Polynomial-Time Sum-of-Squares Can Robustly Estimate Mean and Covariance of Gaussians Optimally

Pravesh K. Kothari PRAVESHK@CS.CMU.EDU
Peter Manohar PMANOHAR@CS.CMU.EDU
Brian Hu Zhang BHZHANG@CS.CMU.EDU
Carnegie Mellon University

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

In this work, we revisit the problem of estimating the mean and covariance of an unknown d-dimensional Gaussian distribution in the presence of an ε-fraction of adversarial outliers. The work of Diakonikolas et al. (2016) gave a polynomial time algorithm for this task with optimal $\tilde{O}(\varepsilon)$ error using $n = \mathrm{poly}(d, 1/\varepsilon)$ samples.

On the other hand, Kothari and Steurer (2017) introduced a general framework for robust moment estimation via a canonical sum-of-squares relaxation that succeeds for the more general class of *certifiably subgaussian* and *certifiably hypercontractive* (Bakshi and Kothari, 2020) distributions. When specialized to Gaussians, this algorithm obtains the same $\tilde{O}(\varepsilon)$ error guarantee as Diakonikolas et al. (2016) but incurs a super-polynomial sample complexity ($n = d^{O(\log 1/\varepsilon)}$) and running time ($n^{O(\log(1/\varepsilon))}$). This cost appears inherent to their analysis as it relies only on sum-of-squares certificates of upper bounds on directional moments while the analysis in Diakonikolas et al. (2016) relies on *lower bounds* on directional moments inferred from algebraic relationships between moments of Gaussian distributions.

We give a new, simple analysis of the *same* canonical sum-of-squares relaxation used in Kothari and Steurer (2017) and Bakshi and Kothari (2020) and show that for Gaussian distributions, their algorithm achieves the same error, sample complexity and running time guarantees as of the specialized algorithm in Diakonikolas et al. (2016). Our key innovation is a new argument that allows using moment lower bounds without having sum-of-squares certificates for them. We believe that our proof technique will likely be useful in designing new robust estimation algorithms.

Keywords: Robust estimation, sum-of-squares, mean estimation

Contents

1. Introduction

Designing estimation algorithms for estimating basic parameters of probability distributions from samples is a foundational computational problem in machine learning. However, natural estimation algorithms, such as taking the sample mean for population mean, can be brittle – even a single outlier in the data can lead to an arbitrarily large estimation error. In the 1960s, Tukey and Huber began systematic efforts to build *robust estimators* that can tolerate minor deviations of the input from the chosen model, such as the injection of a small constant fraction of adversarially chosen outliers into the sample. While this effort has led to a burgeoning body of work called *robust statistics*, the algorithms from this line of work typically require exponential time in the underlying dimension to succeed and are thus inefficient in high-dimensional settings.

In 2016, two papers (Diakonikolas et al., 2016; Lai et al., 2016) pioneered a systematic effort to build *computationally efficient* robust estimators. Since their work, the study of *algorithmic robust statistics* has evolved into an active area, that, in addition to yielding concrete solutions to basic robust estimation problems, has led to the synthesis of truly new algorithmic ideas (often improving even the classical, non-robust algorithms) that identify and clarify general principles for efficient robust estimation.

A key insight from this line of work has been a general blueprint for robust estimation using the *sum-of-squares* (SoS) method. A sequence of works in 2018 gave a canonical sum-of-squares relaxation and a rounding algorithm that gives the nearly statistically optimal outlier-robust estimation of *moments* (Kothari and Steurer, 2017) and robust clustering (Kothari and Steinhardt, 2017; Hopkins and Li, 2018) of spherical mixtures of a broad class of probability distributions. Since then, this framework has been refined and expanded to obtain state-of-the-art robust estimation algorithms for problems such as outlier-robust regression (Klivans et al., 2018; Bakshi and Prasad, 2021), clustering of non-spherical mixtures (Bakshi and Kothari, 2020; Bakshi et al., 2020a), heavy tailed estimation (Hopkins, 2018; Cherapanamjeri et al., 2020), list-decodable regression and subspace recovery (Karmalkar et al., 2019; Bakshi and Kothari, 2020; Raghavendra and Yau, 2020a,b) and robust learning of a mixture of arbitrary Gaussians (Liu and Moitra, 2021; Bakshi et al., 2020b).

In addition, algorithms from the SoS-based robust estimation framework have the advantage of abstracting out natural analytic properties of the statistical model in question and yielding robust estimators for all distributions that satisfy such properties in a blackbox way. For example, the algorithms for robust estimation of moments (Kothari and Steurer, 2017) and clustering (Hopkins and Li, 2018; Kothari and Steinhardt, 2017) apply to all *certifiably subgaussian* distributions, that, informally speaking, are distributions that admit "sum-of-squares certificates" of the property of having subgaussian low order moments. As another example, the covariance estimation algorithm of Bakshi and Kothari (2020) applies to all distributions that admit sum-of-squares certificates of bounds on moments of degree-2 polynomials (*certifiable hypercontractivity*). Such properties are already known to hold for a broad class of distributions and verifying them for a new family immediately generalizes such results. In fact, one can interpret the analysis in the sum-of-squares framework as identifying structural properties (certificates of appropriate analytic properties) of the distribution families that can be exploited for the design of efficient robust estimation algorithms.[1]

Robust Mean Estimation for Gaussians. While the SoS-based framework above typically achieves the best known recovery guarantees (among polynomial time algorithms), a striking excep-

1. See this recent talk for this perspective and its applications to weakening distributional assumptions in robust estimation.

tion so far has been the task of robustly estimating the mean and covariance of an unknown Gaussian distribution. In this problem, the algorithm is given input data $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{R}^d$ that is obtained by *arbitrarily and adversarially* corrupting εn points in an i.i.d. sample $X = \{x_1, x_2, \ldots, x_n\}$ from an unknown Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. The algorithm of Diakonikolas et al. (2016) obtains estimates $\hat{\mu}, \hat{\Sigma}$ so that the total variation distance $d_{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) \leqslant \tilde{O}(\varepsilon)$. This is optimal up to logarithmic factors in ε in the bound (and there is evidence (Diakonikolas et al., 2017) that such a loss might be necessary for polynomial time algorithms). Their algorithm requires $n = \mathrm{poly}(d, 1/\varepsilon)$ samples and $\mathrm{poly}(n)$ running time. On the other hand, the best known SoS-based algorithm for the problem is obtained by specializing the analysis in Bakshi and Kothari (2020) for mean and covariance estimation of *certifiably hypercontractive* distributions to the case of Gaussians. This analysis yields the same error bound of $\tilde{O}(\varepsilon)$ on the total variation error but requires super-polynomially many $d^{O(\log 1/\varepsilon)}$ samples and $n^{O(\log 1/\varepsilon)}$ running time.

There is an important technical bottleneck in the analysis of the canonical SoS algorithm for obtaining the stronger guarantees in Diakonikolas et al. (2016). The analysis in Kothari and Steurer (2017) (and extensions in Bakshi and Kothari (2020)) only uses upper bounds on the higher moments of distributions. On the other hand, the stronger analysis in Diakonikolas et al. (2016) implicitly relies on a non-trivial *lower bound* on moments of arbitrary subsets of the original sample of size $(1 - \varepsilon)n$. The best known sum-of-squares certificates for such a lower bound property appear to require an exponential cost in $O(\log 1/\varepsilon)$ in both sample complexity and running time. And, it is plausible (though, still unproven) that such a cost is necessary! This state of affairs leads us to the main motivating question of this work:

Can the canonical SoS based algorithm give a robust estimate with $\tilde{O}(\varepsilon)$ total variation error for the mean and covariance of Gaussian distributions in polynomial time and samples? Or is the SoS framework for moment estimation weaker, when specialized to Gaussian distributions?

In this work, we give a new analysis of the canonical sum-of-squares-based algorithm for robust mean estimation for Gaussians (that only has subgaussian upper bounds on 4th moment as constraints) that recovers the polynomial running time and sample complexity guarantees of Diakonikolas et al. (2016) and same error up to $\mathrm{poly} \log 1/\varepsilon$ factors. Our key innovation (that we explain later in this section) is a new argument that works around the issue of finding efficient sum-of-squares certificates for moment lower bounds and yet manages to prove the stronger guarantee. We believe that this new technique will likely find further applications in efficient robust estimation.

1.1. Our results

Formally, our algorithms work in the following strong contamination model for corrupted samples used in several prior works on robust estimation, beginning with Diakonikolas et al. (2016) and Lai et al. (2016).

Definition 1 (Strong contamination model) *Let D be a distribution on $\mathbb{R}^d$ and let $X = \{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample from D. In the strong contamination model, an ε-corrupted sample is obtained by replacing any adversarially chosen εn points from X with arbitrary outliers to obtain $Y = \{y_1, y_2, \ldots, y_n\}$.*

Our main result is an analysis of the following canonical SoS relaxation for mean and covariance estimation along with a simple rounding (used in Kothari and Steurer (2017) and Bakshi and Kothari (2020)) for estimating the mean and covariance of an unknown distribution.

Algorithm 2 (Mean and spectral norm covariance estimation)

Input: Parameter $\varepsilon \in (0,1)$, and corrupted samples $y_1, \dots, y_n \in \mathbb{R}^d$.

Operation: Find a degree-12 pseudo-expectation $\tilde{\mathbb{E}}$ (solution to the SoS semidefinite programming relaxation) in variables $x'_1, \dots, x'_n \in \mathbb{R}^d$, $w_1, \dots, w_n \in \mathbb{R}$, $\mu' := \mathbb{E}_i x'_i$, $\Sigma' := \mathbb{E}_i (x'_i - \mu')(x'_i - \mu')^\top$ satisfying the following set of constraints:

(1) **Booleanity of intersection Variables:** $w_i^2 = w_i$ for every $i \in [n]$,

(2) **Size of intersection:** $\sum_{i=1}^n w_i = (1-\varepsilon)n$,

(3) **Intersection constraints:** $w_i x'_i = w_i y_i$ for every $i \in [n]$,

(4) **Certifiable Subgaussianity of 4th moments:** $\frac{1}{n}\sum_{i=1}^n \left(\langle x'_i - \mu', v\rangle^2 - v^\top \Sigma' v\right)^2 \leqslant (2 + \tilde{O}(\varepsilon))(v^\top \Sigma' v)^2$ for every $v \in \mathbb{R}^d$.

Output: $\hat{\mu} = \tilde{\mathbb{E}}[\mu']$, $\hat{\Sigma} = \tilde{\mathbb{E}}[\Sigma']$.

The constraints of the program encode the task of finding a set of points X' that intersects the input sample Y in $(1-\varepsilon)n$ points (encoded by the first 3 sets of constraints) such that the empirical 4-th moments of X' are bounded above by at most ~ 2 times the square of the 2nd moments (the last set of constraints). The last set of constraints, though apparently infinitely many (one for every $v \in \mathbb{R}^d$) admit a succinct representation via techniques of Kothari and Steurer (2017) and Hopkins and Li (2018) (see Appendix B, or Fleming et al. (2019), Chapter 4 for an exposition). The intended solution for this polynomial system is $X' = X$ – the unknown, true i.i.d. sample. (And then setting $w_i = \mathbf{1}(x_i = y_i)$, and μ', Σ' to be the empirical mean/covariance of X.) It is easy to check that X satisfies the last set of constraints – the only property of i.i.d. Gaussian samples that we enforce.

We prove the following formal guarantees on Algorithm 2.

Theorem 3 (Mean and spectral norm covariance estimation) *Algorithm 2 takes as input an ε-corrupted sample of size n from a Gaussian distribution with mean μ and covariance Σ and in $\mathrm{poly}(n)$-time, outputs estimates $\hat{\mu} \in \mathbb{R}^d, \hat{\Sigma} \in \mathbb{R}^{d\times d}$ with the following guarantee. If $\Sigma \succeq 2^{-\mathrm{poly}(d)} I$, and $n \geqslant \tilde{O}(d^2 \log^5(1/\delta)/\varepsilon^2)$, with probability at least $1-\delta$ over the draw of the original uncorrupted sample X, the estimates $\hat{\mu}, \hat{\Sigma}$ satisfy:*

(1) (Mean estimation) $\left\|\Sigma^{-1/2}(\hat{\mu} - \mu)\right\|_2 \leqslant \tilde{O}(\varepsilon)$, *and*

(2) (Covariance estimation in spectral norm) $(1 - \tilde{O}(\varepsilon))\Sigma \preceq \hat{\Sigma} \preceq (1 + \tilde{O}(\varepsilon))\Sigma$.

Remark 4 (Computational Model and Numerical Issues) *Our algorithm succeeds in the standard word RAM model of computation. In this model, the input sample Y is given to the algorithm after "truncating" the real numbers to rational numbers with $\mathrm{poly}(d)$ bits of precision. The running time of our algorithm is polynomial in the size of the bit representation of the input. The assumption on the smallest eigenvalue of Σ in the statement above is entirely an artefact of the numerical issues as the truncation of Y to rational numbers, in general, does not allow recovering eigenvalues of Σ that are not representable in polynomially many bits of precision. Such an assumption is required (but sometimes not stated explicitly) by all prior works on robust estimation when implemented in the standard word RAM model of computation.*

We note that it is possible (though, requires additional steps in the algorithm) to remove the assumption on the smallest eigenvalue of Σ if we instead assume that the unknown Σ has rational entries. Such an assumption is clearly necessary as algorithms in the word RAM model can only output $\hat{\Sigma}$ with rational entries. We omit the description of such a method and instead choose to make an assumption that the smallest eigenvalue of Σ can be written down in $\mathrm{poly}(d)$ bits.

Theorem 3 shows that the algorithm of Kothari and Steurer (2017), when analyzed for Gaussian distributions, achieves the information-theoretically optimal $\tilde{O}(\varepsilon)$ error guarantee using $n = \mathrm{poly}(d, 1/\varepsilon)$ samples and $\mathrm{poly}(n)$ running time. This shows that the analysis of Kothari and Steurer (2017), which is tight for the more general class of certifiable subgaussian and certifiable hypercontractive distributions, can be improved in the specific case of Gaussians.

The guarantees achieved by Theorem 3 are weaker than the guarantees of the algorithm in Diakonikolas et al. (2016), whose estimate $\hat{\Sigma}$ is additionally close to Σ in relative Frobenius error. We show that by analyzing the degree-12 SoS relaxation of the following program (that replaces the certifiable subgaussianity constraints by certifiable hypercontractivity constraints on degree-2 polynomials), we can upgrade the guarantees of Theorem 3 to achieve the stronger Frobenius norm guarantee of Diakonikolas et al. (2016). We note that this program (with additional higher-degree certifiable hypercontractivity constraints) was analyzed in Bakshi and Kothari (2020) to obtain similar guarantees on the mean and covariance estimation of the more general class of all certifiably hypercontractive distributions, but needed $n = d^{O(\log 1/\varepsilon)}$ samples and $n^{O(\log 1/\varepsilon)}$ running time. Our contribution is obtaining a sharper analysis of the same program for the case of Gaussian distributions.

Algorithm 5 (Frobenius norm covariance estimation)

Input: Parameter $\varepsilon \in (0, 1)$, and corrupted samples $y_1, \ldots, y_n \in \mathbb{R}^d$.

