List-decodable Covariance Estimation

Misha Ivkov[*]
mishai@stanford.edu

Pravesh K. Kothari [†]
praveshk@cs.cmu.edu

June 22, 2022

Abstract

We give the first polynomial time algorithm for *list-decodable covariance estimation*. For any $\alpha > 0$, our algorithm takes input a sample $Y \subseteq \mathbb{R}^d$ of size $n \geqslant d^{\mathrm{poly}(1/\alpha)}$ obtained by adversarially corrupting an $(1-\alpha)n$ points in an i.i.d. sample X of size n from the Gaussian distribution with unknown mean μ_* and covariance Σ_*. In $n^{\mathrm{poly}(1/\alpha)}$ time, it outputs a constant-size list of $k = k(\alpha) = (1/\alpha)^{\mathrm{poly}(1/\alpha)}$ candidate parameters that, with high probability, contains a $(\hat{\mu}, \hat{\Sigma})$ such that the total variation distance $TV(\mathcal{N}(\mu_*, \Sigma_*), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) < 1 - O_\alpha(1)$. This is the statistically strongest notion of distance and implies multiplicative spectral and relative Frobenius distance approximation with dimension independent error. Our algorithm works more generally for $(1-\alpha)$-corruptions of any distribution D that possesses low-degree sum-of-squares certificates of two natural analytic properties: 1) anti-concentration of one-dimensional marginals and 2) hypercontractivity of degree 2 polynomials.

Prior to our work, the only known results for estimating covariance in the list-decodable setting were for the special cases of list-decodable linear regression and subspace recovery [KKK19, RY19, BK21, RY20b]. The best-known algorithms for both these problems only yield a weak recovery guarantee that needs super-polynomial time for any sub-constant (in dimension d) target error for the parameters in natural norms. As a corollary, our result yields the first polynomial time *exact* algorithm for list-decodable linear regression and subspace recovery that, in particular, obtain $2^{-\mathrm{poly}(d)}$ error in polynomial-time in the underlying dimension. List-decodable setting also generalizes the problem of robust clustering non-spherical mixtures in the strong contamination model [BK20b, DHKK20] and the state of the art [BK20b] for this latter problem needs $d^{k^{O(k)}}$ samples and tolerates an $\varepsilon \ll k^{-O(k)}$ fraction outliers. Our result implies an algorithm with an improved running time and sample bound of $d^{\mathrm{poly}(k)}$ that handles a larger $\varepsilon \ll 1/\mathrm{poly}(k)$ fraction of outliers.

[*] Carnegie Mellon University & Stanford University.

[†] Carnegie Mellon University. Supported by NSF CAREER Award #2047933, a Alfred P. Sloan Research Fellowship and a Google Research Scholar Award.

Contents

1 Introduction

Can we accurately estimate the mean and covariance of a high-dimensional probability distribution D from a input sample with outliers? What properties of D allow the *robust estimation* of such basic parameters to be statistically and computationally tractable?

When outliers form a small constant (say $\leqslant 10\%$) fraction of the input data, we now have a good first-cut understanding of efficient robust estimation of basic parameters of distributions. The works [DKK+16, LRV16] invented the first polynomial time algorithms for the problem with dimension-independent error guarantees and invigorated the now active field of *high-dimensional robust statistics*. The ensuing follow-ups provide optimal guarantees for estimating the mean [KS17a, KS17b, HL17, DKS18], covariance and higher moments [KS17b] of a broad class of distributions while tolerating a small constant fraction of outliers. This progress has resulted in new broadly applicable techniques, abstracted out properties of the distributions[1] that make efficient robust estimation possible and even inspired progress on related problems such as finding optimal estimators for mean [Hop20, CHK+20, CFB19] and covariance of heavy-tailed distributions.

In contrast, much less is understood in the setting where a *majority* of the input data are outliers. Since unique recovery of parameters is clearly impossible in this setting, the goal is to compute a dimension-independent constant size *list* of candidate parameters one of which is close to those of the unknown distribution. This model was introduced by Blum, Balcan and Vempala [BBV08] to study an agnostic variant of clustering[2] without separation assumptions on the underlying input data. Indeed, list-decodable learning implies clustering algorithms *without any separation assumptions* and allows *partial cluster recovery* even when outliers obliterate multiple clusters completely.

The recent effort in designing algorithms that tolerate such overwhelming fraction of outliers began with the influential work of Charikar, Steinhardt and Valiant [CSV17]. In addition to the applications above, they argued that list-decodable learning is a natural model for learning from untrusted data and showed applications to semi-verified learning. Their work gave the first non-trivial guarantees for *list-decodable mean estimation* for distributions with *spherical* covariances (i.e. multiples of identity). Subsequent works obtained stronger guarantees for spherical Gaussians [DKS18] and more generally, Poincaré distributions [KS17a] with corollaries [HL17, KS17a] to clustering spherical mixtures at the statistically minimum mean separation. A recent sequence of works have even sped-up these results to almost linear time in certain settings [CMY20, DKK20a, DKK+20b].

List-decodable covariance estimation Despite this progress on mean estimation, the problem of *covariance estimation* in the list-decodable setting has turned out to be significantly more challenging. Prior works [KKK19, RY19] built a framework via the sum-of-squares method for list-decodable learning to make progress in two special cases: *list-decodable linear regression* [KKK19, RY19] (corresponds to the case where the unknown covariance is spherical in a subspace of co-dimension 1) and *list-decodable subspace recovery* [BK20a, BK21] (unknown covariance is spherical in an arbitrary subspace) by introducing the new tool of *certifiable anti-concentration*. However, there is an inherent

[1] In this case, efficiently verifiable certificates of upper bounds on directional moments.

[2] The "inliers" correspond to one of the clusters.

bottleneck in their approach (see Section 2.1) that leads to significantly weak error guarantees even for the special cases they study: the best known algorithms need *super-polynomial* time for any *sub-constant* error in the underlying dimension and do not appear to extend to settings when the unknown covariance has eigenvalues of different scales[3]. Recent works for the special case of clustering non-spherical Gaussian mixtures [BK20b, DHKK20] managed to wriggle out of the difficulty[4] by crucially relying on the input data being generated from a mixture of k Gaussians where every pair is *separated in total variation distance* and the fraction of outliers is at most $\leqslant k^{-\operatorname{poly}(k)}$.

This Work In this work, we design the first polynomial time algorithm for list-decodable covariance estimation. As immediate corollaries, we also obtain the first polynomial time *exact* algorithms for list-decodable linear regression and subspace recovery in $\mathbb{R}^d$ obtaining constant size lists of candidates one of which achieves as small as $2^{-\operatorname{poly}(d)}$ error in any natural norm and a $d^{\operatorname{poly}(k)}$ time algorithm for clustering non-spherical mixtures in the presence of $\varepsilon = 1/\operatorname{poly}(k)$ fraction outliers (the best known prior work needs $d^{k^{\operatorname{poly}(k)}}$ running time and sample complexity and tolerates $\leqslant k^{-O(k)}$-fraction outliers).

Our list-decodable covariance estimation algorithm relies on the coalescence of a number of sophisticated tools developed in robust statistics over the past few years. This includes the algorithmic certificates for basic probabilistic phenomenon such as certifiable subgaussianity [KS17b], certifiable hypercontractivity [KOTZ14, BK20b, DHKK20] and certifiable anti-concentration [KKK19, RY19] and the sum-of-squares framework for robust statistics.

The main idea that allows us to finally obtain an algorithm for all covariances is to abandon the previous “one-shot rounding" approach in related list-decodable learning algorithms and instead settle for a *coarse spectral recovery* guarantee via rounding a sum-of-squares relaxation to obtain a combination of multiplicative approximation for large eigenvalues and additive approximation for small eigenvalues. We then give an iterated pruning method that, that instead of relying on the strong certifiable anti-concentration (the bottleneck in the previous works that holds only for Gaussian-like distributions) property, crucially only needs the significantly milder *Paley-Zygmund* anti-concentration inequality that holds for all subgaussian distributions. Our final algorithm is obtained by an interleaved iteration of coarse spectral recovery, a new “subgaussian restriction” subroutine and the pruning procedure based on mild anti-concentration. We expect our algorithmic primitives to be useful in improving robust estimation algorithms that rely on certifiable anti-concentration to faster methods that apply to broader family of distributions.

1.1 Our Results

We now describe our main results in more detail. Our results formally hold in the following *strong contamination* model for list-decodable learning.

Definition 1.1 (Strong Contamination Model for List-Decodable Learning)**.** In the strong contamination model, a $(1-\alpha)$-corrupted sample of size n from a distribution D is generated by choosing

[3] For e.g., covariances such as $I + \log d \cdot vv^\top - uu^\top$ for unknown orthogonal unit vectors u, v.

[4] With substantial technical effort – see Section 2.2 of the overview in [BK20b] for a discussion.

an i.i.d. sample X of size n from D, and, adversarially switching any $(1-\alpha)n$ points to obtain Y.

Remark 1.2. This is the harshest studied model for robust estimation (and also used in [RY20c]). It generalizes the more commonly studied list-decodable learning model [CSV17, KKK19, RY19, BK21, DKP+21, CMY20, DKK20a, DKK+20b] where the input sample Y is obtained by *adding* $(1-\alpha)n$ outliers to an i.i.d. sample of size αn. In contrast, our model above allows both *adding and deleting* points from an independent sample of size n: our input Y can be generated by first selecting an arbitrarily "biased" subset of αn points from an i.i.d. sample X and then adding $(1-\alpha)n$ outliers. Our motivations for working with the harsher model are natural: we'd like to design algorithms that provide strong recovery guarantees under weakest possible modeling assumptions. A concrete advantage of our choice (see Corollary 1.9) is that the resulting algorithms (unlike standard list-decodable learning) imply significantly improved algorithms that are more sample efficient, faster, handle larger outlier rates for robust clustering of non-spherical mixtures in the *strong contamination model*.

Main Result Our estimation guarantees are in the following notion of distance between parameters. As we explain, this captures the information-theoretically strongest possible estimation guarantees in our setting.

Definition 1.3 (Parameter Distance)**.** We say that the distance $\mathsf{parameter\text{-}distance}((\mu,\Sigma),(\mu',\Sigma'))$ between two sets of mean-covariance pairs is at most Δ if the following three parameter distance bounds hold.

1. **Mahalanobis Mean Closeness:** $\forall v \in \mathbb{R}^d$, $\langle \mu - \mu', v\rangle^2 \leqslant \Delta v^\top(\Sigma + \Sigma')v$,
2. **Multiplicative Spectral Closeness:** $\forall v \in \mathbb{R}^d$, $\frac{1}{\Delta} v^\top \Sigma' v \leqslant v^\top \Sigma v \leqslant \Delta v^\top \Sigma' v$, and,
3. **Relative-Frobenius Closeness:** $\left\| \Sigma^{\dagger/2}\Sigma'\Sigma^{\dagger/2} - I \right\|_F \leqslant \Delta$.

If $\mathsf{parameter\text{-}distance} \leqslant \Delta$, we can conclude that the total variation distance between the corresponding Gaussians is at most $1 - \exp(-\Delta^{O(1)})$ (see Fact 3.24). As a result, obtaining recovery guarantees in $\mathsf{parameter\text{-}distance}$ translates into bounds on the total variation error with no dimension dependence. Total variation is the strongest possible (and arguably, the "right") notion of distance in this context and is the metric of choice in prior works on robust mean and covariance estimation [DKK+16]. Our main result is the following theorem that gives a polynomial time algorithm for list-decodable learning of mean and covariance of an unknown Gaussian distribution.

Theorem 1.4. *For every $d \in \mathbb{N}$, there is a $n^{\mathrm{poly}(1/\alpha)}$ time algorithm*[5] *that takes input a $(1-\alpha)$-corrupted sample of size $n \geqslant d^{\alpha^{-O(1)}}$ from a d-dimensional Gaussian distribution with mean μ_* and covariance Σ_* and outputs a list of $2^{O(1/\alpha^{O(1)})}$-parameters such that with probability at least* 0.99 *over the draw*

[5] Our algorithm works in the standard word RAM model. The running time of our algorithm is polynomial in the total bit complexity of the input *and of the unknown* Σ_*. The dependence on the bit-complexity of Σ_* is necessary, see Section 3.1 for a discussion. We note that our algorithm can also be formalized in the idealized "real RAM" model [BCSS98, EvM20] of real computation that is implicitly used in prior works but we choose not to do this.

of the uncorrupted sample X and the randomness of the algorithm, there is a $(\hat{\mu}, \hat{\Sigma})$ in the list satisfying: $\mathsf{parameter\text{-}distance}((\hat{\mu}, \hat{\Sigma}), (\mu_*, \Sigma_*)) \leqslant \Delta$, *for* $\Delta = \mathrm{poly}(1/\alpha)$. *As a corollary, we obtain that* $d_{\mathsf{TV}}(\mathcal{N}(\mu_*, \Sigma_*), \mathcal{N}(\mu, \Sigma') \leqslant 1 - \exp(-\alpha^{-O(1)})$ *where* $d_{\mathsf{TV}}(\cdot, \cdot)$ *denotes the total variation or statistical distance between two probability distributions.*

Remark 1.5. Observe, that our algorithm needs no assumptions on the unknown covariance Σ_*. In particular, Σ_* can have $\exp(d)$ large condition number and can be rank deficient. In fact, if Σ_* is singular, our algorithm must construct a candidate $\hat{\Sigma}$ with the *same* range space as Σ_* and thus, must recover the low-rank structure in Σ_* exactly. For the numerical issues that arise in obtaining this strong guarantee and how we handle them, we direct the reader to the discussion in Section 3.

Our main algorithmic innovation is a list-decodable learning algorithm for covariance estimation that achieves a *multiplicative spectral approximation* to the unknown covariance with a *dimension-independent* multiplicative factor. Even as a function of α, our guarantees are tight[6] up to constant factors in the exponent of α in Δ. We then use this estimate to obtain the stronger relative Frobenius distance recovery guarantees. Our result for mean estimation then follows by using our estimates to "isotropize" (and thus effectively make the covariance almost spherical) and applying a list-decodable mean estimation algorithm [KS17a] for covariances of bounded spectral norm.

Running Time and List Size: For any constant α, the running time and sample complexity of our algorithm is polynomial in the underlying dimension. The dependence on α of the running time and sample complexity is exponential. This appears necessary. As we explain below (see Corollary 1.9), our list-decodable covariance estimation algorithm implies an algorithm for robustly clustering well-separated (in total variation distance) mixtures of Gaussians. Even in the application to this special-case and to the setting *without any outliers*, known statistical query lower bounds [DKS17, DKP+21] suggest a lower bound of $d^{\Omega(1/\alpha)}$ time that matches our guarantees up to the exponent of the polynomial of $1/\alpha$. In terms of the list-size, our algorithm returns a list of size $2^{\mathrm{poly}(1/\alpha)}$. This is a dimension-independent constant but can likely be improved to the optimal $O(1/\alpha)$ bound.

List-decodable learning of all "reasonable" distributions Our algorithm more generally works for any distribution D on $\mathbb{R}^d$ as long as it satisfies two natural analytic properties of probability distributions identified in the context of robust non-spherical clustering in [BK20b]. Informally speaking, these properties ask for low-degree sum-of-squares certificates of *anti-concentration* and *hypercontractivity of degree 2 polynomials* of the distribution D (we postpone formal definitions to Section 3.3). While certifiable hypercontractivity of degree 2 polynomials is known to be true for uniform distribution on product domains (such as discrete/solid hypercube), we only have verified certifiable anti-concentration property for rotationally invariant distributions such as Gaussian distributions and affine transforms of uniform distribution on the unit sphere [KKK19, RY19, BK20b]. Our algorithm thus succeeds as is (and does not require the knowledge of moments of underlying distribution) for all such distributions. We believe that finding natural analytic properties that govern the success of algorithms adds to our understanding of robust estimation in general.

[6] Given only a $(1-\alpha)$-corrupted sample, we cannot distinguish between the 1-D Gaussians $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, \alpha^2)$ the variances of which are $1/\alpha^2$ multiplicatively far and $\Omega(1/\alpha)$ additively far.

Theorem 1.6 (See Theorem 8.6 for a detailed version)**.** *For any $\alpha > 0$, there is a $n^{\mathrm{poly}(1/\alpha)}$ time algorithm that takes input a $Y \subseteq \mathbb{R}^d$ of size n and outputs a $\mathcal{L}$ list of size $2^{\mathrm{poly}(1/\alpha)}$ of estimates $(\hat{\mu}, \hat{\Sigma})$ with the following guarantee. Suppose there is an i.i.d. sample X of size $n \geqslant n_0 = d^{\mathrm{poly}(1/\alpha)}$ from a certifiably $(C, \alpha^3/2C)$-anti-concentrated distribution D with mean μ_* and covariance Σ_* with C-certifiably hypercontractive degree 2 polynomials such that $|Y \cap X| = \alpha n$. Then, with probability at least 0.99 over the randomness of the algorithm, there exists a candidate $(\hat{\mu}, \hat{\Sigma})$ in the list $\mathcal{L}$ such that* $\textsf{parameter-distance}((\hat{\mu}, \hat{\Sigma}), (\mu_*, \Sigma_*)) \leqslant \mathrm{poly}(1/\alpha)$.

As an immediate consequence of our algorithm for list-decodable covariance estimation, we obtain improved guarantees for the previously studied problems of list-decodable linear regression and subspace recovery and clustering non-spherical mixtures.

Applications to Linear Regression In list-decodable linear regression, we are given a $(1-\alpha)$-corruption of a system of linear equations $\langle x_i, \ell_* \rangle = b_i$ where each x_i is drawn from Gaussian distribution and ℓ_* is an unknown unit vector. Introducing the key new tool of *certifiable anti-concentration*, Karmarkar, Klivans and Kothari [KKK19] and Raghavendra and Yau [RY19] gave an algorithm[7] for this problem with a running time of $n^{O(1/(\eta^4\alpha^4)}$ time to produce a list of size $O(1/\alpha)$ [KKK19] (the list size is a slightly larger bound of $O(1/\alpha^{\log(1/\alpha)})$ in [RY19]) that contains a $\hat{\ell}$ that is $\left\|\hat{\ell} - \ell_*\right\|_2 \leqslant \eta$. This running time was improved to $n^{O(\log(1/\eta)+1/\alpha^4)}$ (at the cost of a larger list size of $\alpha^{O(\log 1/\eta)}$) via a general *error reduction within SoS* method by Bakshi and Kothari [BK20a]. Note that both results assume that the covariance of x_is is known to be I and more importantly, for any target sub-constant error $\eta \to 0$ as $d \to \infty$, the running time required is super-polynomial. This is in fact the consequence of the recovery guarantees being in a norm weaker than total variation.

Observe that the coefficients of the uncorrupted set of equations (x_i, b_i) are distributed as $d+1$-dimensional Gaussian with mean 0 and covariance matrix I restricted to a subspace of co-dimension 1 – namely, the one orthogonal to the vector $(\ell_*, -1)$. Thus, list-decoding linear equations above is equivalent to list-decoding the (kernel of) the covariance. Our multiplicative spectral guarantees above for covariance estimation immediately yields an algorithm that can obtain an error as low as $\eta = 2^{-\mathrm{poly}(d)}$ in polynomial time. In fact, our algorithm is *exact* in the sense that the sample complexity does not depend on the target error η and the estimation error is entirely because of finite numerical precision in computing the output. In addition, unlike prior works, our algorithm does not need to know the covariance of x_is or the length of the unknown vector ℓ_*.

Corollary 1.7 (Exact Algorithm for list-decodable linear regression)**.** *For any $\alpha > 0$ and target error η, there is a $n^{\mathrm{poly}(1/\alpha)}$ $\mathrm{poly}\log(1/\eta)$-time algorithm for list-decodable linear regression that succeeds with probability at least 0.99 whenever $n \geqslant d^{\mathrm{poly}(1/\alpha)}$ and produces a list of $2^{\mathrm{poly}(1/\alpha)}$ candidate vectors such that there is an $\hat{\ell}$ in the list satisfying $\left\|\hat{\ell} - \ell_*\right\|_2 \leqslant \eta$.*

Applications to Subspace Recovery In list-decodable subspace recovery, we are given $(1-\alpha)n$ corrupted samples from $\mathcal{N}(0, \Pi)$ where Π is a projection matrix to a subspace of $\mathbb{R}^d$. In [RY20b], the authors gave an algorithm that takes such a set of points and in $n^{O(1)}$ time and $d^{O(1)}$ samples, finds

[7] We note that these algorithms can handle random additive noise in the equations of variance $\ll \alpha$.

a list of constant size containing a candidate $\hat{\Pi}$ such that $\left\|\hat{\Pi} - \Pi\right\|_F^2 \leqslant O(1/\alpha^5)$. The work [BK21] obtained the stronger guarantee of $\left\|\hat{\Pi} - \Pi\right\|_F^2 \leqslant \eta$ for arbitrarily small η in time $n^{O(\log 1/\eta)/\alpha^4}$ time whenever $n \geqslant d^{O(1/\alpha^4)}$. However, even this improved algorithm requires super-polynomial time to achieve any sub-constant recovery error.

As a direct corollary of our stronger multiplicative spectral approximation guarantee, we immediately obtain the first *exact* algorithm for list-decodable subspace recovery, that, in particular allows achieving even exponentially small errors in polynomial time.

Corollary 1.8 (Exact Algorithm for list-decodable subspace recovery)**.** *For any $\alpha > 0$ and target error η, there is a $n^{\mathrm{poly}(1/\alpha)} \operatorname{poly}\log(1/\eta)$-time algorithm for list-decodable subspace recovery that succeeds with probability at least 0.99 whenever $n \geqslant d^{\mathrm{poly}(1/\alpha)}$ and produces a list of $2^{\mathrm{poly}(1/\alpha)}$ candidate projection matrices such that there is an $\hat{\Pi}$ in the list satisfying $\left\|\hat{\Pi} - \Pi_*\right\|_F \leqslant \eta$.*

Applications to Robust Clustering of Non-Spherical Mixtures Our work immediately improves the best known prior algorithms for robust clustering of non-spherical mixtures in the "small outlier regime". Specifically, the goal in this problem is to take input an ε-corrupted sample from a mixture of k Gaussians with equal weights with means and covariances μ_i, Σ_i such that each pair is Δ-separated in parameter distance (equivalent to separated on total variation distance as discussed above), and output an estimate $\hat{\mu}_i, \hat{\Sigma}_i$ of the parameters of each component that are close in parameter distance (Definition 1.3). Two recent works obtained the first efficient algorithms for solving this problem. Specifically, the algorithm in [BK20b] obtains a $n = d^{k^{O(k)}}$-sample and $n^{k^{O(k)}}$ time algorithm for $\Delta = k^{\mathrm{poly}(k)}$-separated mixtures to obtain $k^{\mathrm{poly}(k)}\varepsilon$-close estimates in parameter distance as long as $\varepsilon \ll k^{-O(k)}$. Their algorithm succeeds more generally for mixtures of all "reasonable distributions" discussed above. The work of [DHKK20] obtains a $d^{F(k)}$ sample and $n^{F(k)}$ time algorithm that tolerates a $\varepsilon \ll 1/F(k)$ of outliers for the same problem when the components are Gaussians with $\Delta = F(k)$-separation where $F(k)$ is at most a poly(k) size tower of exponentials in k. While both algorithms are polynomial time for a fixed k, their running times and sample complexity are exponentially larger than the potentially optimal bound of $d^{\mathrm{poly}(k)}$ (that matches the SQ lower bounds in [DKS17]). Progress in obtaining clustering algorithms for non-spherical mixtures is a key component in the recent resolution of the problem of robust learning of mixtures of arbitrary Gaussians [LM21, BDJ+20].

By combining our list-decodable covariance estimation algorithm (here, our algorithm running in the strong contamination model of list-decodable learning is important) with a clustering algorithm with known approximate parameters (based on the partial clustering framework of [BK20b]) and a verification subroutine from [BK20b], we obtain the following improved algorithm on three fronts: 1) the algorithm applies to arbitrary weighted mixtures of Gaussians, 2) handles as large as $\varepsilon \leqslant O(p_{min}/k)$ fraction outliers (note that $\varepsilon \ll p_{min}$ is information theoretically necessary, and 3) needs sample and running time scaling as $d^{\mathrm{poly}(1/p_{min})} - d^{\mathrm{poly}(k)}$ for the equiweighted case. We present a detailed proof sketch in Section 9.3.

Corollary 1.9 (Improved Algorithms for Clustering Non-Spherical Mixtures, See Theorem 9.3)**.** *Let $d, k \in \mathbb{N}$ and $\varepsilon \ll O(p_{min}/k)$. For any $\eta > 0$, there is an algorithm that takes input an ε-corrupted sample $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{Q}^d$ drawn from $\sum_i p_i \mathcal{N}(\mu_i, \Sigma_i)$ for $p_i \geqslant p_{min}$ for each i and with probability at least 0.99, outputs estimates $\hat{\mu}_i, \hat{\Sigma}_i$ such that $(\hat{\mu}_i, \hat{\Sigma}_i)$ are $O(k\varepsilon)$-close to μ_i, Σ_i for each i. The algorithm needs $n \geqslant n_0 = d^{\mathrm{poly}(k)}/\varepsilon^2$ samples and runs in time $n^{\mathrm{poly}(k/\eta)}$.*

1.2 Comparison with Related Work

Clustering Well-Separated Non-Spherical Mixtures List-decodable (mean and) covariance estimation significantly generalizes the problem of *robust clustering* of non-spherical mixture models. In the *robust clustering* problem, the input is an $1 - \alpha = \varepsilon$-corruption (Definition 1.1) of an i.i.d. sample from a mixture of k distributions (say Gaussians). By viewing any cluster as "inliers" and all the other points as "outliers", this corresponds to the setting of $\alpha = 1/2k$ if $\varepsilon \leqslant 1/2k$. Recent works [BK20b, DHKK20] gave an $n^{\mathrm{poly}(k)}$ time ($n^{f(k)}$ time in [DHKK20] where $f(k)$ is a polynomial size tower of exponentials in k) algorithm for clustering equiweighted *non-spherical* mixtures of k Gaussians by relying on the new tools of certifiable anti-concentration [KKK19, RY19] and certifiable hypercontractivity of degree 2 polynomials.

Such clustering algorithms need two crucial assumptions: 1) every pair of the k components is separated in total variation distance by $1 - \exp(-k^{O(k)})$ and 2) the fraction of outliers $\varepsilon \ll k^{-O(k)}$. In contrast, even when specialized to clustering, our algorithm for list-decodable covariance estimation succeeds (and gives total variation distance guarantee) without any separation assumptions and handles as large as $1 - 1/k$ fraction outliers – enough to obliterate all but one clusters. Indeed, this *agnostic* clustering application was the main motivation in the initial work of Balcan, Blum and Vempala [BBV08] that defined and studied the list-decodable learning model.

The significantly more general list-decodable setting makes the approach in [BK20b, DHKK20] inapplicable. Let us briefly explain why: the key idea in [BK20b, DHKK20] is to give a *sum-of-squares proof* that if w indicates a subset C of the input points that satisfy some "Gaussian-like" properties, then C cannot simultaneously have a large intersection with two different components. This fact crucially needs (even for its information-theoretic truth) that every pair of components is pairwise well-separated. Indeed, the rounding algorithm in [BK20b, DHKK20] comes up with an approximation to the ground-truth clustering of the input points – a goal that is not meaningful in the setting of list-decodable covariance estimation.

