Robust Sparse Estimation for Gaussians with Optimal Error under Huber Contamination

Ilias Diakonikolas [1] **Daniel Kane** [2] **Sushrut Karmalkar** [1] **Ankit Pensia** [3] **Thanasis Pittas** [1]

Abstract

We study Gaussian sparse estimation tasks in Huber's contamination model with a focus on mean estimation, PCA, and linear regression. For each of these tasks, we give the first sample and computationally efficient robust estimators with optimal error guarantees, within constant factors. All prior efficient algorithms for these tasks incur quantitatively suboptimal error. Concretely, for Gaussian robust k-sparse mean estimation on $\mathbb{R}^d$ with corruption rate $\epsilon > 0$, our algorithm has sample complexity $(k^2/\epsilon^2)\mathrm{polylog}(d/\epsilon)$, runs in sample polynomial time, and approximates the target mean within ℓ_2-error $O(\epsilon)$. Previous efficient algorithms inherently incur error $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$. At the technical level, we develop a novel multi-dimensional filtering method in the sparse regime that may find other applications.

1. Introduction

Robust statistics focuses on developing estimators resilient to a constant fraction of outliers in the sample data (Huber & Ronchetti, 2009; Diakonikolas & Kane, 2023). A data set may have been contaminated by outliers originating from a variety of sources: measurement error, equipment malfunction, data mismanagement, etc. The pivotal question of developing robust estimators in statistics was first posed in the 1960s by Tukey and Huber (Huber, 1964; Tukey, 1960). The standard model for handling outliers, originally formalized in Huber (1964), is defined below.

Definition 1.1 (Huber Contamination Model)**.** Given $0 < \epsilon < 1/2$ and a distribution family $\mathcal{D}$, the algorithm specifies $n \in \mathbb{N}$ and observes n i.i.d. samples from a distribution $P = (1-\epsilon)G + \epsilon B$, where $G \in \mathcal{D}$ and B is arbitrary. We say that G is the distribution of inliers, B the distribution of outliers, and P is the ϵ-corrupted version of G. A set of samples generated in this fashion is called an ϵ-corrupted set of samples from P.

Estimating the parameters of a Gaussian distribution — the prototypical family of distributions in statistics — in the Huber contamination model is a foundational problem in robust statistics (Huber & Ronchetti, 2009). Huber in his foundational work (Huber, 1964) settled the question of robust *univariate* Gaussian mean estimation, i.e., $\mathcal{D} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$. Since then, a large body of work has developed sample-efficient robust estimators for various tasks (Huber & Ronchetti, 2009), that were unfortunately computationally inefficient for high-dimensional tasks. Only in the past decade have the first computationally efficient robust estimators been introduced; see Diakonikolas & Kane (2023) for a book on the topic.

Despite this remarkable progress, perhaps surprisingly, our understanding of this fundamental problem of Gaussian estimation under Huber contamination remains incomplete for *structured* settings.

In many high-dimensional settings, additional structural information about the data can dramatically decrease the sample complexity of estimation. Our focus here is on the structure of *sparsity*. In the context of mean estimation, this corresponds to the regime that at most k out of the d coordinates of the target mean μ are non-zero; our focus is on the practically relevant regime of $k \ll d$. Sparsity has been crucial in improving statistical performance in a myriad of applications (Hastie et al., 2015). We begin with the problem of robust sparse mean estimation.

Definition 1.2 (Robust Sparse Mean Estimation)**.** Given $\epsilon \in (0, 1/2)$, $k \in \mathbb{Z}_+$ and ϵ-corrupted samples from $\mathcal{N}(\mu, \mathbf{I})$ on $\mathbb{R}^d$ under Huber contamination for an unknown k-sparse[1] mean μ, compute an estimate $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2$ is small.

For the above problem, it is known that the information-theoretically optimal error is $\Theta(\epsilon)$. Moreover, the sample complexity of this task is known to be $\mathrm{poly}(k \log d, 1/\epsilon)$ (upper and lower bound); this should be contrasted with the unstructured setting (i.e., dense) whose sample complexity

[*]Equal contribution [1] University of Wisconsin-Madison [2]University of California, San Diego [3]IBM Research. Correspondence to: Sushrut Karmalkar <skarmalkar@wisc.edu>, Ankit Pensia <ankitp@ibm.com >, Thanasis Pittas <pittas@wisc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]We say x is k-sparse if it has at most k non-zero coordinates.

is $\mathrm{poly}(d/\epsilon)$. Hence, we call algorithms with sample complexity $\mathrm{poly}(k, \log d, 1/\epsilon)$ sample-efficient. However, until recently, all known sample-efficient algorithms had running time $d^{\Omega(k)}$ (essentially amounting to brute-force search for the hidden support). The first sample and computationally efficient algorithm for robust sparse mean estimation was given in Balakrishnan et al. (2017), but it incurred an error of $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$.[2] On the other hand, Diakonikolas et al. (2018) gave a computationally-efficient algorithm with error $O(\epsilon)$ for the *dense* setting with sample complexity polynomial in d, thus sample-inefficient. This leads to the question:

Question 1. Is there a sample and computationally efficient robust sparse mean estimator with $O(\epsilon)$ error?

Beyond mean estimation, other important sparse estimation tasks include principal component analysis (PCA) and linear regression, defined below:

Definition 1.3 (Robust Sparse PCA)**.** Given $\epsilon \in (0, 1/2)$, spike strength $\rho > 0$, and a set of ϵ-corrupted samples from $\mathcal{N}(0, \mathbf{I} + \rho vv^\top)$ for an unknown k-sparse unit vector $v \in \mathbb{R}^d$, compute an estimate $\widehat{v}$ such that $\|\widehat{v}\widehat{v}^\top - vv^\top\|_{\mathrm{F}}$ is small.

Robust PCA (in the dense setting) has been studied since Xu et al. (2013), alas with suboptimal error.

Definition 1.4 (Robust Sparse Linear Regression)**.** For $\beta \in \mathbb{R}^d$ and standard deviation $\sigma > 0$, we define $P_{\beta,\sigma}$ to be the joint distribution over (X, y) where $X \sim \mathcal{N}(0, \mathbf{I})$ and $y \sim \mathcal{N}(x^\top\beta, \sigma^2)$. Given $\epsilon \in (0, 1/2)$, $\sigma > 0$, and a set of ϵ-corrupted samples from $P_{\beta,\sigma}$ for an unknown k-sparse $\beta \in \mathbb{R}^d$, compute an estimate $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta\|_2$ is small.

Taking $\rho = \sigma = 1$ for convenience, the optimal errors for both robust sparse PCA and linear regression are still $\Theta(\epsilon)$. Similarly to mean estimation, existing sample and computationally efficient estimators for these problems incur error $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ (Balakrishnan et al., 2017). Focusing on linear regression, the recent work of Diakonikolas et al. (2023b) gave a computationally efficient estimator Definition 1.4 achieving $O(\epsilon)$ error; but since they do not incorporate sparsity, their algorithm inherently requires $\Omega(d)$ samples. For robust PCA, the computational landscape is even less understood: even with $\mathrm{poly}(d)$ samples, no polynomial-time estimator is known that achieves $O(\epsilon)$ error. Thus, we are led to the following question:

Question 2. Are there sample and computationally efficient estimators for robust sparse PCA and robust sparse linear regression that achieve $O(\epsilon)$ error?

We answer both of these questions in the affirmative.

[2]We remark that their algorithm is robust to strong contamination model, which is stronger than Huber contamination model; see Section 1.3 for a thorough discussion.

1.1. Our Results

In what follows, we let $\epsilon_0 \in (0, 1/2)$ be a sufficiently small positive constant. We start with the mean estimation result:

Theorem 1.5 (Robust Sparse Mean Estimation)**.** *For any $\epsilon \in (0, \epsilon_0)$, let T be an ϵ-corrupted set of n samples (in the Huber contamination model) from $\mathcal{N}(\mu, \mathbf{I})$ for an unknown k-sparse mean $\mu \in \mathbb{R}^d$. There exists an algorithm that, given corruption rate $\epsilon \in (0, \epsilon_0)$, failure probability $\delta \in (0, 1)$, sparsity parameter $k \in \mathbb{N}$ and a dataset with $n \geq \frac{k^2 \log d + \log(1/\delta)}{\epsilon^2}\mathrm{polylog}(\frac{1}{\epsilon})$ samples, computes an estimate $\widehat{\mu} \in \mathbb{R}^d$ such that $\|\widehat{\mu} - \mu\|_2 = O(\epsilon)$ with probability at least $1 - \delta$. Moreover, the algorithm runs in $\mathrm{poly}(nd)$-time.*

The error of $O(\epsilon)$ is information-theoretically optimal up to constants[3]. Importantly, Theorem 1.5 is the first sample and computationally efficient algorithm achieving this optimal error guarantee. Moreover, the k^2 dependence in the sample complexity is optimal within the class of computationally efficient algorithms (Diakonikolas et al., 2017; Brennan & Bresler, 2020). For robust sparse PCA, we show:

Theorem 1.6 (Robust Sparse PCA)**.** *For an $\epsilon \in (0, \epsilon_0)$, let T be an ϵ-corrupted set of n samples (in the Huber contamination model) from $\mathcal{N}(0, \mathbf{I} + \rho vv^\top)$ for an unknown k-sparse unit vector $v \in \mathbb{R}^d$ and $\Omega(\epsilon \log(1/\epsilon)) < \rho < 1$. There exists an algorithm that, given corruption rate ϵ, spike strength ρ, a sparsity parameter $k \in \mathbb{N}$, and dataset T with $n := |T| \geq \frac{k^2 \log d}{\epsilon^2}\mathrm{polylog}(1/\epsilon)$ many samples, computes an estimate $\widehat{v} \in \mathbb{R}^d$ such that with probability at least 0.9: (i) $\|\widehat{v}\widehat{v}^\top - vv^\top\|_{\mathrm{F}} = O(\epsilon/\rho)$ and (ii) $\widehat{v}^\top \mathbf{\Sigma} \widehat{v} \geq \left(1 - O\left(\epsilon^2/\rho\right)\right)\|\mathbf{\Sigma}\|_{\mathrm{op}}$ for $\mathbf{\Sigma} := \mathbf{I} + \rho vv^\top$. Moreover, the algorithm runs in $\mathrm{poly}(nd)$-time.*

Similarly, the error guarantee of Theorem 1.6 is optimal (for the considered range of ρ) up to a constant factor, significantly improving on Xu et al. (2013); Balakrishnan et al. (2017); Diakonikolas et al. (2019). Notably, the sample complexity dependence on k^2 is necessary among computationally efficient algorithms, even without outliers (Berthet & Rigollet, 2013a). Even in the dense setting, Theorem 1.6 provides the first polynomial-time algorithm with $O(\epsilon)$ error; in comparison, the only existing algorithm (Diakonikolas et al., 2018) uses quasipolynomial time to get $O(\epsilon)$ error for Definition 1.3. We additionally highlight that the approximation factor of $(1 - O(\epsilon^2/\rho))$ for $\frac{\widehat{v}^\top \mathbf{\Sigma} \widehat{v}}{\|\mathbf{\Sigma}\|_{\mathrm{op}}}$ improves upon the known $1 - O(\epsilon \log(1/\epsilon))$ guarantees achieved without the spike structure (Jambulapati et al., 2020). Finally, for sparse linear regression we show:

Theorem 1.7 (Robust Sparse Linear Regression)**.** *For $\epsilon \in (0, \epsilon_0)$, let T be an ϵ-corrupted set of n samples (in the*

[3]Observe that Theorem 1.5 cannot be extended to identity covariance sub-Gaussian distributions, as the information-theoretic error for this class is $\Theta(\epsilon\sqrt{\log(1/\epsilon)})$.

Huber contamination model) from $P_{\beta,\sigma}$ for an unknown k-sparse regressor $\beta \in \mathbb{R}^d$ and $\|\beta\|_2 = O(\sigma)$, and $\sigma \in \mathbb{R}_+$. There exists an algorithm that, given ϵ, k, and a dataset T with $n := |T| \geq \frac{k^2 \log d}{\epsilon^2}\mathrm{polylog}(1/\epsilon)$ many samples computes an estimate $\widehat{\beta} \in \mathbb{R}^d$ such that $\|\widehat{\beta} - \beta\|_2 = O(\sigma\epsilon)$ with probability at least 0.9. Moreover, the algorithm runs in $\mathrm{poly}(nd)$-time.

Similar to our previous results, the above error is optimal up to a constant, improving upon Balakrishnan et al. (2017); Liu et al. (2020). The dependence on d, ϵ in the sample complexity is similarly nearly optimal for efficient algorithms (Brennan & Bresler, 2020). The restriction on the norm of β is rather mild because of existing algorithm from Liu et al. (2020) which already achieves error $O(\sigma\epsilon\log(1/\epsilon))$ in polynomial time (but with sample complexity depending logarithmically on the initial norm); thus, we could simply use Liu et al. (2020) as a warm start; see Remark 4.2 for further details.

1.2. Our Techniques

At a high-level, we adapt the $O(\epsilon)$ error algorithm of Diakonikolas et al. (2018) to the sparse setting, using ideas from Balakrishnan et al. (2017); Diakonikolas et al. (2019).

We start by explaining the standard filtering algorithms that achieve $\epsilon\sqrt{\log(1/\epsilon)}$ error. Let μ' and $\mathbf{\Sigma}'$ be the empirical mean and the empirical covariance of the (corrupted) data, respectively. Algorithms for robust mean estimation detect outliers by searching for atypical behaviors in $\mathbf{\Sigma}'$. Particularly, if $v^\top \mathbf{\Sigma}' v \geq 1 + C\epsilon\log(1/\epsilon)$ for some direction v, then one can filter points using projections $|v^\top(x-\mu')|^2$, with the guarantee of removing more outliers than inliers (on average). This additional $\log(1/\epsilon)$ factor is necessary here because the ϵ-tail of $(v^\top X)^2$ for $X \sim \mathcal{N}(0, \mathbf{I})$ is at $\log(1/\epsilon)$ (and that of $|v^\top X|$ at $\sqrt{\log(1/\epsilon)}$); without this factor, the algorithm might remove too many inliers. Consequently, when the algorithm stops, there could be directions v with variance $1 + \Theta(\epsilon\log(1/\epsilon))$ such that the ϵn outliers remain $\Omega(\sqrt{\log(1/\epsilon)})$ far from the $v^\top\mu$, leading to a total error of $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ in the algorithm's output.

To improve this error to $O(\epsilon)$, Diakonikolas et al. (2018) makes the following key observation. If there are r (orthogonal) directions $v_1, \ldots, v_r$ all with variance bigger than $1 + C\epsilon$, then (i) either r is small, implying that a brute-force approach can be used to learn the mean optimally in this r dimensional space (in the orthogonal space, the sample mean would already be $O(\epsilon)$ close), or (ii) r is large, in which case, it is unlikely for an inlier to have large projections along r of them *simultaneously* (formalized by the Hanson-Wright inequality), thus permitting us to remove more outliers for large r. Choosing $r = \Theta(\log(1/\epsilon))$, both (i) runs in polynomial time and (ii) removes sufficiently many outliers. The resulting algorithm thus filters until the r-th largest eigenvalue is at most $1 + O(\epsilon)$, thereby decomposing the data into an r-dimensional space V and its complement $V^\perp$ such that the sample mean on $V^\perp$ has error $O(\epsilon)$, while the brute force approach on V also incurs $O(\epsilon)$ error and runs in polynomial time. As a final step, the algorithm adds these two orthogonal estimates.

Adapting this approach to the sparse regime in a sample-efficient manner requires that we filter outliers only along sparse directions v, which immediately hits the roadblock that maximizing $v^\top\mathbf{\Sigma}'v$ over sparse directions v is computationally hard. Thus, robust sparse estimation requires relaxing the objective $v^\top\mathbf{\Sigma}'v = \langle \mathbf{\Sigma}', vv^\top\rangle$ for computational efficiency (while still being sample-efficient). The relaxation of Balakrishnan et al. (2017) maximizes $\langle \mathbf{\Sigma}', \mathbf{A}\rangle$ over PSD matrices $\mathbf{A}$ with unit trace and bounded entry-wise ℓ_1 norm, which is a semidefinite program.[4] If the maximum is larger than $1 + \Omega(\epsilon\log(1/\epsilon))$ with maximizer $\mathbf{A}^*$, one can filter out points x with large score $x^\top\mathbf{A}^*x$. Since the filter relies only on a single "direction" $\mathbf{A}$, this approach is inherently limited to $\epsilon\sqrt{\log(1/\epsilon)}$ error.

Adapting Diakonikolas et al. (2018)'s approach to the relaxation of Balakrishnan et al. (2017) is challenging. Promisingly, it is plausible that one can filter along r orthogonal "directions" $\mathbf{A}_1, \ldots, \mathbf{A}_r$ (sample-efficiently) such that if their average score is $1 + \Omega(\epsilon)$, then one may remove enough outliers. Consequently, at the end of filtering, we can identify r "directions" $\mathbf{A}_1, \ldots, \mathbf{A}_r$ such that all other orthogonal feasible $\mathbf{A}$'s would have small score, i.e., $\langle \mathbf{A}, \mathbf{\Sigma}'\rangle = 1 + O(\epsilon)$. At this point, however, the analogy of $\mathbf{A}$'s being a "direction" breaks down. There is no natural decomposition of the data using $\mathbf{A}$'s into a low-dimensional space V and its orthogonal space $V^\perp$, such that the variance in $V^\perp$ is $1 + O(\epsilon)$ (so that the sample mean is $O(\epsilon)$ close on $V^\perp$).

We instead consider a different relaxation from Diakonikolas et al. (2019) that maximizes $\langle \mathbf{\Sigma}' - \mathbf{I}, \mathbf{A}\rangle$ over k^2-sparse unit Frobenius norm matrices $\mathbf{A}$.[5] Their key observation was that the resulting relaxation is both sample and computationally efficient. Since they filtered along a single $\mathbf{A}$, their algorithm could not achieve $o(\epsilon\sqrt{\log(1/\epsilon)})$ error. However, since their relaxations consider *sparse* matrices $\mathbf{A}$, they naturally lead to a decomposition of coordinates: support of $\mathbf{A}$ and the rest of the coordinates. Inspired by Diakonikolas et al. (2018), we extend Diakonikolas et al. (2019)'s approach as follows: We start with an empty set of coordinates H and find the sparse matrix $\mathbf{A}_1$ that maximizes $\langle \mathbf{\Sigma}' - \mathbf{I}, \mathbf{A}_1\rangle$. If the maximum is larger than $1 + \Omega(\epsilon)$, we add the support of $\mathbf{A}_1$ to H, and proceed to find $\mathbf{A}_2$ that

[4] The relaxation amounts to ignoring the rank constraint and relaxing ℓ_0 norm to ℓ_1 norm.

[5] The relaxation ignores the rank, symmetry, and PSD constraints.

maximizes $\langle (\mathbf{\Sigma}' - \mathbf{I})_{H^\complement}, \mathbf{A}_2 \rangle$, where $(\mathbf{\Sigma}' - \mathbf{I})_{H^\complement}$ is zero on the coordinates in H. We continue until we have either (i) identified r such $\mathbf{A}_i$'s, each with score $1 + \Omega(\epsilon)$, in which case we filter similarly to Diakonikolas et al. (2018); or (ii) we have identified a small set of coordinates, H, such that the sample mean is $O(\epsilon)$ accurate on $H^\complement$. This still leaves the task of estimating the mean on the coordinates in H: although brute force approach is not possible on H, we can invoke the *dense* algorithm from Diakonikolas et al. (2018) on H using fresh samples since $|H| = \tilde{O}(k^2)$.

For sparse PCA, we provide a novel reduction that reduces robust Gaussian PCA to robust (approximate) Gaussian mean estimation. It is crucial here that we maintain the (approximate) Gaussianity in the latter because robust mean estimation for generic subgaussian distributions incurs $\omega(\epsilon)$ error. In fact, even for *dense* robust PCA, our algorithm is the first polynomial-time algorithm to achieve $O(\epsilon)$ error. Recall that our goal is to estimate v from corrupted samples of $X \sim \mathcal{N}(0, \mathbf{I} + \rho v v^\top)$. Given an initial rough approximation w of the spike v, we focus on estimating the correction $z := v - w$. We decompose z as $z := z' + z_\perp$, where z' is parallel to w and $z_\perp \perp w$; the challenge lies in estimating $z_\perp$. Our key observation concerns the conditional distribution of (uncorrupted) samples projected orthogonally to w, conditioned on $w^\top x = a$, which we denote by $X_a^\perp$. It turns out that the distribution of $X_a^\perp$ is Gaussian, with mean proportional to $z_\perp$, and approximately isotropic covariance. Although this insight reduces sparse PCA to (Gaussian) sparse mean estimation, this does not directly lead to an algorithm, because we cannot simulate this conditional sampling exactly (even in the outlier-free setting). We combine our insight with a template from Diakonikolas et al. (2023b), which overcomes similar challenges in linear regression.

1.3. Related Work

Our work lies in the field of robust statistics, initiated in the 1960s (Tukey, 1960; Huber, 1964). We refer the reader to Diakonikolas & Kane (2023) for a comprehensive overview and discuss the most relevant works below. We discuss additional related work in Appendix A.

Focusing on robust sparse estimation, several recent works have developed efficient algorithms in various regimes. Balakrishnan et al. (2017) gave an approach for robust sparse functional estimation and applied it to mean estimation, PCA, and linear regression (among others). While running in polynomial time, the resulting algorithms were not practical because they relied on solving large semidefinite programs. Moreover, their error guarantees are qualitatively suboptimal. Diakonikolas et al. (2019) proposed efficient and practical algorithms for robust mean estimation and PCA. Focusing on sparse mean estimation, subsequent works have proposed further extensions such as Diakonikolas et al. (2022c) for heavy-tailed distributions, Diakonikolas et al. (2022b) for light-tailed distributions with unknown covariance matrix, Cheng et al. (2022) for non-convex first order methods (also see Zhu et al. (2022)), and Diakonikolas et al. (2022a); Zeng & Shen (2022) for list-decodable estimation. Liu et al. (2020) extended the work of Balakrishnan et al. (2017) to sparse linear regression.