Operation: Find a degree-12 pseudo-expectation $\tilde{\mathbb{E}}$ in the variables $x'_1, \ldots, x'_n \in \mathbb{R}^d$, $w_1, \ldots, w_n \in \mathbb{R}$, $\mu' := \mathbb{E}_i x'_i$, $\Sigma' := \mathbb{E}_i (x'_i - \mu')(x'_i - \mu')^\top$ satisfying the following set of constraints:

(1) $w_i^2 = w_i$ for every $i \in [n]$,
(2) $\sum_{i=1}^n w_i = (1 - \varepsilon)n$,
(3) $w_i x'_i = w_i y_i$ for every $i \in [n]$,
(4) $\mathbb{E}_i \langle (x'_i - \mu')(x'_i - \mu')^\top - \Sigma', P \rangle^2 \leqslant (2 + \tilde{O}(\varepsilon)) \|P\|_F^2$ for every symmetric $P \in \mathbb{R}^{d \times d}$.

Output: $\hat{\Sigma} := \tilde{\mathbb{E}}[\Sigma']$.

Theorem 6 (Frobenius norm covariance estimation with $\Sigma \approx \mathbb{I}$) *Algorithm 5 takes as input an ε-corrupted sample of size n from a Gaussian distribution with mean μ and covariance Σ with $(1 - \tilde{O}(\varepsilon))\mathbb{I} \preceq \Sigma \preceq (1 + \tilde{O}(\varepsilon))\mathbb{I}$, and in $\mathrm{poly}(n)$-time, outputs an estimate $\hat{\Sigma} \in \mathbb{R}^{d \times d}$ with the following guarantee. If $n \geqslant \tilde{O}(d^2 \log^5(1/\delta)/\varepsilon^2)$, then with probability at least $1 - \delta$ over the draw of the original uncorrupted sample X, the estimate $\hat{\Sigma}$ satisfies $\|\Sigma^{-1/2} \hat{\Sigma} \Sigma^{-1/2} - \mathbb{I}\|_F \leqslant \tilde{O}(\varepsilon)$.*

We note that Theorem 6 requires that the input distribution has covariance that is close to $\mathbb{I}$ in spectral norm. This is easily achieved by first running Algorithm 2 to derive an estimate $\hat{\Sigma}_0$ that is

close to Σ in spectral norm, and then running Algorithm 5 on inputs that are linearly transformed as $y \mapsto \hat{\Sigma}_0^{-1/2} y$. After this linear transformation, the new, "true" covariance $\Sigma_0^{-1/2}\Sigma\hat{\Sigma}_0^{-1/2}$ satisfies $(1-\tilde{O}(\varepsilon))\mathbb{I} \preceq \hat{\Sigma}_0^{-1/2}\Sigma\hat{\Sigma}_0^{-1/2} \preceq (1+\tilde{O}(\varepsilon))\mathbb{I}$.

We thus obtain the final corollary:

Corollary 7 (Mean and Frobenius norm covariance estimation) *There is an SoS-based algorithm that takes as input an ε-corrupted sample of size n from a Gaussian distribution with mean μ and covariance Σ and in $\mathrm{poly}(n)$-time, outputs estimates $\hat{\mu} \in \mathbb{R}^d, \hat{\Sigma} \in \mathbb{R}^{d\times d}$ with the following guarantee. If $\Sigma \succeq 2^{-\mathrm{poly}(d)} I$, and $n \geqslant \tilde{O}(d^2 \log^5(1/\delta)/\varepsilon^2)$, with probability at least $1-\delta$ over the draw of the original uncorrupted sample X, the estimates $\hat{\mu}, \hat{\Sigma}$ satisfy:*

(1) (Mean estimation) $\left\|\Sigma^{-1/2}(\hat{\mu}-\mu)\right\|_2 \leqslant \tilde{O}(\varepsilon)$, *and*

(2) (Covariance estimation in Frobenius norm) $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbb{I}\|_F \leqslant \tilde{O}(\varepsilon)$.

In particular, $\Delta_{TV}(N(\mu,\Sigma), N(\hat{\mu},\hat{\Sigma})) \leqslant \tilde{O}(\varepsilon)$.

Thus, we obtain the same guarantee[2] on the total variation distance as in Diakonikolas et al. (2016).

Lastly, we note that if we are also given (an estimate of) the covariance as input, then we can estimate the mean using fewer samples.

Corollary 8 *Let $\hat{\Sigma} \in \mathbb{R}^{d\times d}$. Consider the following modification of Algorithm 2, where we replace the SoS variable Σ' with the additional input $\hat{\Sigma}$. This algorithm takes as input an ε-corrupted sample of size n from a Gaussian distribution with mean μ and covariance Σ and in $\mathrm{poly}(n)$-time, outputs an estimate $\hat{\mu} \in \mathbb{R}^d$ with the following guarantee. If $(1-\tilde{O}(\varepsilon))\Sigma \preceq \hat{\Sigma} \preceq (1+\tilde{O}(\varepsilon))\Sigma$, and $n \geqslant O((d+\log(1/\delta))/\varepsilon^2)$, with probability at least $1-\delta$ over the draw of the original uncorrupted sample X, the estimate $\hat{\mu}$ satisfies $\left\|\Sigma^{-1/2}(\hat{\mu}-\mu)\right\|_2 \leqslant \tilde{O}(\varepsilon)$.*

1.2. A brief overview of our key idea

We now give a high level sketch of the key idea used in our proof. First, let's briefly recap the style of analysis in Kothari and Steurer (2017) and Bakshi and Kothari (2020) by focusing on the guarantee for mean estimation. The analysis in these works utilizes the "proofs to algorithms" framework of algorithm design via the sum-of-squares method. The polynomial constraints in the our program (Algorithm 2) encodes finding a set X' of samples that intersects the input corrupted sample Y in $(1-\varepsilon)n$ points and has 4th moments upper bounded in terms of the squared 2nd moments in every direction. The analysis proceeds by using the constraints to derive, via a $O(1)$-degree SoS proof that the error in the so called *Mahalanobis* norm, $\left\|\Sigma^{-1/2}(\mu'-\mu)\right\|_2 \leqslant O(\varepsilon^{3/4})$. Such an inequality implies that any degree $O(1)$-pseudo-expectation $\tilde{\mathbb{E}}$ that satisfies the constraints of our program must also satisfy all consequences obtainable via $O(1)$-degree SoS proofs. As a result, the rounded estimate $\hat{\mu} = \tilde{\mathbb{E}}[\mu']$ also satisfies the inequality above giving us the required guarantee.

The Mahalanobis error bound of $\varepsilon^{3/4}$ here comes from the upper bound on the 4th moments (in general, we can obtain $\sim \sqrt{t}\varepsilon^{1-1/t}$ error bounds by working with upper bounds on t-th moments)

2. Our formal guarantees are not explicit about $\mathrm{polylog}(1/\varepsilon)$ factors in the error, and as a result, formally speaking our error bounds only match that of Diakonikolas et al. (2016) up to $\mathrm{polylog}(1/\varepsilon)$ factors. We believe that our argument in fact gets the same $\mathrm{polylog}(1/\varepsilon)$ dependence, as we rely on the same concentration bounds as in Diakonikolas et al. (2016), which is where the $\mathrm{polylog}(1/\varepsilon)$ factors arise. But, our proofs currently do not explicitly show this.

encoded in our constraints and is polynomially off from the optimal $\tilde{O}(\varepsilon)$ bound we intend to achieve when the unknown distribution is Gaussian. In fact, the analysis and the bounds obtained by Kothari and Steurer (2017) and Bakshi and Kothari (2020) are *information-theoretically optimal*: for any t, there are two distributions with means that are $\sqrt{t}\varepsilon^{1-1/t}$ far, satisfy the $2t$-th moment upper bound condition, and are $\sim \varepsilon$-different in total variation distance. In particular, one can take such a pair of distributions and produce corrupted samples that are statistically indistinguishable!

At this point, one might wonder – how can one hope the *same set of constraints* to yield a tighter guarantee for Gaussians? Indeed, our analysis follows a substantially different path to use the Gaussianity of the underlying input distribution.

Enter Resilience. At a high-level, in retrospect, the key property of Gaussians that Diakonikolas et al. (2016) exploit (which is not satisfied by distributions that constitute the "hard examples" above) is a certain mild anti-concentration property (that we call resilience, following Steinhardt et al. (2017)) inferred from *lower bounds* on moments of subsamples. Specifically, if X is a typical i.i.d. sample from a d-dimensional, 0-mean Gaussian distribution of size $n \gg d/\varepsilon$, and $S \subseteq X$ is *any* subset of size $(1-\varepsilon)n$, then the empirical covariance Σ_S of points in S satisfies $\Sigma_S \geqslant (1-\tilde{O}(\varepsilon))\Sigma_X$. Or, equivalently, that for $\bar{S} = X \setminus S$, it holds that $\Sigma_{\bar{S}} \preceq \mathrm{polylog}(1/\varepsilon)\Sigma_X$. The analysis of Kothari and Steurer (2017) can only infer (via Cauchy-Schwarz inequality) the exponentially worse bound of $\Sigma_{\bar{S}} \leqslant O(1/\sqrt{\varepsilon})\Sigma_X$.

Crucially, resilience of the covariance *cannot* be inferred from 4th moment upper bounds, such as those encoded by our constraints. It can indeed be inferred from an argument that relies on boundedness of $O(\log 1/\varepsilon)$ moments, but if we wanted our sample X to have all of its $\leqslant O(\log 1/\varepsilon)$ moments close to that of the true distribution, we would need $d^{O(\log 1/\varepsilon)}$ (in particular, superpolynomially many) samples.

A key insight of Diakonikolas et al. (2016) is the observation that one can prove the resilience of covariance by a simple Hoeffding + union bound for a sample of size $n \sim d/\varepsilon$. Notice that Hoeffding's inequality itself relies on subgaussianity of all moments of the distribution but the relevant consequence of it – namely resilience – can be "seen" in typical samples of size $\sim d/\varepsilon$.

Resilience is likely not efficiently certifiable. While this is encouraging, using this property within the sum-of-squares framework poses a major issue. Notice that, a priori, verifying that a sample X satisfies resilience requires an exponential search since we need to verify some property for every subset S of size $(1-\varepsilon)n$. Indeed, given a sample of size n – as far as we know, there is no known polynomial time algorithm to output a *certificate* (whether via sum-of-squares or otherwise) of such a property. On the other hand, since the analysis style in Kothari and Steurer (2017) involves deriving a bound on the Mahalanobis distance between μ' and the true mean μ, we would need a low-degree sum-of-squares certificate of resilience in order to plug it into the SoS framework. This is the key technical issue that prevented prior attempts to "SoSize" the argument of Diakonikolas et al. (2016) for mean (and more generally, covariance estimation) for Gaussians.

Circumventing certificates by proving "only in pseudo-expectation". Our main contribution is an argument that allows us to use resilience without requiring an SoS certificate. Notice that, though powerful and elegant, obtaining a low-degree sum-of-squares proof of a bound on the Mahalanobis distance $\left\|\Sigma^{-1/2}(\mu'-\mu)\right\|_2$ is overkill for our purpose! We only need the inequality *after taking pseudo-expectations*. Our key idea, thus, is to directly prove a bound on $\left\|\Sigma^{-1/2}(\tilde{\mathbb{E}}[\mu']-\mu)\right\|_2$ without going through low-degree sum-of-squares proofs.

If, for a second, we pretend that pseudo-expectations are in fact actual probability distributions over solutions X', then this is akin to proving an inequality on the expectation of the solution X' without establishing (the considerably stronger claim) that it holds "pointwise" in the support of the distribution. Thus, our idea above can be summarized as attempting to prove a fact "in pseudo-expectation" without establishing it pointwise in the support of the "pseudo-distribution".

We show that for the purpose of arguing "after taking pseudo-expectations", we can in fact leverage the resilience bound discussed above. Our final argument thus derives some facts "within low-degree sum-of-squares proof system" – with some technical choices that make the composition with facts "after taking expectations" possible. Making this work and extending to covariance estimation in spectral and then Frobenius norms requires some more technical work which, for the purpose of this overview, we omit.

To the best of our knowledge, this is the first example in the SoS proofs to algorithms framework for robust statistics where the difference between facts "derived via low-degree SoS proofs" vs "proved only in pseudo-expectations" appears to make a significant material difference to the results so obtained. We believe that this style of analysis might come in handy in future applications of the SoS method to robust statistics and more generally, problems in statistical estimation.

2. Preliminaries

In this section, we give an overview of the sum-of-squares algorithm and state the concentration properties of Gaussians that we need for our results.

2.1. A crash course in sum-of-squares

We give a brief overview of the sum-of-squares (SoS) algorithm. For a more in-depth survey, see Fleming et al. (2019).

The sum-of-squares algorithm works in the standard word RAM model of computation. We assume that all numerical inputs are rational numbers represented as a pair of integers describing the numerator and the denominator. In order to measure the running time of algorithms, we will need to account for the length of the numbers that arise during the run of the algorithm. The following definition captures the size of the representations of the rational numbers:

Definition 9 (Bit complexity) *The bit complexity of an integer $p \in \mathbb{Z}$ is $1 + \lceil \log_2 p \rceil$. The bit complexity of a rational number p/q where $p, q \in \mathbb{Z}$ is the sum of the bit complexities of p and q.*

We now move to discussing the SoS algorithm. Consider a generic polynomial feasibility problem of the form

$$\text{find} \quad x \in \mathbb{R}^m \quad \text{s.t.} \quad f_i(x) \geqslant 0 \ \ \forall i, \quad g_j(x) = 0 \ \ \forall j \tag{2.1}$$

where f_i and g_j are arbitrary polynomial functions of x with rational coefficients of bit complexity B, and the total number of constraints is $\mathrm{poly}(m)$. Let $\mathcal{P}_{m,k}$ denote the set of polynomials p in m variables with degree at most k. A degree-k pseudo-expectation is an object that mimics a real expectation $\mathbb{E}$ for low-degree polynomials, and is defined as follows.

Definition 10 (Degree-k pseudo-expectation) *A degree-k pseudo-expectation (k even) over m variables is a linear operator $\tilde{\mathbb{E}} \colon \mathcal{P}_{m,k} \to \mathbb{R}$ satisfying:*

(1) (Normalization) $\tilde{\mathbb{E}}[1] = 1$, *and*

(2) (PSDness) $\tilde{\mathbb{E}}[p^2] \geqslant 0$ *for all* $p \in \mathcal{P}_{m,k/2}$.

We say that the PSDness condition is satisfied with error τ if $\tilde{\mathbb{E}}[p^2] \geqslant -\tau\|p\|_2^2$ for each $p \in \mathcal{P}_{m,k/2}$, where $\|p\|_2$ is the ℓ_2-norm of the vector of coefficients of p.

We now define what it means for $\tilde{\mathbb{E}}$ to (approximately) satisfy constraints.

Definition 11 (Satisfying constraints) *For a polynomial g, we say that a degree-k $\tilde{\mathbb{E}}$ satisfies the constraint $\{g = 0\}$ exactly if for every polynomial p of degree $\leqslant k - \deg(g)$, $\tilde{\mathbb{E}}[pg_j] = 0$ and τ-approximately if $|\tilde{\mathbb{E}}[pg_j]| \leqslant \tau\|p\|_2$. We say that $\tilde{\mathbb{E}}$ satisfies the constraint $\{g \geqslant 0\}$ exactly if for every polynomial p of degree $\leqslant k/2 - \deg(g)/2$, it holds that $\tilde{\mathbb{E}}[p^2 g] \geqslant 0$ and τ-approximately if $\tilde{\mathbb{E}}[p^2 g] \geqslant -\tau\|p\|_2^2$.*

We note that in the above two definitions, the requirements on the degree of the polynomial is such that $\tilde{\mathbb{E}}$ is well-defined, e.g., $\tilde{\mathbb{E}}[pg_j]$ is only well-defined when $\deg(pg_j) \leqslant k$.