Learning Arbitrary Gaussian Mixtures Our work is related (but incomparable and complementary, in both results and techniques) to the recent resolution of the problem of *robust learning* of a mixture of k-arbitrary Gaussians [LM21, BDJ+20]. When viewed from our vantage point, these works give a polynomial time algorithm (for any fixed k) to learn the parameters of a mixture of k Gaussians given an ε-corrupted input sample. The algorithms of [LM21, BDJ+20] do not need strong separation assumptions but crucially need that the fraction of outliers is small (at most $\sim \exp(-k!)$). In that setting, their algorithm recovers estimates of the components that are close (within some $\varepsilon^{F(k)}$ in [BDJ+20]) to those of the unknown mixture. On the other hand, our algorithm for list-decodable covariance estimation must handle an overwhelming $1 - \alpha \sim 1 - 1/k$-fraction outliers and list-decodes to an error guarantee of $1 - \theta_\alpha$ in total variation distance where θ_α is a function only of α and bounded away from 0 for all $\alpha > 0$. This is essentially the best possible guarantee in our setting as it is statistically impossible to obtain a total variation error $< 1 - \alpha$.

Indeed, our techniques are significantly (and necessarily so) different from those in [LM21, BDJ+20]. In fact, the algorithms in [LM21, BDJ+20] use robust clustering algorithms from [BK20b, DHKK20] as a first step with their key new algorithmic components coming *after* the clustering step. We note that using our new list-decodable covariance estimation algorithm in lieu of the clustering algorithm in the first step both simplifies and speeds up that step in their proof.

2 Technical Overview

In this section, we give a high-level overview of our algorithm and the main ideas that go into improving on the approaches from prior works.

Let $X \subseteq \mathbb{R}^d$ be an i.i.d. sample from $\mathcal{N}(\mu_*, \Sigma_*)$. Let $Y \subseteq \mathbb{R}^d$ be obtained by taking any $(1-\alpha)$-corruption (i.e. corrupting an $1-\alpha$ fraction of the points) of X. The goal of our algorithm is to take input any such Y and come up with a list of candidate hypotheses $(\hat{\mu}_i, \hat{\Sigma}_i)$ of size some fixed dimension-independent constant, such that there an i satisfying $d_{\mathsf{TV}}(\mathcal{N}(\mu_*, \Sigma_*), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) < 1 - \theta_\alpha$ where θ_α is a function only of α bounded away from 0 for all $\alpha > 0$. By Fact 3.24, it is enough to obtain a list of parameters that contains $(\hat{\mu}, \hat{\Sigma})$ satisfying $\mathsf{parameter\text{-}distance} \leqslant \mathrm{poly}(1/\alpha)$ (see Definition 1.3).

In this overview, we will assume that μ_* is 0. This is essentially without the loss of any generality. If X is an i.i.d. sample from $\mathcal{N}(\mu_* \Sigma_*)$, then, $\frac{x - x'}{\sqrt{2}}$ is distributed as $\mathcal{N}(0, \Sigma_*)$. Thus, if we start by taking a random matching of Y and applying the scaled difference transform above for pairs in the matching, we can simulate access to a $(1 - \alpha^2)$-corrupted sample from $\mathcal{N}(0, \Sigma_*)$. We will further restrict attention to obtaining mutiplicative spectral guarantee in our estimate – the key component of our algorithm that requires the introduction of several new ideas.

The standard approach for list-decodable learning Let's start with the approach in prior works [KKK19, RY19, BK21, RY20b] on list-decodable linear regression and subspace recovery. The algorithms in both those works find and round a solution to the sum-of-squares relaxation of a system of polynomial constraints (see Section 4 for the system we use) that encode the task of finding a subset, say C, of Y of size αn (indicated by 0-1 variables $w_1, w_2, \ldots, w_n$) that satisfies two relevant properties of Gaussian distributions: anti-concentration and hypercontractivity. In order to impose such properties as constraints, we use the standard (see discussion on succinct representation of constraints in Chapter 4 of [FKP19]) technique (from [KS17b, HL17]) of constraint compression by relying on sum-of-squares proofs.

Analysis by relating total variation distance to parameter distance To understand the main idea in their analyses, consider a "real world" solution to the constraint system. Such a solution is simply a subset C of Y of size αn. The conceptual crux of the algorithm in [KKK19, RY19, BK21, RY20b] is the following observation: If $|C \cap X| \geqslant \alpha |Y \cap X|$ – i.e., the set C intersects the "inlier" part of Y that comes from the X in α fraction of its points – then, the empirical parameters of C must be close to that of X. This is a basic statement in statistics that relates a non-trivial bound on the total variation distance (which corresponds to intersection when specialized to uniform distributions on two sets of points) to the closeness of the corresponding parameters. In fact, their analyses can be directly used to infer the following purely information-theoretic result: *if the uniform distributions on C and X are both anti-concentrated and have hypercontractive degree 2 polynomials, and $|C \cap X| \geqslant \alpha|C|$, then, the covariance of C multiplicatively approximates the covariance of X.* The anti-concentration property implies that if $C' \subseteq C$ is a subset of arbitrarily small but fixed constant fraction of C, then, the "variance" $\mathbb{E}_{x\sim C'}\langle x,v\rangle^2$ in any direction v on the subset C' must be within a constant factor of $\mathbb{E}_{x\sim C}\langle x,v\rangle^2$ – the variance in the same direction on the whole subset C. Such a property can be used to show *statistical identifiability* of a small list – we direct the reader to the technical overview sections of [KKK19, BK21, BK20b] that provide an essentially complete proof of such a statistical identifiability result with related discussions. While this is exactly the statement we want, such an argument does not yield an efficient algorithm [8].

Formalizing identifiability in low-degree sum-of-squares proof system In order to obtain efficient algorithms, the works above formalize the above information-theoretic reasoning into the *low-degree sum-of-squares proof system*.

In order to work in the low-degree sum-of-squares proof system, we need to work with sum-of-squares certificates for anti-concentration and hypercontractivity inequalities. Informally speaking, this strategy involves creating Boolean indicator variables $w_1, w_2, \ldots, w_n$ that identify a subset of the input corrupted sample Y of size n and force that the subset of points indicated by w admit SoS certificates of hypercontractivity and anti-concentration (i.e., satisfy the two relevant properties of Gaussian distributions that an "uncorrupted" part of Y is promised to satisfy). Now, notice that there can be multiple solutions to this relaxation even for the unrelaxed polynomial formulation since Y could be a disjoint union of $1/\alpha$ different subsets of size about αn such that each of these subsets provide a feasible assignment for ws. The solution to the relaxation yields a "pseudo-distribution" – for the sake of exposition in this section, the reader can think of a pseudo-distribution as a probability distribution supported on ws that describes subsets of Y that satisfy the two relevant properties of Gaussians we imposed as constraints.

We must now give a rounding algorithm to take such a pseudo-distribution and produce a small list of parameters, one of which is close to the ground truth. For this goal, we might want to replicate the above information-theoretic strategy and argue that the parameters of ws in the pseudo-distribution must be close to that of X. Such a statement would of course require that the *on average*, a subset C indicated by ws in the support of the pseudo-distribution intersects substantially in X (as otherwise, there's no reason for C to have any information about parameters of X). Such a statement does not generically hold for all pseudo-distributions (since we can, in general, have solutions ws that are entirely supported on the "outlier part" in Y). But prior works [KKK19, RY19] show that certain "spread-out-ness" constraints (formulated as minimizing

[8] This, by itself, is not surprising. Statistical identifiability in parameter estimation is often significantly simpler to establish than the task of finding efficient algorithms.

surrogates for entropy of pseudo-distributions) imply that on average, C indicated by w in the support of the pseudo-distribution does intersect in about α fraction of its points with X.

At this point, we might naturally want to replicate the information-theoretic strategy above that infers closeness of parameters of C and X from large intersection between them. But this creates a major technical difficulty in prior works that while tackled with some effort in special cases with weaker notions of error, prevents applications to general covariances. Let us explain this issue a little more:

A concrete way to analyze the pseudo-distribution is to work the following variables w that allow capturing the intersection of ws in the support of the pseudo-distribution with the unknown, uncorrupted $X \cap Y \subseteq Y$. Let $w'_i = w_i \cdot \mathbf{1}(x_i = y_i)$. Then, notice that w'_i is the indicator of indices of points in the intersection $C \cap X$. Following on the information-theoretic strategy above, we'd like to argue that the variance of points indicated by w' in any direction v is multiplicatively close to that of X in the same direction. Instead of "real-world" anti-concentration, this time, we must use a low-degree sum-of-squares certificate for anti-concentration only. The low-degree sum-of-squares certificate for anti-concentration from [KKK19, RY19] allows us to obtain a claim of the following form in degree $O(1/\delta^2)$ (which translates into a running time of $n^{O(1/\delta^2)}$.

$$\frac{1}{n}\sum_i w'_i \langle x_i, v\rangle^2 \geqslant \delta^2 \left(\frac{1}{n}\sum_i w'_i - O(\delta)\right) \frac{1}{n}\sum_i \langle x_i, v\rangle^2 \,. \tag{2.1}$$

Informally speaking, the LHS counts the contribution to the variance in the direction v from the points in $C \cap X$. The RHS, on the other hand, is a scaling of the variance of X in the direction v with the major difference from the real world version is the presence of the additive $-O(\delta)$ slack in the right hand side in the multiplier to the variance.

2.1 Key Bottleneck: Exponential Dependence on Condition Number

The expression above reveals a "gap" between low-degree certificates for anti-concentration inequality vs "real world" anti-concentration: the guarantee above is meaningful only when $\frac{1}{n}\sum_i w'_i \gg \delta$. Further, the sum-of-squares degree dependence of $O(1/\delta^2)$ for such a certificate happens to be *tight*[9] – translating into a running time cost of $n^{O(1/\delta^2)}$. This might appear innocuous – after all, if w was indeed an indicator of $Y \cap X$, the associated $\frac{1}{n}\sum_i w'_i \geqslant \alpha$ so simply choosing $\delta \ll \alpha$ should work. But this is misleading. If we were to analyze, for example, the average mean and covariance under the (pseudo-) distribution, we need that $\frac{1}{n}\sum_i w'_i \geqslant \delta$ to hold *pointwise* in the support of the distribution. But this is of course not enforceable as a constraint. In "real world", we could analyze the distribution by going over all ws in the support of it and splitting into two cases depending on whether a given w satisfies the above large intersection condition. But such an argument involves an if-then statement that is hard to formulate as a low-degree sum-of-squares proof (indeed, a version of this argument is precisely what is used in [BK20b] but its success strongly

[9] The certificates rely on the univariate polynomial approximators for indicator functions of δ-length interval around 0 over standard Gaussian distributions. Such a polynomial can be shown to need degree $O(1/\delta^2)$ by standard techniques in approximation theory. See the recent talk for a research direction on potentially stronger certificates that could escape such lower bounds.

relies on Y being a sample from a Mixture of Gaussians with only a small fraction of outliers) and the source of all the trouble in the list-decodable setting. We remark that this issue of "pointwise" facts provable in low-degree sum-of-squares proof system also arises in the recent work [KMZ22] on using the sum-of-squares relaxation for robust moment estimation to obtain optimal error guarantees for Gaussian distributions.

In order to get around this issue of additive slack and obtain a meaningful anti-concentration inequality from (2.1), we need an *upper-bound* on $\frac{1}{n}\sum_i \langle x_i, v\rangle^2$. For example, if we knew and encoded into our constraint system that the unknown covariance Σ_* has all its eigenvalues at most K, then, we can conclude from (2.1) that:

$$\frac{1}{n}\sum_i w_i \langle x_i, v\rangle^2 + O(K\delta)\,\|v\|_2^2 \geqslant \delta^2 \left(\frac{1}{n}\sum_i w_i\right) \frac{1}{n}\sum_i \langle x_i, v\rangle^2 \,. \tag{2.2}$$

That is, we must incur an additive *slack* in the anti-concentration inequality that scales with the *largest eigenvalue* of Σ_*. Observe that in order for this guarantee to be meaningful, we'd have to choose $\delta \ll 1/\sqrt{K}$ and thus our running time must scale *exponentially in the condition number* Σ_*.

This is the key reason for the weaker error guarantees in the prior results on list-decodable learning, and in fact prevents any meaningful guarantee (even of the sort known for regression and subspace recovery) without a known bound on the condition number of Σ_*. In fact, approaches based on prior works appear to fail even if the unknown covariance matrix has eigenvalues of two different scales, such as $I + (\log d)uu^\top - vv^\top$ for pairwise orthonormal unit vectors u, v.

This appears to be a fundamental issue in using anti-concentration inequalities within the SoS framework. For e.g., in their initial version of the algorithm for clustering non-spherical mixtures [BK20b], the authors made assumptions on the condition number of the covariances (with running time growing exponentially in $\log \kappa$) of the components in order to obtain their guarantees with substantial effort invested into getting around the issue of additive slack in certificates of anticoncentration [10]. They later managed to find an iterative "bootstrapping" technique that crucially relied on the strong separation assumptions available in the clustering setting in order to get an assumption-free robust clustering algorithm.

In the list-decodable setting, there are no such assumptions to work with and as a result the follow-up work [BK21] on list-decodable subspace recovery only obtains the weaker error guarantees discussed before. Our key innovation is a new algorithmic strategy that circumvents the issues with certifiable anti-concentration. Our algorithm is based on iterative use of three algorithmic components (that make an essential use of the *isotropic position*) that may be useful in robust estimation in general. We explain these new components and how they fit together next.

2.2 Coarse Spectral Recovery: Theorem 4.2

Let's depart from the approach above and abandon the goal of recovering multiplicative spectral approximation to Σ_* *in one shot* as in the prior works. Instead, we will shoot for a *coarse spectral*

[10] See also the discussion on the need for *a priori upper bounds* in [DHKK20].

recovery algorithm where we obtain 1) multiplicative approximation for the large eigenvalues, and, 2) additive approximation for all small eigenvalues of Σ_*.

Of course, the algorithm does not know the scale of the eigenvalues of Σ_*. So what does "large" mean? We will say that a quadratic form on v of Σ_* is large if it is at least $\text{poly}(\alpha)$ relative to that of the empirical covariance of the input *corrupted* sample. That is, $v^\top \Sigma_* v \geqslant \text{poly}(\alpha) \mathbb{E}_{y \sim Y} \langle v, y \rangle^2$. Observe that Y has $(1-\alpha)n$ outliers that can be arbitrarily large and can completely drown out all eigenvalues of Σ_*. In that case, coarse spectral recovery will only achieve a vacuous guarantee and we'd have to make progress via a different route.

To obtain such a "multiplicative+additive" guarantee, we introduce a new "bootstrapping" technique that obtains low-degree sum-of-squares certificates of *Frobenius norm* error bounds *restricted to the subspace* where Σ_* has small eigenvalues. Combining with multiplicative spectral recovery bounds on large eigenvalues then yields the required guarantee. This technique requires the use of *certifiable hypercontractivity* of degree 2 polynomials (in our constraint system) in addition to certifiable anti-concentration. This is in contrast to prior works [RY20a, KKK19]) that only need certifiable anti-concentration for such a goal.

We will make "conditional progress" via this method. Specifically, we argue that if every eigenvalue of every candidate $\hat{\Sigma}$ in the list generated in the coarse spectral recovery step is large (compared to that of the corrupted sample), we show that one of them must be a good multiplicative approximation that we desire. As we argued before, the additive loss is a direct consequence of the "slack" term in certificates of anticoncentration. In order to make progress, we will rely on subroutines "outside of the SoS" system.

Naive Pruning of Y When there is a candidate $\hat{\Sigma}$ in the list recovered in coarse spectral recovery that does have small eigenvalues, we will make progress by a new pruning step. Here's the key intuition: Suppose that a candidate $\hat{\Sigma}$ has a small eigenvalue in direction v. Suppose further that $\hat{\Sigma}$ is a "good" candidate (i.e. the one that achieves the multiplicative and additive guarantee w.r.t. the unknown covariance Σ_*). Then, the variance of points in the inlier $X \cap Y$ (i.e. the intersection with the uncorrupted sample) in the direction v must be small because of X being anticoncentrated in all directions including v. Thus, if we *prune* the $y \in Y$ such that the projection in direction v is large (i.e., $\langle v, y \rangle \gg v^\top \hat{\Sigma} v$), we will remove only a small fraction of inliers $X \cap Y$. Thus, our pruning functions as a biased filter that removes mostly outliers along with a small fraction of inliers. Of course, we do not know which candidate is good a priori, so we simply run the process on every candidate and obtain a different pruned subset of Y for each.

How do we make progress in this step? Observe that we have no guarantees on Y to begin with. And for all we know, the variance of Y in direction v is large (compared to $v^\top \hat{\Sigma} v$) because of say, just a single large y. If we were to repeatedly apply coarse spectral recovery with pruning, we'd have run $O(n)$ interleavings and in each such step expand the list by a factor of $\text{poly}(1/\alpha)$ leading to an exponential run time and list-size.

In order to escape this pitfall, we must find a way to guarantee that when we do try to remove points in Y that are too large, we end up chopping off a constant fraction of Y without hurting the inliers. It turns out that this is true if the low-order empirical moments of Y are *subgaussian*.

Pruning with Subgaussianity: The Power of Mild Anti-Concentration (see Section 6) Informally, the pruning step removes too few points only if a small fraction of y contribute most of the variance in a given direction v. But such "fat-tails" cannot exist if low-order moments of Y are subgaussian. In fact, we prove that if Y has subgaussian moments of $O(1)$-degree, then, we can prune away a poly(α) fraction of points in Y while ensuring that for any good candidate $\hat{\Sigma}$ the fraction of inliers in the pruned points is small. Specifically, say Y is normalized to be isotropic: that is $\mathbb{E}_{y \sim Y}\, yy^\top = I$. Then, using the Paley-Zygmund anti-concentration inequality, if the $4th$ moment of Y in the direction v is at most Δ, then, there must be a $\Omega(1/\Delta)$ points in Y such that $|\langle v, y\rangle|$ exceeds say $1/2$. On the other hand, if $\hat{\Sigma}$ were a good candidate and additively approximates all small eigenvalues of Σ_*, then, X itself would be subgaussian and given that $v^\top \hat{\Sigma} v$ is small, the fraction of points x such that $|\langle v, x\rangle|$ is large should be tiny. Thus, if we pruned away points y where $|\langle v, y\rangle|$ is larger than, say, $1/2$, we'd have removed a poly(α) fraction of points in Y while essentially leaving the inliers untouched.

Note that in this step, we are using subgaussianity to infer a mild anti-concentration of the *entire corrupted sample*. We have strong anti-concentration guarantees for the inliers (and the original uncorrupted sample) but a priori, no guarantees for Y that includes a majority of outliers. The key gain from our splitting subroutine is making progress *without needing to certify strong anti-concentration properties* – by resorting to mild anti-concentration that can be inferred only from *moment upper bounds*. Such an anti-concentration, by itself, is not enough to do list-decodable learning, but it's enough for our splitting algorithm to succeed.

Of course, our splitting subroutine relied on the corrupted sample Y being subgaussian – which, of course, it needn't be. Let us describe how we perform a different pruning to ensure this property.

2.3 Subgaussian Restriction: Theorem 5.2

The goal of our subgaussian restriction algorithm is to take input a subset of points Y and prune away some points, if needed, to ensure low-order subgaussianity: $\mathbb{E}_{y \sim Y}\langle y, v\rangle^{2t} \leqslant (C't)^t \left(\mathbb{E}_{y \sim Y}\langle y, v\rangle^2\right)^t$ for some appropriate constant C' and $t = O(1/\alpha)$. Additionally, we must ensure that the points we prune away are overwhelmingly just the outliers.

Observe that the requirement above is *linearly invariant* and significantly stronger than the variant (accomplished as a step in list-decodable mean estimation algorithms such as [KS17a]) where we only want that $\mathbb{E}_{y \sim Y}\langle y, v\rangle^{2t} \leqslant (C't)^t \|v\|_2^2$.

In order to accomplish this goal, we give a new algorithm based on a natural sum-of-squares relaxation that maximizes $\mathbb{E}_{y \sim Y}\langle y, v\rangle^{2t}$ over vectors $v \in \mathbb{R}^d$ (after putting Y into isotropic position). Our algorithm uses the solution (if the relaxation is infeasible, we are sure that Y is already subgaussian as we'd want) to the relaxation to effectively assign to each $y \in Y$, the "weight" $\widetilde{\mathbb{E}}[\langle v, y\rangle^{2t}]$.

We then use a "reverse Markov" inequality to argue that there is a threshold Z such that the set of all $y \in Y$ such that $\widetilde{\mathbb{E}}[\langle v, y\rangle^{2t}] \geqslant Z$ must be significantly larger than what we expect for X and thus, we have found a "outlier-dominated" portion of Y that can be pruned away. We then iterate on the resulting pruned Y.

Observe that unlike our splitting algorithm above, we do not guarantee pruning away non-trivial fraction of points in Y. Instead, our idea can be summarized as saying that if the adversary added outliers so as to maintain subgaussianity, then, our basic splitting procedure above makes progress. If not, then, we must make progress in the subgaussian restriction phase.

2.4 Combining the Subroutines for a Multiplicative Spectral Guarantee: Theorem 7.2

Given the above pieces, here's how we can combine them all together: at all points, our algorithm maintains a list of candidate covariances along with "current witness sets" – these are subsets of Y obtained by applying pruning and splitting steps above assuming that the corresponding candidate was a "good" candidate. If a candidate $\hat{\Sigma}$ has all its eigenvalues large compared to that of its current witness, we can label it "final". If not, we first apply subgaussian restriction to its witness set, put the resulting Y' in isotropic position and run coarse spectral recovery. This gives a new list of candidates. If any of the candidates has a small eigenvalue, the splitting step above prunes away poly(α) fraction of the points from its witness set. On the other hand, we know that for a good candidate, we never prune away too many of the inliers. Thus, once the size of the witness set drops below, say, $\alpha n/2$, we can reject that candidate completely. Thus, for a "good" initial candidate $\hat{\Sigma}$, we must eventually end up with a "witness" set Y' such that $\hat{\Sigma}$ has none of its eigenvalues are small relative to that of the $\mathbb{E}_{y\sim Y'} y'y'^{\top}$. Altogether, this gives us an algorithm with a "recursion depth" of $1/\text{poly}(\alpha)$ and at "generation", we increase the list-size by a factor $1/\text{poly}(\alpha)$. Altogether, the algorithm thus runs in time $n^{\text{poly}(1/\alpha)}$ and produces a list of size $(1/\alpha)^{\text{poly}(1/\alpha)}$.

3 Preliminaries

Throughout this paper, we will use X to denote an i.i.d. (uncorrupted) sample of n points in $\mathbb{R}^d$ and Y to denote its $(1-\alpha)$-corruption. For any finite set S of points, we will $\mathbb{E}_{s\sim S} f(s)$ to denote the empirical average of $f(s)$ as s varies uniformly over S.

3.1 Computational Model and Numerical Inputs

Our algorithms work in the standard word RAM model[11]. We assume that the inputs are rational numbers and the running time of the algorithm is a function total bit complexity of the representation of the input. This does require a moment's thought since a draw from the standard Gaussian, for e.g., is irrational with probability 1. In this work, we will assume that we have access to a *bit-oracle* for the input irrational number that can furnish as many bits of precision as our algorithm desires. The complexity of the algorithm grows as the number of bits of precision it demands increases.

[11] Works in statistical learning theory often (and sometimes implicitly) present algorithmic guarantees in the real RAM model [BCSS98]. This model deems a certain (carefully chosen) list of operations on real numbers to be doable in a single step (this includes arithmetic operations). There are a few different choices considered in prior works for such operations, see the discussion in the paragraph titled "formally modeling real RAM algorithms" on Page 1 in [EvM20]. Depending on the choice of the list of such operations, our algorithm can be implemented in this model. However, a blackbox running-time preserving translation from algorithms that work in the real RAM model to the one in the standard word RAM model is not known (see the recent work of Erickson on a smoothed version of such a statement [EvM20]). So we choose the more direct route of working in the word RAM model in this work.

Numerical Issues in Vanilla Covariance Estimation in Total Variation Distance Following prior works on learning mixtures of Gaussians, our recovery guarantee for parameters of an unknown Gaussian distribution is naturally stated in *total variation* distance – the strongest possible notion of distance in our context. When the covariance $\Sigma_* \succeq 2^{-\operatorname{poly}(d)} I$, our analysis succeeds even on rational truncation of the inputs with essentially no change.

When Σ_* is singular, however, we need some care[12] in dealing with numerical issues. To see this, consider the basic task of estimating the covariance of an unknown Gaussian distribution $\mathcal{N}(0,\Sigma)$ on $\mathbb{R}^d$ from independent samples $x_1, x_2, \ldots, x_n$. This is a basic subroutine (that has nothing to do with robust estimation) in numerical algorithms. Standard matrix concentration results imply that for $n \gg d$, $(1 - O(d/n))\Sigma \preceq \hat{\Sigma} = \frac{1}{n}\sum_{i \leqslant n} x_i x_i^\top \preceq (1 + O(d/n))\Sigma$. Such a multiplicative guarantee in Löwner ordering is necessary to obtain any bound < 1 on the total variation distance $TV(\mathcal{N}(0,\hat{\Sigma}), \mathcal{N}(0,\Sigma))$ between the unknown Gaussian distribution and the one we estimate from samples. When implemented in the word RAM model, the samples must be truncated to rationals and as a result, the estimated $\hat{\Sigma}$ will have rational entries. It is easy to construct examples where such a procedure (and in fact any rational $\hat{\Sigma}$) must be maximally far from the true covariance ($i.e.$1) in total variation distance. For example, let $v = (\sqrt{3/5}, \sqrt{2/5}, 0, \ldots, 0)$ and let $\Sigma = I - vv^\top$. Then, for every $\hat{\Sigma}$ with rational entries, the multiplicative guarantee above fails and, in fact, $TV(\mathcal{N}(0,\hat{\Sigma}), \mathcal{N}(0,\Sigma)) = 1$.

The above example shows that it is provably impossible to output a rational $\widehat{\Sigma}$ even for the basic task of estimating covariance from i.i.d. samples if we desire a multiplicative spectral guarantee. However, this also appears somewhat pathological since the "hardness" here seems to arise entirely from issues in representing the input (without any role of the algorithm itself).

Our Resolution: In order to circumvent this issue, we will make an assumption that is arguably weakest possible: we will assume that there *exists* a matrix with rational entries that satisfies the guarantees we want. That is, thus *some* output – however hard to compute – could have satisfied the requirements of the algorithm. For covariance estimation above, this is essentially equivalent to assuming that the unknown covariance Σ is rational entries. Note that the input sample points will still have irrational entries with probability 1 and will be truncated to rationals.

In this case, it is possible to recover the multiplicative spectral guarantees for the basic covariance estimation task. It is easy to prove (see Proposition B.1) that the smallest non-zero singular value of a $d \times d$ matrix of B-bit rationals is at least $2^{-\operatorname{poly}(Bd)}$. Thus, if our estimate from the truncated samples happens to produce singular values $\ll 2^{-\operatorname{poly}(Bd)}$, we can hope to "round them down" and learn the kernel of the unknown covariance. Note of course that for this to be possible, the algorithm needs to know an *a priori* bound on the bit complexity of the unknown Σ_* as otherwise there is no way to find the right precision for input truncation. The rounding down step needs some care – we formally perform it using the lattice basis reduction algorithm of [LLL82] (see Section B).

