}} = \mu_w - \mu_{\overline{w}} = (\mu_w - \mu_{w^\star}) + (\mu_{w^\star} - \mu_{\overline{w}}) \; . \tag{21}$$

Since $\overline{w}, w^\star \in \Delta_{n,2\epsilon}$ and both only put positive weight on samples in G, it follows from the stability conditions (Definition D.2) that

$$\left|(\mu_{w^\star} - \mu_{\overline{w}})^\top v\right| \le \|\mu_{w^\star} - \mu_{\overline{w}}\|_2 \le \|\mu_{w^\star} - \mu\|_2 + \|\mu - \mu_{\overline{w}}\|_2 \le 2\delta \; . \tag{22}$$

We choose $v = \frac{\mu_w - \mu_{w^\star}}{\|\mu_w - \mu_{w^\star}\|_2}$. From Equations (21) and (22), we have

$$\begin{aligned} \left(\beta \cdot (\mu_{\widehat{w}} - \mu_{\overline{w}})^\top v\right)^2 &= \left((\mu_w - \mu_{w^\star})^\top v + (\mu_{w^\star} - \mu_{\overline{w}})^\top v\right)^2 \\ &\ge \frac{\left((\mu_w - \mu_{w_\star})^\top v\right)^2}{2} - \left((\mu_{\overline{w}} - \mu_{w^\star})^\top v\right)^2 \ge \frac{\|\mu_w - \mu_{w_\star}\|_2^2}{2} - 4\delta^2 \; . \end{aligned} \tag{23}$$

The first inequality in Equation (23) uses the fact that $(x+y)^2 \geq \frac{x^2}{2} - y^2$ for any $x, y \in \mathbb{R}$.

Putting Equations (20) and (23) together for our choice of v, we have

$$\begin{aligned} \|\Sigma_w - I\|_2 &\geq \alpha \left(1 - \tfrac{\delta^2}{\epsilon}\right) - 1 + \frac{\alpha}{\beta} \left(\beta \cdot (\mu_{\overline{w}} - \mu_{\widehat{w}})^\top v\right)^2 \\ &\geq \frac{1-2\epsilon}{1-\epsilon} \left(1 - \tfrac{\delta^2}{\epsilon}\right) - 1 + \frac{1-2\epsilon}{\epsilon} \left(\frac{\|\mu_w - \mu_{w_\star}\|_2^2}{2} - 4\delta^2\right) \\ &\geq \frac{1}{4\epsilon} \|\mu_w - \mu_{w_\star}\|_2^2 - O\left(\tfrac{\delta^2}{\epsilon}\right) \ . \end{aligned} \qquad \square$$

Proof of Theorem D.1. We are now ready to prove our main result of this section.

Proof of Theorem D.1. Fix any weight vector $w \in \Delta_{n,\epsilon}$. We will show that if w is a bad solution, then w cannot be an approximate first-order stationary point.

Let c_1 be the constant in $O(\cdot)$ in Lemma D.3. By Lemma D.3, we know that if $\|\mu_w - \mu\|_2 \geq c_2 \delta$ for a sufficiently large constant c_2, then we have $\|\Sigma_w - I\|_2 \geq \frac{\|\mu_w - \mu\|_2^2}{4\epsilon} - c_1 \delta^2 \geq (\frac{c_2^2}{4} - c_1) \frac{\delta^2}{\epsilon} = \Omega(\frac{\delta^2}{\epsilon})$. Recall that $w^\star$ is the uniform distribution on G. We will show that $f(w) = \|\Sigma_w - I\|_2$ decreases if w moves toward $w^\star$. By Lemma 3.6,

$$\Sigma_{(1-\eta)w + \eta w^\star} - I = (1-\eta)(\Sigma_w - I) + \eta(\Sigma_{w^\star} - I) + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top \ .$$

Using the triangle inequality for the spectral norm, we have

$$\left\|\Sigma_{(1-\eta)w+\eta w^\star} - I\right\|_2 \leq (1-\eta) \|\Sigma_w - I\|_2 + \eta \|\Sigma_{w^\star} - I\|_2 + \eta(1-\eta) \|\mu_w - \mu_{w^\star}\|_2^2 \ .$$