For intuition, it is helpful to observe that a pseudo-expectation is a relaxation of the familiar notion of expectations: it may be useful to think of pseudo-expectation as satisfying $\tilde{\mathbb{E}}[p] = \mathbb{E}_{z\sim D}[p(z)]$ for some distribution D over $\mathbb{R}^m$. Clearly, if D is a distribution over *feasible* solutions of the constraints in (2.1), then $\tilde{\mathbb{E}}$ satisfies all constraints.

We are now ready to define the *sum-of-squares algorithm*.

Fact 12 (Sum-of-Squares algorithm, Shor (1987), Parrilo (2000), Nesterov (2000), and Lasserre (2001)) *There is an algorithm, the* degree-k sum-of-squares algorithm*, with the following properties: The algorithm takes as input $B \in \mathbb{N}$, $\tau > 0$, $k \in \mathbb{N}$, and a problem of the form (2.1) with $\mathrm{poly}(m)$ constraints, each with rational coefficients of bit complexity B. If there is a degree-k pseudo-distribution satisfying (2.1), then the algorithm outputs in $\mathrm{poly}(B, \log\frac{1}{\tau}) \cdot m^{O(k)}$ a degree-k pseudo-expectation $\tilde{\mathbb{E}}$ that τ-approximately satisfies all the constraints in (2.1), and otherwise outputs "infeasible".*

For the purposes of this paper, we can set $\tau = 2^{-m}$ and $B = \mathrm{poly}(m)$. The "total error" that we will incur will be $O(\mathrm{poly}(m, B)2^{-m}) = O(\mathrm{poly}(m)2^{-m})$ which is negligible.

We state some basic known facts about the pseudo-expectations that we use below.

Proposition 13 (see for e.g., Barak and Steurer (2016) and Fleming et al. (2019)) *For any degree-k $\tilde{\mathbb{E}}$, the following Cauchy-Schwarz inequalities hold:*

(1) For any $p, q \in \mathcal{P}_{m,k/2}$, we have $\tilde{\mathbb{E}}[pq]^2 \leqslant \tilde{\mathbb{E}}[p^2]\,\tilde{\mathbb{E}}[q^2]$.

(2) For any $p_1, \ldots, p_n, q_1, \ldots, q_n \in \mathcal{P}_{m,k/2}$ and distribution $\mathcal{D}$ over $[n]$, $\tilde{\mathbb{E}}$ satisfies the polynomial inequality $\mathbb{E}_{i\sim\mathcal{D}}[p_i q_i]^2 \leqslant \mathbb{E}_{i\sim\mathcal{D}}[p_i^2]\,\mathbb{E}_{i\sim\mathcal{D}}[q_i^2]$. In particular, $(p_1+p_2+p_3)^2 \leqslant 3(p_1^2+p_2^2+p_3^2)$.

Definition 14 (SoS proofs of non-negativity) *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a polynomial. We say that h* has a degree-ℓ SoS proof of nonnegativity *if $h = \sum_{i=1}^r h_i^2$ for some polynomials $h_i \in \mathcal{P}_{d,\ell/2}$.*

2.2. Resilience and certifiable subgaussianity of Gaussian moments

We now give a brief overview of the key properties of Gaussian moments that we use.

Our algorithm relies on concentration bounds of Gaussians from Diakonikolas et al. (2016), which prove resilience of the first 4 moments of the Gaussian distribution. We state the bounds for the first two moments below.

Lemma 15 (Resilience of first and second moments; Lemmas 4.4, 4.3 in Diakonikolas et al. (2016)) *Let $x_1, \dots, x_n \sim \mathcal{N}(0, \mathbb{I}_{d\times d})$, and $n \geqslant O((d + \log(1/\delta))/\varepsilon^2)$. Then with probability $1-\delta$, for all $v \in \mathbb{R}^d$ and vectors $a \in [0,1]^n$ such that $\mathbb{E}_i a_i \geqslant 1-\varepsilon$, we have*

$$\big|\mathop{\mathbb{E}}_i a_i \langle x_i, v\rangle\big| \leqslant \tilde{O}(\varepsilon)\|v\|_2 \ , \quad \text{and}$$
$$\big|\mathop{\mathbb{E}}_i a_i [\langle x_i, v\rangle^2 - \|v\|_2^2]\big| \leqslant \tilde{O}(\varepsilon)\|v\|_2^2 \ .$$

To see the importance of Lemma 15, we note that, when combined with Proposition 2 in Steinhardt et al. (2018), Lemma 15 immediately yields an exponential time algorithm to robustly estimate the mean μ of a Gaussian $\mathcal{N}(\mu, \Sigma)$ with *known* covariance Σ, i.e., output $\hat{\mu}$ satisfying (1) in Theorem 3.

The second resilience property we need is an upgrade of the resilience property of the second moment in Lemma 15, as well as the resilience of the fourth moment.

Lemma 16 (Stronger resilience of second and resilience of fourth moments) *Let $x_1, \dots, x_n \sim \mathcal{N}(0, \mathbb{I}_{d\times d})$, and $n \geqslant \tilde{O}(d^2 \log^5(1/\delta)/\varepsilon^2)$. Then with probability $1-\delta$, for all symmetric $P \in \mathbb{R}^{d\times d}$ and vectors $a \in [0,1]^n$ such that $\mathbb{E}_i a_i \geqslant 1-\varepsilon$, we have*

$$\big|\mathop{\mathbb{E}}_i a_i \langle x_i x_i^\top - \mathbb{I}, P\rangle\big| \leqslant \tilde{O}(\varepsilon) \cdot \|P\|_F \ , \quad \text{and}$$
$$\big|\mathop{\mathbb{E}}_i a_i \big[\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 - 2\|P\|_F^2\big]\big| \leqslant \tilde{O}(\varepsilon) \cdot \|P\|_F^2 \ .$$

Lemma 16 follows from Corollary 4.8, Lemma 5.17 and Lemma 5.21 of Diakonikolas et al. (2016); we include a short proof for completeness in Appendix C.

We will also need slightly different forms of the above resilience results. We now state the results in the form that we need, and postpone the proof to Appendix C. The proofs are a straightforward (but somewhat tedious) consequence of the above results from Diakonikolas et al. (2016).

Lemma 17 *Let $x_1, \dots, x_n \sim \mathcal{N}(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ positive definite. Let n be as in Lemma 16, and $\mu_0 = \mathbb{E}_i x_i$ and $\Sigma_0 = \mathbb{E}_i (x_i - \mu_0)(x_i - \mu_0)^\top$ be the sample mean and covariance respectively. Let $X_{ij} = \frac{1}{2}(x_i - x_j)(x_i - x_j)^\top$, and let $a_{ij} \in [0,1]$ for $i, j \in [n]$ and a_i for $i \in [n]$ be such that*

(1) $a_{ij} = a_{ji}$ *for all* i, j,

(2) $\mathbb{E}_{ij} a_{ij} \geqslant 1 - 4\varepsilon$, *and*

(3) $\mathbb{E}_j a_{ij} \geqslant a_i(1-2\varepsilon)$ *for all* i, *and* $a_{ij} \leqslant a_i$ *for all* i *and* j.

Then, with probability $1 - O(\delta)$, for all $v \in \mathbb{R}^d$ and symmetric $P \in \mathbb{R}^{d\times d}$, we have

(1) $|\langle \mu - \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{v^\top \Sigma v}$,

(2) $\left|\mathop{\mathbb{E}}_i a_i\langle x_i-\mu_0, v\rangle\right| \leqslant \tilde{O}(\varepsilon)\cdot\sqrt{v^\top\Sigma_0 v}$,

(3) $\left|\mathop{\mathbb{E}}_i a_i[\langle x_i-\mu_0, v\rangle^2 - v^\top\Sigma_0 v]\right| \leqslant \tilde{O}(\varepsilon)\cdot v^\top\Sigma_0 v$.

(4) $\left|\langle\Sigma_0-\Sigma, P\rangle\right| \leqslant \tilde{O}(\varepsilon)\|\Sigma^{1/2}P\Sigma^{1/2}\|_F$,

(5) $\left|\mathop{\mathbb{E}}_i\left[\langle(x_i-\mu_0)(x_i-\mu_0)^\top-\Sigma_0, P\rangle^2 - 2\|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2\right]\right| \leqslant \tilde{O}(\varepsilon)\cdot\|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2$,

(6) $\left|\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij}-\Sigma_0, P\rangle\right| \leqslant \tilde{O}(\varepsilon)\cdot\|\Sigma^{1/2}P\Sigma^{1/2}\|_F$, *and*

(7) $\left|\mathop{\mathbb{E}}_{ij} a_{ij}\left[\langle X_{ij}-\Sigma_0, P\rangle^2 - 2\|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2\right]\right| \leqslant \tilde{O}(\varepsilon)\cdot\|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2$.

The final property we will need of Gaussians is *certifiable subgaussianity*, which says that certain moment inequalities have low-degree SoS proofs.

Lemma 18 (Certifiable fourth moments of Gaussian samples, Section 5 in Kothari and Steurer (2017)) *Let $\varepsilon, \delta > 0$, and $n \geqslant \tilde{O}((d\log(1/\delta)/\varepsilon)^2)$. Let $x_1, \ldots, x_n \sim \mathcal{N}(0, \Sigma)$ be samples from a d-dimensional Gaussian. Then with probability $1-\delta$,*

$$h(x, v) := (3+\varepsilon)\langle v, \Sigma v\rangle^2 - \mathop{\mathbb{E}}_{i\leftarrow[n]}\langle x_i, v\rangle^4$$

has a degree-4 SoS proof of nonnegativity in v (Definition 14).

3. Mean and Covariance Estimation of Gaussians via SoS

In this section, we prove **??** 3**??** 6. First, we prove Theorem 3. Then, we prove Theorem 6 in Section 3.2. We combine **??** 3**??** 6 to prove Corollary 7 in Appendix A.

3.1. Analyzing the canonical SoS program: proof of Theorem 3

We now prove Theorem 3, restated below.

Theorem *[Restatement of Theorem 3] Algorithm 2 takes as input an ε-corrupted sample of size n from a Gaussian distribution with mean μ and covariance Σ and in $\mathrm{poly}(n)$-time, outputs estimates $\hat{\mu} \in \mathbb{R}^d, \hat{\Sigma} \in \mathbb{R}^{d\times d}$ with the following guarantee. If $\Sigma \succeq 2^{-\mathrm{poly}(d)} I$, and $n \geqslant \tilde{O}(d^2\log^5(1/\delta)/\varepsilon^2)$, with probability at least $1-\delta$ over the draw of the original uncorrupted sample X, the estimates $\hat{\mu}, \hat{\Sigma}$ satisfy:*

(1) (Mean estimation) $\left\|\Sigma^{-1/2}(\hat{\mu}-\mu)\right\|_2 \leqslant \tilde{O}(\varepsilon)$, *and*

(2) (Covariance estimation in spectral norm) $(1-\tilde{O}(\varepsilon))\Sigma \preceq \hat{\Sigma} \preceq (1+\tilde{O}(\varepsilon))\Sigma$.

For convenience, we shall assume without loss of generality that εn is an integer; this can be done by changing ε by at most a constant factor.

The canonical degree-12 SoS relaxation of Algorithm 2 outputs a degree-12 pseudo-expectation $\tilde{\mathbb{E}}$ in the variables $x'_1, \ldots, x'_n \in \mathbb{R}^d$, $w_1, \ldots, w_n \in \mathbb{R}$, satisfying the constraints of Algorithm 2, if such a $\tilde{\mathbb{E}}$ exists. The estimates produced by the algorithm are $\hat{\mu} := \tilde{\mathbb{E}}[\mu']$ and $\hat{\Sigma} := \tilde{\mathbb{E}}[\Sigma']$.

Let $x_1, \dots, x_n \sim \mathcal{N}(\mu, \Sigma)$. Let $\mu_0 = \mathbb{E}_i\, x_i$ be the sample mean, and let $\Sigma_0 = \mathbb{E}_i (x_i - \mu_0)(x_i - \mu_0)^\top$ be the sample covariance. Fix $\varepsilon \in (0, 1)$, and let $y_1, \dots, y_n$ be an ε-corruption of $x_1, \dots, x_n$.

By Lemma 17, with probability $1 - \delta$, the following inequalities hold for any $a_1, \dots, a_n \in [0, 1]$ with $\sum_i a_i \geqslant (1 - 2\varepsilon)n$ and $v \in \mathbb{R}^d$:

$$|\langle \mu - \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{v^\top \Sigma v}\ , \tag{3.1}$$

$$\left|\mathbb{E}_i\, a_i \langle x_i - \mu_0, v\rangle\right| \leqslant \tilde{O}(\varepsilon) \cdot \sqrt{v^\top \Sigma_0 v}\ , \tag{3.2}$$

$$\left|\mathbb{E}_i\, a_i [\langle x_i - \mu_0, v\rangle^2 - v^\top \Sigma_0 v]\right| \leqslant \tilde{O}(\varepsilon) \cdot v^\top \Sigma_0 v\ . \tag{3.3}$$

Next, we let $X_{ij} := \frac{1}{2}(x_i - x_j)(x_i - x_j)^\top$, for any $i, j \in [n]$. Let $T \subseteq [0, 1]^{n^2}$ denote the set of $(a_{ij})_{i,j\in[n]}$ such that:

(1) $a_{ij} = a_{ji}$ for all i, j,

(2) $\sum_{i,j=1}^n a_{ij} \geqslant (1 - 4\varepsilon)n$, and

(3) there exist $a_1, \dots, a_n \in [0, 1]$ such that $a_i \geqslant \mathbb{E}_j\, a_{ij} \geqslant a_i(1 - 2\varepsilon)$ for all $i \in [n]$.

By Lemma 17 (setting $P = vv^\top$), with probability $1 - \delta$, the following inequalities hold for any $(a_{ij})_{i,j\in[n]} \in T$:

$$\left|v^\top(\Sigma_0 - \Sigma)v\right| \leqslant \tilde{O}(\varepsilon) v^\top \Sigma_0 v\ , \tag{3.4}$$

$$\left|\mathbb{E}_i [(\langle x_i - \mu_0, v\rangle^2 - v^\top \Sigma_0 v)^2 - 2(v^\top \Sigma_0 v)^2]\right| \leqslant \tilde{O}(\varepsilon) \cdot (v^\top \Sigma_0 v)^2\ , \tag{3.5}$$

$$\left|\mathbb{E}_{ij}\, a_{ij} [v^\top (X_{ij} - \Sigma_0)v]\right| \leqslant \tilde{O}(\varepsilon) v^\top \Sigma_0 v\ , \tag{3.6}$$

$$\left|\mathbb{E}_{ij}\, a_{ij} [(v^\top (X_{ij} - \Sigma_0)v)^2 - 2(v^\top \Sigma_0 v)^2]\right| \leqslant \tilde{O}(\varepsilon)(v^\top \Sigma_0 v)^2\ . \tag{3.7}$$

We proceed with the rest of the proof, assuming that the above resilience conditions hold. From this point on, we will no longer need to use the randomness of the x_i's.

