Robust Sparse Mean Estimation via Sum of Squares

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin-Madison

Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego

Sushrut Karmalkar SKARMALKAR@WISC.EDU
University of Wisconsin-Madison

Ankit Pensia ANKITP@CS.WISC.EDU
University of Wisconsin-Madison

Thanasis Pittas PITTAS@WISC.EDU
University of Wisconsin-Madison

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study the problem of high-dimensional sparse mean estimation in the presence of an ϵ-fraction of adversarial outliers. Prior work obtained sample and computationally efficient algorithms for this task for identity-covariance subgaussian distributions. In this work, we develop the first efficient algorithms for robust sparse mean estimation without a priori knowledge of the covariance. For distributions on $\mathbb{R}^d$ with "certifiably bounded" t-th moments and sufficiently light tails, our algorithm achieves error of $O(\epsilon^{1-1/t})$ with sample complexity $m = (k\log(d))^{O(t)}/\epsilon^{2-2/t}$. For the special case of the Gaussian distribution, our algorithm achieves near-optimal error of $\tilde{O}(\epsilon)$ with sample complexity $m = O(k^4\mathrm{polylog}(d))/\epsilon^2$. Our algorithms follow the Sum-of-Squares based, proofs to algorithms approach. We complement our upper bounds with Statistical Query and low-degree polynomial testing lower bounds, providing evidence that the sample-time-error tradeoffs achieved by our algorithms are qualitatively the best possible.

Keywords: robust statistics, sparse estimation, sum of squares, statistical query model

1. Introduction

High-dimensional robust statistics Hampel et al. (1986); Huber and Ronchetti (2009) aims to design estimators that are tolerant to a *constant fraction* of outliers, independent of the dimension. Early work in this field, see, e.g., Tukey (1960); Huber (1964); Tukey (1975), developed sample-efficient robust estimators for various basic tasks, alas with runtime exponential in the dimension. During the past five years, a line of work in computer science, starting with Diakonikolas et al. (2016); Lai et al. (2016), has developed the first *computationally efficient* robust high-dimensional estimators for a range of tasks. This progress has led to a revival of robust statistics from an algorithmic perspective, see, e.g., Diakonikolas and Kane (2019); Diakonikolas et al. (2021a) for recent surveys.

Throughout this work, we focus on the following standard contamination model.

Definition 1 (Strong Contamination Model) *Fix a parameter $0 < \epsilon < 1/2$. We say that a set of m points is an ϵ-corrupted set of samples from a distribution D if it is generated as follows: First, a set S of m points is sampled i.i.d. from D. Then an adversary observes S, replaces any ϵm of points in S with any vectors they like to obtain the set T. We say that T is an ϵ-corrupted version of S.*

Here we study high-dimensional robust statistics tasks in the presence of *sparsity constraints*. Leveraging sparsity in high-dimensional datasets is a fundamental and practically important problem (see, e.g., Hastie et al. (2015) for a textbook on the topic). We focus on arguably the most fundamental such problem, that of *robust sparse mean estimation*. Specifically, we are given an ϵ-corrupted set of samples from a structured distribution D, whose unknown mean $\mu = \mathbf{E}_{X\sim D}[X] \in \mathbb{R}^d$ is k-sparse (i.e., supported on at most k coordinates), and we want to compute a good approximation $\widehat{\mu}$ of μ. Importantly, in the sparse setting, we have access to much fewer samples compared to the dense case — namely $\mathrm{poly}(k, \log d)$ instead of $\mathrm{poly}(d)$. Consequently, the design and analysis of algorithms for robust sparse estimation requires additional ideas, as compared to the standard (dense) setting Diakonikolas et al. (2016).

Prior work on robust sparse mean estimation Balakrishnan et al. (2017); Li (2018); Diakonikolas et al. (2019a); Cheng et al. (2021) focused on the case that the covariance matrix of the inliers is *known or equal to the identity*. For identity covariance distributions with sufficiently good concentration (specifically, subgaussian concentration), the aforementioned works give efficient algorithms for robust k-sparse mean estimation that use $\mathrm{poly}(k, \log(d), 1/\epsilon)$ samples and achieve near-optimal ℓ_2-error of $\tilde{O}(\epsilon)$. On the other hand, if the covariance matrix of the inlier distribution is *unknown* and spectrally bounded by the identity, the techniques in these works can at best achieve error of $O(\sqrt{\epsilon})$, even for the special case of the Gaussian distribution. One can of course use a robust covariance estimation algorithm to reduce the problem to the setting of known covariance. The issue is that the covariance matrix is not necessarily sparse, and therefore naive attempts of robustly estimating the covariance (e.g., with respect to Frobenius or Mahalanobis distance) would require $\mathrm{poly}(d)$ samples.

Motivated by these drawbacks of prior work, in this paper we aim to design computationally efficient algorithms for robust sparse mean estimation, using $\mathrm{poly}(k, \log(d), 1/\epsilon)$ samples, *that achieve near-optimal error guarantees without a priori knowledge of the covariance matrix*. Our main contribution is a comprehensive picture of the tradeoffs between sample complexity, running time, and error guarantee for a range of inlier distributions. In more detail, for distributions with appropriate tail bounds and "certifiably bounded" t-th moments in sparse directions (see Definition 13), we give an efficient algorithm that achieves error $O(\epsilon^{1-1/t})$. For the special case of the Gaussian distribution, we give an algorithm with near-optimal error of $\tilde{O}(\epsilon)$. For both settings, we establish Statistical Query (SQ) lower bounds (and low-degree polynomial testing lower bounds) which give evidence that the error-sample-time tradeoffs achieved by our algorithms are qualitatively the best possible.

1.1. Our Results

We start by recalling prior results for the dense robust mean estimation of bounded moment distributions. We say that a distribution D on $\mathbb{R}^d$ with mean μ has t-th central moments bounded by M if for all unit vectors v it holds $\mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right] \le M$. Although it is information-theoretically possible to robustly estimate the mean of such distributions, in the ℓ_2-norm, up to error $O(M^{1/t}\epsilon^{1-1/t})$ using $O(d/\epsilon^{2-2/t})$ samples (see Appendix G for the proof), all known efficient algorithms require the following stronger assumption.

Definition 2 (Certifiably (M,t)-Bounded Central Moments) *We say that a distribution D on $\mathbb{R}^d$ with mean μ has t-th central moments* certifiably *bounded by M if $M\|v\|_2^t - \mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]$ can be written as a sum of square polynomials in $v = (v_1, \ldots, v_d)$ of degree $O(t)$.*

Prior works Kothari and Steurer (2017); Hopkins and Li (2018) gave efficient algorithms for dense robust mean estimation of distributions with certifiably bounded central moments. Their algorithms incur sample complexities at least $m = \mathrm{poly}(d^t)/\epsilon^2$, have running times $\mathrm{poly}((md)^t)$, and guarantee ℓ_2-error of $O(M^{1/t}\epsilon^{1-1/t})$. [1]

We now turn our attention to the sparse setting, which is the focus of the current work. In prior work, the term "sparse mean estimation" refers to the task of computing a $\widehat{\mu}$ such that $\widehat{\mu} - \mu$ is small in ℓ_2-norm, *assuming that μ is k-sparse*. We note that estimating a sparse vector in the ℓ_2-norm is a special case of estimating an arbitrary vector in the $(2, k)$-norm, defined below (see Fact 9).

Definition 3 ((2, k)-norm) *We define the $(2, k)$-norm of a vector x to be the maximum correlation with any k-sparse unit vector, i.e., $\|x\|_{2,k} \stackrel{\mathrm{def}}{=} \max_{\|v\|_2=1, v:k-\mathrm{sparse}} \langle v, x\rangle$.*

We henceforth focus on this more general formulation; we will use the term "sparse mean estimation" to mean that the error guarantees are defined with respect to the $(2, k)$-norm.

For distributions with (M, t)-bounded central moments, the information-theoretically optimal error for robust sparse mean estimation is $O(M^{1/t}\epsilon^{1-1/t})$ and can be obtained with $(k \log(d/k))/\epsilon^{2-2/t}$ samples (see Appendix G for the simple proof).

Our first result is a computationally efficient robust sparse mean estimation algorithm that applies to any distribution D with certifiably bounded t-th moments in k-sparse directions (Definition 13) and light tails. In particular, we assume that D has subexponential tails, i.e., for some universal constant c, for all unit vectors v and all $p \in \mathbb{N}$ it holds $\mathbf{E}_{X\sim D}\left[|\langle v, X - \mathbf{E}_{X\sim D}[X]\rangle|^p\right]^{1/p} \le cp$. (In fact, our algorithmic result holds as long the distribution D has bounded $\mathrm{poly}(t \log d)$ moments along coordinate axes; see Section 3.1 and Appendix B.2.) Our algorithm achieves error $O(M^{1/t}\epsilon^{1-1/t})$ with $m = \mathrm{poly}((k \log d)^t)/\epsilon^{2-2/t}$ samples and $\mathrm{poly}((md)^t)$ running time.

Theorem 4 (Robust Sparse Mean Estimation for Certifiably Bounded Moments) *Let t be a power of two, D be a distribution on $\mathbb{R}^d$ with unknown mean μ, and $\epsilon < \epsilon_0$ for a sufficiently small constant $\epsilon_0 > 0$. Suppose that D has t-th moments certifiably bounded in k-sparse directions by M (cf. Definition 13) and subexponential tails. There is an algorithm which, given ϵ, M, t, k, and an ϵ-corrupted set of $m = (tk \log d)^{O(t)} \max(1, M^{-2})/\epsilon^{2-2/t}$ samples from D, runs in time $\mathrm{poly}((md)^t)$, and returns a vector $\widehat{\mu}$ satisfying $\|\widehat{\mu} - \mu\|_{2,k} \le O(M^{1/t}\epsilon^{1-1/t})$ with high probability.*

It is natural to ask which distributions have such "certifiably bounded moments in k-sparse directions". In the dense case, Kothari and Steinhardt (2017) showed that Definition 2 is satisfied by σ-Poincaré distributions. (A distribution D is σ-Poincaré if for every differentiable $f : \mathbb{R}^d \to \mathbb{R}$, $\mathbf{Var}_{X\sim D}[f(X)] \le \sigma^2\, \mathbf{E}_{X\sim D}[\|\nabla f(X)\|_2^2]$). We show in Appendix B.1 that this class also has certifiably bounded moments in k-sparse directions, in the sense of Definition 13. Combining this with the fact that the tails of σ-Poincaré distributions are inherently subexponential, Theorem 4 is applicable.

We complement our upper bound with a Statistical Query (SQ) lower bound (and low-degree testing lower bound), which gives evidence that the factor $k^{O(t)}$ in the sample complexity of Theorem 4 might be necessary for efficient algorithms.

We remind the reader that SQ algorithms Kearns (1998) do not draw samples from the data distribution, but instead have access to an oracle that can return the expectation of any bounded function (up to a desired additive error). Specifically, an SQ algorithm is able to perform adaptive queries to a $\mathrm{STAT}(\tau)$ oracle, which we define below.

1. For simplicity of the exposition, we will not account for bit complexity. In essence, we assume that the bit complexity of all relevant parameters is bounded by $\mathrm{poly}(md)$.

Definition 5 (STAT Oracle) *Let D be a distribution on $\mathbb{R}^d$. A statistical query is a bounded function $f : \mathbb{R}^d \to [-1, 1]$. For $\tau > 0$, the $\mathrm{STAT}(\tau)$ oracle responds to the query f with a value v such that $|v - \mathbf{E}_{X\sim D}[f(X)]| \leq \tau$. We call τ the tolerance of the statistical query.*

An SQ lower bound is an unconditional lower bound showing that for any SQ algorithm, either the number of queries q must be large or the tolerance of some query, τ, must be small. The standard interpretation of SQ lower bounds hinges on the fact that simulating a query to $\mathrm{STAT}(\tau)$ using i.i.d. samples may require $\Omega(1/\tau^2)$ many samples. Thus, an SQ lower bound stating that any SQ algorithm either makes r queries or needs tolerance τ is interpreted as a tradeoff between runtime $\Omega(r)$ and sample complexity $\Omega(1/\tau^2)$.

Recall that a distribution D is subgaussian if there exists an absolute constant c such that for all unit vectors v it holds that $\mathbf{E}_{X\sim D}\left[|\langle v, X - \mathbf{E}_{X\sim D}[X]\rangle|^p\right]^{1/p} \leq c\sqrt{p}$. We show the following (see Theorem 75 for a detailed formal statement).

Theorem 6 (SQ Lower Bound for Subgaussian Distributions, Informal Statement) *Fix $t \in \mathbb{N}$ with $t \geq 2$ and assume that $d \geq k^2$ for k sufficiently large. Any SQ algorithm that obtains error $o(\epsilon^{1-1/t})$ for robust sparse mean estimation of a subgaussian distribution (with t-th moments certifiably bounded in k-sparse directions) either requires $d^{\Omega(\sqrt{k})}$ statistical queries or makes at least one query with tolerance $k^{-\Omega(t)}$.*

For the statement of our low-degree testing lower bound, see Appendix F.3. Informally speaking, Theorem 6 shows that any SQ algorithm that returns a $\widehat{\mu}$ satisfying $\|\widehat{\mu} - \mu\|_{2,k} = o(\epsilon^{1-1/t})$ requires runtime exponential in k, unless it uses queries of tolerance $k^{-\Omega(t)}$ — requiring $k^{\Omega(t)}$ samples for simulation. We briefly remark that Theorem 6 also has implications for the dense setting: By taking $k = \sqrt{d}$, Theorem 6 suggests that $d^{\Omega(t)}$ samples may be necessary to efficiently obtain $o(\epsilon^{1-1/t})$ error in the dense setting; this qualitatively matches the algorithmic results of Kothari and Steurer (2017).

Interestingly, Theorem 6 does not apply when the inlier distribution is Gaussian, i.e., the SQ-hard instance of Theorem 6 is *not* a Gaussian distribution. In fact, our next result shows that it *is* possible to achieve the near-optimal error of $\tilde{O}(\epsilon)\sqrt{\|\Sigma\|_2}$ for $\mathcal{N}(\mu, \Sigma)$, using $(k^4/\epsilon^2)\mathrm{polylog}(d/\epsilon)$ samples.

Theorem 7 (Robust Sparse Gaussian Mean Estimation) *Let $k, d \in \mathbb{Z}_+$ with $k \leq d$ and $\epsilon < \epsilon_0$ for a sufficiently small constant $\epsilon_0 > 0$. Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix. There exists an algorithm which, given ϵ,k, and an ϵ-corrupted set of samples from $\mathcal{N}(\mu, \Sigma)$ of size $m = O((k^4/\epsilon^2)\log^5(d/(\epsilon)))$, runs in time $\mathrm{poly}(md)$, and returns an estimate $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_{2,k} \leq \tilde{O}(\epsilon)\sqrt{\|\Sigma\|_2}$ with high probability.*

Information-theoretically, $O(k\log(d/k))/\epsilon^2$ samples suffice to obtain $O(\epsilon)$ error (see Theorem 87). Prior work has given evidence that $\Omega(k^2)$ samples might be necessary for efficient algorithms to obtain dimension-independent error Diakonikolas et al. (2017b); Brennan and Bresler (2020). Perhaps surprisingly, here we establish an SQ lower bound suggesting that the $\Omega(k^4)$ sample complexity of our algorithm might be inherent for efficient algorithms to achieve error $o(\epsilon^{1/2})$ (see Theorem 73 for a formal statement):

Theorem 8 (SQ Lower Bound for Gaussian Sparse Mean Estimation, Informal Statement) *Let $0 < c < 1$ and assume that $d \geq k^2$ for k sufficiently large. Any SQ algorithm that performs robust sparse mean estimation of Gaussians with $\Sigma \preceq I$ up to error $o(\sqrt{\epsilon})$ does one of the following: It either requires $d^{\Omega(ck^c)}$ queries or makes at least one query with tolerance $O(k^{-2+2c})$.*

The intuitive interpretation of Theorem 8 is that any SQ algorithm for this task either has runtime $d^{\mathrm{poly}(k)}$ or uses $\Omega(k^4)$ samples (a similar hardness holds for low-degree polynomial tests; see Theorem 85).

1.2. Overview of Techniques

To establish Theorems 4 and 7, we use the *sum-of-squares* framework, i.e., solve a *sum-of-squares* (SoS) SDP relaxation of a system of polynomial inequalities.

1.2.1. ROBUST SPARSE MEAN ESTIMATION WITH BOUNDED MOMENTS

Identifiability in the Presence of Outliers Similar to Kothari and Steurer (2017), our starting point is a set of polynomial constraints (Definition 17) in which the variables try to identify the uncorrupted samples. The program has a vector of variables for each sample in the set (we will refer to these variables as "ghost samples"), and enforces that these ghost samples match a $(1-\epsilon)$-fraction of the data. The constraints also enforce that the uniform distribution over the ghost samples has t-th moments certifiably bounded by M *in k-sparse directions* (Definition 13). A key property of an SoS relaxation is that it satisfies any polynomial inequality that is true subject to the constraints of the original polynomial system, *as long as* this inequality has an "SoS proof" of degree t (i.e., the difference of the two sides is a sum of square polynomials, where each polynomial has degree at most t). We give an SoS proof of the fact that the mean of the ghost samples is close to the true mean in k-sparse directions (proof of identifiability; see Section 4.1).

Sampling Preserves Certifiably Bounded Moments For our identifiability proof (and to ensure feasibility of our program), we require that the uniform distribution over the uncorrupted samples satisfies certifiably bounded t-th moments in k-sparse directions. However, we initially know only that the distribution from which these samples are drawn has t-th moments certifiably bounded by M in k-sparse directions. Given that the underlying inlier distribution satisfies certifiably bounded moments, we need to show that the property transfers to the empirical distribution. In the dense case, it is relatively easy to prove such a concentration result, using $\mathrm{poly}(d)$ samples for all $v \in \mathbb{R}^d$, via spectral concentration inequalities. On other other hand, establishing the analogous statement in the sparse setting with $\mathrm{poly}(k)$ samples requires an alternate approach.

The first step towards showing this is Lemma 15, which states that for all polynomials $r(v) = \sum_{I\in[d]^t} r_I \prod_{j\in[t]} v_{I_j}$ and k-sparse vectors v, the inequality $r(v)^2 \le k^t \max_{I\in[d]^t} r_I^2$ has an SoS proof. Applying this to the true and empirical moments, $p(v) := \mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]$ and $\widehat{p}(v) = \mathbf{E}_{i\sim[m]}\left[\langle v, X_i-\overline{\mu}\rangle^t\right]$, we see that there is an SoS proof of the following statement $(p(v)-\widehat{p}(v))^2 \le k^t \|\mathbf{E}_{i\sim[m]}[(X_i-\mu)^{\otimes t}] - \mathbf{E}_{X\sim\mathcal{D}}[(X-\mu)^{\otimes t}]\|_\infty^2$.

In Appendix B.2, we establish concentration of the ℓ_∞-norm of the aforementioned tensor. We note that our result only requires $O(\log(d))$ moments to be bounded for concentration; prior work in the sparse setting assumes that the distribution has known covariance and is additionally either subgaussian or subexponential.

1.2.2. ACHIEVING NEAR-OPTIMAL ERROR FOR GAUSSIAN INLIERS

The first polynomial-time algorithm for robustly learning an arbitrary Gaussian (in the dense setting) was given in Diakonikolas et al. (2016). Specifically, that work showed how to robustly estimate the mean in ℓ_2-norm and the covariance in Mahalanobis norm up to an error of $\tilde{O}(\epsilon)$ using $\tilde{O}(d^2/\epsilon^2)$

samples. It is not clear how to directly adapt the approach of Diakonikolas et al. (2016) to the sparse setting, while achieving the desired sample complexity of $\mathrm{poly}(k, \log(d), 1/\epsilon)$. Instead, our starting point will be the recent work Kothari et al. (2022), which gave nearly matching guarantees for the dense Gaussian setting using the SoS method. The difficulty of matching the Diakonikolas et al. (2016) guarantees using sum-of-squares lies in the fact that standard SoS approaches typically require concentration of degree-t polynomials to obtain error $O(\epsilon^{1-1/t})$; the parameter t would need to be roughly $\sqrt{\log(1/\epsilon)}$ to get error of $\tilde{O}(\epsilon)$. Kothari et al. (2022) was able to achieve this result using SoS certifiability of bounded moments only up to degree four.

We now explain how we adapt the approach of Kothari et al. (2022) to the sparse setting. Assume that the covariance of the inliers is spectrally bounded, namely that $\|\Sigma\|_2 \leq 1$. For the dense result obtained in Kothari et al. (2022), it suffices to show SoS proofs of *multiplicative* concentration inequalities for Gaussian polynomials of degree up to four. In the absence of sparsity constraints, this is achieved by standard spectral matrix concentration. Unfortunately, the technique from the previous section only gives us an *additive* concentration inequality. This qualitative difference is significant and makes it challenging to obtain a guarantee scaling with $\|\Sigma\|_2$. To circumvent this issue, we add independent noise to each sample generated as $\mathcal{N}(0, I)$, which ensures that $I \preceq \Sigma \preceq 2I$, while keeping the mean unaffected. We thus obtain an efficient estimator with $\tilde{O}(\epsilon)$ error for the case that $I \preceq \Sigma \preceq 2I$ (Section 5.2). On the other hand, if $\|\Sigma\|_2$ is much smaller than 1, then the right error guarantee is $\tilde{O}(\epsilon)\sqrt{\|\Sigma\|_2}$. In Appendix D.4, we use Lepskii's method Lepskii (1991) to obtain an error guarantee that scales with $\|\Sigma\|_2$, as desired. This is done roughly as follows: We first obtain a rough estimate of $\|\Sigma\|_2$ that is within $\mathrm{poly}(d)$ factor away from the true value (by taking the median of $\|X_i - X_j\|_2$, where the X_i are samples). We next run our robust estimation algorithm after convolving the data with noise at various scales σ and getting a corresponding estimate. With high probability, whenever our candidate upper bound, σ, is bigger than $\sqrt{\|\Sigma\|_2}$, we get a point within distance $\tilde{O}(\epsilon)\sigma$ of the true mean. We then find the smallest value of σ such that the output is consistent with the larger values of σ and return the corresponding estimate.

1.2.3. STATISTICAL QUERY AND LOW-DEGREE TESTING LOWER BOUNDS

Our SQ lower bounds leverage the framework of Diakonikolas et al. (2017b) which showed the following: Let A be a one-dimensional distribution matching its first m moments with $\mathcal{N}(0, 1)$. Then the task of distinguishing between (i) $\mathcal{N}(0, I)$ and (ii) the d-dimensional distribution that coincides with A in an unknown k-sparse direction but is standard Gaussian in all perpendicular directions, requires either $q = d^{\mathrm{poly}(k)}$ queries or tolerance $\tau < \frac{1}{k^{(m+1)/2}}$ in the SQ model. The robust sparse mean estimation problems that we consider can be phrased in this form; the challenge is to construct the appropriate moment-matching distributions.

In Theorem 8, we establish a lower bound of $\Omega(k^4)$ on the sample complexity of any efficient SQ algorithm that robustly estimates a sparse mean within ℓ_2-error $o(\sqrt{\epsilon})$. Interestingly, this lower bound nearly matches the sample complexity of our algorithm (Theorem 7). We view this information-computation tradeoff as rather surprising. Recall that in the (easier) case where the covariance of the inliers is known to be the identity, $O(k^2 \log d)$ samples are sufficient for efficient algorithms Balakrishnan et al. (2017), and there is evidence that this sample size is also necessary for efficient algorithms Diakonikolas et al. (2017b); Brennan and Bresler (2020).

To prove our SQ lower bound in this case, we need to construct a univariate density A that matches (i) $m = 3$ moments with $\mathcal{N}(0, 1)$, and (ii) A is ϵ-corruption of $\mathcal{N}(\Theta(\sqrt{\epsilon}), 1)$. To achieve this, we leverage a lemma from Diakonikolas et al. (2019b) that lets A have a Gaussian inlier component

with mean $\Theta(\sqrt{\epsilon})$ and variance slightly smaller than 1. A suitable outlier component can then correct the first three moments of the overall mixture, so that they match the first three moments of $\mathcal{N}(0,1)$.

A more sophisticated choice of A is required to establish our Theorem 6. Specifically, we need to select $A = (1-\epsilon)G + \epsilon B$, where (i) A matches its first t moments with $\mathcal{N}(0,1)$, (ii) G is an explicit subgaussian distribution, and (iii) $\mathbf{E}_{X\sim G}[X] = \Omega(\epsilon^{1-1/t})$. For G, we start with a shifted Gaussian, $\mathcal{N}(\Theta(\epsilon^{1-1/t}),1)$, that we modify by adding a degree-t polynomial $p(x)$ in $[-1,1]$. Since we modify the Gaussian only on $[-1,1]$, the distribution continues to be subgaussian. By imposing the moment-matching conditions and expanding $p(x)$ in the basis of Legendre polynomials, we show that such a $p(\cdot)$ exists, so that (i)-(iii) above hold. We also show that the constructed distributions have SoS certifiable bounded t-th moments, and hence fall into the class of distributions for which our upper bounds apply (see Appendix F.2).

Finally, by exploiting the relationship between the SQ model and low-degree polynomial tests from Brennan et al. (2021), we also obtain quantitatively similar lower bounds against low-degree polynomial tests. The information-theoretic characterization of error and sample complexity appear in Appendix G.

1.3. Prior and Related Work

After the early works Diakonikolas et al. (2016); Lai et al. (2016), the field of algorithmic robust statistics has seen a plethora of research activity. Focusing on the dense setting, prior work has obtained computationally-efficient algorithms for a variety of problems, including mean estimation Diakonikolas et al. (2017a); Cheng et al. (2018); Depersin and Lecue (2019); Dong et al. (2019); Diakonikolas et al. (2020c), covariance and higher moment estimation Diakonikolas et al. (2016); Kothari and Steurer (2017); Cheng et al. (2019), linear regression Klivans et al. (2018); Diakonikolas et al. (2019c); Pensia et al. (2020); Bakshi and Prasad (2021), learning with a majority of outliers and clustering mixture models Kothari and Steinhardt (2017); Hopkins and Li (2018); Diakonikolas et al. (2018b, 2020a); Bakshi et al. (2020a); Liu and Moitra (2020); Bakshi et al. (2020b); Diakonikolas et al. (2020b, 2021b,c), and stochastic convex optimization Prasad et al. (2020); Diakonikolas et al. (2018a). We remark that some of these algorithms also leverage the SoS method.

Finally, we discuss results that leverage sparsity to improve sample complexity for computationally-efficient algorithms. Balakrishnan et al. (2017) presented the first computationally-efficient algorithms for a range of sparse estimation tasks including mean estimation. However, their estimation algorithm crucially relies on the fact that the inlier distribution is Gaussian with identity covariance. As opposed to the convex programming approach of Balakrishnan et al. (2017), Diakonikolas et al. (2019a) proposed a spectral algorithm for sparse robust mean estimation of identity-covariance Gaussians. Recently, Cheng et al. (2021) proposed a non-convex formulation and showed that any approximate stationary point (that can be obtained by efficient first-order algorithms) suffices. We reiterate that none of these algorithms give $o(\sqrt{\epsilon})$ error when the covariance of the inliers is unknown.

Finally, we mention that median-of-means preprocessing has been applied to achieve $O(\sqrt{\epsilon})$ error for robust mean estimation in near-linear time Depersin and Lecue (2019); Diakonikolas et al. (2020c); Hopkins et al. (2020); Lei et al. (2020). However, median-of-means preprocessing does not obtain $o(\sqrt{\epsilon})$ error, even when the inliers are Gaussian with identity covariance, and is thus not applicable to our setting.

1.4. Organization

The structure of this paper is as follows: In Section 2, we define the necessary notation and record basic facts about the SoS framework. In Section 3, we define the notion of *certifiably bounded central moments* in k-sparse directions, and show that this property is preserved under sampling. In Section 4, we give an SoS algorithm for robust sparse mean estimation under certifiably bounded central moments in sparse directions, establishing Theorem 4. In Section 5, we give an efficient estimator that achieves near-optimal error for Gaussian distributions with unknown covariance, establishing Theorem 7. Due to space constraints, we prove our SQ lower bounds for the previous two settings in Appendix E, establishing Theorems 6 and 8. For clarity of the exposition, some technical proofs are deferred to the appendix.

2. Preliminaries

Basic Notation. We use I_d to denote the $d \times d$ identity matrix. We use $\mathbb{N}$ to denote natural numbers and $\mathbb{Z}_+$ to denote positive integers. For $n \in \mathbb{Z}_+$ we denote $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$ and use $\mathcal{S}^{d-1}$ for the d-dimensional unit sphere. We denote by $\mathbf{1}(\mathcal{E})$ the indicator function of the event $\mathcal{E}$.

For a random variable X, we use $\mathbf{E}[X]$ for its expectation. We use $\mathcal{N}(\mu, \Sigma)$ to denote the Gaussian distribution with mean μ and covariance matrix Σ. We let ϕ denote the pdf of the one-dimensional standard Gaussian. When D is a distribution, we use $X \sim D$ to denote that the random variable X is distributed according to D. For a vector v, we let $\|v\|_2$ denote its ℓ_2-norm. We use $\langle v, u \rangle$ for the inner product of the vectors u, v. For a matrix A, we use $\|A\|_2, \|A\|_\infty$ to denote the spectral and entry-wise infinity-norm respectively. We denote the fact that A is PSD (positive semidefinite) by $A \succeq 0$. We write $A \preceq B$ when $B - A$ is PSD. We will use $\otimes$ to denote the standard Kronecker product. For any sequence $a_1, \ldots, a_m \in \mathbb{R}^n$, we will also use $\mathbf{E}_{i \sim [m]}[a_i]$ to denote $\frac{1}{m} \sum_{i \in [m]} a_i$. For any vector $(a_1, \ldots, a_d)$ and an ordered tuple $T \in [d]^t$ we define $a_T := \prod_{i \in T} a_i$. We will use $\mathcal{U}_k(d) := \left\{ v \in \mathbb{R}^d : \|v\|_2 = 1, \|v\|_0 = k \right\}$ to denote the set of unit k-sparse vectors of dimension d. We will omit the dimension from this notation when it is clear from the context. $\mathbb{R}[x_1, \ldots, x_d]_{\leq t}$ will denote the set of all degree t polynomials in $x_1, \ldots, x_d$ which have degree at most t.

The following fact (proved in Appendix A for completeness) can be used to translate bounds from the $(2, k)$-norm to the usual ℓ_2-norm when the underlying mean μ is sparse:

Fact 9 *Let $h_k : \mathbb{R}^d \to \mathbb{R}^d$ denote the function where $h_k(x)$ is defined to truncate x to its k largest coordinates in magnitude and zero out the rest. For all $\mu \in \mathcal{U}_k$, $\|h_k(x) - \mu\|_2 \leq 3\|x - \mu\|_{2,k}$.*

SoS Preliminaries. The following notation and preliminaries are specific to the SoS part of this paper. We refer the reader to Barak and Steurer (2016) for a complete treatment of basic definitions about the SoS hierarchy and SoS proofs. Here, we review the basics.