Huber's model is the prototypical contamination model in robust statistics (Huber, 1964). Since Diakonikolas et al. (2016); Lai et al. (2016), much of the literature has focused on developing efficient robust algorithms in the *strong contamination model*, which is stronger than Huber's model. Interestingly, the information-theoretic optimal error rate for Gaussian estimation tasks considered in our work is $\Theta(\epsilon)$ in both models. However, all computationally efficient algorithms developed for the strong contamination model incur a larger error of $(\epsilon\sqrt{\log(1/\epsilon)})$ for Gaussians. In fact, Diakonikolas et al. (2017) gives evidence that removing the extra $\log(1/\epsilon))$ factor under strong contamination model is computationally hard. Since our focus is on developing efficient algorithms with error $O(\epsilon)$, we need to restrict our attention to the Huber contamination model.

2. Preliminaries

Notation. We denote $[n] := \{1, \ldots, n\}$. For $w : \mathbb{R}^d \to [0, 1]$ and a distribution P, we use P_w to denote the weighted by w version of P, i.e., the distribution with pdf $P_w(x) = w(x)P(x)/\mathbf{E}_{X\sim P}[w(X)]$. We use $\mu_P, \mathbf{\Sigma}_P$ for the mean and covariance of P. When the vector μ is clear from the context, we use $\overline{\mathbf{\Sigma}}_P$ to denote the second moment matrix of P centered with respect to μ, i.e., $\overline{\mathbf{\Sigma}}_P := \mathbf{E}_{X\sim P}[(X-\mu)(X-\mu)^\top]$. We use $\|\cdot\|_2$ for ℓ_2 norm of vectors and $\|\cdot\|_0$ for the number of non-zero entries in a vector. For a (square) matrix $\mathbf{A}$, we use $\mathrm{tr}(\cdot)$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_{\mathrm{F}}$ for trace, operator, and Frobenius norm. We use $\langle \mathbf{A}, \mathbf{B} \rangle := \mathrm{tr}(\mathbf{A}^\top \mathbf{B}) = \sum_{i,j} A_{j,i} B_{i,j}$ for the inner product between matrices. If $H \subset [d]$ and $v \in \mathbb{R}^d$, we denote by $(v)_H$ the vector x restricted to the entries in H. We use $\mathrm{polylog}()$ to denote a quantity that is poly-logarithmic in its arguments and $\tilde{O}$, $\tilde{\Omega}()$, $\tilde{\Theta}$ to hide such factors.

Definition 2.1 (Sparse Euclidean Norm)**.** For $x \in \mathbb{R}^d$ and $k \in [d]$, we define $\|x\|_{2,k} := \sup_{v:\|v\|_2 \le 1, \|v\|_0 \le k} v^\top x$.

Definition 2.2 (Sparse Frobenius and Operator Norm)**.** For a $d \times d$ matrix $\mathbf{A}$, for $i \in [d]$, let A_i denote the rows of $\mathbf{A}$. We define $\|\mathbf{A}\|_{\mathrm{F},k,k} := \sqrt{\max_{S\subseteq[d]:|S|=k} \sum_{i\in S} \|A_i\|_{2,k}^2}$. For a matrix $\mathbf{A}$, we define $\|\mathbf{A}\|_{\mathrm{op},k} := \sup_{v:\|v\|_0 \le k, \|v\|_2 \le 1} \|\mathbf{A}v\|_2$.

Note the alternative variational definition (proven in Appendix B.1):

Fact 2.3 (Variational definition)**.** $\|\mathbf{A}\|_{\mathrm{F},k,k} = \max_{\mathbf{B}} \langle \mathbf{A}, \mathbf{B} \rangle$ *where the maximum is taken over all matri-*

ces $\mathbf{B}$ *with* $\|\mathbf{B}\|_{\mathrm{F}}=1$ *that have* k *non-zero rows, each of which has at most* k *non-zero elements.*

Moreover, a maximizer $\mathbf{B}$ *of this variational formulation can be found in* $\mathrm{poly}(d,k)$ *time given* $\mathbf{A}$.

We also note the following inequality between (F,k,k) and $\|\cdot\|_{\mathrm{op},k}$ norms:

Fact 2.4. $\|\mathbf{A}\|_{\mathrm{op},k} \le \|\mathbf{A}\|_{\mathrm{F},k,k}$.

2.1. Deterministic Conditions on Inliers

Recall that for a distribution D and a weight function $w:\mathbb{R}^d \to [0,1]$, the distribution D_w denotes the weighted (and appropriately normalized) version of D using w. We further use μ_{D_w} to denote the mean of D_w.

Definition 2.5 ((ϵ,α,k)-goodness)**.** For $\epsilon \in (0,1/2)$, $\alpha > 0$ and $k \in \mathbb{N}$, we say that a distribution G on $\mathbb{R}^d$ is (ϵ,α,k)-good with respect to $\mu \in \mathbb{R}^d$, if the following are satisfied:

(1) For all $w:\mathbb{R}^d \to [0,1]$ with $\mathbf{E}_{X\sim G}[w(X)] \ge 1-\alpha$:

- (1.a) (Mean) $\|\mu_{G_w}-\mu\|_{2,k} \lesssim \alpha\sqrt{\log(1/\alpha)}$.
- (1.b) (Covariance) $\|\overline{\mathbf{\Sigma}}_{G_w}-\mathbf{I}\|_{\mathrm{op},k} \lesssim \alpha\log(\frac{1}{\alpha})$, where $\overline{\mathbf{\Sigma}}_{G_w} := \frac{1}{\mathbf{E}_{X\sim G}[w(X)]}\mathbf{E}_{X\sim G}[w(X)(X-\mu)(X-\mu)^\top]$.

(2) (Tails of *sparse* degree-2 polynomials) If $\mathbf{A}\in\mathbb{R}^{d\times d}$ is a matrix with at most k^2 non-zero elements, $\|\mathbf{A}\|_{\mathrm{F}} \le \sqrt{\log(1/\epsilon)}$ and $\|\mathbf{A}\|_{\mathrm{op}} \le 1$, then the polynomial $p(x) := (x-\mu)^\top\mathbf{A}(x-\mu)-\mathrm{tr}(\mathbf{A})$ satisfies:

- (a) $\mathbf{E}_{X\sim G}[p(X)\mathbb{1}(p(X)>100\log(1/\epsilon))] \le \epsilon$.
- (b) $\Pr_{X\sim G}[p(X)>10\log(1/\epsilon)] \le \epsilon$.
- (c) $\mathbf{E}_{X\sim G}[p(X)\mathbb{1}(h(x)>100\log(1/\epsilon))] \le \epsilon$ for all $h(x)$ of the form $h(x)=\beta+v^\top(x-\mu)$ where $|\beta|\le 1$ and v is k-sparse and unit norm.

(3) $\Pr_{X\sim G}[|v^\top(X-\mu)| \ge 40\log(1/\epsilon))] \le \epsilon$, for all k-sparse unit norm vectors $v\in\mathbb{R}^d$.

We will focus on regime $\alpha = \Theta(\epsilon/\log(1/\epsilon))$. We show in Appendix E (cf. Lemma B.10) that if G is the uniform distribution on a set of $(\frac{k^2}{\epsilon^2})\mathrm{polylog}(\frac{d}{\epsilon})$ i.i.d. samples from $\mathcal{N}(\mu,\mathbf{I})$, then, with high probability, G is $(\epsilon,\Theta(\epsilon/\log(1/\epsilon)),k)$-good with respect to μ.

2.2. Certificate Lemma

The following lemma shows that if the covariance with respect to the weighted distribution P_w is close to the identity along k-sparse directions, then the mean of P_w is close to μ in $(2,k)$-norm. Its proof is similar to prior work but we include it for completeness in Appendix E.

Lemma 2.6 (Certificate Lemma)**.** *Let* $0<\alpha<\epsilon<1/4$. *Let* $P=(1-\epsilon)G+\epsilon B$ *be a mixture of distributions, where* G *satisfies Conditions (1.a) and (1.b) of Definition 2.5 with respect to* $\mu\in\mathbb{R}^d$. *Let* $w:\mathbb{R}^d\to[0,1]$ *be such that* $\mathbf{E}_{X\sim G}[w(X)] > 1-\alpha$. *If* $\|\mathbf{\Sigma}_{P_w}-\mathbf{I}\|_{\mathrm{op},k} \le \lambda$, *then*

$$\|\mu_{P_w}-\mu\|_{2,k} \lesssim \alpha\sqrt{\log\left(\frac{1}{\alpha}\right)}+\sqrt{\lambda\epsilon}+\epsilon+\sqrt{\alpha\epsilon\log\left(\frac{1}{\alpha}\right)}.$$

Motivated by Lemma 2.6, the algorithm starts with weights $w(x)=1$ for all data and aims to iteratively down-weight outliers until $\|\mathbf{\Sigma}_{P_w}-\mathbf{I}\|_{\mathrm{op},k} = O(\epsilon)$; We also need to ensure inliers are not too much downweighted, in the sense $\mathbf{E}_{X\sim G}[w(X)]>1-\alpha$ with $\alpha=\Theta(\epsilon/\log(1/\epsilon))$). If achieved, the total error will be $O(\epsilon)$, as desired. As it will turn out, we will be able to ensure only that a large subspace has small sparse operator norm, not the entire $\mathbb{R}^d$. While we can estimate μ there using Lemma 2.6, we shall use the following estimator on the complement subspace (with resulting sample complexity scaling with the subspace's rank).

Fact 2.7 (Dense Mean Estimation (Diakonikolas et al., 2018; 2023b))**.** *There is a polynomial-time algorithm that, given parameters* $\epsilon\in(0,\epsilon_0), \delta\in(0,1)$ *and* $n \ge \frac{C}{\epsilon^2}(d+\log(1/\delta))\mathrm{polylog}(d/\epsilon)$ *samples, for a large constant* C, *from an* ϵ*-corrupted version of* $\mathcal{N}(\mu,\mathbf{I})$ *in the Huber contamination model, computes an estimate* $\widehat{\mu}$ *such that* $\|\widehat{\mu}-\mu\|_2 = O(\epsilon)$ *with probability at least* $1-\delta$.

2.3. Down-weighting Filter

The filtering step of the algorithm is the following standard procedure: Rescale every x by $w(x)\in[0,1]$ times a non-negative score $\tilde{\tau}(x)\ge 0$, whose role is to quantify our belief about how much of an outlier x is. Let G and B denote the uniform distribution over the inliers and outliers, respectively. The uniform distribution of the entire data is denoted by $P=(1-\epsilon)G+\epsilon B$. If s is a known bound to the weighted scores of inliers, i.e., $\mathbf{E}_{X\sim G}[w(x)\tilde{\tau}(x)] \le s$, then Algorithm 1 checks whether the average score over the entire dataset is abnormally large, i.e., $\mathbf{E}_{X\sim P}[w(x)\tilde{\tau}(x)] > s\beta$ (where $\beta>1$ is a parameter), and if so, down-weighs each point x proportionally to its $\tilde{\tau}(x)$.

The filter guarantees that it removes roughly β-times more mass from outliers than inliers. Given the preceding discussion after Lemma 2.6, we will eventually use $\beta = \Theta(\epsilon/\alpha) = \Theta(\log(1/\epsilon))$. The filter is now standard (see, e.g., Dong et al. (2019); Diakonikolas et al. (2023b) for proofs).

Lemma 2.8 (Filtering Guarantee)**.** *Let* $P=(1-\epsilon)G+\epsilon B$ *be a mixture of distributions supported on* n *points and* $\beta>1$. *If* $(1-\epsilon)\,\mathbf{E}_{X\sim G}[w(X)\tilde{\tau}(X)] < s$, *then the new weights* $w'(x)$ *output by Algorithm 1 satisfy:*

$$(1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}[w(X)-w'(X)] < \frac{\epsilon}{\beta-1}\mathop{\mathbf{E}}_{X\sim B}[w(X)-w'(X)],$$

and the filter can be implemented in $O(n\log(\frac{\tilde{\tau}_{\max}}{s}))$*-time.*

Algorithm 1 Down-weighting Filter

1: **Input**: Distribution P on n points, weights $w(x)$, scores $\tilde{\tau}(x) \geq 0$, threshold $s > 0$, parameter $\beta > 0$. **Output**: New weights $w'(x)$.
2: Initialize $w'(x) \leftarrow w(x)$.
3: $\ell_{\max} \leftarrow \max_x \frac{\tilde{\tau}_{\max}}{es}$, where $\tilde{\tau}_{\max} := \max_{x \in \mathrm{support}(P)} \tilde{\tau}(x)$.
4: **for** $i = 1, \ldots, \ell_{\max}$ **do**
5: **if** $\mathbf{E}_{X \sim P}[w'(X)\tilde{\tau}(X)] > s\beta$ **then**
6: $w'(x) \leftarrow w'(x)(1 - \tilde{\tau}(x)/\max_{x: w(x)>0} \tilde{\tau}(x))$.
7: **end if**
8: **end for**
9: **return** w'.

3. Robust Sparse Mean Estimation

In this section, we sketch our proof of Theorem 1.5.

As mentioned in Section 1.2, the algorithm (stated in Algorithm 2) consists of two parts, summarized below:

- (First phase) First, a loop that iteratively finds sparse and orthogonal maximizers $\mathbf{A}_i, \ldots, \mathbf{A}_r$ of $\mathbf{\Sigma}_w - \mathbf{I}$ for $r = \log(1/\epsilon)$, which are used to filter outliers, until the "average variance" along these $\mathbf{A}_i$'s drops to $O(\epsilon)$ (cf. Line 7)
- (Second phase) After the loop, the algorithm identifies a set H of $k^2 r$-many coordinates informed by the final $\mathbf{A}_i$'s (cf. Line 12). Algorithm then splits the space $\mathbb{R}^d$ into $[d] \setminus H$ and H, and finds an $O(\epsilon)$ approximation of μ for both subspaces separately. For the former, it uses the empirical mean in those coordinates (which ought to be accurate because of Lemma 2.6), and for the latter, it employs a dense mean estimator (cf. Fact 2.7).

In the rest of the section, we formalize this high-level sketch. Throughout the section, we will use the notation $P, \mu_w, \mathbf{\Sigma}_w, \tilde{p}, \tilde{\tau}$ defined in Lines 6 and 9 of the pseudocode. Starting with the first phase, we formally define $\mathbf{A}_1, \ldots, \mathbf{A}_r$ mentioned in the previous paragraph, and we also define what we informally referred to as "average variance along the $\mathbf{A}_i$'s". In particular, $\mathbf{A}_1, \ldots, \mathbf{A}_r$ will be the matrices in Definition 3.1 below for $\mathbf{B} = \mathbf{\Sigma}_w - \mathbf{I}$, and the "average variance along the $\mathbf{A}_i$'s" will be $\frac{1}{r} g_r(\mathbf{\Sigma}_w - \mathbf{I})$.

Definition 3.1. For any matrix $\mathbf{B}$, we define $h_i(\mathbf{B})$, $\mathbf{A}_i$, and H_i for $i \in [r]$ recursively as follows.

- For $i = 1$, $h_1(\mathbf{B}) := \|\mathbf{B}\|_{\mathrm{F},k,k} = \max_{\mathbf{A} \in \mathcal{S}} \langle \mathbf{A}, \mathbf{B} \rangle$ where $\mathcal{S}$ is the set of matrices $\mathbf{A}$ that have $\|\mathbf{A}\|_{\mathrm{F}} = 1$ and have at most k-non-zero rows, each of which has at most k non-zero entries. Let $\mathbf{A}_1$ be the matrix achieving the maximum. The set $H_1 \subseteq [d]$ denotes the rows and columns in which $\mathbf{A}_1$ has non-zero elements.
- For $i \in \{2, \ldots, r\}$, we recursively define $h_i(\mathbf{B})$, $\mathbf{A}_i$, and H_i as follows: $h_i(\mathbf{B}) = \|\mathbf{B}'\|_{\mathrm{F},k,k}$ where $\mathbf{B}'$ is $\mathbf{B}$ after deleting (zeroing out) the rows and columns from $H_1 \cup \cdots \cup H_{i-1}$. Similarly, $\mathbf{A}_i := \mathrm{argmax}_{\mathbf{A} \in \mathcal{S}} \langle \mathbf{A}, \mathbf{B}' \rangle$ and H_i is the non-zero rows and columns of $\mathbf{A}_i$.
- Finally, we define $g_r(\mathbf{B}) := \sum_{i=1}^r h_i(\mathbf{B})$.

Observe that the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_r$ can be computed efficiently using Fact 2.3.

We now explain why we informally call $\frac{1}{r} g_r(\mathbf{\Sigma}_w - \mathbf{I})$ the "average variance along the $\mathbf{A}_i$'s": For each i, $h_i(\mathbf{B})$ represents the mean of the degree-two polynomial $(x - \mu_w)^\top \mathbf{A}_i (x - \mu_w) - \mathrm{tr}(\mathbf{A}_i)$, representing a variance-like quantity of x along $\mathbf{A}_i$. Formally, (see (22) for the details):

$$g_r(\mathbf{\Sigma}_w - \mathbf{I}) = \sum_{i=1}^r \mathop{\mathbf{E}}_{X \sim P_w} [(X - \mu_w)^\top \mathbf{A}_i (X - \mu_w) - \mathrm{tr}(\mathbf{A})] \tag{1}$$

Algorithm 2 Robust Sparse Mean Estimation

1: **Input**: Set of points $T_0 = \{x_i\}_{i \in [n]}$ and $\epsilon > 0$.
2: **Output**: A vector $\widehat{\mu} \in \mathbb{R}^d$.
3: Let C be a sufficiently large constant, and $r := \log(\frac{1}{\epsilon})$.
4: $T \leftarrow$ PREPROCESSING(T_0, ϵ, k). ▶{cf. Fact B.11}
5: Initialize $w(x) \leftarrow \mathbb{1}(\|x - \mu_T\|_2 \leq 10\sqrt{d}\log(d/\epsilon))$.
6: Let P be the uniform distribution on the set T, P_w be the weighted by w version of P (with pdf $P_w(x) = P(x)/\mathbf{E}_{X \sim P}[w(X)]$), $\mu_w := \mathbf{E}_{X \sim P_w}[w(X)]$ and $\mathbf{\Sigma}_w := \mathbf{E}_{X \sim P_w}[(X - \mu_w)(X - \mu_w)^\top]$ the weighted mean and covariance of P. [6]
7: **while** $\frac{1}{r} g_r(\mathbf{\Sigma}_w - \mathbf{I}) > C\epsilon$ **do**
8: Let $\mathbf{A}_1, \ldots, \mathbf{A}_r$ be the matrices from Definition 3.1 for $\mathbf{B} = \mathbf{\Sigma}_w - \mathbf{I}$. (Also see Fact 2.3 for efficient computation.)
9: Define $\tilde{p}(x) := (x - \mu_w)^\top \mathbf{A}(x - \mu_w) - \mathrm{tr}(\mathbf{A})$, and $\tilde{\tau}(x) = \tilde{p}(x)\mathbb{1}(\tilde{p}(x) > 200\log(\frac{1}{\epsilon}))$ for $\mathbf{A} = \sum_{i \in [r]} \mathbf{A}_i$.
10: Update $w \leftarrow$ DOWNWEIGHTINGFILTER$(P, w, \tilde{\tau}, s = \epsilon, \beta = \log(1/\epsilon))$.
11: **end while**
12: For $i \in [r]$, let $H_i \subseteq [d]$ be sets defined in Definition 3.1 and form $H := \bigcup_{i=1}^r H_i$ (cf. Definition 3.1).
13: Run the dense mean estimator from Fact 2.7 on a fresh data restricted to the coordinates in H, to obtain $\widehat{\mu}_1 \in \mathbb{R}^d$ that is zero in every coordinate in $[d] \setminus H$ and satisfies $\|(\widehat{\mu}_1 - \mu)_H\|_{2,k} = O(\epsilon)$.
14: Let $\widehat{\mu}_2$ be the vector that is equal to $\mathbf{E}_{P_w}[X]$ in the coordinates in $[d] \setminus H$ and zero in the coordinates in H.
15: **Return** $\widehat{\mu} = \widehat{\mu}_1 + \widehat{\mu}_2$.

[6] The weights $w(x)$ may change in the course of the algorithm; $\mu_w, \mathbf{\Sigma}_w$ will denote the quantity based on the latest weights.

3.1. Proof Overview of Theorem 1.5

The proof of correctness consists of the following claims:

1. For any iteration of line 7, if $w(x), w'(x)$ denote the weights before and after the iteration:
   (a) ($\log(1/\epsilon)$ more outliers than inliers are removed) $\mathop{\mathbf{E}}_{X\sim G}[w(X)-w'(X)] < \frac{2\epsilon}{\log(\frac{1}{\epsilon})} \mathop{\mathbf{E}}_{X\sim B}[w(X)-w'(X)]$.
   (b) (Non-trivial mass is removed) $\mathbf{E}_{X\sim P}[w(X) - w'(X)] = \tilde{\Omega}(\epsilon/d)$.
2. After the loop ends, $\|(\widehat{\mu}_2 - \mu)_{[d]\setminus H}\|_{2,k} = O(\epsilon)$.[7]

Item 1b means that the algorithm terminates after $\tilde{O}(d/\epsilon)$ iterations (since no outliers are left at that point), Item 1a means that $\mathbf{E}_{X\sim G}[w(X)] \geq 1-3\epsilon/\log(1/\epsilon)$ is an invariant condition throughout the loop, and Item 2 states that the empirical mean is accurate on the coordinates from $[d] \setminus H$.

Before proving these claims, we show how they imply Theorem 1.5. We decompose μ into $\mu_1 + \mu_2$ for $\mu_1 = (\mu)_H$ and $\mu_2 := (\mu)_{[d]\setminus H}$. By triangle inequality and definition of $\widehat{\mu}$, we have $\|\widehat{\mu} - \mu\|_{2,k} \leq \|\widehat{\mu}_1 - \mu_1\|_{2,k} + \|\widehat{\mu}_2 - \mu_2\|_{2,k}$. We bound each of these terms by $O(\epsilon)$. The inequality $\|\widehat{\mu}_1 - \mu_1\|_{2,k} \leq O(\epsilon)$ corresponds to the guarantee of the dense estimator (Fact 2.7), run on a fresh dataset, when restricted to coordinates in H. Fact 2.7 gives an estimator with $O(\epsilon)$ error and sample complexity $\frac{1}{\epsilon^2}(|H| + \log(1/\delta))\mathrm{polylog}(d/\epsilon)$. Since $|H| \leq k^2 \log(1/\epsilon)$, the setting satisfies the assumptions of Theorem 1.5. The term $\|\widehat{\mu}_2 - \mu_2\|_{2,k}$ is equal to $\|(\widehat{\mu}_2 - \mu)_{[d]\setminus H}\|_{2,k}$ and thus $O(\epsilon)$ by Item 2.