We know that $\|\Sigma_{w^\star} - I\|_2 \leq \delta^2/\epsilon$ by the stability conditions in Definition D.2. Using Lemma D.3 and that $f(w) = \|\Sigma_w - I\|_2 = \Omega(\frac{\delta^2}{\epsilon})$, we can show that for all $0 < \eta < 1$,

$$\begin{aligned} f((1-\eta)w + \eta w^\star) &= \left\|\Sigma_{(1-\eta)w+\eta w^\star} - I\right\|_2 \\ &\approx \|(1-\eta)(\Sigma_w - I) + \eta(\Sigma_{w^\star} - I)\|_2 \\ &\leq \|(1-\eta)(\Sigma_w - I)\|_2 \ + \ \|\eta(\Sigma_{w^\star} - I)\|_2 \\ &= (1-\eta) f(w) + \eta f(w^\star) < f(w) \ . \end{aligned} \tag{24}$$

The last inequality requires $(\frac{1}{2} - 4\epsilon) \|\Sigma_w - I\|_2 \geq (4c_1 + 1)\frac{\delta^2}{\epsilon}$, which holds if $\epsilon \leq 1/10$ and we choose $c_2^2 \geq 164\, c_1 + 40$.

It follows immediately that w cannot be a stationary point of f. Let $u = \frac{w^\star - w}{\|w^\star - w\|_2}$ and $h = \eta \|w^\star - w\|_2$. When $h \to 0$, we have $w + hu = (1-\eta)w + \eta w^\star \in \Delta_{n,\epsilon}$ because $w, w^\star \in \Delta_{n,\epsilon}$ and $\Delta_{n,\epsilon}$ is convex. Moreover, we have $\|w^\star - w\|_2 = O(\sqrt{\epsilon/n})$ and therefore

$$u^\top \nabla f(w) = \lim_{h \to 0} \tfrac{f(w+hu) - f(w)}{h} \leq \lim_{\eta \to 0} \tfrac{-(\eta/2) f(w)}{\eta \|w^\star - w\|_2} \leq -\tfrac{\Omega(\delta^2/\epsilon)}{\|w^\star - w\|_2} \leq -\Omega(n^{1/2} \delta^2 \epsilon^{-3/2}) \ .$$

By Definition 2.3, we know w cannot be a γ-stationary point of f for some $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2})$. □

E Structural Results for Robust PCA

In this section, we consider (non-sparse) robust PCA with spiked covariance. In this model, the good samples are drawn from a centered subgaussian distribution with covariance $\Sigma = I + \rho v v^\top$ where $0 < \rho \leq 1$ and $v \in \mathbb{R}^d$ is a unit vector.

We show that our idea in Section 4 for robust sparse PCA can be applied to the non-sparse case. This leads to a new objective function (25) and a simple analysis showing a similar structural result. Consider the following optimization problem:

$$\min_w f(w) = \|M_w - I\|_{\star,2} \quad \text{subject to } \ w \in \Delta_{n,\epsilon} \ , \tag{25}$$

where $M_w = \sum_i w_i X_i X_i^\top$ and $\|A\|_{\star,2}$ is the Ky Fan 2-norm of A (i.e., the sum of the largest two singular values of A).

Because $(M_w - I)$ is always symmetric, we can equivalently define

$$f(w) = \max_{\|u\|_2 = \|v\|_2 = 1, u \perp v} \left|u^\top (M_w - I) u\right| + \left|v^\top (M_w - I) v\right| \ .$$

Note that $f(w)$ is convex in w.

We give some intuition for this objective function. When $w = w^\star$, the uniform distribution on the good samples, $(M_w - I)$ is very close to $\rho v v^\top$. To find a good weight vector with $M_w - I \approx \rho v v^\top$, we minimize the sum of (the absolute values of) the largest two eigenvalues of $(\Sigma_w - I)$. We expect the top eigenvalue to be close to ρ, and consequently, the second largest eigenvalue must be small, which would ensure that $(M_w - I - \rho v v^\top) \approx 0$.

We show that any approximate stationary point w of $f(w)$ yields a good solution for robust PCA. In particular, the algorithm simply outputs the top eigenvector u of $(M_w - I)$.