3.1.1. Feasibility

Let us now argue that the constraints in Algorithm 2 are feasible. Set $x'_i = x_i$ for each $i \in [n]$, and let $w_i = 1$ if $y_i = x_i$ and 0 otherwise. Constraints (1), (2), (3) of Algorithm 2 are clearly satisfied, so it remains to argue that constraint (4) is satisfied. By Eq. (3.5) (with $a_i = 1$ for all i) constraint (4) is satisfied. Hence, the constraints in Algorithm 2 are feasible. In particular, Algorithm 2 will output in $\mathrm{poly}(n)$ time a degree-12 pseudo-expectation $\tilde{\mathbb{E}}$ in the variables $x'_1, \dots, x'_n, w_1, \dots, w_n$, satisfying the constraints of Algorithm 2. From this point on, we shall think of the pseudo-expectation $\tilde{\mathbb{E}}$ as fixed.

In light of the above, we summarize our notation in the box below.

Notation:

- μ, Σ, the true mean/covariance of the Gaussian $\mathcal{N}(\mu, \Sigma)$
- μ_0, Σ_0, the sample mean/covariance of the true samples $x_1, \dots, x_n$
- μ', Σ', the SoS variables for the mean/covariance

- $\hat{\mu} = \tilde{\mathbb{E}}[\mu'], \hat{\Sigma} = \tilde{\mathbb{E}}[\Sigma']$, the estimates for the mean/covariance
- $y_1, \ldots, y_n$, the ε-corruption of the true samples $x_1, \ldots, x_n$
- $x'_1, \ldots, x'_n$, the SoS variables for the samples
- $w_1, \ldots, w_n$, the SoS variables for the indicators $\mathbf{1}(x'_i = y_i)$

3.1.2. GUARANTEES FOR THE MEAN

We now analyze the estimate $\hat{\mu} := \tilde{\mathbb{E}}[\mu'] = \tilde{\mathbb{E}}[\mathbb{E}_i x'_i]$ for the mean μ, where $\tilde{\mathbb{E}}$ is a degree-12 pseudo-expectation satisfying the constraints in Algorithm 2. We need to show that $\hat{\mu}$ satisfies $|\langle \hat{\mu} - \mu_0, v \rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{v^\top \Sigma v}$.

The key ingredient in the proof is the following lemma, which we prove in Section 4.

Lemma 19 *Let $x_1, \ldots, x_n \in \mathbb{R}^d$. Suppose that there is some $\Sigma \in \mathbb{R}^{d \times d}$ such that for all $v \in \mathbb{R}^d$ and $a \in [0,1]^d$ with $\sum_{i=1}^n a_i \geqslant (1-2\varepsilon)n$, we have*

$$|\mathop{\mathbb{E}}_i a_i \langle x_i - \mu_0, v \rangle| \leqslant \tilde{O}(\varepsilon) \cdot \sqrt{v^\top \Sigma_0 v} \quad \text{and} \quad |\mathop{\mathbb{E}}_i a_i [\langle x_i - \mu_0, v \rangle^2 - v^\top \Sigma_0 v]| \leqslant \tilde{O}(\varepsilon) v^\top \Sigma_0 v \ .$$

Let $y_1, \ldots, y_n$ be any ε-corruption of $x_1, \ldots, x_n$, let $\tilde{\mathbb{E}}$ be a degree-6 pseudo-expectation in the variables $x'_1, \ldots, x'_n \in \mathbb{R}^d$ and $w_1, \ldots, w_n \in \mathbb{R}$. Let $\mu' = \mathbb{E}_i x'_i$. Suppose that

(1) $\tilde{\mathbb{E}}$ satisfies $w_i^2 = w_i$ for every $i \in [n]$,

(2) $\tilde{\mathbb{E}}$ satisfies $\sum_{i=1}^n w_i = (1-\varepsilon)n$,

(3) $\tilde{\mathbb{E}}$ satisfies $w_i x'_i = w_i y_i$ for every $i \in [n]$,

(4) $\tilde{\mathbb{E}}[\mathbb{E}_i \langle x'_i - \mu', v \rangle^2] \leqslant v^\top \hat{\Sigma} v$ for every $v \in \mathbb{R}^d$

Then, for every $v \in \mathbb{R}^d$, it holds that:

$$|\langle \hat{\mu} - \mu_0, v \rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{v^\top \Sigma_0 v} + \sqrt{O(\varepsilon) \cdot v^\top (\hat{\Sigma} - \Sigma_0) v + \tilde{O}(\varepsilon^2) v^\top (\hat{\Sigma} + \Sigma_0) v} \ .$$

We now finish the proof, assuming Lemma 19. We first observe that the hypotheses of Lemma 19 hold. Indeed, the two resilience conditions of Lemma 19 follow by Eqs. (3.2) and (3.3). Second, $\tilde{\mathbb{E}}$ is a degree-12 pseudo-expectation (and so is also degree-6) with the required properties: (1) – (3) clearly hold, and (4) follows from the definition of Σ', as $\hat{\Sigma} = \tilde{\mathbb{E}}[\Sigma']$. As the hypotheses of Lemma 19 are satisfied, we thus conclude that

$$|\langle \hat{\mu} - \mu_0, v \rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{v^\top \Sigma_0 v} + \sqrt{O(\varepsilon) \cdot v^\top (\hat{\Sigma} - \Sigma_0) v + \tilde{O}(\varepsilon^2) v^\top (\hat{\Sigma} + \Sigma_0) v} \ . \tag{3.8}$$

Suppose that the estimate for the covariance $\hat{\Sigma}$ satisfies the desired conclusion, i.e., that $|v^\top (\hat{\Sigma} - \Sigma) v| \leqslant \tilde{O}(\varepsilon) v^\top \Sigma v$ for all $v \in \mathbb{R}^d$ (we will prove this next). Then, Eqs. (3.4) and (3.8) imply that

$$|\langle \hat{\mu} - \mu_0, v \rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{v^\top \Sigma v} \ .$$

Finally, by Eq. (3.1), we conclude that

$$|\langle \hat{\mu} - \mu, v \rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{v^\top \Sigma v} \ ,$$

assuming that $\hat{\Sigma}$ satisfies its desired property. By choosing v appropriately, this implies (1) in Theorem 3.

We note here that this also proves Corollary 8. Indeed, this is because in Corollary 8 we are already given a good estimate $\hat{\Sigma}$ of Σ as input, and so the assumption that we have such an estimate is satisfied. As we have only used the resilience of the first two moments (Lemma 15) in the analysis above, the number of samples n that we need in Corollary 8 thus comes from Lemma 15, i.e., we only need $n \geqslant O((d + \log(1/\delta))/\varepsilon^2)$.

3.1.3. SPECTRAL GUARANTEES ON THE COVARIANCE

We now analyze the estimate $\hat{\Sigma} := \tilde{\mathbb{E}}[\Sigma']$ for the covariance, where $\tilde{\mathbb{E}}$ is a degree-12 pseudo-expectation satisfying Algorithm 2. First, we observe that the polynomial $\Sigma' := \mathbb{E}_i (x'_i - \mu')(x'_i - \mu')^\top$ is equal to $\mathbb{E}_{ij} X'_{ij}$ where $X'_{ij} := \frac{1}{2}(x'_i - x'_j)(x'_i - x'_j)^\top$, and similarly we also have $\Sigma_0 = \mathbb{E}_{ij} X_{ij}$, where $X_{ij} := \frac{1}{2}(x_i - x_j)(x_i - x_j)^\top$.

Let $T \subseteq [0,1]^{n^2}$ denote the set of $(a_{ij})_{i,j \in [n]}$ such that:

(1) $a_{ij} = a_{ji}$ for all i, j,

(2) $\sum_{i,j=1}^n a_{ij} \geqslant (1 - 4\varepsilon)n$, and

(3) there exist $a_1, \dots, a_n \in [0,1]$ such that $\mathbb{E}_j a_{ij} \geqslant a_i(1 - 2\varepsilon)$ for all i, and $a_{ij} \leqslant a_i$ for all i and j.

The key ingredient here is the following lemma, which is very similar to Lemma 19 that appeared in the case of mean estimation.

Lemma 20 *Let $X_1, \dots, X_{n^2} \in \mathbb{R}^{d \times d}$, and let $\Sigma_0 := \mathbb{E}_{ij} X_{ij}$. Let $T \subseteq [0,1]^{n^2}$. Suppose that, for all $v \in \mathbb{R}^d$ and $a \in T$ such that $\sum_{i,j} a_{ij} \geqslant (1 - 4\varepsilon)n^2$, we have*

$$|\mathop{\mathbb{E}}_{ij} a_{ij} v^\top (X_{ij} - \Sigma_0) v| \leqslant \tilde{O}(\varepsilon) \cdot v^\top \Sigma_0 v \quad \text{and} \quad |\mathop{\mathbb{E}}_{ij} a_{ij} [(v^\top (X_{i,j} - \Sigma_0) v)^2 - 2(v^\top \Sigma_0 v)^2]| \leqslant \tilde{O}(\varepsilon)(v^\top \Sigma_0 v)^2 \ .$$

Let $Y_1, \dots, Y_{n^2}$ be any $(2\varepsilon - \varepsilon^2)$-corruption of $X_1, \dots, X_{n^2}$, let $\tilde{\mathbb{E}}$ be a degree-6 pseudo-expectation in the variables $X'_1, \dots, X'_{n^2} \in \mathbb{R}^{d \times d}$ and $w_1, \dots, w_{n^2} \in \mathbb{R}$. Let $\Sigma' = \mathbb{E}_{ij} X'_{ij}$. Suppose that

(1) $\tilde{\mathbb{E}}$ satisfies $w_{ij}^2 = w_{ij}$ for every $i, j \in [n]$,

(2) $\tilde{\mathbb{E}}$ satisfies $\sum_{i,j=1}^N w_{ij} = (1 - \varepsilon)^2 n^2$,

(3) $\tilde{\mathbb{E}}$ satisfies $w_{ij} X'_{ij} = w_{ij} Y_{ij}$ for every $i, j \in [n]$,

(4) $\tilde{\mathbb{E}}[\mathbb{E}_{ij} (v^\top (X'_{ij} - \Sigma') v)^2] \leqslant (2 + \tilde{O}(\varepsilon))\, \tilde{\mathbb{E}}[(v^\top \Sigma' v)^2]$ for every $v \in \mathbb{R}^d$, and

(5) $a \in T$, where a is the vector with $a_{ij} := \tilde{\mathbb{E}}[w_{ij}] \mathbf{1}(X_{ij} = Y_{ij})$ for each $i, j \in [n]$.

Then, for every $v \in \mathbb{R}^d$, the following hold:

$$\tilde{\mathbb{E}}(v^\top (\Sigma' - \Sigma_0) v)^2 \leqslant O(\varepsilon)(\tilde{\mathbb{E}}(v^\top \Sigma' v)^2 + (v^\top \Sigma_0 v)^2) \ ,$$
$$|\langle \hat{\Sigma} - \Sigma_0, v \rangle| \leqslant \tilde{O}(\varepsilon) v^\top \Sigma_0 v + \sqrt{\tilde{\mathbb{E}}[\mathop{\mathbb{E}}_{ij}[(1 - w'_{ij}) \cdot v^\top (X'_{ij} - \Sigma_0) v]^2]} \ ,$$

where $w'_{ij} := w_{ij}\mathbf{1}(X_{ij} = Y_{ij})$, $\hat{\Sigma} := \tilde{\mathbb{E}}[\Sigma']$, *and*

$$\tilde{\mathbb{E}}[\mathop{\mathbb{E}}_{ij}[(1 - w'_{ij}) \cdot v^\top (X'_{ij} - \Sigma_0)v]^2] \leqslant O(\varepsilon) \cdot (\tilde{\mathbb{E}}(v^\top \Sigma' v)^2 - (v^\top \Sigma_0 v)^2) + \tilde{O}(\varepsilon^2) \cdot (\tilde{\mathbb{E}}(v^\top \Sigma' v)^2 + (v^\top \Sigma_0 v)^2) \ .$$

As before, we postpone the proof of Lemma 20 to Section 4, and use it to finish the proof.

We apply Lemma 20 as follows. First, we note that Σ_0 defined in Lemma 20 is the same as the sample mean Σ_0. Next, let T be the subset of vectors $(a_{ij})_{i,j\in[n]}$ defined earlier. We see that Eqs. (3.6) and (3.7) imply that the X_{ij}'s defined satisfy the resilience conditions in Lemma 20.

Now, we let $Y_{ij} = \frac{1}{2}(y_i - y_j)(y_i - y_j)^\top$, and let $X'_{ij} = \frac{1}{2}(x'_i - x'_j)(x'_i - x'_j)^\top$. We note that Σ' as defined in Lemma 20 is the same polynomial as Σ' defined earlier. We observe that the Y_{ij}'s must be a $(2\varepsilon - \varepsilon^2)$-corruption of the X_{ij}'s. Moreover, if we let $w_{ij} := w_i w_j$, then the pseudo-expectation defined by $\tilde{\mathbb{E}}$ on the polynomials X'_{ij} and w_{ij} is a degree-6 pseudo-expectation, and additionally satisfies properties (1) – (3). To see that (4) holds, we observe the following polynomial equality:

$$\mathop{\mathbb{E}}_{ij}(v^\top (X'_{ij} - \Sigma')v)^2] = \frac{1}{2}(\mathop{\mathbb{E}}_{i}\langle x'_i - \mu', v\rangle^4 + (v^\top \Sigma' v)^2) \ .$$

Combining with constraint (4) in Algorithm 2 and taking pseudo-expectations shows that property (4) holds.

Finally, property (5) in Lemma 20 holds, as $(a_{ij})_{i,j\in[n]} \in T$ because it satisfies the required properties with respect to the vector $a_i = \tilde{\mathbb{E}}[w_i]\mathbf{1}(x_i = y_i)$ for each i.