[12] We thank Sam Hopkins and Daniel Kane for discussions on computational models for statistical learning algorithms that motivated our formalization in this section.

3.2 Sum-of-Squares Preliminaries

We refer the reader to the monograph [FKP19] and the lecture notes [BS16] for a detailed exposition of the sum-of-squares method and its usage in average-case algorithm design. A *degree-ℓ pseudo-distribution* is a finitely-supported function $D : \mathbb{R}^n \to \mathbb{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x) f(x)^2 \geqslant 0$ for every polynomial f of degree at most $\ell/2$. We define the *pseudo-expectation* of a function f on $\mathbb{R}^d$ with respect to a pseudo-distribution D, denoted $\widetilde{\mathbb{E}}_{D(x)} f(x)$, as $\widetilde{\mathbb{E}}_{D(x)} f(x) = \sum_x D(x) f(x)$.

The degree-ℓ pseudo-moment tensor of a pseudo-distribution D is the tensor $\mathbb{E}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes \ell}$ with entries corresponding to pseudo-expectations of monomials of degree at most ℓ in x. The set of all degree-ℓ moment tensors of degree d pseudo-distributions is also closed and convex.

Definition 3.1 (Constrained pseudo-distributions)**.** Let D be a degree-ℓ pseudo-distribution over $\mathbb{R}^n$. Let $\mathcal{A} = \{f_1 \geqslant 0, f_2 \geqslant 0, \ldots, f_m \geqslant 0\}$ be a system of m polynomial inequality constraints. We say that D *satisfies the system of constraints $\mathcal{A}$ at degree r* (satisfies it η-approximately, respectively), if for every $S \subseteq [m]$ and every sum-of-squares polynomial h with $\deg h + \sum_{i \in S} \max\{\deg f_i, r\}$, $\widetilde{\mathbb{E}}_D\, h \cdot \prod_{i \in S} f_i \geqslant 0$ ($\widetilde{\mathbb{E}}_D\, h \cdot \prod_{i \in S} f_i \geqslant \|h\|_2 \prod_{i \in S} \left\|f_i\right\|_2$ where $\|h\|_2$ for any polynomial h is the Euclidean norm of its coefficient vector. We say that D satisfies (similarly for approximately satisfying) $\mathcal{A}$ (without mentioning degree) if D satisfies $\mathcal{A}$ at degree r.

Basic Facts about Pseudo-Distributions.

Fact 3.2 (Hölder's Inequality for Pseudo-Distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $x \in \mathbb{R}^d$. Fix $t \in \mathbb{N}$. Then, for any degree dt pseudo-distribution $\tilde{\zeta}$, $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[f^{t-1} g] \leqslant (\widetilde{\mathbb{E}}_{\tilde{\zeta}}[f^t])^{\frac{t-1}{t}} (\widetilde{\mathbb{E}}_{\tilde{\zeta}}[g^t])^{1/t}$.*

Observe that the special case of $t = 2$ corresponds to the Cauchy-Schwarz inequality. The following idea of *reweighted* pseudo-distributions follows immediately from definitions and was first formalized and used in [BKS17]).

Fact 3.3 (Reweightings [BKS17])**.** *Let D be a pseudo-distribution of degree k satisfying a set of polynomial constraints $\mathcal{A}$ in variable x. Let p be a sum-of-squares polynomial of degree t such that $\widetilde{\mathbb{E}}[p(x)] \neq 0$. Let D' be the pseudo-distribution defined so that for any polynomial f, $\widetilde{\mathbb{E}}_{D'}[f(x)] = \widetilde{\mathbb{E}}_D[f(x)p(x)]/\widetilde{\mathbb{E}}_D[p(x)]$. Then, D' is a pseudo-distribution of degree $k - t$ satisfying $\mathcal{A}$.*

Sum-of-squares proofs A *sum-of-squares proof* that the constraints $\{f_1 \geqslant 0, \ldots, f_m \geqslant 0\}$ imply the constraint $\{g \geqslant 0\}$ consists of polynomials $(p_S)_{S \subseteq [m]}$ such that $g = \sum_{S \subseteq [m]} p_S \cdot \Pi_{i \in S} f_i$.

We say that this proof has *degree ℓ* if for every set $S \subseteq [m]$, the polynomial $p_S \Pi_{i \in S} f_i$ has degree at most ℓ and write:

$$\{f_i \geqslant 0 \mid i \leqslant r\} \underset{\ell}{\vdash} \{g \geqslant 0\}\,. \tag{3.1}$$

Fact 3.4 (Soundness)**.** *If D satisfies $\mathcal{A}$ for a degree-ℓ pseudo-distribution D and there exists a sum-of-squares proof $\mathcal{A} \underset{r'}{\vdash} \mathcal{B}$, then D satisfies $\mathcal{B}$ at degree $rr' + r'$.*

Definition 3.5 (Total bit complexity of Sum-of-Squares Proofs). Let $p_1, p_2, \ldots, p_m$ be polynomials in indeterminate x with rational coefficients. For a polynomial p with rational coefficients, we say that $\{p_i \geqslant 0\}$ derives $\{p \geqslant 0\}$ in degree k and total bit complexity B if $p = \sum_i q_i^2 + \sum_i r_i p_i$ where each q_i^2, r_i are polynomials with rational coefficients of degree at most k and $k - deg(p_i)$ for every i, and the total number number of bits required to describe all the coefficients of all the polynomials q_i, r_i, p_i is at most B.

There's an efficient separation oracle for moment tensors of pseudo-distributions that allows approximate optimization of linear functions of pseudo-moment tensors approximately satisfying constraints. The *degree-ℓ sum-of-squares algorithm* optimizes over the space of all degree-ℓ pseudo-distributions that approximately satisfy a given set of polynomial constraints:

Fact 3.6 (Efficient Optimization over Pseudo-distributions [Sho87, Par00, Nes00, Las01]). *Let $\eta > 0$. There exist an algorithm that for $n, m \in \mathbb{N}$ runs in time $(n+m)^{O(\ell)} \operatorname{poly}\log 1/\eta$, takes input an explicitly bounded and satisfiable system of m polynomial constraints $\mathcal{A}$ in n variables with rational coefficients and outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ η-approximately.*

Basic Sum-of-Squares Proofs

Fact 3.7 (Operator norm Bound). *Let A be a symmetric $d \times d$ matrix with rational entries with numerators and denominators upper-bounded by 2^B and v be a vector in $\mathbb{R}^d$. Then, for every $\varepsilon \geqslant 0$,*

$$\vdash_{2}^{v} \left\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 + \varepsilon \right\}$$

The total bit complexity of the proof is $\operatorname{poly}(B, d, \log 1/\varepsilon)$.

Fact 3.8 (SoS Hölder's Inequality). *Let f_i, g_i for $1 \leqslant i \leqslant s$ be indeterminates. Let p be an even positive integer. Then,*

$$\vdash_{p^2}^{f,g} \left\{ \left(\frac{1}{s} \sum_{i=1}^{s} f_i g_i^{p-1} \right)^p \leqslant \left(\frac{1}{s} \sum_{i=1}^{s} f_i^p \right) \left(\frac{1}{s} \sum_{i=1}^{s} g_i^p \right)^{p-1} \right\} .$$

The total bit complexity of the sos proof is $s^{O(p)}$.

Observe that using $p = 2$ yields the SoS Cauchy-Schwarz inequality.

Fact 3.9 (SoS Almost Triangle Inequality). *Let $f_1, f_2, \ldots, f_r$ be indeterminates. Then,*

$$\vdash_{2t}^{f_1, f_2, \ldots, f_r} \left\{ \left(\sum_{i \leqslant r} f_i \right)^{2t} \leqslant r^{2t-1} \left(\sum_{i=1}^{r} f_i^{2t} \right) \right\} .$$

The total bit complexity of the sos proof is $r^{O(t)}$.

Fact 3.10 (SoS AM-GM Inequality, see Appendix A of [BKS15]). *Let $f_1, f_2, \ldots, f_m$ be indeterminates. Then,*

$$\left\{ f_i \geqslant 0 \mid i \leqslant m \right\} \vdash_{m}^{f_1, f_2, \ldots, f_m} \left\{ \left(\frac{1}{m} \sum_{i=1}^{m} f_i \right)^m \geqslant \Pi_{i \leqslant m} f_i \right\} .$$

The total bit complexity of the sos proof is $\exp(O(m))$.

Fact 3.11 (Univariate SoS Proofs). *Let $p = \sum_{i \leqslant k} \alpha_i x^i$ be a univariate polynomial of degree k with rational coefficients α_i with numerators and denominators upper-bounded by 2^B for some $B \in \mathbb{N}$. For every $\varepsilon \geqslant 0$: $\vdash^{x}_{d} \left\{ p(x) \geqslant 0 \right\}$ and the total bit complexity of the SoS proof is upper-bounded by $\mathrm{poly}(B, \log 1/\varepsilon)$.*

Fact 3.12 (Cancellation within SoS, Constant RHS [BK20b]). *Suppose A is indeterminate and $t \geqslant 1$. Then,*

$$\left\{ A^{2t} \leqslant 1 \right\} \vdash^{A}_{2t} \left\{ A^2 \leqslant 1 \right\}$$

Further, the total bit complexity of the SoS proof is at most $2^{O(t)}$.

Lemma 3.13 (Cancellation within SoS [BK20b]). *Suppose A and C are indeterminates and $t \geqslant 1$. Then,*

$$\left\{ A \geqslant 0 \cup A^t \leqslant C A^{t-1} \right\} \vdash^{A,C}_{2t} \left\{ A^{2t} \leqslant C^{2t} \right\} .$$

Further, the total bit complexity of the SoS proof is at most $2^{O(t)}$.

Fact 3.14 (Frobenius-Operator Norm Bounds in SoS [BK20b]). *Suppose $A \in \mathbb{Q}^{d \times d}$ have entries of bit complexity at most B. Let Q be a $d \times d$ matrix valued indeterminate. Then*

$$\vdash^{Q}_{2} \left\{ \|AQ\|_F^2 \leqslant \left\| A^\top A \right\|_{op} \|Q\|_F^2 \right\}$$

The total bit complexity of the SoS proof is at most $O(B^2 d^2)$.

Fact 3.15 (Contraction and Frobenius Norms, Lemma 9.1 in [BK20b]). *Let A, B be $d \times d$ matrix-valued indeterminates. Let β be a scalar-valued indeterminate. Then,*

$$\left\{ \beta (v^\top A^\top A v)^t \leqslant \Delta \|v\|_2^{2t} \right\} \vdash \left\{ \beta \|AQ\|_F^{2t} \leqslant \Delta t^t \|Q\|_F^{2t} \right\} ,$$

and,

$$\left\{ \beta (v^\top A^\top A v)^t \leqslant \Delta \|v\|_2^{2t} \right\} \vdash \left\{ \beta \|QA\|_F^{2t} \leqslant \Delta t^t \|Q\|_F^{2t} \right\} .$$

Fact 3.16 (See Lemma A.5 in [KS17b]). *Let $\mathcal{A}$ be a set of polynomial equality axioms in variable x such that:*

$$\mathcal{A} \vdash^{2t}_{x,u} \left\{ p(x,u) \geqslant 0 \right\} ,$$

for a polynomial p with total degree at most $2t$. Then, for any pseudo-distribution D of degree $2t$ on x satisfying $\mathcal{A}$,

$$\vdash^{2t}_{u} \left\{ \widetilde{\mathbb{E}}_{D(x)} p(x,u) \geqslant 0 \right\} .$$

3.3 Analytic Properties of Probability Distributions

Certifiable Anti-Concentration

Fact 3.17 (Univariate Approximator to Interval Indicator (see Lemma A.1 in [KKK19])). *For each $\delta > 0$, there is a univariate polynomial p_δ and a sum-of-squares polynomial S_δ of degrees $\leqslant s(\delta) = O(1/\delta^2)$ both with rational coefficients with numerators and denominators upper bounded by $2^{O(s(\delta))}$ satisfying:*

1. *$p_\delta(x) = p_\delta(-x)$ for every x. Thus the non-zero coefficients of p_δ are on even-power monomials in x.*

2. *$p_\delta(x) \geqslant 1/2$ for all x such that $|x| \leqslant \delta$.*

3. *$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\, p_\delta^2(x) + S_\delta(x) = C\delta$ for an absolute constant $C > 0$.*

Corollary 3.18. *Let $\delta > 0$, and $x \in \mathbb{R}^d$. Let R, Σ be $d \times d$ symmetric matrix-valued indeterminates. Let $q_{\delta,\Sigma}(x, v)$ be the following polynomial in d-dimensional vector valued indeterminate v (parameterized by x).*

$$q_{\delta,\Sigma}(x,v) = (v^\top \Sigma v)^{s(\delta)/2} p_\delta(\frac{\langle x, v\rangle}{\sqrt{v^\top \Sigma v}})$$

Then, $q_{\delta,\Sigma}(x, v)$ satisfies:

1. $\Big|\frac{\Sigma,v}{2s(\delta)}\ \left\{\langle x,v\rangle^2 (v^\top \Sigma v)^{s(\delta)-1} + \delta^2 q_{\delta,\Sigma}(x,v)^2 - \delta^2 (v^\top\Sigma v)^{s(\delta)} = SoS(v,R)\right\}$.

2. $\left\{R^2 = \Sigma\right\} \Big|\frac{v}{2s(\delta)}\ \left\{\mathbb{E}_{x\sim\mathbb{N}(0,I)}\, q_{\delta,\Sigma}(Rx,v)^2 \leqslant C\delta(v^\top\Sigma v)^{s(\delta)}\right\}$.

Here, $SoS(v, R)$ denotes a sum-of-squares polynomial in indeterminates v and R. Further, the total bit complexity of both the SoS proofs above is at most $d^{O(s(\delta))}$.

We provide a proof of the above corollary for completeness in Section A of the Appendix.

Definition 3.19 (Certifiable Anti-Concentration)**.** A distribution D on $\mathbb{R}^d$ with mean 0 and covariance Σ_* of rational entries with numerator and denominators upper-bounded by 2^B is said to be $s(\delta)$-certifiably (C, δ)-anti-concentrated if for q_{δ,Σ_*} defined in Corollary 3.18 satisfies:

1. $\Big|\frac{v}{4s}\ \left\{\langle x,v\rangle^2 (v^\top \Sigma_* v)^{s(\delta)-1} + \delta^2 q_{\delta,\Sigma_*}(x,v)^2 - \delta^2 (v^\top\Sigma v)^{s(\delta)} = SoS(v,\Pi)\right\}$,

2. $\Big|\frac{v}{4s}\ \left\{\mathbb{E}_{x\sim\mathbb{N}(0,I)}\, q_{\delta,\Sigma_*}(\Pi x,v)^2 \leqslant C\delta(v^\top\Sigma_* v)^{s(\delta)}\right\}$, and

the total bit complexity of each of two SoS proofs above is at most poly$(B, s(\delta))$. A set $X \subseteq \mathbb{R}^d$ is said to be $s(\delta)$-certifiably (C, δ)-anti-concentrated if the uniform distribution on X is $s(\delta)$-certifiably (C, δ)-anti-concentrated.

Fact 3.20 (Certifiable Anti-concentration of Gaussians and Spherically Symmetric Distributions, Theorem 6.2 in [BK21])**.** *Gaussian distribution (with arbitrary covariances) and more generally, affine transforms of any spherically symmetric random variable H on $\mathbb{R}^d$ with sub-exponentially distributed $\|H\|_2^2$ is $s(\delta)$-certifiably (C, δ)-anti-concentrated for $s(\delta) \leqslant O(1/\delta^2)$ and $C = O(1)$.*

Certifiable Hypercontractivity of Degree 2 Polynomials Next, we define *certifiable hypercontractivity* of degree-2 polynomials that formulates (within SoS) the fact that higher moments of degree-2 polynomials of distributions (such as Gaussians) can be bounded in terms of appropriate powers of their 2nd moment.

Definition 3.21 (Certifiable Hypercontractivity)**.** An isotropic distribution $\mathcal{D}$ on $\mathbb{R}^d$ is said to be h-certifiably C-hypercontractive if there is a degree-h sum-of-squares proof of the following unconstrained polynomial inequality in $d \times d$ matrix-valued indeterminate Q:

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} (x^\top Q x)^h \leqslant (Ch)^h \left(\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left((x^\top Q x) - \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [x^\top Q x] \right)^2 \right)^{h/2} .$$

A set of points $X \subseteq \mathbb{R}^d$ is said to be C-certifiably hypercontractive if the uniform distribution on X is h-certifiably C-hypercontractive.

Remark 3.22. Certifiable hypercontractivity is sometimes also defined (such as in [BK20b]) with the RHS above being $h/2$-th power of the 2nd moment of $x^\top Q x$ instead of variance as in the above definition. In that case, an additional property (called "certifiable bounded variance") is needed to obtain the statement in terms of the variance on the RHS above. We choose the simpler formulation with the RHS above directly stated in terms of the variance of $x^\top Q x$.

Observe that the definition above is invariant under linear transforms of the the random variable x. It can also be shown to be invariant under affine transforms of x (see Lemma 2.3 in [BDJ+20]). Hypercontractivity is an important notion in high-dimensional probability and analysis on product spaces [O'D14]. Kauers, O'Donnell, Tan and Zhou [KOTZ14] showed certifiable hypercontractivity of Gaussians and more generally product distributions with subgaussian marginals. Certifiable hypercontractivity strictly generalizes the better known *certifiable subgaussianity* property (formalized and studied first in [KS17b]) that is the special case of certifiable hypercontractivity of (squares of) linear polynomials, or, equivalently, when $Q = vv^\top$ for a vector-valued indeterminate v.

Analytic Properties Under Sampling The following lemma can be proven via similar, standard techniques as in several prior works [KKK19, BK21, RY19, RY20b].

Fact 3.23 (Certifiable Anti-concentration and Hypercontractivity Under Sampling (see for e.g. Section 8 in [BK20b]))**.** *Let D be a $s(\delta)$-certifiably (C, δ)-anti-concentrated distribution with mean μ_* and covariance Σ_* with B bit rational entries and $2t$-certifiably C-hypercontractive degree 2 polynomials on $\mathbb{R}^d$ for every $t \in \mathbb{N}$. Let X be an i.i.d. sample from D with $n \geqslant n_0 = O(d^{s(\delta)})$ and let $\tilde{X}$ be obtained by truncating each entry of each $x \in X$ to a rational number of* $\mathrm{poly}(Bd)$ *bits. Then, with probability at least $1 - 1/d$ over the draw of X, 1) X and $\tilde{X}$ are $s(\delta)$-certifiably $(2C, \delta)$-anti-concentrated with $s(\delta)$-certifiably $2C$-hypercontractive degree 2 polynomials, 2) $(\mathbb{E}_{x \sim X}\, x - \mu_*)(\mathbb{E}_{x \sim X}\, x - \mu_*)^\top \preceq 0.01\Sigma_*$, 3) $\mathbb{E}_{x \sim X}(x - \mu_*)(x - \mu_*)^\top \in [0.99, 1.01]\Sigma_*$ and 4)$\left\| \Sigma_*^{\dagger/2}\, \mathbb{E}_{x \sim x}(x - \mu_*)(x - \mu_*)^\top \Sigma_*^{\dagger/2} \right\|_F \leqslant 0.1$.*

Further, if all entries of Σ_* are B-bit rational numbers, then, all the above facts are true for $\mathrm{poly}(d)$*-bit precision truncations of points in an i.i.d. sample X.*

Total Variation vs Parameter Distance for Gaussians The total variation distance (a.k.a. statistical distance) between any two probability density functions p, q on $\mathbb{R}^d$ is defined by $d_{\mathsf{TV}}(p, q) = \frac{1}{2} \int |p(x) - q(x)| dx$. Then, $0 \leqslant d_{\mathsf{TV}}(p, q) \leqslant 1$ for all probability density functions p, q.

The following fact relates the total variation distance between a pair of Gaussians and an appropriate notion of distance between their parameters. A relationship of this form was recently proved by [DMR18] but their bounds are only meaningful in the regime where the total variation distance is at most some absolute constant $\ll 1$. Instead, we use the following result established in the recent works on clustering [BK20b, DHKK20] that gives a meaningful parameter distance translation in the regime where the total variation distance is close to but bounded away from 1.

Fact 3.24 (TV vs Parameter Distance for Gaussians, see Prop. A.1 in [BK20b])**.** *Fix $\Delta > 0$ and let μ, μ' and $\Sigma, \Sigma' \succ 0$ satisfy:*

1. ***Mean Closeness:*** *for all $v \in \mathbb{R}^d$, $\langle \mu - \mu', v \rangle_2^2 \leqslant \Delta^2 v^\top (\Sigma + \Sigma') v$.*
2. ***Spectral Closeness:*** *for all $v \in \mathbb{R}^d$ $\frac{1}{\Delta^2} v^\top \Sigma v \leqslant v^\top \Sigma' v \leqslant \Delta^2 v^\top \Sigma(r') v$.*
3. ***Relative Frobenius Closeness:*** $\left\| \Sigma^{\dagger/2} \Sigma' \Sigma^{\dagger/2} - I \right\|_F^2 \leqslant \Delta^2 \cdot \left\| \Sigma^\dagger \Sigma' \right\|_2^2$.

Then, $d_{\mathsf{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')) \leqslant 1 - \exp(-O(\Delta^2 \log \Delta))$.

4 Coarse Spectral Recovery

In this section, we present the first component of our algorithm for list-decodable covariance estimation. This subroutine produces a list of candidate covariances that includes a candidate that multiplicatively approximates the spectrum of the unknown Σ_* *when restricted to the subspace with sufficiently large eigenvalues* (compared to that of the corrupted sample) while giving an additive error guarantee on all small eigenvectors of Σ_*.

Formally, our algorithm succeeds whenever we are given an $(1 - \alpha)$-corruption Y of a *good* set X of points that we define next. Recall that for any finite set X, we use the notation $\mathbb{E}_{x \sim X}$ to mean average over uniform draw of x from X.

Definition 4.1 (Good Set)**.** For $d \in \mathbb{N}$, we say that a subset $X \subseteq \mathbb{Q}^d$ is a (C, δ)-good set with mean $\mathbb{E}_{x \sim X}\, x = \mu_*$ and 2nd moment $\mathbb{E}_{x \sim X}\, x x^\top = \Sigma_*$[13] if X satisfies the following for $s(\delta) = O(1/\delta^2)$:

1. **Small Mean:** $\mu_* \mu_*^\top \preceq 0.1\, \mathbb{E}_{x \sim X} (x - \mu_*)(x - \mu_*)^\top$.
2. **Anti-Concentration:** For $\beta \geqslant \delta$, $v \in \mathbb{R}^d$, $\mathbb{P}_{x \sim X}[\langle x, v \rangle^2 \leqslant \frac{\beta}{2} v^\top \Sigma_* v] \leqslant \beta$.
3. **Certifiable Anti-Concentration:** X is s-certifiably (C, δ)-anti-concentrated for $s = s(\delta)$.
4. **Hypercontractivity:** X has $2s(\delta)$-certifiably C-hypercontractive degree 2 polynomials.

The following theorem is the main result of this section.

[13] In our application of this subroutine, we can ensure that X is a sample from a mean 0 distribution. We invite the reader to think of the mean of X to be exactly 0 in a first reading. In this case, the 2nd moment of X is the covariance of X. We continue to use the same notation for covariance and 2nd moment as in the small mean case (that is satisfied whp by a large enough random sample from a zero-mean distribution) the 2nd moment spectrally approximates the covariance within a factor of 1.01 which is enough for our guarantees for covariance recovery.

Theorem 4.2 (Coarse Spectral Recovery)**.** *Let $1 \geqslant \alpha, \nu > 0$. For every $t \in \mathbb{N}$, there is an algorithm that takes input a collection of n points $Y \subseteq \mathbb{Q}^d$ such that $\frac{1}{n}\sum_i y_i y_i^\top = (1 \pm 2^{-d})I$ and outputs a list of positive semidefinite matrices $\widehat{\Sigma_1}, \widehat{\Sigma}_2, \ldots, \widehat{\Sigma}_k \in \mathbb{Q}^{d \times d}$ for $k = O(1/\alpha^{t+2}) \cdot \log(1/\nu)$ with the following guarantee:*

*For $\delta = \alpha^3/2C$, suppose there is a (C, δ)-*good *set of points $X = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{Q}^d$ satisfying $\mathbb{E}_{x \sim X}\, xx^\top = \Sigma_*$ such that $|Y \cap X| \geqslant \alpha n$. Then, with probability at least $1 - \nu$ over only the randomness of the algorithm, there is an $i \leqslant k$ such that:*

$$\Sigma_* \preceq \hat{\Sigma}_i \preceq O\left(\frac{1}{\alpha^{6t+18}}\right)\Sigma_* + O(\alpha^{2t-28})I \,. \tag{4.1}$$

For $t = 20$, this gives a list $\mathcal{L}$ of size $O(1/\alpha^{22})$ containing a candidate $\hat{\Sigma}_i$ satisfying:

$$\Sigma_* \preceq \hat{\Sigma}_i \preceq O\left(\frac{1}{\alpha^{138}}\right)\Sigma_* + O(\alpha^{12})I \,. \tag{4.2}$$

The algorithm runs in time $(Bn)^{O(1/\alpha^{12})}O(\log 1/\nu)$ where B is the bit complexity of entries of y_is.

4.1 Algorithm

Our algorithm approximately solves and rounds a sum-of-squares relaxation of an appropriate polynomial system. Our polynomial constraint system encodes finding a set of n points $X' \subseteq \mathbb{R}^{d \times n}$ such that X' (intended to be variables for X) satisfies the properties of the original sample X for some covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ (intended to be Σ_*). Our polynomial system has the following indeterminates.

1. x_i' for $1 \leqslant i \leqslant n$: d-dimensional vector valued indeterminates forming X'.
2. R, Σ, U, Z: $d \times d$ matrix-valued indeterminates. Here, Σ encodes the empirical covariance matrix of X, R stands for a matrix square root of Σ and U forces R to be positive semidefinite.
3. w_i for $1 \leqslant i \leqslant n$: scalar indeterminates encoding that Y intersects X' in $\geqslant \alpha n$ points.

We impose the following constraints on the indeterminates above (categorized for exposition).

$$\text{Covariance Constraints: } \mathcal{A}_1 = \left\{ \begin{aligned} R &= UU^\top \\ R^2 &= \Sigma \\ \left(\frac{8}{\alpha^2}I - \Sigma\right) &= ZZ^\top \end{aligned} \right\} \tag{4.3}$$

$$\text{Subset Constraints: } \mathcal{A}_2 = \left\{ \begin{array}{lr} \forall i \in [n] & w_i^2 = w_i \\ & \sum_{i \in [n]} w_i = \alpha n \\ \forall i \in [n] & w_i(y_i - x_i') = 0 \end{array} \right\} \tag{4.4}$$

$$\text{Parameter Constraints: } \mathcal{A}_3 = \left\{ \frac{1}{n}\sum_{i=1}^{n} w_i x_i' x_i'^\top = \Sigma \right\} \tag{4.5}$$

$$\text{Cert. Anti-Concentration} : \mathcal{A}_4 = \left\{ \quad \frac{1}{n}\sum_{i=1}^{n} q^2_{\delta,\Sigma}\left(x'_i, v\right) \leqslant C\delta \left(v^\top \Sigma v\right)^{s(\delta)} \quad \right\} \tag{4.6}$$

$$\text{Cert. Hypercontractivity} : \mathcal{A}_5 = \left\{ \forall j \leqslant s(\delta), \quad \frac{1}{n}\sum_{i=1}^{n} \left({x'_i}^\top Q x'_i - \frac{1}{n}\sum_{i \leqslant n} {x'_i}^\top Q x'_i \right)^{2j} \leqslant (2Cj)^{2j} \left\|RQR\right\|_F^{2j} \right\} \tag{4.7}$$

4.2 Algorithm

We next describe our algorithm.