We now sketch the proofs of Items 1a, 1b and 2 used above.

Proof of Item 2 Once Item 1b is shown, it implies that $\mathbf{E}_{X\sim G}[w(X)] \geq 1-3\epsilon/\log(1/\epsilon)$ and thus Item 2 becomes a straightforward application of the Certificate Lemma 2.6 with $\alpha = 3\epsilon/\log(1/\epsilon)$ and $\lambda = C\epsilon$ to agree with the stopping condition of line 7 (for this we need to show that the condition of line 7 implies $\|(\mathbf{\Sigma}_w - \mathbf{I})_{([d]\setminus H)\times([d]\setminus H)}\|_{\mathrm{op},k} = O(\epsilon)$; see Claim C.4 in Appendix C for the details).

Proof of Item 1a We want to use Lemma 2.8 with $\beta = \log(1/\epsilon)$ and $s = \epsilon$. To apply the lemma, we need to show that $\mathbf{E}_{X\sim G}[w(X)\tilde{\tau}(X)] \leq \epsilon$. This looks like Item (2)a of the goodness conditions (Definition 2.5) but the difference is that $\tilde{\tau}(x)$ centers the point around μ_w instead of μ that is used in $\tau(x)$. While these centering issues are easily dealt with when $\mathbf{A}$ is PSD and no sparsity constraints are present, our setting requires additional technical work. We defer the full proof to Appendix C, sketching the steps here.

[7] Recall $(x)_H$ denotes the vector x restricted to $H \subset [d]$.

Let $\tilde{\tau}(x) = \tilde{p}(x)\mathbb{1}(\tilde{p}(x) > 200\log(1/\epsilon))$, where $\tilde{p}(x) = (x-\mu_w)^\top \mathbf{A}(x-\mu_w) - \mathrm{tr}(\mathbf{A})$; the algorithm uses $\tilde{\tau}(x)$ as scores. Define $\tau(x) = p(x)\mathbb{1}(p(x) > 100\log(1/\epsilon))$, with $p(x) = (x-\mu)^\top \mathbf{A}(x-\mu) - \mathrm{tr}(\mathbf{A})$, to be the ideal scores appearing in the deterministic condition (that center data around the true μ). Denote the difference of these polynomials by $\Delta p(x) := \tilde{p}(x) - p(x) = \delta_\mu^\top \mathbf{A}\delta_\mu + (x-\mu)^\top \mathbf{A}\delta_\mu + (x-\mu)^\top \mathbf{A}^\top \delta_\mu$ for $\delta_\mu := \mu - \mu_w$. Using triangle inequalities, we get

$$\begin{aligned}
\mathop{\mathbf{E}}_{X\sim G}[w(X)\tilde{\tau}(X)] \quad &\leq \quad |\delta_\mu^\top \mathbf{A}\delta_\mu| \\
&+ \mathop{\mathbf{E}}_{X\sim G}[(X-\mu)^\top \mathbf{A}\delta_\mu \mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \\
&+ \mathop{\mathbf{E}}_{X\sim G}[(X-\mu)^\top \mathbf{A}^\top \delta_\mu \mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \\
&+ \mathop{\mathbf{E}}_{X\sim G}[p(X)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))]. \qquad (2)
\end{aligned}$$

We need to bound all three terms above by $O(\epsilon)$. For the first term, we take advantage of the sparsity of $\mathbf{A}$ to establish:

Claim 3.2. *Let $\mathbf{A} = \sum_{\ell\in[r]} \mathbf{B}^{(\ell)}$ where each $\mathbf{B}^{(\ell)}$ is a square matrix with Frobenius norm equal to one, k non-zero rows, each of which has k non-zero entries. Then, for any vectors u, v, it holds $|u^\top \mathbf{A} v| \leq r\|u\|_{2,k}\|v\|_{2,k}$.*

This means that $|\delta_\mu^\top \mathbf{A}\delta_\mu| \leq \log(1/\epsilon)\|\delta_\mu\|_{2,k}^2$, which can eventually be bounded by ϵ using Lemma 2.6 and the preprocessing of Line 4. The second term in (2) may be broken into two terms by considering the cases $\Delta p(X) \leq 100\log(1/\epsilon)$ and $\Delta p(X) > 100\log(1/\epsilon)$. The latter case is a very low-probability event by Item (3), eventually bounding the relevant term by ϵ. For the former case, we can use that $p(X) > 200\log(1/\epsilon)$ is a low-probability event (by Item (2)b of Definition 2.5). The third term uses an identical argument. Finally, the third term in (2) is similarly split into two by taking cases for $\Delta p(X)$, and bounding each one using either Item (2)a or Item (2)c of the Definition 2.5.

Proof of Item 1b By design, the down-weighting filter only removes mass when $\mathbf{E}_{X\sim P}[w(X)\tilde{\tau}(X)] \geq s\beta =: \epsilon\log(1/\epsilon)$ (cf. line 5 of Algorithm 1). Thus, we first need to show that this is true throughout the loop of Line 7. To do so, we write $\mathbf{E}_{X\sim P}[w(X)\tilde{\tau}(X)] = \mathbf{E}_{X\sim P}[w(X)\tilde{p}(X)] - \mathbf{E}_{X\sim P}[w(X)\tilde{\tau}(X)\mathbb{1}(\tilde{p}(X) \leq 200\log(1/\epsilon))]$. The first term is already roughly $g_r(\mathbf{\Sigma}_w - \mathbf{I})$ by (1), up to a normalization of $\mathbf{E}[w(x)] \approx 1$, and hence at least $\epsilon\log(1/\epsilon)$ by Line 7). Showing that the second term is less than $g_r(\mathbf{\Sigma}_w - \mathbf{I})/2$ involves a multi-step argument similar to the ones in the previous paragraph, which can be found in Appendix C. Once this is established, Item 1b follows easily: First, by design of the down-weighting filter, $\mathbf{E}_{X\sim P}[w(X) - w'(X)] = \mathbf{E}_{X\sim P}[w(X)\tilde{\tau}(X)]/\max_x \tilde{\tau}(X)$. We have already shown that the numerator is $\Omega(\epsilon)$. The denominator is $O(d\,\mathrm{polylog}(d/\epsilon))$ since $\|x\|_2 = \sqrt{d}\log(d/\epsilon)$ for all points in the dataset, by Gaussian concentration.

4. Robust Sparse PCA and Linear Regression

4.1. Robust PCA

In this section, we show Theorem 1.6 via a novel reduction to mean estimation (which also implies new results for the dense setting). A natural first attempt is to consider the following existing reduction to mean estimation (see, e.g., Diakonikolas et al. (2019)): the mean of $\mathrm{vec}(XX^\top - \mathbf{I})$ for $X \sim D$, where vec denotes the operator that converts matrices to vectors by stacking its rows, is exactly $\rho vv^\top$, which is also $\mathrm{poly}(k)$ sparse. However, the distribution of $\mathrm{vec}(XX^\top - \mathbf{I})$ is not Gaussian but rather a second power of a Gaussian, thus existing mean estimators would only yield $O(\epsilon \log(1/\epsilon))$ error. We propose a different reduction, which does not lose this $\log(1/\epsilon)$ factor. Let w be a unit vector that is an $(\epsilon\sqrt{\log(1/\epsilon)})$-approximation of the spike v (e.g., using Balakrishnan et al. (2017)). Denote the projection of X onto the subspace orthogonal to w by $\mathrm{Proj}_{w^\perp}(X)$.[8] Our key idea is that $\mathrm{Proj}_{w^\perp}(X)$ conditioned on $w^\top x = \alpha$ can give information about $\mathrm{Proj}_{w^\perp}(v - w)$, i.e., the correction $v - w$ in that subspace. We prove the following simple claim in Appendix D.

Claim 4.1. *Let $X \sim \mathcal{N}(0, \mathbf{I} + \rho vv^\top)$ be a random variable from the spiked covariance model and w be a unit vector. Let $Z = \mathrm{Proj}_{w^\perp}(X)$, the projection of X onto the subspace perpendicular to w. For $\alpha \in \mathbb{R}$, let G_α denote the distribution of Z conditioned on $w^\top X = \alpha$. Then the distribution G_α is equal to $\mathcal{N}(\tilde{\mu}, \tilde{\mathbf{\Sigma}})$ with*

$$\tilde{\mu} = \frac{\rho(w^\top v)\alpha}{1 + \rho(w^\top v)^2}\bar{v} \quad \textit{and}$$
$$\tilde{\mathbf{\Sigma}} = \mathbf{I} + \frac{\rho}{1 + \rho(w^\top v)^2}\bar{v}\bar{v}^\top,$$

where $\bar{v} := \mathrm{Proj}_{w^\perp}(v) = v - (w^\top v)w$.

We use the result above to estimate $\mathrm{Proj}_{w^\perp}(v)$, and our final estimate, $\widehat{v}$, shall be $\widehat{v}_1 + \widehat{v}_2$, where $\widehat{v}_1$ estimates $\mathrm{Proj}_w(v) = (w^\top v)w$ and $\widehat{v}_2$ estimates $\mathrm{Proj}_{w^\perp}(v)$. Importantly, the mean of G_α is a scaled version of $\mathrm{Proj}_{w^\perp}(v)$, and thus if $z \approx \mu_{G_\alpha}$, we could use $\widehat{v}_2 = z(\rho\alpha(w^\top v))/(1 + \rho(w^\top v)^2)$. However, since $w^\top v$ is unknown, we need to estimate it from data; we also need it to estimate $\widehat{v}_1$. Note that $1 + \rho(w^\top v)^2$ is the variance of $X^\top w$. Thus we can use one-dimensional (robust) variance estimation algorithm to find y such that $|y - (w^\top v)^2| = O(\epsilon/\rho)$. This leads to Algorithm 3.

We show that the final error, i.e. $\|\widehat{v} - v\|_2 \le \|z\frac{1+\rho y}{\rho\sqrt{y}\alpha} - \mathrm{Proj}_{w^\perp}(v)\|_2 + |\sqrt{y} - w^\top v|$, is $O(\epsilon/\rho)$ using the guarantees of the two aforementioned estimators (i.e., that $|y - (w^\top v)^2| = O(\epsilon)$ and $\|z - \mu_{G_\alpha}\|_2 = O(\epsilon)$); see Claim D.3.

[8] For a vector u, we use $\mathrm{Proj}_{u^\perp}(\cdot)$ to denote the projection operator on the null space of u.

Algorithm 3 Reduction from PCA to Mean Estimation

1: Find unit vector w such that $\|ww^\top - vv^\top\|_{\mathrm{F}} = O(\epsilon\sqrt{\log(1/\epsilon)}/\rho)$.
  ▶{For example, using Balakrishnan et al. (2017)}
2: Find y: $|y - (w^\top v)^2| = O(\epsilon/\rho)$.
  ▶{by robustly estimating the variance of $(w^\top x)$ (see Claim D.4)}
3: Fix an $\alpha = \Omega(1)$ and find z with $\|z - \mu_{G_\alpha}\|_2 = O(\epsilon)$, where G_α is the conditional distribution from Claim 4.1.
  ▶{e.g., using Algorithm 2 (see Claim D.4 for details)}
4: Return $\widehat{v} = z\frac{1+\rho y}{\rho\sqrt{y}\alpha} + w\sqrt{y}$.

To complete an overview of the proof of Theorem 1.6, we need to explicitly show how to obtain z in Line 3 of Algorithm 3 using our sparse mean estimator, Theorem 1.5. An obvious issue is that we cannot simulate samples from G_α using $X \sim \mathcal{N}(0, \mathbf{I} + \rho vv^\top)$ by rejection sampling—let alone that is at most $O(\epsilon)$ corrupted given ϵ-corrupted X—because the probability that a sample has $w^\top x = \alpha$ is zero. To overcome this, we use insights from Diakonikolas et al. (2023b) and relax this procedure by instead conditioning on samples to be in a thin interval I around α. The resulting pseudocode is as follows:

1. Draw α uniformly from $[-(1+\rho), 1+\rho]$ and define the interval $I := [\alpha - \ell, \alpha + \ell]$ for $\ell = 1/\log(1/\epsilon)$.
2. $T' = \{\mathrm{Proj}_{w^\perp}(x) : x \in T \text{ and } w^\top x \in I\}$.
3. Let z be the output of Algorithm 2 run on T'.

Although conditioning on an interval increases the probability and thus permits efficient rejection sampling, the downside is that the conditional distribution on $w^\top x \in I$ is no longer a Gaussian distribution, but instead a continuous mixture of Gaussians:

$$G_I(z) = \frac{\int_{\alpha' \in I} G_{\alpha'}(z)\,\mathrm{Pr}_{X\sim\mathcal{N}(0,\mathbf{I}+\rho vv^\top)}[w^\top X = \alpha']\mathrm{d}\alpha'}{\mathrm{Pr}_{X\sim\mathcal{N}(0,\mathbf{I}+\rho vv^\top)}[w^\top X \in I]}.$$

The mean of the mixture, μ_{G_I}, may shift away from μ_{G_α}, and thus we need the length ℓ of I to be small enough so that the shift is $O(\epsilon)$. Moreover, G_I is not Gaussian, and thus Theorem 1.5 is not applicable in a black-box manner. However, for small ℓ, it is close enough to a Gaussian so that the deterministic conditions of Definition 2.5 hold with respect to μ_{G_α} (the details are deferred to Appendix D).

To ensure applicability of our mean estimator, it remains to show that the fraction of outliers in the conditional dataset T' is $O(\epsilon)$. This is why the center α of the interval needs to be chosen randomly (as in Diakonikolas et al. (2023b)); otherwise, the outlier distribution could happen to have all

outliers x satisfying $w^\top x = \alpha$. To show that T' is $O(\epsilon)$-corrupted, it suffices to check that the probability of an outlier x satisfying $w^\top x \in I$ divided by the probability that an inlier x' having $w^\top x' \in I$ is at most $O(1)$. Since I is chosen independently of everything, we can imagine that the outlier x is fixed and only I is drawn randomly. Let us examine only the case where $w^\top x \in [-2(1+\rho), 2(1+\rho)]$ (because otherwise $w^\top x \notin I$). The probability that $w^\top x \in I$ is then the ratio of the length of I to the length of the interval $[-2(1+\rho), 2(1+\rho)]$, i.e., $O(\ell/(1+\rho))$. Regarding the inliers x', we can use the same trick to imagine that I is fixed and the inlier x' is drawn from $\mathcal{N}(0, I + \rho vv^\top)$. Note that $w^\top x' \sim \mathcal{N}(0, \tilde{\sigma}^2)$ with $\tilde{\sigma}^2 := 1 + \rho(w^\top v)^2$. Since $I \subseteq [-3\tilde{\sigma}^2, 3\tilde{\sigma}^2]$, the Gaussian distribution behaves approximately uniformly there and thus the probability that $w^\top x \in I$ is $\Omega(\ell/\tilde{\sigma})$, which is also $\Omega(\ell/(1+\rho))$ by using $|w^\top v|^2 = \Omega(1)$.

4.2. Robust Sparse Linear Regression

We conclude with Theorem 1.7, which follows by a reduction to mean estimation from Diakonikolas et al. (2023b). Their reduction seamlessly extends to the sparse setting considered in this paper, thus we describe it only briefly.

As a first step, using Liu et al. (2020) as preprocessing, we may assume that $\|\beta\|_2 \lesssim \sigma\epsilon\log(1/\epsilon)$. Analogously to Claim 4.1, Diakonikolas et al. (2023b, Claim 4.1) shows the following: let Q_a denote the conditional distribution of X, conditioned on $y = \alpha$ for $(X, y) \sim P_{\beta,\sigma}$ in Definition 1.4, then $Q_a \sim \mathcal{N}((\alpha/\sigma_y^2)\beta, \mathbf{I} - \beta\beta^\top/\sigma_y^2)$ for $\sigma_y^2 := \sigma^2 + \|\beta\|_2^2$. Since Q_a is an approximately isotropic Gaussian with a *sparse* mean, we can hope to estimate it with $O(\epsilon)$ error in a sample-efficient way by Theorem 1.5. Finally, to obtain Theorem 1.7, one needs to use similar tricks as in the last section (such as conditioning on a random interval of an appropriate length instead of a fixed point). Fortunately, all of these approximations suffice to get $O(\epsilon)$ error as in Diakonikolas et al. (2023b). The final algorithm is given in Algorithm 4.

Algorithm 4 Robust Linear Regression

Input: Set of points $T = \{(x_i, y_i)\}_{i\in[n]}$ and $\epsilon > 0$.
Output: A vector $\widehat{v} \in \mathbb{R}^d$.
Find $\widehat{\sigma}_y$ such that $|\widehat{\sigma}_y^2 - \sigma_y^2| = O(\sigma_y^2\epsilon\log(1/\epsilon))$.
Draw $\alpha \in \mathbb{R}$ uniformly at random from $[-\widehat{\sigma}_y, \widehat{\sigma}_y]$.
Define $I = [\alpha - \ell, \alpha + \ell]$ for $\ell := \widehat{\sigma}_y/\log(1/\epsilon)$.
$T' \leftarrow \{x : (x, y) \in T, y \in I\}$.
Let $\widehat{\beta}_I$ be the output of Algorithm 2 on T'.
Return $\widehat{\beta} := (\widehat{\sigma}_y^2/\alpha)\widehat{\beta}_I$.

Remark 4.2. We further expand on the norm constraint of $\|\beta\|_2 = O(\sigma)$ in Theorem 1.7. The algorithm in (Liu et al., 2020) obtains an error of $O(\sigma\epsilon\log(1/\epsilon))$ but their sample complexity scales multiplicatively with $\log(\|\beta\|_2/\epsilon\sigma)$. If we do not assume a norm constraint on β, the sample complexity in Theorem 1.7 would also have an extra multiplicative term of $\log(\|\beta\|_2/\epsilon\sigma)$ if we use (Liu et al., 2020) as a warm-start. However, this factor of $\log(\|\beta\|_2/\epsilon\sigma)$ does not appear in the information-theoretical rate and can potentially be removed using either a tighter analysis of (Liu et al., 2020) or a different (computationally-efficient) algorithm.

5. Discussion

In this paper, we presented the first computationally-efficient algorithms that achieve the information-theoretic optimal error under Huber contamination for various sparse estimation tasks. We now discuss some immediate open problems. Starting with mean estimation, our algorithm (Theorem 1.5) needs to know the covariance matrix of the inlier distribution; developing a sample and computationally-efficient algorithm for an unknown covariance matrix remains an important open problem. More broadly, one could consider robust covariance estimation in the sparse operator norm or the (F, k, k) norm. For robust PCA (Theorem 1.6), our algorithm works for a somewhat restricted range of the spike parameter ρ. Removing this spike assumption and, more broadly, developing a sparse PCA algorithm for the gap-free setting (similar to Jambulapati et al. (2020); Kong et al. (2020); Diakonikolas et al. (2023a) in the dense setting) remains open.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Balakrishnan, S., Du, S. S., Li, J., and Singh, A. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pp. 169–212, 2017.

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

Berthet, Q. and Rigollet, P. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780, 2013b.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5.

Brennan, M. and Bresler, G. Reducibility and statistical-

computational gaps from secret leakage. In *Conference on Learning Theory*, pp. 648–847. PMLR, 2020.

Brennan, M. S., Bresler, G., Hopkins, S., Li, J., and Schramm, T. Statistical query algorithms and low degree tests are almost equivalent. In *Conference on Learning Theory*, pp. 774–774. PMLR, 2021.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 1–37, 2011.

Canonne, C., Hopkins, S. B., Li, J., Liu, A., and Narayanan, S. The full landscape of robust mean testing: Sharp separations between oblivious and adaptive contamination. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2023.

Chen, Y., Caramanis, C., and Mannor, S. Robust sparse regression under adversarial corruption. In *International conference on machine learning*, pp. 774–782. PMLR, 2013.

Cheng, Y., Diakonikolas, I., Ge, R., and Woodruff, D. P. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory, COLT 2019*, pp. 727–757, 2019.

Cheng, Y., Diakonikolas, I., Kane, D. M., Ge, R., Gupta, S., and Soltanolkotabi, M. Outlier-robust sparse estimation via non-convex optimization. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Croux, C. and Haesbroeck, G. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618, 2000.

Croux, C., Filzmoser, P., and Fritz, H. Robust sparse principal component analysis. *Technometrics*, 55(2):202–214, 2013.

Diakonikolas, I. and Kane, D. M. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robust estimators in high dimensions without the computational intractability. In *Proceedings of FOCS'16*, pp. 655–664, 2016.

Diakonikolas, I., Kane, D. M., and Stewart, A. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pp. 73–84, 2017.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018*, pp. 2683–2702, 2018.

Diakonikolas, I., Karmalkar, S., Kane, D., Price, E., and Stewart, A. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Diakonikolas, I., Kane, D. M., Karmalkar, S., Pensia, A., and Pittas, T. List-Decodable Sparse Mean Estimation via Difference-of-Pairs Filtering. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022a.

Diakonikolas, I., Kane, D. M., Karmalkar, S., Pensia, A., and Pittas, T. Robust sparse mean estimation via sum of squares. In *Conference on Learning Theory*, pp. 4703–4763. PMLR, 2022b.

Diakonikolas, I., Kane, D. M., Lee, J. C. H., and Pensia, A. Outlier-Robust Sparse Mean Estimation for Heavy-Tailed Distributions. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022c.

Diakonikolas, I., Kane, D. M., Pensia, A., and Pittas, T. Nearly-linear time and streaming algorithms for outlier-robust PCA. In *International Conference on Machine Learning*, 2023a.

Diakonikolas, I., Kane, D. M., Pensia, A., and Pittas, T. Near-optimal algorithms for gaussians with huber contamination: Mean estimation and linear regression. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023b.

Dong, Y., Hopkins, S. B., and Li, J. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *Advances in Neural Information Processing Systems*, 32:6067–6077, 2019.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN 1498712169, 9781498712163.

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

Huber, P. J. and Ronchetti, E. M. *Robust Statistics*. John Wiley & Sons, 2009.

Jambulapati, A., Li, J., and Tian, K. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Kong, W., Somani, R., Kakade, S., and Oh, S. Robust meta-learning for mixed linear regression with small batches. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Lai, K. A., Rao, A. B., and Vempala, S. Agnostic estimation of mean and covariance. In *Proceedings of FOCS'16*, 2016.

Li, J. Z. *Principled approaches to robust machine learning and beyond*. PhD thesis, Massachusetts Institute of Technology, 2018.

Liu, L., Shen, Y., Li, T., and Caramanis, C. High dimensional robust sparse regression. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

Moshksar, K. On the absolute constant in hanson-wright inequality. *arXiv preprint arXiv:2111.00557*, 2021.