Thus, by Lemma 20, we have

$$|\langle \hat{\Sigma} - \Sigma_0, vv^\top \rangle| \leqslant \tilde{O}(\varepsilon) \cdot v^\top \Sigma_0 v + \sqrt{R} \ ,$$

where

$$R := \tilde{\mathbb{E}}[\mathop{\mathbb{E}}_{i}[\mathop{\mathbb{E}}_{j}[(1 - w'_{ij}) \cdot v^\top (X_{ij} - \Sigma_0)v]^2] \leqslant O(\varepsilon) \cdot (\tilde{\mathbb{E}}[(v^\top \Sigma' v)^2] - (v^\top \Sigma_0 v)^2) + \tilde{O}(\varepsilon^2) \cdot (\tilde{\mathbb{E}}[(v^\top \Sigma' v)^2] + (v^\top \Sigma_0 v)^2) \ .$$

Write $\Sigma' = A + B$, where $B = \mathbb{E}_{ij}(1 - w'_{ij})X'_{ij}$ and $A = \mathbb{E}_{ij}\, w'_{ij} X'_{ij} = \mathbb{E}_{ij}\, w'_i X_{ij}$; the latter equality holds because the following polynomial equalities are satisfied by $\tilde{\mathbb{E}}$:

$$w'_{ij}X'_{ij} = w_i w_j \mathbf{1}(x_i = y_i)\mathbf{1}(x_j = y_j) \cdot \frac{1}{2}(x'_i - x'_j)(x'_i - x'_j)^\top$$
$$= w_i w_j \mathbf{1}(x_i = y_i)\mathbf{1}(x_j = y_j) \cdot \frac{1}{2}(y_i - y_j)(y_i - y_j)^\top = w_i w_j \mathbf{1}(x_i = y_i)\mathbf{1}(x_j = y_j) \cdot \frac{1}{2}(x_i - x_j)(x_i - x_j)^\top \ .$$

Additionally, let $A_v := v^\top A v$ and $B_v := v^\top B v$, and $\Sigma_v = v^\top \Sigma_0 v$. We have that

$$\tilde{\mathbb{E}}[A_v^2] = \tilde{\mathbb{E}}[(\mathop{\mathbb{E}}_{ij} w'_{ij} v^\top X_{ij} v)^2] = \mathop{\mathbb{E}}_{i_1,j_1}\mathop{\mathbb{E}}_{i_2,j_2}\tilde{\mathbb{E}}[w'_{i_1 j_1} w'_{i_2 j_2}] \cdot v^\top X_{i_1 j_1} v \cdot v^\top X_{i_2 j_2} v$$
$$\leqslant \mathop{\mathbb{E}}_{i_1,j_1}\mathop{\mathbb{E}}_{i_2,j_2}\sqrt{\tilde{\mathbb{E}}[w'_{i_1,j_1}]\,\tilde{\mathbb{E}}[w'_{i_2,j_2}]} \cdot v^\top X_{i_1 j_1} v \cdot v^\top X_{i_2 j_2} v \text{ (as } \tilde{\mathbb{E}}[{w'_{ij}}^2] = \tilde{\mathbb{E}}[w'_{ij}])$$
$$= \tilde{\mathbb{E}}[\mathop{\mathbb{E}}_{i,j}\sqrt{\tilde{\mathbb{E}}[w'_{ij}]} v^\top X_{ij} v]^2 \leqslant (1 + \tilde{O}(\varepsilon))\Sigma_v^2 \text{ (by Eq. (3.6) applied to } a_{ij} = \sqrt{\tilde{\mathbb{E}}[w'_{ij}]})$$

We now bound R. In this notation, we have

$$R = \tilde{\mathbb{E}}[(B_v - \mathop{\mathbb{E}}_{ij}[1 - w'_{ij}] \cdot \Sigma_v)^2] \leqslant O(\varepsilon) \cdot (\tilde{\mathbb{E}}[(A_v + B_v)^2] - \Sigma_v^2) + \tilde{O}(\varepsilon^2) \cdot (\tilde{\mathbb{E}}[(A_v + B_v)^2] + \Sigma_v^2) \ . \tag{3.9}$$

First, we have

$$\tilde{\mathbb{E}}[(B_v - \mathop{\mathbb{E}}_{ij}[1 - w'_{ij}] \cdot \Sigma_v)^2] = \tilde{\mathbb{E}}[B_v^2 + \mathop{\mathbb{E}}_{ij}[1 - w'_{ij}]^2 \cdot \Sigma_v^2 - 2B_v \mathop{\mathbb{E}}_{ij}[1 - w'_{ij}] \cdot \Sigma_v] \geqslant \tilde{\mathbb{E}}[B_v^2] - 4\varepsilon\Sigma_v \tilde{\mathbb{E}}[B_v] \ , \tag{3.10}$$

as $\Sigma_v \geqslant 0$ and $\tilde{\mathbb{E}}$ satisfies $B_v \geqslant 0$ because B_v is a sum-of-squares polynomial. As $\tilde{\mathbb{E}}[A_v^2] \leqslant (1 + \tilde{O}(\varepsilon))\Sigma_v^2$ and $\tilde{\mathbb{E}}[A_v B_v]^2 \leqslant \tilde{\mathbb{E}}[A_v^2]\,\tilde{\mathbb{E}}[B_v^2]$ (by Proposition 13), it follows that

$$\tilde{\mathbb{E}}[(A_v + B_v)^2] \leqslant \tilde{\mathbb{E}}[B_v^2] + 2\sqrt{\tilde{\mathbb{E}}[A_v^2]\,\tilde{\mathbb{E}}[B_v^2]} + (1 + \tilde{O}(\varepsilon))\Sigma_v^2 \leqslant \tilde{\mathbb{E}}[B_v^2] + 2\Sigma_v\sqrt{\tilde{\mathbb{E}}[B_v^2]} + (1 + \tilde{O}(\varepsilon))\Sigma_v^2 \ . \tag{3.11}$$

Combining Eqs. (3.9) to (3.11) thus yields

$$\tilde{\mathbb{E}}[B_v^2] - 4\varepsilon\Sigma_v\,\tilde{\mathbb{E}}[B_v] \leqslant R \leqslant O(\varepsilon)(\tilde{\mathbb{E}}[B_v^2] + 2\Sigma_v\sqrt{\tilde{\mathbb{E}}[B_v^2]} + \tilde{O}(\varepsilon)\Sigma_v^2) + \tilde{O}(\varepsilon^2) \cdot \Sigma_v^2 \ .$$

Rearranging, applying $\tilde{\mathbb{E}}[B_v] \leqslant \sqrt{\tilde{\mathbb{E}}[B_v^2]}$, and solving for $\tilde{\mathbb{E}}[B_v^2]$ yields

$$\begin{aligned} &\tilde{\mathbb{E}}[B_v^2] \leqslant \tilde{O}(\varepsilon^2) \cdot \Sigma_v^2 \\ &\implies R \leqslant O(\varepsilon)(\tilde{\mathbb{E}}[B_v^2] + 2\Sigma_v\sqrt{\tilde{O}(\varepsilon^2)\Sigma_v^2} + \tilde{O}(\varepsilon)\Sigma_v^2) = \tilde{O}(\varepsilon^2)\Sigma_v^2 \\ \implies &|v^\top(\hat{\Sigma} - \Sigma_0)v| \leqslant \tilde{O}(\varepsilon) \cdot v^\top\Sigma_0 v + \sqrt{R} = \tilde{O}(\varepsilon) \cdot v^\top\Sigma_0 v \ . \end{aligned}$$

This is the desired spectral norm guarantee, only with Σ_0 in place of Σ. Using Eq. (3.4) and the triangle inequality, we have $|v^\top(\hat{\Sigma} - \Sigma)v| \leqslant \tilde{O}(\varepsilon)v^\top\Sigma v$, and so we thus have the desired spectral norm guarantee. This finishes the proof, as we have shown that $\hat{\mu}$ satisfies its desired property, assuming that $\hat{\Sigma}$ has this property.

3.2. Relative Frobenius guarantees on the covariance: proof of Theorem 6

We now prove Theorem 6, restated below.

Theorem *[Restatement of Theorem 6] Algorithm 5 takes as input an ε-corrupted sample of size n from a Gaussian distribution with mean μ and covariance Σ with $(1 - \tilde{O}(\varepsilon))\mathbb{I} \preceq \Sigma \preceq (1 + \tilde{O}(\varepsilon))\mathbb{I}$, and in $\mathrm{poly}(n)$-time, outputs an estimate $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ with the following guarantee. If $n \geqslant \tilde{O}(d^2 \log^5(1/\delta)/\varepsilon^2)$, then with probability at least $1 - \delta$ over the draw of the original uncorrupted sample X, the estimate $\hat{\Sigma}$ satisfies $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbb{I}\|_F \leqslant \tilde{O}(\varepsilon)$.*

Let $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \Sigma)$, where $(1 - \tilde{O}(\varepsilon))\mathbb{I} \preceq \Sigma \preceq (1 + \tilde{O}(\varepsilon))\mathbb{I}$. Fix $\varepsilon \in (0, 1)$, and let $y_1, \ldots, y_n$ be an ε-corruption of $x_1, \ldots, x_n$. Let $\mu_0 = \mathbb{E}_i\, x_i$ be the sample mean, and let $\Sigma_0 = \mathbb{E}_i(x_i - \mu_0)(x_i - \mu_0)^\top$ be the sample covariance.

We observe that for every symmetric $P \in \mathbb{R}^{d\times d}$, it holds that

$$|\|P\|_F - \|\Sigma^{1/2}P\Sigma^{1/2}\|_F| \leqslant \tilde{O}(\varepsilon)\min(\|P\|_F, \|\Sigma^{1/2}P\Sigma^{1/2}\|_F) \ , \tag{3.12}$$

as $(1 - \tilde{O}(\varepsilon))\mathbb{I} \preceq \Sigma \preceq (1 + \tilde{O}(\varepsilon))\mathbb{I}$, using the standard inequality $\|AB\|_F \leqslant \|A\|_2\|B\|_F$.

For each $i, j \in [n]$, let $X_{ij} := \frac{1}{2}(x_i - x_j)(x_i - x_j)^\top$. Let $T \subseteq [0, 1]^{n^2}$ denote the set of $(a_{ij})_{i,j\in[n]}$ such that:

(1) $a_{ij} = a_{ji}$ for all i, j,

(2) $\sum_{i,j=1}^{n} a_{ij} \geqslant (1-4\varepsilon)n$, and

(3) there exist $a_1, \dots, a_n \in [0,1]$ such that $\mathbb{E}_j a_{ij} \geqslant a_i(1-2\varepsilon)$ for all i, and $a_{ij} \leqslant a_i$ for all i and j.

By Lemma 17, with probability $1-\delta$ the following hold for any $(a_{ij}) \in T$ and symmetric $P \in \mathbb{R}^{d\times d}$:

$$\big|\langle \Sigma_0 - \Sigma, P\rangle\big| \leqslant \tilde{O}(\varepsilon)\|\Sigma^{1/2} P \Sigma^{1/2}\|_F \ , \tag{3.13}$$

$$\big|\mathop{\mathbb{E}}_i \big[\langle (x_i-\mu_0)(x_i-\mu_0)^\top - \Sigma_0, P\rangle^2 - 2\|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2\big]\big| \leqslant \tilde{O}(\varepsilon)\cdot \|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2 \ , \tag{3.14}$$

$$\big|\mathop{\mathbb{E}}_{ij} a_{ij}[\langle X_{ij}, P\rangle - \langle \Sigma_0, P\rangle]\big| \leqslant \tilde{O}(\varepsilon)\cdot \|\Sigma^{1/2}P\Sigma^{1/2}\|_F \ , \tag{3.15}$$

$$\big|\mathop{\mathbb{E}}_{ij} a_{ij}\big[\langle X_{ij}-\Sigma_0, P\rangle^2 - 2\|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2\big]\big| \leqslant \tilde{O}(\varepsilon)\cdot \|\Sigma^{1/2}P\Sigma^{1/2}\|_F^2 \ . \tag{3.16}$$

From the above, feasibility is simple: set $x'_i = x_i$ for all i, $w_i = \mathbf{1}(x_i = y_i)$, and observe that now $\mu' = \mu_0$, $\Sigma' = \Sigma_0$, and constraint (4) in Algorithm 5 is satisfied by Eq. (3.14) as $\|\Sigma^{1/2}P\Sigma^{1/2}\|_F \leqslant (1+\tilde{O}(\varepsilon))\|P\|_F$. Thus, Algorithm 5 will output in $\mathrm{poly}(n)$ time a degree-12 pseudo-expectation $\tilde{\mathbb{E}}$ satisfying the constraints in Algorithm 5.

Covariance estimation in Frobenius norm. We now analyze the output $\hat{\Sigma} := \tilde{\mathbb{E}}[\Sigma']$ of the algorithm. We observe that $\mathbb{E}_{ij} X_{ij}$ is equal to the sample covariance Σ_0. Let $Y_{ij} := \frac{1}{2}(y_i - y_j)(y_i-y_j)^\top$, and let $X'_{ij} := \frac{1}{2}(x'_i - x'_j)(x'_i - x'_j)^\top$. Similarly, we have that the SoS variable $\Sigma' := \mathbb{E}_i (x'_i - \mu')(x'_i-\mu')^\top$ is equal to $\mathbb{E}_{ij} X'_{ij}$.

The key ingredient in the proof is the following technical lemma, which we prove in Section 4. This lemma is similar to Lemmas 19 and 20.

Lemma 21 *Let $X_1, \dots, X_{n^2} \in \mathbb{R}^{d\times d}$, and let $\Sigma_0 := \mathbb{E}_{ij} X_{ij}$. Let $T \subseteq [0,1]^{n^2}$. Suppose that, for all symmetric $P \in \mathbb{R}^{d\times d}$ and $a \in T$, we have*

$$|\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij}-\Sigma_0, P\rangle| \leqslant \tilde{O}(\varepsilon)\cdot\|P\|_F \quad \text{and} \quad |\mathop{\mathbb{E}}_{ij} a_{ij}[\langle X_{ij}-\Sigma_0, P\rangle^2 - 2\|P\|_F^2]| \leqslant \tilde{O}(\varepsilon)\|P\|_F^2 \ .$$

Let $Y_1, \dots, Y_{n^2}$ be any $(2\varepsilon-\varepsilon^2)$-corruption of $X_1, \dots, X_{n^2}$, and let $\tilde{\mathbb{E}}$ be a degree-6 pseudo-expectation in the variables $X'_1, \dots, X'_{n^2} \in \mathbb{R}^{d\times d}$ and $w_1, \dots, w_{n^2} \in \mathbb{R}$. Let $\Sigma' = \mathbb{E}_{ij} X'_{i,j}$. Suppose that

(1) $\tilde{\mathbb{E}}$ satisfies $w_{ij}^2 = w_{ij}$ for every $i, j \in [n]$,

(2) $\tilde{\mathbb{E}}$ satisfies $\sum_{i,j=1}^{n} w_{ij} = (1-\varepsilon)^2 n^2$,

(3) $\tilde{\mathbb{E}}$ satisfies $w_{ij} X'_{ij} = w_{ij} Y_{ij}$ for every $i, j \in [n]$,

(4) $\tilde{\mathbb{E}}[\mathbb{E}_{ij}\langle X'_{ij} - \Sigma', v\rangle^2] \leqslant (2+\tilde{O}(\varepsilon))\|P\|_F^2$ for every symmetric $P \in \mathbb{R}^{d\times d}$, and

(5) $a \in T$, where a is the vector with $a_{ij} := \tilde{\mathbb{E}}[w_{ij}]\mathbf{1}(X_{ij} = Y_{ij})$ for each $i, j \in [n]$.

Then, for every symmetric $P \in \mathbb{R}^{d\times d}$, it holds that

$$|\langle \hat{\Sigma} - \Sigma_0, P\rangle| \leqslant \tilde{O}(\varepsilon)\|P\|_F \ ,$$

where $\hat{\Sigma} = \tilde{\mathbb{E}}[\Sigma']$.

We now apply Lemma 21. We observe that by Eqs. (3.15) and (3.16), and using Eq. (3.12), the resilience condition in Lemma 21 is satisfied by the X_{ij}'s. We also observe that the pseudo-expectation $\tilde{\mathbb{E}}$, in the variables X'_{ij} and w_{ij} with $w_{ij} := w_i w_j$, is a degree-6 pseudo-expectation, and trivially satisfies properties (1) – (3). Property (4) follows as $\tilde{\mathbb{E}}$ satisfies constraint (4) in Algorithm 5 and the following polynomial equality holds:

$$\mathop{\mathbb{E}}_{ij}\langle X'_{ij} - \Sigma', v\rangle^2 = \frac{1}{2}\mathop{\mathbb{E}}_{i}\langle (x'_i - \mu')(x'_i - \mu')^\top, P\rangle^2 + \frac{1}{2}\langle \Sigma', v\rangle^2 \ .$$

Property (5) follows by using the vector a with $a_i = \tilde{\mathbb{E}}[w_i]\mathbf{1}(x_i = y_i)$ to show membership of $(a_{ij})_{i,j\in[n]}$ in T.