Algorithm 4.3 (List-Decoding for Coarse Spectral Recovery)**.**

Given: $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{Q}^d$ such that $\frac{1}{n}\sum_{i=1}^{n} y_i y_i^\top = (1 \pm 2^{-d})I$ and $\alpha, \eta > 0$.

Output: A list $\mathcal{L}$ of $O(1/\alpha^{t+2})$ positive semidefinite matrices in $\mathbb{Q}^{d \times d}$ for $t = 20$.

Operation:

1. For $\delta = \alpha^3/2C$, find a pseudo-distribution $\tilde{\zeta}$ of degree $O(s(\delta) + 2t + 1)$ that approximately satisfies the constraint system $\mathcal{A}$ and minimizes $\|\tilde{\mathbb{E}}[w]\|_2$ with an error $\leqslant 2^{-(Bd)^{O(s(\delta))}}$.
2. For any multiset $S \subseteq [n]$ of size $2t + 1$ such that $\tilde{\mathbb{E}}[w_S] = \tilde{\mathbb{E}}[\Pi_{i \in S} w_i] > 0$, let $\tilde{\Sigma}_S = \frac{\tilde{\mathbb{E}}[w_S \Sigma]}{\tilde{\mathbb{E}}[w_S]}$.
3. For $O(1/\alpha^{t+2})$ times: 1) pick a multiset $S \subseteq [n]$ of size $2t + 1$ with probability proportional to $\tilde{\mathbb{E}}[w_S]$ and 2) add $O(\frac{1}{\alpha^8})\tilde{\Sigma}_S$ to $\mathcal{L}$.
4. Return $\mathcal{L}$.

4.3 Deriving Key Properties Via Low-Degree Sum-of-Squares Proofs

The goal of the next few lemmas is to derive a key consequence of our constraint system $\mathcal{A}$. Informally speaking, we show that if X' – the algorithm's "guess" for the unknown X – intersects with X non-trivially, then, the quadratic form of the empirical 2nd moments of X and X' on any vector v must be close. The closeness is quantified by an error term with a multiplicative part and an additive part.

Notation 4.1. *Let $w(X') = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot \mathbf{1}(x_i = y_i)$ be the linear polynomial in the indeterminates w_is. Note that $w(X')$ measures the fraction of points X' has in common with the (unknown) good set X.*

Lemma 4.4 (Spectral Recovery Guarantee – Lower Bound)**.** *Under the hypothesis of Theorem 4.2,*

$$\mathcal{A} \vdash_{4s}^{\Sigma, w, v, X'} \left\{ \frac{1}{\delta^2}(v^\top \Sigma v)(v^\top \Sigma_* v)^{s-1} + C\delta (v^\top \Sigma_* v)^s \geqslant w(X')(v^\top \Sigma_* v)^s \right\}.$$

Further, if entries of y_is have bit complexity $\leqslant B$, then, the bit-complexity of the SoS proof is $(Bd/\delta)^{O(st)}$.

Proof. One of the key properties of w_i is that for *any* polynomial $h(\cdot)$ of degree at most $O(s)$, it is true that

$$\mathcal{A}_2 \vdash^{w,X'}_{O(s)} \left\{ w_i \mathbf{1}(x_i = y_i) h(x_i') = w_i \mathbf{1}(x_i = y_i) h(y_i) = w_i \mathbf{1}(x_i = y_i) h(x_i) \right\}. \tag{4.8}$$

So, recalling the first certifiable anti-concentration constraint (Corollary 3.18) on x_i

$$\vdash^{v}_{O(s)} \left\{ \langle x_i, v\rangle^2 (v^\top \Sigma_* v)^{s-1} + \delta^2 q_{\delta,\Sigma_*}(x_i, v)^2 \geqslant \delta^2 (v^\top \Sigma_* v)^s \right\}$$

and applying Equation (4.8) after multiplying by $w_i \mathbf{1}(x_i = y_i)$ gives

$$\mathcal{A}_4 \cup \mathcal{A}_2 \vdash^{w,\Sigma,x_i',v}_{O(s)} \left\{ w_i \mathbf{1}(x_i = y_i) \langle x_i', v\rangle^2 (v^\top \Sigma_* v)^{s-1} + \delta^2 w_i \mathbf{1}(x_i = y_i) q_{\delta,\Sigma_*}(x_i, v)^2 \geqslant \delta^2 w_i \mathbf{1}(x_i = y_i)(v^\top \Sigma_* v)^s \right\}.$$

Using next that $\mathcal{A}_2 \vdash^{w}_{2} \left\{ w_i \mathbf{1}(x_i = y_i) \leqslant w_i \right\}$ (and that this is also at most 1) on the left hand side components and averaging over i transforms this equation to

$$\mathcal{A}_4 \cup \mathcal{A}_2 \vdash^{w,\Sigma,v,X'}_{O(s)} \left\{ \frac{1}{n} \sum_{i=1}^{n} w_i \langle x_i', v\rangle^2 (v^\top \Sigma_* v)^{s-1} + \frac{\delta^2}{n} \sum_{i=1}^{n} q_{\delta,\Sigma_*}(x_i, v)^2 \geqslant \frac{\delta^2}{n} \sum_{i=1}^{n} w_i \mathbf{1}(x_i = y_i)(v^\top \Sigma_* v)^s \right\}.$$

Now we wish to simplify each of these three terms:

- By definition of Σ, we have $\mathcal{A}_3 \vdash^{v}_{2} \left\{ \frac{1}{n} \sum_{i=1}^{n} w_i \langle x_i', v\rangle^2 = v^\top \Sigma v \right\}$.
- By certifiable anti-concentration of the true samples (Corollary 3.18) we obtain
$$\vdash^{v}_{O(s)} \left\{ \frac{1}{n} \sum_{i=1}^{n} q_{\delta,\Sigma}(x_i, v)^2 \leqslant C\delta (v^\top \Sigma_* v)^s \right\}.$$
- Finally, by definition we have $\frac{1}{n} \sum_{i=1}^{n} w_i \mathbf{1}(x_i = y_i) = w(X')$.

Putting these three facts together yields the desired conclusion:

$$\mathcal{A}_4 \cup \mathcal{A}_2 \vdash^{w,\Sigma,v,X'}_{O(s)} \left\{ \frac{1}{\delta^2} (v^\top \Sigma v)(v^\top \Sigma_* v)^{s-1} + C\delta (v^\top \Sigma_*)^s \geqslant w(X')(v^\top \Sigma_* v)^s \right\}. \tag{4.9}$$

□

Our next lemma proves an upper-bound version of the spectral guarantee. Note that

Lemma 4.5 (Spectral Recovery – Upper Bound)**.** *Under the hypothesis of Theorem 4.2, for any $t \in \mathbb{N}$,*

$$\mathcal{A}_4 \cup \mathcal{A}_2 \cup \mathcal{A}_1 \vdash^{w,X',\Sigma,v}_{O(st)} \left\{ w(X')^{2ts} (v^\top \Sigma v)^{2s} \leqslant 2^{2s} \left(\frac{1}{\delta^{4s}} (v^\top \Sigma_* v)^{2s} + \left(\frac{4C^t \delta^t}{\alpha^2} \right)^{2s} \|v\|_2^{4s} \right) \right\}. \tag{4.10}$$

Further, if entries of y_is have bit complexity $\leqslant B$, then, the bit-complexity of the SoS proof is $(Bd/\delta)^{O(st)}$.

Proof. We begin similarly to the proof of Lemma 4.4: in particular, by swapping the roles of x_i, x_i' in the proof we obtain

$$\mathcal{A}_4 \cup \mathcal{A}_2 \vdash^{w,\Sigma,v,X'}_{O(s)} \left\{ \frac{1}{\delta^2}(v^\top \Sigma_* v)(v^\top \Sigma v)^{s-1} \geqslant (w(X') - C\delta)(v^\top \Sigma v)^s \right\}. \tag{4.11}$$

Now, note that we would like $w(X')^t$ to appear. To do so, we use that

$$w(X')^t - (C\delta)^t = (w(X') - C\delta)\left(\sum_{i=0}^{t-1} w(X')^i (C\delta)^{t-1-i}\right)$$

and the fact that from $\vdash^{w}_{2} \{0 \leqslant w(X') \leqslant 1\}$, $C\delta \leqslant \frac{1}{2}$ it follows that

$$\vdash^{w}_{O(t)} \left\{ 0 \leqslant \sum_{i=0}^{t-1} w(X')^i (C\delta)^{t-1-i} \leqslant \sum_{i=0}^{t-1} (C\delta)^{t-1-i} \leqslant \frac{1}{1 - C\delta} \leqslant 2 \right\}.$$

Therefore, multiplying both sides of Equation (4.11) by $\sum_{i=0}^{t-1} w(X')^i (C\delta)^{t-1-i}$ yields

$$\begin{aligned}\mathcal{A}_4 \cup \mathcal{A}_2 \vdash^{w,\Sigma,v,X'}_{O(s+t)} \Bigg\{ \frac{2}{\delta^2}(v^\top \Sigma_* v)(v^\top \Sigma v)^{s-1} &\geqslant \left(\sum_{i=0}^{t-1} w(X')^i (C\delta)^{t-1-i}\right) \cdot \frac{1}{\delta^2}(v^\top \Sigma_* v)(v^\top \Sigma v)^{s-1} \\ &\geqslant (w(X')^t - (C\delta)^t)(v^\top \Sigma v)^s \Bigg\}\end{aligned} \tag{4.12}$$

We may now rearrange this to

$$\mathcal{A}_4 \cup \mathcal{A}_2 \vdash^{w,\Sigma,v,X'}_{O(s+t)} \left\{ \left(\frac{2}{\delta^2}(v^\top \Sigma_* v) + (C\delta)^t (v^\top \Sigma v)\right)(v^\top \Sigma v)^{s-1} \geqslant w(X')^t (v^\top \Sigma v)^s \right\}.$$

Multiplying both sides by $w(X')^{t(s-1)}$ and applying Cancellation within SoS (Lemma 3.13) with $A = w(X')^t (v^\top \Sigma v)$ and $C = \frac{2}{\delta^2}(v^\top \Sigma_* v) + (C\delta)^t (v^\top \Sigma v)$ then brings us closer to our end goal by proving

$$\mathcal{A}_4 \cup \mathcal{A}_2 \vdash^{w,\Sigma,v,X'}_{O(st)} \left\{ w(X')^{2ts}(v^\top \Sigma v)^{2s} \leqslant \left(\frac{2}{\delta^2}(v^\top \Sigma_* v) + (C\delta)^t (v^\top \Sigma v)\right)^{2s} \right\}.$$

Now, applying the SoS Almost-Triangle Inequality (Fact 3.9) on the right hand side separates these latter terms:

$$\mathcal{A}_4 \cup \mathcal{A}_2 \vdash^{w,\Sigma,v,X'}_{O(st)} \left\{ w(X')^{2ts}(v^\top \Sigma v)^{2s} \leqslant 2^{2s}\left(\frac{2^{2s}}{\delta^{4s}}(v^\top \Sigma_* v)^{2s} + (C\delta)^{2ts}(v^\top \Sigma v)^{2s}\right) \right\}.$$

To finish, recall that $\mathcal{A}_1 \vdash^{\Sigma}_{2} \left\{ \frac{8}{\alpha^2} I - \Sigma = ZZ^\top \right\}$ so in particular $v^\top \Sigma v \leqslant \frac{4}{\alpha^2}\|v\|_2^2$. Therefore, plugging this in gives the final desired bound of

$$\mathcal{A}_4 \cup \mathcal{A}_2 \cup \mathcal{A}_1 \vdash^{w,\Sigma,v,X'}_{O(st)} \left\{ w(X')^{2ts}(v^\top \Sigma v)^{2s} \leqslant 2^{2s}\left(\frac{2^{2s}}{\delta^{4s}}(v^\top \Sigma_* v)^{2s} + \left(\frac{8C^t\delta^t}{\alpha^2}\right)^{2s}\|v\|_2^{4s}\right) \right\}.$$

□

Bootstrapping Spectral Recovery via Frobenius Recovery Our spectral recovery guarantees (from previous two lemmas) is actually insufficient to ensure that a rounded candidate multiplicatively approximates *all* the large eigenvalues of the unknown Σ_*. One of the key innovations in our analysis is a "boostrapping" trick that relies on a stronger *Frobenius norm* guarantee *restricted to the subspace of small eigenvalues of* Σ_* which we prove below. We invite the reader to think of $\mathcal{P}$ as the (unknown) projector to the subspace of small eigenvalues of the (unknown) Σ_*.

Lemma 4.6 (Frobenius Recovery)**.** *Under the hypothesis of Theorem 4.2, for any projection matrix $\mathcal{P}$ to a subspace of $\mathbb{R}^d$, and for any $t \in \mathbb{N}$,*

$$\mathcal{A} \Big|\frac{w,\Sigma}{O(st+s^2)} \left\{ w(X')^{2t+1} \left\| \mathcal{P}(\Sigma - \Sigma_*)\mathcal{P} \right\|_F^2 \leqslant O(s^2) \left(\left(\frac{4C^t \delta^t}{\alpha^2} \right)^2 + \frac{4}{\delta^4} \left\| \mathcal{P}\Sigma_*\mathcal{P} \right\|_{op}^2 \right) \right\} . \tag{4.13}$$

Further, if entries of y_is have bit complexity $\leqslant B$, then, the bit-complexity of the SoS proof is $(Bd/\delta)^{O(st)}$.

We will use the following consequence of certifiable hypercontractivity in the proof of Lemma 4.6.

Lemma 4.7 (Frobenius bound)**.** *Under the hypothesis of Theorem 4.2, for any $h \in \mathbb{N}$, and $d \times d$ matrix-valued indeterminate Q,*

$$\mathcal{A}_5 \Big|\frac{Q,w,X'}{O(h^2)} \left\{ w(X')^{2h} \left\langle \Sigma - \Sigma_*, Q \right\rangle^{2h} \leqslant w(X')^{2h-1} \cdot 2^{2h} (Ch)^{2h} \left(\left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F^{2h} + \left\| RQR \right\|_F^{2h} \right) \right\}$$

Further, if entries of y_is have bit complexity $\leqslant B$, then, the bit-complexity of the SoS proof is $(Bd)^{O(h^2)}$.

Proof. We begin by rewriting

$$\begin{aligned} w(X')\langle \Sigma - \Sigma_*, Q\rangle &= \frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\langle \Sigma - \Sigma_*, Q\rangle \\ &= \frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\langle \Sigma - x_i' x_i'^\top, Q\rangle + \frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\langle x_i' x_i'^\top - \Sigma_*, Q\rangle \\ &= \frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\langle \Sigma - x_i' x_i'^\top, Q\rangle + \frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\langle x_i x_i^\top - \Sigma_*, Q\rangle \end{aligned}$$

where in the last step we applied $\mathcal{A}_2 \Big|\frac{w}{2} \left\{ w_i \mathbf{1}(x_i = y_i)\langle x_i' x_i'^\top, Q\rangle = w_i \mathbf{1}(x_i = y_i)\langle x_i x_i^\top, Q\rangle \right\}$ ala the remark at the beginning of Lemma 4.4 (passing through y_i).

Therefore, by the SoS Almost Triangle Inequality (Fact 3.9) it follows that

$$w(X')^{2h}\langle \Sigma - \Sigma_*, Q\rangle^{2h} \leqslant 2^{2h} \left(\left(\frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\langle \Sigma - x_i' x_i'^\top, Q\rangle \right)^{2h} + \left(\frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\langle x_i x_i^\top - \Sigma_*, Q\rangle \right)^{2h} \right) \tag{4.14}$$

so it suffices to bound each of these terms separately.

Beginning with the first term, we apply SoS Hölder's inequality (Fact 3.8) with $f_i = \langle \Sigma - x_i' x_i'^\top, Q\rangle$ and $g_i = w_i \mathbf{1}(x_i = y_i)$ (which is idempotent, that is $g_i^k = g_i$) to obtain

$$\mathcal{A} \left|\frac{w,\Sigma,X',Q}{O(h^2)}\right. \left\{ \left(\frac{1}{n} \sum_{i=1}^n w_i \mathbf{1}(x_i = y_i) \langle \Sigma - x_i' x_i'^\top, Q\rangle \right)^{2h} \leqslant \left(\frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\right)^{2h-1} \left(\frac{1}{n}\sum_{i=1}^n \langle x_i' x_i'^\top - \Sigma, Q\rangle^{2h}\right) \right.$$
$$\left. \leqslant w(X')^{2h-1} (Ch)^{2h} ||RQR||_F^{2h} \right\} \quad (4.15)$$

where we used the certifiable hypercontractivity ($\mathcal{A}_5$) of X'. Similarly, we may bound that

$$\mathcal{A} \left|\frac{w,\Sigma,X',Q}{O(h^2)}\right. \left\{ \left(\frac{1}{n} \sum_{i=1}^n w_i \mathbf{1}(x_i = y_i) \langle x_i x_i^\top - \Sigma_*, Q\rangle \right)^{2h} \leqslant \left(\frac{1}{n}\sum_{i=1}^n w_i \mathbf{1}(x_i = y_i)\right)^{2h-1} \left(\frac{1}{n}\sum_{i=1}^n \langle x_i' x_i'^\top - \Sigma_*, Q\rangle^{2h}\right) \right.$$
$$\left. \leqslant w(X')^{2h-1} (Ch)^{2h} \left\| \Sigma_*^{\frac{1}{2}} Q \Sigma_*^{\frac{1}{2}} \right\|_F^{2h} \right\} \quad (4.16)$$

by instead using the true anticoncentration of X. Finally, plugging Equations (4.15) and (4.16) into Equation (4.14) yields the desired

$$\mathcal{A} \left|\frac{w,\Sigma,X',Q}{O(h^2)}\right. \left\{ w(X')^{2h} \langle \Sigma - \Sigma_*, Q\rangle^{2h} \leqslant w(X')^{2h-1} 2^{2h} (Ch)^{2h} \left(||RQR||_F^{2h} + \left\| \Sigma_*^{\frac{1}{2}} Q \Sigma_*^{\frac{1}{2}} \right\|_F^{2h} \right) \right\}.$$

The bit complexity bound on the SoS proof follows by accounting the bounds for each of the elementary inequalities used in the argument above. □

We now go on to prove Lemma 4.6.

Proof of Lemma 4.6. From the conclusion of Lemma 4.7, setting $h = 2s$ and letting $\mathcal{P}$ be fixed gives:

$$\mathcal{A}_5 \left|\frac{\Sigma,R,m}{O(st+s^2)}\right. \left\{ w(X')^{4s} \langle \mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}, Q\rangle^{4s} = w(X')^{4s} \langle \Sigma - \Sigma_*, \mathcal{P}Q\mathcal{P}\rangle^{4s} \right.$$
$$\left. \leqslant 2^{4s} w(X')^{4s-1} (2Cs)^{4s} \left(\left\| \Sigma_*^{1/2} \mathcal{P}Q\mathcal{P} \Sigma_*^{1/2} \right\|_F^{4s} + \|R\mathcal{P}Q\mathcal{P}R\|_F^{4s} \right) \right\}. \quad (4.17)$$

Multiplying throughout by the SoS polynomial $w(X')^{4s(t-1)+1}$ yields:

$$\mathcal{A}_5 \left|\frac{\Sigma,R,m}{O(st+s^2)}\right. \left\{ w(X')^{4st+1} \langle \mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}, Q\rangle^{4s} = w(X')^{4st+1} \langle \Sigma - \Sigma_*, \mathcal{P}Q\mathcal{P}\rangle^{4s} \right.$$
$$\left. \leqslant 2^{4s} w(X')^{4st} (2Cs)^{4s} \left(\left\| \Sigma_*^{1/2} \mathcal{P}Q\mathcal{P} \Sigma_*^{1/2} \right\|_F^{4s} + \|R\mathcal{P}Q\mathcal{P}R\|_F^{4s} \right) \right\}. \quad (4.18)$$

As before, let's analyze the two terms on the RHS separately.

Using cyclic properties of Frobenius norm, we know that $||\Sigma_*^{1/2}\mathcal{P}Q\mathcal{P}\Sigma_*^{1/2}||_F^2 = ||\mathcal{P}\Sigma_*\mathcal{P}Q||_F^2$. Therefore, applying Lemma 3.14, we have that:

$$\mathcal{A} \vdash^{\Sigma,w}_{O(st+s^2)} \left\{ w(X')^{4st} \left\| \Sigma_*^{1/2}\mathcal{P}Q\mathcal{P}\Sigma_*^{1/2} \right\|_F^{4s} \leqslant w(X')^{4st} \left(\|\mathcal{P}\Sigma_*\mathcal{P}\|_{op}^{4s} \, \|Q\|_F^{4s} \right) \right\} . \tag{4.19}$$

For the second term, we start from the guarantee of Lemma 4.5 applied to the vector $\mathcal{P}v$:

$$\begin{aligned} \mathcal{A} \vdash^{R,w}_{O(st+s^2)} \Bigg\{ w(X')^{2st}(v^\top \mathcal{P}\Sigma\mathcal{P}v)^{2s} &\leqslant 2^{2s} \left(\left(\frac{8C^t\delta^t}{\alpha^2} \right)^{2s} ||v||_2^{4s} + \frac{2^{2s}}{\delta^{4s}} (v^\top \mathcal{P}^\top \Sigma_* \mathcal{P}v)^{2s} \right) \\ &\leqslant 2^{2s} ||v||_2^{4s} \left(\left(\frac{4C^t\delta^t}{\alpha^2} \right)^{2s} + \frac{2^{2s}}{\delta^{4s}} ||\mathcal{P}\Sigma_*\mathcal{P}||_{op}^{2s} \right) \Bigg\} . \end{aligned} \tag{4.20}$$

From Contraction within SoS (Fact 3.15), we may derive that

$$\left\{ \beta(v^\top A^\top A v)^t \leqslant \Delta ||v||_2^{2t} \right\} \vdash \left\{ \beta^2 ||AQA^\top||_F^{2t} \leqslant \Delta^2 t^{2t} ||Q||_F^{2t} \right\} .$$

Applying this with $\beta = w(X')^{2st}$, $A = R\mathcal{P}$, $t = 2s$, and $\Delta = 2^{2s}\left(\left(\frac{8C^t\delta^t}{\alpha^2} \right)^{2s} + \frac{2^{2s}}{\delta^{4s}} \, ||\mathcal{P}\Sigma_*\mathcal{P}||_{op}^{2s} \right)$ yields

$$\mathcal{A} \vdash^{R,w,Q}_{O(st+s^2)} \left\{ w(X')^{4st} \, \|R\mathcal{P}Q\mathcal{P}R\|_F^{4s} \leqslant 2 \cdot (4s)^{4s} \left(\left(\frac{8C^t\delta^t}{\alpha^2} \right)^{4s} + \frac{2^{4s}}{\delta^{8s}} \, \|\mathcal{P}\Sigma_*\mathcal{P}\|_{op}^{4s} \right) \|Q\|_F^{4s} \right\} \tag{4.21}$$

where we also crucially used an application of the SoS Almost-Triangle Inequality to expand Δ^2.

Plugging back the estimates from (4.19) and (4.21) into (4.18) gives:

$$\mathcal{A}_5 \vdash^{\Sigma,w,Q}_{O(st+s^2)} \left\{ w(X')^{4st+1} \, \langle \mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}, Q \rangle^{4s} \leqslant (O(s))^{4s} \left(\left(\frac{8C^t\delta^t}{\alpha^2} \right)^{4s} + \frac{2^{4s}}{\delta^{8s}} \, \|\mathcal{P}\Sigma_*\mathcal{P}\|_{op}^{4s} \right) \|Q\|_F^{4s} \right\} . \tag{4.22}$$

Note that as $0 \leqslant w(X') \leqslant 1$ we may shrink the LHS by multiplying it through further by $w(X')^{2s-1}$. Substituting $Q = \mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}$ and multiplying throughout by the SoS polynomial $w(X')^{4st+2s}$ now yields:

$$\mathcal{A}_5 \vdash^{\Sigma,w,Q}_{O(st+s^2)} \left\{ w(X')^{8st+4s} \, \|\mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}\|_F^{8s} \leqslant w(X')^{4st+2s}(O(s))^{4s} \left(\left(\frac{8C^t\delta^t}{\alpha^2} \right)^{4s} + \frac{2^{4s}}{\delta^{8s}} \, \|P\Sigma_*P\|_{op}^{4s} \right) \|\mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}\|_F^{4s} \right\} . \tag{4.23}$$

We now apply Lemma 3.13 (Cancellation within SoS) with $A = w(X')^{4st+2s} \, \|\mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}\|_F^{4s}$ and the SoS Almost-Triangle Inequality to obtain that:

$$\mathcal{A} \vdash^{w,\Sigma}_{O(st+s^2)} \left\{ w(X')^{16st+8s} \, \|\mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}\|_F^{16s} \leqslant (O(s))^{16s} \left(\left(\frac{8C^t\delta^t}{\alpha^2} \right)^{16s} + \frac{2^{16s}}{\delta^{32s}} \, \|\mathcal{P}\Sigma_*\mathcal{P}\|_{op}^{16s} \right) \right\} . \tag{4.24}$$

We finally apply Cancellation with Constant RHS (Lemma 3.12) to conclude that:

$$\mathcal{A} \vdash^{w,\Sigma}_{O(st+s^2)} \left\{ w(X')^{2t+1} \, \|\mathcal{P}(\Sigma - \Sigma_*)\mathcal{P}\|_F^{2} \leqslant (O(s))^{2} \left(\left(\frac{8C^t\delta^t}{\alpha^2} \right)^{2} + \frac{4}{\delta^{4}} \, \|\mathcal{P}\Sigma_*\mathcal{P}\|_{op}^{2} \right) \right\} . \tag{4.25}$$

□

4.4 Analysis of Rounding

We first show that setting $X' = X$ (and using the naturally induced assignments for other indeterminates) yields a feasible solution for the relaxation.

Lemma 4.8 (Feasibility of the Relaxation)**.** *For $\delta \leqslant \frac{\alpha^2}{2C}$, suppose there exists a (C,δ)-good set $X \subseteq \mathbb{R}^d$ of size n such that $x_i = y_i$ for αn different $i \leqslant n$. Then, there is a pseudo-distribution of degree $O(s(\delta)+q)$ consistent with the constraint system $\mathcal{A}$ and as a consequence, Step 1 of Algorithm 4.3 succeeds.*

Proof. We give a solution to the constraint system $\mathcal{A}$ to prove that the constraints are satisfiable. We set $X' = X$ and $w_i = 1$ if and only if $x_i = y_i$. Since X is a (C,δ)-good set, we immediately obtained that $\mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4, \mathcal{A}_5$ are satisfied. Therefore, it remains to check $\mathcal{A}_1$.