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

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. ISBN 978-1-108-62777-1 978-1-108-49802-9.

Wang, T., Berthet, Q., and Samworth, R. J. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, pp. 1896–1930, 2016.

Xu, H., Caramanis, C., and Sanghavi, S. Robust pca via outlier pursuit. *Advances in neural information processing systems*, 23, 2010.

Xu, H., Caramanis, C., and Mannor, S. Outlier-robust pca: The high-dimensional case. *IEEE Trans. on Information Theory*, 59(1):546–572, 2013.

Zeng, S. and Shen, J. List-decodable sparse mean estimation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Zhu, B., Jiao, J., and Steinhardt, J. Robust estimation via generalized quasi-gradients. *Information and Inference: A Journal of the IMA*, 11(2):581–636, 2022. doi: 10.1093/imaiai/iaab018.

Supplementary Material

The supplementary material is structured as follows: Appendix A discusses additional related work, Appendix B includes omitted preliminaries, Appendix C provides the full proof of Theorem 1.5 for robust sparse mean estimation, Appendix D provides the proof of Theorem 1.6 for PCA, and finally, Appendix E includes omitted proofs from Appendix B.

A. Additional Related Work

Our work lies in the field of robust statistics, initiated in the 1960s (Tukey, 1960; Huber, 1964). We refer the reader to Diakonikolas & Kane (2023) for a comprehensive overview and discuss the most relevant works below.

Robust PCA Principal Component Analysis has been studied extensively in the outlier-robust setting using a variety of algorithmic approaches, such as robustly estimating the covariance matrix first and maximizing certain robust variance measures (see Croux & Haesbroeck (2000); Xu et al. (2010); Candès et al. (2011) and the references therein). Xu et al. (2013) gave the first efficient algorithm that overcomes prior work's challenges stemming from high-dimensions.

Robust Sparse Estimation In high-dimensional statistics, sample sizes that scale with the dimension d can quickly become overwhelming. However, a smaller sample size is possible under additional structural assumptions such as sparsity. In the context of mean estimation of distributions with light tails, the folklore sample size of d is replaced by k when the mean is known to be k-sparse (Hastie et al., 2015). Similar improvement is known for the robust version of the problem, where ϵ-fraction of the samples is corrupted (Balakrishnan et al., 2017; Diakonikolas et al., 2019; Cheng et al., 2019; Diakonikolas et al., 2022b). Focusing on sparse mean estimation, subsequent works have proposed further extensions such as Diakonikolas et al. (2022c) for heavy-tailed distributions, Diakonikolas et al. (2022b) for light-tailed distributions with unknown covariance matrix, Cheng et al. (2022) for non-convex first order methods (also see Zhu et al. (2022)), and Diakonikolas et al. (2022a); Zeng & Shen (2022) for list-decodable estimation. Liu et al. (2020) extended the work of Balakrishnan et al. (2017) to sparse linear regression.

PCA, has also been studied under sparsity. For the uncorrupted case, Berthet & Rigollet (2013b); Wang et al. (2016) provided optimal information-theoretic bounds as well as evidence through reductions to planted clique problem that efficient algorithms might require quadratically more samples. Croux et al. (2013) provided a sparse adaptation of earlier techniques for robust sparse PCA that showed improved performance in simulations. Balakrishnan et al. (2017) and Diakonikolas et al. (2019) studied theoretically the formulation of the problem as stated in this paper. Finally, guarantees have been developed for robust sparse linear regression too (see Chen et al. (2013) for early work on this problem). Balakrishnan et al. (2017) gave sample-efficient and poly-time algorithm for the task but with an error that scales with $\|\beta\|_2$, the norm of the unknown regressor. Liu et al. (2020) removed this dependence on $\|\beta\|_2$, resulting in nearly optimal error.

Huber Contamination The Huber contamination model of Definition 1.1 is the prototypical model under which the study of robust statistics was initiated. Since then, stronger models have been used, such as the "total variation model" where the samples come i.i.d. from a distribution that is $O(\epsilon)$-away from the original one in TV-distance, or the so called "strong contamination model", where a set of samples are drawn i.i.d. from the original distribution and then a computationally unbounded adversary is allowed to inspect them and edit arbitrarily ϵ-fraction of them, potentially breaking independence between samples. Information-theoretically, for all of these models, the optimal error for Gaussian k-sparse mean estimation is $\Theta(\epsilon)$ using $k\log(d)/\epsilon^2$ (see, e.g., Diakonikolas & Kane (2023)). However, the different models play a role when computational efficiency is considered.[9] Prior works on robust sparse mean estimation that obtain $O(\epsilon\sqrt{\log(1/\epsilon)})$ error (such as Balakrishnan et al. (2017); Diakonikolas et al. (2019)) succeed under the strong contamination model (and thus also Huber contamination model). However, there is evidence that with $\mathrm{poly}(k)$ samples, even in the total variation model, it is computationally hard to remove the $\sqrt{\log(1/\epsilon)}$ factor from the error. This evidence comes in the form of Statistical Query (SQ) lower bounds (Diakonikolas et al., 2017) (which transfers to the low-degree polynomials model due to the equivalence between them (Brennan et al., 2021)). Finally, we emphasize that we can not relax the Gaussianity assumption to generic sub-Gaussianity, since, even under univariate Huber contamination, the information-theoretic optimal error is $\epsilon\sqrt{\log(1/\epsilon)}$ for sub-Gaussian distributions.

[9] For other tasks, there even may be statistical differences: (Canonne et al., 2023) has shown that the sample complexity for these two models may be different (for testing problems).

Robust Sparse Estimation with Unknown Covariance One generalization of Definition 1.2 is to consider Gaussian k-sparse mean estimation when the covariance $\mathbf{\Sigma}$ is not necessarily identity and unknown to the algorithm. The information-theoretic limit for this case remains unchanged, apart from the fact that now the error naturally needs to scale with the size of the covariance, i.e., it becomes $O(\epsilon)\sqrt{\|\mathbf{\Sigma}\|_{\mathrm{op}}}$. However, we do not know of a polynomial time algorithm to achieve $O(\epsilon)\sqrt{\|\mathbf{\Sigma}\|_{\mathrm{op}}}$ error, while the currently best known polynomial-time algorithm (Diakonikolas et al., 2022b) achieves a larger error of $\epsilon\mathrm{polylog}(1/\epsilon)\sqrt{\|\mathbf{\Sigma}\|_{\mathrm{op}}}$ error (i.e., off by a $\mathrm{polylog}(1/\epsilon)$ factor) with $\mathrm{poly}(k/\epsilon)$ samples. Achieving the optimal $O(\epsilon)\sqrt{\|\mathbf{\Sigma}\|_{\mathrm{op}}}$ error in polynomial time is not even known in the dense setting: the current fastest algorithm runs in quasi-polynomial time (Diakonikolas et al., 2018).

B. Preliminaries

This section contains additional preliminaries and omitted facts and proofs.

Additional Notation If $U \subseteq [d] \times [d]$ is a set of pairs of indices such that for every $(i, j) \in U$, (j, i) is also in U, then for any matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ we denote by $(\mathbf{A})_U$ the matrix restricted to the entries from U. We use $x \lesssim y$ to denote that $x \leq Cy$ for some absolute constant C. We use the notation $a \gg b$ to mean that $a > Cb$ where C is some sufficiently large constant.

In the next few subsections, we state some well-known facts without proof, and some useful lemmata.

B.1. Miscellaneous Facts

Fact 2.3 (Variational definition)**.** $\|\mathbf{A}\|_{\mathrm{F},k,k} = \max_{\mathbf{B}}\langle \mathbf{A}, \mathbf{B}\rangle$ *where the maximum is taken over all matrices* $\mathbf{B}$ *with* $\|\mathbf{B}\|_{\mathrm{F}}{=}1$ *that have* k *non-zero rows, each of which has at most* k *non-zero elements.*

Moreover, a maximizer $\mathbf{B}$ *of this variational formulation can be found in* $\mathrm{poly}(d, k)$ *time given* $\mathbf{A}$.

Proof. Let $\mathbf{M}$ be the square matrix that is equal to 1 for each (i, j) for which $\mathbf{A}^2_{i,j}$ is witnessed in $\max_{S\subseteq[d]:|S|=k} \sum_{i\in S} \|A_i\|^2_{2,k}$, and 0 otherwise. Also, let $[\mathbf{A} \odot \mathbf{M}]_{i,j} := \mathbf{A}_{i,j}\mathbf{M}_{i,j}$.

Then,

$$\|\mathbf{A}\|_{\mathrm{F},k,k} = \sqrt{\max_{S\subseteq[d]:|S|=k} \sum_{i\in S} \|A_i\|^2_{2,k}} = \|\mathbf{A} \odot \mathbf{M}\|_{\mathrm{F}} = \max_{\mathbf{V},\|\mathbf{V}\|_{\mathrm{F}}=1} \sum_{i,j} \mathbf{A}_{i,j}\mathbf{M}_{i,j}\mathbf{V}_{i,j}.$$

Since $\mathbf{M}_{i,j}$ is non-zero only on k rows, and k elements in each of these rows, the expression above is equivalent to $\max_{\mathbf{B}}\langle \mathbf{A}, \mathbf{B}\rangle$ where the maximum is taken over all matrices $\mathbf{B}$ with $\|\mathbf{B}\|_{\mathrm{F}}{=}1$ that have k non-zero rows, each of which has at most k non-zero elements. Given $\mathbf{A}$, we can construct the mask $\mathbf{M}$, and setting $\mathbf{B} := (\mathbf{A} \cdot \mathbf{M})/\|\mathbf{A} \cdot \mathbf{M}\|_{\mathrm{F}}$ achieves the maximum value. □

Fact 2.4. $\|\mathbf{A}\|_{\mathrm{op},k} \leq \|\mathbf{A}\|_{\mathrm{F},k,k}$.

Proof. This is true because $vv^\top$ and $-vv^\top$ for any k-sparse unit vector v has Frobenius norm 1 and has at most k non-zero rows, each of which has at most k non-zero entries. □

Fact B.1 (Cover of the Sphere)**.** *Let* $r > 0$. *Let* $B_R = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$. *There exists a set* $\mathcal{C} \subseteq B_R$ *such that* $|\mathcal{C}| \leq (1 + 2R/\eta)^d$ *and for every* $v \in B_R$ *we have that* $\min_{y\in\mathcal{C}} \|y - v\|_2 \leq \eta$.

Fact B.2. *For any square matrix* $\mathbf{A}$ *and positive-semidefinite matrix* $\mathbf{B}$, $\mathrm{tr}(\mathbf{AB}) \leq \|\mathbf{A}\|_{\mathrm{op}}\mathrm{tr}(\mathbf{B})$.

Fact B.3 (see, for example, Diakonikolas & Kane (2023))**.** *For* $x, y \in \mathbb{R}^d$ *with* y *being a* k*-sparse vector, we have that* $\|t_k(x) - y\|_2 \leq \sqrt{6}\|x - y\|_{2,k}$, *where* $\|\cdot\|_{2,k}$ *denotes the sparse Euclidean norm (Definition 2.1) and* $t_k(x)$ *the operator that sets all but the* k *coordinates with the largest absolute values to zero.*

Fact B.4. *For unit vectors* $w, v \in \mathbb{R}^d$, *we have that* $\|ww^\top - vv^\top\|_{\mathrm{F}} = \Theta(\|w - v\|_2)$.

B.2. Probability Facts

Fact B.5 (Gaussian Norm Concentration)**.** *For every* $0 \le \beta \le \sigma\sqrt{d}$ *we have that*

$$\Pr_{X\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}[|\|X\|_2 - \sigma\sqrt{d}| > \beta] \le 2\exp\left(-\frac{\beta^2}{16\sigma^2}\right) .$$

Fact B.6. *For any* $d \times d$ *matrix* $\mathbf{A}$*, it holds* $\mathrm{Var}_{X\sim\mathcal{N}(0,\mathbf{I})}[X^\top \mathbf{A} X] = \|\mathbf{A}\|_{\mathrm{F}}^2 + \mathrm{tr}(\mathbf{A}^2)$.

Definition B.7 (Sub-Gaussian and Sub-gamma Random Variables)**.** A one-dimensional random variable Y is sub-Gaussian if $\|Y\|_{\psi_2} := \sup_{p\ge 1} p^{-1/2}\,\mathbf{E}\left[|Y|^p\right]$ is finite. We say that $\|Y\|_{\psi_2}$ is the sub-Gaussian norm of Y. A random vector X in $\mathbb{R}^d$ is sub-Gaussian if for every $v \in \mathcal{S}^{d-1}$, $\|v^\top X\|_{\psi_2}$ is finite. The sub-Gaussian norm of the vector is defined to be

$$\|X\|_{\psi_2} := \sup_{v\in\mathcal{S}^{d-1}} \|v^\top X\|_{\psi_2} .$$

We call a centered one-dimensional random variable Y a $(\nu,\alpha)_+$ sub-gamma if $\mathbf{E}[\exp(\lambda Y)] \le \nu^2\lambda^2/2$ for all $0 \le \lambda \le 1/\alpha$. We call $\|Y\|_{\psi_1} := \sup_{p\ge 1} p^{-1}\,\mathbf{E}[|Y|^p]$ the sub-gamma norm of Y.

Lemma B.8 (Properties of Sub-gamma Random Variables (Wainwright, 2019; Boucheron et al., 2013))**.** *The class of sub-gamma random variables satisfy the following:*

1. *(Wainwright, 2019, Proposition 2.9) If* Y *is a centered* $(\nu,\alpha)_+$ *sub-gamma random variable, then with probability* $1-\delta$*,* $Y \lesssim \nu\sqrt{\log(1/\delta)} + \alpha\log(1/\delta)$.
2. *(Boucheron et al., 2013, Theorem 2.3) If* Y *is a centered random variable satisfying that for all* $\delta \in (0,1)$*,* $Y \le \nu\sqrt{\log(1/\delta)} + \alpha\log(1/\delta)$*, then* Y *is* $(\nu',\alpha')_+$ *sub-gamma with* $\nu' \lesssim \nu+\alpha$ *and* $\alpha' \lesssim \alpha$.
3. *(Wainwright, 2019, Section 2.1.3) Let* $Y_1,\ldots,Y_k$ *be* k *centered independent* $(\nu,\alpha)_+$ *sub-gamma random variables. Then* $\sum_{i=1}^k Y_i$ *is a* $(\nu\sqrt{k},\alpha)_+$ *sub-gamma random variable.*

We also use the standard Hanson-Wright inequality (Vershynin, 2018):

Fact B.9 (Hanson-Wright Inequality)**.** *For* $X \sim \mathcal{N}(0,\mathbf{I})$ *in* $\mathbb{R}^d$ *and for every square* $d\times d$ *matrix* $\mathbf{A}$ *and scalar* $t \ge 0$*, the following holds:*

$$\Pr[|X^\top\mathbf{A}X - \mathbf{E}[X^\top\mathbf{A}X]| > t] \le 2\exp\left(-0.1\min\left(\frac{t^2}{\|\mathbf{A}\|_{\mathrm{F}}^2}, \frac{t}{\|\mathbf{A}\|_{\mathrm{op}}}\right)\right) .$$

The constant 0.1 above follows from Moshksar (2021).

B.3. Deterministic Conditions

The correctness of our algorithm will require the inliers to satisfy generic structural properties defined in Definition 2.5.

Recall that for a distribution D and a weight function w, we denote the weighted (and appropriately normalized) version of D using w by D_w. We further use μ_{D_w} to denote the mean of D_w. We restate the conditions and then state formally the lemma showing that Gaussian samples satisfy them with high probability.

Definition 2.5 ((ϵ,α,k)-goodness)**.** For $\epsilon \in (0,1/2)$, $\alpha > 0$ and $k \in \mathbb{N}$, we say that a distribution G on $\mathbb{R}^d$ is (ϵ,α,k)-good with respect to $\mu \in \mathbb{R}^d$, if the following are satisfied:

(1) For all $w : \mathbb{R}^d \to [0,1]$ with $\mathbf{E}_{X\sim G}[w(X)] \ge 1-\alpha$:

- (1.a) (Mean) $\|\mu_{G_w} - \mu\|_{2,k} \lesssim \alpha\sqrt{\log(1/\alpha)}$.
- (1.b) (Covariance) $\|\overline{\mathbf{\Sigma}}_{G_w} - \mathbf{I}\|_{\mathrm{op},k} \lesssim \alpha\log(\frac{1}{\alpha})$, where $\overline{\mathbf{\Sigma}}_{G_w} := \frac{1}{\mathbf{E}_{X\sim G}[w(X)]}\mathbf{E}_{X\sim G}[w(X)(X-\mu)(X-\mu)^\top]$.

(2) (Tails of *sparse* degree-2 polynomials) If $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a matrix with at most k^2 non-zero elements, $\|\mathbf{A}\|_{\mathrm{F}} \le \sqrt{\log(1/\epsilon)}$ and $\|\mathbf{A}\|_{\mathrm{op}} \le 1$, then the polynomial $p(x) := (x-\mu)^\top\mathbf{A}(x-\mu) - \mathrm{tr}(\mathbf{A})$ satisfies:

(a) $\mathbf{E}_{X\sim G}[p(X)\mathbb{1}(p(X) > 100\log(1/\epsilon))] \le \epsilon$.
(b) $\Pr_{X\sim G}[p(X) > 10\log(1/\epsilon)] \le \epsilon$.
(c) $\mathbf{E}_{X\sim G}[p(X)\mathbb{1}(h(x) > 100\log(1/\epsilon))] \le \epsilon$ for all $h(x)$ of the form $h(x) = \beta + v^\top(x-\mu)$ where $|\beta| \le 1$ and v is k-sparse and unit norm.

(3) $\Pr_{X\sim G}[|v^\top(X-\mu)| \ge 40\log(1/\epsilon))] \le \epsilon$, for all k-sparse unit norm vectors $v \in \mathbb{R}^d$.

The following lemma (proved in Appendix E) demonstrates that the uniform distribution over a sufficiently large set of samples drawn from $\mathcal{N}(\mu, \mathbf{I}_d)$ satisfies the deterministic conditions with respect to μ.

Lemma B.10 (Sample Complexity of Goodness Conditions)**.** *Let $\epsilon_0 > 0$ be a sufficiently small absolute constant. Let S be a set of n samples drawn i.i.d. from $\mathcal{N}(\mu, \mathbf{I}_d)$. Let G denote the uniform distribution on the points from S. If $\epsilon < \epsilon_0$, $k^2 \le d$ and $n \gg \frac{1}{\min\{\epsilon^2,\alpha^2\}}(k^2\log(d) + \log(1/\delta))\mathrm{polylog}(1/\epsilon)$, then with probability at least $1-\delta$, G is (ϵ, α, k)-good.*

B.4. Certificate Lemma

We restate and prove the following lemma.

Lemma 2.6 (Certificate Lemma)**.** *Let $0 < \alpha < \epsilon < 1/4$. Let $P = (1-\epsilon)G + \epsilon B$ be a mixture of distributions, where G satisfies Conditions (1.a) and (1.b) of Definition 2.5 with respect to $\mu \in \mathbb{R}^d$. Let $w : \mathbb{R}^d \to [0,1]$ be such that $\mathbf{E}_{X\sim G}[w(X)] > 1-\alpha$. If $\|\mathbf{\Sigma}_{P_w} - \mathbf{I}\|_{\mathrm{op},k} \le \lambda$, then*

$$\|\mu_{P_w} - \mu\|_{2,k} \lesssim \alpha\sqrt{\log\left(\frac{1}{\alpha}\right)} + \sqrt{\lambda\epsilon} + \epsilon + \sqrt{\alpha\epsilon\log\left(\frac{1}{\alpha}\right)}.$$

Proof. Let $\rho = \epsilon\,\mathbf{E}_{X\sim B}[w(X)]/\,\mathbf{E}_{X\sim P}[w(X)]$. Recall our notation $P_w(x) = w(x)P(x)/\,\mathbf{E}_{X\sim P}[w(X)], B_w(x) = w(x)B(x)/\,\mathbf{E}_{X\sim B}[w(X)], G_w(x) = w(x)G(x)/\,\mathbf{E}_{X\sim G}[w(X)]$ for the weighted versions of the distributions P, B, G and denote the corresponding covariance matrices by $\mathbf{\Sigma}_{P_w}, \mathbf{\Sigma}_{B_w}, \mathbf{\Sigma}_{G_w}$. We can write:

$$\mathbf{\Sigma}_{P_w} = \rho\mathbf{\Sigma}_{B_w} + (1-\rho)\mathbf{\Sigma}_{G_w} + \rho(1-\rho)(\mu_{G_w} - \mu_{B_w})(\mu_{G_w} - \mu_{B_w})^\top\ . \tag{3}$$

Let v be a k-sparse unit norm vector. Since $v^\top\mathbf{\Sigma}_{P_w}v \le 1+\lambda$, we obtain the following:

$$\begin{aligned} 1+\lambda \ge v^\top\mathbf{\Sigma}_{P_w}v &\ge (1-\rho)v^\top\mathbf{\Sigma}_{G_w}v + \rho(1-\rho)(v^\top(\mu_{B_w} - \mu_{G_w}))^2 \\ &\ge (1-\rho)\left(1-\alpha\log(1/\alpha)\right) + \rho(1-\rho)(v^\top(\mu_{B_w} - \mu_{G_w}))^2\ , \end{aligned}$$

where the second step uses Equation (3) and the last step uses the fact that G satisfies Condition (1.b) of Definition 2.5. The expression above implies the following:

$$(v^\top(\mu_{B_w} - \mu_{G_w}))^2 \le \frac{\lambda + \rho + \alpha\log(1/\alpha)}{\rho(1-\rho)}\ .$$

We can now bound the error $|v^\top(\mu_{P_w} - \mu)|$ as follows:

$$\begin{aligned} |v^\top(\mu_{P_w} - \mu)| &= |v^\top(\mu_{G_w} - \mu) + \rho v^\top(\mu_{B_w} - \mu_{G_w})| \\ &\le |v^\top(\mu_{G_w} - \mu)| + \rho|v^\top(\mu_{B_w} - \mu_{G_w})| \\ &\le \|\mu_{G_w} - \mu\|_{2,k} + \rho|v^\top(\mu_{B_w} - \mu_{G_w})| \\ &\le \alpha\sqrt{\log(1/\alpha)} + \sqrt{\rho}\sqrt{\frac{\lambda + \rho + \alpha\log(1/\alpha)}{1-\rho}}, \end{aligned}$$

where the last inequality uses that G satisfies Condition (1.a). We now use bounds on ρ to simplify the terms. Recall that $\rho = \epsilon\,\mathbf{E}_{X\sim B}[w(X)]/\,\mathbf{E}_{X\sim P}[w(X)] \le \epsilon/(1-\alpha)$. As $\alpha < 1/2$, we get that $\rho < 2\epsilon$. In addition, note that $\rho < 1/2$. Using these, we conclude that

$$\|\mu_{P_w} - \mu\|_{2,k} \le |v^\top(\mu_{P_w} - \mu)| \lesssim \alpha\sqrt{\log(1/\alpha)} + \sqrt{\lambda\epsilon} + \epsilon + \sqrt{\alpha\epsilon\log(1/\alpha)}\ .$$

□

B.5. Useful Procedures from Robust Statistics

The following result is implicit in Diakonikolas et al. (2019), which gives algorithm for k-sparse mean estimation with sub-optimal error of $O(\epsilon\sqrt{\log(1/\epsilon)})$. Since that algorithm relies on a certificate lemma similar to Lemma 2.6, it filters out points until the k-sparse norm of the empirical covariance minus identity becomes $O(\epsilon \mathrm{polylog}(1/\epsilon))$. Even though it does not manage to make the variance as small as would be required for the Certificate Lemma to yield $O(\epsilon)$ error, this is a useful starting point, and we will use it to pre-process the data in order to assume an $O(\epsilon \mathrm{polylog}(1/\epsilon))$ bound on the variance throughout our proofs.