Theorem E.1. *Let $0 < \rho \leq 1$, $0 < \epsilon < \epsilon_0$, and $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (ϵ, δ)-stable (as in Definition E.2) w.r.t. a centered distribution with covariance $I + \rho v v^\top$ for an unknown unit vector $v \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$.*

Let $f(w)$ be the objective function defined in Equation (25). Then, there exists some $w^\star \in \Delta_{n,\epsilon}$ with $f(w^\star) \leq \rho + \delta$. Moreover, given any weight vector $w \in \Delta_{n,\epsilon}$ with $f(w) \leq \rho + O(\delta)$, we can show that for the top eigenvector u of $(M_w - I)$ satisfies that $\left\|u u^\top - v v^\top\right\|_F = O(\delta/\rho)$.

Theorem E.1 requires the following deterministic conditions on the original good samples.

Definition E.2. *A set of n samples $G^\star = (X_i)_{i=1}^n$ is said to be (ϵ, δ)-stable (with respect to a centered distribution with covariance $\Sigma = I + \rho v v^\top$) iff for any weight vector $w \in \Delta_{n,2\epsilon}$, we have*

$$\|M_w - I\|_{\star,2} \leq \delta \ .$$

where $M_w = \sum_i w_i X_i X_i^\top$ and $\|\cdot\|_{\star,2}$ is the Ky Fan 2-norm.

In particular, when the ground-truth distribution is subgaussian, a set of $n = \Omega(d/\delta^2)$ samples satisfies the stability conditions in E.2 with high probability.

Proof of Theorem E.1. Recall that $w^\star$ is the uniform distribution on the remaining good samples. By the stability conditions in Definition E.2, we know

$$f(w^\star) = \|M_w - I\|_{\star,2} \leq \left\|M_w - I - \rho v v^\top\right\|_{\star,2} + \left\|\rho v v^\top\right\|_{\star,2} \leq \delta + \rho \ .$$

Fix a $w \in \Delta_{n,\epsilon}$ with $f(w) \leq \rho + O(\delta)$. Let $A = M_w - I$. We can write

$$A = \lambda v v^\top + B$$

where $v^\top B v = 0$.

Recall that $w_G = \sum_{i \in G} w_i \geq \frac{1-2\epsilon}{1-\epsilon}$. By the stability conditions in Definition E.2, we know that $\left\|M_w - I - \rho v v^\top\right\|_2 \leq \delta$ for all $w \in \Delta_{n,\epsilon}$, and therefore,

$$\begin{aligned}
\lambda = v^\top A v &= v^\top \left(\sum_{i=1}^n w_i X_i X_i^\top - I \right) v \\
&\geq v^\top \left(\sum_{i \in G} w_i X_i X_i^\top - I \right) v \\
&= v^\top \left(\sum_{i \in G} w_i \left(X_i X_i^\top - I - \rho v v^\top \right) \right) v - (1 - w_G) + w_G \rho \\
&\geq w_G \rho - \left\| \sum_{i \in G} w_i (X_i X_i^\top - I - \rho v v^\top) \right\|_2 - (1 - w_G) \\
&\geq \frac{1-2\epsilon}{1-\epsilon} \rho - \frac{1-\epsilon}{1-2\epsilon} O(\delta) - \frac{\epsilon}{1-\epsilon} \geq \rho - O(\delta) \ .
\end{aligned}$$

Consequently, we must have $\|B\|_2 \leq O(\delta)$. Otherwise, $f(w) = \|A\|_{\star,2} \geq \lambda + \|B\|_2$ would be larger than $\rho + O(\delta)$.

By the matrix perturbation theorem, we have

$$\left\|uu^\top - vv^\top\right\|_F \leq \sqrt{2}\left\|uu^\top - vv^\top\right\|_2 = O\left(\frac{\|B\|_2}{\lambda - \lambda_2}\right) = O(\delta/\rho) \ .$$

In the last step, λ_2 is the second largest eigenvalue of A, which is at most $\|A\|_{\star,2} - \lambda = O(\delta)$, so the eigengap is at least $\rho/2$. □