For $\mathcal{A}_1$, we set $R = \Sigma_*^{1/2}$ – the PSD square root of Σ_* and U to be a matrix such that $UU^\top = \Sigma_*^{1/2}$ (which exists). Let us now prove that there is a Z such that $\mathcal{A}_1$ is feasible. To do so, we make use of the isotropic position of Y.

To begin, note via $|X \cap Y| = \alpha n$ that

$$\mathop{\mathbb{E}}_{x \sim X \cap Y} xx^\top = \frac{1}{|X \cap Y|} \sum_{x \in X \cap Y} xx^\top = \frac{1}{\alpha}\left(\frac{1}{n}\sum_{x \in X \cap Y} xx^\top\right) \preceq \frac{1}{\alpha}\left(\frac{1}{n}\sum_{x \in Y} xx^\top\right) = \frac{1}{\alpha} I.$$

Now, recall that true anti-concentration of X tells us that $\mathbb{P}_{x \sim X}[\langle x, v\rangle^2 \leqslant \frac{\beta}{2} v^\top \Sigma_* v] \leqslant \beta$. Therefore, at most $\frac{\alpha}{2}n$ points in X have $\langle x, v\rangle^2 \leqslant \frac{\alpha}{4} v^\top \Sigma^* v$, and hence at least $\alpha n - \frac{\alpha}{2}n = \frac{\alpha}{2}n$ points in $X \cap Y$ have $\langle x, v\rangle^2 \geqslant \frac{\alpha}{4} v^\top \Sigma^* v$. Therefore, this implies that

$$\mathop{\mathbb{E}}_{x \sim X \cap Y} \langle x, v\rangle^2 = \frac{1}{|X \cap Y|} \sum_{x \in X \cap Y} \langle x, v\rangle^2 \geqslant \frac{1}{\alpha n} \cdot \left(\frac{\alpha n}{2} \cdot \frac{\alpha}{4} v^\top \Sigma_* v\right) = \frac{\alpha}{8} v^\top \Sigma_* v$$

and thus we have $\frac{\alpha}{8}\Sigma_* \preceq \frac{1}{\alpha} I$. Rearranging gives that $\frac{8}{\alpha^2} I - \Sigma_* = ZZ^\top$ for some Z and hence we have feasibility of $\mathcal{A}_1$.

This completes the proof. □

Next, we prove that in expectation, $w(X')$ – the normalized intersection of X' and the (unknown) sample X must be at least α^2. This is a consequence of our relaxation hunting for a maximum entropy (or more precisely, max "collision probability") solution.

Lemma 4.9 (High-entropy Pseudo-distributions Intersect X)**.** *Under the hypothesis of Theorem 4.2, let $\tilde{\zeta}$ be a pseudo-distribution of degree $\geqslant O(s(\delta))$ satisfying $\mathcal{A}$ that minimizes $\left\|\tilde{\mathbb{E}}_{\tilde{\zeta}}\, w\right\|_2$. Then, $\tilde{\mathbb{E}}_{\tilde{\zeta}}[w(X')] \geqslant \alpha^2$.*

Proof. Towards a contradiction, we will show that if the conclusion does not hold then there exists a pseudo-distribution with smaller value of $\left\|\tilde{\mathbb{E}}_{\tilde{\zeta}}\, w\right\|_2$.

Suppose $\tau = \frac{1}{\alpha n}\sum_{i=1}^n \tilde{\mathbb{E}}_{\tilde{\zeta}}[w_i \mathbf{1}(x_i = y_i)] = \frac{1}{\alpha n}\sum_{i=1}^n \tilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \cdot \mathbf{1}(x_i = y_i) < \alpha$ where we used that $\mathbf{1}(x_i = y_i)$ is a constant (as opposed to an indeterminate in our polynomial program). To show a contradiction, we will exhibit a pseudodistribution ν that is 1) feasible for our SoS relaxation and 2) has a smaller value of $\left\|\tilde{\mathbb{E}}_\nu[w]\right\|_2$.

Toward this, notice that we can write

$$\left\| \frac{1}{\alpha n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w] \right\|_2^2 = \frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]^2 = \frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \left(\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \mathbf{1}(x_i = y_i)^2 \right)^2 + \frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \left(\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] (1 - \mathbf{1}(x_i = y_i))^2 \right)^2 \tag{4.26}$$

where we used that $\tilde{\zeta}$ satisfies $w_i^2 = w_i$ for every i.

By Cauchy Schwarz inequality for pseudo-distributions, we observe

$$\begin{aligned}
\tau^2 &= \left(\frac{1}{\alpha n} \sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \mathbf{1}(x_i = y_i)^2 \right)^2 \\
&\leqslant \left(\frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \left(\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \mathbf{1}(x_i = y_i) \right)^2 \right) \left(\sum_{i=1}^{n} \mathbf{1}(x_i = y_i)^2 \right) \\
&= \alpha n \left(\frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \left(\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \mathbf{1}(x_i = y_i) \right)^2 \right) \qquad (4.27)
\end{aligned}$$

and similarly

$$\begin{aligned}
(1 - \tau)^2 &= \left(\frac{1}{\alpha n} \sum_{i=1}^{n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] (1 - \mathbf{1}(x_i = y_i))^2 \right)^2 \\
&\leqslant \left(\frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \left(\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] (1 - \mathbf{1}(x_i = y_i)) \right)^2 \right) \left(\sum_{i=1}^{n} (1 - \mathbf{1}(x_i = y_i))^2 \right) \\
&= (1 - \alpha) n \left(\frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \left(\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] (1 - \mathbf{1}(x_i = y_i)) \right)^2 \right). \qquad (4.28)
\end{aligned}$$

We can apply Equations 4.27 and 4.28 to 4.26 to obtain

$$\left\| \frac{1}{\alpha n} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w] \right\|_2^2 \geqslant \frac{\tau^2}{\alpha n} + \frac{(1 - \tau)^2}{(1 - \alpha) n} = \frac{1}{\alpha n} \left(\tau^2 + (1 - \tau)^2 \frac{\alpha}{1 - \alpha} \right). \tag{4.29}$$

Now we proceed to construct our ν achieving lower norm. Let ζ^* be the (true) distribution supported on a single point w, where $w_i = \mathbf{1}(x_i = y_i)$ (i is an inlier). Due to X being a good sample, it follows that ζ^* satisfies the constraint system.

Then, we let $\nu = (\alpha - \tau)\zeta^* + (1 + \tau - \alpha)\tilde{\zeta}$, and set $\lambda = \alpha - \tau$. Note that as $0 \leqslant \tau < \alpha < 1$, this "mix" is indeed a pseudodistribution. From here, notice that

$$\begin{aligned}
\left\| \frac{1}{\alpha n} \widetilde{\mathbb{E}}_{\nu}[w] \right\|_2^2 &= \frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \left(\lambda \, \widetilde{\mathbb{E}}_{\zeta^*}[w_i] + (1 - \lambda) \, \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \right)^2 \\
&= \frac{1}{\alpha^2 n^2} \sum_{i=1}^{n} \left[\lambda^2 \, \widetilde{\mathbb{E}}_{\zeta^*}[w_i]^2 + (1 - \lambda)^2 \, \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]^2 + 2\lambda(1 - \lambda) \, \widetilde{\mathbb{E}}_{\zeta^*}[w_i] \, \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] \right]
\end{aligned}$$

$$= \frac{\lambda^2}{\alpha n} + (1-\lambda^2)\left\|\frac{1}{\alpha n}\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2 + 2\lambda(1-\lambda)\frac{\tau}{\alpha n}.$$

Therefore, it follows that

$$\begin{aligned}\left\|\frac{1}{\alpha n}\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2 - \left\|\frac{1}{\alpha n}\widetilde{\mathbb{E}}_{\nu}[w]\right\|_2^2 &= (2\lambda-\lambda^2)\left\|\frac{1}{\alpha n}\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2 - \frac{\lambda^2}{\alpha n} - 2\lambda(1-\lambda)\frac{\tau}{\alpha n} \\ &\geqslant \frac{\lambda}{\alpha n}\left((2-\lambda)\left(\tau^2 + (1-\tau)^2\frac{\alpha}{1-\alpha}\right) - \lambda - 2\tau(1-\lambda)\right).\end{aligned}$$

Plugging in $\lambda = \alpha - \tau$ yields

$$\left\|\frac{1}{\alpha n}\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2 - \left\|\frac{1}{\alpha n}\widetilde{\mathbb{E}}_{\nu}[w]\right\|_2^2 \geqslant \frac{(1-\tau^2)(\alpha-\tau)^2}{(1-\alpha)\alpha n} > 0$$

when $0 \leqslant \tau < \alpha < 1$. Hence, we have found a pseudodistribution ν satisfying the constraints and of lower norm, contradiction. Therefore it must be the case that $\tau \geqslant \alpha$. □

We will now use the above fact along with our spectral and Frobenius recovery guarantees to derive properties of the rounded solutions. Our rounding prcedure depends on sampling multisets.

Notation 4.2 (Sampling from Pseudo-distribution)**.** *Let $k \in \mathbb{N}$ and $\tilde{\zeta}$ be a pseudodistribution. Consider the distribution $D(k, \tilde{\zeta})$ on multisets $S \subseteq [n]$ of size k chosen via the following process:*

1. *Pick $S = (i_1, i_2, \ldots, i_k) \subseteq [n]$ with probability $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\prod_{j\in S} w_j]/(\alpha n)^k$.*
2. *Output S.*

Further define the conditional distribution $D'(k, \tilde{\zeta}) = D(k, \tilde{\zeta}) \mid x_i = y_i \forall i \in S$.

Note that $\sum_{|S|=k}\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\prod_{j\in S} w_j] = (\alpha n)^k$ to show that this is a well defined probability distribution.

Lemma 4.10 (Analysis of Rounding, Lower-Bound)**.** *Assume the hypothesis of Theorem 4.2. Let $\tilde{\zeta}$ be a pseudo-distribution of degree $\geqslant O(s(\delta))$ consistent with $\mathcal{A}$ minimizing $\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}}\, w\right\|_2$ and fix $\delta = \alpha^3/2C \leqslant \alpha^2/2C$. Then,*

$$\mathop{\mathbb{P}}_{S\sim D'(2t+1,\tilde{\zeta})}\left[\left(\frac{16C^2}{\alpha^8}\right)\frac{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S\Sigma]}{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S]} \succeq \Sigma_*\right] \geqslant \alpha^2/4\,.$$

Proof. We know that for any $S \subseteq [n]$, $\widetilde{\mathbb{E}}[w_S\Sigma] = \widetilde{\mathbb{E}}[w_S RR^\top] \succeq 0$. Next, from Lemma 4.4 we have that

$$\mathcal{A}\,\Big|\frac{\Sigma,w}{4s}\,\left\{(v^\top\Sigma v)(v^\top\Sigma_* v)^{s-1} \geqslant \delta^2(w(X') - C\delta)(v^\top\Sigma_* v)^s\right\}.$$

Let's multiply both sides by the SoS polynomial $w_S = w_S^2$ and take pseudo-expectations with respect to $\tilde{\zeta}$. Then, by Fact 3.16 acting on w_S, Σ, X', we must have that **for every** $v \in \mathbb{R}^d$,

$$(v^\top\widetilde{\mathbb{E}}[w_S\Sigma]v)(v^\top\Sigma_* v)^{s-1} \geqslant \delta^2\left(\widetilde{\mathbb{E}}[w(X')w_S] - C\delta\,\widetilde{\mathbb{E}}[w_S]\right)\left(v^\top\Sigma_* v\right)^s\ .$$

For any S such that $\widetilde{\mathbb{E}}[w_S] > 0$, let $\hat{\Sigma}_S = \widetilde{\mathbb{E}}[w_S \Sigma]/\widetilde{\mathbb{E}}[w_S]$. Then, dividing through by $\widetilde{\mathbb{E}}[w_S]$ yields that for every $v \in \mathbb{R}^d$, we have

$$(v^\top \hat{\Sigma}_S v)(v^\top \Sigma_* v)^{s-1} \geqslant \delta^2 \left(\frac{\widetilde{\mathbb{E}}[w(X')w_S]}{\widetilde{\mathbb{E}}[w_S]} - C\delta \right) (v^\top \Sigma_* v)^s \,. \tag{4.30}$$

Let us analyze the random variable $a_S = \left(\frac{\widetilde{\mathbb{E}}[w(X')w_S]}{\widetilde{\mathbb{E}}[w_S]} - C\delta \right)$ when $S \sim D'(2t+1, \tilde{\zeta})$ (note that the only randomness here is over the choice of S, the pseudo-distribution $\tilde{\zeta}$ is fixed).

Let $G = \{i \in [n] : x_i = y_i\} \subseteq [n]$ be the set of "good" indices. Then, noting that $\widetilde{\mathbb{E}}[w(X')^{2t+1}] = \sum_{S=(i_1,i_2,\ldots,i_{2t+1}) \in G^{2t+1}} [\widetilde{\mathbb{E}}[w_S]]$ (by moving out $\mathbf{1}(x_i = y_i)$), we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S \sim D'(2t+1,\tilde{\zeta})}[a_S] &= \frac{1}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]} \sum_{S \in G^{2t+1}} \widetilde{\mathbb{E}}[w_S] \frac{\widetilde{\mathbb{E}}\left[w(X')w_S\right]}{\widetilde{\mathbb{E}}[w_S]} - C\delta \\
&= \frac{1}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]} \sum_S \widetilde{\mathbb{E}}[w(X')w_S] - C\delta \\
&= \frac{1}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]} \cdot \widetilde{\mathbb{E}}\left[w(X') \sum_S w_S \right] - C\delta \\
&= \frac{\widetilde{\mathbb{E}}[w(X')^{2t+2}]}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]} - C\delta
\end{aligned}$$

We claim then that $\widetilde{\mathbb{E}}[w(X')^{2t+2}] \geqslant \widetilde{\mathbb{E}}[w(X')^{2t+1}]\,\widetilde{\mathbb{E}}[w(X')]$ (which is true in real expectations by monotonicity). Indeed, $\widetilde{\mathbb{E}}[w(X')] \leqslant \widetilde{\mathbb{E}}[w(X')^{2t+2}]^{1/(2t+2)}$ by application of Hölder's inequality (Fact 3.2) with $f = 1$, $g = w(X')$ and $\widetilde{\mathbb{E}}[w(X')^{2t+1}] \leqslant \widetilde{\mathbb{E}}[w(X')^{2t+2}]^{(2t+1)/(2t+2)}$ by application of Hölder's with $f = w(X')$ and $g = 1$. Multiplying these gives the result. Hence, we have

$$\mathop{\mathbb{E}}_{S \sim D'(t,\tilde{\zeta})}[a_S] = \frac{\widetilde{\mathbb{E}}[w(X')^{2t+2}]}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]} - C\delta \geqslant \widetilde{\mathbb{E}}[w(X')] - C\delta \geqslant \alpha^2 - C\delta \geqslant \frac{\alpha^2}{2}.$$

where the second to last step is by Lemma 4.9.

Now, since $a_S \leqslant 1$ for every S, we may apply a Reverse Markov Bound to obtain that

$$\mathop{\mathbb{P}}_{S \sim D'}[a_S \leqslant \alpha^2/4] = \mathop{\mathbb{P}}_{S \sim D'}[1 - a_S \geqslant 1 - \alpha^2/4] \leqslant \frac{1 - \mathbb{E}[a_S]}{1 - \alpha^2/4} \leqslant \frac{1 - \alpha^2/2}{1 - \alpha^2/4}$$

and hence $\mathbb{P}_{S \sim D'}[a_S > \alpha^2/4] \geqslant \frac{\alpha^2/4}{1-\alpha^2/4} \geqslant \frac{\alpha^2}{4}$.

Thus, with probability at least $\alpha^2/4$ over the choice of $S \sim D'$, it must hold that $\widetilde{\mathbb{E}}[w_S w(X')]/\widetilde{\mathbb{E}}[w_S] - C\delta \geqslant \alpha^2/4$. Under this event and letting $\delta = \frac{\alpha^3}{2C}$, it follows by (4.30) that

$$\left(v^\top \frac{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S \Sigma]}{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S]} v \right) (v^\top \Sigma_* v)^{s-1} \geqslant \delta^2 \cdot \frac{\alpha^2}{4} (v^\top \Sigma_* v)^s \geqslant \frac{\alpha^8}{16C^2} (v^\top \Sigma_* v)^s \,.$$

Dividing through by $(v^\top \Sigma_* v)^{s-1}$ (and the upcoming guarantee holds even if this is 0), we obtain $v^\top \frac{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S \Sigma]}{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S]} v \geqslant \frac{\alpha^8}{16C^2} v^\top \Sigma_* v$ for every $v \in \mathbb{R}^d$. Therefore, this implies that $\frac{16C^2}{\alpha^8} \frac{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S \Sigma]}{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_S]} \succeq \Sigma_*$ with probability at least $\alpha^2/4$ over the choice of $S \sim D'(t, \tilde{\zeta})$ and we are done. □

Lemma 4.11 (Analysis of Rounding, Upper-Bound)**.** *Fix $\delta = \alpha^3/2C$. Let $\tilde{\zeta}$ be a pseudo-distribution of degree $\geqslant O(s(\delta)^2)$ consistent with $\mathcal{A}$ minimizing $\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}}\, w\right\|_2$. Let $\mathcal{P}$ be the projection matrix to the subspace spanned by all the eigenvectors of Σ_* with eigenvalues at most $O(\alpha^{6t+8})$. Then, with probability at least $1 - \alpha^2/10$ over the choice of $S \sim D'(2t+1, \tilde{\zeta})$, we have that for $\hat{\Sigma}_S = \widetilde{\mathbb{E}}[w_S \Sigma]/\widetilde{\mathbb{E}}[w_S]$:*

$$\mathcal{P}\hat{\Sigma}_S\mathcal{P} \preceq O(\alpha^{2t-20})I + \mathcal{P}\Sigma_*\mathcal{P}\,. \tag{4.31}$$

Proof of Lemma 4.11. From Lemma 4.6, we have:

$$\mathcal{A}\,\Big|\!\frac{\,w,\Sigma\,}{O(st)}\left\{ w(X')^{2t+1}\left\|\mathcal{P}(\Sigma-\Sigma_*)\mathcal{P}\right\|_F^2 \leqslant O(s^2)\left(\left(\frac{8C^t\delta^t}{\alpha^2}\right)^2 + \frac{4}{\delta^4}\left\|\mathcal{P}\Sigma_*\mathcal{P}\right\|_{op}^2\right)\right\}. \tag{4.32}$$

Using that $s(\delta) = O(1/\delta^2)$, it follows that $O(s^2)\left(\frac{8C^t\delta^t}{\alpha^2}\right)^2 = O(C^{2t})\cdot\delta^{2t-4}\cdot\alpha^{-4}$, and using $\delta = \alpha^3/(2C)$ yields that this term is at most $O(C^4\cdot\alpha^{6t-16})$. The second term is at most $O(1/\delta^8)\cdot||\mathcal{P}\Sigma_*\mathcal{P}||_{op}^2$ using again that $s(\delta) = O(1/\delta^2)$. Thus, altogether, we have:

$$\mathcal{A}\,\Big|\!\frac{\,w,\Sigma\,}{O(st)}\left\{ w(X')^{2t+1}\left\|\mathcal{P}(\Sigma-\Sigma_*)\mathcal{P}\right\|_F^2 \leqslant O(\alpha^{6t-16}) + O\left(\frac{1}{\delta^8}\right)\left\|\mathcal{P}\Sigma_*\mathcal{P}\right\|_2^2\right\}. \tag{4.33}$$

Taking pseudo-expectations with respect to $\tilde{\zeta}$ yields:

$$\widetilde{\mathbb{E}}[w(X')^{2t+1}\left\|\mathcal{P}(\Sigma-\Sigma_*)\mathcal{P}\right\|_F^2] \leqslant O(\alpha^{6t-16}) + O\left(\frac{1}{\delta^8}\right)\left\|\mathcal{P}\Sigma_*\mathcal{P}\right\|_2^2\,. \tag{4.34}$$

Expanding, using that $\tilde{\zeta}$ satisfies $w_S^2 = w_S$ and denoting once more $G = \{i \in [n] \mid x_i = y_i\} \subseteq [n]$ as the "good" set yields:

$$\frac{1}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]}\sum_{S=(i_1,i_2,\ldots,i_{2t+1})\in G^{2t+1}}\widetilde{\mathbb{E}}[w_S]\left\|\mathcal{P}\left(\frac{\widetilde{\mathbb{E}}[w_S\Sigma]}{\widetilde{\mathbb{E}}[w_S]}-\Sigma_*\right)\mathcal{P}\right\|_F^2 \leqslant \frac{O(\alpha^{6t-16}) + O\left(\frac{1}{\delta^8}\right)\left\|\mathcal{P}\Sigma_*\mathcal{P}\right\|_2^2}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]\cdot}\,. \tag{4.35}$$

From Lemma 4.9 and an application of Hölder's, we know that $\widetilde{\mathbb{E}}[w(X')^{2t+1}] \geqslant \alpha^{4t+2}$. Thus, we conclude that

$$\frac{1}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]}\sum_{S=(i_1,i_2,\ldots,i_{2t})\in G^{2t}}\widetilde{\mathbb{E}}[w_S]\left\|\mathcal{P}\left(\frac{\widetilde{\mathbb{E}}[w_S\Sigma]}{\widetilde{\mathbb{E}}[w_S]}-\Sigma_*\right)\mathcal{P}\right\|_F^2 \leqslant O(\alpha^{2t-18}) + O\left(\frac{1}{\alpha^{4t+2}\delta^8}\right)\left\|\mathcal{P}\Sigma_*\mathcal{P}\right\|_2^2\,. \tag{4.36}$$

Since $\mathcal{P}$ is the projection to the subspace where Σ_* has eigenvalues smaller than $O(\alpha^{6t-16}\delta^8) = O(\alpha^{6t+8})$, the second term is $O(\alpha^{2t-18})$. So altogether, the RHS above is $O(\alpha^{2t-18})$. Thus,

$$\frac{1}{\widetilde{\mathbb{E}}[w(X')^{2t+1}]}\sum_{S=(i_1,i_2,\ldots,i_{2t})\in G^{2t}}\widetilde{\mathbb{E}}[w_S]\left\|\mathcal{P}\left(\frac{\widetilde{\mathbb{E}}[w_S]\Sigma}{\widetilde{\mathbb{E}}[w_S]}-\Sigma_*\right)\mathcal{P}\right\|_F^2 \leqslant O(\alpha^{2t-18})\,. \tag{4.37}$$

Note that by definition of D', we have that is exactly

$$\mathop{\mathbb{E}}_{S\sim D'(2t+1,\tilde{\zeta})}\left[\left\|\mathcal{P}(\hat{\Sigma}_S-\Sigma_*)\mathcal{P}\right\|_F^2\right]\leqslant O(\alpha^{2t-18})$$

Hence, by Markov's Inequality, it follows that

$$\mathop{\mathbb{P}}_{S\sim D'(2t+1,\tilde{\zeta})}\left[\left\|\mathcal{P}(\hat{\Sigma}_S-\Sigma_*)\mathcal{P}\right\|_F^2\leqslant O(\alpha^{2t-20})\right]\geqslant 1-\frac{\alpha^2}{10}.$$

For such an S, we must then have that

$$\mathcal{P}\hat{\Sigma}_S\mathcal{P}\preceq O(\alpha^{2t-20})I+\mathcal{P}\Sigma_*\mathcal{P}\,.$$

This completes the proof. □

Putting Things Together We now put the upper and lower bounds above together to prove Theorem 4.2. We will use the following simple bound:

Lemma 4.12 (Splitting on Projections)**.** *Let $\mathcal{P}$ be a projection matrix to a subspace of $\mathbb{R}^d$. Let A be any $d\times d$ PSD matrix. Then, we have:*

$$A\preceq 2\mathcal{P}A\mathcal{P}+2(I-\mathcal{P})A(I-\mathcal{P})\,.$$

Proof. For any vector $v\in\mathbb{R}^d$, we have:

$$\begin{aligned}v^\top Av&=v^\top(\mathcal{P}+I-\mathcal{P})A(\mathcal{P}+I-\mathcal{P})v\\&=v^\top\mathcal{P}A\mathcal{P}v+v^\top(I-\mathcal{P})A(I-\mathcal{P})v+2v^\top\mathcal{P}A(I-\mathcal{P})v\,.\end{aligned}$$

We now have using the Cauchy-Schwarz followed by the AM-GM inequality:

$$\begin{aligned}v^\top\mathcal{P}A(I-\mathcal{P})v&=v^\top\mathcal{P}A^{1/2}A^{1/2}(I-\mathcal{P})v\\&\leqslant\sqrt{v^\top\mathcal{P}A\mathcal{P}v}\sqrt{v^\top(I-\mathcal{P})A(I-\mathcal{P})v}\\&\leqslant\frac{v^\top\mathcal{P}A\mathcal{P}v+v^\top(I-\mathcal{P})A(I-\mathcal{P})v}{2}\,.\end{aligned}$$

□

Proof of Theorem 4.2. From the constraints $\mathcal{A}$ and that the fact that $\tilde{\zeta}$ is consistent with $\mathcal{A}$ along with Fact 3.16, we must have that for every $2t+1$-tuple S,

$$\tilde{\mathbb{E}}[w_S\Sigma]\preceq\tilde{\mathbb{E}}[w_S]\frac{8}{\alpha^2}\,.$$

Thus,

$$\hat{\Sigma}_S=\frac{\tilde{\mathbb{E}}[w_S\Sigma]}{\tilde{\mathbb{E}}[w_S]}\preceq\frac{8}{\alpha^2}\,.$$

Next, from Lemma 4.11, we know that with probability $1-\alpha^2/10$ over sampling $S \sim D'(2t+1,\tilde{\zeta})$ we must have:

$$\mathcal{P}\hat{\Sigma}_S\mathcal{P} \preceq \mathcal{P}\Sigma_*\mathcal{P} + O(\alpha^{2t-20})I\,.$$

Thus, applying Lemma 4.12 and noting that $(I-\mathcal{P})^2 = I-\mathcal{P}$, we have that for such an S:

$$\hat{\Sigma}_S \preceq 2\mathcal{P}\Sigma_*\mathcal{P} + O(\alpha^{2t-20})I + \frac{16}{\alpha^2}(I-\mathcal{P})\,.$$

Next, observe that $(I-\mathcal{P})\Sigma_*(I-\mathcal{P}) \succeq O(\alpha^{6t+8})(I-\mathcal{P})$ since $\mathcal{P}$ is the subspace of all eigenvectors of Σ_* with eigenvalues $\leqslant O(\alpha^{6t+8})$.

Note also that $\Sigma_* = \mathcal{P}\Sigma_*\mathcal{P} + (I-\mathcal{P})\Sigma_*(I-\mathcal{P})$ (for any vector v, $v^\top(I-\mathcal{P})\Sigma_* v = 0$ by orthonormality of eigenvectors and the fact that eigenvectors must lie in either $\mathcal{P}$ or $I-\mathcal{P}$).