Fact B.11. *Let $D \sim \mathcal{N}(\mu, \mathbf{I})$ be a Gaussian distribution on $\mathbb{R}^d$ with unknown mean μ and $\epsilon \leq \epsilon_0$ for some sufficiently small absolute constant $\epsilon_0 > 0$. Let T be a set of n samples from an ϵ-corrupted version of D, according to the Huber contamination model. If $n \gg (k^2 \log(d) + \log(1/\delta))\mathrm{polylog}(1/\epsilon)/\epsilon^2$, the algorithm from (Diakonikolas et al., 2019) run on input T, k, ϵ outputs a subset $T' \subseteq T$ of the original dataset such that the following hold with probability at least $1 - \delta$:*

1. *(Algorithm deletes $\log(1/\epsilon)$ more outliers than inliers) If S denotes the set of inliers in T, we have that $|(T \setminus T') \cap S| \leq \frac{1}{\log(1/\epsilon)}|(T \setminus T') \cap (T \setminus S)|$.*
2. *(Small (F, k, k)-norm) Denoting by $\mathbf{\Sigma}_{T'}$ the covariance matrix of the output set T', we have that $\|\mathbf{\Sigma}_{T'} - \mathbf{I}\|_{F,k,k} = O(\epsilon \log^2(1/\epsilon))$.*
3. *(Estimate of true mean) The empirical mean $\mu_{T'}$ of the output dataset satisfies $\|\mu_{T'} - \mu\|_{2,k} \lesssim \epsilon \log(1/\epsilon)$.*

Proof sketch. (Algorithm deletes $\log(1/\epsilon)$ more outliers than inliers): Algorithm 1 from (Diakonikolas et al., 2019) filters points while ensuring that for every inlier deleted, $\Omega(1)$ outliers are deleted. This is done by eliminating points according to the tail probabilities of the data. If the tail beyond some point T has a constant fraction more mass than it should, then a constant fraction of the points that are eliminated should be outliers. To boost the ratio of inliers to outliers being deleted, one just needs to adjust the threshold in lines 7 and 10 in Algorithm 1 of (Diakonikolas et al., 2019) so that the mass beyond T has to be more than $1 + \log(1/\epsilon)$ times the mass of the inliers. This will imply that the ratio of inliers to outliers deleted can be is boosted to $\log(1/\epsilon)$. The cost that we need to pay for this change is that the stopping condition in Line 4 of Algorithm 1 in Diakonikolas et al. (2019) changes from $\|(\mathbf{\Sigma}_{T'} - \mathbf{I})_U\|_{\mathrm{F}} \leq O(\epsilon \log(1/\epsilon))$ to $\|(\mathbf{\Sigma}_{T'} - \mathbf{I})_U\|_{\mathrm{F}} \leq O(\epsilon \log^2(1/\epsilon))$.

(Small (F,k, k)-norm of output covariance matrix): Running the (modified version of) Algorithm 1 in (Diakonikolas et al., 2019) with sparsity set to $2k$ (instead of k) ensures that upon termination we have that $\|(\mathbf{\Sigma}_{T'} - \mathbf{I})_U\|_{\mathrm{F}} = O(\epsilon \log^2(1/\epsilon))$ for $U \subset [d] \times [d]$ being the set of the $2k$ largest magnitude diagonal entries of $\mathbf{\Sigma}_{T'} - \mathbf{I}$ and the largest magnitude $4k^2 - 2k$ off diagonal entries, with ties broken so that if $(i, j) \in U$ then $(j, i) \in U$. By using the definition of (F, k, k)-norm, this implies that:

$$\|\mathbf{\Sigma}_{T'} - \mathbf{I}\|_{F,k,k} \leq \|(\mathbf{\Sigma}_{T'} - \mathbf{I})_U\|_{\mathrm{F}} \leq O(\epsilon \log^2(1/\epsilon)) \ , \tag{4}$$

(Estimate of the mean): Since the algorithm deleted $\log(1/\epsilon)$ factor more outliers than inliers and initially we have ϵn outliers, it must be the case that the inliers after filtering are at least $(1 - \epsilon/\log(1/\epsilon))n$. Also note that $\sup_{v \in \mathbb{R}^d : \|v\|_2 = 1, \|v\|_0 = k} v^\top (\mathbf{\Sigma}_{T'} - \mathbf{I})v = O(\epsilon \log^2(1/\epsilon))$, which comes from (4) and Fact 2.4. This allows us to use Lemma 2.6, which implies that $\|\mu_{T'} - \mu\|_{2,k} \lesssim \epsilon \log(1/\epsilon)$ (for this application we also used that the inliers in the dataset satisfy the goodness conditions of Definition 2.5 with $\alpha = \epsilon/\log(1/\epsilon)$)).

□

C. Robust Sparse Mean Estimation Under Huber Contamination

In this section we provide the complete proof of correctness, an overview of which was given in Section 3.

Recall Definition 3.1. We record a useful fact about the above definition.

Fact C.1. *Consider the matrix $\mathbf{A} = \sum_{i=1}^r \mathbf{A}_i$ where $\mathbf{A}_i$ for $i \in [r]$ are the matrices in Definition 3.1. Then, it is true that (i) $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{r}$, and (ii) $\|\mathbf{A}\|_{\mathrm{op}} \leq 1$.*

Proof. Proof of (i): Since $\mathbf{A} = \sum_{i=1}^r \mathbf{A}_i$, the $\mathbf{A}_i$'s have their non-zero entries on disjoint coordinates, and $\|\mathbf{A}_i\|_{\mathrm{F}} = 1$ we have that $\|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_{i=1}^r \|\mathbf{A}_i\|_{\mathrm{F}}^2 = r$.

Proof of (ii): Observe that for $i \neq j$, the matrices $\mathbf{A}_i$ and $\mathbf{A}_j$ satisfy that if (k, ℓ) entry of $\mathbf{A}_i$ is non-zero, then the corresponding entry of $\mathbf{A}_j$ must be zero. As a result, by relabeling coordinates, we see that $\mathbf{A}$ can be written as a block matrix:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{O} & \dots & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{A}_2 & \dots & \mathbf{O} & \mathbf{O} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{O} & \mathbf{O} & \dots & \mathbf{A}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \dots & \mathbf{O} & \mathbf{O}, \end{bmatrix}$$

where $\mathbf{O}$ represent a matrix with all entries set to zero; observe that $\mathbf{O}$ above could be of varing dimensions. Therefore, $\mathbf{A}^\top \mathbf{A}$ is equal to

$$\mathbf{A}^\top \mathbf{A} = \begin{bmatrix} \mathbf{A}_1^\top \mathbf{A}_1 & \mathbf{O} & \dots & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{A}_2^\top \mathbf{A}_2 & \dots & \mathbf{O} & \mathbf{O} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{O} & \mathbf{O} & \dots & \mathbf{A}_r^\top \mathbf{A}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \dots & \mathbf{O} & \mathbf{O}, \end{bmatrix}$$

Since this is a block diagonal PSD matrix, the operator norm of $\mathbf{A}^\top \mathbf{A}$ is equal to $\max_i \|\mathbf{A}_i^\top \mathbf{A}_i\|_{\mathrm{op}}$, which we can upper bound as $\|\mathbf{A}_i^\top \mathbf{A}_i\|_{\mathrm{op}} \leq \mathrm{tr}(\mathbf{A}_i^\top \mathbf{A}_i) = \|\mathbf{A}_i\|_{\mathrm{F}}^2 = 1$. Therefore, $\|\mathbf{A}\|_{\mathrm{op}} \leq \sqrt{\|A^\top \mathbf{A}\|_{\mathrm{op}}} \leq 1$. □

We now provide the full proof of Theorem 1.5, where we state each of the three steps as separate claims and prove them individually.

Proof of Theorem 1.5. We start with some notation. Let T be the dataset after preprocessing (line 4). We let P be the uniform distribution over T, G be the uniform distribution over the inliers of T and B the uniform distribution over the outliers. We can write P as mixture $P = (1-\epsilon)G + \epsilon B$. If $w : T \to [0, 1]$ denotes the weights that the algorithm maintains at a given point during its execution, we denote by P_w the weighted by w version of P. By Lemma B.10, we have that with probability at least $1 - \delta$, G is (ϵ, α, k)-good with $\alpha = 3\epsilon/\log(1/\epsilon)$.

We will show Theorem 1.5 via the following steps listed as individual claims below:

Claim C.2. *Consider the setting and notation of Theorem 1.5 and Algorithm 2. The condition* $\mathbf{E}_{X\sim G}[w(X)] \geq 1 - 3\epsilon/\log(1/\epsilon)$ *remains true throughout the execution of the loop in line 7.*

Claim C.3. *Under the setting of Theorem 1.5, the loop of line 7 terminates after* $\tilde{O}(d/\epsilon)$ *time.*

Claim C.4. *Consider the setting and notation of Theorem 1.5 and Algorithm 2. After the loop of line 7 ends, let* w *be the resulting weight function, let* $\mu_w = \mathbf{E}_{X\sim P_w}[X]$ *be the mean of the dataset weighted by* w*, and denote by* H *the set of coordinates as in line 12 of the pseudocode. Then, for every* k*-sparse* $v \in \mathbb{R}^d$*, it holds* $|v^\top(\mu_w - \mu)_{[d]\setminus H}| = O(\epsilon)\|v\|_2$.

We will prove these claims at the end of the current proof. We first use these to show that the algorithm has error $O(\epsilon)$ overall: For every k-sparse vector v of $\mathbb{R}^d$, let v_1 denote the copy of v that has all of its entries in the coordinates from $[d] \setminus H$ zeroed out and v_2 the copy of v that has all of the entries in the coordinates from H zeroed out. Similarly, decompose μ into $\mu_1 + \mu_2$. Then:

$$\left|v^\top(\hat{\mu} - \mu)\right|^2 = \left|v^\top(\hat{\mu}_1 - \mu_1 + \hat{\mu}_2 - \mu_2)\right|^2 \leq 2\left|v_1^\top(\hat{\mu}_1 - \mu_1)\right|^2 + 2\left|v_2^\top(\hat{\mu}_2 - \mu_2)\right|^2 ,$$

where we used the "almost triangle inequality" $(a+b)^2 \leq 2a^2 + 2b^2$. To conclude the proof, we claim that $\left|v_1^\top(\hat{\mu}_1 - \mu_1)\right|^2 \leq \|v_1\|_2^2 O(\epsilon^2)$ and $\left|v_2^\top(\hat{\mu}_2 - \mu_2)\right|^2 \leq \|v_2\|_2^2 O(\epsilon^2)$, which, once established, can be used to conclude that $\left|v^\top(\hat{\mu} - \mu)\right|^2 \leq O(\epsilon^2)(\|v_1\|_2^2 + \|v_2\|_2^2) = O(\epsilon^2)\|v\|_2^2$, i.e., the final error is $O(\epsilon)$.

The claim that $\left|v_1^\top(\hat{\mu}_1 - \mu_1)\right| \le \|v_1\|_2 O(\epsilon)$ follows by the $O(\epsilon)$-error guarantee of the dense robust mean estimator run on a dataset restricted to the coordinates in H. Let $d' := |H|$ be the dimensionality of the restricted dataset. Fact 2.7 states that in order for that estimator to achieve $O(\epsilon)$ error the dataset used needs to be of size a sufficiently large multiple of $\frac{1}{\epsilon^2}(d' + \log(1/\delta))\mathrm{polylog}(d/\epsilon)$. Since $d' = |H| \le rk = \log(1/\epsilon)k$, this quantity is $O(\frac{1}{\epsilon^2}(k + \log(1/\delta)))\mathrm{polylog}(d/\epsilon)$ and thus smaller than the sample complexity mentioned in the statement of Theorem 1.5.

The claim that $\left|v_2^\top(\hat{\mu}_2 - \mu_2)\right| \le \|v_2\|_2 O(\epsilon)$ follows by Claim C.4.

We now prove Claims C.2 to C.4.

Proof of Claim C.2. To show this, it suffices to show that every time the weight function is updated, $\log(1/\epsilon)$ more mass is removed from outliers than inliers. We will show this inductively: Assume it is true for all previous rounds and we will show it for the current round using Lemma 2.8 applied with $\beta = \log(1/\epsilon)$, $s = \epsilon$ once we show that the lemma is applicable. Denote $\lambda' := \|\mathbf{\Sigma}_w - \mathbf{I}\|_{\mathrm{F},k,k}$. In order to show that Lemma 2.8 is applicable, we need to check that $\mathbf{E}_{X\sim G}[w(X)\tilde{\tau}(X)] \le \epsilon$. Regarding that, $\tilde{\tau}(X)$ looks like the thresholded polynomial $\tau(x)$ used in the deterministic Condition (2) for which we know that $\mathbf{E}_{X\sim G}[w(X)\tau(X)] \le \epsilon$. The difference is that $\tilde{\tau}(x)$ centers the point around μ_w instead of μ that is used in $\tau(x)$, thus we need some triangle inequalities and Claim 3.2, which is shown in Appendix C.1, to prove that this difference is not substantial.

Let $\tilde{\tau}(x) = \tilde{p}(x)\mathbb{1}(\tilde{p}(x) > 200\log(1/\epsilon))$, where $\tilde{p}(x) = (x-\mu_w)^\top \mathbf{A}(x-\mu_w) - \mathrm{tr}(\mathbf{A})$ be the scores that the algorithm uses (which center points around μ_w) and $\tau(x) = p(x)\mathbb{1}(p(x) > 100\log(1/\epsilon))$ with $p(x) = (x-\mu)^\top \mathbf{A}(x-\mu) - \mathrm{tr}(\mathbf{A})$ be the ideal scores appearing in the deterministic condition (that center things around the true μ). Denote by $\Delta p(x) := \tilde{p}(x) - p(x) = (\mu-\mu_w)^\top \mathbf{A}(\mu-\mu_w) + (x-\mu)^\top \mathbf{A}(\mu-\mu_w) + (x-\mu)^\top \mathbf{A}^\top(\mu-\mu_w)$ the difference of the two polynomials. We have that

$$
\begin{aligned}
\mathop{\mathbf{E}}_{X\sim G}[w(X)\tilde{\tau}(X)] &\le \mathop{\mathbf{E}}_{X\sim G}[\tilde{\tau}(X)] \\
&= \mathop{\mathbf{E}}_{X\sim G}[\tilde{p}(X)\mathbb{1}(\tilde{p}(X) > 200\log(1/\epsilon))] \\
&= \mathop{\mathbf{E}}_{X\sim G}[(\Delta p(X) + p(X))\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \\
&= \mathop{\mathbf{E}}_{X\sim G}[\Delta p(X)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \\
&\quad + \mathop{\mathbf{E}}_{X\sim G}[p(X)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \\
&\le |(\mu-\mu_w)^\top \mathbf{A}(\mu-\mu_w)| \\
&\quad + \mathop{\mathbf{E}}_{X\sim G}[(X-\mu)^\top \mathbf{A}(\mu-\mu_w)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \\
&\quad + \mathop{\mathbf{E}}_{X\sim G}[(X-\mu)^\top \mathbf{A}^\top(\mu-\mu_w)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \\
&\quad + \mathop{\mathbf{E}}_{X\sim G}[p(X)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \,. \qquad (5)
\end{aligned}
$$

We claim that each of the four terms above is at most $O(\epsilon)$. Denote $\lambda := \max_{v\in\mathbb{R}^d:\|v\|_2=1,\|v\|_0=k} v^\top(\mathbf{\Sigma}_w - \mathbf{I})v$—not to be confused with $\lambda' := \|\mathbf{\Sigma}_w - \mathbf{I}\|_{\mathrm{F},k,k}$. For the first term in (5), we have that

$$
\begin{aligned}
|(\mu-\mu_w)^\top \mathbf{A}(\mu-\mu_w)| &\le r\|\mu-\mu_w\|_{2,k}^2 && \text{(using Claim 3.2)} \\
&\lesssim r(\epsilon\lambda + \epsilon^2) && \text{(using Lemma 2.6)} \\
&\le r\epsilon\lambda' + r\epsilon^2 && (\lambda \le \lambda' \text{ by Fact 2.4}) \\
&\le r\epsilon^2\log^2(1/\epsilon) && (\text{using } \lambda' \lesssim \epsilon\log^2(1/\epsilon)) \\
&\le \epsilon, && (6)
\end{aligned}
$$

applicable as follows: in the second line we applied Lemma 2.6 with $\alpha = 3\epsilon/\log(1/\epsilon)$ (the requirement that $\mathbf{E}_{X\sim G}[w(X)] \ge 1-\alpha$ of that lemma is satisfied by inductive hypothesis), and the last line uses that $r := \log(1/\epsilon)$.

We now move to the second term in (5). The bound for the third term is identical, thus we will omit it. We have that

$$
\mathop{\mathbf{E}}_{X\sim G}[(X-\mu)^\top \mathbf{A}(\mu-\mu_w)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))]
$$

$$
\begin{aligned}
&\leq \mathop{\mathbf{E}}_{X\sim G}[(X-\mu)^\top \mathbf{A}(\mu-\mu_w)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X), \Delta p(X) \leq 100\log(1/\epsilon))] \\
&\quad + \mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top \mathbf{A}(\mu-\mu_w)|\mathbb{1}(\Delta p(X) > 100\log(1/\epsilon))] \\
&\leq \mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top \mathbf{A}(\mu-\mu_w)|\mathbb{1}(p(X) > 100\log(1/\epsilon))] \\
&\quad + \mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top \mathbf{A}(\mu-\mu_w)|\mathbb{1}(\Delta p(X) > 100\log(1/\epsilon))] \,. \qquad (7)
\end{aligned}
$$

We now work with the two terms individually. We start with the first term of (7). In what follows we define the vectors $u_i := \mathbf{A}_i(\mu-\mu_w)/\|\mathbf{A}_i(\mu-\mu_w)\|_{2,k}$ to shorten notation. We have the following series of inequalities (see below for step by step explanations)

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top \mathbf{A}(\mu-\mu_w)|\mathbb{1}(p(X) > 100\log(1/\epsilon))] \\
&\leq \sum_{i=1}^{r} \mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top \mathbf{A}_i(\mu-\mu_w)|\mathbb{1}(p(X) > 100\log(1/\epsilon))] && (8)\\
&\lesssim \sum_{i=1}^{r} \sqrt{\mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top \mathbf{A}_i(\mu-\mu_w)|^2]}\sqrt{\Pr_{X\sim G}[p(X) > 100\log(1/\epsilon)]} && (9)\\
&= \sum_{i=1}^{r} \|\mathbf{A}_i(\mu-\mu_w)\|_{2,k}\sqrt{\mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top u_i|^2]}\sqrt{\Pr_{X\sim G}[p(X) > 100\log(1/\epsilon)]} && (10)\\
&\leq \sum_{i=1}^{r} \|\mathbf{A}_i\|_{\mathrm{F}}\|\mu-\mu_w\|_{2,k}\sqrt{\mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top u_i|^2]}\sqrt{\Pr_{X\sim G}[p(X) > 100\log(1/\epsilon)]} && (11)\\
&\lesssim r(\sqrt{\lambda'\epsilon}+\epsilon)\sqrt{\Pr_{X\sim G}[p(X) > 100\log(1/\epsilon)]} && (12)\\
&\lesssim r(\sqrt{\lambda'\epsilon}+\epsilon)\sqrt{\epsilon} && (13)\\
&\lesssim \epsilon^{1.5}\log^2(1/\epsilon) && (14)\\
&\leq \epsilon \,. && (15)
\end{aligned}
$$

(8) follows from the definition of $A = \sum_{i=1}^{r} \mathbf{A}_i$ and the triangle inequality. (9) uses the Cauchy–Schwarz inequality. (10) is a re-writing using $u_i := \mathbf{A}_i(\mu-\mu_w)/\|\mathbf{A}_i(\mu-\mu_w)\|_{2,k}$, where the point to note is that u_i is k-sparse (because $\mathbf{A}_i$ has only k non-zero rows) and unit norm. (11) uses the inequality $\|\mathbf{C}v\|_{2,k} \leq \|\mathbf{C}\|_{\mathrm{F}}\|v\|_{2,k}$. This is true since $\mathbf{C}$ is a matrix with at most k nonzero rows with at most k nonzero entries in each row and has Frobenius norm 1. An application of Claim 3.2 with $r = 1$ then gives us what we want. Then, (12) uses that $\|\mathbf{A}_i\|_{\mathrm{F}} \leq 1$, $\|\mu-\mu_w\|_{2,k} \lesssim \sqrt{\epsilon\lambda}+\epsilon \leq \sqrt{\epsilon\lambda'}+\epsilon$ by Lemma 2.6 and Fact 2.4, and $\mathbf{E}_{X\sim G}[|(x-\mu)^\top u_i|^2] \leq 1+\tilde{O}(\epsilon) \lesssim 1$ by the deterministic Condition (1.b) (these utilize the fact that u_i is unit-norm k-sparse vector). (13) uses that $\Pr_{X\sim G}[p(X) > 100\log(1/\epsilon)] \leq \epsilon$ by the deterministic Item (2)b. (14) uses that $\lambda' \lesssim \epsilon\log^2(1/\epsilon)$ by the preprocessing step of the algorithm (cf. Fact B.11) and also that $r = \log(1/\epsilon)$.