Definition 10 (SoS Proof) *Let $x_1, \ldots, x_d$ be indeterminates and let $\mathcal{A}$ be a set of polynomial inequalities $\{p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$. An SoS proof of the inequality $r(x) \geq 0$ from axioms $\mathcal{A}$ is a set of polynomials $\{r_S(x)\}_{S \subseteq [m]}$ such that each r_S is a sum of square polynomials and $r(x) = \sum_{S \subseteq [m]} r_S(x) \prod_{i \in S} p_i(x)$. If the polynomials $r_S(x) \cdot \prod_{i \in S} p_i(x)$ have degree at most t for all $S \subseteq [m]$ and the bit complexity of the coefficients of $r_S(x)$ and $p_i(x)$ is bounded by B, we say that this proof is of degree t and bit complexity B and denote it by $\mathcal{A} \vdash_t r(x) \geq 0$. We omit the bit complexity from our notation.*

When we need to emphasize what indeterminates are involved in a particular SoS proof, we denote it by $\mathcal{A} \left|\frac{x}{t}\right. r(x) \geq 0$. *When* $\mathcal{A}$ *is empty, we directly write* $\left|\frac{}{t}\right. r(x) \geq 0$ *and* $\left|\frac{x}{t}\right. r(x) \geq 0$. *We also often refer to* $\mathcal{A}$ *containing polynomial equations* $q(x) = 0$, *by which we mean that* $\mathcal{A}$ *contains both* $q(x) \geq 0$ *and* $q(x) \leq 0$.

We frequently compose SoS proofs without comment – see Barak and Steurer (2016) for basic facts about composition of SoS proofs and bounds on the degree of the resulting proofs. Our algorithm also uses the dual objects to SoS proofs, *pseudoexpectations*.

Definition 11 (Pseudoexpectation) *Let* $x_1, \ldots, x_d$ *be indeterminates. A degree-*t *pseudoexpectation* $\tilde{\mathbf{E}}$ *is a linear map* $\tilde{\mathbf{E}} : \mathbb{R}[x_1, \ldots, x_d]_{\leq t} \to \mathbb{R}$ *from degree-*t *polynomials to* $\mathbb{R}$ *such that* $\tilde{\mathbf{E}}\left[p(x)^2\right] \geq 0$ *for any* p *of degree at most* $t/2$ *and* $\tilde{\mathbf{E}}\left[1\right] = 1$. *If* $\mathcal{A} = \{p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$ *is a set of polynomial inequalities, we say that* $\tilde{\mathbf{E}}$ *satisfies* $\mathcal{A}$ *if for every* $S \subset [d]$, *the following holds:* $\tilde{\mathbf{E}}[s(x)^2 \prod_{i \in S} p_i(x)] \geq 0$ *for all squares* $s(x)^2$ *such that* $s(x)^2 \prod_{i \in S} p_i(x)$ *has degree at most* d.

Pseudoexpectations satisfy several basic inequalities some of which are Cauchy-Schwartz, Hölder and a modified version of the triangle inequality. We will use these extensively. For details, please look at Appendix A.1.

We will also rely on the algorithmic fact that given a satisfiable system $\mathcal{A}$ of m polynomial inequalities in d variables, there is an algorithm which runs in time $(d+m)^{O(t)}$ and computes a pseudoexpectation of degree t satisfying $\mathcal{A}$. More details can be found in Appendix A.1.

3. Certifiably Bounded Moments in Sparse Directions

Our algorithm succeeds whenever the uncorrupted samples have *certifiably bounded moments*. To define this property, we first need to capture the sparsity of vectors using polynomial equations, which we do as follows:

Definition 12 *We use* $\mathcal{A}_{k\text{-sparse}}$ *to denote the following system of equations over* $v_1, \ldots, v_d, z_1, \ldots, z_d$:
$\mathcal{A}_{k\text{-sparse}} := \{z_i^2 = z_i\}_{i \in [d]} \cup \{v_i z_i = v_i\}_{i \in [d]} \cup \left\{\sum_{j=1}^d z_j = k\right\} \cup \left\{\sum_{i=1}^d v_i^2 = 1\right\}$.

A vector $v = (v_1, \ldots, v_d)$ is k-sparse if and only if there exist $z = (z_1, \ldots, z_d)$ such that v, z satisfy $\mathcal{A}_{k\text{-sparse}}$. Here, the z_i's correspond to the support of the vector v. We will also need the notion of the t-th moment of a distribution being certifiably bounded.

Definition 13 ((M,t) Certifiably Bounded Moments in k-sparse Directions) *For an* $M > 0$ *and even* $t \in \mathbb{N}$, *we say that the distribution* D *with mean* μ *satisfies* (M,t) certifiably bounded moments in k-sparse directions *if* $\mathcal{A}_{k-sparse} \left|\frac{v,z}{O(t)}\right. \mathbf{E}_{X \sim D}\left[\langle v, X - \mu\rangle^t\right]^2 \leq M^2$.

An example of such a distribution is implicit in Theorem 1.1 from Kothari and Steinhardt (2017). Their result says that if a distribution D is σ-Poincare, i.e., for all differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$, $\mathbf{Var}_{X \sim D}\left[f(X)\right] \leq \sigma^2 \mathbf{E}_{X \sim D}[\|\nabla f(X)\|_2^2]$, then it has certifiably bounded moments *in every direction* v, i.e., the appropriate inequality follows even ignoring the z constraints in $\mathcal{A}_{k\text{-sparse}}$. It can be seen (see Appendix B.1) that this class also satisfies Definition 13.

3.1. Sampling and Certifiably Bounded Moments in Sparse Directions

In this section we show that sampling from distributions with not too heavy tails preserves the property of certifiably bounded moments. The correctness of the algorithm we develop later will rely solely on this property of the samples. We show the following in Appendix B.3.

Lemma 14 *Let D be a distribution over $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq I$. Suppose that D satisfies $\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{2t}\right. \mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]^2 \leq M^2$ and D has c-subexponential tails, where c is an absolute constant c. Let $S = \{X_1, \ldots, X_m\}$ be a set of m i.i.d. samples from D with $m = (tk(\log d))^{O(t)} \max(1, M^{-2})/\epsilon^2$. Let D' be the uniform distribution over S and $\overline{\mu} := \mathbf{E}_{X\sim D'}[X]$. Then, with probability 0.9, we have that $\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{2t}\right. \mathbf{E}_{X\sim D'}\left[\langle v, X-\overline{\mu}\rangle^t\right]^2 \leq 8M^2$ and $\|\overline{\mu}-\mu\|_{2,k} \leq M^{1/t}\epsilon^{1-1/t}$.*

As a remark, the core concentration lemma used to prove this theorem (Appendix B.2) is in fact applicable to all distributions with bounded $(t^2 \log d)$ moments, not only subexponential distributions.

In the rest of the section, we give an overview of the proof of Lemma 14. The claim $\|\overline{\mu}-\mu\|_{2,k} \leq M^{1/t}\epsilon^{1-1/t}$ follows from a standard Markov inequality. We thus focus on the first claim. A crucial part of this proof is that polynomials over $\mathcal{A}_{k\text{-sparse}}$ are bounded by the square of the maximum coefficient times k^t, i.e., the following:

Lemma 15 (Polynomials of k-sparse vectors are bounded) *Let $p(v_1, \ldots, v_d) = \sum_{T\in[d]^t} a_T v_T$ be a polynomial of degree t, where the coefficients $\{a_T\}_{T\in[d]^t} \subset \mathbb{R}$ are real numbers (not variables of the SoS program), then $\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{2t}\right. p(v_1, \ldots, v_d)^2 \leq k^t \max\{a_T^2 \mid T \in [d]^t\}$.*

Since D has subexponential tails, with m samples the ℓ_∞ norm of the difference between the expected and empirical central moments is $M/\sqrt{k^t}$. We can show that in the setting of Lemma 14, with probability 0.9, $\|\mathbf{E}_{i\sim[m]}[(X_i-\overline{\mu})^{\otimes t}] - \mathbf{E}_{X\sim D}[(X-\mu)^{\otimes t}]\|_\infty \leq M/\sqrt{k^t}$. An application of Lemma 15 to $p(v) = \sum_{T\in[d]^t}\left(\mathbf{E}_{i\sim[m]}\left[X_i-\overline{\mu}\right]_T - \mathbf{E}_{X\sim D}\left[X-\mu\right]_T\right) v_T$ and the SoS triangle inequality completes the proof.

4. Robust Sparse Mean Estimation with Unknown Covariance

Given that the inliers have certifiably bounded moments in k-sparse directions (which happens with high probability because of Lemma 14), we show that our SoS algorithm finds a vector that is within $O(M^{1/t}\epsilon^{1-1/t})$ of the empirical mean of the inliers. In this section, we show the following theorem, which when combined with Lemma 14 shows Theorem 4.

Theorem 16 *Let t be a power of 2 and $\epsilon \leq \epsilon_0$ for a sufficiently small constant ϵ_0. Let $X_1, \ldots, X_m \in \mathbb{R}^d$ such that the uniform distribution $\{X_1, \ldots, X_m\}$ has (M, t) certifiably bounded moments in k-sparse directions (see Definition 13). Given ϵ, k, M, t and any ϵ-corruption of $X_1, \ldots, X_m$, Algorithm 1 runs for time $(md)^{O(t)}$ and returns a vector $\widehat{\mu}$ with $\|\widehat{\mu} - \mathbf{E}_{i\sim[m]}[X_i]\|_{2,k} = O(M^{1/t}\epsilon^{1-1/t})$.*

Additional Notation. To avoid confusion, we fix the following notation for the rest of the paper. We use $X_1, \ldots, X_m$ to denote the inlier points. Their empirical mean and covariance is denoted by $\overline{\mu}$ and $\overline{\Sigma}$ respectively. The points $Y_1, \ldots, Y_m$ are the ϵ-corrupted set of samples. We use $X'_1, \ldots, X'_m$ to denote vector-valued variables of length d for the SoS program and μ', Σ' to denote their empirical

mean and covariance. Finally, $w_1, \dots, w_m$ will be scalar-valued variables of the SoS program.

Our algorithm is based on the system of polynomial inequalities defined in Definition 17 below, which capture the following properties of the uncorrupted samples: (i) $X'_i = Y_i$ for all but ϵm indices, and (ii) The t-th moment of the uniform distribution on $\{X'_i\}_{i=1}^m$ is certifiably bounded in every k-sparse direction. Although the last constraint seems complicated, we show in Appendix A.2 that it can be expressed as $d^{O(t)}$ polynomial constraints. Finally, our algorithm SPARSE-MEAN-EST will solve a semidefinite programming (SDP) relaxation of the polynomial system $\mathcal{A}_{\text{sparse-mean-est}}$.

Definition 17 (Sparse Mean Estimation Axioms $\mathcal{A}_{\text{sparse-mean-est}}$) *Let $Y_1, \dots, Y_m \in \mathbb{R}^d$. Let $t \in \mathbb{N}$ be even and let $\delta, \epsilon > 0$. $\mathcal{A}_{\text{sparse-mean-est}}$ denotes the system of the following constraints.*

1. *Let $\mu' = \frac{1}{m}\sum_{i=1}^m X'_i$.*
2. *Let $\mathcal{A}_{\text{corruptions}} := \{w_i^2 = w_i\}_{i\in[m]} \cup \{w_i(Y_i - X'_i) = 0\}_{i\in[m]} \cup \{\sum_{i\in[m]} w_i = (1-\epsilon)m\}$.*
3. *$X'_1, \dots, X'_m$ satisfy (M, t) certifiably bounded moments in k-sparse directions (Definition 13).*

Algorithm 1 Robust Sparse Mean Estimation

function SPARSE-MEAN-EST($Y_1, \dots, Y_m, t, M, \epsilon, k$)
  Find a pseudo-expectation $\tilde{\mathbf{E}}$ of degree $10t$ which satisfies the system of Definition 17.
  return $\hat{\mu} := \tilde{\mathbf{E}}\,[\mu']$.
end function

4.1. Proof of Theorem 16

We first show that the system given in Definition 17 is feasible: Observe that the following assignments satisfy the constraints: $X'_i = X_i$, $w_i = \mathbf{1}_{(Y_i = X_i)}$, $\mu' = \frac{1}{m}\sum_i X_i$. It is easy to check that the first two constraints are satisfied. The final constraint is satisfied because of the assumption in the theorem and Fact 39.

In what follows, we assume that v is a fixed sparse vector. The proof consists of first showing that $\langle v, \overline{\mu} - \mu'\rangle^{2t} \le O(M^2\epsilon^{2t-2})$ has an SoS proof and then showing that $\tilde{\mathbf{E}}[\mu']$ also satisfies the same inequality as μ' does. We start with the first step. The program variables X'_i have constraints which ensure that a $(1-\epsilon)$ fraction of these will match the data, Y_i. The following standard claim (shown in Appendix C) shows that the program variables match a $(1-2\epsilon)$ fraction of the *uncorrupted* samples X_i. Note that in the claim below the r_i are constants, even though they are not known to the algorithm.

Claim 18 *Let $r_i := \mathbf{1}_{X_i = Y_i}$ and $W_i := w_i r_i$. There exists an SoS proof of $\{W_i^2 = W_i\}_{i=1}^m \cup \{\sum_{i=1}^m (1 - W_i) \le 2\epsilon m\} \cup \{W_i\,(X_i - X'_i) = 0\}_{i=1}^m$ from the axioms $\{W_i = w_i r_i\}_{i=1}^m \cup \mathcal{A}_{\text{corruptions}}$.*

We now work towards an upper bound on $\langle v, \overline{\mu} - \mu'\rangle^{2t}$. Let $r_i := \mathbf{1}_{X_i = Y_i}$ and $W_i := w_i r_i$ as above. We first show that there is an SoS proof for $\langle v, \overline{\mu} - \mu'\rangle^{2t} \le (2\epsilon)^{2t-2}\ \mathbf{E}_{i\sim[m]}\left[\langle v, X_i - X'_i\rangle^t\right]^2$:

$$\mathcal{A}_{\text{sparse-mean-est}} \vdash_{O(t)} \langle v, \overline{\mu} - \mu'\rangle^{2t} = \left(\mathop{\mathbf{E}}_{i\sim[m]}\left[(1 - W_i)\langle v, X_i - X'_i\rangle\right]\right)^{2t}$$

$$\leq \left(\left(\mathop{\mathbf{E}}_{i\sim[m]} [1 - W_i] \right)^{t-1} \mathop{\mathbf{E}}_{i\sim[m]} \left[\langle v, X_i - X_i' \rangle^t \right] \right)^2$$
$$\leq (2\epsilon)^{2t-2} \mathop{\mathbf{E}}_{i\sim[m]} \left[\langle v, X_i - X_i' \rangle^t \right]^2 , \tag{1}$$

where we used SoS Hölder (Fact 36) and Claim 18. Now, to bound $\mathbf{E}_{i\sim[m]} \left[\langle v, X_i - X_i' \rangle^t \right]^2$, first observe that $\langle v, X_i - X_i' \rangle = \langle v, X_i - \overline{\mu} \rangle + \langle v, \overline{\mu} - \mu' \rangle + \langle v, \mu' - X_i' \rangle$. Applying SoS triangle inequality (Fact 37) twice, we see that there is an $O(t)$-degree SoS proof of the following inequality from axioms $\mathcal{A}_{\text{sparse-mean-est}}$.

$$\mathop{\mathbf{E}}_{i\sim[m]} \left[\langle v, X_i - X_i' \rangle^t \right]^2 \leq 3^{2t+2} \left(\mathop{\mathbf{E}}_{i\sim[m]} \left[\langle v, X_i - \overline{\mu} \rangle^t \right]^2 + \langle v, \overline{\mu} - \mu' \rangle^{2t} + \mathop{\mathbf{E}}_{i\sim[m]} \left[\langle v, \mu' - X_i' \rangle^t \right]^2 \right).$$

The first and last term above can be bounded by M^2. To see this, note that the uniform distributions on $\{X_i\}_{i\in[m]}$ as well as the program variables $\{X_i'\}_{i\in[m]}$ have (M, t) certifiable bounded central moments in k-sparse directions. Putting these together, thus far we have shown that

$$\mathcal{A}_{\text{sparse-mean-est}} \vdash_{O(t)} \langle v, \overline{\mu} - \mu' \rangle^{2t} \leq (2\epsilon)^{2t-2} \cdot 3^{2t+2} \left(2M^2 + \langle v, \overline{\mu} - \mu' \rangle^{2t} \right)$$
$$\leq 6^{2t+2} \cdot \epsilon^{2t-2} \cdot \left(M^2 + \langle v, \overline{\mu} - \mu' \rangle^{2t} \right).$$

Rearranging and using the assumption that $\epsilon < 3/1000$ implies $6^{2t+2} \cdot \epsilon^{2t-2} \leq 1/2$, we get that

$$\mathcal{A}_{\text{sparse-mean-est}} \vdash_{O(t)} \langle v, \overline{\mu} - \mu' \rangle^{2t} \leq \epsilon^{2t-2} \cdot \frac{6^{2t+2} \cdot M^2}{1 - 6^{2t+2} \cdot \epsilon^{2t-2}} \leq 6^{2t+3} M^2 \, \epsilon^{2t-2}. \tag{2}$$

Finally, taking pseudoexpectations on both sides of Equation (2) and using Fact 35 (pseudoexpectation Cauchy-Schwartz), we see that $\langle v, \overline{\mu} - \tilde{\mathbf{E}}[\mu'] \rangle \leq O(M^{1/t} \epsilon^{1-1/t})$ for all k-sparse unit vectors, or equivalently $\|\overline{\mu} - \tilde{\mathbf{E}}\mu'\|_{2,k} = O(M^{1/t} \epsilon^{1-1/t})$. This completes the proof of Theorem 16.

5. Achieving Near-optimal Error for Gaussian Inliers

In this section, we show Theorem 7 by exhibiting a $(k^4/\epsilon^2)\text{polylog}(d/\epsilon)$ sample, polynomial time algorithm to estimate the mean of a multivariate Gaussian distribution in k-sparse directions. This follows from a modification of the main result of Kothari et al. (2022).

An important component of the algorithm for the sparse setting is the SoS program given by Definition 19. Our notation is as before, with the addition of Σ', which is a $d \times d$ matrix-valued indeterminate. Additionally, define $\widehat{\mu} = \tilde{\mathbf{E}}[\mu'], \widehat{\Sigma} = \tilde{\mathbf{E}}[\Sigma'], \overline{\mu} = \mathbf{E}_{i\sim[m]}[X_i], \overline{\Sigma} = \mathbf{E}_{i\sim[m]}[(X_i - \overline{\mu})(X_i - \overline{\mu})^T], Y_{ij} = \frac{1}{2}(Y_i - Y_j)(Y_i - Y_j)^T, X_{ij} = \frac{1}{2}(X_i - X_j)(X_i - X_j)^T$.

Definition 19 (Gaussian Sparse Mean Estimation Axioms $\mathcal{A}_{\text{G-sparse-mean-est}}$) *Let $Y_1, \ldots, Y_m \in \mathbb{R}^d$. Let $0 < \epsilon < 1/2$. We define $\mathcal{A}_{\text{G-sparse-mean-est}}$ to be the following constraints.*

1. $\mu' = \frac{1}{m} \sum_{i=1}^m X_i'$ *and* $\Sigma' = \frac{1}{m} \sum_{i=1}^m (X_i' - \mu')(X_i' - \mu')^T$.
2. $\mathcal{A}_{\text{corruptions}} := \{w_i^2 = w_i\}_{i\in[m]} \cup \{w_i(Y_i - X_i') = 0\}_{i\in[m]} \cup \{\sum_{i\in[m]} w_i = (1-\epsilon)m\}$.
3. $\mathcal{A}_{\text{k-sparse}} \vdash_8 \left(\mathbf{E}_{i\sim[m]} \left[\langle v, X_i' - \mu' \rangle^4 \right] - 3(v^T \Sigma' v)^2 \right)^2 \leq \tilde{O}(\epsilon^2)(v^T \Sigma' v)^4$.
4. $\mathcal{A}_{\text{k-sparse}} \vdash_2 (v^T \Sigma' v)^2 \leq 9$.

More simply, $\mathcal{A}_{\text{G-sparse-mean-est}}$ consists of constraints that capture the following: (1) $X'_i = Y_i$ for all but ϵm indices; and (2) The fourth moment of the uniform distribution on $\{X'_i\}_{i\in[m]}$ is bounded in 'k-sparse' directions. The algorithm (Algorithm 2) consists of finding a degree-12 pseudo-expectation that satisfies $\mathcal{A}_{\text{G-sparse-mean-est}}$, and estimates the sparse mean up to an error of $\tilde{O}(\epsilon)$.

Algorithm 2 Robust Sparse Mean Estimation

function SPARSE-MEAN-EST($Y_1, \ldots, Y_m, \epsilon, k$)
  Find a pseudo-expectation $\tilde{\mathbf{E}}$ of degree-12 that satisfies the program of Definition 19.
  Let $\widehat{\mu} = \tilde{\mathbf{E}}[\mu']$ and output $\widehat{\mu}$.
end function

As in Kothari et al. (2022), our result will rely on the notion of 'resilience' from Diakonikolas et al. (2016). However, instead of proving the result for *all* directions, we will instead require this only for *k-sparse* directions.

5.1. Deterministic Conditions on Inliers

We require a set of deterministic conditions similar to that in Kothari et al. (2022). However, instead of proving the relevant conditions for *all* directions, we will instead require that they hold only for *k-sparse* directions. We show that, with high probability, a set of $(k^4/\epsilon^2)\text{polylog}(d/\epsilon)$ samples drawn from $\mathcal{N}(\mu, \Sigma)$ satisfy the following set of conditions.

Lemma 20 *Let T denote the set of all $a \in [0,1]^{m\times m}$ such that (i) $a_{ij} = a_{ji}$ for all $i, j \in [m]$, (ii) $\mathbf{E}_{ij}[a_{ij}] \geq 1 - 4\epsilon$, and (iii) $\mathbf{E}_j[a_{ij}] \geq a_i(1-2\epsilon)$ for all $i \in [m]$ and $a_{ij} \leq a_i$ for all $i, j \in [m]$. Let $X_1, \ldots, X_m \sim \mathcal{N}(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and a positive definite matrix $I_d \preceq \Sigma \preceq 2I_d$. Denote $X_{ij} := (1/2)(X_i - X_j)(X_i - X_j)^T$ and $\overline{\Sigma} := \mathbf{E}_{ij}[X_{ij}]$. If $m > (k^4/\epsilon^2)\text{polylog}(d/\epsilon\gamma)$, then, with probability $1-\gamma$ we have that the following hold for all $v \in \mathcal{U}_k$:*

1. $|\langle v, \overline{\mu} - \mu\rangle| \leq \tilde{O}(\epsilon)\sqrt{v^T\Sigma v}$.
2. $|\mathbf{E}_{i\sim[m]}[a_i\langle v, X_i - \overline{\mu}\rangle]| \leq \tilde{O}(\epsilon)\sqrt{v^T\overline{\Sigma}v}$.
3. $\left|\mathbf{E}_{i\sim[m]}\left[a_i\left(\langle v, X_i - \overline{\mu}\rangle^2 - v^T\overline{\Sigma}v\right)\right]\right| \leq \tilde{O}(\epsilon)v^T\overline{\Sigma}v$.
4. $|v^T(\overline{\Sigma} - \Sigma)v| \leq \tilde{O}(\epsilon)v^T\Sigma v$.
5. $|\mathbf{E}_{i,j\sim[m]}[a_{ij}(v^TX_{ij}v - v^T\overline{\Sigma}v)]| \leq \tilde{O}(\epsilon)v^T\overline{\Sigma}v$.
6. $|\mathbf{E}_{i,j\sim[m]}[a_{ij}((v^TX_{ij}v - v^T\overline{\Sigma}v)^2 - 2(v^T\overline{\Sigma}v)^2)]| \leq \tilde{O}(\epsilon)(v^T\overline{\Sigma}v)^2$.

The proof of this lemma is provided in Appendix D.1. In order to show that the program of Definition 19 is feasible, we first need to argue that after taking enough samples, the empirical fourth moment of Gaussian is certifiably close to its distributional value. This is a consequence of the results of Section 3 and the assumption that $\Sigma \succeq I$. For the proof, see Appendix D.

Lemma 21 *Let $X_1, \ldots, X_m \sim \mathcal{N}(\mu, \Sigma)$ for a k-sparse vector $\mu \in \mathbb{R}^d$ and a $d \times d$ symmetric matrix $I_d \preceq \Sigma \preceq 2I_d$. Let $\overline{\mu}$ and $\overline{\Sigma}$ be the empirical mean and covariance respectively. If the number of samples $m > (k^4/\epsilon^2)\text{poly}\log(d, 1/\gamma, 1/\epsilon)$, then, with probability at least $1-\gamma$, we have that $\mathcal{A}_{k\text{-sparse}} \left\vdash^{v,z}_{8} \left(\mathbf{E}_{i\sim[m]}\left[\langle v, X_i - \overline{\mu}\rangle^4\right] - 3(v^T\overline{\Sigma}v)^2\right)^2 \leq \tilde{O}(\epsilon^2)(v^T\overline{\Sigma}v)^4$.*

As a corollary, we establish the feasibility of the system of Definition 19 in Appendix D.1.3.

5.2. Proof of Theorem 7

In this section we prove Theorem 7 under the assumption that $I \preceq \Sigma \preceq 2I$, since we have to use Lemmata 20 and 21. As explained in Section 1.2.2, this assumption is removed in Appendix D.4, where the proof of Theorem 7 for arbitrary Σ is completed.

We now outline the proof for the case $I \preceq \Sigma \preceq 2I$ (Theorem 59 in Appendix), deferring proofs of intermediate lemmata to Appendix D.2. We first condition on the conclusion of Lemma 20, which holds high probability. Further, by the discussion at the end of Section 5.1, we know that the program is feasible. The first step is to show that our theorem holds given that $\tilde{\mathbf{E}}[\Sigma']$ is a good enough approximation of Σ.

Lemma 22 *Let $Y_1, \dots, Y_m$ be an ϵ-corruption of the set $X_1, \dots, X_m$, satisfying Items 2 and 3 of Lemma 20. Let $\tilde{\mathbf{E}}$ be a degree-6 pseudo-expectation in variables w_i, X_i', Σ', μ' satisfying the system of Definition 19. Denote by $\overline{\mu}, \overline{\Sigma}$ the empirical mean and covariance of $X_1, \dots, X_m$ and let $\widehat{\Sigma} := \tilde{\mathbf{E}}[\Sigma']$. Then, for all $v \in \mathcal{U}_k$ it holds $|\langle v, \widehat{\mu} - \overline{\mu}\rangle| \le \tilde{O}(\epsilon)\sqrt{v^T\overline{\Sigma}v} + \sqrt{O(\epsilon)v^T(\widehat{\Sigma} - \overline{\Sigma})v + \tilde{O}(\epsilon^2)v^T(\widehat{\Sigma} + \overline{\Sigma})v}$.*

It now suffices to show that $|v^T(\widehat{\Sigma} - \overline{\Sigma})v| \le \tilde{O}(\epsilon)v^T\overline{\Sigma}v$ since Lemma 22 combined with Items 1 and 4 of Lemma 20 implies that $|\langle v, \widehat{\mu} - \mu\rangle| \le \tilde{O}(\epsilon)\sqrt{v^T\Sigma v} \le \tilde{O}(\epsilon)$ and thus proves our main theorem. Thus, we focus on showing that $|v^T(\widehat{\Sigma} - \overline{\Sigma})v| \le \tilde{O}(\epsilon)v^T\Sigma v$ for all $v \in \mathcal{U}_k$.

Lemma 23 *Let $Y_1, \dots, Y_m$ be an ϵ-corruption of $X_1, \dots, X_m$ satisfying Items 5 and 6 of Lemma 20. Let $\tilde{\mathbf{E}}$ be a degree-12 pseudo-expectation in variables w_i, X_i', Σ', μ' satisfying the system of Definition 19. Define $Y_{ij} = (1/2)(Y_i - Y_j)(Y_i - Y_j)^T$, $X_{ij} = (1/2)(X_i - X_j)(X_i - X_j)^T$, $X_{ij}' = (1/2)(X_i' - X_j')(X_i' - X_j')^T$, $\widehat{\Sigma} = \tilde{\mathbf{E}}[\Sigma']$, $w_{ij}' = w_i w_j \mathbf{1}(X_{ij} = Y_{ij})$, and $R = \tilde{\mathbf{E}}[\mathbf{E}_{ij}[(1 - w_{ij}')v^T(X_{ij}' - \overline{\Sigma})v]^2]$. Then, for every $v \in \mathcal{U}_k$, we have that, $|v^T(\widehat{\Sigma} - \overline{\Sigma})v| \le \tilde{O}(\epsilon)v^T\overline{\Sigma}v + \sqrt{R}$ and $R \le O(\epsilon)(\tilde{\mathbf{E}}[(v^T\Sigma' v)^2] - (v^T\overline{\Sigma}v)^2) + \tilde{O}(\epsilon)(\tilde{\mathbf{E}}[(v^T\Sigma' v)^2] + (v^T\overline{\Sigma}v)^2)$.*

The final part of the proof (for the setting $I \preceq \Sigma \preceq 2I$) is identical to Kothari et al. (2022) and is provided in Appendix D.3 for completeness. It consists of showing that $R = \tilde{O}(\epsilon^2)(v^T\overline{\Sigma}v)^2$. Finally, we remove the condition $I \preceq \Sigma \preceq 2I$ in Appendix D.4 and obtain the result for general Σ with the error scaling as $\tilde{O}(\epsilon\sqrt{\|\Sigma\|_2})$.

Acknowledgments

Ilias Diakonikolas was supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. Daniel M. Kane was supported by NSF Medium Award CCF-2107547, NSF Award CCF-1553288 (CAREER), a Sloan Research Fellowship, and a grant from CasperLabs. Sushrut Karmalkar was supported by NSF under Grant #2127309 to the Computing Research Association for the CIFellows 2021 Project. Ankit Pensia was supported by NSF grants NSF Award CCF-1652862 (CAREER), DMS-1749857, and CCF-1841190. Thanasis Pittas was supported by NSF Award CCF-1652862 (CAREER). The authors would also like to thank Pravesh Kothari for useful clarifications regarding prior work.

References

G. Aubrun and S. J. Szarek. *Alice and Bob meet Banach*, volume 223. American Mathematical Soc., 2017.

A. Bakshi and A. Prasad. Robust linear regression: Optimal rates in polynomial time. In *ACM SIGACT Symposium on Theory of Computing (STOC 2021)*, pages 102–115. ACM, 2021. doi: 10.1145/3406325.3451001.

A. Bakshi, I. Diakonikolas, S. B. Hopkins, D. Kane, S. Karmalkar, and P. K. Kothari. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 149–159. IEEE, 2020a.

A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari, and S. S. Vempala. Robustly learning mixtures of k arbitrary gaussians. *CoRR*, abs/2012.02119, 2020b. URL https://arxiv.org/abs/2012.02119.

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

B. Barak and D. Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. 1, 2016. URL http://www.sumofsquares.org/public/index.html.

L. Birgé. An alternative point of view on lepski's method. *Lecture Notes-Monograph Series*, 2001.

I. M. Bomze. On standard quadratic optimization problems. *Journal of Global Optimization*, 13(4), 1998.

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

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

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 10 2018.

Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. *CoRR*, abs/1811.09380, 2018. URL http://arxiv.org/abs/1811.09380. Conference version in SODA 2019, p. 2755-2771.

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

Y. Cheng, I. Diakonikolas, D. M. Kane, R. Ge, S. Gupta, and M. Soltanolkotabi. Outlier-robust sparse estimation via non-convex optimization. *CoRR*, abs/2109.11515, 2021.

J. Depersin and G. Lecue. Robust subgaussian estimation of a mean vector in nearly linear time. *CoRR*, abs/1906.03058, 2019.

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

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

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

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

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart. Sever: A robust meta-algorithm for stochastic optimization. *CoRR*, abs/1803.02815, 2018a. URL http://arxiv.org/abs/1803.02815. Conference version in ICML 2019.

I. Diakonikolas, D. M. Kane, and A. Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1047–1060, 2018b. Full version available at https://arxiv.org/abs/1711.07211.