Thus, we must have that with probability at least $1-\alpha^2/10$ over the choice of $S \sim D'(2t+1,\tilde{\zeta})$:

$$\hat{\Sigma}_S \preceq 2\mathcal{P}\Sigma_*\mathcal{P} + O(\alpha^{2t-20})I + O\left(\frac{1}{\alpha^{6t+8}}\right)\cdot\frac{16}{\alpha^2}(I-\mathcal{P})\Sigma_*(I-\mathcal{P}) \preceq O\left(\frac{1}{\alpha^{6t+10}}\right)\Sigma_* + O(\alpha^{2t-20})I\,.$$

Next, from Lemma 4.10, we have that with probability at least $\alpha^2/4$ over the choice of S conditioned on S satisfying $x_i = y_i$ for $i \in S$:

$$O(1/\alpha^8)\hat{\Sigma}_S \succeq \Sigma_*\,.$$

By a union bound, we obtain that with probability at least $\alpha^2/10$ over the choice of $S \sim D'(2t+1,\tilde{\zeta})$:

$$\Sigma_* \preceq O(1/\alpha^8)\hat{\Sigma}_S \preceq O(\frac{1}{\alpha^{6t+18}})\Sigma_* + O(\alpha^{2t-28})I\,. \tag{4.38}$$

Finally, note that since $\tilde{\mathbb{E}}[w(X')^t] \geqslant \alpha^2$, the chance that S satisfies $x_i = y_i$ for every $i \in S$ (as in, our draw from D is also from D') is at least $\tilde{\mathbb{E}}[w(X')^t]/(\alpha^t) \geqslant \alpha^t$. Thus, together, we obtain that with probability at least $\alpha^{t+2}/10$, the candidate $O(1/\alpha^8)\hat{\Sigma}_S$ corresponding to the set S output by the rounding algorithm satisfies (4.38). This completes the proof.

□

5 Subgaussian Restriction

In this section, we describe and analyze a subroutine that effectively allows us assume that the corrupted sample Y, after a pruning step, itself has subgaussian moments with respect to its covariance. To do so, we use a new definition of a good set as compared to the previous section.

Definition 5.1 (Well-behaved Set). For $d \in \mathbb{N}$, we say that a subset $X \subseteq \mathbb{Q}^d$ with mean μ_* and covariance Σ_* is a (C,δ,t)-well behaved set if:

1. **Small Mean:** $\mu_*\mu_*^\top \preceq 0.1\Sigma_*$.

2. **Anticoncentration:** X is $(\delta, C\delta)$-anticoncentrated.

3. **Subgaussianity:** For indeterminate v and all $s \leqslant 2t$,

$$\vdash^{v}_{O(s)} \left\{ \mathop{\mathbb{E}}_{x \sim X} \langle x, v\rangle^{2s} \leqslant (Cs)^{2s} \cdot \left(\mathop{\mathbb{E}}_{x \sim X} \langle x, v\rangle^{2} \right)^{s} \right\}.$$

With this definition in mind, we will prove the following main theorem:

Theorem 5.2 (Subgaussian Restriction). *Fix $1 \geqslant \alpha > 0$ and $d, n, t \in \mathbb{N}$. Let $\tau > 0$. Let X be (C, δ, t)-well behaved set and let Y be an arbitrary collection of n points in $\mathbb{R}^d$ such that $|Y \cap X| \geqslant \alpha n$ and $\frac{1}{n}\sum_i y_i y_i^{\top} = (1 \pm 2^{-d})I$. Then, for any $0 < \tau \leqslant O(\alpha^{2t+11})$ there is a $Y' \subseteq Y$ with $|X \cap Y'| \geqslant \max(\alpha|Y'|, (\alpha - \alpha^{10})n)$ and satisfying*

$$\vdash^{v}_{O(t)} \left\{ \mathop{\mathbb{E}}_{y' \sim Y'} \langle y', v\rangle^{2t} \leqslant \frac{1}{\tau} \left(\frac{16Ct}{\alpha^2} \right)^{t} \left(\mathop{\mathbb{E}}_{y' \sim Y'} \langle y', v\rangle^{2} \right)^{t} \right\}.$$

Further, given α, Y as above, a $Y' \subseteq Y$ satisfying the above properties can be found in time $(Bn)^{O(t)}$ where B is the bit complexity of entries of Y.

Our algorithm to establish the above theorem uses a subroutine that shows that if Y is not subgaussian with the parameters above, then, we can find a subset of Y that is "outlier-heavy", that is, contains significantly bigger fraction of outliers than in all of Y. Removing these points can only "increase" the density of the inliers so cannot hurt us.

We will use the following simple inequality that relates the tails of a distribution to its mean.

Lemma 5.3 (Expectation vs Tail). *Let Z be a non-negative real-valued random variable. Then,*

$$\mathbb{E}[Z] \leqslant \frac{1}{2} + \int_{1/2}^{\infty} \mathbb{P}[Z \geqslant L] dL.$$

Proof. $\mathbb{E}[Z] = \int_0^{\infty} \mathbb{P}[Z \geqslant L]dL = \int_0^{1/2} \mathbb{P}[Z \geqslant L]dL + \int_{1/2}^{\infty} \mathbb{P}[Z \geqslant L]dL \leqslant 1/2 + \int_{1/2}^{\infty} \mathbb{P}[Z \geqslant L]dL.$ □

We will use the above bound to show that if Y is not certifiably subgaussian, then, the outliers must make an outsized contribution to one of a few natural sections of the tail of Y.

Lemma 5.4 (Outliers Must Dominate Some Portion of Tail). *Fix $1 \geqslant \alpha > 0$ and $d, n, t \in \mathbb{N}$. Let $\tau > 0$ and X be (C, δ, t)-well behaved set.*

Let Y be an arbitrary collection of n points in $\mathbb{R}^d$ such that $|Y \cap X| \geqslant \alpha' n$ for $\alpha' \geqslant \alpha/2$ and $\frac{1}{n}\sum_i y_i y_i^{\top} = (1 \pm 2^{-d})I$. Suppose there is a pseudo-distribution $\tilde{\zeta}$ of degree $\geqslant 4t$ over d-dimensional vector-valued indeterminate v such that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\mathbb{E}_{y\sim Y}\langle v, y\rangle^{2t}] > \Gamma$ for $\Gamma = \frac{1}{\tau}(16Ct/\alpha^2)^t$. Then, there is a $L > 0$ such that $\mathbb{P}_{y\sim Y}[\tilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{4t}] \geqslant L] > \Gamma/4L$. In contrast, for every $L > 0$, $\mathbb{P}_{x\sim X}[\tilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x, v\rangle^{4t}] \geqslant L] \leqslant \tau^2\Gamma^2/L$.

Proof. Let's prove the first claim. Note that $\Gamma > 1$ and suppose for the sake of contradiction that for every $L > 0$, $\mathbb{P}_{y\sim Y}[\tilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{4t}] \geqslant L] \leqslant \Gamma/4L$. Observe that by Cauchy-Schwarz inequality for pseudo-distributions and the fact that $\tilde{\zeta}$ has degree $\geqslant 4t$, we must have that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{4t}] \geqslant \left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{2t}]\right)^2$. Thus, whenever $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{2t}] \geqslant L$, $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{4t}] \geqslant L^2$.

Applying Lemma 5.3 to the random variable $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{2t}]$ as y is chosen uniformly at random from Y (notice the randomness here is simply y being chosen uniformly from Y), we have:

$$\begin{aligned}\mathop{\mathbb{E}}_{y\sim Y}[\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{2t}]] &= 1/2 + \int_{1/2}^{\infty} \mathbb{P}[\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{2t}] \geqslant L]dL \\ &\leqslant 1/2 + \int_{1/2}^{\infty} \mathbb{P}[\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{4t}] \geqslant L^2]dL \\ &\leqslant 1/2 + \int_{1/2}^{\infty} \frac{\Gamma}{4L^2}dL \leqslant 1/2 + \Gamma/2 < \Gamma\,,\end{aligned}$$

since $\Gamma > 1$. This is a contradiction and proves the first claim.

Next, let $\Sigma_* = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$. From the argument in Lemma 4.8, we have that $\Sigma_* \preceq \frac{8}{\alpha^2}I$.

By C-certifiable $4t$-subgaussianity of X and the fact that $\tilde{\zeta}$ is a pseudo-distribution of degree $\geqslant 4t$, we must have that $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\mathbb{E}_{x\sim X}[\langle x, v\rangle^{4t}]] \leqslant (2Ct)^{2t}\left(\mathbb{E}_{x\sim X}\langle x, v\rangle^2\right)^{2t} = (2Ct)^{2t}(\frac{8}{\alpha^2})^{2t} = (\tau\Gamma)^2$. By Markov's inequality, $\mathbb{P}_{x\sim X}[\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle x, v\rangle^{4t}] > L] \leqslant \tau^2\Gamma^2/L$.

□

We can now describe the subgaussian restriction algorithm.

Algorithm 5.5 (Subgaussian Restriction Algorithm)**.**

Given: A set of points $Y = \{y_1, y_2, \dots, y_n\} \subseteq \mathbb{Q}^d$ and $\alpha, \tau > 0$.

Output: A subset $Y' \subseteq Y$ that is certifiably subgaussian:

$$\vdash_{2t}^{v} \left\{ \mathop{\mathbb{E}}_{y\sim Y'} \langle y, v\rangle^{2t} \leqslant \frac{1}{\tau}\left(\frac{16Ct}{\alpha^2}\right)^t \left(\mathop{\mathbb{E}}_{y\sim Y'} \langle y, v\rangle^2\right)^t \right\}.$$

Operation: Initialize $Y' = Y$. While true, do:

1. **Isotropize:** By a linear transformation of all $y \in Y'$, ensure that $\mathbb{E}_{y\sim Y'} yy^\top = (1 \pm 2^{-d})I$.
2. **Subgaussianity Check:** Find a degree $4t$ pseudo-distribution $\tilde{\zeta}$ over d-dimensional vector-valued indeterminate v satisfying $\|v\|_2^2 = 1$ and maximizing $\mathbb{E}_{y\sim Y'}[\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{2t}]]$. If the objective value is $\leqslant \Gamma$, halt and return Y'.
3. **Find Outlier-Heavy Level Set:** If not, by binary search, find L such that $\mathbb{P}_{y\sim Y'}[\widetilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y\rangle^{4t}] \geqslant L] > \Gamma/4L$. Such an L is guaranteed to exist by Lemma 5.4. Let R be the set of all such points in Y.
4. **Prune:** Set $Y' = Y \setminus R$. Go to step 1.

Proof of Theorem 5.2. Starting with $Y' = Y$, Algorithm 5.5 repeatedly (approximately) isotropizes Y' and whenever Y' does not satisfy the desired subgaussianity condition, it finds an "outlier-heavy" subset of Y' to prune away.

Let's analyze this algorithm. We will prove the following two facts about the run of the algorithm to analyze it.

1. In each iteration either we halt and return or remove at least one point from Y'. This immediately implies that the algorithm terminates in n iterations and thus, in time $(Bn)^{O(t)}$ (where B is the bit complexity of entries of Y).

2. Let $R \subseteq Y'$ be the subset of points removed in any iteration. Then, at most $\alpha^{10}|R|$ of the points belong to $X \cap Y$. Notice that this immediately implies that the total number of "inliers" (i.e. points in $X \cap Y$) removed are at most $\alpha^{10} n$.

We will prove both the claims by induction over the iterations of the algorithm. Consider any iteration starting with Y'. Let $|X \cap Y'| = \alpha'$. Then, by inductive assumption, we know that $\alpha' \geqslant (\alpha - \alpha^{10}) \geqslant \alpha/2$. By an argument similar to the one in the proof of Lemma 4.8, we can infer that (after the isotropic linear transformation), the transformed $\Sigma_* \preceq \frac{8}{\alpha^2} I$. If Y' does not pass the subgaussianity check, then, by duality of pseudo-distributions and SoS proofs, we can find a pseudo-distribution $\tilde{\zeta}$ on v such that $\mathbb{E}_{y' \sim Y'} \tilde{\mathbb{E}}_{\tilde{\zeta}}[\langle v, y' \rangle^{2t}] > \Gamma$. Thus, by Lemma 5.4, the binary search to find L for an outlier-dense level set must succeed. So we must remove at least one point from Y' (proving (1)). We know then that the fraction of R contained in $X \cap Y'$ is at most $\frac{\frac{\tau^2 \Gamma^2}{L}}{\frac{\alpha/2 \cdot \Gamma}{8L}}$ where we used that $|Y'| \geqslant \frac{\alpha}{2} \cdot n$. Rearranging this yields that we need $\frac{128 C t \tau}{\alpha^{2t+1}} \leqslant \alpha^{10}$ implying that we can take $\tau = O(\alpha^{2t+11})$ to finish (2). This completes the proof.

□

6 Splitting via Paley-Zygmund Anti-Concentration

In this section, we describe and analyze a subroutine that takes input a corrupted sample Y and prunes away a constant fraction of Y. Crucially, our splitting algorithm requires only a mild anti-concentration property that can be inferred from only moment *upper bounds*. This is in contrast to strong (and certifiable) anticoncentration needed in our coarse spectral recovery algorithm.

Our algorithm below first performs a simple check on a candidate obtained from coarse spectral recovery step. If a candidate passes the check, we obtain a certificate that a coarse spectral estimate in fact must in fact be a multiplicative spectral estimate. If not, our algorithm efficiently splits Y into two approximately balanced parts such that the most of $X \cap Y$ is included on one side. Thus, we either "finish" by finding a good candidate covariance or can recurse and make progress be decreasing the fraction of the sample that is corrupted.

Our algorithm itself is simple:

Algorithm 6.1 (Spectral Splitting)**.**
1. **Input:** $Y \subseteq \mathbb{R}^d$, a vector $v \in \mathbb{R}^d$.
2. **Operation:** Return Y_2 – the set of all points y_i such that $\langle y_i, v \rangle^2 > 0.5$.

We first prove that the splitting algorithm always makes progress.

Lemma 6.2 (Splitting Algorithm: Progress)**.** *Let Y be a set of points in $\mathbb{R}^d$ such that 1) $\frac{1}{n}\sum_i y_i y_i^\top = (1 \pm 2^{-d})I$ and 2) $\frac{1}{n}\sum_{i=1}^n \langle y_i, v\rangle^{2t} \leqslant \Delta \|v\|_2^{2t}$ for $\Delta = \frac{1}{\tau}\left(\frac{16Ct}{\alpha'^2}\right)^t$ and all $v \in \mathbb{R}^d$. Let $Y_2 = \{y \in Y \mid \langle y, v\rangle^2 > 1/2\}$. Then, $|Y_2| > O(\alpha^{10})|Y|$.*

Next, we prove that if Y intersects with the original uncorrupted sample X appreciably, then the splitting algorithms prunes away only a small fraction of points from $Y \cap X$.

Lemma 6.3 (Splitting Algorithm: Correctness)**.** *Suppose that X is a set of n points satisfying $2t$-certifiable C-subgaussianity: $\frac{1}{n}\sum_{i=1}^n \langle x_i, v\rangle^{2t} \leqslant (Ct)^t \left(\frac{1}{n}\sum_{i=1}^n \langle x_i, v\rangle^2\right)^t$ for $t = O(1/\alpha)$. For some $\alpha, \tau > 0$, suppose Y is a set of n points in $\mathbb{R}^d$ satisfying $|Y \cap X| \geqslant \alpha' n \geqslant \alpha n/2$. Let $\widehat{\Sigma}$ be a PSD matrix such that $\Sigma_* \preceq \widehat{\Sigma}$ and suppose there is a unit vector v such that $v^\top \widehat{\Sigma} v \leqslant \eta$ for $\eta < O(\alpha^6)$. Then, there is a* poly(n) *time algorithm to split Y into $Y_1 \cup Y_2$ such that $|X \cap Y_1| \geqslant (\alpha' - O(\alpha^{12})|Y|$.*

We will use the following basic consequence of the Paley-Zygmund anti-concentration inequality:

Lemma 6.4 (Mild Anti-Concentration via Paley-Zygmund)**.** *Let Y be a set of n points in $\mathbb{R}^d$ such that $\frac{1}{n}\sum_i y_i y_i^\top = (1 \pm 2^{-d})I$. Suppose further that $\frac{1}{n}\sum_{i=1}^n \langle y_i, v\rangle^4 \leqslant \Delta \|v\|_2^4$ for every $v \in \mathbb{R}^d$. Then, for any $v \in \mathbb{R}^d$, the fraction of y_is such that $|\langle y_i, v\rangle| > \frac{1}{2}$ is at least $\frac{1}{4\Delta}$.*

Proof. Observe that the contribution of y_is such that $|\langle y_i, v\rangle| \leqslant \frac{1}{2}$ to $\frac{1}{n}\sum_i \langle y_i, v\rangle^2 \leqslant \frac{1}{4}$. Thus, by Cauchy-Schwarz inequality, we have:

$$\begin{aligned}\frac{1}{2}\|v\|_2^2 \leqslant \frac{1}{n}\sum_i \langle y_i, v\rangle^2 \mathbf{1}(|\langle y_i, v\rangle| > 1/2) &\leqslant \sqrt{\frac{1}{n}\sum_i \langle y_i, v\rangle^4}\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbf{1}(|\langle y_i, v\rangle| > 1/2)} \\ &\leqslant \sqrt{\Delta}\|v\|_2^2 \sqrt{\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}(|\langle y_i, v\rangle| > 1/2)\right)}.\end{aligned}$$

Rearranging yields that:

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}(|\langle y_i, v\rangle| > 1/2) \geqslant \frac{1}{4\Delta}.$$

□

Proof of Lemma 6.2. Let us analyze Y_2. By Hölder's inequality, $\frac{1}{n}\sum_{i=1}^n \langle y_i, v\rangle^4 \leqslant \left(\frac{1}{n}\sum_{i=1}^n \langle y_i, v\rangle^{2t}\right)^{2/t} \leqslant \frac{1}{\tau^{2/t}}\left(\frac{16Ct}{\alpha'^2}\right)^2 \|v\|_2^4$. Thus, by Lemma 6.4, we must have that at least a $O(\alpha^4)\tau^{2/t}/t^2$ fraction of y_is satisfy $\langle y_i, v\rangle^2 \geqslant \frac{1}{2}$. This fraction, for $t = O(1/\alpha)$ and $\tau = O(\alpha^{2t+11})$ is at least $O(\alpha^{10})$. Thus, $|Y_2| \geqslant O(\alpha^{10})|Y|$ as desired. □

Proof of Lemma 6.3. Observe that $v^\top \Sigma_* v \leqslant v^\top \widehat{\Sigma} v \leqslant \eta$. Thus, using C-subgaussianity of 4th moments of X we have that

$$\mathop{\mathbb{E}}_{x_i \sim X}[\langle x_i, v\rangle^4] \leqslant (2C)^2 (v^\top \Sigma_* v)^2 \leqslant 4C^2\eta^2.$$

So, if a κ fraction of x_i had $\langle x_i, v\rangle^2 \geqslant \frac{1}{2}$ we would have that

$$\left(\frac{1}{2}\right)^2 \kappa \leqslant 4C^2\eta^2$$

which upon rearranging yields that the fraction of x_is such that $\langle x_i, v\rangle^2 \geqslant 1/2$ is at most $16C^2\eta^2$. Thus, Y_1 contains at least $\alpha' - O(\eta^2) = \alpha' - O(\alpha^{12})$ fraction of the points in Y as desired. □

7 List-Decodable Covariance Estimation with Spectral Accuracy

In this section, we put the three components from the previous sections together to obtain an algorithm for list-decodable covariance estimation with multiplicative spectral recovery guarantee.

Algorithm 7.1 (List-Decodable Covariance Estimation with Spectral Recovery Guarantee)**.**

Given: $Y = \{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{Q}^d$ such that $\frac{1}{n}\sum_{i=1}^n y_i y_i^\top = (1 \pm 2^{-d})I$ and $\alpha > 0$.

Output: A list $\mathcal{L}$ of positive semidefinite matrices in $\mathbb{Q}^{d\times d}$.

Operation: Maintain a list $\mathcal{L}'$ of candidates with "witness subsets" of Y. Initialize $\mathcal{L}'$ with $(Y, I, 0)$. During the course of the algorithm, some candidates in $\mathcal{L}'$ will become "final".

1. Set $t_1 = O(1/\alpha)$, $t_2 = 20$, and $\tau = O(\alpha^{2t_1+11})$. For $g = 0, 1, \ldots,$ do:
2. **Process g-th Generation Candidates:** While there is $(Y^{(i)}, \widehat{\Sigma}^{(i)}, g)$ in $\mathcal{L}'$ that is not marked final, remove $(Y^{(i)}, \hat{\Sigma}^{(i)}, g)$ from $\mathcal{L}'$ and run the following steps:
   (a) **Subgaussian Restriction:** Run the Subgaussian Restriction Algorithm (Algorithm 5.5) to find $Y' \subseteq Y^{(i)}$ satisfying
   $\left|\frac{v}{2t}\right. \left\{\mathbb{E}_{y\sim Y'}\langle y, v\rangle^{2t} \leqslant \frac{1}{\tau}\left(\frac{16Ct}{\alpha^2}\right)^t \left(\mathbb{E}_{y\sim Y'}\langle y, v\rangle^2\right)^t\right\}$ for $t = t_1$ and $\tau = \tau$ above.
   (b) **Isotropization:** By a linear transform, ensure that Y' satisfies $\mathbb{E}_{y'\sim Y'}\, y'y'^\top = I$.
   (c) **Coarse Spectral Recovery:** Apply Coarse Spectral Recovery Algorithm (Algorithm 4.3) to input Y' with $t = t_2$, fraction of inliers set to $\alpha/2$ and failure probability $\nu = \alpha^{10}/100$. If the algorithm outputs infeasible, go back to the beginning of the loop. Otherwise:
   (d) **Check Certificate of Spectral Approximation:** For each of the candidates $\hat{\Sigma}_i$s produced, check if the minimum eigenvalue of $\hat{\Sigma}_i$ is at least η for $\eta = 2\alpha^6$. If yes, add $(\hat{\Sigma}_i, Y', g+1)$ to $\mathcal{L}'$ after undoing the isotropic transformation from Step (b) above and label the candidate "final".
   (e) **Apply Spectral Splitting:** For each candidate $\hat{\Sigma}_i$ that is not labeled final, find an eigenvector v of $\hat{\Sigma}_i$ with eigenvalue $< \eta$ and apply Spectral Splitting Algorithm (Algorithm 6.1) with respect to $\hat{\Sigma}_i$ to Y' to obtain Y''. If $|Y''| < \alpha n/2$, reject the

```
            candidate and continue to loop. Otherwise, add $(\hat{\Sigma}_i, Y'', g+1)$ to the list $\mathcal{L}$ after
            undoing the isotropic linear transformation on both $Y_i'$ and $\hat{\Sigma}_i$ from Step 2 above.
    3. When the for loop exits, add all the candidate covariances $\hat{\Sigma}$ from $\mathcal{L}'$ to $\mathcal{L}$.
```

We first explain the idea of the algorithm and the analysis. We start with the input corrupted sample Y along with the 0-th generation candidate I and apply the subgaussian restriction procedure. Intuitively, the goal of the algorithm is to make progress on getting a good candidate covariance in the list as the generations g progress. Theorem 5.2 ensures that in this process, we only increase the density of inliers (i.e. points intersecting with the original good set of points X) and that the number of inliers is at least $(\alpha - \alpha^{10})n$.. We then make the resulting Y approximately isotropic and apply the coarse spectral recovery algorithm to obtain a list of $O(1/\alpha^2)$ size that is guaranteed to contain an $\hat{\Sigma}_i$ such that $\Sigma_* \preceq \hat{\Sigma}_i \preceq \mathrm{poly}(1/\alpha) + \alpha^6 I$ (where we have set $\eta = \alpha^6$). If all eigenvalues of every candidate $\hat{\Sigma}_i$ are at least $2\alpha^6$, then, the coarse guarantee implies a multiplicative spectral approximation (see Lemma 7.7) as we'd like and so we are done (our algorithm labels such candidates "final").

Otherwise, there must be a $\hat{\Sigma}_i$ that has an eigenvalue smaller than $2\alpha^6$. In this case, because Y satisfies certifiable subgaussianity (as a result of our subgaussian restriction subroutine), spectral splitting can prune out $\geqslant O(\alpha^{10})$ fraction of points from Y while removing at most $O(\alpha^{12})n$ points from $X \cap Y$. We call the pruned set Y' a "witness subset" for $\hat{\Sigma}_i$. If $\hat{\Sigma}_i$ happened to be a "candidate good estimate" for the unknown covariance, then, the witness subset is sufficient to work with from this point on in the algorithm. Of course we do not know whether $\hat{\Sigma}_i$ is the "right" candidate and in general, the witness subset is different for different candidates $\hat{\Sigma}_i$. Thus, our algorithm maintains the the current witness subset for each potential candidate $\hat{\Sigma}_i$ in our list. In the subsequent runs of the algorithm, we repeat the algorithm on witness subset for each candidate that is not marked final. At any point of time, each member $(\hat{\Sigma}^{(i)}, Y^{(i)})$ in the list $\mathcal{L}'$ is obtained by a sequence of subgaussian-restrictions, coarse-spectral-recovery and spectral splitting applied to the initial input set Y. And each time a candidate is processed, we must decrease the size of its witness set by at at least $1 - O(\alpha^{10})$ factor while increasing the list size by $O(1/\alpha^{22})$. Since we know that for the "correct candidate", we never throw away more than $\alpha n/2$ inliers, if the size of a witness set drops below $\alpha n/2$, we can comfortably reject the corresponding candidate. Thus the number of generations in the algorithm cannot be more than $\tilde{O}(1/\alpha^{10})$ giving us the bound on the list-size.

We now formally argue the guarantees of the algorithm. The following theorem summarizes the guarantees of Algorithm 7.1.

Theorem 7.2. *Let $1 \geqslant \alpha > 0$. Suppose $X = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{Q}^d$ is a* good set *(Definition 4.1) of n points satisfying $\mathbb{E}_{x \sim X} xx^\top = \Sigma_*$ such that $|Y \cap X| \geqslant \alpha n$. Let $Y \subseteq \mathbb{Q}^d$ be a set of n points such that $\frac{1}{n}\sum_i y_i y_i^\top = I$. Then, Algorithm 7.1 on input Y, 1) runs in time $(Bn/\alpha)^{O(1/\alpha^{12})}$, 2) outputs a list $\mathcal{L}$ of size $\alpha^{\tilde{O}(1/\alpha^6)}$, with the guarantee that with probability at least* 0.99 *only over the randomness of the algorithm, 3) $\mathcal{L}$ contains a $\hat{\Sigma}$ satisfying $\Sigma_* \preceq \hat{\Sigma} \preceq O(1/\alpha^{150})\Sigma_*$.*

We prove the theorem in the following sequence of lemmas.

The first set of claims below analyze the size of the list $\mathcal{L}$ output by the algorithm.