We now move to the second term of (7).

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim G}[|(X-\mu)^\top \mathbf{A}(\mu-\mu_w)|\mathbb{1}(\Delta p(X) > 100\log(1/\epsilon))] \\
&\leq r(\sqrt{\lambda'\epsilon}+\epsilon)\sqrt{\Pr_{X\sim G}[\Delta p(X) > 100\log(1/\epsilon)]} && \text{(similar to steps (8-12))}\\
&\lesssim \epsilon^{1.5}\log^2(1/\epsilon) \lesssim \epsilon \,, && (16)
\end{aligned}
$$

where we bounded $\Pr_{X\sim G}[\Delta p(X) > 100\log(1/\epsilon)]$ as follows: First, $\Pr_{X\sim G}[\Delta p(x) > 100\log(1/\epsilon)] \leq \Pr_{X\sim G}[|(X-\mu)^\top \mathbf{A}(\mu-\mu_w)| > 99\log(1/\epsilon)] + \Pr_{X\sim G}[|(X-\mu)^\top \mathbf{A}^\top(\mu-\mu_w)| > 99\log(1/\epsilon)]$, where this uses that $\Delta p(x) = (\mu-\mu_w)^\top \mathbf{A}(\mu-\mu_w) + (x-\mu)^\top \mathbf{A}^\top(\mu-\mu_w) + (x-\mu)^\top \mathbf{A}(\mu-\mu_w)$, and that $|(\mu-\mu_w)^\top \mathbf{A}(\mu-\mu_w)| \leq 1 < \log(1/\epsilon)$ by (6). We will only focus on the first term, since the second probability is bounded identically:

$$
\Pr_{X\sim G}[|(X-\mu)^\top \mathbf{A}(\mu-\mu_w)| > 99\log(1/\epsilon)] \leq \Pr_{X\sim G}\left[\left|(X-\mu)^\top \frac{\mathbf{A}(\mu-\mu_w)}{\|\mathbf{A}(\mu-\mu_w)\|_2}\right| > \frac{99\log(1/\epsilon)}{\|\mathbf{A}(\mu-\mu_w)\|_2}\right] \qquad (17)
$$

$$\leq \Pr_{X\sim G}\left[\left|(X-\mu)^\top \frac{\mathbf{A}(\mu-\mu_w)}{\|\mathbf{A}(\mu-\mu_w)\|_2}\right| > 99\log(1/\epsilon)\right] \tag{18}$$

$$\leq \epsilon \tag{19}$$

where (17) divides by $\mathbf{A}(\mu-\mu_w)$ both sides (18) uses that $\|\mathbf{A}(\mu-\mu_w)\|_2 \leq \sum_{i=1}^r \|\mathbf{A}_i(\mu-\mu_w)\|_2 = \sum_{i=1}^r \|\mathbf{A}_i(\mu-\mu_w)\|_{2,k} \leq \sum_{i=1}^r \|\mathbf{A}_i\|_{\mathrm{F}}\|\mu-\mu_w\|_{2,k} \leq r\|\mu-\mu_w\|_{2,k} \lesssim r(\sqrt{\lambda'\epsilon}+\epsilon) \leq 1$, where the first step is a triangle inequality, the second step uses the fact that $\mathbf{A}_i$ has only k-nonzero rows, the next step uses Cauchy-Schwartz, then we use that there are r terms in the sum, $\|\mathbf{A}_i\|_{\mathrm{F}} \leq 1$, and finally that $\|\mu-\mu_w\|_{2,k} \lesssim \sqrt{\epsilon\lambda'}+\epsilon$ by Lemma 2.6 and finally that $r=\log(1/\epsilon), \lambda' = O(1)$ by our preprocessing step inside the algorithm (Fact B.11). (19) uses Item (3) of Definition 2.5, which is indeed applicable because $\mathbf{A}(\mu-\mu_w)$ is rk-sparse since $\mathbf{A}$ has at most rk non-zero rows.

We are now done with all the terms in (7) and can now move to the last term of (5):

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim G}[p(X)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X))] \\
&\leq \mathop{\mathbf{E}}_{X\sim G}[p(X)\mathbb{1}(p(X) > 200\log(1/\epsilon) - \Delta p(X), \Delta p(X) < 100\log(1/\epsilon))] \\
&\quad + \mathop{\mathbf{E}}_{X\sim G}[p(X)\mathbb{1}(\Delta p(X) > 100\log(1/\epsilon))]
\end{aligned}$$

Now, the first term above is $\mathbf{E}_{X\sim G}[p(X)\mathbb{1}(p(X) > 100\log(1/\epsilon))] = \mathbf{E}_{X\sim G}[\tau(x)]$ which is at most $O(\epsilon)$ by our deterministic Condition (2). The $\mathbf{E}_{X\sim G}[p(X)\mathbb{1}(\Delta p(X) > 100\log(1/\epsilon))]$ is bounded by ϵ using Item (2)c of the deterministic condition Condition (2) (the application is very similar to the application of Item (3) of Definition 2.5 in equations (17)-(19)). □

Proof of Claim C.3. We will show that in every iteration of the while loop of line 7, $\mathbf{E}_{X\sim P}[w(X)]$ is reduced by at least $\tilde{\Omega}(\epsilon/d)$, where $w(x)$ are the weights that the algorithm maintains and P is the uniform distribution over the input dataset. Since initially $\mathbf{E}_{X\sim P}[w(X)] = 1$, this would mean that after at most $\tilde{O}(d/\epsilon)$ iterations, the weight from all the outliers would have been reduced to zero. We will finally show that this would trigger the stopping condition of line 7 and cause the algorithm to terminate.

We start with the first claim, that $\mathbf{E}_{X\sim P}[w(X)]$ gets non-trivially reduced in every round. Fix an iteration of the algorithm and denote by $w(x)$ the weights at the start of that round and by $w'(x)$ the weights after that round ends. The mass removed is, by design of the filtering algorithm (see line 6 of Algorithm 1)

$$\mathop{\mathbf{E}}_{X\sim P}[w(X) - w'(X)] = \frac{\mathbf{E}_{X\sim P}[w(X)\tilde{\tau}(X)]}{\max_{x:w(x)>0}\tilde{\tau}(X)} . \tag{20}$$

We will show that the denominator is $O(d\log^2(d/\epsilon))$, and that the numerator is $\Omega(\epsilon\log(1/\epsilon))$.

Regarding the denominator, let μ_T denote the vector from line 5 of the algorithm (the vector used in the naïve pruning step). For every point with $w(x) > 0$ it holds $\tilde{\tau}(x) \leq |(x-\mu_w)^\top \mathbf{A}(x-\mu_w)| \leq \|\mathbf{A}\|_{\mathrm{op}}\|x-\mu_w\|_2^2 \leq \|x-\mu_w\|_2^2 \lesssim \|x-\mu\|_2^2 + \|\mu-\mu_T\|_2^2 + \|\mu_T-\mu_w\|_2^2 \lesssim d\log^2(d/\epsilon)$, where the fact that every point in the dataset is within $10\sqrt{d}\log(d/\epsilon)$ from μ_T (by the pruning done in line 5 of the algorithm).

Regarding the numerator, let $\lambda'' := g_r(\mathbf{\Sigma}_w - \mathbf{I})$. We will show that the numerator is $\Omega(\lambda'')$. Note that as long as the while loop of line 7 has not been terminated, $\lambda'' \gg \log(1/\epsilon)\epsilon$ by design. We rewrite the numerator:

$$\mathop{\mathbf{E}}_{X\sim P}[w(X)\tilde{\tau}(X)] = \mathop{\mathbf{E}}_{X\sim P}[w(X)\tilde{p}(X)] - \mathop{\mathbf{E}}_{X\sim P}[w(X)\tilde{p}(X)\mathbb{1}(\tilde{p}(X) \leq 200\log(1/\epsilon))] \tag{21}$$

We first show that, after re-normalizing, the first term at least $0.5g_r(\mathbf{\Sigma}_w - \mathbf{I})$:

$$\begin{aligned}
\mathop{\mathbf{E}}_{X\sim P}[w(X)\tilde{p}(X)] &= \mathop{\mathbf{E}}_{X\sim P}[w(X)] \mathop{\mathbf{E}}_{X\sim P_w}[\tilde{p}(X)] \\
&= \mathop{\mathbf{E}}_{X\sim P}[w(X)] \mathop{\mathbf{E}}_{X\sim P_w}[\langle \mathbf{A}, (X-\mu_w)(X-\mu_w)^\top - \mathbf{I}\rangle] \\
&= \mathop{\mathbf{E}}_{X\sim P}[w(X)] \left\langle \mathbf{A}, \mathop{\mathbf{E}}_{X\sim P_w}[(X-\mu_w)(X-\mu_w)^\top] - \mathbf{I}\right\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \mathop{\mathbf{E}}_{X\sim P}[w(X)]\langle \mathbf{A}, \mathbf{\Sigma}_w - \mathbf{I}\rangle \\
&= \mathop{\mathbf{E}}_{X\sim P}[w(X)] \sum_{i=1}^{r} \langle \mathbf{A}_i, \mathbf{\Sigma}_w - \mathbf{I}\rangle && \text{(see Definition 3.1)} \\
&= \mathop{\mathbf{E}}_{X\sim P}[w(X)] \sum_{i=1}^{r} h_r(\mathbf{\Sigma}_w - \mathbf{I}) && \text{(see Definition 3.1)} \\
&= \mathop{\mathbf{E}}_{X\sim P}[w(X)] g_r(\mathbf{\Sigma}_w - \mathbf{I}) && (22) \\
&\geq 0.5\lambda'' ,
\end{aligned}
$$

where we used Claim C.2 to obtain that $\mathbf{E}_{X\sim P}[w(X)] \geq 1/2$.

In the reminder, we show that the second term in Equation (21) is at most $0.002\lambda''$. First, we decompose into inliers and outliers:

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim P}[w(X)\tilde{p}(X)\mathbb{1}(\tilde{p}(X) \leq 200\log(1/\epsilon))] \\
&\quad = (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)\tilde{p}(X)\mathbb{1}(\tilde{p}(X) \leq 200\log(1/\epsilon))] \\
&\quad + \epsilon \mathop{\mathbf{E}}_{X\sim B}[w(X)\tilde{p}(X)\mathbb{1}(\tilde{p}(X) \leq 200\log(1/\epsilon))]
\end{aligned}
$$

We again treat each term individually. For the second term, (the one due to outliers), we have the following due to the indicator function

$$
\epsilon \mathop{\mathbf{E}}_{X\sim B}[w(X)\tilde{p}(X)\mathbb{1}(\tilde{p}(X) \leq 200\log(1/\epsilon))] \leq \epsilon(200\log(1/\epsilon))) \ll \lambda'' ,
$$

where we used that by design of line 7 of the algorithm $\lambda'' > Cr\epsilon$ with $r := \log(1/\epsilon)$ and C being a sufficiently large constant.

For the term due to inliers, we can say the following: Denote by $\Delta p(x) = \tilde{p}(x) - p(x) = (\mu - \mu_w)^\top \mathbf{A}(\mu - \mu_w) + (x - \mu)^\top \mathbf{A}(\mu - \mu_w) + (x - \mu)^\top \mathbf{A}^\top(\mu - \mu_w)$. Then, we will establish the following bounds (see below for explanations of each step):

$$
\begin{aligned}
&(1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)\tilde{p}(X)\mathbb{1}(\tilde{p}(X) \leq 200\log(1/\epsilon))] && (23) \\
&\leq (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)\tilde{p}(X)] && (24) \\
&= (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)p(X)] + (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)\Delta p(X)] && (25) \\
&\leq \epsilon + (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)\Delta p(X)] && (26) \\
&\leq \epsilon + (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)(\mu - \mu_w)^\top \mathbf{A}(\mu - \mu_w)] \\
&\quad + (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)(x-\mu)^\top \mathbf{A}(\mu - \mu_w)] + (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w(X)(x-\mu)^\top \mathbf{A}^\top(\mu - \mu_w)] && (27) \\
&\leq O(\epsilon) + (\mu - \mu_w)^\top \mathbf{A}(\mu - \mu_w) + (\mu_G - \mu)^\top \mathbf{A}(\mu - \mu_w) + (\mu_G - \mu)^\top \mathbf{A}^\top(\mu - \mu_w) && (28) \\
&= O(\epsilon) && (29) \\
&\leq 0.002\lambda'' && (30) \\
& && (31)
\end{aligned}
$$

We explain the steps below: (24) uses that the indicator only removes non-negative terms. (25) decomposes into inliers and outliers. (26) bounds the average value of the polynomial over inliers as follows (we will use the notation $p_{\mathbf{A}}(x) = (x-\mu)^\top \mathbf{A}(x-\mu) - \mathrm{tr}(\mathbf{A})$ for clarity):

$$
\begin{aligned}
\mathop{\mathbf{E}}_{X\sim G}[w(X)p_{\mathbf{A}}(X)] &= \mathop{\mathbf{E}}_{X\sim G}[w(X)p_{\mathbf{A}}(X)] \\
&= \mathop{\mathbf{E}}_{X\sim G}[p_{\mathbf{A}}(X)] - \mathop{\mathbf{E}}_{X\sim G}[(1-w(X))p_{\mathbf{A}}(X)]
\end{aligned}
$$

$$= \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[p_{\mathbf{A}}(X)] + O(\epsilon) - \mathop{\mathbf{E}}_{X\sim G}[(1-w(X))p_{\mathbf{A}}(X)]$$
$$= O(\epsilon) + \mathop{\mathbf{E}}_{X\sim G}[(1-w(X))p_{-\mathbf{A}}(X)] \tag{32}$$

where in the second to last line we used that for all degree-2 polynomials g with at most k^2 terms $\left|\mathbf{E}_{X\sim G}[g(X)] - \mathbf{E}_{Z\sim\mathcal{N}(\mu,\mathbf{I})}[g(Z)]\right| \le \epsilon\sqrt{\mathrm{Var}_{Z\sim\mathcal{N}(\mu,\mathbf{I})}[g(Z)]}$, which can be found in which can be found in Lemma 4.3 in (Diakonikolas et al., 2019). (32) line we renamed $-\mathbf{A}$ to $\mathbf{A}'$. Then, decomposing the remaining term into the large and small part using indicator functions, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim G}[(1-w(X))p_{-\mathbf{A}}(X)] &= \mathop{\mathbf{E}}_{X\sim G}[(1-w(X))p_{-\mathbf{A}}(X)\mathbb{1}(p_{-\mathbf{A}}(X)\le 100\log(1/\epsilon))]\\ &+ \mathop{\mathbf{E}}_{X\sim G}[(1-w(X))p_{-\mathbf{A}}(X)\mathbb{1}(p_{-\mathbf{A}}(X)\ge 100\log(1/\epsilon))]\\ &\lesssim \mathop{\mathbf{E}}_{X\sim G}[(1-w(X))100\log(1/\epsilon)] + \epsilon\\ &\lesssim \epsilon\end{aligned}$$

where in the second to last step we used that $\mathbf{E}_{X\sim G}[1-w(X)] \lesssim \epsilon/\log(1/\epsilon)$ by Claim C.2, and also that $\mathbf{E}_{X\sim G}[(1-w(X))p_{-\mathbf{A}}(X)\mathbb{1}(p_{-\mathbf{A}}(X)\ge 100\log(1/\epsilon))] \le \epsilon$ by the deterministic condition (2).

Back to explaining the sequence of bounds in (23)-(30), the bound used in (29) can be proved exactly as in (6) and (30) uses that $\lambda'' > Cr\epsilon$ because of line 7 of the algorithm.

We have thus completed the argument that $\tilde{\Omega}(\epsilon/d)$ mass is removed in every round of the loop of line 7. To conclude the proof of Claim C.3, it remains to show that after $\tilde{O}(d/\epsilon)$ iterations the algorithm will necessarily terminate. First note that after that many iterations, there cannot be any outliers left. Moreover, by Claim C.2 a large fraction of inliers still remains in the dataset (i.e., $\mathbf{E}_{X\sim G}[w(X)] \ge 1 - 3\epsilon/\log(1/\epsilon)$). We claim that under this setting, $g_r(\mathbf{\Sigma}_w - \mathbf{I}) = \mathbf{E}_{X\sim G}[w(X)\tilde{p}(X)] \lesssim \epsilon\log(1/\epsilon)$ causing the stopping condition of line 7 to trigger. We show the bound as follows:

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim G}[w(X)\tilde{p}(X)] &\le O(\epsilon) + (\mu-\mu_w)^\top\mathbf{A}(\mu-\mu_w) + (\mu_{G_w}-\mu)^\top\mathbf{A}(\mu-\mu_w) + (\mu_{G_w}-\mu)^\top\mathbf{A}^\top(\mu-\mu_w)\\ &\lesssim \epsilon + r\|\mu-\mu_w\|_{2,k}^2 + r\|\mu-\mu_w\|_{2,k}\|\mu_{G_w}-\mu_w\|_{2,k}\end{aligned}$$

where the first inequality is as in (24)-(28). Finally to bound each term by $O(\epsilon)$ one can use the deterministic Condition (1.a) and (6).

□

Proof of Claim C.4. The claim follows by Lemma 2.6. We show how to apply this lemma: By the definition of the stopping condition of the algorithm, upon termination we have that $g_r(\mathbf{\Sigma}_w - \mathbf{I}) \le O(r\epsilon)$, or, equivalently $\frac{1}{r}\sum_{i=1}^r h_i(\mathbf{\Sigma}_w - \mathbf{I}) \le O(\epsilon)$. This means that h_r, as the smallest term in the sum, must be $O(\epsilon)$. Since $h_r(\mathbf{\Sigma}_w - \mathbf{I})$ is the (F, k, k) norm of the matrix after deletion of the elements in $([d]\setminus H)\times([d]\setminus H)$ (cf. line 12 of pseudocode), it follows that $\|(\mathbf{\Sigma}_w - \mathbf{I})_{([d]\setminus H)\times([d]\setminus H)}\|_{\mathrm{F},k,k} \le h_r(\mathbf{\Sigma}_w - \mathbf{I}) \le O(\epsilon)$. Combining with Fact 2.4, we have that $\sup_{\|v\|_2=1,\|v\|_0=k} v^\top((\mathbf{\Sigma}_w - \mathbf{I})_{([d]\setminus H)\times([d]\setminus H)})v = O(\epsilon)$. We also recall that the dataset is (ϵ,α,k)-stable with $\alpha = 3\epsilon/\log(1/\epsilon)$ and, because of Claim C.2, $\mathbf{E}_{X\sim G}[w(X)] \ge 1-\alpha$ holds when exiting the main loop of the algorithm. All of these mean that we can apply Lemma 2.6 with $\lambda = O(\epsilon)$ and $\alpha = 3\epsilon/\log(1/\epsilon)$ to obtain that $\left|v_2^\top(\hat{\mu}_2 - \mu_2)\right| \le \|v_2\|_2^2 O(\epsilon)$ for any vector k-sparse vector v_2 supported on $[d]\setminus H$. □

□

C.1. Proof of Claim 3.2

We restate and prove the following inequality

Claim 3.2. *Let $\mathbf{A} = \sum_{\ell\in[r]}\mathbf{B}^{(\ell)}$ where each $\mathbf{B}^{(\ell)}$ is a square matrix with Frobenius norm equal to one, k non-zero rows, each of which has k non-zero entries. Then, for any vectors u, v, it holds $|u^\top\mathbf{A}v| \le r\|u\|_{2,k}\|v\|_{2,k}$.*

Proof. First, $u^\top \mathbf{A} v = \sum_\ell u^\top \mathbf{B}^{(\ell)} v$. Consider a single matrix $\mathbf{B}^{(\ell)}$ from the sum and denote by $b_i^{(\ell)}$ for $i \in [d]$ the rows of $\mathbf{B}^{(\ell)}$ (only k of them are non-zero and each has at most k non-zero elements). We have the following:

$$\begin{aligned} u^\top \mathbf{B}^{(\ell)} v &= u^\top \mathbf{B}^{(\ell)} v \\ &= \sum_{i,j} u_i v_j (\mathbf{B}^{(\ell)})_{ij} \\ &= \sum_i u_i \sum_j v_j (\mathbf{B}^{(\ell)})_{ij} \\ &\leq \sum_i u_i \|v\|_{2,k} \|b_i^{(\ell)}\|_2 && (33) \\ &\leq \|u\|_{2,k} \|v\|_{2,k} \|\mathbf{B}^{(\ell)}\|_{\mathrm{F}} && (34) \\ &= \|u\|_{2,k} \|v\|_{2,k} && (35) \end{aligned}$$

where the step in (33) used Cauchy–Schwarz inequality along with the fact that at most k-entries are non zero in the i-th row of $\mathbf{B}^{(\ell)}$, and (34) used Cauchy–Schwarz again along with the fact that there are at most k non-zero rows in $\mathbf{B}^{(\ell)}$.

Finally, summing over all terms in (35) concludes the proof. □

D. Robust Principal Component Analysis

We state and prove a more detailed version of Claim 4.1 below:

Claim D.1. *Let $X \sim \mathcal{N}(0, \mathbf{I} + \rho v v^\top)$ be a random variable from the spiked covariance model and $Z = \mathrm{Proj}_{w^\top}(X)$ the projection of X onto the subspace perpendicular to w. For $\alpha \in \mathbb{R}$, let G_α denote the distribution of Z conditioned on $w^\top X = \alpha$, and for an interval $I \subset \mathbb{R}$, let G_I denote the distribution of Z conditioned on $w^\top X \in I$. We also use the notation $\phi(z; \mu, \mathbf{\Sigma})$ for the pdf of $\mathcal{N}(\mu, \mathbf{\Sigma})$. Then, the pdfs of the two aforementioned distributions are:*

1. *$G_\alpha(z) = \phi(z; \tilde{\mu}, \tilde{\mathbf{\Sigma}})$ with $\tilde{\mu} = \frac{\rho(w^\top v)\alpha}{1+\rho(w^\top v)^2} \bar{v}$ and $\tilde{\mathbf{\Sigma}} = \mathbf{I} + \frac{\rho}{1+\rho(w^\top v)^2} \bar{v}\bar{v}^\top$, where $\bar{v} := (v - (w^\top v) w)$.*
2. *$G_I(z) = \frac{1}{\Pr_{X \sim \mathcal{N}(0, \mathbf{I}+\rho v v^\top)}[w^\top X \in I]} \int_{\alpha' \in I} G_{\alpha'}(z) \Pr_{X \sim \mathcal{N}(0, I + \rho v v^\top)}[w^\top X = \alpha'] \mathrm{d}\alpha'$.*

Proof. Let $e_1, \ldots, e_d$ denote the standard orthonormal basis of $\mathbb{R}^d$. By a rotation of the space, we can assume without loss of generality that w is aligned with e_d, i.e., $w = \|w\|_2 e_d$. We denote by $\overline{v} = (v_1, \ldots, v_{d-1})$ the projection of v to the subspace orthogonal to w and by $v_d = w^\top v$ the projection of v in the direction of w. This way the subspace perpendicular to w is the one spanned by the first $d-1$ basis elements. The first part of the claim follows from Fact D.2 below applied with $y_1 = (x_1, \ldots, x_{d-1}), y_2 = \alpha, \mu_1 = 0, \mu_2 = 0, \mathbf{\Sigma}_{11} = \mathbf{I} + \rho \bar{v}\bar{v}^\top, \mathbf{\Sigma}_{12} = \rho v_d \bar{v}, \mathbf{\Sigma}_{21} = \rho v_d \bar{v}^\top$ and $\mathbf{\Sigma}_{22} = 1 + \rho v_d^2$ (and the fact that $v_d = w^\top v$ and $\overline{v} = v - (w^\top v) w$). The second claim follows by the law of total probability.