I. Diakonikolas, S. Karmalkar, D. Kane, E. Price, and A. Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *Advances in Neural Information Processing Systems 33, NeurIPS 2019*, 2019a.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 2745–2754, 2019b.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proc. 30th Annual Symposium on Discrete Algorithms (SODA)*, pages 2745–2754, 2019c.

I. Diakonikolas, S. B. Hopkins, D. Kane, and S. Karmalkar. Robustly learning any clusterable mixture of gaussians. *CoRR*, abs/2005.06417, 2020a. URL https://arxiv.org/abs/2005.06417.

I. Diakonikolas, D. Kane, and D. Kongsgaard. List-decodable mean estimation via iterative multi-filtering. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020b.

I. Diakonikolas, D. M. Kane, and A. Pensia. Outlier Robust Mean Estimation with Subgaussian Rates via Stability. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, 2020c.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustness meets algorithms. *Commun. ACM*, 64(5):107–115, 2021a.

I. Diakonikolas, D. Kane, D. Kongsgaard, J. Li, and K. Tian. List-decodable mean estimation in nearly-pca time. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 10195–10208, 2021b.

I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian. Clustering mixture models in almost-linear time via list-decodable mean estimation. *CoRR*, abs/2106.08537, 2021c.

Y. Dong, S. B. Hopkins, and J. Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *CoRR*, abs/1906.11366, 2019. URL http://arxiv.org/abs/1906.11366. Conference version in NeurIPS 2019.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of STOC'13*, pages 655–664, 2013. Full version in Journal of the ACM, 2017.

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

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

S. B. Hopkins. Clustering and sum of squares proofs: Six blog posts on unsupervised learning. 2018.

S. B. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

S. B. Hopkins, J. Li, and F. Zhang. Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, 2020.

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

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

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

A. Klivans, P. Kothari, and R. Meka. Efficient algorithms for outlier-robust regression. In *Proc. 31st Annual Conference on Learning Theory (COLT)*, pages 1420–1430, 2018.

P. K. Kothari and J. Steinhardt. Better agnostic clustering via relaxed tensor norms. *CoRR*, abs/1711.07465, 2017.

P. K. Kothari and D. Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017.

P. K. Kothari, P. Manohar, and B. H. Zhang. Polynomial-time sum-of-squares can robustly estimate mean and covariance of gaussians optimally. In *International Conference on Algorithmic Learning Theory*, pages 638–667. PMLR, 2022.

D. Kunisky, A. S. Wein, and A. S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

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

J. B. Lasserre. New positive semidefinite relaxations for nonconvex quadratic programs. In *Advances in Convex Analysis and Global Optimization*, pages 319–331. Springer, 2001.

M. Laurent. *Sums of squares, moment matrices and optimization over polynomials*, pages 155–270. Number 149 in The IMA Volumes in Mathematics and its Applications Series. Springer Verlag, Germany, 2009. ISBN 9780387096858.

Z. Lei, K. Luh, P. Venkat, and F. Zhang. A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory, COLT 2020*, 2020.

O. V. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

J. Li. *Principled Approaches to Robust Machine Learning and Beyond*. PhD thesis, Massachusetts Institute of Technology, 2018.

A. Liu and A. Moitra. Settling the robust learnability of mixtures of gaussians. *CoRR*, abs/2011.03622, 2020.

G. Lugosi and S. Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393 – 410, 2021.

Y. Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, pages 405–440. Springer, 2000.

A. Pensia, V. Jog, and P. Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *CoRR*, abs/2009.12976, 2020.

A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3): 601–627, July 2020. ISSN 13697412. doi: 10.1111/rssb.12364.

N.Z. Shor. Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences*, 1987.

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

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

Appendix A. Omitted Background

Basic Notation We use $\mathbb{N}$ to denote natural numbers and $\mathbb{Z}_+$ to denote positive integers. For $n \in \mathbb{Z}_+$ we denote $[n] := \{1, \ldots, n\}$. We denote by $\mathbf{1}(\mathcal{E})$ the indicator function of the event $\mathcal{E}$. For $a_1(x), \ldots, a_d(x)$ polynomials in x and an ordered tuple $T \in [d]^t$, we use $a_T(x)$ to define the polynomial $a_T(x) := \prod_{i \in T} a_i(x)$. We denote by $\mathbb{R}[x_1, \ldots, x_d]_{\leq t}$ the class of real-valued polynomials of degree at most t in variables $x_1, \ldots, x_d$. We use $\mathrm{poly}(\cdot)$ to indicate a quantity that is polynomial in its arguments. Similarly, $\mathrm{polylog}(\cdot)$ denotes a quantity that is polynomial in the logarithm of its arguments. For an ordered set of variables $V = \{x_1, \ldots, x_n\}$, we will denote $p(V)$ to mean $p(x_1, \ldots, x_n)$.

Linear Algebra Notation We use I_d to denote the $d \times d$ identity matrix. We will drop the subscript when it is clear from the context. We typically use small case letters for deterministic vectors and scalars. We will specify the dimensionality unless it is clear from the context. We denote by $e_1, \ldots, e_d$ the vectors of the standard orthonormal basis, i.e., the j-th coordinate of e_i is equal to $\mathbf{1}_{\{i=j\}}$, for $i, j \in [d]$. We use $\mathcal{S}^{d-1}$ to denote the d-dimensional unit sphere. For a vector v, we let $\|v\|_2$ denote its ℓ_2-norm. We call a vector k-sparse if it has at most k non-zero coordinates. We define the set of k-sparse d-dimensional unit-norm vectors as $\mathcal{U}_k^d := \{x \in \mathbb{R}^d : x \text{ is } k\text{-sparse}, \|x\|_2 = 1\}$. We will often drop the superscript when it is clear from the context. We use $\langle v, u \rangle$ for the inner product of the vectors u, v. For a matrix A, we use $\|A\|_F, \|A\|_2, \|A\|_\infty$ to denote the Frobenius, spectral, and entry-wise infinity-norm. We denote the trace of A by $\mathrm{tr}(A)$ and the number of nonzero entries in A by $\|A\|_0$. For two matrices $A, B \in \mathbb{R}^{m \times d}$, we define the inner product $\langle A, B \rangle := \mathrm{tr}(A^T B)$. For a matrix $A \in \mathbb{R}^{d \times d}$, we use $A^\flat$ to denote the flattened vector in $\mathbb{R}^{d^2}$, and for a $v \in \mathbb{R}^{d^2}$, we use $v^\sharp$ to denote the unique matrix A such that $A^\flat = v^\sharp$. We say a symmetric matrix A is PSD (positive semidefinite) and write $A \succeq 0$ if $x^T A x \geq 0$ for all vectors x. We write $A \preceq B$ when $B - A$ is PSD. We will use $\cdot^{\otimes s}$ to denote the standard Kronecker product.

Probability Notation We use capital letters for random variables. For a random variable X, we use $\mathbf{E}[X]$ for its expectation. We use $\mathcal{N}(\mu, \Sigma)$ to denote the Gaussian distribution with mean μ and covariance matrix Σ. We let ϕ denote the pdf of the one-dimensional standard Gaussian. When D is a distribution, we use $X \sim D$ to denote that the random variable X is distributed according to D. When S is a set, we let $\mathbf{E}_{X \sim S}[\cdot]$ denote the expectation under the uniform distribution over S. For any sequence $a_1, \ldots, a_m \in \mathbb{R}^d$, we will also use $\mathbf{E}_{i \sim [m]}[a_i]$ to denote $\frac{1}{m} \sum_{i \in [m]} a_i$. For a real-valued random variable X and $p \geq 1$, we use $\|X\|_{L_p}$ to denote its L_p norm, i.e., $\|X\|_{L_p} := (\mathbf{E}[|X|^p])^{1/p}$.

The following fact can be used to translate bounds from the $(2, k)$-norm to the usual ℓ_2-norm when the underlying mean μ is sparse:

Fact 9 *Let $h_k : \mathbb{R}^d \to \mathbb{R}^d$ denote the function where $h_k(x)$ is defined to truncate x to its k largest coordinates in magnitude and zero out the rest. For all $\mu \in \mathcal{U}_k$, $\|h_k(x) - \mu\|_2 \leq 3\|x - \mu\|_{2,k}$.*

Proof Let $\|x - \mu\|_{2,k} = b$. Let $S^* := \mathrm{supp}(\mu)$ and $S' := \mathrm{supp}(h_k(x))$. Then, $\|(\mu - h_k(x))_{S^*}\|_2 \leq B$ and $\|(x)_{S' \setminus S^*}\|_2 = \|(\mu - h_k(x))_{S' \setminus S^*}\|_2 \leq b$.

If $h_k(x) = x$, then we are done because $\|(\mu - x)_{(S' \setminus S^*) \cup S^*}\|_2 \leq 2b$. If not, then $|S'| = k$. Since $|S^*| \leq k$, $|S' \setminus S^*| \geq |S^* \setminus S'|$. Since S' contains the indices for the k largest in magnitude entries of x, for any $i \in S' \setminus S^*$ and $j \in S^* \setminus S'$, $|x_i| \geq |x_j|$. Since $\|(x)_{S' \setminus S^*}\|_2 \leq b$, at least one coordinate $j \in S' \setminus S^*$ must satisfy $(x_j)^2 \leq b^2 / |S' \setminus S^*|$. Therefore, for every $i \in S^* \setminus S'$ we have

$(x_i)^2 \leq b^2/|S' \setminus S^*|$. Adding these up we get the following upper bound on $\|(x)_{S^* \setminus S'}\|_2$.

$$\|(x)_{S^* \setminus S'}\|_2^2 = \sum_{i \in S^* \setminus S'} (x)_i^2 \leq b^2 \cdot \frac{|S^* \setminus S'|}{|S' \setminus S^*|} \leq b^2.$$

Finally, we have that

$$\|\mu - h_k(x)\|_2^2 = \|(\mu - x)_{S' \cap S^*}\|_2^2 + \|(\mu)_{S^* \setminus S'}\|_2^2 + \|(x)_{S' \setminus S^*}\|_2^2 \leq 6b^2,$$

where the bound on $\|(\mu)_{S^* \setminus S'}\|^2$ follows by a triangle inequality and the fact $\|(\mu - x)_{S^*}\|_2 \leq b$. ■

A.1. SoS Preliminaries

The following notation and preliminaries are specific to the SoS part of this paper. We refer the reader to Barak and Steurer (2016) for a complete treatment of basic definitions about the SoS hierarchy and SoS proofs. Here we review the basics. Our algorithms will work under the condition that the numerical precision of all the numbers involved is controlled. To describe these conditions formally, we use the standard notion of bit complexity, defined below for completeness.

Definition 24 (Bit complexity) *The bit complexity of an integer $z \in \mathbb{Z}$ is $1 + \lceil \log_2 z \rceil$. The bit complexity of a rational number r/t is the sum of the individual bit complexities of r and t. The bit complexity of a vector is the sum of the bit complexities of its coordinates and the bit complexity of a set of vectors is the sum of the bit complexities of the set's elements.*

Definition 25 (Symbolic polynomial) *A degree-t symbolic polynomial p is a collection of indeterminates $\widehat{p}(\alpha)$, one for each multiset $\alpha \subseteq [d]$ of size at most t. We think of it as representing a polynomial $p : \mathbb{R}^d \to \mathbb{R}$ whose coefficients are themselves indeterminates via $p(x) = \sum_{\alpha \subseteq [d], |\alpha| \leq t} \widehat{p}(\alpha) x^\alpha$.*

Definition 26 (SoS Proof) *Let $x_1, \ldots, x_d$ be indeterminates and let $\mathcal{A}$ be a set of polynomial inequalities $\{p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$. An SoS proof of the inequality $r(x) \geq 0$ from axioms $\mathcal{A}$ is a set of polynomials $\{r_S(x)\}_{S \subseteq [m]}$ such that each r_S is a sum of square polynomials and $r(x) = \sum_{S \subseteq [m]} r_S(x) \prod_{i \in S} p_i(x)$. If the polynomials $r_S(x) \cdot \prod_{i \in S} p_i(x)$ have degree at most t for all $S \subseteq [m]$, we say that this proof is of degree t and denote it by $\mathcal{A} \left|\frac{}{t}\right. r(x) \geq 0$. The bit complexity of the SoS proof is the sum of the bit complexities of the coefficients of the polynomials r_S and p_i.*

When we need to emphasize what indeterminates are involved in a particular SoS proof, we denote it by $\mathcal{A} \left|\frac{x}{t}\right. r(x) \geq 0$. When $\mathcal{A}$ is empty, we directly write $\left|\frac{}{t}\right. r(x) \geq 0$ and $\left|\frac{x}{t}\right. r(x) \geq 0$. We also often refer to $\mathcal{A}$ containing polynomial equations $q(x) = 0$, by which we mean that $\mathcal{A}$ contains both $q(x) \geq 0$ and $q(x) \leq 0$.

We frequently compose SoS proofs without comment — see Barak and Steurer (2016) for basic facts about composition of SoS proofs and bounds on the degree of the resulting proofs.

Our algorithm also uses the dual objects to SoS proofs, commonly called *pseudoexpectations*.

Definition 27 (Pseudoexpectation) *Let $x_1, \ldots, x_d$ be indeterminates. A degree-t pseudoexpectation $\tilde{\mathbf{E}}$ is a linear map $\tilde{\mathbf{E}} : \mathbb{R}[x_1, \ldots, x_d]_{\leq t} \to \mathbb{R}$ from degree-t polynomials to $\mathbb{R}$ such that $\tilde{\mathbf{E}}\left[p(x)^2\right] \geq 0$ for any p of degree at most $t/2$ and $\tilde{\mathbf{E}}\left[1\right] = 1$. If $\mathcal{A} = \{p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$*

is a set of polynomial inequalities, we say that $\tilde{\mathbf{E}}$ satisfies $\mathcal{A}$ if for every $S \subset [m]$, the following holds: $\tilde{\mathbf{E}}[s(x)^2 \prod_{i \in S} p_i(x)] \geq 0$ for all squares $s(x)^2$ such that $s(x)^2 \prod_{i \in S} p_i(x)$ has degree at most t.

We say that a pseudoexpectation is τ-approximate if it satisfies all the conditions up to slack τ, i.e., $\tilde{\mathbf{E}}[p(x)^2] \geq -\tau \|p\|_2^2$ for any p of degree at most $t/2$ and $\tilde{\mathbf{E}}[s(x)^2 \prod_{i \in S} p_i(x)] \geq -\tau \|s^2\|_2 \prod_{i \in S} \|p_i\|_2$ for all sets S and polynomials $s(x)$ such that $s(x)^2 \prod_{i \in S} p_i(x)$ has degree at most t, where $\|p\|_2$ denotes the ℓ_2-norm of the vector of coefficients of p.

We will also rely on the algorithmic fact that given a satisfiable system $\mathcal{A}$ of m polynomial inequalities in d variables, there is an algorithm which runs in time $(dm)^{O(t)}$ and computes a pseudoexpectation of degree t approximately satisfying $\mathcal{A}$.

Theorem 28 (The SoS Algorithm Shor (1987); Lasserre (2001); Nesterov (2000); Bomze (1998)) *Let $\mathcal{A}$ be a satisfiable system of m polynomial inequalities in variables $x_1, \ldots, x_d$, each with coefficients having bit complexity at most B and degree at most t. Suppose that $\mathcal{A}$ contains an inequality of the form $\|x\|_2^2 \leq M$, with M having bit complexity at most B. There is an algorithm which takes $t \in \mathbb{Z}_+$, τ, and B and returns in time $\mathrm{poly}(B, \log(1/\tau), d^t, m^t)$ a degree-t pseudo-expectation $\tilde{\mathbf{E}}$ which satisfies $\mathcal{A}$ up to error τ.*

All of our SoS proofs will be of bit complexity $\mathrm{poly}(m^t, d^t)$. We thus apply Theorem 28 with $B = \mathrm{poly}(m^t, d^t)$ and $\tau = 2^{-\mathrm{poly}(tB)}$ to ensure that the total error that we incur is at most $O(2^{-md})$. Since this error is negligible, we will not treat it explicitly in the remainder of the paper.

Pseudoexpectations satisfy several basic inequalities, some of which are Cauchy-Schwartz, Hölder and a modified version of the triangle inequality. We will use these extensively. See Appendix A.1 for details.

Fact 29 (Moments of Gaussian) *For* any $v \in \mathbb{R}^d$ *and any $s \in \mathbb{N}$, the moments of $\mathcal{N}(\mu, \Sigma)$ are $\mathbf{E}_{X \sim \mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu \rangle^{2s}\right] = (2s-1)!! \, \mathbf{E}_{X \sim \mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu \rangle^2\right]^s$.*

For two polynomials p, q, the notation $p \preceq q$ means that $q - p$ is a sum of square polynomials.

Fact 30 *Any degree-t polynomial $r(x)$ in d variables which is a sum of square polynomials, can always be written as a sum of at most $d^{t/2}$ square polynomials.*

Proof *Let $r(x) = \sum_j q_j(x)^2$. Observe that $q_j(x) = \langle u_j, m(x) \rangle$ where $m(x)$ is the vector of all possible monomials up to degree $t/2$ of the variables $x_1, \ldots, x_d$ and u_j is the vector containing the coefficients used for each of them in the polynomial q_j. Let $\sum_j u_j u_j^T = U$, then $r(x) = m(x)^T U \, m(x)$. Note U is a positive semidefinite matrix. It therefore has an eigen-decomposition of at most $d^{t/2}$ vectors $v_1, \ldots, v_{d^{t/2}}$ with eigenvalues $\lambda_1, \ldots, \lambda_{d^{t/2}} \geq 0$. This means that we can write $r(x) = \sum_{j=1}^{d^{t/2}} \lambda_j m(x)^T v_j v_j^T m(x) = \sum_{j=1}^{d^{t/2}} h_j(x)^2$ where $h_j(x) = \sqrt{\lambda_j} \, \langle v_j, m(x) \rangle$.* ■

Definition 31 (Symbolic polynomial) *A degree-t symbolic polynomial p is a collection of indeterminates $\widehat{p}(\alpha)$, one for each multiset $\alpha \subseteq [d]$ of size at most t. We think of it as representing a polynomial $p : \mathbb{R}^d \to \mathbb{R}$ whose coefficients are themselves indeterminates via $p(x) = \sum_{\alpha \subseteq [d], |\alpha| \leq t} \widehat{p}(\alpha) x^\alpha$.*

The following fact is a simple corollary of the fundamental theorem of algebra:

Fact 32 *For any univariate degree d polynomial $p(x)$, with $p(x) \geq 0$ for all $x \in \mathbb{R}$, $\left|\frac{x}{t}\right. \{p(x) \geq 0\}$.*

This can be extended to univariate polynomial inequalities over intervals of $\mathbb{R}$.

Fact 33 (Fekete and Markov-Lukács, see Laurent (2009)) *For any univariate degree d polynomial $p(x) \geq 0$ for $x \in [a,b]$, $\{x \geq a, x \leq b\} \left|\frac{x}{d}\right. \{p(x) \geq 0\}$.*

We will rely on the following algorithmic fact (Lasserre, 2001; Nesterov, 2000; Bomze, 1998; Kothari and Steurer, 2017).

Theorem 34 (The SoS Algorithm) *There is an algorithm which takes a natural number t and a satisfiable system of m polynomial inequalities $\mathcal{A}$ in variables $x_1, \ldots, x_d$ with coefficients having at most $\mathrm{poly}(m,d)$ bit complexity, containing an inequality of the form $\|x\|_2^2 \leq M$ for some real number M also having bit complexity $\mathrm{poly}(m,d)$ and returns in time $(d+m)^{O(t)}$ a degree-t pseudo-expectation $\tilde{\mathbf{E}}$ which satisfies $\mathcal{A}$ up to error 2^{-d}.*

We did not define what it means for $\tilde{\mathbf{E}}$ to satisfy $\mathcal{A}$ up to error 2^{-d}. The idea is that 2^{-d} slack is added to each constraint. Since the coefficients in all the SoS proofs in this paper have magnitude at most $d^{O(1)}$, these 2^{-d} errors are negligible and we will not treat them explicitly. See Barak and Steurer (2016) for further discussion. We will also need the following facts about SoS proofs:

We record some additional facts that we will use in our proofs.

Fact 35 (Cauchy-Schwarz for Pseudoexpectations) *Let f, g be polynomials of degree at most t. Then, for any degree-$2t$ pseudoexpectation $\tilde{\mathbf{E}}$, $\tilde{\mathbf{E}}[fg] \leq \sqrt{\tilde{\mathbf{E}}[f^2]}\sqrt{\tilde{\mathbf{E}}[g^2]}$. Consequently, for every squared polynomial p of degree t, and k a power of two, $\tilde{\mathbf{E}}[p^k] \geq (\tilde{\mathbf{E}}[p])^k$ for every $\tilde{\mathbf{E}}$ of degree-$2tk$.*

Fact 36 (SoS Cauchy-Schwartz and Hölder (see, e.g., Hopkins (2018))) *Let $f_1, g_1, \ldots, f_n, g_n$ be indeterminates over $\mathbb{R}$. Then,*

$$\left|\frac{f_1,\ldots,f_n,g_1,\ldots,g_n}{2}\right. \left\{ \left(\frac{1}{n}\sum_{i=1}^n f_i g_i\right)^2 \leq \left(\frac{1}{n}\sum_{i=1}^n f_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n g_i^2\right) \right\} .$$

The total bit complexity of the SoS proof is $\mathrm{poly}(n)$. Moverover, if $p_1, \ldots, p_n$ are indeterminates, for any $t \in \mathbb{Z}_+$ that is a power of 2, we have that

$$\{w_i^2 = w_i \mid i \in [n]\} \left|\frac{p_1,\ldots,p_n}{O(t)}\right. \left(\sum_i w_i p_i\right)^t \leq \left(\sum_{i\in[n]} w_i\right)^{t-1} \cdot \sum_{i\in[n]} p_i^t \quad \text{and}$$

$$\{w_i^2 = w_i \mid i \in [n]\} \left|\frac{p_1,\ldots,p_n}{O(t)}\right. \left(\sum_i w_i p_i\right)^t \leq \left(\sum_{i\in[n]} w_i\right)^{t-1} \cdot \sum_{i\in[n]} w_i p_i^t .$$

The total bit complexity of the SoS proof is $\mathrm{poly}(n^t)$.

Fact 37 (SoS Triangle Inequality) *If k is a power of two, $\left|\frac{a_1,a_2,\ldots,a_n}{k}\right. \left\{\left(\sum_i a_i\right)^k \leq n^k \left(\sum_i a_i^k\right)\right\}$. The total bit complexity of the SoS proof is $\mathrm{poly}(n^k)$.*

We will apply the above facts in a way so that the final bit complexity of these SoS proofs will be bounded by $\mathrm{poly}(m^t, d^t)$.

A.2. Quantifier Elimination

In this section, we describe a set of constraints that guarantee that the variables of a given SoS program satisfy a certain polynomial inequality *for all* (possibly infinite) values of some subset of the given variables, i.e., essentially leave a desired subset of the variables free. This is particularly useful to us since we would like to ensure that our samples have certifiably bounded moments in *all* k-sparse directions. Concretely, let V be the set of variables, let $F \subseteq V$ be the set of free variables, $\mathcal{A}$ be a set of polynomial constraints on F, and let $b \in \mathbb{R}[V]$. Suppose we like to ensure that $b(V) \geq 0$ for all values of F that satisfy $\mathcal{A}$. The basic idea here is to observe that it is enough to ensure that there is an SoS proof of this inequality in the variables F, and that this proof can be obtained by ensuring that a certain list of polynomials exist whose coefficients satisfy specific equalities. Hence it is sufficient to add a list of polynomial equality constraints. These constraints will become clearer in the following discussion.

We will need the following notation: if $a_1(x), \ldots, a_d(x)$ are polynomials in x and $T \in [d]^t$ is an ordered tuple, $a_T(x)$ is defined to be $a_T(x) := \prod_{i \in T} a_i(x)$. Also, let $d, t \in \mathbb{N}$ and $V := \{x_1, \ldots, x_d\}$ be formal variables and let $b \in \mathbb{R}[x_1, \ldots, x_d]$ of degree at most t.

We are now ready to provide the details below. Define the following:

1. Let $F \subset V$ denote the subset of variables that we would like to leave free.
2. Let $\mathcal{A} = \{a_1, \ldots, a_r\} \subset \mathbb{R}[F]$ be a set of polynomials in F of degree at least 1. Suppose the variables F satisfy $\{a(F) = 0 \mid a \in \mathcal{A}\}$.

Consider an assignment π to the variables $V \setminus F$. We define $b_\pi(F)$ to be the polynomial that is obtained by assigning the variables in $V \setminus F$ in $b(V)$ according to the assignment π. We know from Definition 26 that $\{a \geq 0 \mid a \in \mathcal{A}\} \left|\frac{F}{t}\right. b_\pi(F) \geq 0$, if and only if

$$b_\pi(F) = \sum_{T \subset [r], |T| \leq t} a_T(F) q_T(F),$$

where each q_T is a sum of D square polynomials, where by Fact 30 we can assume that $D < |F|^{O(t)}$. If the constraints are instead $\{a = 0 \mid a \in \mathcal{A}\}$, then the condition can be changed to

$$b_\pi(F) = \sum_{i \in [r]} a_i(F) p_i(F) + q(F), \tag{3}$$

where each p_i is an arbitrary polynomial in F and q is a sum of at most D square polynomials in F for $D < |F|^{O(t)}$, and the degree of each term on the right-hand side is at most t. In the context of our paper, $\mathcal{A}$ above will be a set of polynomial equalities which are satisfied only by sparse vectors.

Definition 38 (Quantifier Elimination) *Let $d, t \in \mathbb{N}$ and $V := \{x_1, \ldots, x_d\}$ be formal variables and let $b \in \mathbb{R}[x_1, \ldots, x_d]$ of degree at most t. Let $F \subset V$ and $\mathcal{A} = \{a_1, \ldots, a_r\} \subset \mathbb{R}[F]$ polynomial axioms of degree at least 1. We define **cons**$_F(\mathcal{A}, \{b\}, t)$ to be the set of equality constraints that equate the coefficients of F of the polynomials in Equation (3), where the coefficients may involve polynomials of $V \setminus F$. This is done by introducing variable vectors $\{P_i \mid i \in [r]\}$ for the coefficients of p_i and $\{Q_j \mid j \in [D]\}$ for the coefficients of q in Equation (3) (where $D < |F|^{O(t)}$) and equating the coefficients of the LHS and RHS when both sides are interpreted to be polynomials in F. This*

leads to at most $|F|^{O(t)}$ *many equality constraints in the variables* $\{x_i \mid i \in V \setminus F\}$, $\{P_i \mid i \in [r]\}$, $\{Q_j \mid j \in [D]\}$, *and each* P_i *and* Q_j *is of dimension at most* $|F|^{O(t)}$. [2]

The following fact from Kothari and Steurer (2017) allows us to effectively use the constraints defined above.

Fact 39 *In the setting of Definition 38, for any fixed* $F \subset V$ *and fixed assignment* π *to* $V \setminus F$, *we can extend this assignment to a solution of* $\mathbf{cons}_F(\mathcal{A}, \{b\}, t)$ *iff* $\mathcal{A} \vdash_t^F \{b_\pi(F) \geq 0\}$, *where* $b_\pi \in \mathbb{R}[F]$ *is obtained by assigning the variables in* $V \setminus F$ *in* $b(V)$ *according to the assignment* π.

Fact 40 *Consider the setting in Definition 38. Let* $V' = \{P_i \mid i \in [r]\} \cup \{Q_j \mid j \in [D]\}$. *Let* π *be an assignment to* F *that satisfies* $a_i \in \mathcal{A}$, *i.e.,* $a_i(\pi(F)) = 0$ *for each* $i \in [r]$. *Let* $b_\pi(V \setminus F)$ *be the polynomial in* $V \setminus F$ *that is obtained by assigning the variables in* F *in* $b(V)$ *according to the assignment* π. *Then* $\mathbf{cons}_F(\mathcal{A}, \{b\}, t) \vdash_t^{V \setminus F, V'} b_\pi(V \setminus F) \geq 0$.

Proof Consider the polynomial $h(F, V') = \sum_{i=1}^r a_i(F)p_i(F) + \sum_{j=1}^D q_j^2(F)$, where $\{P_i\}$ and $\{Q_j\}$ are coefficients of p_i and q_j respectively. Note that $\mathbf{cons}_F(\mathcal{A}, \{b\}, t)$ is a set of polynomial equality constraints in the variables $(V \setminus F) \cup V'$ that enforce the coefficients of the two polynomials $b(F, V \setminus F)$ and $h(F, V')$, when expanded in the monomial basis in F, to be equal. That is, for each $S \in [|F|^t]$, $\mathbf{cons}_F(\mathcal{A}, \{b\}, t)$ contains the constraint $c_S(V \setminus F, V') = 0$, where $b(F, V \setminus F) - h(F, V') = \sum_{S \in [|F|^t]} c_S(V \setminus F, V')F_S$ and $c_S(V \setminus F, V')$ is a polynomial in $V \setminus F$ and V'.

Our goal is to show that the inequality $b_\pi(V \setminus F) \geq 0$ has an SoS proof subject to $\mathbf{cons}_F(\mathcal{A}, \{b\}, t)$. We show this below. Observe that,

$$b(F, V \setminus F) = (b(F, V \setminus F) - h(F, V')) + h(F, V') = \sum_{S \in [|F|]^t} c_S(V \setminus F, V')F_S + h(F, V').$$

Let $f = \pi(F)$. Since the assignment π satisfies the a_i's, we see that $h(f, V') = \sum_{i=1}^r a_i(f)p_i(f) + \sum_{j=1}^D q_j^2(f) = \sum_{j=1}^D q_j^2(f)$. Hence,

$$b(f, V \setminus F) = \sum_{S \in [|F|]^t} f_S c_S(V \setminus F, V') + \sum_{j=1}^D q_j^2(f).$$

This is a valid SoS proof from the axioms $\mathbf{cons}_F(\mathcal{A}, \{b\}, t)$. ■

Appendix B. Omitted Proofs from Section 3

Lemma 15 (Polynomials of k-sparse vectors are bounded) *Let* $p(v_1, \ldots, v_d) = \sum_{T \in [d]^t} a_T v_T$ *be a polynomial of degree* t, *where the coefficients* $\{a_T\}_{T \in [d]^t} \subset \mathbb{R}$ *are real numbers (not variables of the SoS program), then* $\mathcal{A}_{k\text{-sparse}} \vdash_{2t}^{v,z} p(v_1, \ldots, v_d)^2 \leq k^t \max\{a_T^2 \mid T \in [d]^t\}$.

2. Note that if there is an SoS proof of b subject to $\mathcal{A}$ having bounded bit-complexity, then there is a solution to $\mathbf{cons}_F(\mathcal{A}, \{b\}, t)$ which has bounded ℓ_2 norm.

Proof

$$\begin{aligned}
\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{2t}\right. \left(\sum_{T\in[d]^t} a_T v_T\right)^2 &= \left(\sum_{T\in[d]^t} a_T z_T v_T\right)^2 \\
&\le \left(\sum_{T\in[d]^t} a_T^2 z_T^2\right)\left(\sum_{T\in[d]^t} v_T^2\right) \\
&\le \left(\max_{T\in[d]^t}(a_T)^2\right)\left(\sum_{T\in[d]^t} z_T^2\right)\left(\sum_{T\in[d]^t} v_T^2\right) \\
&= \left(\max_{T\in[d]^t}(a_T)^2\right)\left(\sum_{i=1}^d z_i^2\right)^t\left(\sum_{i=1}^d v_i^2\right)^t \\
&= k^t \max_{T\in[d]^t}(a_T)^2 ,
\end{aligned}$$

where the first line uses $\{v_i z_i = v_i\}_{i\in[d]}$, the second line uses SoS Cauchy-Schwartz (Fact 36), the third line uses that $0 \preceq \sum_{T\in[d]^t}(\max_{T\in[d]^t}(a_T)^2 - a_T^2)z_T^2$, the fourth line uses the rewriting $\sum_{T\in[d]^t} z_T = (\sum_{i\in[d]} z_i)^t$ and the same equality for the v_T's, and the last line uses the axioms $\{z_i^2 = z_i\}_{i\in[d]} \cup \{\sum_{i=1}^d z_i = k\} \cup \{\sum_{i=1}^d v_i^2 = 1\}$. ■

B.1. Certifiability for σ-Poincaré Distributions

Previous work has shown that σ-Poincaré distributions have certifiably bounded moments. In this section we show that this implies that σ-Poincaré distributions also have certifiably bounded moments in k-sparse directions. At the end of the section, we demonstrate that certifiability of the moments in k-sparse directions does not always imply the same condition for all (possibly dense) directions.