Lemma 7.3 (Size of witness sets in g-th generation)**.** *Let $(\hat{\Sigma}_i, Y^{(i)}, g)$ be a g-th generation candidate that is not marked final. Then, $|Y^{(i)}| \leqslant (1 - O(\alpha^{10}))^g n$.*

Proof. We prove this by induction on the generation iterator g. For the base case, observe that at the beginning, $g = 0$ and $|Y^{(0)}| = |Y| = n$. Next, for the inductive case, observe that a g-th generation candidate for $g \geqslant 1$ is obtained by taking a 1) $g-1$-th generation candidate $(\hat{\Sigma}_i, Y^{(i)})$, 2) applying subgaussian restriction to $Y^{(i)}$, applying coarse-spectral recovery to obtain a list of g-th generation candidates by applying Algorithm 4.3 to input $Y^{(i)}$. By Theorem 5.2, $Y^{(i)}$ satisfies

$$\Big|\frac{v}{O(t)} \left\{ \mathop{\mathbb{E}}_{y \sim Y^{(i)}} \langle y, v \rangle^{2t} \leqslant \frac{1}{\tau} \left(\frac{16Ct}{\alpha^2} \right)^t \left(\mathop{\mathbb{E}}_{y \sim Y^{(i)}} \langle y, v \rangle^2 \right)^t \right\} .$$

If a g-th generation candidate covariance $\hat{\Sigma}_j$ is not marked final, then, there must be an unit length eigenvector v of $\hat{\Sigma}_j$ with an eigenvalue of at most $\eta = \alpha^6 \mathbb{E}_{y \sim Y^{(i)}} \langle y, v \rangle^2$. Thus, the assumptions of Lemma 6.2 are met and the splitting algo must prune away at least $O(\alpha^{10})|Y^{(i)}|$ points from $Y^{(i)}$ before producing a g-th generation candidate $(\hat{\Sigma}_j, Y^{(j)}, g)$. This completes the proof. □

Lemma 7.4 (Bounding the Number of Generations). *The maximum value of g during the run of the algorithm is $O(\log 1/\alpha)/\alpha^{10}$.*

Proof. From Lemma 7.3, the size of the witness set drops as $(1 - O(\alpha^{10}))^g$ in the g-th generation. If $g > O(\log 1/\alpha)/\alpha^{10}$, then, the above size is $\leqslant \alpha n/2$ in which case, Step 2(e) of the algorithm exits the loop disallowing further generations. This completes the proof. □

As an immediate corollary, we obtain a bound on the size of the list obtained by the algorithm above:

Lemma 7.5 (List Size Bound). *The size of the list $\mathcal{L}$ of covariances output by the algorithm is at most $\alpha^{\tilde{O}(1/\alpha^{10})}$.*

Proof. Every candidate in the list $\mathcal{L}$ corresponds to a candidate from $\mathcal{L}'$ that the algorithm marks "final". Each candidate in $\mathcal{L}'$ marked final is at most of $O(\log 1/\alpha)/\alpha^{10}$ generation from Lemma 7.4. Each g-th generation candidate in $\mathcal{L}'$ produces at most $O(1/\alpha^{20})$ $g+1$-th generation candidates from the guarantees of Theorem 4.2. This immediately yields the upper bound on the list size as desired. □

Next, we bound the running time of the algorithm.

Lemma 7.6 (Running Time). *The running time of Algorithm 7.1 is $(Bn/\alpha)^{O(1/\alpha^{12})}$.*

Proof. Given our parameters, the running time of each iteration is dominated by the running time of coarse spectral recovery. The number of iterations is upper bounded by $\alpha^{-\tilde{O}(1/\alpha^{11})}$ by an argument similar to the one bounding the size of the list output by the algorithm. This gives the final running time bound as desired. □

Finally, we prove the correctness – that one of the candidates in the list gives a multiplicative approximation to the unknown covariance. Our proof will rely on the following simple observation that we will use to infer that if all eigenvalues of every candidate $\hat{\Sigma}$ are not too small relative to its witness set then the set of candidates must contain a multiplicative spectral approximation to the unknown covariance Σ_*.

Lemma 7.7 (Certificates of Spectral Recovery)**.** *Suppose $\widehat{\Sigma}$ satisfies $\Sigma_* \preceq \widehat{\Sigma} \preceq O\left(\frac{1}{\alpha^{150}}\right)\Sigma_* + \eta I$. Further, suppose that for all unit vectors v, $v^\top \widehat{\Sigma} v > 2\eta$. Then,*

$$\Sigma_* \preceq \widehat{\Sigma} \preceq O\left(\frac{1}{\alpha^{150}}\right)\Sigma_* \,.$$

Proof. Let v be a unit vector such that $\lambda = v^\top \widehat{\Sigma} v > 2\eta$. Then, we have:

$$\frac{v^\top \Sigma v}{v^\top \Sigma_* v} \leqslant O\left(\frac{\lambda}{(\lambda-\eta)\alpha^{150}}\right) \leqslant O\left(1/\alpha^{150}\right) \,.$$

This completes the proof. □

Lemma 7.8 (Correctness)**.** *Let $1 \geqslant \alpha > 0$. Suppose $X = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{Q}^d$ is a* good *set of n points satisfying $\mathbb{E}_{x\sim X}\, xx^\top = \Sigma_*$ such that $|Y \cap X| \geqslant \alpha n$. Let $Y \subseteq \mathbb{Q}^d$ be a set of n points such that $\frac{1}{n}\sum_i y_i y_i^\top = I$. Then, Algorithm 7.1 on input Y outputs a list of $\alpha^{-\tilde{O}(1/\alpha^{10})}$ covariance matrices such that there is a candidate $\hat{\Sigma}$ in the list satisfying:*

$$\Sigma_* \preceq \hat{\Sigma} \preceq O\left(\frac{1}{\alpha^{150}}\right)\Sigma_* \,. \tag{7.1}$$

Proof. Let us call a $g+1$-th generation candidate $(\hat{\Sigma}_j, Y^{(j)}, g)$ good if it satisfies 1) $\Sigma_* \preceq \hat{\Sigma}_i \preceq O(1/\alpha^{150})\Sigma_* + O(\alpha^{16})\,\mathbb{E}_{y\sim Y^{(i)}}\, yy^\top$, 2) $|Y^{(i)} \cap X| \geqslant (\alpha - g \cdot O(\alpha^{10}))n$, 3) $|Y^{(i)}| \leqslant (1 - O(\alpha^{10}))^g n$. Here $(\hat{\Sigma}_i, Y^{(i)}, g)$ is the g-th generation candidate processing of which generated $\hat{\Sigma}_j$ as a candidate in the g-th iteration of the while loop in Algorithm 7.1.

We will prove the following by induction on g: suppose there is a g-th generation candidate that is good and not marked final. Then, there is a $g+1$th generation candidate that is good.

We first observe that this claim is enough to complete the proof. To see why, observe that the number of generations g is no more than $O(\log 1/\alpha)/\alpha^{10}$. So $|Y^{(i)} \cap X| \geqslant (\alpha/2)n$ for all g encountered in the run of the algorithm. Now consider a good candidate $(\hat{\Sigma}_i, Y^{(i)}, 1)$ in generation $g = 1$ and let's track the sequence of good candidates guaranteed by the inductive claim above for each of the generations $g > 1$ starting with $(\hat{\Sigma}_i, Y^{(i)}, 1)$. Let g_* be the largest g such that the good candidate in generation g is not marked final. Since $g_* \leqslant \tilde{O}(1/\alpha^{10})$ and $\nu = \alpha^{10}/100$ and the assumptions on such $Y^{(i)}$ for Algorithm 4.3 succeeding are met, each run of of Algorithm 4.3 along such a path succeed with probability at least $1-\nu$. By a union bound, and that $g_* \leqslant \tilde{O}(1/\alpha^{10})$, all the runs succeed with probability at least 0.99. Let's condition on this event. Then, in iteration g_*, starting with such a good candidate, the coarse spectral recovery algorithm must produce a generation $g_* + 1$ candidate that is marked final. In which case, we must have that all eigenvalues of $\hat{\Sigma}_j$ are at least $\eta = O(\alpha^6)$ relative to $\mathbb{E}_{y\sim Y^{(j)}}\, yy^\top$. Since $|Y^{(j)} \cap X| \geqslant \alpha/2$, by Lemma 7.7, we can conclude that $\hat{\Sigma}_j \preceq O(1/\alpha^{150})\Sigma_*$. This completes the proof modulo the inductive claim.

Let us now prove the inductive claim to finish the proof.

For the base case, observe that the first iteration runs with the only 0-th generation candidate in the list $\mathcal{L}'$, namely, $(\frac{1}{\alpha^2}I, Y, 0)$ and $|Y \cap X| \geqslant \alpha n$. Let us now analyze the steps of the algorithm when $Y^{(i)} = Y$ is processed in $g = 1$st iteration. In the first (subgaussian restriction) step, we apply Algorithm 5.2 which, from Lemma 5.2, allows us to obtain a $Y' \subseteq Y$ such that Y' satisfies certifiable subgaussianity and Y' satisfies $|Y' \cap X| \geqslant (\alpha - \alpha^{10})n$. Thus, our coarse spectral recovery algorithm (Algorithm 4.3) gets input a $(1-\alpha')$ corrupted sample for $\alpha' \geqslant \alpha - \alpha^{10}$ fraction of Y and as a result of Theorem 4.2, with probability at least $1-\nu$, returns a list of candidates one of which, say $\hat{\Sigma}_j$, satisfies the first conclusion of the lemma. If this candidate is not marked final, then, by Lemma 6.3, we know that $(\hat{\Sigma}_j, Y^{(j)})$ is added to $\mathcal{L}'$ with $Y^{(j)}$ satisfying (2) and (3) as desired. The analysis of the inductive case is entirely analogous. □

8 List-Decodable Mean and Covariance Estimation

In this section, we prove that given a good spectral estimate of the covariance (i.e., guaranteed by Theorem 7.2) and an $(1-\alpha)$-corruption of a good set of points X, we can obtain a list of $O(1/\alpha^{\mathrm{poly}(1/\alpha)})$-candidates that contains an estimate of the mean that is accurate within $\mathrm{poly}(1/\alpha)$-Mahalanobis distance and covariance that is accurate in (stronger) relative Frobenius distance. The stronger guarantee immediately implies our main theorem on list-decoding mean and covariance for Gaussian distributions with a total variation error guarantee.

We first start by describing the strong relative Frobenius error guarantee for covariance estimation.

8.1 Covariance Recovery in Relative Frobenius Error

Our algorithm builds on ideas in the prior work [BK21] on list-decodable subspace recovery (which can be thought of as the special case where the unknown covariance is allowed to have eigenvalues that are either 0 or 1).

Our notion of *good* set for Frobenius error guarantee (under additional spectral closeness hypothesis) is significantly weaker:

Definition 8.1 (Good set for relative Frobenius Recovery for known spectral approximation)**.** We say that a subset $X \subseteq \mathbb{R}^d$ is a C-good with mean $\mathbb{E}_{x \sim X} x = \mu_*$ and 2nd moment $\mathbb{E}_{x\sim X} xx^\top = \Sigma_*$ if 1) $\mu_* \mu_*^\top \preceq 0.1\, \mathbb{E}_{x \sim X}(x-\mu_*)(x-\mu_*)^\top$ and 2) X has $O(1)$-certifiably C-hypercontractive degree 2 polynomials.

Theorem 8.2 (List-decoding covariances with relative Frobenius error guarantee)**.** *There is an algorithm that takes input $Y \subseteq \mathbb{R}^d$ of size n, runs in time $n^{O(1)}$ and outputs a list of PSD matrices $\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_k$ for $k = O(1/\alpha^4)$ with the following guarantees. Let X be a C-*good *set of n points in $\mathbb{R}^d$ such that $\mathbb{E}_{x\sim X} xx^\top = \Sigma_*$ satisfying $I \preceq \Sigma_* = \mathbb{E}_{x \sim X} xx^\top \preceq O(1/\alpha^{150})I$. Suppose Y be an $(1-\alpha)$-corruption of X, i.e., $Y \subseteq \mathbb{R}^d$ of size n satisfying $|Y \cap X| = \alpha n$. Then, there is an i such that $\left\| \Sigma_*^{-1/2}(\widehat{\Sigma}_i - \Sigma_*)\Sigma_*^{-1/2} \right\|_F^2 \leqslant O(1/\alpha^{304})$.*

Algorithm Our algorithm solves the SoS relaxation of the constraints $\mathcal{A}_1' \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4$ defined in Section 4 – we have dropped the anti-concentration constraints $\mathcal{A}_5$ and will modify $\mathcal{A}_1$ to replace the third constraint with $(\Sigma - I) = VV^\top$ and $(O(\frac{1}{\alpha^{150}})I - \Sigma) = ZZ^\top$ for matrix valued indeterminates V and Z that encode the additional information that $I \preceq \Sigma \preceq O(1/\alpha^{150})I$.

The proof of Lemma 4.8 extends to show that setting $X' = X$ and w_i to be the indicator of i such that $y_i = x_i$ along with appropriate values to U, V, Z gives a feasible solution to the polynomial constraints $\mathcal{A}_1' \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4$.

Our full algorithm is as follows:

1. Find a pseudo-distribution $\tilde{\zeta}$ of degree $O(1)$ consistent with $\mathcal{A} = \mathcal{A}_1' \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4$ and minimizing $\left\|\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w]\right\|_2^2$.
2. **Rounding:** Repeat $O(1/\alpha^4)$ times:
   (a) choose (i_i, i_2) with probability proportional to $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2}]$.
   (b) Output $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2} \Sigma]/\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2}]$.

The main claim in our analysis is the following lemma (analogous to Lemma 4.6).

Lemma 8.3 (Deriving Frobenius Error Bounds within Low-Degree SoS)**.**

$$\mathcal{A} \left|\frac{w,\Sigma}{O(1)}\right. \left\{ w(X')^2 \left\| \Sigma_*^{-1/2} (\Sigma - \Sigma_*) \Sigma_*^{-1/2} \right\|_F^2 \leqslant O(1/\alpha^{300}) \right\}. \tag{8.1}$$

Proof. From the conclusion of Lemma 4.7 (with $h = 1$):

$$\mathcal{A}_5 \left|\frac{\Sigma,R,Q,w}{O(1)}\right. \left\{ w(X')^4 \left\langle \Sigma_*^{-1/2}(\Sigma - \Sigma_*)\Sigma_*^{-1/2}, Q \right\rangle^4 = w(X')^4 \left\langle \Sigma - \Sigma_*, \Sigma_*^{-1/2} Q \Sigma_*^{-1/2} \right\rangle^4 \right.$$
$$\left. \leqslant O(1) w(X')^4 \left(\|Q\|_F^4 + \left\| R \Sigma_*^{-1/2} Q \Sigma_*^{-1/2} R \right\|_F^4 \right) \right\}. \tag{8.2}$$

For the second term, we start from $\mathcal{A}_1'$ using that $I \preceq \Sigma_*$ and thus $\left\|\Sigma_*^{-1/2}\right\|_2 \leqslant 1$, we must have:

$$\mathcal{A} \left|\frac{R,w}{O(1)}\right. \left\{ w(X')^2 (v^\top \Sigma_*^{-1/2} R^\top R \Sigma_*^{-1/2} v)^2 = w(X')^2 (v^\top \Sigma_*^{-1/2} \Sigma \Sigma_*^{-1/2} v)^2 \right.$$
$$\left. \leqslant O(1/\alpha^{600}) w(X')^2 \left\| \Sigma_*^{-1/2} v \right\|_2^4 \leqslant O(1/\alpha^{600}) w(X')^2 \|v\|_2^4 \right\}. \tag{8.3}$$

Using Contraction within SoS (Fact 3.15) with $\beta = w(X')^2$, $A = R$, $t = 2$, and $\Delta = O(1/\alpha^{600})$ twice:

$$\mathcal{A} \left|\frac{R,w,Q}{O(h)}\right. \left\{ w(X')^4 \left\| R \Sigma_*^{-1/2} Q \Sigma_*^{-1/2} R \right\|_F^4 \leqslant O(1/\alpha^{600}) \|Q\|_F^4 \right\} \tag{8.4}$$

Plugging back the estimate from (4.21) in (4.18) gives:

$$\mathcal{A}_5 \left|\frac{\Sigma,w,Q}{O(1)}\right. \left\{ w(X')^4 \left\langle \Sigma_*^{-1/2}(\Sigma - \Sigma_*)\Sigma_*^{-1/2}, Q \right\rangle^4 \leqslant O(1/\alpha^{600}) \|Q\|_F^4 \right\}. \tag{8.5}$$

Substituting $Q = \Sigma_*^{-1/2}(\Sigma - \Sigma_*)\Sigma_*^{-1/2}$ and multiplying by the SoS polynomial $w(X')^4$ yields:

$$\mathcal{A}_5 \,\Big|\!\frac{\Sigma,w}{O(1}\, \left\{ w(X')^8 \left\| \Sigma_*^{-1/2}(\Sigma - \Sigma_*)\Sigma_*^{-1/2} \right\|_F^8 \leqslant w(X')^4 \cdot O(1/\alpha^{600}) \left\| \Sigma_*^{-1/2}(\Sigma - \Sigma_*)\Sigma_*^{-1/2} \right\|_F^4 \right\} . \tag{8.6}$$

We now apply Lemma 3.13 (Cancellation within SoS) with $A = w(X')^4 \left\|\Sigma - \Sigma_*\right\|_F^4$ to obtain that:

$$\mathcal{A} \,\Big|\!\frac{w,\Sigma}{O(1)} \left\{ w(X')^{16} \left\| (\Sigma_*^{-1/2}\Sigma - \Sigma_*)\Sigma_*^{-1/2} \right\|_F^{16} \leqslant O(1/\alpha^{2400}) \right\} . \tag{8.7}$$

We finally apply Cancellation with Constant RHS (Lemma 3.12) to conclude that:

$$\mathcal{A} \,\Big|\!\frac{w,\Sigma}{O(1)} \left\{ w(X')^2 \left\| \Sigma_*^{-1/2}(\Sigma - \Sigma_*)\Sigma_*^{-1/2} \right\|_F^2 \leqslant O(1/\alpha^{300}) \right\} . \tag{8.8}$$

□

Proof. Observe that by Lemma 3.16, we know that $I \preceq \frac{\widetilde{\mathbb{E}}[w_{i_1} w_{i_2} \Sigma_*]}{\widetilde{\mathbb{E}}[w_{i_1} w_{i_2}]} \preceq O(1/\alpha^{150}) I$. Next, let $G \subseteq [n]$ be the set of indices i such that $x_i = y_i$ (unknown to the algorithm). Taking pseudo-expectations with respect to $\tilde{\zeta}$ of the conclusion of Lemma 8.3, we obtain that:

$$\sum_{i_1, i_2 \in G} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2} \left\| \Sigma_*^{-1/2}(\Sigma - \Sigma_*)\Sigma_*^{-1/2} \right\|_F^2 \leqslant O(1/\alpha^{300})] .$$

Dividing both sides by $\widetilde{\mathbb{E}}[w(X')^2]$, using that $\sum_{i_1, i_2 \in G} \widetilde{\mathbb{E}}[w_{i_1} w_{i_2}] = \widetilde{\mathbb{E}}[w(X')^2]$ and the conclusion of Lemma 4.9 along with Cauchy-Schwarz inequality for pseudo-distributions that yields that $\widetilde{\mathbb{E}}[w(X')^2] \geqslant \widetilde{\mathbb{E}}[w(X')]^2 \geqslant \alpha^4$, we obtain:

$$\frac{1}{\widetilde{\mathbb{E}}[w(X')^2]} \sum_{i_1, i_2 \in G} \widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2}] \left\| \Sigma_*^{-1/2} \left(\frac{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2} \Sigma]}{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2}]} - \Sigma_* \right) \Sigma_*^{-1/2} \right\|_F^2 \leqslant O(1/\alpha^{304}) .$$

The left hand side can now be interpreted as expectation over the choice of i_1, i_2 with probability proportional to $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2}]$ conditioned on $i_1, i_2, \in G$. By Markov's inequality, with probability at least 0.99, a draw from this distribution of (i_1, i_2) must satisfy:

$$\left\| \Sigma_*^{-1/2} \left(\frac{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2} \Sigma]}{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2}]} - \Sigma_* \right) \Sigma_*^{-1/2} \right\|_F^2 \leqslant O(1/\alpha^{304}) .$$

Further, since $\widetilde{\mathbb{E}}[w(X')] \geqslant \alpha^2$, the probability that (i_1, i_2) chosen with probability proportional to $\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2}]$ satisfy that $i_1, i_2 \in G$ is at least α^4. Thus, altogether, the chance that a random draw of (i_1, i_2) yields an estimate $\frac{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2} \Sigma]}{\widetilde{\mathbb{E}}_{\tilde{\zeta}}[w_{i_1} w_{i_2}]}$ satisfying the relative Frobenius error bound above is at least $O(\alpha^4)$. Thus, repeating the sampling process $O(1/\alpha^4)$ times is sufficient to ensure that the list contains a candidate close in relative Frobenius distance as desired with probability at least 0.99.

□

8.2 Mean Estimation Given Spectral Approximation to Covariance

Given a multiplicative spectral approximation to the covariance, one can apply algorithms from prior works (or a significantly simpler variant of our list-decoding algorithm for covariance estimation above) to obtain good estimates of the mean. We only note the consequence here.

Technically speaking, Theorem 1.2 in [KS17a] works for the weaker model of list-decodable estimation where Y must contain as a subset αn i.i.d. points from a certifiably subgaussian distribution D. For our stronger model, we observe that any subst of size αn of a good set (Definition 4.1) satisfies $O(C/\alpha)$-certifiable subgaussianity. To do this, we only need the following basic observation (applied to centered version of a good set):

Lemma 8.4. *Suppose $X \subseteq \mathbb{Q}^d$ is a set of n points with mean 0 such that $\mathbb{P}_{x\sim X}[|\langle x, v\rangle| \leqslant \alpha/100 \mathbb{E}_{x\sim X}\langle x, v\rangle^2] \leqslant \alpha/2$. Let $Z \subseteq X$ be any subset of X of size αn. Then, $\mathbb{E}_{z\in Z} zz^\top \succeq (\alpha/200) \mathbb{E}_{x\sim X} xx^\top$.*

Proof. Fix $v \in \mathbb{R}^d$. Apply the anti-concentration of the set X to conclude that at most $\alpha n/2$ points in X can satisfy $\langle x, v\rangle^2 \leqslant \alpha/100 \mathbb{E}_{x\sim X}\langle x, v\rangle^2$. Thus, at least $\alpha n/2$ points in Z satisfy $\langle x, v\rangle^2 > \alpha/100 \mathbb{E}_{x\sim X}\langle x, v\rangle^2$ and as a result, $\mathbb{E}_{z\sim Z} zz^\top \succeq \alpha/100 \mathbb{E}_{x\sim X} xx^\top$. □

The above lemma immediately yields that for any subset $Z \subseteq X$ of size at least αn, whenever X is $2t$-certifiably C-subgaussian, Z itself is $2t$-certifiably $O(C/\alpha)$-subgaussian:

$$\vdash_{2t} \left\{ \mathop{\mathbb{E}}_{z\sim Z}\langle z, v\rangle^{2t} \leqslant \frac{1}{\alpha} \mathop{\mathbb{E}}_{x\sim X}\langle x, v\rangle^{2t} \leqslant \frac{1}{\alpha}(Ct)^t \left(\mathop{\mathbb{E}}_{x\sim X}\langle x, v\rangle^2\right)^t \leqslant \frac{1}{\alpha}(200Ct/\alpha)^t \left(\mathop{\mathbb{E}}_{z\sim Z}\langle z, v\rangle^2\right)^t \right\}.$$

We can thus apply Theorem 1.2 in [KS17a] (at the cost of loss of an additional factor of $O(1/\alpha)$ in the error).

Theorem 8.5 (List-decoding mean estimation given spectral approximation to covariance, Theorem 1.2 in [KS17a])**.** *There is an algorithm that takes input $Y \subseteq \mathbb{R}^d$ of size n, runs in time $n^{O(\log(1/\alpha))}$ and outputs a list of d-dimensional vectors $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k$ for $k = O(1/\alpha)$ with the following guarantees. Let X be a C-*good *set of n points in $\mathbb{R}^d$ such that $\mathbb{E}_{x\sim X} x = \mu_*$ and $\mathbb{E}_{x\sim X} xx^\top = \Sigma_*$ satisfying $I \preceq \Sigma_* = \mathbb{E}_{x\sim X} xx^\top \preceq O(1/\alpha^{150})I$. Suppose Y be an $(1-\alpha)$-corruption of X, i.e., $Y \subseteq \mathbb{R}^d$ of size n satisfying $|Y \cap X| = \alpha n$. Then, there is an i such that for every $u \in \mathbb{R}^d$, $\left\|\widehat{\mu}_i - \mu_*, u\right\|_2^2 \leqslant O(\log(1/\alpha)/\alpha^{152})u^\top\Sigma_* u^2$.*

8.3 Proof of Main Theorem

We now have all the components to prove Theorem 1.6.

Theorem 8.6 (Main Theorem)**.** *Fix $\alpha > 0$. For any B, there is a $(Bn)^{\tilde{O}(1/\alpha^{12})}$ time algorithm that takes input a collection of n points $Y \subseteq \mathbb{Q}^d$ with entries of bit-complexity at most* poly(Bd) *and either "rejects" or outputs a list of parameters $\{(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i\leqslant k}$ for $k = \alpha^{-\mathrm{poly}(1/\alpha)}$ with the following guarantee: suppose that for some absolute constant $C > 0$, D is a distribution on $\mathbb{R}^d$ with mean μ_* and covariance Σ_* with rational entries of bit complexity $\leqslant B$ that is 1) $s(\delta)$-certifiably (C, δ)-anti-concentrated for $\delta = O(\alpha^3)$, $s(\delta) = O(1/\delta^2)$ and*

2) has $2t$-certifiable C-hypercontractive degree 2 polynomials for $t \geqslant O(1/\alpha)$. Suppose Y is a $\mathrm{poly}(d)$*-bit rational truncation of an ε-corrupted sample from D of size $n \geqslant n_0 = d^{\tilde{O}(1/\alpha^6)}$.*

Then, with probability at least 0.99 *over the draw of X and over the random choices of the algorithm, the algorithm does not reject and outputs a list of parameters of size k such that there exists an i such that for every $u \in \mathbb{R}^d$:*

$$\langle \hat{\mu}_i - \mu_*, u \rangle \leqslant \tilde{O}(\log(1/\alpha)/\alpha^{152})\sqrt{u^\top \Sigma_* u}\,,$$

$$\Sigma_* \preceq \hat{\Sigma} \preceq O(1/\alpha^{152})\Sigma_*\,,$$

and,

$$\left\| \Sigma_*^{\dagger/2}(\hat{\Sigma} - \Sigma_*)\Sigma_*^{\dagger/2} \right\|_F \leqslant O(1/\alpha^{304})\,.$$

Combined with the characterization of total variation distance in terms of the three parameter distance bounds (Proposition A.1 in [BK20b]), we immediately obtain Theorem 1.6.

Corollary 8.7 (List-decoding Gaussian with total variation error guarantee)**.** *There is a $n^{\mathrm{poly}(1/\alpha)}$ time algorithm that takes input a $(1-\alpha)$-corrupted sample of size $n \geqslant d^{\alpha^{-O(1)}}$ from a d-dimensional Gaussian distribution with mean μ_* and covariance Σ_* and outputs a list of $2^{O(1/\alpha^{O(1)})}$-parameters such that there is a $(\hat{\mu}, \hat{\Sigma})$ in the list satisfying:*

$$d_{\mathsf{TV}}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu_*, \Sigma_*)) \leqslant 1 - \exp(-O(1/\alpha^{304}))\,.$$

Proof of Theorem 8.6. Let X be an i.i.d. sample from D and let $\tilde{X}$ is a truncation of D to $\mathrm{poly}(Bd)$ bits. If $\Sigma_* \succeq 2^{-\mathrm{poly}(Bd)}I$, then, Fact 3.23 shows that $\tilde{X}$ has small mean (i.e. $\mathbb{E}_{x\sim\tilde{X}}\, x = \tilde{\mu}$ such that $\tilde{\mu}\tilde{\mu}^\top \preceq 0.1\,\mathbb{E}_{x\sim\tilde{X}}(x-\tilde{\mu})(x-\tilde{\mu})^\top$), covariance in $[0.99, 1.01]\cdot\Sigma_*$, satisfies the properties of good set required in Definition 4.1.