We thus have by Lemma 21 that $|\langle \hat{\Sigma} - \Sigma_0, P\rangle| \leqslant \tilde{O}(\varepsilon)\|P\|_F$ for all symmetric $P \in \mathbb{R}^{d\times d}$. Using Eq. (3.13), it follows that $|\langle \hat{\Sigma} - \Sigma, P\rangle| \leqslant \tilde{O}(\varepsilon)\|P\|_F$. Hence,

$$\begin{aligned}|\langle \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbb{I}, P\rangle| &= |\langle \hat{\Sigma} - \Sigma, \Sigma^{-1/2}P\Sigma^{-1/2}\rangle| \\ &\leqslant \tilde{O}(\varepsilon)\|\Sigma^{-1/2}P\Sigma^{-1/2}\|_F \leqslant \tilde{O}(\varepsilon)\|P\|_F \ .\end{aligned}$$

Setting $P = \frac{\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbb{I}}{\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbb{I}\|_F}$, we conclude that

$$\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbb{I}\|_F \leqslant \tilde{O}(\varepsilon) \ ,$$

as required.

4. A Generic Estimation Lemma

Lemmas 19 to 21 are special cases of a generic technical result, which we now state and prove.

Lemma 22 *Let $x_1, \ldots, x_n \in \mathbb{R}^d$, and let $\mu_0 := \mathbb{E}_i\, x_i$. Let $V(\mu_0, v)$ for $v \in \mathbb{R}^d$ be a degree-2 polynomial in μ_0, and let $S \subseteq \mathbb{R}^d$ be a set such that $V(\mu_0, v) \geqslant 0$ for all $v \in S$ and $\mu_0 \in \mathbb{R}^d$.*

Let $T \subseteq [0,1]^n$. Suppose that, for all $v \in \mathbb{R}^d$ and $a \in T$ such that $\sum_i a_i \geqslant (1-\varepsilon)n$, we have

$$|\mathop{\mathbb{E}}_i a_i\langle x_i - \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\cdot\sqrt{V(\mu_0, v)} \quad \text{and} \quad |\mathop{\mathbb{E}}_i a_i[\langle x_i - \mu_0, v\rangle^2 - V(\mu_0, v)]| \leqslant \tilde{O}(\varepsilon)V(\mu_0, v) \ . \tag{4.1}$$

Let $y_1, \ldots, y_n$ be any ε-corruption of $x_1, \ldots, x_n$, let $\tilde{\mathbb{E}}$ be a degree-6 pseudo-expectation in the variables $x'_1, \ldots, x'_n \in \mathbb{R}^d$ and $w_1, \ldots, w_n \in \mathbb{R}$. Let $\mu' = \mathbb{E}_i\, x'_i$. Suppose that

(1) $\tilde{\mathbb{E}}$ satisfies $w_i^2 = w_i$ for every $i \in [n]$,

(2) $\tilde{\mathbb{E}}$ satisfies $\sum_{i=1}^n w_i = (1-\varepsilon)n$,

(3) $\tilde{\mathbb{E}}$ satisfies $w_i x'_i = w_i y_i$ for every $i \in [n]$,

(4) $\tilde{\mathbb{E}}[\mathbb{E}_i\langle x_i' - \mu', v\rangle^2] \leqslant (1+\tilde{O}(\varepsilon))\,\tilde{\mathbb{E}}[V(\mu', v)]$ *for every* $v \in S$*, and*

(5) $a \in T$*, where* a *is the vector with* $a_i := \tilde{\mathbb{E}}[w_i]\mathbf{1}(x_i = y_i)$ *for each* $i \in [n]$*.*

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

$$\tilde{\mathbb{E}}\langle \mu' - \mu_0, v\rangle^2 \leqslant O(\varepsilon)(\tilde{\mathbb{E}}\, V(\mu', v) + V(\mu_0, v)) \ ,$$
$$|\langle \hat{\mu} - \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{V(\mu_0, v)} + \sqrt{\tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i' - \mu_0, v\rangle]^2]} \ ,$$

where $\hat{\mu} := \tilde{\mathbb{E}}[\mu']$*, and*

$$\tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i' - \mu_0, v\rangle]^2] \leqslant O(\varepsilon)\cdot(\tilde{\mathbb{E}}\, V(\mu', v) - V(\mu_0, v)) + \tilde{O}(\varepsilon^2)\cdot(\tilde{\mathbb{E}}\, V(\mu', v) + V(\mu_0, v)) \ .$$

One should think of V as the variance of the distribution from which the x_i's are drawn, along the direction $v \in \mathbb{R}^d$. We now turn to the proof of Lemma 22.

Proof [Proof of Lemma 22] For each $i \in [n]$, let $w_i' = w_i \cdot \mathbf{1}(x_i = y_i)$. One should think of w_i as indicating that the algorithm "thinks" that $x_i = y_i$; the variable w_i' then indicates that the algorithm correctly "thinks" that $x_i = y_i$.

We now notice that the constraints ${w_i'}^2 = w_i'$, $w_i' x_i' = w_i' x_i$, and $\sum_i w_i' \geqslant (1-2\varepsilon)n$ are all satisfied by $\tilde{\mathbb{E}}$. Indeed, e.g., the fact that $w_i' x_i' = w_i' x_i$ is satisfied is consistent with the logic that if the algorithm thinks that $x_i = y_i$, then it chooses $x_i' = y_i = x_i$, and therefore $x_i' = x_i$.

We then have

$$\begin{aligned}
|\langle \hat{\mu} - \mu_0, v\rangle| &= |\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle x_i' - x_i, v\rangle| \\
&= |\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i w_i'\langle x_i' - x_i, v\rangle + \tilde{\mathbb{E}}\mathop{\mathbb{E}}_i(1-w_i')\langle x_i' - x_i, v\rangle| \\
&= 0 + |\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i(1-w_i')\langle x_i' - \mu_0 + \mu_0 - x_i, v\rangle| \quad \text{(because } \tilde{\mathbb{E}} \text{ satisfies } w_i' x_i' = w_i' x_i) \\
&\leqslant |\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i(1-w_i')\langle x_i' - \mu_0, v\rangle| + |\mathop{\mathbb{E}}_i\tilde{\mathbb{E}}[1-w_i']\langle x_i - \mu_0, v\rangle| \ .
\end{aligned}$$

One should notice that in the calculation above, we split the estimation error into the term when the algorithm "thinks" correctly and the error term, and then we "center" the error term about the sample mean μ_0.

We now apply the robustness assumption (4.1) to the second error term. Let $a_i = \tilde{\mathbb{E}}[w_i']$ for each i. We have that $\sum_{i=1}^n a_i \geqslant (1-2\varepsilon)n$ and $a_i \in [0,1]$ because $\tilde{\mathbb{E}}[{w_i'}^2] = \tilde{\mathbb{E}}[w_i']$, and $a \in T$ by assumption. Hence, again by assumption, we have that

$$|\mathop{\mathbb{E}}_i\tilde{\mathbb{E}}[w_i']\langle x_i - \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{V(\mu_0, v)} \quad \text{and} \quad |\mathop{\mathbb{E}}_i\langle x_i - \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{V(\mu_0, v)} \ .$$

Thus, $|\mathbb{E}_i\,\tilde{\mathbb{E}}[1-w_i']\langle x_i - \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{V(\mu_0, v)}$. We note here that the robustness assumption we apply is an inequality that holds "outside" the pseudo-expectation $\tilde{\mathbb{E}}$.

It thus remains to bound the first error term: $|\tilde{\mathbb{E}}\,\mathbb{E}_i(1-w_i')\langle x_i' - \mu_0, v\rangle|$. We do this by using constraint (4) to control its second moments.

First, by applying the Cauchy-Schwarz inequality, we have

$$|\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i(1-w_i')\langle x_i' - \mu_0, v\rangle| \leqslant \sqrt{\tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i' - \mu_0, v\rangle]^2]} \ ,$$

and that

$$\begin{aligned}
&\tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i'-\mu_0,v\rangle]^2] \\
&\leqslant \tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(1-w_i')^2]\cdot\mathop{\mathbb{E}}_i[(1-w_i')^2\langle x_i'-\mu_0,v\rangle^2]] && \text{(by Item (2) in Proposition 13)} \\
&= \tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[1-w_i']\cdot\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i'-\mu_0,v\rangle^2]] && \text{(as } \tilde{\mathbb{E}} \text{ satisfies } {w_i'}^2 = w_i') \\
&\leqslant 2\varepsilon\cdot\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i'-\mu_0,v\rangle^2] && \text{(as } \tilde{\mathbb{E}} \text{ satisfies } \mathop{\mathbb{E}}_i[1-w_i'] \leqslant 2\varepsilon)
\end{aligned}$$

Note that here we crucially need that $\tilde{\mathbb{E}}$ is a degree-6 pseudo-expectation, as $\mathbb{E}_i[(1-w_i')^2]\cdot\mathbb{E}_i[(1-w_i')^2\langle x_i'-\mu_0,v\rangle^2]$ is a degree-6 polynomial in the SoS variables $x_1',\dots,x_n'$ and $w_1,\dots,w_n$.

We thus need to control the second moments $\tilde{\mathbb{E}}\,\mathbb{E}_i[(1-w_i')\langle x_i'-\mu_0,v\rangle^2]$. Using constraint (4), we have that

$$\begin{aligned}
&\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i'-\mu_0,v\rangle^2] \\
&= \tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle x_i'-\mu_0,v\rangle^2 - \tilde{\mathbb{E}}\mathop{\mathbb{E}}_i w_i'\langle x_i'-\mu_0,v\rangle^2 \\
&= \tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle x_i'-\mu_0,v\rangle^2 - \mathop{\mathbb{E}}_i\tilde{\mathbb{E}}\, w_i'\langle x_i-\mu_0,v\rangle^2 && \text{(as } \tilde{\mathbb{E}} \text{ satisfies } w_i'x_i' = w_i'x_i) \\
&\leqslant \tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle x_i'-\mu'+\mu'-\mu_0,v\rangle^2 - (1-\tilde{O}(\varepsilon))V(\mu_0,v) && \text{(by Eq. (4.1), setting } a_i = \tilde{\mathbb{E}}[w_i']) \\
&= \tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle x_i'-\mu'\rangle^2 + \tilde{\mathbb{E}}\langle\mu'-\mu_0,v\rangle^2 + \tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle x_i'-\mu',v\rangle\langle\mu'-\mu_0,v\rangle - (1-\tilde{O}(\varepsilon))V(\mu_0,v) \\
&\leqslant \tilde{\mathbb{E}}[(1+\tilde{O}(\varepsilon))V(\mu',v)] + \tilde{\mathbb{E}}\langle\mu'-\mu_0,v\rangle^2 + 0 - (1-\tilde{O}(\varepsilon))V(\mu_0,v) && \text{(by constraint (4))} \\
&= \tilde{\mathbb{E}}\,V(\mu',v) - V(\mu_0,v) + \tilde{O}(\varepsilon)(\tilde{\mathbb{E}}\,V(\mu',v) + V(\mu_0,v)) + \tilde{\mathbb{E}}\langle\mu'-\mu_0,v\rangle^2 \ .
\end{aligned}$$

Again, we remark that Eq. (4.1), used above to *lower bound* the second moment, is an inequality that holds "outside" $\tilde{\mathbb{E}}$.

Finally, we upper bound $\tilde{\mathbb{E}}\langle\mu'-\mu_0,v\rangle^2$. We compute

$$\begin{aligned}
&\tilde{\mathbb{E}}\langle\mu'-\mu_0,v\rangle^2 \\
&= \tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(w_i'+(1-w_i'))\langle x_i'-x_i,v\rangle]^2] \\
&= \tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i'-x_i,v\rangle]^2] && \text{(as } \tilde{\mathbb{E}} \text{ satisfies } w_i'x_i' = w_i'x_i) \\
&\leqslant \tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(1-w_i')^2]\cdot\mathop{\mathbb{E}}_i[\langle x_i'-x_i,v\rangle^2]] \leqslant 2\varepsilon\,\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle x_i'-x_i,v\rangle^2 \\
&= 2\varepsilon\,\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle(x_i'-\mu')+(\mu'-\mu_0)+(\mu_0-x_i),v\rangle^2 \\
&\leqslant 6\varepsilon\,\tilde{\mathbb{E}}\mathop{\mathbb{E}}_i\langle x_i'-\mu',v\rangle^2 + 6\varepsilon\mathop{\mathbb{E}}_i\langle x_i-\mu_0,v\rangle^2 + 6\varepsilon\,\tilde{\mathbb{E}}\langle\mu'-\mu_0,v\rangle^2 && \text{(by Proposition 13)} \\
&\leqslant 6\varepsilon(1+\tilde{O}(\varepsilon))(\tilde{\mathbb{E}}\,V(\mu',v)+V(\mu_0,v)) + 6\varepsilon\,\tilde{\mathbb{E}}\langle\mu'-\mu_0,v\rangle^2 \ ,
\end{aligned}$$

and so it follows that $\tilde{\mathbb{E}}\langle\mu'-\mu_0,v\rangle^2 \leqslant O(\varepsilon)(\tilde{\mathbb{E}}\,V(\mu',v)+V(\mu_0,v))$.

Putting everything together, we conclude that:

$$\tilde{\mathbb{E}}[\mathop{\mathbb{E}}_i[(1-w_i')\langle x_i'-\mu_0,v\rangle]^2] \leqslant O(\varepsilon)\cdot(\tilde{\mathbb{E}}\,V(\mu',v)-V(\mu_0,v)) + \tilde{O}(\varepsilon^2)\cdot(\tilde{\mathbb{E}}\,V(\mu',v)+V(\mu_0,v))$$

and that

$$|\langle \hat{\mu} - \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\sqrt{V(\mu_0, v)} + \sqrt{\tilde{\mathbb{E}}[\mathbb{E}_i[(1-w_i')\langle x_i' - \mu_0, v\rangle]^2]} \ ,$$

for every v in S. ■

Acknowledgements

Pravesh K. Kothari is Supported by NSF CAREER Award #2047933 and an award from the Google Research Scholar program. Peter Manohar is supported in part by an ARCS Scholarship, NSF Graduate Research Fellowship (under Grant No. DGE1745016) and NSF CCF-1814603. Brian Hu Zhang is supported by the National Science Foundation under grants IIS-1718457, IIS-1901403, and CCF-1733556, and the ARO under award W911NF2010081.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *CoRR*, abs/2005.02970, 2020.

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: optimal rates in polynomial time. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 102–115. ACM, 2021.

Ainesh Bakshi, Ilias Diakonikolas, Samuel B. Hopkins, Daniel Kane, Sushrut Karmalkar, and Pravesh K. Kothari. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pages 149–159. IEEE, 2020a.

Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. Robustly learning mixtures of k arbitrary gaussians. *CoRR*, abs/2012.02119, 2020b. URL https://arxiv.org/abs/2012.02119.

Boaz Barak and David Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares, 2016.

Yeshwanth Cherapanamjeri, Samuel B. Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: regression, covariance estimation, and beyond. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 601–609. ACM, 2020.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 655–664. IEEE Computer Society, 2016.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 73–84. IEEE Computer Society, 2017.

Noah Fleming, Pravesh Kothari, and Toniann Pitassi. Semialgebraic proofs and efficient algorithm design. *Foundations and Trends in Theoretical Computer Science*, 14(1-2):1–221, 2019.

Samuel B. Hopkins. Sub-gaussian mean estimation in polynomial time. *CoRR*, abs/1809.07425, 2018. URL http://arxiv.org/abs/1809.07425.