Lemma 41 *If D is a σ-Poincaré distribution over $\mathbb{R}^d$ with mean μ, then for some constant C_t depending on t, we have that $\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{O(t)}\right. \mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]^2 \le C_t^2\sigma^{2t}$. The bit complexity of the proof is a factor of at most some $\mathrm{poly}(t)$ more than the bit complexity of the polynomial $\mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]^2 - C_t^2\sigma^{2t}$.*

Proof Previous work focused on the notion of *certifiably bounded moments* in the absence of sparsity constraints, i.e., $\{\sum_i v_i^2 = 1\} \left|\frac{v}{t}\right. M^2 \ge \mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]^2$. The following claim implies that if a distribution has certifiably bounded moments, then it also satisfies Definition 13.

Claim 42 (Proofs transfer to unit k-sparse vectors) *For every polynomial $p : \mathbb{R}^d \to \mathbb{R}$, if there is a proof of $\left\{\sum_i v_i^2 = 1\right\} \left|\frac{v}{t}\right. p(v_1,\dots,v_d) \ge 0$ with bit complexity B, then there is a proof of $\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{t}\right. p(v_1,\dots,v_d) \ge 0$ with bit complexity at most B.*

Proof To show $\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{t}\right. p(v_1,\dots,v_d) \ge 0$, it suffices to demonstrate that there exists a set of polynomials $\{r_c(v,z)\}_{c\in\mathcal{A}_{k\text{-sparse}}}$ and a sum of square polynomials $Q(\cdot)$ such that:

$$p(v_1,\dots,v_d) = \sum_{c\in\mathcal{A}_{k\text{-sparse}}} r_c(v,z)c(v,z) + Q(v,z),$$

where the polynomials $r_c(v,z)\cdot c(v,z)$ and $Q(v,z)$ have degree at most t. However, we know that $p(v) = q(v,z) + (\sum_j v_j^2 - 1)q'(v,z)$ for some polynomial q' of degree t and some sum of square polynomials q also of degree t. Setting $r_{\{\sum_j v_j^2 - 1\}} = q'$, $Q = q$, and $r_c = 0$ for all $c \neq \{\sum_j v_j^2 - 1\}$ proves our claim. ■

The following lemma, implicit in Theorem 1.1 from Kothari and Steinhardt (2017), says that if a distribution D is σ-Poincaré, i.e., it holds $\mathbf{Var}_{X\sim D}[f(X)] \leq \sigma^2 \mathbf{E}_{X\sim D}\left[\|\nabla f(X)\|_2^2\right]$ for all differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$, then it has certifiably bounded moments in every (possibly dense) direction.

Lemma 43 (Kothari and Steinhardt (2017)) *If D is a σ-Poincaré distribution over $\mathbb{R}^d$ with mean μ, then there exists some constant C_t depending only on t, such that*

$$\left\{\sum_{i=1}^d v_i^2 = 1\right\} \Big|\frac{v}{O(t)}_{X\sim D} \left[\langle v, X-\mu\rangle^t\right] \leq C_t\sigma^t \,.$$

Moreover, the bit complexity of this proof is at most $\mathrm{poly}(t,b)$, where b is the bit complexity of the coefficients of the polynomial $C_t\sigma^t - \mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]$.

Combining Claim 42 and Lemma 43, and using the fact that for any polynomials A, B, $\{0 < A < B\} \vdash A^2 < B^2$, completes the proof of Lemma 41. ■

Regarding the difference between the two definitions of certifiably bounded moments, one for the dense setting (Definition 2) and one for the sparse setting (Definition 13), we note that there exist distributions that satisfy Definition 13 but do not have certifiably bounded moments in every direction (with a dimension-independent M): Let ξ be the Rademacher random variable and define D to be the distribution of the random variable $X = (X_1, \ldots, X_d)$, where each $X_i = \xi$. Let μ and Σ be the mean and covariance matrix of D. Since the operator norm of Σ is $\sqrt{d}$, it follows that there exists a unit vector v^* (we can take $v^* = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$) such that for any even t, $\mathbf{E}_{X\sim D}[\langle v^*, X-\mu\rangle^t]^2 \geq d^t$. Thus the distribution D does not satisfy Definition 2 with any dimension-independent bound. However, we have that $\mathcal{A}_{k\text{-sparse}} \Big|\frac{v,z}{O(t)} \mathbf{E}_{X\sim D}[\langle v, X-\mu\rangle^t]^2 \leq k^t$ by noting that $\mathbf{E}_{X\sim D}[\langle v, X-\mu\rangle^t] = \mathbf{E}[(\sum_i v_i\xi)^t] = (\sum_i v_i)^t$ and applying Lemma 15.

B.2. Concentration Inequalities for SoS-sparse-certifiability

The goal of this section is to understand the sample complexity required for the result of Section 3.1. Throughout this section, we let $\|X\|_{L_p}$ denote the L_p-norm of the real-valued random variable X, which is defined as $(\mathbf{E}[|X|^p])^{1/p}$. We begin by showing the following concentration result that will be useful in the subsequent proofs.

Lemma 44 *Let P be a random variable over $\mathbb{R}^d$ with mean μ and suppose that for all $s \in [1,\infty)$, P has its s^{th} moment bounded by $(f(s))^s$ for a non-decreasing function $f : [1,\infty) \to \mathbb{R}_+$, in the direction e_j, i.e., suppose that for all $j \in [d]$ and $X \sim P$: $\|\langle e_j, X-\mu\rangle\|_{L_s} \leq f(s)$. Let S be a set of m i.i.d. samples of P. For some sufficiently large absolute constant $C > 0$, we have the following:*

1. *(t-th moment tensor) If* $m > C\frac{1}{\delta^2}\left(t\log d + \log(1/\gamma)\right)\left(f(t^2\log d + t\log(1/\gamma))\right)^{2t}$*, then, with probability* $1-\gamma$*, the* t^{th} *central moment tensor of* P *is bounded in* ℓ_∞ *by* δ*, i.e.*

$$\left\|\mathop{\mathbf{E}}_{S}\left[(X-\mu)^{\otimes t}\right] - \mathop{\mathbf{E}}_{X\sim P}\left[(X-\mu)^{\otimes t}\right]\right\|_\infty \leq \delta. \tag{4}$$

2. *(Absolute moments) If* $m > C\log(d/\gamma)(f(t\log(d/\gamma))/f(t))^{2t}$*, then, with probability* $1-\gamma$*, for all* $i \in [d]$ *and for all* $r \in [t]$*:*

$$\left[\mathop{\mathbf{E}}_{S}[|(X-\mu)_i|^r]\right]^{\frac{1}{r}} \leq 2f(t). \tag{5}$$

3. *(Sample Mean) If* $m > C(1/\delta^2)\log(d/\gamma)(f(\log(d/\gamma)))^2$*, then, with probability* $1-\gamma$*,*

$$\|\mathop{\mathbf{E}}_{S}[X] - \mu\|_\infty \leq \delta. \tag{6}$$

Proof It suffices to consider the case when $\mu = 0$.

Part 1 For any ordered tuple $T \in [d]^t$, we define $p_T : \mathbb{R}^d \to \mathbb{R}$ as $p_T(x) := \prod_{j\in T} x_j$. Let $Y \sim P$. It suffices to show the following:

$$\forall T \in [d]^t : \left|\frac{1}{m}\sum_{i=1}^m \left(p_T(X_i) - \mathbf{E}[p_T(Y)]\right)\right| \leq \tau.$$

Define $Z_{T,i} := p_T(X_i) - \mathbf{E}[p_T(Y)]$ for $i \in [m]$ and $Z_T = \frac{1}{m}\sum_{i=1}^m Z_{T,i}$. Let $s \in \mathbb{Z}_+$. We will control the s-th moment of Z_T using the bound on the s-th moment of $Z_{T,i}$ and independence of $(Z_{T,i})_{i=1}^m$. Recall that X_i has the same distribution as Y.

$$\|Z_{T,i}\|_{L_s} = \|p_T(X_i) - \mathbf{E}[p_T(Y)]\|_{L_s} \leq 2\|p_T(Y)\|_{L_s},$$

where we use triangle inequality and Jensen's inequality. We use Y_j to denote the j-th coordinate of Y. Using the moment bounds on $p_T(Y)$ and Hölder inequality, we get the following:

$$\|p_T(Y)\|_{L_s}^s = \mathbf{E}\left[\left(\prod_{j\in T} Y_j\right)^s\right] = \mathbf{E}\left[\prod_{j\in T}\left(Y_j^s\right)\right] \leq \prod_{j\in T}\left(\mathbf{E}\left[Y_j^{st}\right]\right)^{\frac{1}{t}} \leq \prod_{j\in T}(f(st))^s = (f(st))^{st}, \tag{7}$$

where the first inequality above uses the Cauchy-Schwarz inequality for products of t variables and the second inequality uses the assumption on the moments of Y_j. Thus, $\|Z_{T,i}\|_{L_s} \leq 2\|p_T(Y)\|_{L_s} \leq 2(f(st))^t$.

We will use the following inequalities:

Fact 45 *(Marcinkiewicz-Zygmund's inequality) Let* $W_1, \ldots, W_m, W$ *be identical and independent centered random variables on* $\mathbb{R}$ *with a finite* s*-th moment for* $s \geq 2$*. Then,*

$$\left\|\frac{1}{m}\sum_{i=1}^m W_i\right\|_{L_s} \leq \frac{3\sqrt{s}}{\sqrt{m}}\|W\|_{L_s}.$$

Fact 46 *For a random variable X, we have that w.p. $1-\gamma$, $|X - \mathbf{E}[X]| \le e\|X - \mathbf{E}[X]\|_{L_{\log(1/\gamma)}}$.*

Proof Let $Y = X - \mathbf{E}[X]$. We have the following:

$$\Pr\left[|Y| \ge e\|Y\|_{L_{\log(1/\gamma)}}\right] \le \frac{\mathbf{E}[|Y|^{\log(1/\gamma)}]}{e^{\log(1/\gamma)}\,\mathbf{E}[|Y|^{\log(1/\gamma)}]} = \frac{1}{e^{\log(1/\gamma)}} = \gamma.$$

■

Using Fact 45 and the moment bounds in (7), we get that for any $T \in [d]^t$,

$$\|Z_T\|_{L_s} \lesssim \frac{\sqrt{s}}{\sqrt{m}}(f(st))^t.$$

Using Fact 46 with the above claim, we have that with probability $1-\gamma'$,

$$|Z_T| \le e\|Z_T\|_{L_{\log(1/\gamma')}} \lesssim \frac{\sqrt{\log(1/\gamma')}}{\sqrt{m}}(f(t\log(1/\gamma')))^t.$$

Taking a union bound over $T \in [d]^t$ ordered tuples and taking $\gamma' = \gamma/d^t$, we get that with probability $1-\gamma$,

$$\forall T \in [d]^t : |Z_T| \lesssim \frac{\sqrt{t\log d + \log(1/\gamma)}}{\sqrt{m}}(f(t^2\log d + t\log(1/\gamma)))^t\,.$$

This completes the proof of the first claim.

Part 2 Let $Y := (Y_1, \ldots, Y_d)$ be distributed as P. Using monotonicity of L_p norms, it suffices to bound, for all $i \in [d]$, $\left[\mathbf{E}_S[|(X-\mu)_i|^t]\right]^{\frac{1}{t}}$.

Recall that we assume $\mu = 0$ without loss of generality. For an $i \in [d]$, $j \in [m]$, let $Z_{i,j} := |(X_j)_i|^t$ and $Z_i := \frac{1}{m}\sum_{j=1}^m Z_{i,j}$. By assumption, we have the following for all $r \ge 1$:

$$\mathbf{E}[|Z_{i,j}|^r] = \mathbf{E}[|Y_i|^{rt}] \le (f(rt))^{rt}\,.$$

Thus $\|Z_{i,j}\|_{L_r} \le f(rt)^t$. In particular, for all $r \ge 1$, we have $|\,\mathbf{E}[Z_i]| = |\,\mathbf{E}[Z_{i,j}]| \le \|[Z_{i,j}\|_{L_r} \le (f(rt))^t$, where the first inequality follows from the monotonicity of L_p-norms. Thus we have that $\|Z_{i,j} - \mathbf{E}[Z_{i,j}]\|_{L_r} \le 2(f(rt))^t$.

Applying Fact 45, we have that for all $r \ge 1$

$$\|Z_i - \mathbf{E}[Z_i]\|_{L_r} \lesssim \sqrt{\frac{r}{m}}(f(rt))^t.$$

Applying Fact 46, we have that, with probability $1-\gamma'$, we have that

$$|Z_i - \mathbf{E}[Z_i]| \lesssim \sqrt{\frac{\log(1/\gamma')}{m}}\left(f(t\log(1/\gamma'))\right)^t.$$

Taking $\gamma' = \gamma/d$ with a union bound, we have the following:

$$\forall i \in [d]: \quad Z_i \le (f(t))^t\left(1 + C\sqrt{\frac{\log(d/\gamma)}{m}}\left(\frac{f(t\log(d/\gamma)}{f(t)}\right)^t\right),$$

where C is a large enough constant. The bound follows by noting that $Z_i^{1/t} = \left[\mathbf{E}_S[|(X-\mu)_i|^t]\right]^{\frac{1}{t}}$.

Part 3 For $i \in [d]$, $j \in [m]$, let $Z_{i,j} := (X_j)_i$ and $Z_i := \frac{1}{m}\sum_{j=1}^m Z_{i,j}$.

We have that $\|Z_{i,j}\|_{L_s} \le f(s)$. Applying Fact 45, we get the following: with probability $1-\gamma/d$,

$$\|Z_i - \mathbf{E}[Z_i]\|_{L_{\log(d/\gamma)}} \lesssim \frac{\sqrt{\log(d/\gamma)}}{\sqrt{m}} f(\log(d/\gamma)).$$

Applying a union bound, we get the following: with probability $1-\gamma$,

$$\|\mathop{\mathbf{E}}_S[X] - \mu\|_\infty \lesssim \frac{\sqrt{\log(d/\gamma)}}{\sqrt{m}} f(\log(d/\gamma)).$$

■

Using the above result, we are now ready to prove the concentration result that was required in Section 3.1.

Lemma 47 *Let D be a distribution over $\mathbb{R}^d$ with mean μ. Suppose that for all $s \in [1,\infty)$, D has it's s^{th} moment bounded by $(f(s))^s$ for some non-decreasing function $f : [1,\infty) \to \mathbb{R}_+$, in the direction e_j, i.e. suppose that for all $j \in [d]$ and $X \sim D$:*

$$\|\langle e_j, X-\mu\rangle\|_{L_s} \le f(s).$$

Let $X_1, \ldots, X_m$ be m i.i.d. samples from D and define $\overline{\mu} := \sum_{i=1}^m X_i$. Then with probability $1-\gamma$, we have that

$$\left\| \mathop{\mathbf{E}}_{i\sim[m]}[(X_i-\overline{\mu})^{\otimes t}] - \mathop{\mathbf{E}}_{X\sim D}[(X-\mu)^{\otimes t}] \right\|_\infty \le \delta,$$

when

$$m \ge \max\left(\frac{1}{\delta^2}, 1\right) C\,(t\log(d/\gamma))\left(2f(t^2\log(d/\gamma))\right)^{2t} \max\left(1, \frac{1}{f(t)^{2t}}\right).$$

Proof We can safely assume that $\delta \le 1$. Let $S := \{X_1, \ldots, X_m\}$. The goal is to bound the following:

$$\left\| \mathop{\mathbf{E}}_{X\sim S}\left[(X-\overline{\mu})^{\otimes t}\right] - \mathop{\mathbf{E}}_{X\sim D}\left[(X-\mu)^{\otimes t}\right] \right\|_\infty.$$

We first add and subtract μ in the first term. To prove our lemma, we will bound each entry indexed by an ordered tuple $T \in [d]^t$ of the resulting tensor. We will use the following guarantees on our samples: (i) $\left\|\mathbf{E}_S\left[(X-\mu)^{\otimes t}\right] - \mathbf{E}_{X\sim D}\left[(X-\mu)^{\otimes t}\right]\right\|_\infty \le \delta_1$, (ii) $\max_{i\in d}\max_{r\le t}(\mathbf{E}_S|X_i-\mu_i|^r)^{1/r} \le \delta_2$, and (iii) $\|\mu-\overline{\mu}\|_\infty \le \delta_3$, which appear in Lemma 44, for some values of $\delta_1, \delta_2, \delta_3$ to be defined later. We begin with the following decomposition:

$$\begin{aligned}&\left| \mathop{\mathbf{E}}_{X\sim S}\left[(X-\mu+\mu-\overline{\mu})^{\otimes t}\right]_T - \mathop{\mathbf{E}}_{X\sim D}\left[(X-\mu)^{\otimes t}\right]_T \right| \\ &\quad = \left| \mathop{\mathbf{E}}_{X\sim S}\left[\prod_{q\in T}(X-\mu+\mu-\overline{\mu})_q\right] - \mathop{\mathbf{E}}_{X\sim D}\left[(X-\mu)^{\otimes t}\right]_T \right|. \end{aligned} \tag{8}$$

We can expand $\prod_{q\in T}(X-\mu+\mu-\overline{\mu})_q = \sum_{Q\subseteq T}\prod_{q\in Q}(X-\mu)_q\prod_{q\in T\setminus Q}(\mu-\overline{\mu})_q = \prod_{q\in T}(X-\mu)_q + \sum_{Q\subsetneq T}\prod_{q\in Q}(X-\mu)_q\prod_{q\in T\setminus Q}(\mu-\overline{\mu})_q$, and apply the triangle inequality to get

$$\left|\mathop{\mathbf{E}}_{X\sim S}\left[(X-\overline{\mu})^{\otimes t}\right]_T - \mathop{\mathbf{E}}_{X\sim D}\left[(X-\mu)^{\otimes t}\right]_T\right| \le \left\|\mathop{\mathbf{E}}_{X\sim S}\left[(X-\mu)^{\otimes t}\right] - \mathop{\mathbf{E}}_{X\sim D}\left[(X-\mu)^{\otimes t}\right]\right\|_\infty + \left|\mathop{\mathbf{E}}_{X\sim S}\left[\sum_{Q\subsetneq T}\left[\prod_{q\in Q}(X-\mu)_q\prod_{q\in T\setminus Q}(\mu-\overline{\mu})_q\right]\right]\right|. \quad (9)$$

By assumption, the first term in upper bounded by δ_1. We will now focus on the second term. For a particular $Q\subsetneq T$, we get the following using Holder's inequality:

$$\begin{aligned}\left|\mathop{\mathbf{E}}_{X\sim S}\left[\prod_{q\in Q}(X-\mu)_q\prod_{q\in T\setminus Q}(\mu-\overline{\mu})_q\right]\right| &\le \|\mu-\overline{\mu}\|_\infty^{|T\setminus Q|}\mathop{\mathbf{E}}_{X\sim S}\left[\prod_{q\in Q}|(X-\mu)_q|\right]\\ &\le (\delta_3)^{|T\setminus Q|}\prod_{q\in Q}\left[\mathop{\mathbf{E}}_{X\sim S}|(X-\mu)_q|^{|Q|}\right]^{\frac{1}{|Q|}}\\ &\le \delta_3^{|T\setminus Q|}\delta_2^{|Q|}.\end{aligned}$$

Using the fact that $|\{Q: Q\subsetneq T\}|\le 2^t$ and $|T\setminus Q|\ge 1$, we get the following bound on the second term in (9),

$$\left|\mathop{\mathbf{E}}_{X\sim S}\left[\sum_{Q\subsetneq T}\left[\prod_{q\in Q}(X-\mu)_q\prod_{q\in T\setminus Q}(\mu-\overline{\mu})_q\right]\right]\right| \le 2^t\delta_3\max(1,\delta_2^{t-1},\delta_3^{t-1}).$$

This leads to the following bound on the expression in (8):

$$\left|\mathop{\mathbf{E}}_{X\sim S}\left[(X-\mu+\mu-\overline{\mu})^{\otimes t}\right]_T - \mathop{\mathbf{E}}_{X\sim D}\left[(X-\mu)^{\otimes t}\right]_T\right| \le \delta_1 + \delta_3 2^t\max(1,\delta_2^{t-1},\delta_3^{t-1}). \quad (10)$$

We can choose $\delta_1=\delta/2$, $\delta_2 = 2f(t)$ and $\delta_3 = 2^{-t}\max(1,\delta_2)^{-t+1}\delta/2$, we get that the expression in (10) is upper bounded by δ by noting that $\delta_3\le 1$ (since $\delta\le 1$) and $2^t\delta_3\max(1,\delta_2,\delta_3)^{t-1}\le 2^t\delta_3\max(1,\delta_2)^{-t+1}) \le (\delta/2)$. By Lemma 44, we get that the total sample complexity is at most

$$m = \frac{1}{\delta^2}\left(t\log(d/\gamma)\right)\left(Cf(t^2\log(d/\gamma))\right)^{2t}\max\left(1,\frac{1}{f(t)^{2t}}\right), \quad (11)$$

where we perform the following crude upper bounds on the sample complexity guarantee in Lemma 44 for the ease of presentation:

$$\begin{aligned}&\frac{1}{\delta^2}(t\log(d/\gamma))f(t^2\log d/\gamma) + \log(d/\gamma)\left(\frac{f(\log(d/\gamma))}{f(t)}\right)^{2t}\\ &\qquad + \frac{1}{\delta^2}(\log(d/\gamma))f(\log d/\gamma)^2 2^{8t}(\max(1,2f(t))^{2t-2}\\ &\le (t\log d/\gamma)(10f(t^2\log d/\gamma))^{2t}\Big(\frac{1}{\delta^2}+\frac{1}{f(t)^{2t}}+\frac{1}{\delta^2}\max\left(1,\frac{1}{f(t)}\right)^{2t-2}\Big)\end{aligned}$$

$$\leq \frac{1}{\delta^2}(t\log d/\gamma)(10f(t^2\log d/\gamma))^{2t}\Big(1+\frac{1}{f(t)^{2t}}+\max\Big(1,\frac{1}{f(t)}\Big)^{2t-2}\Big)$$
$$\leq \frac{1}{\delta^2}(t\log d/\gamma)(10f(t^2\log d/\gamma))^{2t}\max\Big(1,\frac{1}{f(t)^{2t}},\frac{1}{f(t)^{2t-2}}\Big)$$
$$\leq \frac{1}{\delta^2}(t\log d/\gamma)(10f(t^2\log d/\gamma))^{2t}\max\Big(1,\frac{1}{f(t)^{2t}}\Big).$$

■

B.3. Proof of Lemma 14

Lemma 14 *Let D be a distribution over $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq I$. Suppose that D satisfies $\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{2t}\right. \mathbf{E}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]^2 \leq M^2$ and D has c-subexponential tails, where c is an absolute constant c. Let $S=\{X_1,\ldots,X_m\}$ be a set of m i.i.d. samples from D with $m = (tk(\log d))^{O(t)}\ \max(1,M^{-2})/\epsilon^2$. Let D' be the uniform distribution over S and $\overline{\mu} := \mathbf{E}_{X\sim D'}[X]$. Then, with probability 0.9, we have that $\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{2t}\right. \mathbf{E}_{X\sim D'}\left[\langle v, X-\overline{\mu}\rangle^t\right]^2 \leq 8M^2$ and $\|\overline{\mu}-\mu\|_{2,k} \leq M^{1/t}\epsilon^{1-1/t}$.*

Proof Suppose for now that with m samples, the ℓ_∞ norm of the difference between the expected and empirical t-th tensors of D is $M/\sqrt{k^t}$, i.e.,

$$\left\|\mathop{\mathbf{E}}_{i\sim[m]}[(X_i-\overline{\mu})^{\otimes t}] - \mathop{\mathbf{E}}_{X\sim D}[(X-\mu)^{\otimes t}]\right\|_\infty \leq \frac{M}{\sqrt{k^t}}.$$

Let $p(v_1,\ldots,v_d) := \sum_{T\in[d]^t}(\mathbf{E}_{i\sim[m]}[X_i-\overline{\mu}]_T - \mathbf{E}_{X\sim D}[X-\mu]_T)v_T$. An easy corollary of Lemma 15 is its application to $p(v_1,\ldots,v_d)$. Combining these two steps we have that:

$$\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{2t}\right. \left(\mathop{\mathbf{E}}_{i\sim[m]}\left[\langle v, X_i-\overline{\mu}\rangle^t\right] - \mathop{\mathbf{E}}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]\right)^2$$
$$\leq k^t\left\|\mathop{\mathbf{E}}_{i\sim[m]}[(X_i-\overline{\mu})^{\otimes t}] - \mathop{\mathbf{E}}_{X\sim D}[(X-\mu)^{\otimes t}]\right\|_\infty^2 \leq M^2\,. \tag{12}$$

To prove bounded central moments of the uniform distribution over the samples, observe that,

$$\mathcal{A}_{k\text{-sparse}} \left|\frac{v,z}{2t}\right. \mathop{\mathbf{E}}_{i\sim[m]}\left[\langle v, X_i-\overline{\mu}\rangle^t\right]^2$$
$$= \left(\mathop{\mathbf{E}}_{i\sim[m]}\left[\langle v, X_i-\overline{\mu}\rangle^t\right] - \mathop{\mathbf{E}}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right] + \mathop{\mathbf{E}}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]\right)^2$$
$$\leq 2^2\left(\mathop{\mathbf{E}}_{i\sim[m]}\left[\langle v, X_i-\overline{\mu}\rangle^t\right] - \mathop{\mathbf{E}}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]\right)^2 + 2^2\mathop{\mathbf{E}}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]^2$$
$$\leq 4\left(M^2 + \mathop{\mathbf{E}}_{X\sim D}\left[\langle v, X-\mu\rangle^t\right]^2\right)$$
$$\leq 8M^2\,,$$

where the third line uses SoS triangle inequality (Fact 37), the fourth line uses Equation (12) and the last one uses our assumption that D has certifiably bounded moments.

We now calculate the sample complexity for the first claim. Since the distribution is sub-exponential, we have that for $Y = (Y_1, \dots, Y_d) \sim D$, $\|Y\|_{L_s} \le cs$, i.e., $f(x) = O(x)$. Lemma 47 with $\delta = M/\sqrt{k^t}$ and $f(x) = O(x)$ implies that the sample complexity is at most the following:

$$C \max\left(1, \frac{k^t}{M^2}\right)(t\log(d/\gamma))(ct^2\log(d/\gamma))^{2t} \max\left(1, \frac{1}{(ct)^{2t}}\right) \lesssim (kt\log(d/\gamma))^{O(t)} \max(1, M^{-2}).$$

We now focus on the second claim. It suffices to show that $\|\overline{\mu} - \mu\|_\infty \le M^{1/t}\epsilon^{1-1/t}/\sqrt{k}$ since this implies the desired $\|\overline{\mu} - \mu\|_{2,k} \le M^{1/t}\epsilon^{1-1/t}$. Part 3 of Lemma 44 implies that this requires samples at most $CkM^{-2/t}\epsilon^{2-2/t}\log(d/\gamma)^3$ since $f(x) = O(x)$. Finally, we note that $\max(M^{-2}, M^{-2/t}, 1) = \max(M^{-2}, 1)$. ■

Appendix C. Omitted Proof from Section 4

We provide the proof of Claim 18 below that was omitted from Section 4.

Claim 18 *Let $r_i := \mathbf{1}_{X_i = Y_i}$ and $W_i := w_i r_i$. There exists an SoS proof of $\{W_i^2 = W_i\}_{i=1}^m \cup \{\sum_{i=1}^m (1 - W_i) \le 2\epsilon m\} \cup \{W_i(X_i - X_i') = 0\}_{i=1}^m$ from the axioms $\{W_i = w_i r_i\}_{i=1}^m \cup \mathcal{A}_{\text{corruptions}}$.*

Proof Since $r_i = \mathbf{1}_{X_i = Y_i}$, then $\sum_i r_i = (1-\epsilon)m, r_i^2 = r_i$ and $r_i(X_i - Y_i) = 0$ for all $i \in [m]$. We see that

1. $W_i^2 = w_i^2 r_i^2 = w_i r_i = W_i$.
2. $\mathcal{A}_{\text{corruptions}} \vdash W_i(X_i - X_i') = w_i r_i (X_i - Y_i + Y_i - X_i') = r_i w_i (Y_i - X_i) + w_i r_i (Y_i - X_i) = 0$.
3. Additionally, since $\{W_i^2 = W_i\} \vdash (1 - W_i)^2 = 1 - 2W_i + W_i^2 = 1 - W_i$, and using the fact that $\{x^2 = x\} \vdash_{O(1)} \{x > 0, x < 1\}$, we see

   $$\mathcal{A}_{\text{corruptions}} \vdash_{O(1)} 1 - W_i \le 2(1 - W_i) = 1 - w_i r_i + 1 - w_i r_i \le (1 - w_i) + (1 - r_i).$$

   A sum over $i \in [m]$ gives us $\vdash_{O(1)} \sum_i (1 - W_i) \le 2\epsilon m$.

■

Appendix D. Omitted Proofs from Section 5

This section contains the omitted proofs from Section 5. We begin by proving that inliers satisfy deterministic conditions with high probability in Appendix D.1. We prove the proofs of estimation lemmata (Lemmata 22 and 23) in Appendix D.2. Remaining technical details are provided in Appendix D.3.

D.1. Deterministic Conditions on Inliers

In this section, we prove that the deterministic conditions required in Section 5.1 hold with high probability. In particular, we provide the proofs of Lemmata 20 and 53.

D.1.1. PROOF OF LEMMA 20

We prove Lemma 20 in this section. To this end, we need the series of lemmata below.

Lemma 48 (Li (2018)) *Let $k \in \mathbb{Z}_+$ with $k \leq d$, $0 < \epsilon \leq 1/2$ and $0 < \gamma < 1$. Let $X_1, \ldots, X_m \sim \mathcal{N}(0, \Sigma)$. There exists an absolute constant C such that, if*

$$m > C \frac{\min(d, k^2) + \log \binom{d^2}{k^2} + \log(1/\gamma)}{\epsilon^2 \log(1/\epsilon)}$$

then, with probability at least $1 - \gamma$, we have that for any choice of weights $a_i \in [0, 1]$ with $\mathbf{E}_{i \sim [m]}[a_i] \geq 1 - \epsilon$, the following two inequalities hold for all vectors $v \in \mathcal{U}_k$:

1. $\left|\mathbf{E}_{i \sim [m]}[a_i \langle v, X_i \rangle]\right| \leq O\left(\epsilon \sqrt{\log(1/\epsilon)}\right) \sqrt{v^T \Sigma v}$.
2. $\left|\mathbf{E}_{i \sim [m]}\left[a_i \langle v, X_i \rangle^2\right] - v^T \Sigma v\right| \leq O(\epsilon \log(1/\epsilon)) v^T \Sigma v$.

The result in Li (2018) is for $\Sigma = I_d$. This version follows by taking a union bound over the support and re-normalizing the distribution. We also require a similar property for the fourth moment of the inliers.