Fact D.2. *If $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \right)$, then $y_1 | y_2 \sim \mathcal{N}(\bar{\mu}, \bar{\mathbf{\Sigma}})$, with $\bar{\mu} = \mu_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(y_2 - \mu_2)$ and $\bar{\mathbf{\Sigma}} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$.*

□

Claim D.3. *Under the assumptions of Theorem 1.6 and assuming that the estimators in lines 1,2 and 3 exist, we have that $\|\hat{v} - v\|_2 = O(\epsilon/\rho)$.*

Proof. Let $e_1, \ldots, e_d$ denote the standard orthonormal basis of $\mathbb{R}^d$. By a rotation of the space, we can assume without loss of generality that w is aligned with e_d, i.e., $w = \|w\|_2 e_d$. We denote by $\overline{v} = (v_1, \ldots, v_{d-1})$ the projection of v to the subspace orthogonal to w and by $v_d = w^\top v$ the projection of v in the direction of w.

Some useful observations for later on are the following: Note that, by Fact B.4, the fact that we start with $\|ww^\top - vv^\top\|_{\mathrm{F}} = O(\epsilon\sqrt{\log(1/\epsilon)}/\rho)$ in the algorithm implies that $\|w - v\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)}/\rho)$, which also means that $1 \geq w^\top v \geq 1 - O(\epsilon^2 \log(1/\epsilon)/\rho^2)$. Since $w^\top v = v_d$ (by our rotation assumption) and $\|v\|_2 = 1$, the previous discussion means that:

$$\|\bar{v}\|_2 = \sqrt{1 - (w^\top v)^2} = O(\epsilon\sqrt{\log(1/\epsilon)}/\rho) \text{ and } 1 - O(\epsilon^2 \log(1/\epsilon)/\rho^2) \leq v_d \leq 1\,. \tag{36}$$

Now, let $x \sim \mathcal{N}(0, \mathbf{I} + \rho v v^\top)$ be a random vector coming from our spiked covariance model. We want to consider the distribution of $(x_1, \ldots, x_{d-1})$ conditioned on $w^\top x = \alpha$ (note that by our rotation assumption this is equivalent to conditioning on $x_d = \alpha$). By Claim 4.1, this conditional distribution is $\mathcal{N}\left(\frac{\rho v_d \alpha}{1+\rho v_d^2}\overline{v}, \mathbf{I} + \frac{\rho}{1+\rho v_d^2}\overline{v}\overline{v}^\top\right)$.

The error of our final estimator is

$$\|\hat{v} - v\|_2 \leq \left\| z\frac{1+\rho y}{\rho\sqrt{y}\alpha} - \bar{v}\right\|_2 + |\sqrt{y} - v_d| \,.$$

The second term is $O(\epsilon/\rho)$ by line 2 of the pseudocode. To see this, note that $|y - v_d^2| = O(\epsilon/\rho)$ implies $|\sqrt{y} - v_d| = \frac{|y - v_d^2|}{\sqrt{y}+v_d} = O(\epsilon/\rho)$ by using $v_d \geq 1/2$. For the first term, we have the following:

$$\begin{aligned}
\left\| z\frac{1+\rho y}{\rho\sqrt{y}\alpha} - \bar{v}\right\|_2 &\leq \left\| z\frac{1+\rho y}{\rho\sqrt{y}\alpha} - z\frac{1+\rho v_d^2}{\rho v_d \alpha}\right\|_2 + \left\| z\frac{1+\rho v_d^2}{\rho v_d \alpha} - \bar{v}\right\|_2 \\
&\leq \|z\|_2 \left|\frac{1+\rho y}{\rho\sqrt{y}\alpha} - \frac{1+\rho v_d^2}{\rho v_d \alpha}\right| + \frac{1+\rho v_d^2}{\rho v_d \alpha}\left\| z - \bar{v}\frac{\rho\alpha v_d}{1+\rho v_d^2}\right\|_2 && (37) \\
&\lesssim \rho\left|\frac{1+\rho y}{\rho\sqrt{y}\alpha} - \frac{1+\rho v_d^2}{\rho v_d \alpha}\right| + O(\epsilon/\rho) \,, && (38)
\end{aligned}$$

where the first line is a triangle inequality, and the last line used the following: First, the factor $(1+\rho v_d^2)/(\rho v_d \alpha)$ is $O(1/\rho)$ because of $\alpha = \Theta(1)$, $\rho = O(1)$, and $1 \geq v_d \geq 1/3$ (by (36)). Also, by the mean estimation guarantee (line 3 of the pseudocode), we have that $\|z - \bar{v}\frac{\rho\alpha v_d}{1+\rho v_d^2}\|_2 = O(\epsilon)$, which bounds the last term in (37). Lastly, the previous two imply that $\|z\|_2 \leq O(\rho)$:

$$\begin{aligned}
\|z\|_2 &\leq \left\| z - \bar{v}\frac{\rho\alpha v_d}{1+\rho v_d^2}\right\|_2 + \left\|\bar{v}\frac{\rho\alpha v_d}{1+\rho v_d^2}\right\|_2 \\
&\lesssim \epsilon + \rho\|\bar{v}\|_2 && \text{(mean estimation guarantee, } \alpha = \Theta(1), v_d \leq 1) \\
&\lesssim \epsilon + \rho O(\epsilon\sqrt{\log(1/\epsilon)}/\rho) && \text{(by (36))} \\
&\lesssim \epsilon\sqrt{\log(1/\epsilon)} \lesssim \rho \,. && \text{(by assumption)}
\end{aligned}$$

This is because z is $O(\epsilon)$-close to $\bar{v}\frac{\rho\alpha v_d}{1+\rho v_d^2}$, whose norm can be checked to be $O(\epsilon\sqrt{\log(1/\epsilon)})$ using that $\alpha = \Theta(1), v_d = \Theta(1)$ and $\|\bar{v}\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)}/\rho)$ (by (36)) and finally $\epsilon\sqrt{\log(1/\epsilon)} \lesssim \rho$ by assumption.

We now bound the remaining term in (38). We know that $|y - v_d^2| = O(\epsilon/\rho)$, thus we also have $|\sqrt{y} - v_d| = O(\epsilon/\rho)$. Let us write $\sqrt{y} = v_d + \eta$, $y = v_d^2 + \eta'$ for some $|\eta| = O(\epsilon/\rho)$, $|\eta'| = O(\epsilon/\rho)$. Using this and doing some algebra, the term in (38) is

$$\begin{aligned}
\rho\left|\frac{1+\rho y}{\rho\sqrt{y}\alpha} - \frac{1+\rho v_d^2}{\rho v_d \alpha}\right| &= \frac{\rho}{\rho\alpha}\left|\frac{1+\rho v_d^2 + \rho\eta'}{v_d + \eta} - \frac{1+\rho v_d^2}{v_d}\right| \\
&\leq \left|\frac{\rho\eta' v_d - \eta - \rho v_d^2 \eta}{v_d(v_d+\eta)}\right| && (\alpha = \Theta(1)) \\
&\lesssim \epsilon/\rho \,,
\end{aligned}$$

where in the last step we used $\rho = O(1)$, $|v_d| = \Theta(1)$ (from (36)) and $|\eta| = O(\epsilon/\rho)$, $|\eta'| = O(\epsilon/\rho)$ to show that every term is $O(\epsilon/\rho)$.

□

Claim D.4. *There exists a computationally efficient estimator that uses $O(1/\epsilon)$ ϵ-corrupted samples and achieves the guarantee in line 2 of Algorithm 3. There exists a computationally efficient estimator that uses $O((k^2\log(d) + \mathrm{polylog}(1/\epsilon))/\epsilon^2)$ ϵ-corrupted samples from $\mathcal{N}(0, \mathbf{I} + \rho v v^\top)$ and achieves the guarantee in line 3 of Algorithm 3.*

Proof sketch. The first estimator exists since it is known that robustly estimating the variance of a Gaussian in one dimension can be done with $O(\epsilon)$ error (see, e.g., Diakonikolas et al. (2018) which is for high-dimensions but here we only need

it for one dimension). Concretely, consider $(w^\top x)$ for $x \sim \mathcal{N}(0, \mathbf{I} + \rho vv^\top)$. Then $(w^\top x) \sim \mathcal{N}(0, 1 + \rho(w^\top v)^2)$. By robustly estimating its variance, we can obtain y' such that $|y' - (1 + \rho(w^\top v)^2)| = O(\epsilon(1+\rho)) = O(\epsilon)$; here we used that $\rho = O(1)$ by assumption. Then $y := (y' - 1)/\rho$ satisfies $|y - (w^\top v)^2| = O(\epsilon/\rho)$, which is the guarantee mentioned in line 2 of Algorithm 3.

The estimator for line 3 of Algorithm 3 is the following: Let T be a set of ϵ-corrupted samples from $\mathcal{N}(0, \mathbf{I} + \rho vv^\top)$ in the Huber contamination model.

1. Draw α uniformly from $[-(1+\rho), 1+\rho] \setminus [-0.1, 0.1]$.[10]
2. Define the interval $I = [\alpha - \ell, \alpha + \ell]$ for $\ell = 1/\log(1/\epsilon)$.
3. $T' = \{x \in T : w^\top x \in I\}$.
4. $T'' = \{\mathrm{Proj}_{w^\perp}(x) : x \in T'\}$ (Project samples to subspace orthogonal to w).
5. Let z be the output of Algorithm 2 run on T''.
6. Output the vector obtained from z after zeroing out all coordinates except the $2k$ ones with the largest absolute value.

The correctness of Algorithm 2 relies on the following two: (i) the fraction of outliers in T' continues to be $O(\epsilon)$ and (i.e., we are still in the Huber contamination model with approximately the same corruption level) (ii) the inliers satisfy the deterministic conditions from Definition 2.5.

We start with (i), where we sketch the proof. We argue that the probability of an outlier x having $w^\top x \in I$ divided by the probability that an inlier having $w^\top x \in I$ is at most $O(1)$. This would ensure that the fraction of outliers does not blow up by more than a constant factor. We start with upper bounding the probability for outliers. Since I is chosen randomly and independently of anything else we can imagine that the outlier x is fixed and then I is chosen randomly. Let us only examine the case where $w^\top x \in [-2(1+\rho), 2(1+\rho)]$ (because otherwise $w^\top x \notin I$). The probability that $w^\top x \in I$ is then the ratio of the length of I to the length of the interval $[-2(1+\rho), 2(1+\rho)]$, which is $O(\ell/(1+\rho))$. We now argue about the inliers. For this, we can imagine that I is fixed and the inlier x is drawn randomly from the inlier distribution. We note that $w^\top x \sim \mathcal{N}(0, \tilde{\sigma}^2)$ with $\tilde{\sigma}^2 := 1 + \rho(w^\top v)^2$. Since $I \subseteq [-3\tilde{\sigma}^2, 3\tilde{\sigma}^2]$ the Gaussian distribution behaves approximately uniformly there and thus the probability that $w^\top X \in I$ is $\Omega(\ell/\tilde{\sigma}^2)$, which is also $\Omega(\ell/(1+\rho))$.

We next show (ii). The inliers follow the distribution G_I from Claim 4.1. We want to show that G_I satisfies the deterministic conditions of Definition 2.5 with respect to μ_{G_α}, the mean of the distribution $X \sim \mathcal{N}(0, \mathbf{I} + \rho vv^\top)$ conditioned on $w^\top X = \alpha$, where α is the center of the interval I. We sketch the argument for why each of the conditions holds. Condition (1.a) and Condition (1.b) rely only on three properties: (i) $\|\mu_{G_I} - \mu_\alpha\|_2 = O(\epsilon)$, (ii) $\|\mathbf{I} - \mathbf{\Sigma}_{G_I}\|_{\mathrm{op}} = O(\epsilon)$ and (iii) $O(1)$-subgaussianity. We can check that these hold. For the first one, we have that

$$\|\mu_{G_{\alpha'}} - \mu_{G_\alpha}\|_2 \lesssim \ell \frac{\rho(w^\top v)}{1 + \rho(w^\top v)^2} \|v - w(w^\top v)\|_2 \lesssim \ell\rho(\epsilon\sqrt{\log(1/\epsilon)}/\rho) \lesssim \epsilon\,, \qquad \text{(using } \ell = 1/\log(1/\epsilon))$$

where the last line uses that $\alpha = O(1)$, $(w^\top v)^2 \le 1$ and $\|v - w(w^\top v)\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)}/\rho)$ by the guarantee of the estimator in line 1. We move to the covariance property. Let α_1 and α_2 denote the bounds of the interval I, i.e., $I = [\alpha_1, \alpha_2]$. We have that $\mathbf{E}_{X\sim G_{\alpha'}}[XX^\top] = \mathbf{I} + \frac{\rho}{1+\rho(w^\top v)^2}\bar{v}\bar{v}^\top + \frac{\rho^2(\alpha')^2}{(1+\rho(w^\top v)^2)}\bar{v}\bar{v}^\top$ where $\bar{v} := v - (w^\top v)w$. Thus, as a convex combination, $\mathbf{E}_{X\sim G_I}[XX^\top] = \mathbf{I} + \frac{\rho}{1+\rho(w^\top v)^2}\bar{v}\bar{v}^\top + \frac{\xi\rho^2(\alpha_1)^2 + (1-\xi)\rho^2(\alpha_2)^2}{(1+\rho(w^\top v)^2)}\bar{v}\bar{v}^\top$ for some $\xi \in [0,1]$.

$$\begin{aligned}
\|\mathbf{\Sigma}_{G_I} - \mathbf{I}\|_{\mathrm{op}} &\le \left\| \frac{\rho}{1+\rho(w^\top v)^2}\bar{v}\bar{v}^\top + \frac{\xi\rho^2(\alpha_1)^2 + (1-\xi)\rho^2(\alpha_2)^2}{(1+\rho(w^\top v)^2)}\bar{v}\bar{v}^\top \right\| \\
&\lesssim \rho\|\bar{v}\|_2^2 + \rho^2 \max(\alpha_1^2, \alpha_2^2)\|\bar{v}\|_2^2 \\
&\lesssim \rho(\epsilon\sqrt{\log(1/\epsilon)}/\rho)^2 + \rho^2(\epsilon\sqrt{\log(1/\epsilon)}/\rho)^2 \\
&\lesssim \epsilon^2\log(1/\epsilon)/\rho + (\epsilon\sqrt{\log(1/\epsilon)})^2 \lesssim \epsilon\,. \qquad \text{(using } \rho \ge \epsilon\log(1/\epsilon))
\end{aligned}$$

[10] We draw α from $[-(1+\rho), 1+\rho] \setminus [-0.1, 0.1]$ because we need $|\alpha| = \Omega(1)$ in Claim D.3.

For the subgaussianity, we have that

$$\mathop{\mathbf{E}}_{X\sim G_I}[|v^\top(X-\mu_{G_I})|^p]^{1/p} \le \max_{\alpha'\in I} \mathop{\mathbf{E}}_{X\sim G_{\alpha'}}[|v^\top(X-\mu_{G_I})|^p]^{1/p} \le \max_{\alpha'\in I} \mathop{\mathbf{E}}_{X\sim G_{\alpha'}}[|v^\top(X-\mu_{G_{\alpha'}})|^p]^{1/p} + \|\mu_{G_{\alpha'}} - G_I\|_2$$
$$\lesssim \sqrt{p} + \ell\frac{\rho\alpha(w^\top v)}{1+\rho(w^\top v)^2}\|v - w(w^\top v)\|_2 \lesssim \sqrt{p} + \ell O(\epsilon\sqrt{\log(1/\epsilon)}) = \sqrt{p} + O(1) \ .$$

We now move to the remaining deterministic conditions about thresholded polynomials. Let us use the notation $p_a(x) = (x-\mu_\alpha)^\top \mathbf{A}(x-\mu_\alpha) - \mathrm{tr}(\mathbf{A})$. Our goal is to show that $\mathbf{E}_{X\sim S}[p_\alpha(x)\mathbb{1}(p_\alpha(x) > 100\log(1/\epsilon))] \le \epsilon$, where S is a set of n i.i.d. samples from G_I. As it can be seen by the proof Lemma B.10, the key element for proving that condition is the concentration in (40). Thus it suffices to ensure that it holds for samples from G_I. Since G_I is a mixture of the $G_{\alpha'}$ distributions, it suffices to show that the concentration holds for $G_{\alpha'}$ for all $\alpha' \in I$. Let $\Delta p(x) = p_\alpha(x) - p_{\alpha'}(x) = (\mu_{\alpha'} - \mu_\alpha)^\top \mathbf{A}(\mu_{\alpha'} - \mu_\alpha) + 2(x-\mu_{\alpha'})^\top \mathbf{A}(\mu_{\alpha'} - \mu_\alpha)$ be the difference of the two polynomials. By considering the cases where $|\Delta p(x)| \le 50\log(1/\epsilon))$ and $|\Delta p(x)| > 50\log(1/\epsilon))$, we have that

$$\Pr_{X\sim G_{\alpha'}}[|p_\alpha(X)| \ge 100\log(1/\epsilon)] \le \Pr_{X\sim G_{\alpha'}}[|p_{\alpha'}(X)| \ge 50\log(1/\epsilon)] + \Pr_{X\sim G_{\alpha'}}[|\Delta p(X)| > 50\log(1/\epsilon)] \ . \tag{39}$$

The first term is bounded using Item (2)a of Lemma B.10 (although the lemma has been proved for Gaussians with identity covariance, it can be checked that it goes through for $X \sim G_{\alpha'}$ which is Gaussian with covariance $I + O(\epsilon)$). The second is bounded using Gaussian concentration ($\Delta p(X)$ is a linear polynomial).

The previous discussion means that our mean estimator (Algorithm 2) is applicable and satisfies the same guarantee as in Theorem 1.5, i.e., yields z with $\|z - \mu_{G_I}\|_{2,k} = O(\epsilon)$ (which since $\|\mu_{G_I} - \mu_\alpha\|_2 = O(\epsilon)$ from earlier also means that $\|z - \mu_\alpha\|_{2,k} = O(\epsilon)$). However the guarantee in line 3 of Algorithm 3 that we are trying to prove uses ℓ_2-norm instead of the $(2,k)$-norm. We explain below how one can get from the one norm bound to the other: We start by focusing on the "warm start" estimate w in line 1 of Algorithm 3. First, $\|w-v\|_{2,k} \le \|w-v\|_2 = \Theta(\|ww^\top - vv^\top\|_{\mathrm{F}}) = O(\epsilon\sqrt{\log(1/\epsilon)}/\rho)$ (where first step is Fact B.4, and the second step is because $ww^\top - vv^\top$ is rank-2 matrix). Then, since v is k-sparse, we can assume without loss of generality by Fact B.3 that w is also k-sparse. Using Claim D.1, this means that μ_{G_I} (the mean that we are trying to robustly estimate) is $2k$-sparse, because every μ_{G_α} is a scaled version of $v - (w^\top v)w$ and both v, w are k-sparse. The mean estimator satisfies $\|z - \mu_{G_I}\|_{2,k} = O(\epsilon)$, and by Claim D.1, we can keep the largest $2k$-coordinates in z to obtain a z' with $\|z' - \mu_{G_I}\|_2 = O(\epsilon)$.

Finally, regarding the last part of Theorem 1.6, we have that

$$\frac{\hat{v}\mathbf{\Sigma}\hat{v}}{\|\mathbf{\Sigma}\|_{\mathrm{op}}} = \frac{1+\rho(v^\top\hat{v})^2}{1+\rho} \ge \frac{1+\rho(1-O(\epsilon^2/\rho^2))}{1+\rho} \ge 1 - O\left(\frac{\epsilon^2}{\rho(1+\rho))}\right) \ .$$

□

E. Omitted Proofs from Appendix B.3

We restate and prove the following:

Lemma B.10 (Sample Complexity of Goodness Conditions)**.** *Let $\epsilon_0 > 0$ be a sufficiently small absolute constant. Let S be a set of n samples drawn i.i.d. from $\mathcal{N}(\mu, \mathbf{I}_d)$. Let G denote the uniform distribution on the points from S. If $\epsilon < \epsilon_0$, $k^2 \le d$ and $n \gg \frac{1}{\min\{\epsilon^2,\alpha^2\}}(k^2\log(d) + \log(1/\delta))\mathrm{polylog}(1/\epsilon)$, then with probability at least $1-\delta$, G is (ϵ,α,k)-good.*

Proof. The proof of Conditions (1.a) and (1.b) can be found in prior work (see, e.g., Li (2018)). Item (3) uses that for all k-sparse unit vectors v, and $T \ge 6$, $\Pr_{X\sim G}[|v^\top(X-\mu)| \ge T] \le 3\mathrm{erfc}\left(\frac{T}{\sqrt{2}}\right) + \frac{\epsilon^2}{T^2}$, which can be found in Lemma 4.3 in Diakonikolas et al. (2019).

We prove the remaining conditions below.