Samuel B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1021–1034. ACM, 2018.

Sushrut Karmalkar, Adam R. Klivans, and Pravesh Kothari. List-decodable linear regression. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7423–7432, 2019.

Adam R. Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1420–1430. PMLR, 2018.

Pravesh K. Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. 2017.

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

Kevin A. Lai, Anup B. Rao, and Santosh S. Vempala. Agnostic estimation of mean and covariance. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 665–674. IEEE Computer Society, 2016.

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

Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaussians. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 518–531. ACM, 2021.

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

Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. California Institute of Technology, 2000.

Prasad Raghavendra and Morris Yau. List decodable subspace recovery. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 3206–3226. PMLR, 2020a.

Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 161–180. SIAM, 2020b.

Naum Z Shor. Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences*, 25:1–11, 1987.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *CoRR*, abs/1703.04940, 2017.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 45:1–45:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

Appendix A. Proof of Corollary 7

We prove Corollary 7. Let $x_1, \dots, x_{2n}$ be drawn from $\mathcal{N}(\mu, \Sigma)$, and let $y_1, \dots, y_{2n}$ be an ε-corruption of $x_1, \dots, x_{2n}$. Note that $y_1, \dots, y_n$ is an ε-corruption of $x_1, \dots, x_n$, and $y_{n+1}, \dots, y_{2n}$ is an ε-corruption of $x_{n+1}, \dots, x_{2n}$.

First, we run the algorithm in Theorem 3 on $y_1, \dots, y_n$: this yields an estimate $\hat{\mu}$ satisfying Item (1) of Corollary 7, and an estimate Σ_1 of Σ satisfying $|v^\top(\Sigma_1 - \Sigma)v| \leqslant \tilde{O}(\varepsilon)v^\top \Sigma v$ for all $v \in \mathbb{R}^d$.

Next, we run the algorithm in Theorem 6 on the transformed samples $\Sigma_1^{-1/2}y_{n+1}, \dots, \Sigma_1^{-1/2}y_{2n}$. We observe that these samples are an ε-corruption of $\Sigma_1^{-1/2}x_{n+1}, \dots, \Sigma_1^{-1/2}x_{2n}$, which are drawn from $\mathcal{N}(\mu, \Sigma_2)$, where $\Sigma_2 = \Sigma^{-1/2}\Sigma\Sigma^{-1/2}$. By our guarantee on Σ_1, we must have $(1 - \tilde{O}(\varepsilon))\mathbb{I} \preceq \Sigma_2 \preceq (1 + \tilde{O}(\varepsilon))\mathbb{I}$. Hence, the output of the algorithm in Theorem 6 is Σ_3 where Σ_3 satisfies $\|\Sigma_2^{-1/2}\Sigma_3\Sigma_2^{-1/2} - \mathbb{I}\|_F \leqslant \tilde{O}(\varepsilon)$.

Our final estimate for Σ is $\hat{\Sigma} := \Sigma_1^{1/2}\Sigma_3\Sigma_1^{1/2}$. We have that

$$\begin{aligned}
&\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbb{I}\|_F = \|\Sigma^{-1/2}(\Sigma_1^{1/2}\Sigma_3\Sigma_1^{1/2})\Sigma^{-1/2} - \mathbb{I}\|_F \\
&= \|(\Sigma^{-1/2}\Sigma\Sigma^{-1/2})^{-1/2}\Sigma_3(\Sigma^{-1/2}\Sigma\Sigma^{-1/2})^{-1/2} - \mathbb{I}\|_F = \|\Sigma_2^{-1/2}\Sigma_3\Sigma_2^{-1/2} - \mathbb{I}\|_F \leqslant \tilde{O}(\varepsilon) \ ,
\end{aligned}$$

where we use the following proposition. This finishes the proof of Corollary 7, as by Corollary 2.14 in Diakonikolas et al. (2016) we have the desired bound on the total variation distance.

Proposition 23 *Let $A, B \in \mathbb{R}^{d\times d}$ be symmetric, PSD matrices, with B invertible. Then for any invertible $C \in \mathbb{R}^{d\times d}$, it holds that*

$$\|B^{-1/2}AB^{-1/2} - \mathbb{I}\|_F = \|(CBC^\top)^{-1/2}CAC^\top(CBC^\top)^{-1/2} - \mathbb{I}\|_F$$

Proof Recall that for a symmetric matrix $M \in \mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1, \dots, \lambda_d$, $\|M\|_F = \sum_{i=1}^d \lambda_i^2$. It thus suffices to show that $B^{-1/2}AB^{-1/2}$ and $(CBC^\top)^{-1/2}CAC^\top(CBC^\top)^{-1/2}$ are equivalent up to an orthogonal change of basis.

Let $D = (CBC^\top)^{-1/2}CB^{1/2}$. We then have that $DB^{-1/2}AB^{-1/2}D^\top = (CBC^\top)^{-1/2}CAC^\top(CBC^\top)^{-1/2}$, so it remains to show that D is orthogonal, i.e., $DD^\top = D^\top D = \mathbb{I}$. We have

$$\begin{aligned}
DD^\top &= (CBC^\top)^{-1/2}CB^{1/2}B^{1/2}C^\top(CBC^\top)^{-1/2} = (CBC^\top)^{-1/2}\left((CBC^\top)^{1/2}\right)^2(CBC^\top)^{-1/2} = \mathbb{I} \\
D^\top D &= B^{1/2}C^\top(CBC^\top)^{-1/2}(CBC^\top)^{-1/2}CB^{1/2} = B^{1/2}C^\top(C^{-1})^\top B^{-1}C^{-1}CB^{1/2} = \mathbb{I} \ ,
\end{aligned}$$

which finishes the proof. ■

Appendix B. Quantifier Elimination in Sum-of-Squares

In this section we will justify why the SoS relaxations of **??** 2**??** 5, which are written as a family of infinitely many constraints, can be solved efficiently. The programs have the form

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^m \\ \text{s.t.} \quad & f_i(x) \geqslant 0 \qquad \forall i \\ & g_j(x) = 0 \qquad \forall j \\ & h(x, v) \geqslant 0 \quad \forall v \in \mathbb{R}^d. \end{aligned} \tag{B.1}$$

with $\mathrm{poly}(m)$ constraints f_i, g_j. As such, we need a way to express constraints of the form "$h(x, v) \geqslant 0$ for all $v \in \mathbb{R}^d$" within degree-k SoS. This will follow from the following result:

Lemma 24 (Quantifier elimination in SoS, e.g., Section 4.3.4 in Fleming et al. (2019)) *Suppose that there exists some $x^* \in \mathbb{R}^m$ such that $f_i(x^*) \geqslant 0$ for all i, $g_j(x^*) = 0$ for all j, and $h(x^*, \cdot)$ has a degree-k SoS proof of nonnegativity. Then a degree-k pseudoexpectation satisfying all constraints in (B.1) can be found by solving a semidefinite program of size $m^{O(k)}$.*

Intuitively, this is true because "$h(x^*, \cdot)$ has a degree-k SoS proof of nonnegativity" is equivalent to a particular moment matrix of size $m^{O(k)}$ being PSD, which can be expressed within SoS. We will use this result in two forms.

Lemma 25 (Lemma 4.27 in Fleming et al. (2019)) *If h is a quadratic form, then h has a degree-2 SoS proof of nonnegativity if and only if $h(v) \geqslant 0$ for every v.*

In **??** 2**??** 5, the constraint (4) is indeed a quadratic form in v (or P), so we are done. For constraint (5) in Algorithm 2, we require the certifiable hypercontractivity of Gaussians (Lemma 18), which we restate here:

Lemma *[Restatement of Lemma 18] Let $\varepsilon, \delta > 0$, and $n \geqslant \tilde{O}((d \log(1/\delta)/\varepsilon)^2)$. Let $x_1, \ldots, x_n \sim \mathcal{N}(0, \Sigma)$ be samples from a d-dimensional Gaussian. Then with probability $1 - \delta$,*

$$h(x, v) := (3 + \varepsilon)\langle v, \Sigma v\rangle^2 - \mathop{\mathbb{E}}_{i \leftarrow [n]} \langle x_i, v\rangle^4$$

has a degree-4 SoS proof of nonnegativity in v (Definition 14).

By rearranging, the constraint (5) in Algorithm 2 is exactly the condition that $h(x', v) \geqslant 0$ for all v in the above lemma, so the lemma states that a degree-4 SoS proof exists with high probability for the *true* samples $x'_i = x_i$. Thus, the conditions of Lemma 24 are satisfied, and we are done.

Appendix C. Deferred Proofs from Section 2.2

C.1. Proof of Lemma 16

Proof The first statement is

$$\left\| \mathop{\mathbb{E}}_i a_i [x_i x_i^\top - \mathbb{I}] \right\|_F \leqslant \tilde{O}(\varepsilon)$$

which is Corollary 4.8 in Diakonikolas et al. (2016). The second statement is

$$\left|\mathop{\mathbb{E}}_i a_i[\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 - 2\|P\|_F^2]\right| \leqslant \tilde{O}(\varepsilon)\|P\|_F^2.$$

By convexity, we may assume that $a_i \in \{0,1\}$ for all i. Let S be the set of indices i for which $a_i > 0$. We have:

$$\left|\mathop{\mathbb{E}}_i a_i[\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 - 2\|P\|_F^2]\right| \leqslant \left|\mathop{\mathbb{E}}_{i\sim[n]}\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 - 2\|P\|_F^2\right| + \varepsilon\left|\mathop{\mathbb{E}}_{i\sim[n]\setminus S}\langle x_i x_i^\top - \mathbb{I}, P\rangle^2\right|$$

We bound the two terms separately. Condition on the "good event" in Lemma 5.17 of Diakonikolas et al. (2016). Then,

$$\left|\mathop{\mathbb{E}}_{i\sim[n]}\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 - 2\|P\|_F^2\right| \leqslant O(\varepsilon)\|P\|_F^2$$

follows from Item 3 of Definition 5.15 in Diakonikolas et al. (2016) with $p(x) = \langle xx^T - \mathbb{I}, P\rangle/(\sqrt{2}\|P\|_F)$. The fact that

$$\varepsilon\left|\mathop{\mathbb{E}}_{i\sim[n]\setminus S}\langle x_i x_i^\top - \mathbb{I}, P\rangle^2\right| \leqslant \tilde{O}(\varepsilon)$$

follows from Lemma 5.21 of Diakonikolas et al. (2016) with the same choice of p. Combining these bounds completes the proof. ■

C.2. Proof of Lemma 17

The statements in Lemma 17 are similar to those in Lemma 15 and Lemma 16, so it should be reasonable to believe that they should hold. The proofs are tedious but ultimately mostly brute force.

All the statements are invariant to linear transformations, so assume WLOG that $\mu = 0$ and $\Sigma = \mathbb{I}$. Condition on the conclusions of Lemmas 15 and 16, which hold with high probability for the chosen n. Let $z_i = x_i - \mu_0$ for notational simlpicity.

In the proof, instead of the stated conditions on a, we will use instead normalized vectors, namely, we will assume that $\mathbb{E}_{ij} a_{ij} = \mathbb{E}_i a_i = 1$, $\mathbb{E}_j a_{ij} = a_i \leqslant 1 + O(\varepsilon)$, and $a_{ij} \leqslant a_i(1 + O(\varepsilon))$. Since this amounts to nothing but scaling the coefficients by $1 + O(\varepsilon)$, the conclusions of Lemmas 15 and 16 hold verbatim.

(1) $|\langle \mu_0, v\rangle| \leqslant \tilde{O}(\varepsilon)\|v\|_2$

Proof Set $a_i = 1$ for all i in Lemma 15. ■

(2) $\left|\mathop{\mathbb{E}}_i a_i\langle x_i - \mu_0, v\rangle\right| \leqslant \tilde{O}(\varepsilon)\cdot\sqrt{v^\top\Sigma_0 v}$

Proof By (1) above and Lemma 15, we have

$$\left|\mathop{\mathbb{E}}_i a_i\langle x_i - \mu_0, v\rangle\right| \leqslant \left|\mathop{\mathbb{E}}_i a_i\langle x_i, v\rangle\right| + \left|\mathop{\mathbb{E}}_i a_i\langle \mu_0, v\rangle\right| \leqslant \tilde{O}(\varepsilon)\|v\|_2$$

But $\|v\|_2 = (1 \pm \tilde{O}(\varepsilon))\sqrt{v^\top\Sigma_0 v}$ by (4) (with $P = vv^\top$), so we are done. ■

The following intermediate results will be useful in the remaining proofs.

Lemma 26 $\|P\mu_0\|_2 \leqslant \tilde{O}(\varepsilon)\|P\|_F$, $|\mathbb{E}_i\, a_i\langle x_i\mu_0^\top, P\rangle| \leqslant \tilde{O}(\varepsilon^2)\|P\|_F$, *and* $|\mu_0^\top P\mu_0| \leqslant \tilde{O}(\varepsilon^2)\|P\|_F$.

Proof The first inequality is $\|P\mu_0\|_2 \leqslant \|P\|_2\|\mu_0\|_2 \leqslant \tilde{O}(\varepsilon)\|P\|_F$ by Lemma 15 and $\|P\|_2 \leqslant \|P\|_F$.

The second is $|\mathbb{E}_i\, a_i\langle x_i\mu_0^\top, P\rangle| \leqslant \tilde{O}(\varepsilon)\|P\mu_0\|_2 \leqslant \tilde{O}(\varepsilon^2)\|P\|_F$, by Lemma 15 the first inequality.

The third is the second when $a_i = 1$ for all i. ■

(3) $\left|\mathop{\mathbb{E}}_i a_i\langle z_i z_i^\top - \mathbb{I}, P\rangle\right| \leqslant \tilde{O}(\varepsilon)\|P\|_F$

(The statement in the lemma follows from setting $P = vv^\top$ and applying (4) below, but we will need this more generic statement later, so this is the one we prove.)