Lemma 49 *Let $k \in \mathbb{Z}_+$ with $k \leq d$, $0 < \epsilon \leq 1/2$, $0 < \gamma < 1$. Let $m > C(k^4/\epsilon^2) \log^4(d/(\epsilon\gamma))$ for a sufficiently large constant C and $X_1, \ldots, X_m \sim \mathcal{N}(0, \Sigma)$. Then, with probability at least $1 - \gamma$, for any weights $a_i \in [0, 1]$ with $\mathbf{E}_{i \sim [m]}[a_i] \geq 1 - \epsilon$ it holds*

$$\left| \mathop{\mathbf{E}}_{i \sim [m]} \left[a_i \left(\left(\langle v, X_i \rangle^2 - v^T \Sigma v \right)^2 - 2 (v^T \Sigma v)^2 \right) \right] \right| \leq \tilde{O}(\epsilon) (v^T \Sigma v)^2$$

for all vectors $v \in \mathcal{U}_k$.

Proof We first show the condition in the case where there are no weights ($a_i = 1$, for all $i \in [m]$).

Lemma 50 *If $m > C(k^4/\epsilon^2) \log^4(d/\gamma)$ for a sufficiently large constant C, then a set of m samples from $\mathcal{N}(0, \Sigma)$ for $I_d \preceq \Sigma \preceq 2I_d$, with probability at least $1 - \gamma$, satisfies least $1 - \gamma$*

$$\left| \mathop{\mathbf{E}}_{i \sim [m]} \left[\left(\langle v, X_i \rangle^2 - v^T \Sigma v \right)^2 - 2 (v^T \Sigma v)^2 \right] \right| \leq O(\epsilon) (v^T \Sigma v)^2$$

for all $v \in \mathcal{U}_k$.

Proof We want to show concentration of polynomials of the form $(\langle v, x \rangle^2 - v^T \Sigma v)^2$ for v, a k-sparse vector. Let S be a set of m samples from $\mathcal{N}(0, \Sigma)$. First, let $u = (vv^T)^\flat$ (i.e., the vector having as elements all the products $v_i v_j$). This is a k^2-sparse vector. Define M as the $d^2 \times d^2$ matrix with $M_{(ij),(k\ell)} = \mathbf{E}_{X \sim S}[(X_i X_j - \Sigma_{ij})(X_k X_\ell - \Sigma_{k\ell})] - 2\Sigma_{ij}\Sigma_{k\ell}$ for all $i, j, k, \ell \in [d]$. We note the rewriting:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X \sim S}[(\langle v, X \rangle^2 - v^T \Sigma v)^2] - 2(v^T \Sigma v)^2 \\
&= \mathop{\mathbf{E}}_{X \sim S} \left[\Big(\sum_{i,j \in [d]} v_i v_j (X_i X_j - \Sigma_{ij}) \Big)^2 \right] - 2 \Big(\sum_{i,j \in [d]} v_i v_j \Sigma_{ij} \Big)^2
\end{aligned}$$

$$= \mathop{\mathbf{E}}_{X\sim S}\Big[\sum_{i,j\in[d]} v_i v_j (X_i X_j - \Sigma_{ij}) \sum_{k,\ell\in[d]} v_k v_\ell (X_k X_\ell - \Sigma_{k\ell}) - 2\sum_{i,j\in[d]} v_i v_j \Sigma_{ij} \sum_{k,\ell\in[d]} v_k v_\ell \Sigma_{ij}\Big]$$
$$= \mathop{\mathbf{E}}_{X\sim S}\Big[\sum_{i,j\in[d]} u_{ij} (X_i X_j - \Sigma_{ij}) \sum_{k,\ell\in[d]} u_{k\ell} (X_k X_\ell - \Sigma_{k\ell}) - 2\sum_{i,j\in[d]} u_{ij} \Sigma_{ij} \sum_{k,\ell\in[d]} u_{k\ell} \Sigma_{ij}\Big]$$
$$= u^T M u$$

Hence, it is sufficient to show that $u^T M u \le \tilde{O}(\epsilon)$ for all $u \in \mathcal{U}_{k^2}(d^2)$. For a $Q \subset [d^2]$, we denote by M_Q the $Q \times Q$ submatrix of M. We have that

$$\sup_{u\in\mathcal{U}_{k^2}(d^2)} u^T M u = \sup_{|Q|\le k^2} \|M_Q\|_2 \le \sup_{|Q|\le k^2} \|M_Q\|_F \ ,$$

Thus, it suffices for every element of M_Q to be $O(\epsilon)$, which holds if

$$\left| \mathop{\mathbf{E}}_{X\sim S}[p(x)] - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(0,\Sigma)}[p(x)] \right| \le \frac{\epsilon}{k^2} \ , \tag{13}$$

for the polynomial $p(x) := (x_i x_j - \Sigma_{ij})(x_k x_\ell - \Sigma_{k\ell}) - 2\Sigma_{ij}\Sigma_{k\ell}$. To this end, we use the following concentration inequality, which is a consequence of Gaussian Hypercontractivity:

Fact 51 (see, e.g., Corollary 5.49 in Aubrun and Szarek (2017)) *Let $Z_1, \dots, Z_m$ be independent $\mathcal{N}(0,1)$ variables and let $X = h(Z_1, \dots, Z_m)$, where h is a polynomial of total degree at most q. Then, for any $t \ge (2e)^{q/2}$,*

$$\Pr\left[|X - \mathbf{E}[X]| \ge t\sqrt{\mathbf{Var}[X]}\right] \le \exp\left(-\frac{q}{2e} t^{2/q}\right).$$

Note that we can still apply this lemma for polynomials $h(Z')$ of Gaussians $Z' \sim \mathcal{N}(0,\Sigma)$ with covariance $\Sigma \ne I$ by noting that $Z' = \sqrt{\Sigma} Z$ where $Z \sim \mathcal{N}(0, I)$ and replacing $h(Z'_1, \dots, Z'_m)$ by $h'(Z_1, \dots, Z_m) = h(Z'_1, \dots, Z'_m) = h((\sqrt{\Sigma}Z)_1, \dots, (\sqrt{\Sigma}Z)_m)$ in Fact 51.

We apply the above to the appropriate degree $q = 4$ polynomial of Equation (13), i.e., $h(X_1, \dots, X_m) = \frac{1}{m}\sum_{i=1}^m (p(X_i) - \mathbf{E}_{X\sim\mathcal{N}(0,\Sigma)}[p(X)])$. We note that in our case $\mathbf{Var}[h(X)] = \mathbf{Var}[p(X)]/m = O(1/m)$. $\mathbf{Var}[p(X)]$ is bounded by a constant since it is a degree 4 polynomial with constant coefficients of Gaussian variables with bounded covariance. We thus obtain that for $m > C(k^4/\epsilon^2)\log^4(1/\gamma')$ samples Equation (13) holds with probability $1 - \gamma'$. Using $\gamma' = \gamma/d^4$ and a union bound over (i, j, k, ℓ) yields the final sample complexity. We have thus shown Lemma 50 with $O(\epsilon)$ in the RHS. Assuming $I_d \preceq \Sigma$, this implies the final claim.

■

Having Lemma 50 in hand, we use it to complete the proof of Lemma 49. By convexity, it suffices to assume $a_i \in \{0, 1\}$. Let I be the set of indices such that $a_i = 1$. For a given $v \in \mathcal{U}_k$, define $p_v(x) = (\langle v, x\rangle^2 - v^T\Sigma v)$. Let J^* be the set of $2\epsilon m$ indices with greatest $(p_v^2(X_i) - 2(v^T\Sigma v)^2)$ and define $J_1^* = \{i : p_v^2(X_i) \ge c\log^2(1/\epsilon)(v^T\Sigma v)^2\}$. If $m > C\log(1/\gamma')/\epsilon^2$, with probability $1 - \gamma'$, we have the following:

1. $\|X_i\|_2 = O(\sqrt{d\log(m/\gamma')})$ for all $i \in [m]$.
2. $\frac{1}{m}|\{p^2(X_i) > c\log^2(1/\epsilon)(v^T\Sigma v)^2\}| \le 2\epsilon$.

3. $J_1^* \subseteq J^*$.

4. $\left|\frac{1}{m}\sum_{i\notin J_1^*}(p^2(X_i) - 2(v^T\Sigma v)^2)\right| = O(\epsilon\log^2(1/\epsilon))(v^T\Sigma v)^2$.

The above claims can be shown like in Appendix B.1 of Diakonikolas et al. (2016) (see Equations (44), (45), (46) of the first arxiv version of that paper; concretely, the second item follows from Fact 51, the third follows from the second, and the last is shown in Claim B.4 of Diakonikolas et al. (2016)).

Fix any $I \subseteq [m]$ with $|I| = (1-2\epsilon)m$. Partition $[m]\setminus I$ into $J^+ \cup J^-$, where $J^+ = \{i \notin I : p^2(X_i) \geq 2(v^T\Sigma v)^2\}$, and $J^- = \{i \notin I : p^2(X_i) < 2(v^T\Sigma v)^2\}$. We will show that $(1/|I|)|\sum_{i\in I}(p_v^2(X_i) - 2(v^T\Sigma v)^2)| = \tilde{O}(\epsilon)(v^T\Sigma v)^2$. We first show the upper bound

$$\begin{aligned}
\frac{1}{|I|}\sum_{i\in I}(p_v^2(X_i) - 2(v^T\Sigma v)^2) &\leq \frac{1}{|I|}\sum_{i\in I\cup J^+}(p_v^2(X_i) - 2(v^T\Sigma v)^2) - \frac{1}{|I|}\sum_{i\in J^-}(p_v^2(X_i) - 2(v^T\Sigma v)^2) \\
&\leq \left|\frac{1}{|I|}\sum_{i=1}^{m}(p_v^2(X_i) - 2(v^T\Sigma v)^2)\right| + 2\frac{1}{|I|}\left|\sum_{i\in J^-}^{m}(p_v^2(X_i) - 2(v^T\Sigma v)^2)\right| \\
&\leq O(\epsilon)(v^T\Sigma v)^2 + \frac{|J^-|}{|I|}O((v^T\Sigma v)^2) \\
&= O(\epsilon)(v^T\Sigma v)^2\,,
\end{aligned}$$

where we used Lemma 50. We now focus on the other direction. We note that the lower bound is achieved when $I = [m]\setminus J^*$. Thus we obtain the following using Items 3 and 4:

$$\begin{aligned}
&\frac{1}{|I|}\sum_{i\in I}(p_v^2(X_i) - 2(v^T\Sigma v)^2) \\
&\geq \frac{1}{(1-2\epsilon)m}\left(\sum_{i\in m}(p_v^2(X_i) - 2(v^T\Sigma v)^2) - \sum_{i\in J^*}(p_v^2(X_i) - 2(v^T\Sigma v)^2)\right) \\
&= \frac{1}{(1-2\epsilon)m}\left(\sum_{i\notin J_1^*}(p_v^2(X_i) - 2(v^T\Sigma v)^2) - \sum_{i\in J\setminus J_1^*}(p_v^2(X_i) - 2(v^T\Sigma v)^2)\right) \\
&\geq -\left|\frac{O(1)}{m}\left(\sum_{i\notin J_1^*}(p_v^2(X_i) - 2(v^T\Sigma v)^2)\right)\right| - \frac{O(1)}{m}\sum_{i\in J\setminus J_1^*}p_v^2(X_i) \\
&\geq -O(\epsilon\log^2(1/\epsilon)(v^T\Sigma v)^2) - \frac{O(1)|J|}{m}c\log^2(1/\epsilon)(v^T\Sigma v)^2 \\
&\geq -O(\epsilon\log^2(1/\epsilon))(v^T\Sigma v)^2\,.
\end{aligned}$$

Note that this holds for a fixed v, and all subsets I with $|I| = (1-2\epsilon)m$. The last step is a cover argument. To that end, we first state that the desired expression is Lipschitz with respect to v:

Claim 52 *Conditioned on the event of Item 1, for any unit-norm $u, v \in \mathbb{R}^d$ and $i \in [m]$ we have that $|p_v(X_i)^2 - 2(v^T\Sigma v)^2 - (p_u(X_i)^2 - 2(u^T\Sigma u)^2)| \lesssim \|u-v\|_2(R^2 + \|\Sigma\|_2^2)^2$, where $R = O(\sqrt{d\log(m/\gamma)})$.*

Proof We first claim the following for the difference between the polynomials without the squares:

$$|p_v(X_i) - p_u(X_i)| \le |\langle v, X_i\rangle^2 - \langle u, X_i\rangle^2| + |v^T\Sigma v - u^T\Sigma u| \le 2\|v-u\|_2(R^2 + \|\Sigma\|_2) \; .$$

The first line uses the triangle inequality. For the second term of the following line we use that $|v^T\Sigma v - u^T\Sigma u| = |v^T\Sigma(v-u) - (u-v)^T\Sigma u| \le \|\Sigma\|_2(\|v\|_2\|v-u\|_2 + \|u\|_2\|v-u\|_2) \lesssim \|\Sigma\|_2\|u-v\|_2$. The second term is bounded using the same argument but with $X_i(X_i)^T$ in place of Σ and using that $\|X_i\|_2 = O(R)$.

We can now complete our proof.

$$\begin{aligned}|p_v(X_i)^2 - 2(v^T\Sigma v)^2 - (p_u(X_i)^2 - 2(u^T\Sigma u))| &\le |p_v(X_i)^2 - p_u(X_i)^2| + 2|(v^T\Sigma v)^2 + (u^T\Sigma u)^2| \\ &\lesssim \max\{|p_v(X_i)|, |p_u(X_i)|\}|p_v(X_i) - p_u(X_i)| + \max\{v^T\Sigma v, u^T\Sigma u\}|v^T\Sigma v - u^T\Sigma u| \\ &\lesssim \|v-u\|_2(R^2 + \|\Sigma\|_2)^2 \; ,\end{aligned}$$

where the first line uses the triangle inequality, the second line uses $|a^2 - b^2| \le 2\max\{|a|, |b|\}|a-b|$, and the last one uses the bound $|p_v(X_i)| \le \|X_i\|^2 + \|\Sigma\|_2 \le R^2 + \|\Sigma\|_2$. ■

Recalling that $\|\Sigma\|_2 \le 2$, the RHS of Claim 52 is essentially $\|v-u\|_2 R^4$. In order for it to become $O(\epsilon)$, we would like our cover S of k-sparse unit vectors to have accuracy $O(\epsilon/R^4)$, which results in a set of size $|S| = \binom{d}{k}O(R^4/\epsilon)^k$ vectors. We thus choose the probability of failure $\gamma' = \gamma/|S|$, which means that we need

$$m > C\frac{\log(|S|/\gamma)}{\epsilon^2} = \frac{\log\binom{d}{k} + k\log(d\log(m/\gamma)/\epsilon) + \log(1/\gamma)}{\epsilon^2} \; .$$

The sample complexity of Lemma 50 scales with k^4 and dominates the sample complexity of this paragraph. This completes the proof of Lemma 49. ■

We now have all the ingredients to prove Lemma 20, which we restate below for convenience.

Lemma 20 *Let T denote the set of all $a \in [0,1]^{m\times m}$ such that (i) $a_{ij} = a_{ji}$ for all $i, j \in [m]$, (ii) $\mathbf{E}_{ij}[a_{ij}] \ge 1 - 4\epsilon$, and (iii) $\mathbf{E}_j[a_{ij}] \ge a_i(1-2\epsilon)$ for all $i \in [m]$ and $a_{ij} \le a_i$ for all $i, j \in [m]$. Let $X_1, \dots, X_m \sim \mathcal{N}(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and a positive definite matrix $I_d \preceq \Sigma \preceq 2I_d$. Denote $X_{ij} := (1/2)(X_i - X_j)(X_i - X_j)^T$ and $\overline{\Sigma} := \mathbf{E}_{ij}[X_{ij}]$. If $m > (k^4/\epsilon^2)\mathrm{polylog}(d/\epsilon\gamma)$, then, with probability $1-\gamma$ we have that the following hold for all $v \in \mathcal{U}_k$:*

1. $|\langle v, \overline{\mu} - \mu\rangle| \le \tilde{O}(\epsilon)\sqrt{v^T\Sigma v}$.
2. $|\mathbf{E}_{i\sim[m]}[a_i\langle v, X_i - \overline{\mu}\rangle]| \le \tilde{O}(\epsilon)\sqrt{v^T\overline{\Sigma}v}$.
3. $\left|\mathbf{E}_{i\sim[m]}\left[a_i\left(\langle v, X_i - \overline{\mu}\rangle^2 - v^T\overline{\Sigma}v\right)\right]\right| \le \tilde{O}(\epsilon)v^T\overline{\Sigma}v$.
4. $|v^T(\overline{\Sigma} - \Sigma)v| \le \tilde{O}(\epsilon)v^T\Sigma v$.
5. $|\mathbf{E}_{i,j\sim[m]}[a_{ij}(v^TX_{ij}v - v^T\overline{\Sigma}v)]| \le \tilde{O}(\epsilon)v^T\overline{\Sigma}v$.
6. $|\mathbf{E}_{i,j\sim[m]}[a_{ij}((v^TX_{ij}v - v^T\overline{\Sigma}v)^2 - 2(v^T\overline{\Sigma}v)^2)]| \le \tilde{O}(\epsilon)(v^T\overline{\Sigma}v)^2$.

Proof Without loss of generality, we assume $\mu = 0$. We let $Z_i = X_i - \overline{\mu}$. We condition on the events of Lemmata 48 and 49. For simplicity, we also assume that the a_i''s are scaled so that $\mathbf{E}_{ij}[a_{ij}] = \mathbf{E}_i[a_i'] = 1$ (since this consists of only a scaling of $1 + O(\epsilon)$, it is without loss of generality). We show the individual claims below:

1. Proof of $|\langle v, \overline{\mu}\rangle| \le \tilde{O}(\epsilon)\sqrt{v^T\Sigma v}$: Use Lemma 48 with $a_i = 1$.

2. Proof of $|\mathbf{E}_{i\sim[m]}[a_i'\langle v, X_i - \overline{\mu}\rangle]| \le \tilde{O}(\epsilon)\sqrt{v^T\overline{\Sigma}v}$: By Item 1 and Lemma 48, we have that

$$\left|\mathop{\mathbf{E}}_{i\sim[m]}[a_i'\langle v, X_i - \overline{\mu}\rangle]\right| \le \left|\mathop{\mathbf{E}}_{i\sim[m]}[a_i'\langle v, X_i\rangle]\right| + \left|\mathop{\mathbf{E}}_{i\sim[m]}[a_i'\langle v, \overline{\mu}\rangle]\right| \le \tilde{O}(\epsilon)\sqrt{v^T\Sigma v} \le \frac{\tilde{O}(\epsilon)\sqrt{v^T\overline{\Sigma}v}}{\sqrt{1-\tilde{O}(\epsilon)}} ,$$

   where the last inequality uses Item 4.

3. Proof of $\left|\mathbf{E}_{i\sim[m]}\left[a_i'\left(\langle v, X_i - \overline{\mu}\rangle^2 - v^T\overline{\Sigma}v\right)\right]\right| \le \tilde{O}(\epsilon)v^T\overline{\Sigma}v$: We have the following inequalities.

$$\begin{aligned}
&\left|\mathop{\mathbf{E}}_{i\sim[m]}\left[a_i'\left(\langle v, X_i - \overline{\mu}\rangle^2 - v^T\overline{\Sigma}v\right)\right]\right| \\
&= \left|\mathop{\mathbf{E}}_{i\sim[m]}\left[a_i'\langle v, X_i\rangle^2 + a_i'\langle v, \overline{\mu}\rangle^2 - 2a_i'\langle v, X_i\rangle\langle v, \overline{\mu}\rangle - a_i'v^T\overline{\Sigma}v\right]\right| \\
&\le \left|(1 \pm \tilde{O}(\epsilon))v^T\Sigma v + \tilde{O}(\epsilon^2)v^T\Sigma v \pm 2\mathop{\mathbf{E}}_{i\sim[m]}[a_i'\langle v, X_i\rangle]\tilde{O}(\epsilon)\sqrt{v^T\Sigma v} - v^T\overline{\Sigma}v\right| \\
&= |v^T(\overline{\Sigma} - \Sigma)v| + \tilde{O}(\epsilon)v^T\Sigma v + \tilde{O}(\epsilon^2)v^T\Sigma v \\
&= \tilde{O}(\epsilon)v^T\overline{\Sigma}v ,
\end{aligned}$$

   where the second line uses Lemma 48 along with Item 1 and $\mathbf{E}_i[a_i] = 1$, the next line uses Lemma 48 and the last line relates $v^T\Sigma v$ to $v^T\overline{\Sigma}v$ using Item 4.

4. Proof of $|v^T(\overline{\Sigma} - \Sigma)v| \le \tilde{O}(\epsilon)v^T\Sigma v$: Repeating the steps from the proof of Item 3, we can show $\left|\mathbf{E}_{i\sim[m]}\left[a_i'\left(\langle v, X_i - \overline{\mu}\rangle^2 - v^T\Sigma v\right)\right]\right| \le \tilde{O}(\epsilon)v^T\Sigma v$. Taking $a_i' = 1$ gives Item 4.

5. Proof of $|\mathbf{E}_{i,j\sim[m]}[a_{ij}(v^TX_{ij}v - v^T\overline{\Sigma}v)]| \le \tilde{O}(\epsilon)v^T\overline{\Sigma}v$: We clarify that we will often use the following: Whenever we have a term of the form $\mathbf{E}_{ij}[a_{ij}b_i]$ with $b_i \ge 0$, we will use that $\mathbf{E}_{ij}[a_{ij}b_i] \le \mathbf{E}_i[a_i'b_i]$ (since $0 < a_{ij} \le a_i'$).

$$\begin{aligned}
&\left|\mathop{\mathbf{E}}_{i,j\sim[m]}\left[a_{ij}(v^TX_{ij}v - v^T\overline{\Sigma}v)\right]\right| \\
&\quad= \left|\mathop{\mathbf{E}}_{i,j\sim[m]}[a_{ij}\frac{1}{2}\langle v, X_i - \overline{\mu} - (X_j - \overline{\mu})\rangle^2 - v^T\overline{\Sigma}v]\right| \\
&\quad\lesssim \left|\mathop{\mathbf{E}}_{i}[a_i'\frac{1}{2}\langle v, X_i - \overline{\mu}\rangle^2] - \frac{1}{2}v^T\overline{\Sigma}v\right| + \left|\mathop{\mathbf{E}}_{j}\left[a_j'\frac{1}{2}\langle v, X_j - \overline{\mu}\rangle^2\right] - \frac{1}{2}v^T\overline{\Sigma}v\right| \\
&\quad+ \left|\mathop{\mathbf{E}}_{ij}\left[a_{ij}\langle v, X_i - \overline{\mu}\rangle\langle v, X_j - \overline{\mu}\rangle\right]\right| \\
&\quad\le \tilde{O}(\epsilon)v^T\overline{\Sigma}v ,
\end{aligned}$$

   since each of the first two terms is $\tilde{O}(\epsilon)$ using Items 2 to 4 and the same holds for the last term: Using Cauchy–Schwarz, this term becomes $|\mathbf{E}_{ij}[a_{ij}\langle v, X_i - \overline{\mu}\rangle\langle v, X_j - \overline{\mu}\rangle]| \le \sqrt{|\mathbf{E}_{ij}[a_{ij}\langle v, X_i - \overline{\mu}\rangle^2]\,\mathbf{E}_{ij}[a_{ij}\langle v, X_j - \overline{\mu}\rangle^2]|}$ and then applying again Items 2 to 4 bounds it by $\tilde{O}(\epsilon)v^T\overline{\Sigma}v$.

6. Proof of $|\mathbf{E}_{i,j\sim[m]}[a_{ij}((v^T X_{ij} v - v^T\overline{\Sigma}v)^2 - 2(v^T\Sigma v)^2)]| \leq \tilde{O}(\epsilon)(v^T\overline{\Sigma}v)^2$: Using that $\mathbf{E}_{i,j}[a_{ij}] = 1$ and some algebraic manipulations, we have that:

$$\begin{aligned}
&\left|\mathop{\mathbf{E}}_{i,j\sim[m]}[a_{ij}((v^T X_{ij} v - v^T\overline{\Sigma}v)^2 - 2(v^T\Sigma v)^2)]\right| \\
&= \left|\mathop{\mathbf{E}}_{i,j\sim[m]}\left[a_{ij}\left(\frac{\langle v, X_i - X_j\rangle^2 - 2v^T\overline{\Sigma}v}{2}\right)^2\right] - 2(v^T\Sigma v)^2\right| \\
&= \left|\frac{1}{4}\mathop{\mathbf{E}}_{i,j}\left[a_{ij}\left(\left(\langle v, X_i\rangle^2 - v^T\overline{\Sigma}v + \langle v, X_j\rangle^2 - v^T\overline{\Sigma}v - 2\langle v, X_i\rangle\langle v, X_j\rangle\right)^2\right)\right] - 2(v^T\Sigma v)^2\right| \\
&= \left|\frac{1}{4}\mathop{\mathbf{E}}_{ij\sim[m]}\left[a_{ij}\left(A^2 + B^2 + C^2 + 2AB + 2AC + 2BC\right)\right] - 2(v^T\overline{\Sigma}v)^2 \pm \tilde{O}(\epsilon)\right| , \qquad (14)
\end{aligned}$$

where we replaced $(v^T\Sigma v)^2$ by $(v^T\overline{\Sigma}v)^2 \pm \tilde{O}(\epsilon)$ using Item 4 and the fact that $I \preceq \Sigma \preceq 2I$ and we let $A := \langle v, X_i\rangle^2 - v^T\overline{\Sigma}v$, $B := \langle v, X_j\rangle^2 - v^T\overline{\Sigma}v$, $C := 2\langle v, X_i\rangle\langle v, X_j\rangle$. We work with each term individually. We have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{ij\sim[m]}[A^2] &= \mathop{\mathbf{E}}_{i\sim[m]}[a_i'(\langle v, X_i\rangle^2 - v^T\overline{\Sigma}v)^2] \\
&= \mathop{\mathbf{E}}_{i\sim[m]}[a_i'(\langle v, X_i\rangle^2 - v^T\Sigma v + v^T(\Sigma - \overline{\Sigma})v)^2] \\
&\leq \mathop{\mathbf{E}}_{i\sim[m]}[a_i'(\langle v, X_i\rangle^2 - v^T\Sigma v)^2] + \mathop{\mathbf{E}}_{i\sim[m]}[a_i'](v^T(\Sigma - \overline{\Sigma})v)^2 \\
&\quad + 2\mathop{\mathbf{E}}_{i\sim[m]}[a_i'(\langle v, X_i\rangle^2 - v^T\Sigma v)](v^T(\Sigma - \overline{\Sigma})v) \\
&\leq 2(v^T\Sigma v)^2 + \tilde{O}(\epsilon)(v^T\Sigma v)^2 ,
\end{aligned}$$

since the first term is $(2 + \tilde{O}(\epsilon))(v^T\Sigma v)^2$ and the other two $\tilde{O}(\epsilon)(v^T\Sigma v)^2$: the first term is bounded because of Lemma 49, the second because of Item 4 and the last one because of both (an application of Cauchy-Schwarz is needed there). Similarly, we have that $\mathbf{E}_{ij\sim[m]}[B^2] \leq (2 \pm \tilde{O}(\epsilon))(v^T\Sigma v)^2$. Furthermore, $\mathbf{E}_{ij\sim[m]}[C^2] = (4 \pm \tilde{O}(\epsilon))(v^T\Sigma v)^2$ and that all cross terms involving AB, AC, BC are $\tilde{O}(\epsilon)$. Putting these together we get that the right-hand side of Equation (14) is $\tilde{O}(\epsilon)$. Using Lemma 49 one last time we have that this is at most $\tilde{O}(\epsilon)(v^T\overline{\Sigma}v)^2$.

This completes the proof of Lemma 20. ■

D.1.2. PROOF OF LEMMA 21

Lemma 21 *Let $X_1, \ldots, X_m \sim \mathcal{N}(\mu, \Sigma)$ for a k-sparse vector $\mu \in \mathbb{R}^d$ and a $d \times d$ symmetric matrix $I_d \preceq \Sigma \preceq 2I_d$. Let $\overline{\mu}$ and $\overline{\Sigma}$ be the empirical mean and covariance respectively. If the number of samples $m > (k^4/\epsilon^2)\mathrm{poly}\log(d, 1/\gamma, 1/\epsilon)$, then, with probability at least $1 - \gamma$, we have that $\mathcal{A}_{k\text{-sparse}} \vdash_8^{v,z} \left(\mathbf{E}_{i\sim[m]}\left[\langle v, X_i - \overline{\mu}\rangle^4\right] - 3(v^T\overline{\Sigma}v)^2\right)^2 \leq \tilde{O}(\epsilon^2)(v^T\overline{\Sigma}v)^4$.*

Proof We have the following by the SoS triangle inequality (Fact 37):

$$\mathcal{A}_{k\text{-sparse}} \vdash_8^{v,z} \left(\mathop{\mathbf{E}}_{i\sim[m]}\left[\langle v, X_i - \overline{\mu}\rangle^4\right] - 3(v^T\overline{\Sigma}v)^2\right)^2$$

$$= \left(\mathop{\mathbf{E}}_{i\sim[m]} \left[\langle v, X_i - \overline{\mu}\rangle^4\right] - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu\rangle^4\right] + \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu\rangle^4\right] - 3(v^T\overline{\Sigma}v)^2 \right)^2$$
$$\leq 4\left(\mathop{\mathbf{E}}_{i\sim[m]} \left[\langle v, X_i - \overline{\mu}\rangle^4\right] - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu\rangle^4\right] \right)^2 + 4\left(\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu\rangle^4\right] - 3(v^T\overline{\Sigma}v)^2 \right)^2 .$$

We will bound from above each of the two terms above separately. Focusing on the first term, we first define $\delta' := \epsilon$. Then, similarly to Equation (12), we use Lemma 15 and Lemma 47 with $\delta = \delta'/k^2$ and $t = 4$ to get that

$$\mathcal{A}_{k\text{-sparse}} \overset{v,z}{\underset{8}{\vdash}} \left(\mathop{\mathbf{E}}_{i\sim[m]} \left[\langle v, X_i - \overline{\mu}\rangle^4\right] - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu\rangle^4\right] \right)^2$$
$$\leq k^4 \left\| \mathop{\mathbf{E}}_{i\sim[m]} [(X_i - \overline{\mu})^{\otimes 4}] - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} [(X - \mu)^{\otimes 4}] \right\|_\infty^2$$
$$\leq k^4\delta^2 \leq (\delta')^2 \lesssim (\delta')^2(v^T\overline{\Sigma}v)^4 = O(\epsilon^2)(v^T\overline{\Sigma}v)^4 ,$$

where in the last line we used $\Sigma \succeq I_d$ combined with Item 4 from Lemma 20. The sample complexity of $(k^4/\epsilon^2)\log^5(d/(\epsilon\gamma))$ comes from Lemma 20 and Lemma 47, with $f(s) \leq \sqrt{Cs}$ and $\delta = \epsilon/k^2$.