Proof of Condition (2): We will use the notation $g_{\mathbf{A}}(x) = (x-\mu)^\top\mathbf{A}(x-\mu)$ (i.e., we do not include the centering $\mathrm{tr}(\mathbf{A})$ in the polynomial and also we write the matrix A in the subscript for extra clarity) and $\tau_{\mathbf{A}}(x) = (g_{\mathbf{A}}(x) - \mathrm{tr}(\mathbf{A}))\mathbb{1}(g_{\mathbf{A}}(x) - \mathrm{tr}(\mathbf{A}) > 100\log(1/\epsilon))$. The proof will consist of the following steps (the last two steps involve a cover argument):

(1) First, we show that $\mathbf{E}_{X\sim\mathcal{N}(\mu,I)}[\tau_{\mathbf{A}}(X)] \lesssim \epsilon^4$ for every $\mathbf{A}$ of the form mentioned in Condition (2).

(2) We then show that $\tau_{\mathbf{A}}(X) - \mathbf{E}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A}}(X)]$ for $X \sim \mathcal{N}(\mu, \mathbf{I})$ is a sub-gamma random variable for every $\mathbf{A}$ of the form mentioned in Condition (2).

(3) Then, we show that for any fixed $\mathbf{A}$ of that form, if $X_1, \ldots, X_n$ are i.i.d. samples from $\mathcal{N}(\mu, \mathbf{I})$, with probability $1-\delta$, we have $\frac{1}{n}\sum_{i=1}^n \tau(X_i) - \mathbf{E}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau(X)] \lesssim \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + \frac{1}{n}\|\mathbf{A}\|_{\mathrm{op}}\log(1/\delta)$.

(4) Finally, with probability $1-\delta'$, $\mathbf{E}_{X\sim S}[\tau_{\mathbf{A}}(X)] \lesssim \epsilon^4 + \sqrt{\frac{\log(1/\epsilon)(k^2\log(d)+\log(1/\delta'))}{n}} + \frac{k^2\log(d/\epsilon)+\log(1/\delta')}{n}$ holds *simultaneously* for all matrices $\mathbf{A}$ of the form mentioned in Condition (2).

We now prove the claims above.

Proof of Item (1): Using the Hanson-Wright inequality (Fact B.9), we have that

$$\Pr_{X\sim\mathcal{N}(\mu,\mathbf{I})}[|g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A})| > t] \le 2\exp\left(-0.1\min\left(\frac{t^2}{\|\mathbf{A}\|_{\mathrm{F}}^2}, \frac{t}{\|\mathbf{A}\|_{\mathrm{op}}}\right)\right) . \tag{40}$$

Setting $t = 100\log(1/\epsilon)$, $\|\mathbf{A}\|_{\mathrm{F}} \le \sqrt{\log(1/\epsilon)}$, and $\|\mathbf{A}\|_{\mathrm{op}} \le 1$, the above becomes $\Pr_{X\sim\mathcal{N}(\mu,\mathbf{I})}[g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}) > 100\log(1/\epsilon)] \lesssim \epsilon^{10}$. This allows us to upper bound $\mathbf{E}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A}}(X)]$ by $O(\epsilon^4)$ as follows,

$$\begin{aligned}
\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A}}(X)] &= \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[(g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}))\mathbb{1}(g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}) > t)] \\
&\le \sqrt{\Pr_{X\sim\mathcal{N}(\mu,\mathbf{I})}[g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}) > t] \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[(g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}))^2]} \qquad \text{(by Cauchy-Schwarz)} \\
&= \sqrt{\Pr_{X\sim\mathcal{N}(\mu,\mathbf{I})}[g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}) > t]}\sqrt{\mathrm{Var}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[g_{\mathbf{A}}(X)]} \\
&\lesssim \epsilon^5\sqrt{\|\mathbf{A}\|_{\mathrm{F}} + \mathrm{tr}(A^2)} \le \epsilon^5\sqrt{\|\mathbf{A}\|_{\mathrm{F}} + \|\mathbf{A}\|_{\mathrm{F}}^2} \lesssim \epsilon^5\log(1/\epsilon) \lesssim \epsilon^4 ,
\end{aligned} \tag{41}$$

where the final inequality uses Fact B.6 and then it uses the fact that $\mathrm{tr}(\mathbf{A}^2) \le \|\mathbf{A}\|_{\mathrm{F}}^2$, which can be seen as follows: Let A_i denote the rows of $\mathbf{A}$ and $\tilde{A}_i$ the rows, then $\mathrm{tr}(\mathbf{A}^2) = \sum_{i=1}^d A_i^\top \tilde{A}_i \le \sum_{i=1}^d \|A_i\|_2\|\tilde{A}_i\|_2 \le \sum_{i=1}^d (\|A_i\|_2^2 + \|\tilde{A}_i\|_2^2)/2 \le \|\mathbf{A}\|_{\mathrm{F}}^2$.

Proof of Item (2): Re-writing Equation (40), we see,

$$|g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A})| \lesssim \|\mathbf{A}\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + \|\mathbf{A}\|_{\mathrm{op}}\log(1/\delta) . \tag{42}$$

Moreover,

$$\begin{aligned}
\tau_{\mathbf{A}}(X) - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A}}(X)] &\le \tau_{\mathbf{A}}(X) \qquad (\tau_{\mathbf{A}}(X) \ge 0 \text{ by definition}) \\
&= (g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}))\mathbb{1}(g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}) > 100\log(1/\epsilon)) \\
&\le |g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A})| \\
&\lesssim \|\mathbf{A}\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + \|\mathbf{A}\|_{\mathrm{op}}\log(1/\delta) \qquad \text{(by (42))}
\end{aligned}$$

which means that the random variable $\tau_{\mathbf{A}}(x) - \mathbf{E}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A}}(X)]$ is also $(\nu,\beta)_+$-sub-gamma with $\nu = \|\mathbf{A}\|_{\mathrm{F}}$ and $\beta = \|\mathbf{A}\|_{\mathrm{op}}$.

Proof of Item (3) By Lemma B.8, Item (2) implies that for i.i.d. samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \mathbf{I})$, the average $\frac{1}{n}\sum_{i=1}^n \tau_{\mathbf{A}}(X_i) - \mathbf{E}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A}}(X)]$ is $(\|\mathbf{A}\|_{\mathrm{F}}/\sqrt{n}, \|\mathbf{A}\|_{\mathrm{op}}/n)$-sub-gamma or, equivalently, with probability $1-\delta$ it holds:

$$\frac{1}{n}\sum_{i=1}^n \tau_{\mathbf{A}}(X_i) - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A}}(X)] \lesssim \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + \frac{1}{n}\|\mathbf{A}\|_{\mathrm{op}}\log(1/\delta) . \tag{43}$$

Proof of Item (4): The above is about a fixed choice of the matrix $\mathbf{A}$ from the set $\mathcal{V} := \{\mathbf{A} \in \mathbb{R}^{d\times d} : \|\mathbf{A}\|_{\mathrm{F}} \leq \sqrt{\log(1/\epsilon)}, \|\mathbf{A}\|_{\mathrm{op}} = 1, \mathbf{A} \text{ has at most } k^2 \text{ non-zero elements}\}$. To show that concentration holds for all $\mathbf{A}$ in $\mathcal{V}$, we take a cover set $\mathcal{V}_\eta \subset \mathcal{V}$. For every $\eta > 0$ let $\mathcal{V}_\eta$ be an η-cover of $\mathcal{V}$, i.e., a set $\mathcal{V}_\eta$ such that for every $\mathbf{A} \in \mathcal{V}$ there exists an $\mathbf{A}' \in \mathcal{V}_\eta$ with $\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}} \leq \eta$. The next claim shows that the size of V_η is not too large.

Claim E.1. *The size of $\mathcal{V}_\eta$ is upper bounded by $(6\sqrt{\log(1/\epsilon)}/\eta)^{k^2}\binom{d}{k^2}$.*

Proof. If we look at the flattened version of the matrix as a vector in $\mathbb{R}^{d\times d}$, there are $\binom{d}{k^2}$-many ways to select which are the k^2 non-zero elements, and for each such choice of the non-zero elements there exists a cover of size $(3\sqrt{\log(1/\epsilon)}/\eta)^{k^2}$ by Fact B.1. The union of all of these sets covers has size at most $(3\sqrt{\log(1/\epsilon)}/\eta)^{k^2}\binom{d}{k^2}$ but may not necessarily be a subset of $\mathcal{V}$. However, by Exercise 4.2.9 in Vershynin (2018) there exists a $\mathcal{V}_\eta \subseteq \mathcal{V}$ with size same as before but by replacing η by $\eta/2$. □

We will choose $\eta = 0.0001\epsilon^4/d^4$ (the reason for this choice will be clear later on). By setting $\delta = \delta'/(6\sqrt{\log(1/\epsilon)}/\eta)^{k^2}\binom{d}{k^2}$ and by a union bound, we can ensure that (43) holds for all $A \in \mathcal{V}_\eta$ *simultaneously*. We will now show that, by continuity, the upper bound of (43) with some additional error terms holds for all $A \in \mathcal{V}$ simultaneously.

We will need the following notation: let $\mathbf{A} \in \mathcal{V}$ be an arbitrary element of $\mathcal{V}$ and $\mathbf{A}' \in \mathcal{V}_\eta$ satisfy $\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}} \leq \eta$. Denote by $S = \{X_1, \ldots, X_n\}$ the set of i.i.d. samples from $\mathcal{N}(\mu, \mathbf{I})$. Also denote by $\Delta g(X) := (g_{\mathbf{A}}(x) - \mathrm{tr}(\mathbf{A})) - (g_{\mathbf{A}'}(x) - \mathrm{tr}(\mathbf{A}'))$ the difference between the two polynomials. The goal is to upper bound $\mathbf{E}_{X\sim S}[\tau_{\mathbf{A}}(X)]$ which we do in steps as follows: First, we rewrite $(g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A})) = (g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}')) + \Delta g(X)$, to get

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim S}[\tau_{\mathbf{A}}(X)] &= \mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}))\mathbb{1}(g_{\mathbf{A}}(X) - \mathrm{tr}(\mathbf{A}) > 100\log(1/\epsilon))] \\ &\leq \mathop{\mathbf{E}}_{X\sim S}[|\Delta g(X)|] + \mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}'))\mathbb{1}(g_{\mathbf{A}'}(x) - \mathrm{tr}(\mathbf{A}') > 100\log(1/\epsilon) - \Delta g(X))] \,. \end{aligned} \tag{44}$$

We will now bound each of the terms above. For the first one, we show that the following bound holds with probability $1 - \delta$ over the random selection of S

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim S}[|\Delta g(X)|] &\leq \mathop{\mathbf{E}}_{X\sim S}[|g_{\mathbf{A}}(X) - g_{\mathbf{A}'}(X)|] + |\mathrm{tr}(\mathbf{A} - \mathbf{A}')| \\ &= \left|\left\langle \mathbf{A} - \mathbf{A}', \frac{1}{n}\sum_{i=1}^n (X_i - \mu)(X_i - \mu)^\top \right\rangle\right| + |\mathrm{tr}(\mathbf{A} - \mathbf{A}')| \\ &\leq \|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}} \frac{1}{n}\sum_{i=1}^n \|X_i - \mu\|_2^2 + d\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}} \qquad (45) \\ &\lesssim \eta d, \qquad (46)\end{aligned}$$

where, in (45) we used that $\langle \mathbf{B}, \mathbf{C}\rangle \leq \|\mathbf{B}\|_{\mathrm{F}}\|\mathbf{C}\|_{\mathrm{F}}$ and $\|xx^\top\|_{\mathrm{F}} = \|x\|_2^2$, and in (46) we used that $\|\mathbf{A} - \mathbf{A}'\|_{\mathrm{F}} \leq \eta$ and that $\sum_{i=1}^n \|X_i - \mu\|_2^2 \leq d + O(\log(1/\delta)/n) = O(d)$ by Gaussian norm concentration Fact B.5 combined with the last part of Lemma B.8.

We now move to the second term in the RHS of (44).

$$\begin{aligned}&\mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}'))\mathbb{1}(g_{\mathbf{A}'}(x) - \mathrm{tr}(\mathbf{A}') > 100\log(1/\epsilon) - \Delta g(X))] \\ &\quad= \mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}'))\mathbb{1}(g_{\mathbf{A}'}(x) - \mathrm{tr}(\mathbf{A}') > 100\log(1/\epsilon) - \Delta g(X))\mathbb{1}(\Delta g(X) < \log(1/\epsilon))] \\ &\qquad+ \mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}'))\mathbb{1}(g_{\mathbf{A}'}(x) - \mathrm{tr}(\mathbf{A}') > 100\log(1/\epsilon) - \Delta g(X))\mathbb{1}(\Delta g(X) > \log(1/\epsilon))] \\ &\quad\leq \mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}'))\mathbb{1}(g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}') > 99\log(1/\epsilon))] \\ &\qquad+ \mathop{\mathbf{E}}_{X\sim S}[|g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}')|\mathbb{1}(\Delta g(X) > \log(1/\epsilon))] \,. \end{aligned} \tag{47}$$

The first term is almost the same as $\mathbf{E}_{X\sim S}[\tau_A(X)]$,[11] which by Equation (41) and (43) we know that is upper bounded as follows:

$$\mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X)-\mathrm{tr}(\mathbf{A}'))\mathbb{1}(g_{\mathbf{A}'}(X)-\mathrm{tr}(\mathbf{A}')>99\log(1/\epsilon))] \lesssim \epsilon^4 + \frac{\|\mathbf{A}\|_{\mathrm{F}}\sqrt{\log(1/\delta)}}{\sqrt{n}} + \frac{\|\mathbf{A}\|_{\mathrm{op}}\log(1/\delta)}{n} \tag{48}$$

$$\lesssim \epsilon^4 + \frac{\sqrt{\log(1/\epsilon)\log(1/\delta)}}{\sqrt{n}} + \frac{\log(1/\delta)}{n}, \tag{49}$$

where the last line used the assumptions $\|\mathbf{A}\|_{\mathrm{F}} \le \sqrt{\log(1/\epsilon)}$ and $\|\mathbf{A}\|_{\mathrm{op}} \le 1$.

For the second term in (47) we claim that

$$\mathop{\mathbf{E}}_{X\sim S}[|g_{\mathbf{A}'}(X)-\mathrm{tr}(A')|\mathbb{1}(\Delta g_{\mathbf{A}}(X)>\log(1/\epsilon))] = 0 \tag{50}$$

whenever $\|x-\mu\|_2 \le 10\sqrt{d}$ for all $x \in S$ because of the indicator: This is because $\Delta g(x) = \langle \mathbf{A}-\mathbf{A}', (x-\mu)(x-\mu)^\top\rangle + \mathrm{tr}(\mathbf{A}-\mathbf{A}') \le \|\mathbf{A}-\mathbf{A}'\|_{\mathrm{F}}\|x-\mu\|_2^2 + \mathrm{tr}(\mathbf{A}-\mathbf{A}') \le \eta 100d + d\|\mathbf{A}-\mathbf{A}'\|_{\mathrm{F}} \le 101\eta d$ and if $\eta < 0.0001\log(1/\epsilon)/d$, then the previous quantity is less than $\log(1/\epsilon)$. The event that for all $x \in S$, $\|x-\mu\|_2 \le 10\sqrt{d}$ happens with overwhelming probability: $\Pr_{X_1,\ldots,X_i\sim\mathcal{N}(\mu,I)}[\exists X_i \in S : \|X_i-\mu\|_2 > 10\sqrt{d}] \le ne^{-d/100}$ by Gaussian norm concentration (Fact B.5) and a union bound.

Combining (44),(46),(47),(43),(50), and choosing $\delta = \delta'/(6\sqrt{\log(1/\epsilon)}/\eta)^{k^2}\binom{d}{k^2}$, and $\eta = 0.0001\epsilon^4/d^4$, we conclude that with probability $1-\delta'-ne^{-d/100}$, the following holds simultaneously for all matrices A in the cover:

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim S}[\tau_A(X)] &\lesssim \epsilon^4 + \sqrt{\frac{\log(1/\epsilon)\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} + d\eta \\ &\lesssim \epsilon^4 + \sqrt{\frac{\log(1/\epsilon)\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}\end{aligned} \tag{51}$$

Using $n \gg \epsilon^{-2}\mathrm{polylog}(1/\epsilon)(k^2\log(d/\epsilon)+\log(1/\delta'))$ we can make the RHS less than ϵ. Finally, in order to have $ne^{-d/100} \le \delta'$ we need $d \gg \mathrm{polylog}(1/(\epsilon\delta'))$ but we can assume that this is true without loss of generality by padding the samples with additional Gaussian coordinates (if the goodness conditions hold for the padded data, they continue to hold for the original data).

Proof of Item (2)c in Condition (2): Qualitatively, the proof goes through the same arguments as the one for Item (2)a, thus we will not be that detailed and instead we will focus mostly on the few differences. Denote $g_{\mathbf{A}}(x) = (x-\mu)^\top\mathbf{A}(x-\mu)$, $h_{\beta,v} = \beta + v^\top(x-\mu)$, and $\tau_{\mathbf{A},\beta,v}(x) = (g_{\mathbf{A}}(x)-\mathrm{tr}(\mathbf{A}))\mathbb{1}(h(x)>100\log(1/\epsilon))$. For $t>1$, by Gaussian concentration $\Pr_{X\sim\mathcal{N}(\mu,\mathbf{I})}[h(x)>t] \le e^{-\Omega(t^2)}$. Thus, for $t = 100\log(1/\epsilon)$

$$\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,I)}[\tau_{\mathbf{A}}(X)] \le \sqrt{\Pr_{X\sim\mathcal{N}(\mu,\mathbf{I})}[h(x)>t]}\sqrt{\mathrm{Var}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[g_{\mathbf{A}}(X)]} \le \epsilon^5\|\mathbf{A}\|_{\mathrm{F}} \lesssim \epsilon^4 .$$

Similarly to Equation (42), $\tau_{\mathbf{A},\beta,v}(X) - \mathbf{E}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A},\beta,v}(X)]$ is $(\|\mathbf{A}\|_{\mathrm{F}}, \|\mathbf{A}\|_{\mathrm{op}})$-sub-gamma. Therefore, the average of n i.i.d. samples is $(\|\mathbf{A}\|_{\mathrm{F}}/\sqrt{n}, \|\mathbf{A}\|_{\mathrm{op}}/n)$-sub-gamma. We now let $\mathcal{V}_\eta$ be the cover set of Claim E.1, $\mathcal{C}_\eta$ be the cover set of the k-sparse unit ball, which has size at most $(3/\eta)^k\binom{d}{k}$, and finally let $\mathcal{V}'_\eta = \mathcal{V}_\eta \times \mathcal{C}_\eta$ be the product of the two. We choose $\eta = 0.0001(\epsilon/d)^{10}$ and probability of failure $\delta = 1/|\mathcal{V}'_\eta|$. By a union bound, we have that

$$\frac{1}{n}\sum_{i=1}^n \tau_{\mathbf{A}',\beta',v'}(X_i) - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\mathbf{I})}[\tau_{\mathbf{A}',\beta',v'}(X)] \lesssim \frac{1}{\sqrt{n}}\|\mathbf{A}'\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + \frac{1}{n}\|\mathbf{A}'\|_{\mathrm{op}}\log(1/\delta) . \tag{52}$$

holds simultaneously for all $\mathbf{A}', \beta', v'$ inside the cover set. Now let arbitrary $\mathbf{A}, \beta, v$. We can show that (52) will still hold, with some additional error terms. First, let $\Delta g(x) = (g_{\mathbf{A}}(x)-\mathrm{tr}(\mathbf{A})) - (g_{\mathbf{A}'}(x)-\mathrm{tr}(\mathbf{A}')$ and $\Delta h(x) = \beta - \beta' + (v-v')^\top(x-\mu)$. Also, denote by $S = \{X_1,\ldots,X_n\}$ the set of n samples.

$$\mathop{\mathbf{E}}_{X\sim S}[\tau_{\mathbf{A},\beta,v}(X)] \le \mathop{\mathbf{E}}_{X\sim S}[|\Delta g(X)|] + \mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X)-\mathrm{tr}(\mathbf{A}'))\mathbb{1}(h'(x)>100\log(1/\epsilon)-\Delta h(X))]$$

[11] The only difference is that the constant is 99 instead of 100 but this does not affect the conclusion.

The first term is at most $O(\eta d)$ as in Equation (46). We bound the second term as follows:

$$\mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}'))\mathbb{1}(h'(x) > 100\log(1/\epsilon) - \Delta h(X))] \tag{53}$$

$$\leq \mathop{\mathbf{E}}_{X\sim S}[(g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}'))\mathbb{1}(h'(x) > 99\log(1/\epsilon))] + \mathop{\mathbf{E}}_{X\sim S}[|g_{\mathbf{A}'}(X) - \mathrm{tr}(\mathbf{A}')|\mathbb{1}(\Delta h(X) > \log(1/\epsilon))] \tag{54}$$

The first term above is bounded as in Equation (52) (the only change is that the constant 100 is now 99 but that should only affect the constant in the RHS of Equation (52)). For the second term, we note that with probability $1 - ne^{-d/100}$ we have $\|x-\mu\|_2 \leq 10\sqrt{d}$ for all $x \in S$. Since $\Delta h(x) = \beta - \beta' + (v - v')^\top(x-\mu)$ and we have designed the cover such that $|\beta - \beta'|$ and $\|v - v'\|_2$ are at most $\eta \leq 0.0001(\epsilon/d)^{10}$, we have that $\mathbb{1}(\Delta h(X) > \log(1/\epsilon)) = 0$ for all $X \in S$ under that event. Thus, with probability $1 - ne^{-d/100}$ over the dataset S, the second term in Equation (54) is zero. Putting everything together, we have the same bound as in Equation (51).

Proof of Item (2)b in Condition (2): Using (40) with $t = 10\log(1/\epsilon)$, $\|\mathbf{A}\|_{\mathrm{F}} \leq \sqrt{\log(1/\epsilon)}$ and $\|\mathbf{A}\|_{\mathrm{op}} = 1$, we obtain that $\Pr_{X\sim\mathcal{N}(\mu,\mathbf{I})}[p(X) > 20\log(1/\epsilon)] \lesssim \epsilon^2$. Now by a basic application of Chernoff bounds we have that $|\Pr_{X\sim S}[p(X) > 20\log(1/\epsilon)] - \Pr_{X\sim\mathcal{N}(\mu,\mathbf{I})}[p(X) > 20\log(1/\epsilon)]| \leq \eta$ with probability $1 - e^{-\Omega(\eta^2 n)}$, where we will use $\eta = \epsilon$. Thus, if $n \gg \epsilon^{-2}\log(1/\delta)$, we have that $\Pr_{X\sim S}[p(X) > 20\log(1/\epsilon)] \lesssim \epsilon$ with probability at least $1-\delta$.

□