Proof We have

$$\mathop{\mathbb{E}}_i a_i\langle (x_i-\mu_0)(x_i-\mu_0)^\top, P\rangle = \mathop{\mathbb{E}}_i a_i\langle x_i x_i^\top, P\rangle + \langle \mu_0\mu_0^\top, P\rangle - 2\mathop{\mathbb{E}}_i a_i\langle x_i\mu_0^\top, P\rangle$$

The first term is $\mathbb{E}_i\, a_i\langle x_i x_i^\top, P\rangle = \langle \mathbb{I}, P\rangle \pm \tilde{O}(\varepsilon)\|P\|_F$ by Lemma 16, and the other terms are $\pm\tilde{O}(\varepsilon)\|P\|_F$ by Lemma 26. ■

(4) $\left|\langle \Sigma_0 - \mathbb{I}, P\rangle\right| \leqslant \tilde{O}(\varepsilon)\|P\|_F$

Proof Set $a_i = 1$ for all i in (3). ■

(5) $\left|\mathop{\mathbb{E}}_i\langle z_i z_i^\top - \Sigma_0, P\rangle^2 - 2\|P\|_F^2\right| \leqslant \tilde{O}(\varepsilon)\cdot\|P\|_F^2$

Proof We have:

$$\mathop{\mathbb{E}}_i a_i\langle z_i z_i^\top - \Sigma_0, P\rangle^2 = \mathop{\mathbb{E}}_i a_i\langle z_i z_i^\top - \mathbb{I}, P\rangle^2 + \langle \Sigma_0 - \mathbb{I}, P\rangle^2 - 2\mathop{\mathbb{E}}_i a_i\langle z_i z_i^\top - \mathbb{I}, P\rangle\langle \Sigma_0 - \mathbb{I}, P\rangle$$

For the second term, we have $\langle \Sigma_0 - \mathbb{I}, P\rangle^2 \leqslant \tilde{O}(\varepsilon^2)\|P\|_F^2$ by (4). For the third term, we have

$$\left|\mathop{\mathbb{E}}_i a_i\langle z_i z_i^\top - \mathbb{I}, P\rangle\langle \Sigma_0 - \mathbb{I}, P\rangle\right| \leqslant \tilde{O}(\varepsilon)\|P\|_F\left|\mathop{\mathbb{E}}_i a_i\langle z_i z_i^\top - \mathbb{I}, P\rangle\right| \leqslant \tilde{O}(\varepsilon^2)\|P\|_F^2$$

by (4) and then (3). That only leaves the first term. We have:

$$\begin{aligned}
\mathop{\mathbb{E}}_i a_i\langle (x_i-\mu_0)(x_i-\mu_0)^\top - \mathbb{I}, P\rangle^2 &= \mathop{\mathbb{E}}_i a_i\big\langle (x_i x_i^\top - \mathbb{I}) + \mu_0\mu_0^\top - 2x_i\mu_0^\top, P\big\rangle^2 \\
&= \mathop{\mathbb{E}}_i a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 + \langle \mu_0\mu_0^\top, P\rangle^2 + 4\mathop{\mathbb{E}}_i a_i\langle x_i\mu_0^\top, P\rangle^2 \\
&\quad + \langle \mu_0\mu_0^\top, P\rangle\mathop{\mathbb{E}}_i a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle - 2\langle \mu_0\mu_0^\top, P\rangle\mathop{\mathbb{E}}_i a_i\langle x_i\mu_0^\top, P\rangle \\
&\quad - 2\mathop{\mathbb{E}}_i a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle\langle x_i\mu_0^\top, P\rangle
\end{aligned}$$

We analyze each term separately. The first term is $(2 \pm \tilde{O}(\varepsilon))\|P\|_F^2$ by Lemma 16, so it suffices to show that all remaining terms are small. The second term is $\tilde{O}(\varepsilon^4)\|P\|_F^2$ by Lemma 26. The third

term is $|\mathbb{E}_i\, a_i\langle x_i, P\mu_0\rangle^2| = (1 \pm \tilde{O}(\varepsilon))\|P\mu_0\|_2^2 \leqslant \tilde{O}(\varepsilon^2)\|P\|_F^2$ by Lemmas 15 and 26. The fourth term is $\tilde{O}(\varepsilon^3)\|P\|_F^2$ by Lemmas 16 and 26. The fifth term is $\tilde{O}(\varepsilon^4)\|P\|_F^2$ by Lemma 26. For the final term, we have

$$\left|\mathop{\mathbb{E}}_i a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle\langle x_i\mu_0^\top, P\rangle\right| \leqslant \sqrt{\left(\mathop{\mathbb{E}}_i a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle^2\right)\left(\mathop{\mathbb{E}}_i a_i\langle x_i\mu_0^\top, P\rangle^2\right)}$$

by Cauchy-Schwarz. The first term is $\mathbb{E}_i\, a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 = (2 \pm \tilde{O}(\varepsilon))\|P\|_F^2$, and the second term is $\tilde{O}(\varepsilon^2)\|P\|_F^2$ as argued above. Combining these yields $\left|\mathbb{E}_i\, a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle\langle x_i\mu_0^\top, P\rangle\right| \leqslant \tilde{O}(\varepsilon)\|P\|_F^2$, as needed. ■

The proofs of the remaining two bounds will make repeated use of the following generic technique, roughly speaking. Suppose that resilience (Lemmas 15 and 16) gives us $|\mathbb{E}_i\, a_i z_i| = \tilde{O}(\varepsilon)B$, and we want to argue that $|\mathbb{E}_{ij}\, a_{ij} z_i z_j| = \tilde{O}(\varepsilon)B^2$. This is not immediate, because $a_{ij} \neq a_i a_j$ in general. Instead, we write $\mathbb{E}_{ij}\, a_{ij} z_i z_j = \mathbb{E}_i\, z_i\, \mathbb{E}_j\, a_{ij} z_j$, and apply resilience to the inner expectation (noting that the vector whose jth entry is a_{ij}/a_i is, by construction, a valid resilience vector) to find $|\mathbb{E}_j\, a_{ij} z_j| \leqslant a_i\tilde{O}(\varepsilon)B$ for each i, so that $|\mathbb{E}_{ij}\, a_{ij} z_i z_j| \leqslant B|\mathbb{E}_i\, z_i\tilde{O}(\varepsilon)|$. In this expression, the $\tilde{O}(\varepsilon)$ may depend on i. Let b_i be the term hidden by the $\tilde{O}(\varepsilon)$ for each i, and let $a_i' := 1 - b_i + \mathbb{E}\, b_i$. Note that $\mathbb{E}_i\, a_i' = 1$ and $a_i' = 1 \pm \tilde{O}(\varepsilon)$ for all i, so a' is a valid input to the resilience condition. Thus, we have

$$|\mathop{\mathbb{E}}_i b_i z_i| \leqslant (1 + \mathop{\mathbb{E}}_i b_i)|\mathop{\mathbb{E}}_i z_i| + |\mathop{\mathbb{E}}_i a_i z_i|$$

Now applying resilience to each of the two terms separately gives $|\mathbb{E}_i\, b_i z_i| \leqslant \tilde{O}(\varepsilon)B$, so $|\mathbb{E}_{ij}\, a_{ij} z_i z_j| \leqslant \tilde{O}(\varepsilon)B^2$, as desired.

The following intermediate result will also be useful:

Lemma 27 $\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij}, P\rangle = \mathop{\mathbb{E}}_i a_i\langle x_i x_i^\top, P\rangle \pm \tilde{O}(\varepsilon)\|P\|_F$

Proof We have

$$\begin{aligned}\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij}, P\rangle &= \frac{1}{2}\mathop{\mathbb{E}}_{ij} a_{ij}\langle (x_i - x_j)(x_i - x_j)^\top, P\rangle \\ &= \mathop{\mathbb{E}}_i a_i\langle x_i x_i^\top, P\rangle - \mathop{\mathbb{E}}_{ij} a_{ij}\langle x_i x_j^\top, P\rangle\end{aligned}$$

where we use the symmetry of P and the a_{ij}s. It thus remains to bound the last term. Write $P = \sum_k \lambda_k v_k v_k^\top$ for orthonormal vectors v_k. Note that $\|\sum_k \lambda_k v_k\|_2 = \sqrt{\sum_k \lambda_k^2} = \|P\|_F$ by Pythagorean theorem. Then:

$$\mathop{\mathbb{E}}_{ij} a_{ij}\langle x_i x_j^\top, P\rangle = \mathop{\mathbb{E}}_i \sum_k \lambda_k\langle v_k, x_i\rangle \mathop{\mathbb{E}}_j a_{ij}\langle v_k, x_j\rangle = \mathop{\mathbb{E}}_i a_i(\pm\tilde{O}(\varepsilon))\left\langle \sum_k \lambda_k v_k, x_i\right\rangle = \tilde{O}(\varepsilon)\|P\|_F$$

by applying Lemma 15 twice using the generic technique.

■

We now prove the last two results in Lemma 17.

(6) $\left|\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \Sigma_0, P\rangle\right| \leqslant \tilde{O}(\varepsilon) \cdot \|P\|_F$

Proof By Lemma 27, we have:

$$\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \Sigma_0, P\rangle = \mathop{\mathbb{E}}_{i} a_i\langle x_i x_i^\top - \Sigma_0, P\rangle \pm \tilde{O}(\varepsilon^2)\|P\|_F$$

It thus only remains to bound the first term. We have

$$\big|\mathop{\mathbb{E}}_{i} a_i\langle x_i x_i^\top - \Sigma_0, P\rangle\big| \leqslant \big|\mathop{\mathbb{E}}_{i} a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle\big| + \big|\mathop{\mathbb{E}}_{i} a_i\langle \Sigma_0 - \mathbb{I}, P\rangle\big| \leqslant \tilde{O}(\varepsilon)\|P\|_F$$

by applying Lemma 16 and (4). ∎

(7) $\big|\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \Sigma_0, P\rangle^2 - 2\|P\|_F^2\big| \leqslant \tilde{O}(\varepsilon)\cdot\|P\|_F^2$

Proof We follow the same structure as the proof of (5) above. We have:

$$\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \Sigma_0, P\rangle^2 = \mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \mathbb{I}, P\rangle^2 + \langle \Sigma_0 - \mathbb{I}, P\rangle^2 - 2\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \mathbb{I}, P\rangle\langle \Sigma_0 - \mathbb{I}, P\rangle$$

For the second term, we have $\langle \Sigma_0 - \mathbb{I}, P\rangle^2 \leqslant \tilde{O}(\varepsilon^2)\|P\|_F^2$ by (4). For the third term, we have

$$\begin{aligned}\big|\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \mathbb{I}, P\rangle\langle \Sigma_0 - \mathbb{I}, P\rangle\big| &\leqslant \tilde{O}(\varepsilon)\|P\|_F\big|\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \mathbb{I}, P\rangle\big| \\ &= \tilde{O}(\varepsilon)\|P\|_F\big|\mathop{\mathbb{E}}_{i} a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle \pm \tilde{O}(\varepsilon)\|P\|_F\big| \\ &\leqslant \tilde{O}(\varepsilon^2)\|P\|_F^2\end{aligned}$$

by Lemmas 16 and 27. That only leaves the first term. We have

$$\begin{aligned}\mathop{\mathbb{E}}_{ij} a_{ij}\langle X_{ij} - \mathbb{I}, P\rangle^2 &= \mathop{\mathbb{E}}_{ij} a_{ij}\Big\langle \frac{1}{2}(x_i x_i^\top - \mathbb{I}) + \frac{1}{2}(x_j x_j^\top - \mathbb{I}) - x_i x_j^\top, P\Big\rangle^2 \\ &= \frac{1}{2}\mathop{\mathbb{E}}_{i} a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 + \mathop{\mathbb{E}}_{ij} a_{ij}\langle x_i x_j^\top, P\rangle^2 + \frac{1}{2}\mathop{\mathbb{E}}_{ij} a_{ij}\langle x_i x_i^\top - \mathbb{I}, P\rangle\langle x_j x_j^\top - \mathbb{I}, P\rangle \\ &\quad - 2\mathop{\mathbb{E}}_{ij} a_{ij}\langle x_i x_i^\top - \mathbb{I}, P\rangle\langle x_i x_j^\top, P\rangle\end{aligned}$$

The first term is $(1 \pm \tilde{O}(\varepsilon))\|P\|_F^2$ by Lemma 16. For the second term, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{ij} a_{ij}\langle x_i x_j^\top, P\rangle^2 &= \mathop{\mathbb{E}}_{i}\mathop{\mathbb{E}}_{j} a_{ij} x_i^\top P x_j x_j^\top P x_i \\ &= \mathop{\mathbb{E}}_{i}\mathop{\mathbb{E}}_{j} a_{ij}\langle x_j x_j^\top, P x_i x_i^\top P\rangle \\ &= \mathop{\mathbb{E}}_{i} a_i\big[\langle \mathbb{I}, P x_i x_i^\top P\rangle \pm \tilde{O}(\varepsilon)\|P x_i x_i^\top P\|_F\big] \\ &= \mathop{\mathbb{E}}_{i} a_i(1 \pm \tilde{O}(\varepsilon))\langle x_i x_i^\top, P^2\rangle \\ &= \langle \mathbb{I}, P^2\rangle \pm \tilde{O}(\varepsilon)\|P^2\|_F \\ &= (1 \pm \tilde{O}(\varepsilon))\|P\|_F^2\end{aligned}$$

where we use Lemma 16 twice (the second time exploiting the fact that $a_i(1 \pm \tilde{O}(\varepsilon))$ is still a valid resilience vector), and the last line uses $\langle \mathbb{I}, P^2\rangle = \|P\|_F^2$ and $\|P^2\|_F \leqslant \|P\|_F^2$.

Thus, it only remains to show that the other two terms are small. For the third term, we have

$$\begin{aligned}\big|\mathop{\mathbb{E}}_{ij} a_{ij}\langle x_i x_i^\top - \mathbb{I}, P\rangle\langle x_j x_j^\top - \mathbb{I}, P\rangle\big| &= \big|\mathop{\mathbb{E}}_{i}\langle x_i x_i^\top - \mathbb{I}, P\rangle \mathop{\mathbb{E}}_{j} a_{ij}\langle x_j x_j^\top - \mathbb{I}, P\rangle\big| \\ &\leqslant \|P\|_F \big|\mathop{\mathbb{E}}_{i} a_i(\pm\tilde{O}(\varepsilon))\langle x_i x_i^\top - \mathbb{I}, P\rangle\big| \\ &\leqslant \tilde{O}(\varepsilon)\|P\|_F^2.\end{aligned}$$

by the generic technique. For the final term, we have

$$\begin{aligned}\big|\mathop{\mathbb{E}}_{ij} a_{ij}\langle x_i x_i^\top - \mathbb{I}, P\rangle\langle x_i x_j^\top, P\rangle\big| &= \big|\mathop{\mathbb{E}}_{i} a_{ij}\langle x_i x_i^\top - \mathbb{I}, P\rangle\big(\mathop{\mathbb{E}}_{j} a_{ij}\langle x_j, Px_i\rangle\big)\big| \\ &= \big|\mathop{\mathbb{E}}_{i} a_i(\pm\tilde{O}(\varepsilon))\langle x_i x_i^\top - \mathbb{I}, P\rangle\|Px_i\|_2\big| \\ &\leqslant \sqrt{\big(\mathop{\mathbb{E}}_{i} a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle^2\big)\big(\mathop{\mathbb{E}}_{i} a_i\tilde{O}(\varepsilon^2)\|Px_i\|_2^2\big)}\end{aligned}$$

The first term is $\mathbb{E}_i\, a_i\langle x_i x_i^\top - \mathbb{I}, P\rangle^2 = (2 \pm \tilde{O}(\varepsilon))\|P\|_F^2$. For the second term, we have

$$\begin{aligned}\big|\mathop{\mathbb{E}}_{i} a_i\tilde{O}(\varepsilon^2)\|Px_i\|_2^2\big| &= \tilde{O}(\varepsilon^2)\big|\mathop{\mathbb{E}}_{i}\langle x_i x_i^\top, P^2\rangle\big| \\ &\leqslant \tilde{O}(\varepsilon^2)\big|\mathop{\mathbb{E}}_{i} a_i\langle x_i x_i^\top - \mathbb{I}, P^2\rangle\big| + \tilde{O}(\varepsilon^2)\langle\mathbb{I}, P^2\rangle \\ &\leqslant \tilde{O}(\varepsilon^2)\|P\|_F^2\end{aligned}$$

where the last line applies Lemma 16 to the first term (noting again that $\langle\mathbb{I}, P^2\rangle = \|P\|_F^2$), and the inequality $\|P^2\|_F \leqslant \|P\|_F^2$ to the second. Combining these yields $\big|\mathbb{E}_{ij}\, a_{ij}\langle x_i x_i^\top - \mathbb{I}, P\rangle\langle x_i x_j^\top, P\rangle\big| \leqslant \tilde{O}(\varepsilon)\|P\|_F^2$, which is what we needed. ■