We similarly bound the second term:

$$\mathcal{A}_{k\text{-sparse}} \overset{v,z}{\underset{8}{\vdash}} \left(\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu\rangle^4\right] - 3(v^T\overline{\Sigma}v)^2 \right)^2$$
$$= \left(\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu\rangle^4\right] - \mathop{\mathbf{E}}_{X\sim\mathcal{N}(\overline{\mu},\overline{\Sigma})} \left[\langle v, X - \overline{\mu}\rangle^4\right] \right)^2$$
$$= \left(\mathop{\mathbf{E}}_{Y\sim\mathcal{N}(0,\Sigma)} \left[\langle v, Y\rangle^4\right] - \mathop{\mathbf{E}}_{Y\sim\mathcal{N}(0,\overline{\Sigma})} \left[\langle v, Y\rangle^4\right] \right)^2$$
$$\leq k^4 \left\| \mathop{\mathbf{E}}_{Y\sim\mathcal{N}(0,\Sigma)} [Y^{\otimes 4}] - \mathop{\mathbf{E}}_{Y\sim\mathcal{N}(0,\overline{\Sigma})} [Y^{\otimes 4}] \right\|_\infty^2 ,$$

where we used the specific form of Gaussian moments (Fact 29) for the first equality. In order to bound all elements of the tensor, we use the following lemma, which is shown in Appendix D.1.4.

Lemma 53 *Let $X_1, \ldots, X_m \sim \mathcal{N}(\mu, \Sigma)$ where $I \preceq \Sigma \preceq 2I$, and denote $\overline{\mu} = \mathbf{E}_{i\sim[m]}[X_i]$, $\overline{\Sigma} = \mathbf{E}_{i\sim[m]}[(X_i - \overline{\mu})(X_i - \overline{\mu})^T]$. For any even integer t and $\tau < 1$, if $m > C(1/\tau^2)t^{2t+1}4^t \log(d/\gamma)$ for some absolute constant C, it holds*

$$\left\| \mathop{\mathbf{E}}_{Y\sim\mathcal{N}(0,\Sigma)} [Y^{\otimes t}] - \mathop{\mathbf{E}}_{Y\sim\mathcal{N}(0,\overline{\Sigma})} [Y^{\otimes t}] \right\|_\infty \leq \tau ,$$

with probability $1 - \gamma$.

Using the above with $t = 4$ and $\tau = \delta'/k^2$ with $\delta' = \tilde{O}(\epsilon)$, we get that

$$\mathcal{A}_{k\text{-sparse}} \overset{v,z}{\underset{8}{\vdash}} \left(\mathop{\mathbf{E}}_{X\sim\mathcal{N}(\mu,\Sigma)} \left[\langle v, X - \mu\rangle^4\right] - 3(v^T\overline{\Sigma}v)^2 \right)^2 \leq (\delta')^2 \lesssim (\delta')^2(v^T\overline{\Sigma}v)^4 = \tilde{O}(\epsilon^2)(v^T\overline{\Sigma}v)^4 .$$

This completes the proof of Lemma 21. ■

D.1.3. FEASIBILITY OF DEFINITION 19

Corollary 54 *Under the conditions of Lemma 21, $\mathcal{A}_{\text{G-sparse-mean-est}}$ in Definition 19 is feasible with high probability.*

Proof The pseudo-distribution that is defined to be the uniform distribution on inliers (i.e., $X'_i = X_i$) satisfies the constraints of the program. The first three conditions are trivially satisfied by choosing the w_i's to be the indicators of whether the i-th sample is an inlier. The second to last constraint is satisfied if and only if the inequality $\left(\mathbf{E}_{i\sim[m]}\left[\langle v, X_i - \overline{\mu}\rangle^4\right] - 3(v^T\overline{\Sigma}v)^2\right)^2 \le \tilde{O}(\epsilon^2)(v^T\overline{\Sigma}v)^4$ has an SoS proof. By Lemma 21, we know that this is indeed the case.

We now focus on the last constraint. We need to show an SoS proof of $(v^T\overline{\Sigma}v)^2 < 9$. We will show an SoS proof of $(v^T\overline{\Sigma}v - v^T\Sigma v)^2 \le O(\epsilon^2)$. By techniques similar to the ones used in Lemma 21, we see that

$$\mathcal{A}_{k\text{-sparse}} \vdash (v^T\overline{\Sigma}v - v^T\Sigma v)^2 \le k^2\|\overline{\Sigma} - \Sigma\|_\infty \,.$$

Since $m > C(k^4/\epsilon^2)\log^5(d/\gamma)$ for large enough constant C, we have the following with high probability:

$$\mathcal{A}_{k\text{-sparse}} \vdash (v^T\overline{\Sigma}v - v^T\Sigma v)^2 \le k^2\|\overline{\Sigma} - \Sigma\|_\infty \le O(\epsilon^2) \,.$$

Finally, to get an upper bound on $(v^T\overline{\Sigma}v)^2$, we apply the SoS triangle inequality, as shown below

$$\begin{aligned}\mathcal{A}_{k\text{-sparse}} \vdash (v^T\overline{\Sigma}v)^2 &= (v^T\overline{\Sigma}v - v^T\Sigma v + v^T\Sigma v)^2 \\ &\le 2(v^T\Sigma v)^2 + 2(v^T\overline{\Sigma}v - v^T\Sigma v)^2 \le 8 + 2O(\epsilon^2) \le 8 + O(\epsilon^2) \le 9 \,,\end{aligned}$$

where we use that $\Sigma \preceq 2I$ and ϵ is chosen to be small enough. ■

D.1.4. PROOF OF LEMMA 53

This section contains the proof of the following result that was used in Section 5.1.

Lemma 53 *Let $X_1, \ldots, X_m \sim \mathcal{N}(\mu, \Sigma)$ where $I \preceq \Sigma \preceq 2I$, and denote $\overline{\mu} = \mathbf{E}_{i\sim[m]}[X_i]$, $\overline{\Sigma} = \mathbf{E}_{i\sim[m]}[(X_i - \overline{\mu})(X_i - \overline{\mu})^T]$. For any even integer t and $\tau < 1$, if $m > C(1/\tau^2)t^{2t+1}4^t\log(d/\gamma)$ for some absolute constant C, it holds*

$$\left\| \mathop{\mathbf{E}}_{Y\sim\mathcal{N}(0,\Sigma)}[Y^{\otimes t}] - \mathop{\mathbf{E}}_{Y\sim\mathcal{N}(0,\overline{\Sigma})}[Y^{\otimes t}] \right\|_\infty \le \tau \,,$$

with probability $1 - \gamma$.

Our proof uses the following standard concentration inequality and Isserlis' theorem.

Claim 55 *Let $X^{(1)}, \ldots, X^{(N)} \sim \mathcal{N}(0, \Sigma)$ and $\Sigma_N := \frac{1}{N}\sum_{i=1}^N X^{(i)}X^{(i)^T}$. Let $\epsilon', \delta \in (0, 1)$. Then, with probability at least $1 - \delta$, it holds that $\|\Sigma_N - \Sigma\|_\infty \le \epsilon'\|\Sigma\|_2$ provided that $N > C\log(d/\delta)/\epsilon'^2$ for a sufficiently large universal constant C.*

Proof For $k, \ell \in [d]$, the random variable $\frac{1}{N}\sum_{i=1}^{d} X_k^{(i)} X_\ell^{(i)}$ is sub-exponential with Orlicz norm $\|\frac{1}{N} X_k^{(i)} X_\ell^{(i)}\|_{\psi_1} \leq (C/N)\|\Sigma\|_2$ for some $C > 0$. Therefore, by Bernstein's inequality, if N is a large enough multiple of $\log(1/\delta')/\epsilon'^2$, we have that

$$\Pr\left[\left|\frac{1}{N} X_k^{(i)} X_\ell^{(i)} - \Sigma_{k\ell}\right| \leq \epsilon' \|\Sigma\|_2\right] \leq \delta' \,.$$

By using $\delta' = \delta/d^2$ and taking a union bound over all d^2 elements of the matrix, the result follows. ■

Fact 56 (Isserlis' theorem) *Let $(X_1, \ldots, X_t) \sim \mathcal{N}(0, \Sigma)$. Then,*

$$\mathbf{E}[X_1 \cdots X_t] = \sum_{p \in P_t^2} \prod_{\{i,j\} \in p} \mathbf{E}[X_i X_j] \,,$$

where P_t^2 is the set of all pairings of $[t]$.

We are now ready to prove Lemma 53.

Proof [Proof of Lemma 53] We fix $\ell_1, \ldots, \ell_t \in [d]$ and examine the $(\ell_1, \ldots, \ell_t)$-th entry $\left(\mathbf{E}_{Y \sim \mathcal{N}(0,\Sigma)}[Y^{\otimes t}]\right)_{\ell_1,\ldots,\ell_t}$ $= \mathbf{E}_{Y \sim \mathcal{N}(0,\Sigma)}[Y_{\ell_1} \ldots Y_{\ell_t}]$. Using Fact 56, we can write it as a sum of products of elements of Σ:

$$\left(\mathop{\mathbf{E}}_{Y \sim \mathcal{N}(0,\Sigma)}[Y^{\otimes t}]\right)_{\ell_1,\ldots,\ell_t} = \sum_{p \in P_t^2} \prod_{\{i,j\} \in p} \Sigma_{\ell_i \ell_j} \,.$$

We note that each product has at most $(t/2)$-many factors. The same decomposition holds for each entry of the tensor $\mathbf{E}_{Y \sim \mathcal{N}(0,\overline{\Sigma})}[Y^{\otimes t}]$ by replacing Σ with $\overline{\Sigma}$. Therefore,

$$\left|\left(\mathop{\mathbf{E}}_{Y \sim \mathcal{N}(0,\Sigma)}[Y^{\otimes t}] - \mathop{\mathbf{E}}_{Y \sim \mathcal{N}(0,\overline{\Sigma})}[Y^{\otimes t}]\right)_{\ell_1,\ldots,\ell_t}\right| \leq \sum_{p \in P_t^2} \left| \prod_{\{i,j\} \in p} \Sigma_{\ell_i \ell_j} - \prod_{\{i,j\} \in p} \overline{\Sigma}_{\ell_i \ell_j} \right| . \tag{15}$$

We now focus on a single term of the right-hand side. Assuming that we have an approximation $\|\Sigma - \overline{\Sigma}\|_\infty \leq \delta$ for some $\delta < 1$ to be defined later, we can write $\overline{\Sigma}_{ij} = \Sigma_{ij} + \delta_{ij}$ with $|\delta_{ij}| \leq \delta$. Plugging this gives

$$\left| \prod_{\{i,j\} \in p} \Sigma_{\ell_i \ell_j} - \prod_{\{i,j\} \in p} \overline{\Sigma}_{\ell_i \ell_j} \right| \leq \delta \|\Sigma\|_\infty^{t/2-1} 2^{t/2} \,,$$

where we used that $\prod_{\{i,j\} \in p} (\Sigma_{\ell_i \ell_j} + \delta_{\ell_i \ell_j})$ produces one term which cancels out with $\prod_{\{i,j\} \in p} \Sigma_{\ell_i \ell_j}$ and every one of the $(2^{t/2} - 1)$ remaining ones, is at most $\delta \|\Sigma\|_\infty^{t/2-1}$, because $\delta < 1$, and $\|\Sigma\|_\infty \geq 1$.

Combining the above with Equation (15), we have that

$$\left\| \mathop{\mathbf{E}}_{Y \sim \mathcal{N}(0,\Sigma)}[Y^{\otimes t}] - \mathop{\mathbf{E}}_{Y \sim \mathcal{N}(0,\overline{\Sigma})}[Y^{\otimes t}] \right\|_\infty \lesssim t^t \delta \|\Sigma\|_\infty^{t/2} 2^{t/2} \,,$$

since a crude upper bound on the number of matchings of $[t]$ is $t!$. Imposing that the right-hand side is at most τ, we find that it is sufficient to have

$$\delta \leq \frac{\tau}{t^t 2^{t/2} \|\Sigma\|_\infty^{t/2-1}},$$

which is indeed less than 1 since $\tau \leq 1$ and $\|\Sigma\|_\infty \geq 1$.

Therefore, we use Claim 55 with $\epsilon' \leq \delta/\|\Sigma\|_2$ and $\delta = \gamma/d^t$, in order to do a union bound over the d^t elements of the tensor. Substituting these parameters yields the claimed sample complexity. ■

D.2. Proof of Estimation Lemmata

We recall the following general result from prior work (note that our theorem statement is slightly different from the one in Kothari et al. (2022), this is a minor correction and doesn't change the overall correctness of the proof).

Lemma 57 (Lemma 22 in Kothari et al. (2022)) *Let $X_1, \ldots, X_m \in \mathbb{R}^d$ and $\overline{\mu} = \mathbf{E}_{i\sim[m]}[X_i]$. Let $V(\overline{\mu}, v), V'(\mu', v)$ be degree-2 polynomials in v and $\overline{\mu}$ and v and μ' respectively and let $S \subseteq \mathbb{R}^d$ be a set such that $V(\overline{\mu}, v) \geq 0$ for all $v \in S$ and $\overline{\mu} \in \mathbb{R}^d$. Let $T \subseteq [0,1]^m$. Suppose that for every $v \in S$ and $a \in T$ such that $\sum_i a_i \geq (1-\epsilon)m$, we have the following two*

$$\left| \mathop{\mathbf{E}}_{i\sim[m]}[a_i \langle v, X_i - \overline{\mu}\rangle] \right| \leq \tilde{O}(\epsilon)\sqrt{V(\overline{\mu}, v)}\,, \tag{16}$$

$$\left| \mathop{\mathbf{E}}_{i\sim[m]}[a_i(\langle v, X_i - \overline{\mu}\rangle^2 - V(\overline{\mu}, v))] \right| \leq \tilde{O}(\epsilon) V(\overline{\mu}, v)\,. \tag{17}$$

Let $Y_1, \ldots, Y_m$ be any ϵ-corruption of $X_1, \ldots, X_m$, let $\tilde{\mathbf{E}}$ be a degree-6 pseudo-expectation in the variables $X'_1, \ldots, X'_m \in \mathbb{R}^d$ and $w_1, \ldots, w_n \in \mathbb{R}$. Let $\mu' = \mathbf{E}_{i\sim[m]}[X'_i]$. Suppose that

- *$\tilde{\mathbf{E}}$ satisfies $w_i^2 = w_i$ for all $i \in [m]$.*
- *$\tilde{\mathbf{E}}$ satisfies $\sum_{i\in[m]} w_i = (1-\epsilon)m$.*
- *$\tilde{\mathbf{E}}$ satisfies $w_i X'_i = w_i Y_i$ for all $i \in [m]$.*
- *$\tilde{\mathbf{E}}[\mathbf{E}_{i\sim[m]}[\langle v, X'_i - \mu'\rangle^2]] \leq (1 + \tilde{O}(\epsilon))\tilde{\mathbf{E}}[V'(\mu', v)]$ for every $v \in S$.*
- *$a \in T$, where a is the vector with $a_i = \tilde{\mathbf{E}}[w_i]\mathbf{1}(X_i = Y_i)$ for $i \in [m]$*

Then, for every $v \in S$, the following hold:

$$\tilde{\mathbf{E}}[\langle v, \mu' - \overline{\mu}\rangle^2] \leq O(\epsilon)\left(\tilde{\mathbf{E}}[V'(\mu', v)] + V(\overline{\mu}, v)\right)\,,$$

$$|\langle v, \hat{\mu} - \overline{\mu}\rangle| \leq \tilde{O}(\epsilon)\sqrt{V(\overline{\mu}, v)} + \sqrt{\tilde{\mathbf{E}}\left[\mathop{\mathbf{E}}_{i\sim[m]}[(1-w'_i)\langle v, X'_i - \overline{\mu}\rangle]^2\right]}\,,$$

where $\hat{\mu} := \tilde{\mathbf{E}}[\mu']$ and

$$\tilde{\mathbf{E}}\left[\mathop{\mathbf{E}}_{i\sim[m]}[(1-w'_i)\langle v, X'_i - \overline{\mu}\rangle]^2\right] \leq O(\epsilon)\left(\tilde{\mathbf{E}}[V'(\mu', v)] - V(\overline{\mu}, v)\right)$$

$$+ \tilde{O}(\epsilon)\left(\tilde{\mathbf{E}}[V'(\mu', v)] + V(\overline{\mu}, v)\right),$$

where $w_i' = w_i \mathbf{1}(Y_i = X_i)$.

We now prove how Lemma 22 and Lemma 23 follow from Lemma 57.

Lemma 22 *Let $Y_1, \ldots, Y_m$ be an ϵ-corruption of the set $X_1, \ldots, X_m$, satisfying Items 2 and 3 of Lemma 20. Let $\tilde{\mathbf{E}}$ be a degree-6 pseudo-expectation in variables w_i, X_i', Σ', μ' satisfying the system of Definition 19. Denote by $\overline{\mu}, \overline{\Sigma}$ the empirical mean and covariance of $X_1, \ldots, X_m$ and let $\widehat{\Sigma} := \tilde{\mathbf{E}}[\Sigma']$. Then, for all $v \in \mathcal{U}_k$ it holds $|\langle v, \widehat{\mu} - \overline{\mu}\rangle| \le \tilde{O}(\epsilon)\sqrt{v^T\overline{\Sigma}v} + \sqrt{O(\epsilon)v^T(\widehat{\Sigma} - \overline{\Sigma})v + \tilde{O}(\epsilon^2)v^T(\widehat{\Sigma} + \overline{\Sigma})v}$.*

Proof This is a corollary of Lemma 57 with $V(\overline{\mu}, v) := v^T \mathbf{E}_{i\sim[m]}[(X_i - \overline{\mu})(X_i - \overline{\mu})^T]v$, $V'(\mu', v) := v^T \mathbf{E}_{i\sim[m]}[(X_i' - \mu')(X_i' - \mu')^T]v$, and the set S chosen to be the set of all k-sparse unit vectors of $\mathbb{R}^d$.

We now check that the assumptions of Lemma 57 are true. The assumption of Equations (16) and (17) holds because of Items 2 and 3 of Lemma 20. The first three conditions about the pseudo-expectation $\tilde{\mathbf{E}}$ hold because $\tilde{\mathbf{E}}$ satisfies the program of Definition 19 and the last one holds trivially since $\tilde{\mathbf{E}}[\mathbf{E}_{i\sim[m]}[\langle v, X_i' - \mu'\rangle^2]] = \tilde{\mathbf{E}}[v^T\Sigma' v] = \tilde{\mathbf{E}}[V'(\mu', v)]$. Finally, $a_i' \in [0, 1]$ since $\tilde{\mathbf{E}}$ satisfies $w_i^2 = w_i$ and $\sum w_i \ge 1 - \epsilon$. ■

Lemma 23 *Let $Y_1, \ldots, Y_m$ be an ϵ-corruption of $X_1, \ldots, X_m$ satisfying Items 5 and 6 of Lemma 20. Let $\tilde{\mathbf{E}}$ be a degree-12 pseudo-expectation in variables w_i, X_i', Σ', μ' satisfying the system of Definition 19. Define $Y_{ij} = (1/2)(Y_i - Y_j)(Y_i - Y_j)^T$, $X_{ij} = (1/2)(X_i - X_j)(X_i - X_j)^T$, $X_{ij}' = (1/2)(X_i' - X_j')(X_i' - X_j')^T$, $\widehat{\Sigma} = \tilde{\mathbf{E}}[\Sigma']$, $w_{ij}' = w_i w_j \mathbf{1}(X_{ij} = Y_{ij})$, and $R = \tilde{\mathbf{E}}[\mathbf{E}_{ij}[(1 - w_{ij}')v^T(X_{ij}' - \overline{\Sigma})v]^2]$. Then, for every $v \in \mathcal{U}_k$, we have that, $|v^T(\widehat{\Sigma} - \overline{\Sigma})v| \le \tilde{O}(\epsilon)v^T\overline{\Sigma}v + \sqrt{R}$ and $R \le O(\epsilon)(\tilde{\mathbf{E}}[(v^T\Sigma' v)^2] - (v^T\overline{\Sigma}v)^2) + \tilde{O}(\epsilon)(\tilde{\mathbf{E}}[(v^T\Sigma' v)^2] + (v^T\overline{\Sigma}v)^2)$.*

Proof We use Lemma 57 with the following substitutions: $S := \{(vv^T)^\flat \mid v \in \mathcal{U}_k\}$, that is, let S be the set of all d^2-dimensional vectors that are obtained by flattening matrices vv^T for v k-sparse unit vectors. We let the differences of pairs $(X_{ij})^\flat := (\frac{1}{2}(X_i - X_j)(X_i - X_j)^T)^\flat$ for $i, j \in [m]$ play the role of the vectors $X_i, \ldots, X_j$ that appear in the statement of Lemma 57 and $\mathbf{E}_{ij}[(X_{ij})^\flat]$ play the role of $\overline{\mu}$. We choose $V(\mathbf{E}_{ij}[(X_{ij})^\flat], (vv^T)^\flat) := (v^T\overline{\Sigma}v)^2 = (v^T \mathbf{E}_{ij\sim[m]}[X_{ij}]v)^2 = \langle (vv^T)^\flat, \mathbf{E}_{ij}[(X_{ij})^\flat]\rangle^2$ (from this re-writing it is seen that this is a degree-2 polynomial in its arguments). Similarly, we let V' be as V but where X_i are replaced with X_i', i.e., the program variables. Thus, the assumption of Equations (16) and (17) now corresponds to Items 5 and 6 of Lemma 20. For example, to see the correspondence for the case of Item 5, we note that for $u = (vv^T)^\flat \in S$, the LHS in Equation (16) becomes

$$\begin{aligned}\mathop{\mathbf{E}}_{ij}[a_{ij}\langle u, (X_{ij})^\flat - \mathop{\mathbf{E}}_{ij}[(X_{ij})^\flat]\rangle] &= \mathop{\mathbf{E}}_{ij}[a_{ij}\langle (vv^T)^\flat, ((1/2)(X_i - X_j)(X_i - X_j)^T)^\flat - \mathop{\mathbf{E}}_{ij}[(X_{ij})^\flat]\rangle] \\ &= \mathop{\mathbf{E}}_{ij}[a_{ij}(v^T X_{ij} v - v^T\overline{\Sigma}v)],\end{aligned}$$

which is equal to the LHS in Item 5 of Lemma 20.

It remains to show that the rest of the assumptions used in Lemma 57 hold. We use $w_{ij} := w_i w_j$ in place of the w_i's appearing in Lemma 57. Note that by requiring $\tilde{\mathbf{E}}$ to be degree-12 pseudo-expectation in the variables X'_i, w_i, it follows that $\tilde{\mathbf{E}}$ is degree-6 in the new variables X'_{ij}, w_{ij}. We want to check that

1. $\tilde{\mathbf{E}}$ satisfies $w_{ij}^2 = w_{ij}$ for all $i, j \in [m]$.
2. $\tilde{\mathbf{E}}$ satisfies $\sum_{i,j\in[m]} w_{ij} = (1-\epsilon)^2 m^2$.
3. $\tilde{\mathbf{E}}$ satisfies $w_{ij} X'_{ij} = w_{ij} Y_{ij}$ for every $i, j \in [m]$.
4. $\tilde{\mathbf{E}}[\mathbf{E}_{i,j\sim[m]}[(v^T(X'_{ij} - \Sigma')v)^2]] \le (2 + \tilde{O}(\epsilon))\tilde{\mathbf{E}}[(v^T\Sigma' v)^2]$.

The first three follow immediately from the constraints of the program of Definition 19. The last one is verified below.

Claim 58 *Let $\tilde{\mathbf{E}}$ be a degree-4 pseudo-expectation in X'_{ij}, Σ' (as defined in Lemma 23) satisfying Definition 19. Then*

$$\tilde{\mathbf{E}}\left[\mathop{\mathbf{E}}_{i,j\sim[m]}[(v^T(X'_{ij} - \Sigma')v)^2]\right] \le (2 + \tilde{O}(\epsilon))\tilde{\mathbf{E}}\left[(v^T\Sigma' v)^2\right] .$$

Proof Since $\tilde{\mathbf{E}}$ satisfies the system of Definition 19, by taking pseudoexpectations on the second to last constraint of the program, we see,

$$\tilde{\mathbf{E}}\left[\left(\mathop{\mathbf{E}}_{i\sim[m]}\left[\langle v, X'_i - \mu'\rangle^4\right] - 3(v^T\Sigma' v)^2\right)^2\right] \le \tilde{O}(\epsilon^2)\tilde{\mathbf{E}}\left[(v^T\Sigma' v)^4\right] .$$

Applying Cauchy-Schwarz for pseudoexpectations (Fact 35), we get,

$$\left(\tilde{\mathbf{E}}\left[\mathop{\mathbf{E}}_{i\sim[m]}\left[\langle v, X'_i - \mu'\rangle^4\right] - 3(v^T\Sigma' v)^2\right]\right)^2 \le \tilde{O}(\epsilon^2)\tilde{\mathbf{E}}\left[(v^T\Sigma' v)^4\right] .$$

We know from Fact 33 that $\{0 \le x \le 9\} \vdash \{81x^2 - x \ge 0\}$. Letting $x = (v^T\Sigma' v)^2$ we see that $v^T\Sigma' v = \sum_{ij}\langle X'_i - X'_j, v\rangle^2 > 0$ and the last constraint of the program implies $x \le 9$. Taking pseudoexpectations, we see that $\tilde{\mathbf{E}}[(v^T\Sigma' v)^4] \le O(1)\cdot\tilde{\mathbf{E}}[(v^T\Sigma' v)^2] \le O(1)\tilde{\mathbf{E}}[(v^T\Sigma' v)^2]^2$, where the final inequality follows from the fact that $\tilde{\mathbf{E}}[(v^T\Sigma' v)^2]$ is bounded between constants. Taking square roots of the previous inequality, since all terms involved are powers of two, and hence positive, this implies

$$\tilde{\mathbf{E}}\left[\mathop{\mathbf{E}}_{i\sim[m]}\left[\langle v, X'_i - \mu'\rangle^4\right]\right] \le (3 + \tilde{O}(\epsilon))\tilde{\mathbf{E}}\left[(v^T\Sigma' v)^2\right] .$$

Hence, we have that

$$\tilde{\mathbf{E}}\left[\mathop{\mathbf{E}}_{i\sim[m]}[\langle v, X'_i - \mu'\rangle^4]\right] \le (3 + \tilde{O}(\epsilon))\tilde{\mathbf{E}}[(v^T\Sigma' v)^2] . \tag{18}$$

We also have the following polynomial equality

$$\mathcal{A}_{\text{G-sparse-mean-est}} \vdash_4 \mathop{\mathbf{E}}_{i,j\sim[m]}[(v^T(X'_{ij} - \Sigma')v)^2] = \frac{1}{2}\left(\mathop{\mathbf{E}}_{i\sim[m]}[\langle v, X'_i - \mu'\rangle^4] + (v^T\Sigma' v)^2\right) . \tag{19}$$

Taking pseudo-expectations in Equation (19) and combining with Equation (18), yields that

$$\tilde{\mathbf{E}}\left[\mathop{\mathbf{E}}_{i,j\sim[m]}[(v^T(X'_{ij}-\Sigma')v)^2]\right] = \frac{1}{2}\left(\tilde{\mathbf{E}}\left[\mathop{\mathbf{E}}_{i\sim[m]}[\langle v, X'_i-\mu'\rangle^4]\right] + \tilde{\mathbf{E}}[(v^T\Sigma' v)^2]\right)$$
$$\leq \frac{1}{2}\left(4+\tilde{O}(\epsilon)\right)\tilde{\mathbf{E}}[(v^T\Sigma' v)^2]\,,$$

which is the claimed bound. ■

This completes the proof of Lemma 23. ■

D.3. Omitted Details from Section 5.2

We first state the guarantees of our algroithm under the case $I_d \preceq \Sigma \preceq 2I_d$.

Theorem 59 *Let $k, d \in \mathbb{Z}_+$ with $k \leq d$ and $\epsilon < \epsilon_0$ for a sufficiently small constant $\epsilon_0 > 0$. Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ such that $I_d \preceq \Sigma \preceq 2I_d$. Let $m > C(k^4/\epsilon^2)\log^5(d/(\gamma\epsilon))$ for a sufficiently large constant $C > 0$. There exists an algorithm which, given ϵ,k, and an ϵ-corrupted set of samples from $\mathcal{N}(\mu,\Sigma)$ of size m, it runs in time $\mathrm{poly}(md)$, and returns an estimate $\widehat{\mu}$ such that, with probability $1-\gamma$, $\widehat{\mu}$ satisfies $\|\widehat{\mu}-\mu\|_{2,k} \leq \tilde{O}(\epsilon)$.*

We complete the proof of Theorem 59 as in Kothari et al. (2022) by using the estimation lemmata proved above. By Lemma 23 we have that

$$|\langle \widehat{\Sigma} - \Sigma^*, vv^T\rangle| \leq \tilde{O}(\epsilon)v^T\Sigma_0 v + \sqrt{R},$$

and additionally

$$R := \tilde{\mathbf{E}}[\mathop{\mathbf{E}}_{ij\sim[m]}\left[(1-w'_{ij})\cdot v^T(X_{ij}-\Sigma^*)v\right]^2]$$
$$\leq O(\epsilon)\left(\tilde{\mathbf{E}}[(v^T\Sigma' v)^2] - (v^T\Sigma^* v)\right) + \tilde{O}(\epsilon^2)\left(\tilde{\mathbf{E}}((v^T\Sigma' v)^2) + (v^T\Sigma^* v)^2\right).$$

We can write $\Sigma' = A + B$ with $B = \mathbf{E}_{ij}[(1-w'_{ij})X'_{ij}]$ and $A = \mathbf{E}_{ij}[w'_{ij}X_{ij}] = \mathbf{E}_{ij}[w'_{ij}X'_{ij}]$. The latter equality follows by the definitions of the quantities. We will use the notation $M_v := v^T M v$ for any matrix M (in particular, we will use this for $M \in \{A, B, \Sigma^*\}$). We have that

$$\tilde{\mathbf{E}}[A_v^2] = \tilde{\mathbf{E}}[(\mathop{\mathbf{E}}_{ij\sim[m]}[w'_{ij}v^T X_{ij} v])^2] = \mathop{\mathbf{E}}_{i_1,j_1}\mathop{\mathbf{E}}_{i_2,j_2}\tilde{\mathbf{E}}[w'_{i_1j_1}w'_{i_2j_2}]\cdot v^T X_{i_1j_1}v \cdot v^T X_{i_2j_2}v$$
$$\leq \mathop{\mathbf{E}}_{i_1,j_1}\mathop{\mathbf{E}}_{i_2,j_2}\sqrt{\tilde{\mathbf{E}}[w'_{i_1j_1}]\tilde{\mathbf{E}}[w'_{i_2j_2}]}\cdot v^T X_{i_1j_1}v \cdot v^T X_{i_2j_2}v$$
$$= \left(\mathop{\mathbf{E}}_{i,j\sim[m]}\left[\sqrt{\tilde{\mathbf{E}}[w'_{ij}]}v^T X_{ij} v\right]\right)^2 \leq (1+\tilde{O}(\epsilon))\Sigma_v^2\,,$$

where the final inequality follows from the resilience condition (Lemma 20) with $a_{ij} = \sqrt{\tilde{\mathbf{E}}[w'_{ij}]}$ (note that $(a_{ij})_{i,j}$ satisfy the required properties of that lemma with $a_i = \tilde{\mathbf{E}}[w_i]\mathbf{1}(X_i = Y_i)$ for all i). We can now rewrite upper bound R in terms of A, B, Σ.

$$R = \tilde{\mathbf{E}}[(B_v - \mathop{\mathbf{E}}_{ij}[1-w'_{ij}]\cdot\Sigma_v)^2]$$

$$\leq O(\epsilon)(\tilde{\mathbf{E}}[(A_v + B_v)^2] - \Sigma_v^2) + \tilde{O}(\epsilon^2)(\tilde{\mathbf{E}}[(A_v + B_v)^2] + \Sigma_v^2).$$

By expanding the square, we can also lower bound R as follows:

$$\tilde{\mathbf{E}}[(B_v - \mathop{\mathbf{E}}_{ij}[1 - w'_{ij}] \cdot \Sigma_v)^2] \geq \tilde{\mathbf{E}}[B_v^2] - 4\epsilon\Sigma_v\tilde{\mathbf{E}}[B_v],$$

as $\Sigma_v \geq 0$ and $\tilde{\mathbf{E}}$ satisfies $B_v \geq 0$. As $\tilde{\mathbf{E}}[A_v^2] \leq (1 + \tilde{O}(\epsilon))\Sigma_v^2$ and $\tilde{\mathbf{E}}[A_v B_v] \leq \sqrt{\tilde{\mathbf{E}}[A_v^2]\tilde{\mathbf{E}}[B_v^2]}$ by pseudoexpectation Cauchy-Schwartz (Fact 35),

$$\begin{aligned}\tilde{\mathbf{E}}[(A_v + B_v)^2] &\leq \tilde{\mathbf{E}}[B_v^2] + \sqrt{\tilde{\mathbf{E}}[A_v^2]\tilde{\mathbf{E}}[B_v^2]} + (1 + \tilde{O}(\epsilon))\Sigma_v^2 \\ &\leq \tilde{\mathbf{E}}[B_v^2] + 2\Sigma_v\sqrt{\tilde{\mathbf{E}}[B_v^2]} + (1 + \tilde{O}(\epsilon))\Sigma_v^2.\end{aligned}$$

Thus we can combine upper and lower bounds on R to get,

$$\tilde{\mathbf{E}}[B_v^2] - 4\epsilon\Sigma_v\tilde{\mathbf{E}}[B_v] \leq O(\epsilon)\left(\tilde{\mathbf{E}}[B_v^2] + 2\Sigma_v\sqrt{\tilde{\mathbf{E}}[B_v^2]} + \tilde{O}(\epsilon)\Sigma_v^2\right) + \tilde{O}(\epsilon^2)\Sigma_v^2.$$

Rearranging, applying $\tilde{\mathbf{E}}[B_v] \leq \sqrt{\tilde{\mathbf{E}}[B_v^2]}$ and solving for $\tilde{\mathbf{E}}[B_v^2]$ yields $\tilde{\mathbf{E}}[B_v^2] \leq \tilde{O}(\epsilon^2)\Sigma_v^2$. This implies an upper bound on R of $\tilde{O}(\epsilon^2)\Sigma_v^2$. This, in turn implies

$$|v^T(\widehat{\Sigma} - \Sigma^*)v| \leq \tilde{O}(\epsilon)v^T\Sigma^*v + \sqrt{R} = \tilde{O}(\epsilon)v^T\Sigma^*v.$$

This is the desired guarantee with Σ^* instead of Σ. Using property 4 in Lemma 20 and the triangle inequality, we get the desired spectral norm guarantee in terms of Σ. This finishes the proof (we have already shown in the main body that $\widehat{\mu}$ satisfies this property given that $\widehat{\Sigma}$ does).

D.4. Achieving Error Scaling with $\sqrt{\|\Sigma\|_2}$

In this section, we complete the proof of Theorem 7, which we restate below:

Theorem 60 *Let $k, d \in \mathbb{Z}_+$ with $k \leq d$ and $\epsilon < \epsilon_0$ for a sufficiently small constant $\epsilon_0 > 0$. Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix. There exists an algorithm which, given ϵ,k, and an ϵ-corrupted set of samples from $\mathcal{N}(\mu, \Sigma)$ of size $m = O((k^4/\epsilon^2)\log^5(d/(\epsilon)))$, runs in time $\mathrm{poly}(md)$, and returns an estimate $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_{2,k} \leq \tilde{O}(\epsilon)\sqrt{\|\Sigma\|_2}$ with high probability.*

Thus far, we have obtained an estimator that is $\tilde{O}(\epsilon)$-accurate given samples from $\mathcal{N}(\mu, \Sigma)$ with $I_d \preceq \Sigma \preceq 2I_d$. Note that the assumption $I_d \preceq \Sigma$ can trivially be removed by having a pre-processing step that adds a zero-mean identity covariance Gaussian noise to all samples (since a zero-mean noise does not affect the mean). However, when $\Sigma \preceq \sigma^2 I_d$ with σ much smaller than 1, the optimal error rate is $\tilde{O}(\epsilon)\sigma$, which is much better than $\tilde{O}(\epsilon)$. If σ is known to the algorithm in advance, the simple normalization step that is shown in Algorithm 3 with $\tilde{\sigma} = \sigma$ suffices to yield the desired error of $\tilde{\sigma}\tilde{O}(\epsilon)$. In other words, we have so far obtained an estimator $\mathrm{RobustMean}(S, \tilde{\sigma}, \epsilon, k)$ that is guaranteed to return a vector within $\tilde{\sigma}\tilde{O}(\epsilon)$ from the true mean with probability $1 - \gamma$ (given that the number of samples is as specified in Theorem 59), so long as $\tilde{\sigma} \geq \sigma$.

Theorem 61, known as Lepskii's method Lepskii (1991); Birgé (2001), states that even in the case where the only known bounds for σ are $\sigma \in [A, B]$ for some A, B, a near-optimal error can still be achieved by running $\mathrm{RobustMean}(S, \tilde{\sigma}, \epsilon, k)$ below.

Algorithm 3 Improved estimator when σ is known.

function ROBUSTMEAN($S = \{x_1, \ldots, x_m\}, \tilde{\sigma}, \epsilon, k$)
  Let $e_1, \ldots, e_m \sim \mathcal{N}(0, I_d)$.
  Let $\tilde{S} = \{x_i/\tilde{\sigma} + e_i \ : \ i \in [m]\}$.
  $\tilde{\mu} \leftarrow$ GAUSSIAN-SPARSE-MEAN-EST($\tilde{S}, \epsilon, k$). ▷ Algorithm 2
return $\tilde{\sigma}\tilde{\mu}$
end function

Theorem 61 *Let $\mu \in \mathbb{R}^d$, $A, B > 0$, $\sigma \in [A, B]$, and a non-decreasing function $r : \mathbb{R}^+ \to \mathbb{R}^+$. Suppose $\mathrm{Alg}(\tilde{\sigma}, \gamma')$ is a black-box algorithm which is guaranteed to return a vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \le r(\tilde{\sigma})$, with probability at least $1 - \gamma'$, whenever $\tilde{\sigma} \ge \sigma$. Then, Algorithm 4, returns $\widehat{\mu}^{(\widehat{J})}$ such that, with probability at least $1 - \gamma$, it holds $\|\widehat{\mu}^{(\widehat{J})} - \mu\|_2 \le 3r(2\sigma)$. Moreover, Algorithm 4 calls Alg at most $O(\log(B/A))$ times.*

Proof For $j = 0, 1, \ldots, \log(B/A)$, denote by $\mathcal{E}_j$ the event that $\|\widehat{\mu}^{(j)} - \mu\|_2 \le r(\tilde{\sigma}_j)$. Let J be the index corresponding to the value of the unknown parameter σ, i.e., $\tilde{\sigma}_{J+1} \le \sigma \le \tilde{\sigma}_J$. Conditioned on the event $\cap_{j=0}^{J} \mathcal{E}_j$, we have that $\|\widehat{\mu}^{(j)} - \mu\|_2 \le r(\tilde{\sigma}_j)$ for all $j = 0, 1, \ldots, J$. Using the triangle inequality, this gives that $\|\widehat{\mu}^{(J)} - \widehat{\mu}^{(j)}\|_2 \le r(\tilde{\sigma}_J) + r(\tilde{\sigma}_j)$. This means that the stopping condition of the while loop in Algorithm 4 is satisfied during round J and thus, if $\widehat{\mu}^{(\widehat{J})}$ denotes the vector returned by the algorithm, we have that $\widehat{J} \ge J$ and

$$\|\widehat{\mu}^{(\widehat{J})} - \widehat{\mu}^{(J)}\|_2 \le r(\tilde{\sigma}_{\widehat{J}}) + r(\tilde{\sigma}_J) \le 2r(\tilde{\sigma}_J) \le 2r(2\sigma) \ ,$$

where the first inequality uses the condition of the while loop, the second uses that r is non-decreasing and $\tilde{\sigma}_{\widehat{J}} \le \tilde{\sigma}_J$, and the last one uses that J was defined to be such that $\tilde{\sigma}_{J+1} \le \sigma \le \tilde{\sigma}_J$ so multiplying σ by 2 makes it greater than $\tilde{\sigma}_J$. Using the triangle inequality once more, we get $\|\widehat{\mu}^{(\widehat{J})} - \mu\|_2 \le 3r(2\sigma)$. Finally, by a union bound on the events $\mathcal{E}_j$, the probability of error is upper bounded by $\sum_{j=0}^{J} \gamma' \le \gamma$. ■

Algorithm 4 Adaptive search for σ

Input: $A, B, r(\cdot)$
Denote $\tilde{\sigma}_j := B/2^j$ for $j = 0, 1, \ldots, \log(B/A)$ and set $\gamma' := \gamma/\log(B/A)$.
$\gamma' := \gamma/\log(B/A)$.
$J \leftarrow 0$.
while $\tilde{\sigma}_j \ge A$ *and* $\|\widehat{\mu}^{(J)} - \widehat{\mu}^{(j)}\|_2 \le r(\tilde{\sigma}_J) + r(\tilde{\sigma}_j)$ *for all* $j = 0, 1, \ldots, J-1$ **do**
  $J \leftarrow J + 1$.
  $\widehat{\mu}^{(J)} \leftarrow \mathrm{Alg}(\tilde{\sigma}_J, \gamma')$.
end
$\widehat{J} \leftarrow J - 1$
return $\widehat{\mu}^{(\widehat{J})}$

In our setting, we use the following claim to get estimates for A and B such that B/A is at most polynomial in d.

Claim 62 *Let $S = \{Y_1, \ldots, Y_m\}$ be an ϵ-corrupted set from $\mathcal{N}(\mu, \Sigma)$. Then we can obtain estimates A and B such that $B/A = \mathrm{poly}(d)$ and with probability $1 - \exp(-m)$, $\|\Sigma\|_2 \in [A, B]$.*

Proof Suppose that m is even and define $m' := m/2$. Let $T = \{Z_1, \ldots, Z_{m'}\}$, where $Z_i = (Y_i - Y_{m'+i})/\sqrt{2}$. Note that T is an 2ϵ-corrupted set of m' points from $\mathcal{N}(0, \Sigma)$. Let $X \sim \mathcal{N}(0, \Sigma)$. We know that there exist constants $0 < c_1 < c_2$ such that $\Pr(\|X\|_2^2 \in [c_1 \mathrm{tr}(\Sigma)/d, c_2 \mathrm{tr}(\Sigma)]) \geq 3/4$, which follows by anti-concentration of the Gaussian and Markov's inequality. Thus, the Chernoff bound implies that with probability at least $1 - \exp(-cm)$, at least 60% percent of the points have squared norm lying in $[c_1 \mathrm{tr}(\Sigma)/d, c_2 \mathrm{tr}(\Sigma)]$. Since $\epsilon < 0.1$, we have that with same probability, the empirical median of squared norms also lies in the same range. Assume that this event holds for the remainder of the proof. Let $D = \mathrm{Median}_{z \in T}(\|z\|_2^2)$. We have that $c_1 \|\Sigma\|_2/d \leq c_1 \mathrm{tr}(\Sigma)/d \leq D \leq c_2 \mathrm{tr}(\Sigma) \leq c_2 d \|\Sigma\|_2$. Let $A = D/(c_2 d)$ and $B = dD/c_1$. ■

Putting everything together, we get our final theorem for Gaussian sparse mean estimation with unknown covariance.

Proof (Proof of Theorem 60) Let S be an ϵ-corrupted set from $\mathcal{N}(\mu, \Sigma)$ of size m as specified in the theorem statement. The algorithm is the following: We first obtain rough bounds A, B for $\|\Sigma\|_2$ using the estimator of Claim 62. We then use the procedure of Algorithm 4 with $\mathrm{Alg}(\tilde{\sigma}, \gamma)$ being the $\mathrm{RobustMean}(S, \tilde{\sigma}, \epsilon, k)$ from Algorithm 3, which is guaranteed to succeed with probability $1 - \gamma'$, where $\gamma' = \gamma/(c' \log d)$ for a large enough constant c'. By Theorem 59, it suffices to use $C(k^4/\epsilon^2) \log^5(d/(\gamma\epsilon))$ samples for a large constant C. By Theorem 59, the black-box mean estimator $\mathrm{RobustMean}$ satisfies the guarantees required by Theorem 61 with $\sigma = \sqrt{\|\Sigma\|_2}$, $r(\tilde{\sigma}) = \tilde{\sigma}\tilde{O}(\epsilon)$, and A, B given by those found using the estimator of Claim 62. Therefore, the final guarantee is that Algorithm 4 attains error $3r(2\sigma) = \sqrt{\|\Sigma\|_2}\tilde{O}(\epsilon)$ with probability at least $1 - \gamma$. Since Lepskii's method only calls the black-box estimator $\log(B/A) = O(\log(d))$ times, the computational complexity increases only by a logarithmic factor. ■

Appendix E. Statistical Query Lower Bounds

We begin by summarizing the necessary background and then move to showing our results on Gaussians and distributions with bounded t-th moments in Appendices E.2 and E.3 respectively. We refer the reader to Appendix F.3 for the implications of the lower bounds of this section to hardness against low-degree polynomial tests.

E.1. Background

STATISTICAL QUERY LOWER BOUNDS FRAMEWORK

We start with the basic definitions and facts from Feldman et al. (2013); Diakonikolas et al. (2017b) that we will use later. Although we are interested in proving hardness of estimation problems, we will focus on simpler hypothesis testing (or decision) problems.

Definition 63 (Decision Problem over Distributions) *Let D be a fixed distribution and $\mathcal{D}$ be a family of distributions. We denote by $\mathcal{B}(\mathcal{D}, D)$ the decision (or hypothesis testing) problem in which the input distribution D' is promised to satisfy either (a) $D' = D$ or (b) $D' \in \mathcal{D}$, and the goal is to distinguish between the two cases.*

Definition 64 (Pairwise Correlation) *The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \mathbb{R}^d \to \mathbb{R}_+$ with respect to a distribution with density $D : \mathbb{R}^d \to \mathbb{R}_+$, where the support of D contains the supports of D_1 and D_2, is defined as $\chi_D(D_1, D_2) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} D_1(x) D_2(x)/D(x)\,\mathrm{d}x - 1$.*

Definition 65 *We say that a set of s distributions $\mathcal{D} = \{D_1, \ldots, D_s\}$ over $\mathbb{R}^d$ is (γ, β)-correlated relative to a distribution D if $|\chi_D(D_i, D_j)| \le \gamma$ for all $i \neq j$, and $|\chi_D(D_i, D_j)| \le \beta$ for $i = j$.*

Definition 66 (Statistical Query Dimension) *For $\beta, \gamma > 0$, a decision problem $\mathcal{B}(\mathcal{D}, D)$, where D is a fixed distribution and $\mathcal{D}$ is a family of distributions, let s be the maximum integer such that there exists a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ such that $\mathcal{D}_D$ is (γ, β)-correlated relative to D and $|\mathcal{D}_D| \ge s$. The* statistical query dimension *with pairwise correlations (γ, β) of $\mathcal{B}$ is defined to be s, and is denoted by $\mathrm{SD}(\mathcal{B}, \gamma, \beta)$.*

A lower bound on the SQ dimension of a decision problem implies a lower bound on the complexity of any SQ algorithm for the problem via the following standard result.

Lemma 67 *Let $\mathcal{B}(\mathcal{D}, D)$ be a decision problem, where D is the reference distribution and $\mathcal{D}$ is a class of distributions. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{B}, \gamma, \beta)$. For any $\gamma' > 0$, any SQ algorithm for $\mathcal{B}$ requires queries of tolerance at most $\sqrt{\gamma + \gamma'}$ or makes at least $s\gamma'/(\beta - \gamma)$ queries.*

SPARSE NON-GAUSSIAN COMPONENT ANALYSIS

We will focus on a specific kind of decision problem given by Problem 68 below.

Problem 68 (Sparse Non-Gaussian Component Analysis) *Let A be a distribution on $\mathbb{R}$. For a unit vector v, we denote by $P_{A,v}$ the distribution with the density $P_{A,v}(x) := A(v^T x)\phi_{\perp v}(x)$, where $\phi_{\perp v}(x) = \exp\left(-\|x - (v^T x)v\|_2^2/2\right)/(2\pi)^{(d-1)/2}$, i.e., the distribution that coincides with A on the direction v and is standard Gaussian in every orthogonal direction. We define the following hypothesis testing problem:*

- *H_0: The underlying distribution is $\mathcal{N}(0, I_d)$.*
- *H_1: The underlying distribution is $P_{A,v}$, for some unit vector v that is k-sparse.*

Specializing the result of Lemma 67 for the sparse non-Gaussian component analysis, gives the following SQ lower bound. The proof is standard and is deferred to Appendix F.

Corollary 69 *Let $k, d, m \in \mathbb{Z}_+$ with $k \le \sqrt{d}$. For any distribution A on $\mathbb{R}$ that matches its first m moments with $\mathcal{N}(0, 1)$, any constant $0 < c < 1$, and any SQ algorithm $\mathcal{A}$ that solves the hypothesis testing Problem 68, $\mathcal{A}$ either makes $\Omega(d^{ck^c/8} k^{-(m+1)(1-c)})$ many queries or makes at least one query with tolerance at most $2^{(m/2+1)} k^{-(m+1)(1/2-c/2)} \sqrt{\chi^2(A, \mathcal{N}(0,1))}$.*

When proving our main results, we will apply Corollary 69 to different choices of A to get Theorem 8 and Theorem 6.

FROM ESTIMATION TO HYPOTHESIS TESTING

Our lower bounds will be for estimating the unknown sparse mean in ℓ_2-error[3]. To establish these results, we prove a stronger claim: We consider a hypothesis testing version of the robust sparse mean recovery (Problem 71). We first prove that this is an easier task than the corresponding estimation problem (Problem 70) in Claim 72. We then show hardness of the hypothesis testing problem in the SQ model.

Problem 70 (Robust Sparse Mean Estimation) *Fix $\rho > 0$. Let $\mathcal{D}$ be a family of distributions such that the mean of each distribution D in $\mathcal{D}$ is k-sparse and has norm at most ρ. Given access to the mixture distribution $(1-\epsilon)D + \epsilon B$, for some (unknown) $D \in \mathcal{D}$ and some arbitrary distribution B, the goal is to find a vector $u \in \mathbb{R}^d$ such that $\|u - \mathbf{E}_{X\sim D}[X]\|_2 < \rho/2$.*

Problem 71 (Robust Sparse Mean Hypothesis Testing) *Fix $\rho > 0$. Let $\mathcal{D}$ be a family of distributions such that the mean of each distribution D in $\mathcal{D}$ is k-sparse and has norm* exactly *ρ. We define the following hypothesis testing problem:*

- *H_0: The underlying distribution is $\mathcal{N}(0, I_d)$.*
- *H_1: The underlying distribution is $(1-\epsilon)D+\epsilon B$, for a $D\in\mathcal{D}$ and an arbitrary distribution B.*

Claim 72 (Reduction) *Given an algorithm $\mathcal{A}$ that solves Problem 70 for some $\mathcal{D}$, then there exists another algorithm that solves Problem 71 for $\mathcal{D}'$, where $\mathcal{D}'$ is the set of all distributions in $\mathcal{D}$ that have norm exactly ρ.*

Proof The algorithm is the following: Let u be the estimate returned by $\mathcal{A}$. If $\|u\|_2 < \rho/2$, then return H_0, otherwise return H_1. Since, in both the null and alternative hypothesis, u is guaranteed to be within $\rho/2$ of the true mean, the correctness follows. ■

E.2. SQ Lower Bound for Robust Gaussian Sparse Mean Estimation with Unknown Covariance

We consider the task of robust sparse mean estimation of a Gaussian distribution, $\mathcal{N}(\mu, \Sigma)$, where μ is k-sparse and Σ is unknown and bounded, $\Sigma \preceq I$. Information-theoretically $O((k\log(d/k))/\epsilon^2)$ samples suffice to obtain an estimate $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 = O(\epsilon)$. The polynomial-time algorithm of Balakrishnan et al. (2017) uses $O((k^2 \log d)/\epsilon^2)$ samples and can be shown to achieve error $O(\sqrt{\epsilon})$ for robust sparse mean estimation in this setting. The main result of this section is an SQ lower bound roughly stating that any SQ algorithm that achieves error $o(\sqrt{\epsilon})$ either uses super-polynomially many number of queries or uses a single query that requires k^4 samples to simulate.

Theorem 73 (Formal version of Theorem 7) *Let $k, d \in \mathbb{Z}_+$ with $k \leq \sqrt{d}$, $0 < c < 1, 0 < \epsilon < 1/2$, and $c_1 = 1/10001$. Let $\mathcal{A}$ be an SQ algorithm that is given access to a distribution of the form $(1-\epsilon)\mathcal{N}(c_1\sqrt{\epsilon}v, I_d - (1/3)vv^T) + \epsilon B$, where v is some unit k-sparse vector of $\mathbb{R}^d$ and B is some arbitrary noise distribution. If the output of $\mathcal{A}$ is a vector u such that $\|u - c_1\sqrt{\epsilon}v\|_2 \leq c_1\sqrt{\epsilon}/4$, then $\mathcal{A}$ does one of the following:*

3. Recall that estimating a k-sparse vector in ℓ_2-norm is an easier problem than estimating an arbitrary vector in $(2, k)$-norm.

- *Makes* $\Omega(d^{ck^c/8}k^{-4+4c})$ *queries,*
- *or makes at least one query with tolerance* $O(k^{-2+2c}e^{O(1/\epsilon)})$.

Proof First, we note that there exists a one-dimensional distribution which is an ϵ-corrupted version of a Gaussian with mean $c_1\sqrt{\epsilon}$ and matches the first three moments with $\mathcal{N}(0,1)$.

Lemma 74 (Lemma E.2 of Diakonikolas et al. (2019b)) *Let* $\mu = c_1\sqrt{\epsilon}$ *with* $c_1 = 1/10001$. *For any* $0 < \epsilon < 1$, *there exists a distribution* B *on* $\mathbb{R}$ *such that the mixture* $A = (1-\epsilon)\mathcal{N}(\mu, 2/3) + \epsilon B$ *matches the first three moments with* $\mathcal{N}(0,1)$ *and* $\chi^2(A, \mathcal{N}(0,1)) = e^{O(1/\epsilon)}$.

We now follow the argument of Appendix E.1. We consider Problems 70 and 71 specialized to the case where $\mathcal{D}$ is the family of distributions $\mathcal{D} = \{(1-\epsilon)\mathcal{N}(c_1\sqrt{\epsilon}v, I_d - (1/3)vv^T) + \epsilon B'\}_{v\in\mathcal{U}_k}$, where $\mathcal{U}_k$ is the set of k-sparse unit vectors and B' denotes a distribution whose one-dimensional projection along v coincides with B and every orthogonal projection is standard Gaussian, i.e., $B' = P_{B,v}$. Given the reduction of Claim 72, in order to prove Theorem 73, it remains to show that Problem 71 is hard in the SQ model. To this end, we note that this is the same problem as Problem 68 with the distribution A being that of Lemma 74. An application of Corollary 69 completes the proof of Theorem 73.

■

E.3. SQ Lower Bound for Robust Sparse Mean Estimation of Distributions with Bounded t-th Moment

In this section, we will show that any SQ algorithm to obtain error $o(\epsilon^{1-1/t})$ either uses super-polynomially many queries or uses queries with tolerance $k^{-\Omega(t)}$. In order to state our results formally, we define the following distribution class: let $\mathcal{P}_{k,t}$ be the class of all distributions P that satisfy the following:

1. The mean of the distribution P, μ, is k-sparse, and $\|\mu\|_2 \le 1$.
2. P has subgaussian tails, i.e., there is a constant c such that for for all unit vectors v and $i \in \mathbb{N}$, $(\mathbf{E}_{X\sim P}[|v^T(X-\mu)|^i])^{1/i} \le c\sqrt{i}$.
3. For a large constant C, there is an SoS proof of the following inequality:

$$\{\|v\|_2^2 = 1\} \vdash_{O(t)}^{v} \mathop{\mathbf{E}}_{X\sim P}[\langle X-\mu, v\rangle^t]^2 \le (Ct)^t \ .$$

We prove the following:

Theorem 75 (Formal version of Theorem 6) *Let* $d, k, t \in \mathbb{Z}_+$ *with* $k \le \sqrt{d}$, *let* $0 < \epsilon = (O(t))^{-t}$, $0 \le C < 1/2000$, $0 < c < 1$, *and* $\delta = C\epsilon^{1-1/t}/t$. *Let* $\mathcal{A}$ *be an SQ algorithm that, given access to a distribution of the form* $(1-\epsilon)P + \epsilon B$, *where* $P \in \mathcal{P}_{k,t}$ *(defined above) and* B *is arbitrary,* $\mathcal{A}$ *is guaranteed to find a vector* $\hat{\mu}$ *such that* $\|\hat{\mu} - \mathbf{E}_{X\sim P}[X]\|_2 \le \delta$. *Then* $\mathcal{A}$ *does one of the following:*

- *Makes* $\Omega(d^{ck^c/8}k^{-(t+1)(1-c)})$ *queries.*

- *Makes at least one query with tolerance* $O\left(k^{-(t+1)(1/2-c/2)}2^{(t/2+1)}e^{O(\delta^2/\epsilon^2)}\right)$.

The rest of the section is dedicated to proving Theorem 75. We first show the existence of a one-dimensional distribution A that matches the first t moments with $\mathcal{N}(0,1)$ and is an ϵ-corruption of a distribution with mean $\Omega(\frac{1}{t}\epsilon^{1-1/t})$ and bounded t-th moments. At a high level, we follow the structure of (Diakonikolas et al., 2017b, Proposition 5.2) and (Diakonikolas et al., 2018b, Lemma 5.5). In particular, (Diakonikolas et al., 2018b, Lemma 5.5) establishes an analogous result to Lemma 76 below but in the large ϵ setting, i.e., $\epsilon \to 1$, and thus it is not applicable here. We also show that the family of hard distributions in Theorem 75 has certifiably bounded moments. We defer this analysis to Appendix F.2.

Lemma 76 *Fix an $t \in \mathbb{Z}_+$ and $\epsilon = O(t)^{-t}$. There exists a distribution A over $\mathbb{R}$ such that the following holds:*

1. *There exist two distributions Q_1 and Q_2 such that $A = (1-\epsilon)Q_1 + \epsilon Q_2$.*
2. *A matches first t moments with $\mathcal{N}(0,1)$.*
3. *$\mathbf{E}_{X\sim Q_1}[X] = \delta$, where $\delta = \frac{1}{2000}\frac{1}{t}\epsilon^{1-1/t}$.*
4. *For all $i \geq 1$, $\left(\mathbf{E}_{X\sim Q_1}[|X-\delta|^i]\right)^{1/i} = O(\sqrt{i})$.*
5. *$\chi^2(A, \mathcal{N}(0,1)) < \exp\left(O(\delta^2/\epsilon^2)\right)$.*

Proof Let $G(x)$ be the pdf of the standard normal $\mathcal{N}(0,1)$. Thus $G(x-\delta)$ represents the pdf of $\mathcal{N}(\delta,1)$. We will choose A of the following form:

$$Q_1(x) = G(x-\delta) + \frac{1}{1-\epsilon}p(x)\mathbf{1}_{[-1,1]}(x), \quad Q_2(x) = G(x-\delta'),$$

where $p(\cdot)$ is a degree t polynomial (to be chosen below) and $\delta' = -(1-\epsilon)\delta/\epsilon$. To ensure that Q_1 is a valid distribution and has mean δ, the following suffices since $|\delta| \leq 0.1$ and $\epsilon \leq 0.1$:

1. $\int_{-1}^{1} p(x)\mathrm{d}x = 0$,
2. $\max_{x\in[-1,1]} |p(x)| \leq 0.1$,
3. $\int_{-1}^{1} p(x)x\mathrm{d}x = 0$.

Let P_i be the i-th Legendre polynomial. We will choose p to be of the following form for $a_i \in \mathbb{R}$:

$$p(x) = \sum_{i=0}^{t} a_i P_i(x),$$

where $a_0 = a_1 = 0$.

Fact 77 *Let P_i be the i-th Legendre polynomial. We have the following:*

L.1 *$P_0(x) = 1$ and $P_1(x) = x$.*

L.2 $\int_{-1}^{1} P_i(x)P_j(x)dx = \frac{2}{2i+1}\delta_{i,j}$.

L.3 $\max_{x\in[-1,1]} |P_i(x)| \le 1$.

L.4 $\{P_i\}_{i=0}^{k}$ *form a basis of polynomials of degree up to* k.

Fact 78 *Let* h_i *be the* i*-th normalized probabilist's polynomials and let* $X \sim \mathcal{N}(0,1)$.

H.1 $\mathbf{E}[h_i(X)h_j(X)] = \delta_{i,j}$.

H.2 $\mathbf{E}[h_i(X+\mu)] = \frac{1}{\sqrt{i!}}\mathbf{E}[H_{e_i}(X+\mu)] = \frac{\mu^i}{\sqrt{i!}}$.

H.3 $\{h_i\}_{i=0}^{k}$ *form a basis of polynomials of degree up to* k.

Using L.1 and L.2 we have that $\int_{-1}^{1} p(x)\mathrm{d}x = 0$ and $\int_{-1}^{1} p(x)x\mathrm{d}x = 0$. Using L.3, we have that $\max_{x\in[-1,1]} |p(x)| \le \sum_{i=1}^{t} |a_i|$. We will now ensure that it is possible to match moments while keeping $\sum_i |a_i|$ small.

Recall that in order to match the first t moments of A with $\mathcal{N}(0,1)$, we need to ensure the following holds for all $i \in \{0,\dots,t\}$:

$$(1-\epsilon)\int_{-\infty}^{\infty} x^i G(x-\delta)dx + \int_{-1}^{1} x^i p(x)dx + \epsilon\int_{-\infty}^{\infty} x^i G(x-\delta')dx = \int_{-\infty}^{\infty} x^i G(x)dx.$$

Equivalently, letting $X \sim \mathcal{N}(0,1)$, we need the following for all $i \in \{0,\dots,t\}$:

$$\int_{-1}^{1} x^i p(x)dx = \mathop{\mathbf{E}}_{X\sim\mathcal{N}(0,1)}[X^i - (1-\epsilon)(X+\delta)^i - \epsilon(X+\delta')^i].$$

By L.4, it suffices to ensure the following for all $i \in \{0,\dots,t\}$:

$$\int_{-1}^{1} P_i(x)p(x)dx = \mathop{\mathbf{E}}_{X\sim\mathcal{N}(0,1)}[P_i(X) - (1-\epsilon)P_i(X+\delta) - \epsilon P_i(X+\delta')]. \tag{20}$$

Since $\int_{-1}^{1} p(x)\mathrm{d}x = 0$, $P_0(x) = 1$, and $P_1(x) = x$, we have that Equation (20) holds for $i = 0$ and $i = 1$ as both sides are zero. Note that for any $i \in \{0,\dots,t\}$, the left-hand side above can be calculated using L.2:

$$\int_{-1}^{1} P_i(x)p(x)dx = \sum_{j=0}^{t}\int_{-1}^{1} a_j P_i(x)P_j(x) = \frac{2a_i}{2i+1}. \tag{21}$$

We will now bound the expression on the right-hand side in Equation (20) to show that a_i are small. Let h_i be the i-th normalized probabilist's Hermite polynomials. Using H.3, we can write $P_i(x) = \sum_{j=0}^{i} b_{i,j}h_j(x)$ for some $b_{i,j} \in \mathbb{R}$. We now calculate the right-hand side of Equation (20) as follows for a fixed $i \in \{0,\dots,t\}$:

$$\mathop{\mathbf{E}}_{X\sim\mathcal{N}(0,1)}[P_i(X) - (1-\epsilon)P_i(X+\delta) - \epsilon P_i(X+\delta')]$$

$$= \sum_{j=0}^{i} b_{i,j} \left(\mathop{\mathbf{E}}_{X\sim\mathcal{N}(0,1)}[h_j(X)] - (1-\epsilon) \mathop{\mathbf{E}}_{X\sim\mathcal{N}(0,1)}[h_j(X+\delta) - \epsilon h_j(X+\delta')] \right)$$
$$= \sum_{j=0}^{i} b_{i,j} \left(0 - (1-\epsilon)\frac{\delta^j}{\sqrt{j!}} - \epsilon\frac{(\delta')^j}{\sqrt{j!}} \right)$$
$$= \sum_{j=0}^{i} \frac{-1}{\sqrt{j!}} b_{i,j} \left((1-\epsilon)\delta^j + \epsilon(\delta')^j \right) ,$$

where the second line uses H.2. From the proof of (Diakonikolas et al., 2018b, Claim 5.6), we have that $\sum_{j=0}^{i} b_{i,j}^2 = O((2i)^i)$. We are now ready to calculate the upper bound on $|a_i|$ using Equation (21):

$$\begin{aligned}
|a_i| &= \left(\frac{2i+1}{2}\right) \left| \sum_{j=0}^{i} \frac{-1}{\sqrt{j!}} b_{i,j} ((1-\epsilon)\delta^j + \epsilon(\delta')^j) \right| \\
&\le 2i \sum_{j=0}^{i} \frac{1}{\sqrt{j!}} |b_{i,j}| \left((1-\epsilon)|\delta|^j + \epsilon|\delta'|^j \right) \\
&\le 4i \sum_{j=0}^{i} \frac{1}{\sqrt{j!}} |b_{i,j}| \epsilon \left|\delta'\right|^j \\
&\le 8i^2 \epsilon \max(|\delta'|, |\delta'|^i) \max_{j\in[i]} |b_{i,j}| \\
&\le (2i)^{i+4} \epsilon \max(|\delta'|, |\delta'|^i) ,
\end{aligned}$$

where the third line uses that $\delta' = -(1-\epsilon)\delta/\epsilon$ thus $(1-\epsilon)|\delta|^j = |\delta'|^j \epsilon(\epsilon/(1-\epsilon))^{j-1} \le |\delta'|^j \epsilon$. Thus, we get the following:

$$\begin{aligned}
\max_{x\in[-1,1]} |p(x)| &\le \sum_{i=1}^{t} |a_i| \le \sum_{i=1}^{t} (2i)^{i+4} \epsilon \max(|\delta'|, |\delta'|^i) \le (2t)^{t+5} \epsilon \max(|\delta'|, |\delta'|^t) \\
&\le \epsilon |100t|^t \max(|\delta'|, |\delta'|^t) , \qquad (22)
\end{aligned}$$

where we bounded the sum by t times its last term. We would like to show that the last expression in Equation (22) is less than 0.1 when $\delta = C\epsilon^{1-1/t}/t$ for some constant C. Note that this choice of δ implies that $|\delta'| \ge 0.5(\delta/\epsilon) = 0.5C(\epsilon^{-1/t}/t)$, which is larger than 1 when $\epsilon = (O(t))^{-t}$. Thus the last expression in Equation (22) is at most $\epsilon(100t|\delta'|)^t \le \epsilon 100^t t^t \delta^t/\epsilon^t \le \epsilon 100^t t^t C^t \epsilon^{t-1}/(\epsilon^t t^t) = (100C)^t$, which is less than 0.1 if $C \le 0.0005$.

Finally, the bound on the t-moment of Q_1 centered around δ follows by combining the moment bounds of $\mathcal{N}(\delta, 1)$ and noting that $p(\cdot)$ modifies the Gaussian only on the interval $[-1, 1]$.

It remains to bound the χ^2-divergence between our distribution A and $\mathcal{N}(0, 1)$.

$$\begin{aligned}
1 + \chi^2(A, \mathcal{N}(0,1)) &= \int_{-\infty}^{\infty} \frac{1}{G(x)} ((1-\epsilon)G(x-\delta) + p(x)\mathbf{1}_{[-1,1]} + \epsilon G(x-\delta'))^2 \mathrm{d}x \\
&\le 9 \left(\int_{-\infty}^{\infty} \frac{G^2(x-\delta)}{G(x)} \mathrm{d}x + \int_{-1}^{1} \frac{p^2(x)}{G(x)} \mathrm{d}x + \epsilon^2 \int_{-\infty}^{\infty} \frac{G^2(x-\delta')}{G(x)} \mathrm{d}x \right) .
\end{aligned}$$

Working with each term separately, the first one is bounded as

$$\int_{-\infty}^{\infty} \frac{G^2(x-\delta)}{G(x)} \mathrm{d}x \leq 1 + \chi^2(\mathcal{N}(\delta,1),\mathcal{N}(0,1)) = e^{\delta^2} ,$$

the last one is similarly bounded above by $\epsilon^2 e^{\delta'^2}$ and for the first one we have that

$$\int_{-1}^{1} \frac{p^2(x)}{G(x)} \mathrm{d}x \leq \left(\max_{x\in[-1,1]} |p(x)| \right) \max_{x\in[-1,1]} \frac{1}{G(x)} = O(1) .$$

Given $(\delta')^2 = \Theta(\delta^2/\epsilon^2)$, all three terms are at most $\exp(O(\delta^2/\epsilon^2)$, therefore we have that $\chi^2(A,\mathcal{N}(0,1)) = \exp\left(O(\delta^2/\epsilon^2)\right)$. ■

E.3.1. PROOF OF THEOREM 75

Proof [Proof of Theorem 75] We will prove Theorem 75 using Lemma 76 with the argument of Appendix E.1. Let A be the distribution from Lemma 76. We consider Problems 70 and 71 with $\mathcal{D} = \{P_{A,v}\}_{v\in\mathcal{U}_k}$ (using the notation from Problem 68). Using the notation of Lemma 76, we see that every $P_{A,v}$ in this choice of $\mathcal{D}$ is of the following form $P_{A,v} = (1-\epsilon)P_{Q_1,v} + \epsilon P_{Q_2,v}$, where $P_{Q_1,v}$ belongs to $\mathcal{P}_{k,t}$ as defined in the beginning of this section: (i) its mean is k-sparse (since v is k-sparse), (ii) it satisfies subgaussian tail bounds (since Q_1 has subgaussian tails, see Lemma 76), and (iii) it has t-certifiably bounded moments (Claim 81). Problem 70 is then the same as Problem 68. By the reduction of Claim 72, it remains to show the SQ-hardness of the latter problem. We then use Corollary 69. ■

As a note, by simply replacing the set S of the k-sparse direction of Fact 80 by an analogous set of dense 2^{d^c} vectors (see, e.g., (Diakonikolas et al., 2017b, Lemma 3.7)) we can get an analog of the previous theorem for the dense case.

Theorem 79 (SQ Lower Bound in Dense Case) *Let $t \in \mathbb{Z}_+$, $0 < \epsilon = (O(t))^{-t}$, $C < 1/2000$, $0 < c < 1/2$,and $\delta = C\epsilon^{1-1/t}/t$. Any SQ algorithm that, given access to a distribution of the form $(1-\epsilon)P + \epsilon N$ where P is a distribution with $\mathbf{E}_{X\sim P}[|v^T X|^i]^{1/i} = O(\sqrt{i})$ for every $i \leq t$ and every $v \in \mathcal{S}^{d-1}$ and finds a vector $\hat{\mu}$ such that $\|\hat{\mu} - \mathbf{E}_{X\sim P}[X]\|_2 \leq \delta$ does one of the following:*

- *Makes $2^{\Omega(d^c)} d^{-(t+1)(1/2-c)}$ queries.*
- *Makes at least one query with tolerance $\left(O(d)^{-(t+1)(1/4-c/2)} e^{O(\delta^2/\epsilon^2)}\right)$.*

Appendix F. Omitted Details from Appendix E

We start by providing additional background of Appendix E. In Appendix F.2, we show that the hard instance in Theorem 75 has SoS-certifiable bounded moments. Finally, we present the lower bounds against low-degree polynomial tests in Appendix F.3.

F.1. Omitted Background

We provide the proof of Corollary 69 below.

Proof [Proof of Corollary 69] Problem 68 is a decision problem in the sense of Definition 63, where $D = \mathcal{N}(0, I_d)$ and $\mathcal{D} = \{P_{A,v}\}_{v \in \mathcal{U}_k}$, where $\mathcal{U}_k$ is the set of all k-sparse unit vectors. We calculate the SQ-dimension of $\mathcal{B}(\mathcal{D}, D)$ as follows: Let $\mathcal{D}_D$ be the subset of $\mathcal{D}$ defined as $\mathcal{D}_D = \{P_{A,v}\}_{v \in S}$, for S being the set from the following fact:

Fact 80 (Lemma 6.7 in Diakonikolas et al. (2017b)) *Fix a constant $0 < c < 1$ and let $k \leq \sqrt{d}$. There exists a set S of k-sparse unit vectors on $\mathbb{R}^d$ of cardinality $|S| = \lfloor d^{ck^c/8} \rfloor$ such that for each pair of distinct vectors $v, v' \in S$ we have that $|v^T v'| \leq 2k^{c-1}$.*

Using Lemma 3.4 from Diakonikolas et al. (2017b), for $v, v' \in S$ we have that

$$\chi_D(P_{A,v}, P_{A,v'}) \leq |v^T v'|^{m+1} \chi^2(A, \mathcal{N}(0,1)) = (2k^{c-1})^{m+1} \chi^2(A, \mathcal{N}(0,1)) \,,$$

therefore the statistical dimension is $\mathrm{SD}(\mathcal{B}, \gamma, \beta) = \Omega(d^{ck^c/8})$ for $\gamma = 2^{m+1} k^{(c-1)(m+1)} \chi^2(A, \mathcal{N}(0,1))$ and $\beta = \chi^2(A, \mathcal{N}(0,1))$. An application of Lemma 67 with $\gamma' = \gamma$ yields that any SQ algorithm, either makes at least one query of tolerance at most $2^{(m/2+1)} k^{-(m+1)(1/2-c/2)} \sqrt{\chi^2(A, \mathcal{N}(0,1))}$ or makes at least the following number of queries:

$$\begin{aligned} \Omega(d^{ck^c/8}) \frac{2^{m+1} k^{(c-1)(m+1)} \chi^2(A, \mathcal{N}(0,1))}{\chi^2(A, \mathcal{N}(0,1)) - 2^{m+1} k^{(c-1)(m+1)} \chi^2(A, \mathcal{N}(0,1))} &\geq \Omega(d^{ck^c/8}) 2^{m+1} k^{(c-1)(m+1)} \\ &\geq \Omega\left(d^{ck^c/8} k^{-(m+1)(1-c)} \right) \,. \end{aligned}$$

This completes the proof. ■

F.2. SoS Certifiability of Hard Instances

In this section, we will show that the hard instances in our proof have SoS certifiable bounded t-th moments.

Claim 81 *Fix a $t \in \mathbb{N}$ such that t is a power of 2. Denote by $G(x)$ the pdf of $\mathcal{N}(0,1)$. Let $Q_1(x), Q_2(x)$ be the distributions from the proof of Lemma 76, and define $Q := (1-\epsilon)Q_1 + \epsilon Q_2$ where $\delta = 1/(2000t)\epsilon^{1-1/t}$, $\delta' = -(1-\epsilon)\delta/\epsilon$, and $|\delta'| \geq 1$. Recall that the first t moments of Q has moments match with $\mathcal{N}(0,1)$. Let P_1, P_2, P the distributions that have Q_1, Q_2, and Q, respectively, in the u direction and are standard Gaussian in all perpendicular directions. Let $\mathcal{A} := \{\sum_i v_i^2 = 1\}$, and define $\mu := \mathbf{E}_{X \sim P_1}[X]$. Then P_1 has SoS certifiably bounded moments, i.e., there exists a constant $C > 0$ such that*

$$\mathcal{A} \vdash_{O(t)} \mathop{\mathbf{E}}_{X \sim P_1} [\langle v, X - \mu \rangle^t]^2 \leq (Ct)^t \,.$$

Proof We will use the following claim that depends on the SoS triangle inequality (Fact 37).

Claim 82 *Let P be a distribution over $\mathbb{R}^d$ with mean μ and define $p_i(v) = \mathbf{E}[\langle v, X\rangle^i]$ and $p'_i(v) = \mathbf{E}[\langle v, X-\mu\rangle^i]$. Let $\mathcal{A} = \{\sum_i v_i^2 = 1\}$. There exists a $C > 0$ such that the following holds: Let t be a power of 2. If $\mathcal{A} \vdash_{O(t)} p'_i(v)^2 \le R$ for some $R \ge 0$ and all $i \in [t]$, then $\mathcal{A} \vdash_{O(t)} p_t(v)^2 \le C^t R \max(1, \|\mu\|_2^{2t})$. Similarly, if $\mathcal{A} \vdash_{O(t)} p_i(v)^2 \le R$ for some $R \ge 0$ and all $i \in [t]$, then $\mathcal{A} \vdash_{O(t)} p'_t(v)^2 \le C^t R \max(1, \|\mu\|_2^{2t})$.*

Proof For $i \in \{0, 1, \dots, t\}$, let $C_i = \binom{t}{i}$. We have the following polynomial equalities:

$$p'_t(v) = \sum_{i=0}^{t} (-1)^i C_i p_i(v) \langle \mu, v\rangle^{t-i},$$
$$p_t(v) = \sum_{i=0}^{t} C_i p'_i(v) \langle \mu, v\rangle^{t-i}.$$

Applying SoS triangle inequality (Fact 37), using the fact that $\mathcal{A} \vdash_{2i}^{v} \langle v, \mu\rangle^{2i} \le \|\mu\|_2^{2i}$, and $C_i \le 2^t$, we get the desired claim. ■

Note that the mean of P_1 is $u\delta$ and $\|u\delta\|_2 \le 1$, and thus Claim 82 implies that it suffices to show that $\mathcal{A} \vdash_{O(t)} \mathbf{E}_{X\sim P_1}[\langle v, X\rangle^i]^2 \le (O(t))^t$ for all $i \in [t]$.

Note that P_2 is $\mathcal{N}(u\delta', I)$ and P matches the moments of $\mathcal{N}(0, I)$ up to degree t in every direction. Thus Fact 29 implies the following: for all $i \in [t]$

$$\mathcal{A} \vdash_{O(t)} (Ct)^t - (\mathop{\mathbf{E}}_{X\sim P}[\langle v, X\rangle^i])^2 \ge 0, \tag{23}$$
$$\mathcal{A} \vdash_{O(t)} (C't)^t - (\mathop{\mathbf{E}}_{X\sim P_2}[(\langle v, X\rangle - \delta'\langle u, v\rangle)^i])^2 \ge 0. \tag{24}$$

Suppose for now that P_2 satisfies the following, which we will establish shortly: there exists a constant C'' such that for all $i \in [t]$,

$$\mathcal{A} \vdash_{O(t)} \epsilon^2 \mathop{\mathbf{E}}_{X\sim P_2}[\langle v, X\rangle^i]^2 \le (C''t)^t. \tag{25}$$

To show $\mathcal{A} \vdash_{O(t)} \mathbf{E}_{X\sim P_1}[\langle v, X\rangle^i]^2 \le (O(t))^t$, we proceed as follows:

$$\begin{aligned}
\mathcal{A} \vdash_{O(t)} \mathop{\mathbf{E}}_{X\sim P_1}[\langle v, X\rangle^i]^2 &= (1/(1-\epsilon))^2 (\mathop{\mathbf{E}}_{X\sim P}[\langle v, X\rangle^i] - \epsilon \mathop{\mathbf{E}}_{X\sim P_2}[\langle v, X\rangle^i])^2 \\
&\le 2/(1-\epsilon)^2 \left(\mathop{\mathbf{E}}_{X\sim P}[\langle v, X\rangle^i]^2 + \epsilon^2 \mathop{\mathbf{E}}_{X\sim P_2}[\langle v, X\rangle^i]^2\right) \\
&\le (O(t))^t,
\end{aligned}$$

where the first inequality uses SoS triangle inequality (Fact 37) and the second inequality uses Equation (23) for the first term and Equation (25) for the second term. Thus it only remains to show that Equation (25) holds to complete the proof.

Note that the mean of P_2 has norm δ' and $|\delta'| \ge 1$. Claim 82 and Equation (24) imply the following:

$$\mathcal{A} \vdash_{O(t)} \epsilon^2 \mathop{\mathbf{E}}_{X\sim P_2}[\langle v, X\rangle^i]^2 \le (CC')^t \epsilon^2 |\delta'|^{2t}.$$

By definition, $\epsilon^2|\delta'|^{2t} \le \epsilon^2(\delta/\epsilon)^{2t} \le \epsilon^2(\epsilon^{-1/t})^{2t} = 1$. This completes the proof. ■

F.3. Implications for Low-Degree Polynomial Algorithms

We can get quantitatively similar lower bounds in the low-degree model of computation using its connection with the SQ model Brennan et al. (2021). The result of this section, roughly speaking, is that any polynomial algorithm for sparse non-Gaussian component analysis where the non-Gaussian component matches m moments with $\mathcal{N}(0,1)$, either uses more than k^{m+1} samples or has degree more than $k^{\Omega(1)}$ (which in the worst case requires $d^{k^{\Omega(1)}}$ monomial terms that need to be computed). Plugging $m = 3$ yields an analog of Theorem 73 and letting m be equal to the number of bounded moments of the inliers' distribution, i.e., t, gives an analog of Theorem 75.

Brennan et al. (2021) uses a slightly different version of hypothesis testing problems, where in the alternative hypothesis, the ground truth is chosen according to a probability measure.

Problem 83 (Non-Gaussian Component Hypothesis Testing with Uniform Prior) *Let a distribution A on $\mathbb{R}$. For a unit vector v, we denote by $P_{A,v}$ the distribution with density $P_{A,v}(x) := A(v^T x)\phi_{\perp v}(x)$, where $\phi_{\perp v}(x) = \exp\left(-\|x-(v^Tx)v\|_2^2/2\right)/(2\pi)^{(d-1)/2}$, i.e., the distribution that coincides with A on the direction v and is standard Gaussian in every orthogonal direction. Let S be the set of nearly orthogonal vectors from Fact 80. Let $\mathcal{S} = \{P_{A,v}\}_{u\in S}$. We define the simple hypothesis testing problem where the null hypothesis is $\mathcal{N}(0, I_d)$ and the alternative hypothesis is $P_{A,v}$ for some v uniformly selected from S.*

We now describe the model in more detail. We will consider tests that are thresholded polynomials of low-degree, i.e., output H_1 if the value of the polynomial exceeds a threshold and H_0 otherwise. We need the following notation and definitions. For a distribution D over $\mathcal{X}$, we use $D^{\otimes n}$ to denote the joint distribution of n i.i.d. samples from D. For two functions $f : \mathcal{X} \to \mathbb{R}$, $g : \mathcal{X} \to R$ and a distribution D, we use $\langle f, g\rangle_D$ to denote the inner product $\mathbf{E}_{X\sim D}[f(X)g(X)]$. We use $\|f\|_D$ to denote $\sqrt{\langle f, f\rangle_D}$. We say that a polynomial $f(x_1, \ldots, x_n) : \mathbb{R}^{n\times d} \to \mathbb{R}$ has sample-wise degree (r, ℓ) if each monomial uses at most ℓ different samples from $x_1, \ldots, x_n$ and uses degree at most r for each of them. Let $\mathcal{C}_{r,\ell}$ be linear space of all polynomials of sample-wise degree (r, ℓ) with respect to the inner product defined above. For a function $f : \mathbb{R}^{n\times d} \to \mathbb{R}$, we use $f^{\le r,\ell}$ to be the orthogonal projection onto $\mathcal{C}_{r,\ell}$ with respect to the inner product $\langle\cdot,\cdot\rangle_{D_0^{\otimes n}}$. Finally, for the null distribution D_0 and a distribution P, define the likelihood ratio $\overline{P}^{\otimes n}(x) := P^{\otimes n}(x)/D_0^{\otimes n}(x)$.

Definition 84 (n-sample τ-distinguisher) *For the hypothesis testing problem between two distributions D_0 (null distribution) and D_1 (alternate distribution) over $\mathcal{X}$, we say that a function $p : \mathcal{X}^n \to \mathbb{R}$ is an n-sample τ-distinguisher if $|\mathbf{E}_{X\sim D_0^{\otimes n}}[p(X)] - \mathbf{E}_{X\sim D_1^{\otimes n}}[p(X)]| \ge \tau\sqrt{\mathbf{Var}_{X\sim D_0^{\otimes n}}[p(X)]}$. We call τ the* advantage *of the polynomial p.*

Note that if a function p has advantage τ, then the Chebyshev's inequality implies that one can furnish a test $p' : \mathcal{X}^n \to \{D_0, D_1\}$ by thresholding p such that the probability of error under the null distribution is at most $O(1/\tau^2)$. We will think of the advantage τ as the proxy for the inverse of the probability of error (see Theorem 4.3 in Kunisky et al. (2019) for a formalization of this intuition under certain assumptions) and we will show that the advantage of all polynomials up to a certain

degree is $O(1)$. It can be shown that for hypothesis testing problems of the form of Problem 83, the best possible advantage among all polynomials in $\mathcal{C}_{r,\ell}$ is captured by the low-degree likelihood ratio (see, e.g., Brennan et al. (2021); Kunisky et al. (2019)):

$$\left\| \mathop{\mathbf{E}}_{v\sim\mathcal{U}(S)} \left[\left(\overline{P}_{A,v}^{\otimes n} \right)^{\leq r,\ell} \right] - 1 \right\|_{D_0^{\otimes n}} ,$$

where in our case $D_0 = \mathcal{N}(0, I_d)$.

To show that the low-degree likelihood ratio is small, we use the result from Brennan et al. (2021) stating that a lower bound for the SQ dimension translates to an upper bound for the low-degree likelihood ratio. Therefore, given that we have already established in previous section that $\mathrm{SD}(\mathcal{B}(\{P_{A,v}\}_{v\in S}, \mathcal{N}(0, I_d)), \gamma, \beta) = \Omega(d^{ck^c/8})$ for $\gamma = 2^{m+1}k^{(c-1)(m+1)}\chi^2(A, \mathcal{N}(0,1))$ and $\beta = \chi^2(A, \mathcal{N}(0,1))$, we obtain the corollary:

Theorem 85 *Let $0 < c < 1$. Consider the hypothesis testing Problem 83 where A matches m moments with $\mathcal{N}(0,1)$. For any $d, k, m \in \mathbb{Z}_+$ such that $k \leq \sqrt{d}$ and $ck^c \geq \Omega(m \log k)$, any $n \leq k^{(1-c)(m+1)}/(2^{m+1}\chi^2(A, \mathcal{N}(0, I_d))$ and any even integer $\ell \leq (ck^c \log d)/(32m \log k)$, we have that*

$$\left\| \mathop{\mathbf{E}}_{u\sim S} \left[(\bar{P}_{A,u}^{\otimes n})^{\leq \infty, \Omega(\ell)} \right] - 1 \right\|^2_{\mathcal{N}(0,I_d)^{\otimes n}} \leq 1 .$$

The interpretation of this result is that unless the number of samples used n is greater than $k^{(1-c)(m+1)}/(2^{m+1}\chi^2(A, \mathcal{N}(0, I_d))$, any polynomial of degree roughly up to $k^c \log d$ fails to be a good test (note that any polynomial of degree ℓ has sample-wise degree at most (ℓ, ℓ)). Using the lower bounds for the SQ dimension, we also obtain lower bounds for the low-degree polynomial tests for problems in Theorems 73 and 75 with qualitatively similar guarantees.

Finally, the connection to the estimation problem is again done via the reduction of Claim 72, which also works in the low-degree model family of algorithms.

Remark 86 (Reduction within low-degree polynomial class) *Let $\mathcal{A}$ be a low-degree polynomial algorithm for Problem 70 with degree ℓ. Then the reduction in Claim 72 gives us an algorithm $\mathcal{A}'$ for Problem 71 which can be implemented as a polynomial test of degree 2ℓ.*

Appendix G. Information-Theoretic Error and Sample Complexity

Theorem 87 (Sample Complexity of Robust Sparse Mean Estimation with Bounded Moments) *Let C be a sufficiently large constant and c a sufficiently small positive constant. There is a (computationally inefficient) algorithm that, given any $\epsilon < c$ and an ϵ-corrupted set of size $n > C(k \log d)/\epsilon^{2-2/t}$ from any distribution with k-sparse mean and t-th moments bounded by M, finds a $\widehat{\mu}$, such that $\|\widehat{\mu} - \mathbf{E}_{X\sim D}[X]\|_2 = O(M^{1/t}\epsilon^{1-1/t})$, with probability at least 0.9.*

Proof Let S be the given data set of cardinality n and $\mu = \mathbf{E}_{X\sim D}[X]$. For a unit vector v, let S_v be the projection of the points along v, that is $S_v = \{v^T x : x \in S\}$. Note that we have assumed that in any k-sparse direction v, the t-th moment is at most M.

Let $\mathcal{C}$ be a $1/2$-net of the unit-norm k-sparse vectors (which we denote by $\mathcal{U}_k$). The cardinality of $\mathcal{C}$ is bounded by $\binom{d}{k}5^k$ since there are at most $\binom{d}{k}$ ways to select the non-zero coordinates and a $(1/2)$-net of $\mathbb{R}^k$ has size at most 5^k.

For $\tau < 1$, let f_τ be the real-valued function on univariate sets that computes the τ-trimmed mean of the given data set as in Lugosi and Mendelson (2021). From that, it is implied that for any unit vector v and $\tau = \Theta(\epsilon + \log(1/\gamma')/n)$ (where the parameters are such that $\tau < 1$), with probability $1 - \gamma'$ we have that

$$|f_\tau(S_v) - \mu^T v| = O(M^{1/t}(\epsilon^{1-1/t} + \sqrt{\log(1/\gamma)/n})) \; .$$

Setting $\gamma' = \gamma/|\mathcal{C}|$ and $n \geq C(k \log d + \log(1/\gamma))/\epsilon^{2-2/t}$ and using a union bound, we have that with probability $1 - \gamma$, for each $v \in \mathcal{C}$, $|f_\tau(S_v) - \mu^T v| \leq \delta$, where $\delta = O(M^{1/t}\epsilon^{1-1/t})$. We denote this event by $\mathcal{E}$. We will assume that $\mathcal{E}$ holds for the remainder of the proof. For each $v \in \mathcal{C}$, define $\widehat{\mu}_v := f_\tau(S_v)$ and let the estimate $\widehat{\mu}'$ to be any point with the property $|v^T\widehat{\mu}' - \widehat{\mu}_v| \leq \delta$ for all $v \in \mathcal{C}$ (such a point always exists, since the true mean satisfies that property on $\mathcal{E}$). For that $\widehat{\mu}'$, we have that

$$|v^T(\widehat{\mu}' - \mu)| \leq |v^T\widehat{\mu}' - \widehat{\mu}_v| + |\widehat{\mu}_v - v^T\mu| \leq 2\delta \; ,$$

for every $v \in \mathcal{C}$. We claim that $|v^T(\widehat{\mu}' - \mu)| \leq 4\delta$ for all $v \in \mathcal{U}_k$. To see this, let $v_0 := \arg\max_{v \in \mathcal{U}_k} |v^T(\widehat{\mu}' - \mu)|$ and $w := \arg\min_{x \in \mathcal{C}} \|v_0 - x\|_2$. We have that

$$|v_0^T(\widehat{\mu}' - \mu)| \leq |w^T(\widehat{\mu}' - \mu)| + |(w - v)^T(\widehat{\mu}' - \mu)| \leq |w^T(\widehat{\mu}' - \mu)| + \frac{1}{2}|v_0^T(\widehat{\mu}' - \mu)| \; .$$

Solving for $|v_0^T(\widehat{\mu}' - \mu)|$ shows the claim. Let the final estimate be $\widehat{\mu} = h_k(\widehat{\mu}')$, where h_k is the operator that truncates a vector to its largest k coordinates. Applying Fact 9, we get that $\|\widehat{\mu} - \mu\|_2 = O(\delta)$. ■

Theorem 88 (Sample Complexity of Robust Sparse Mean Estimation of Gaussian) *Let C be a sufficiently large constant and c a sufficiently small positive constant. There is a (computationally inefficient) algorithm that, given any $\epsilon < c$ and any ϵ-corrupted set of size $n > Ck(\log d)/\epsilon^2$ from $\mathcal{N}(\mu, \Sigma)$ with k-sparse μ, finds a $\widehat{\mu}$, such that $\|\widehat{\mu} - \mu\|_2 = O(\sqrt{\|\Sigma\|_2}\epsilon)$, with probability at least* 0.9.

Proof We will use the same notation as the proof of Theorem 87. For each $v \in \mathcal{C}$, define $\widehat{\mu}_v :=$ Median(S_v). Standard results (see, for example, (Lai et al., 2016, Lemma 3.3)) imply that with probability $1 - \exp(-n\epsilon^2)$, $|v^T\mu - \widehat{\mu}_v| \leq \delta$, where $\delta = O(\epsilon\sqrt{\|\Sigma\|_2})$. Let $\mathcal{E}$ be the event where for each $v \in \mathcal{C}$, $|\widehat{\mu}_v - \mu^T v| \leq \delta$. By a union bound, $\mathcal{E}$ happens with probability at least $1 - \gamma$, if $n \geq C(k \log d + \log(1/\gamma))/\epsilon^2)$ for a large enough constant C. Following the same argument as the proof of Theorem 87, we get the desired result. ■

We now state the following folklore results for the information-theoretic lower bound. Although we present the results for univariate distributions, it is easy to see that the same lower bounds also hold for k-sparse distributions for any $k \geq 1$.

Fact 89 (Information-theoretic Lower Bounds) *The following hold:*

- *There exist univariate distributions D_1, D_2 such $D_2 = (1 - \epsilon)D_1 + \epsilon N$ for some N, the t-th moments of both D_1, D_2 are at most M, and $|\mathbf{E}_{X \sim D_1}[X] - \mathbf{E}_{X \sim D_2}[X]| = \Omega(M^{1/t}\epsilon^{1-1/t})$.*

- *There exist Gaussian distributions* $D_1 = \mathcal{N}(\mu_1, \sigma^2), D_2 = \mathcal{N}(\mu_2, \sigma^2)$ *such that* $|\mu_1 - \mu_2| = \Omega(\epsilon\sigma)$ *and* $(1-\epsilon)D_1 + \epsilon N_1 = (1-\epsilon)D_2 + \epsilon N_2$.

Proof By scaling, we focus on the $M = 1$ case. Let D_1 be the Dirac delta at zero and let $D_2 = (1-\epsilon)D_1 + \epsilon N$, where N has all its mass at $\epsilon^{-1/t}$. Then, the means are indeed separated by $\epsilon^{1-1/t}$. For the t-th moment of D_2 we have that $\mathbf{E}_{X\sim D_2}[(X-\mu_2)^t] = (1-\epsilon)\epsilon^{t-1} + \epsilon(\epsilon^{-1/t} - \epsilon^{1-1/t})^t \leq \epsilon + \epsilon(\epsilon^{-1/t}(1-\epsilon))^t \leq \epsilon + (1-\epsilon) \leq 1$.

For the Gaussian distributions, the claim is based on the fact that $\mathrm{dtv}(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) = \Theta(\epsilon)$ whenever $|\mu_1 - \mu_2| = \sigma\epsilon$, thus an additive adversary can make the two distributions look the same. A version of the lower bound can also be found in (Lai et al., 2016, Observation 1.4) and (Chen et al., 2018, Theorem 2.2).

■