Our algorithm works in three steps.

In the first step, we list-decode covariances with spectral guarantee using the Algorithm from Theorem 7.2. Observe that Theorem 7.2 requires that input sample be from a *small-mean* good set X. To meet these guarantees, we work with the "pairwise difference" version of Y.

Assume n is even, randomly permute the points in Y (and assume, for the sake of simplicity that $y_1, y_2, \ldots, y_n$ is the permuted version) and let Y' be the set of $n/2$ points $\frac{y_1-y_2}{\sqrt{2}}, \frac{y_3-y_4}{\sqrt{2}}, \ldots \frac{y_{n-1}-y_n}{\sqrt{2}}$. Then, observe that with probability at least $1-1/n$ over the choice of the random permutation, there are at least $\alpha^2 n/2$ pairs $(2i-1, 2i)$ such that y_{2i-1}, y_{2i} are both in the intersection $Y \cap X$. Without loss of generality, let's say that $y_{2i-1} = x_{2i-1}$ and $y_{2i} = x_{2i}$. Thus, Y' can be thought of as an $(1-\alpha^2/2)$-corruption of randomly paired and $\frac{1}{\sqrt{2}}$-scaled differences, say, $\tilde{X}$ from X. Thus, Theorem 7.2 guarantees that with probability at least 0.99, there is an element in the list of size $\alpha^{-\mathrm{poly}(1/\alpha)}$ output by it that contains a PSD matrix $\hat{\Sigma}$ that satisfies $\Sigma_* \preceq \hat{\Sigma} \preceq O(1/\alpha^{150})\Sigma_*$.

In the 2nd step, we take $\alpha^{-150}\hat{\Sigma}_i$ for each element $\hat{\Sigma}_i$ of the list obtained in the first step and transform Y' by applying the linear transformation $y_i' \to \hat{\Sigma}_i^{-1/2}$. When applied to $\tilde{X}$, this transformation ensures that the covariance of $\tilde{X}$ is sandwiched (in Löwner order) between I and $O(1/\alpha^{150})$. Further, since $\tilde{X}$ satisfies the properties of a good set in Definition 4.1 and certifiable hypercontractivity of degree 2 polynomials is invariant under linear transformation, the linearly transformed $\tilde{X}$ continues to satisfy the requirements of Definition 8.1. Running the algorithm from

Theorem 8.2 for each possible candidate $\hat{\Sigma}_i$ from the list obtained in the first step enlarges the list by a factor of $O(1/\alpha^4)$ and when starting with a good candidate $\hat{\Sigma}_i$ from the first step, Theorem 8.2 guarantees with probability at least 0.99, that there is a candidate $\hat{\Sigma}_j$ in the enlarged list satisfying $\left\|\Sigma_*^{-1/2}\hat{\Sigma}_j\Sigma_*^{-1/2} - I\right\|_F^2 \leqslant O(1/\alpha^{304})$.

In the final step, we take $\alpha^{-150}\hat{\Sigma}_i$ for each element $\hat{\Sigma}_i$ of the list obtained in the first step (we do not need Frobenius guarantees for mean estimation) and transform Y' by applying the linear transformation $y_i' \to \hat{\Sigma}_i^{-1/2}$ and observe that by the same argument as in the analysis of the 2nd step above, the assumptions of Fact 8.5 are met and thus, we obtain an list with estimates of means of size $O(1/\alpha)$ factor larger than the one in Step 1 guaranteed with probability at least 0.99 to contain a candidate $\hat{\mu}_i$ satisfying $\langle \hat{\mu}_i - \mu_*, u\rangle \leqslant O(\log(1/\alpha)/\alpha^{152})\sqrt{u^\top \Sigma_* u}$ for every $u \in \mathbb{R}^d$.

All 3 steps succeed with probability at least 0.9 by a union bound. Returning every possible paired combination of the covariances from Step 2 and means from step 1 then satisfies the requirements of the theorem with a list of size $2^{\mathrm{poly}(1/\alpha)}$.

□

9 Applications

In this section, we derive improved algorithms for list-decodable linear regression, subspace recovery and clustering of non-spherical mixtures as immediate consequences of our algorithm for list-decodable covariance estimation.

9.1 Linear Regression

For list-decodable linear regression, given a accuracy parameter η, the best known prior works [KKK19, RY19] obtain a list containing an η-accurate estimate of the unknown vector with a running time of $n^{O(1/(\eta\alpha)^4)}$ and sample complexity $d^{O(1/(\eta\alpha)^4)}$. An error reduction technique from [BK21] allows improving both the exponents to $O(\log 1/\eta)/\alpha^4$. But all these bounds depend exponentially on the target accuracy η. As a result whenever $\eta \to 0$ as $d \to \infty$, the sample complexity and the running time are both super-polynomial in the underlying dimension d.

Our list-decodable covariance estimation algorithm allows obtaining the first *exact* algorithm for list-decodable linear regression. As a consequence, we can obtain an error of η time $n^{\tilde{O}(1/\alpha^{12})}$ poly $\log(1/\eta)$ and sample complexity $d^{\tilde{O}(1/\alpha^6)}$. In particular, the sample complexity does not depend on the target accuracy and the running time scales polylogarithmically in $1/\eta$. As a result, our algorithm allows obtaining exponentially small error in the underlying dimension d in polynomial time. The size of the list recovered, while still an absolute constant depending only on α, is larger and grows as $\alpha^{-\mathrm{poly}(1/\alpha)}$. Our algorithm works without knowing (any upper bound on) the length of the unknown vector ℓ_* and extends easily to the setting of unknown arbitrary non-spherical covariance (if we simply apply our list-decodable covariance estimation as a preprocessing step). It also succeeds in the strong contamination model for list-decodable learning as opposed to the additive model studied in the prior works.

Corollary 9.1. *For every $\eta > 0$, there is an algorithm that takes input a collection of n equations $Y \subseteq \mathbb{Q}^d \times \mathbb{Q}$, runs in time $n^{\tilde{O}(1/\alpha^{12})} \operatorname{poly}\log(1/\eta)$ and either outputs "reject" or a list $\mathcal{L}$ of size $k = \alpha^{-\operatorname{poly}(1/\alpha)}$ of candidate vectors $\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_k$ with the following guarantees: suppose X is a set of $n \geqslant n_0 = O(d^{O(1/\alpha^6)}/\alpha)$ linear equations $\langle a, \ell_* \rangle = b$ where ℓ_* is an unknown arbitrary vector and $a \sim D$ on $\mathbb{R}^d$ such that D has mean 0, a full-rank covariance Σ_*, is $s(\delta)$-certifiably (C, δ)-anti-concentrated and has $2t$-certifiable C-hypercontractivity of degree 2 polynomials for all t. Suppose Y is an $(1-\alpha)$-corruption of X. Then, with probability at least 0.99 over the draw of X and the random choices of the algorithm, the algorithm does not reject and outputs a list that contains a candidate $\hat{\ell}_k$ satisfying:*

$$\left\| \Sigma_*^{-1/2} (\hat{\ell}_k - \ell_*) \right\|_2 \leqslant \eta \, .$$

Proof Sketch. We observe that if $(a, b) \in \mathbb{Q}^d \times \mathbb{Q}$ are the coefficient vectors and "right-hand-sides" of the equations in X, then, the distribution D' of (a, b) satisfies the conditions in Definition 4.3. Further, the covariance of D' has rank exactly d (in $d+1$ dimensional ambient space) with the kernel in the direction $(\ell_*, -1)$. Thus, finding a multiplicative spectral approximation to the covariance of D' and using the kernel of the estimate to obtain $\hat{\ell}$ immediately gives the required guarantee.

Observe that we only pay (and poly logarithmically so) in the running time for the target accuracy. There is no cost in sample complexity as a function of the target accuracy. □

9.2 Subspace Recovery

A similar argument also upgrades the guarantees for list-decodable subspace recovery obtained in prior works. The best known prior work [BK21] obtained an algorithm that runs in fixed polynomial time (exponent independent of α) and gets an Frobenius estimation error of $O(1/\alpha)$. The independent work [RY20b] obtains a worse error guarantee that grows as $\sqrt{r}$ (where r is the dimension of the unknown subspace). In particular, for obtaining an arbitrary target error $\eta > 0$, the algorithm from [BK21] runs in time $n^{\log(1/\alpha\eta) O(1/\alpha^4)}$ that is superpolynomial for any $\eta \to 0$.

Our result below obtains an algorithm that runs in time $n^{\tilde{O}(1/\alpha^{12})} \operatorname{poly}\log(1/\eta)$. This, in particular, allows obtaining error η as small as 2^{-d} in polynomial time in the dimension d. Our list-size however is $\alpha^{-\operatorname{poly}(1/\alpha)}$ compared to $O(1/\alpha^{\log(1/\alpha) + \log(1/\eta)})$ in [BK21].

Corollary 9.2. *For any $\eta > 0$, there is an algorithm that takes input a collection of n points $Y \subseteq \mathbb{Q}^d$, runs in time $n^{\tilde{O}(1/\alpha^{12})} \operatorname{poly}\log(1/\eta)$ and either outputs "reject" or a list $\mathcal{L}$ of size $k = \alpha^{-\operatorname{poly}(1/\alpha)}$ of candidate projection matrices $\hat{\Pi}_1, \hat{\Pi}_2, \ldots, \hat{\Pi}_k$ with the following guarantee: suppose X is a set of $n \geqslant n_0 = O(d^{O(1/\alpha^6)}/\alpha)$ i.i.d. draws from a distribution D on $\mathbb{R}^d$ such that D has mean 0, covariance Π_* – a projection matrix to a subspace of $\mathbb{R}^d$, is $s(\delta)$-certifiably (C, δ)-anti-concentrated and has $2t$-certifiable C-hypercontractivity of degree 2 polynomials for all t. Suppose Y is an $1-\alpha)$-corruption of X. Then, with probability at least 0.99 over the draw of X and the random choices of the algorithm, the algorithm does not reject and outputs a list that contains a candidate $\hat{\Pi}_k$ satisfying:*

$$\left\| \hat{\Pi}_k - \Pi_* \right\|_F \leqslant \eta \, .$$

9.3 Clustering Non-Spherical Mixtures

Let $M = \sum_i p_i D_i$ be a mixture of $D_1, D_2, \ldots, D_k$ such that for each i, D_i is $s(\delta)$-certifiably C-anti-concentrated distributions with $2t$-certifiably C-hypercontractive degree 2 polynomials for all $t \in \mathbb{N}$ and $p_i \geqslant p_{min}$ for each i. Then, so long as $\varepsilon < p_{min}/2$, an ε corrupted sample from M intersects with an i.i.d. sample from any D_i in at least $p_{min}/4$ points. Thus, we can immediately apply our list-decodable mean and covariance estimation algorithm (Theorem 8.6) with $\alpha = p_{min}/4$ runs in time $d^{\mathrm{poly}(1/p_{min})}$ and get a list of $(1/p_{min})^{\mathrm{poly}(1/p_{min})}$ candidates such that there is a Δ-close (in parameter distance) mean-covariance pair to (μ_i, Σ_i) for every i for $\Delta = \mathrm{poly}(1/p_{min})$. Note that this consequence does not require any separation assumptions.

If the component D_is are guaranteed to have well-separated parameters (as in the main result in [BK20b]), then we can cluster the input corrupted sample Y with at most $O(\varepsilon/p_{min})$ fraction of misclassified points in any cluster. This, in particular, also allows obtaining estimates of the parameters up to $\tilde{O}(\varepsilon/p_{min})$ in parameter-distance.

Let us briefly explain this procedure before supplying a more detailed proof sketch. The main idea is simple: suppose we were given the parameters of each D_i *exactly*. We can then apply a natural clustering procedure based on the parameters. We will argue that this natural clustering procedure continues to function even if we have an estimate of the parameters that is accurate to within $\mathrm{poly}(1/p_{min})$ factor in parameter-distance as long as the pairwise separation (again, in parameter-distance) between the parameters of D_is is at least $1/p_{min}^{O(k)}$. For the sake of keeping the exposition in this section simple, we only describe the algorithm for the case when D_is are Gaussian distributions.

Theorem 9.3 (Robust Clustering of TV-Separated Gaussian Mixtures)**.** *Fix $p_{min} > 0, k \in \mathbb{N}, B \in \mathbb{N}$. For every large enough $d \in \mathbb{N}$, there is an algorithm that takes input an ε-corruption Y of an i.i.d. sample $X = C_1 \cup C_2 \ldots C_k$ of size $n \geqslant n_0 = d^{\mathrm{poly}(1/p_{min})}$ from a d-dimensional mixture $\sum_i p_i D_i(\mu_i, \Sigma_i)$ of Gaussians with parameters μ_i, Σ_i having rational entries of bit complexity at most B and runs in time $(Bn)^{\mathrm{poly}(1/p_{min})}$. If $\varepsilon \leqslant c p_{min}$ for a small enough constant $c > 0$ and parameter-distance$(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j)) \geqslant p_{min}^{-O(k)}$, the algorithm, with probability at least 0.99 over the draw of the original uncorrupted sample X and the random choices of the algorithm, outputs a clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \hat{C}_k$ of Y with the property $\min_{\pi:[k]\to[k]} \max_{i \leqslant k}(1 - |\hat{C}_i \cap C_{\pi(i)}|/|C_{\pi(i)}|) \leqslant \tilde{O}(\varepsilon/p_{min})$.*

Notice that in comparison, the algorithm from [BK20b] is presented only for the equiweighted case (i.e., $p_{min} = 1/k$), needs $n = d^{k^{O(k)}}$ samples and $n^{k^{O(k)}}$ time, and works only when the fraction of outliers $\varepsilon \ll k^{-O(k)}$. The parameters in the algorithm of [DHKK20] are worse (with, roughly speaking, every instance of $k^{O(k)}$ replaced by a $\mathrm{poly}(k)$ size tower of exponential in k for equiweighted mixture of Gaussians).

Proof Sketch. First, we apply the algorithm from Theorem 8.6 to Y with parameter $\alpha = p_{min} - \varepsilon \geqslant 0.99 p_{min}$. By thinking of the uncorrupted points in the any true cluster C_i as the inliers and all the rest of Y as outliers, we observe that Y is an $1 - (p_{min} - \varepsilon)$ corruption of $D_i(\mu_i, \Sigma_i)$ and thus, Theorem 8.6 provides an algorithm that runs in time $(Bn)^{\mathrm{poly}(1/p_{min})}$ and generates a list of size $p_{min}^{-\mathrm{poly}(1/p_{min})}$ such that for every i, there is a $(\hat{\mu}_i, \hat{\Sigma}_i)$ in the list that is $\Delta = \mathrm{poly}(1/p_{min})$-close in parameter-distance to

(μ_i, Σ_i). Call a k-tuple of parameters from this list *good* if for every (μ_i, Σ_i), there is a Δ-close $(\hat{\mu}_i, \hat{\Sigma}_i)$ in the k-tuple. Notice that by the approximate triangle inequality for parameter-distance, every pair of parameters in such a good k-tuple must be Δ^{100k} apart in parameter-distance.

The goal of the algorithm now is to use this list to cluster the input points. Here's how the algorithm proceeds: the algorithm enumerates over all k-tuples of parameters in the list that satisfy the property that every pair in the k-tuple is at a distance of at least Δ^{100k} from each other. We will give a procedure to cluster assuming the k set of parameters are good. Running a cluster verification procedure (Lemma 9.5) similar to the one emploiyed in [BK20b] on each cluster so constructed allows verifying whether each cluster satisfies hypercontractivity and anti-concentration properties finishing the algorithm. Thus, the key remaining piece is to establish that if we chose a subset of k parameters from the list that are good, then we can efficiently construct an approximate clustering of Y.

Let's now describe the clustering procedure assuming a good set of known k parameters, say $\{(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i \leqslant k}$. Our ideas rely on partial cluster recovery procedure employed in [BK20b] that exploits the three kinds of separations possible between mixtures of reasonable distributions (Definition 1.3).

First, suppose there is a unit vector v such that $v^\top \hat{\Sigma}_i v \leqslant \Delta^{O(k)} v^\top \hat{\Sigma}_j v$ for some $1 \leqslant i < j \leqslant k$. Such a vector, if it exists, can be found by going over all pairs i, j, taking the top eigenvector of $\hat{\Sigma}_j^{-\dagger/2} \hat{\Sigma}_i \hat{\Sigma}_j^{-\dagger/2}$ and applying $\hat{\Sigma}_j^{1/2}$ to it. In this case, we will use the following variance clustering procedure. Observe that there is a partition of $[k]$ into $S, \bar{S}$ such that $v^\top \hat{\Sigma}_i v \leqslant \beta$ for all $i \in S$ and $v^\top \hat{\Sigma}_i v \geqslant \Delta^{O(1)} \beta$ for all $i \in \bar{S}$. Since (μ_i, Σ_i) are Δ-close to $(\hat{\mu}_i, \hat{\Sigma}_i)$, by Lemma 9.6, we must thus have that $v^\top \Sigma_i v \leqslant \beta'$ for all $i \in S$ and $v^\top \hat{\Sigma}_i v \geqslant \beta' \Delta^{O(1)}$ where $\beta' = \beta \Delta^{O(1)}$. Our clustering algorithm does the following: for each $y \in Y$, we include y in cluster L if there is an $i \in S$ such that $\frac{\eta}{2\Delta} v^\top \hat{\Sigma}_i v \leqslant \langle y - \hat{\mu}_i, v \rangle^2 \leqslant O(\log 1/\eta) \Delta^2 v^\top \hat{\Sigma}_i v$ for $\eta = \varepsilon$. If there is no such i, we include $y \in R$. We now claim that $| \cup_{i \in S} C_i \cap L| \geqslant | \cup_{i \in S} C_i| - 2\varepsilon n$, and, $| \cup_{i \notin S} C_i \cap R| \geqslant | \cup_{i \notin S} C_i| - 2\varepsilon n$. To see why, observe first that $v^\top \hat{\Sigma}_i v \ll \Delta^{O(1)} v^\top \hat{\Sigma}_j v$ for all $i \in S, j \notin S$. The claim then immediately follows by observing that from Lemma 9.4, at most η fraction of $x \in \cup_{i \notin S}$ get put in L and similarly, at most an η fraction of $x \in \cup_{i \in S}$ get put in R – in particular, we have achieved a partial clustering of the input samples with an error of at most $2\varepsilon n$ points on either side. We can repeat the above "variance clustering" procedure until there's no vector v satisfying $v^\top \hat{\Sigma}_i v \leqslant \Delta^{O(k)} v^\top \hat{\Sigma}_j v$ for some $1 \leqslant i < j \leqslant k$. Thus, in the following, we can assume that for every v and every i, $\Delta^{-O(k)} \frac{1}{k} (\sum_i \hat{\Sigma}_i) v \leqslant v^\top \hat{\Sigma}_i v \leqslant \Delta^{O(k)} \frac{1}{k} (\sum_i \hat{\Sigma}_i) v$.

Next, suppose there is a unit vector v such that $\langle \hat{\mu}_i - \hat{\mu}_j, v \rangle^2 \geqslant \Delta^{O(k)} v^\top \frac{1}{k} \sum_i \hat{\Sigma}_i v) v$. Such a vector, if it exists, can be found by going over all pairs i, j, and checking if $v = (\frac{1}{k} \sum_i \hat{\Sigma}_i)^{-\dagger/2} (\hat{\mu}_i - \hat{\mu}_j)$ satisfies the inequality above. Given such a vector v, we can again partition $[k]$ into two groups, S and $\bar{S}$ such that for every $i \in S, j \notin S$, $\langle \hat{\mu}_i - \hat{\mu}_j, v \rangle^2 \geqslant \Delta^{O(1)} v^\top \frac{1}{k} \sum_i \hat{\Sigma}_i v) v$. We now do a "mean-clustering" procedure as follows: we put $y \in L$ iff there is an $i \in S$ such that $O(\Delta^2 \log 1/\eta) v^\top \Sigma' v \geqslant \langle y - \hat{\mu}_i, v \rangle^2 \geqslant \frac{\eta}{2\Delta} v^\top \Sigma' v$ for $\eta = \varepsilon$. By an analysis similar to the above, we arrive at a partial clustering as before this time ensuring that every group consists of clusters with mean-close parameters.

Finally, suppose there is a pair i, j such that $\left\| \hat{\Sigma}_i - \hat{\Sigma}_j \right\|_F \geqslant \Delta^{O(k)}$. Then, in particular, for $A = \hat{\Sigma}_i - \hat{\Sigma}_j$, it holds that $\mathrm{tr}(A \cdot (\hat{\Sigma}_i - \hat{\Sigma}_j)) \geqslant \Delta^{O(k)}$. As before, we find a partitions of $[k]$ into two groups S and $\bar{S}$ such that for every $i \in S, j \notin S$, $\mathrm{tr}(A \cdot (\hat{\Sigma}_i - \hat{\Sigma}_j) \geqslant \Delta^{O(1)}$. We now apply a "Frobenius clustering" procedure that puts $y \in L$ if there is an $i \in S$ such that $|(y-\hat{\mu}_i)^\top A(y-\hat{\mu}_i)-\mathrm{tr}(A\hat{\Sigma}_i)| \geqslant O(\Delta \log 1/\eta) \left\|\hat{\Sigma}^{1/2}A\hat{\Sigma}^{1/2}\right\|_F$. Using Lemma 9.4 and a similar analysis in the above two cases, we arrive at a partial clustering as before ensuring that every group consists of relative Frobenius close parameters.

At the end of the three modes of clustering, we end up with partial clustering from the original data with at most $O(k\varepsilon n)$ points misclassified in total. Further, within each group, we every pair of clusters that contribute must be within a $\Delta^{O(k)}$ distance in each of the three possible ways of separation and thus, also $\Delta^{O(k)}$-close in parameter-distance. Since every pair of (μ_i, Σ_i)s are $\gg \Delta^{O(k)}$-far in parameter-distance, the resulting groups must in fact be an approximate clustering of the data with at most $O(k\varepsilon n)$ misclassified points as desired.

□

Lemma 9.4 (Approximate Isotropization)**.** *For $\Delta \geqslant 1, \eta > 0$ and two sets of parameters $(\mu, \Sigma), (\mu', \Sigma')$ of d-dimensional Gaussian distributions, let* parameter-distance$((\mu,\Sigma),(\mu',\Sigma')) \leqslant \Delta$*. Let $x \sim \mathcal{N}(\mu, \Sigma)$. Then,*

$$\mathbb{P}[\frac{\eta}{2\Delta} v^\top \Sigma' v \leqslant \langle x-\mu', v\rangle^2 \leqslant O(\Delta^2 \log 1/\eta) v^\top \Sigma' v] \geqslant 1-\eta\,,$$

and,

$$\mathbb{P}[\left|(x-\mu')^\top A(x-\mu') - \mathrm{tr}(A\Sigma')\right| \geqslant O(\Delta \log 1/\eta) \left\|\Sigma'^{1/2} A \Sigma'^{1/2}\right\|_F] \geqslant 1-\eta\,.$$

Proof. We know by subgaussianity and anti-concentration of Gaussian random variables that

$$\mathbb{P}[\eta/2 v^\top \Sigma v \leqslant \langle x-\mu, v\rangle^2 \leqslant O(\log 1/\eta) v^\top \Sigma' v] \geqslant 1-\eta\,.$$

Since parameter-distance$((\mu,\Sigma),(\mu',\Sigma')) \leqslant \Delta$, we also know that $\frac{1}{\Delta} v^\top \Sigma v \leqslant v^\top \Sigma' v \leqslant \Delta v^\top \Sigma v$. Thus,

$$\mathbb{P}[\frac{\eta}{2\Delta} v^\top \Sigma' v \leqslant \langle x-\mu, v\rangle^2 \leqslant O(\Delta \log 1/\eta) v^\top \Sigma' v] \geqslant 1-\eta\,.$$

Further, $\langle \mu-\mu', v\rangle^2 \leqslant \Delta v^\top(\Sigma+\Sigma')v \leqslant (1+\Delta^2) v^\top \Sigma' v$. Thus,

$$\mathbb{P}[\frac{\eta}{2\Delta} v^\top \Sigma' v \leqslant \langle x-\mu', v\rangle^2 \leqslant O(\Delta^2 \log 1/\eta) v^\top \Sigma' v] \geqslant 1-\eta\,.$$

Next, by tail bounds for degree 2 polynomials of hypercontractive distributions, we have:

$$\mathbb{P}[\left|(x-\mu)^\top A(x-\mu) - \mathrm{tr}(A\Sigma)\right| \geqslant O(\log 1/\eta) \left\|\Sigma^{1/2} A \Sigma^{1/2}\right\|_F] \geqslant 1-\eta\,. \tag{9.1}$$

Now, observe that using parameter-distance$((\mu,\Sigma),(\mu',\Sigma')) \leqslant \Delta$, we have:

$$\mathrm{tr}(A(\Sigma-\Sigma')) = \mathrm{tr}(\Sigma'^{1/2} A \Sigma'^{1/2} \cdot (\Sigma'^{\dagger/2} \Sigma \Sigma'^{\dagger/2} - I)) \leqslant \left\|\Sigma'^{1/2} A \Sigma'^{1/2}\right\|_F \left\|I - \Sigma'^{\dagger/2} \Sigma \Sigma'^{\dagger/2})\right\|_F \leqslant \Delta \left\|\Sigma'^{1/2} A \Sigma'^{1/2}\right\|_F\,.$$

Further, again using parameter-distance$((\mu,\Sigma),(\mu',\Sigma')) \leqslant \Delta$, we have:

$$\left\|\Sigma^{1/2} A \Sigma^{1/2}\right\|_F = \left\|\Sigma^{1/2} \Sigma'^{\dagger/2} (\Sigma'^{1/2} A \Sigma'^{1/2}) \Sigma'^{\dagger/2}) \Sigma^{1/2}\right\|_F \leqslant \left\|\Sigma^{1/2} \Sigma'^{\dagger/2}\right\|_2^2 \left\|\Sigma'^{1/2} A \Sigma'^{1/2}\right\|_F \leqslant \Delta \left\|\Sigma'^{1/2} A \Sigma'^{1/2}\right\|_F\,.$$

Finally, since parameter-distance$((\mu,\Sigma),(\mu',\Sigma')) \leqslant \Delta$, we have that $\left\|\Sigma'^{1/2}(\mu-\mu')\right\|_2^2 \leqslant \Delta$ and that $\left\|\Sigma'^{\dagger/2}(x-\mu)^\top\right\|_2^2 \leqslant O(\log 1/\eta)\Delta$ by subgaussianity of $\mathcal{N}(\mu,\Sigma)$. Thus, for $A' = \Sigma'^{1/2} A \Sigma'$,

$$\begin{aligned}&\left|\Sigma'^{\dagger/2}(x-\mu')^\top A' \Sigma'^{\dagger/2}(x-\mu') - \Sigma'^{\dagger/2}(x-\mu)^\top A' \Sigma'^{\dagger/2}(x-\mu)\right| \\ &\leqslant |\Sigma'^{\dagger/2}(\mu-\mu')^\top A' \Sigma'^{\dagger/2}(x-\mu)| + |\Sigma'^{\dagger/2}(x-\mu)^\top A' \Sigma'^{\dagger/2}(\mu-\mu')| + |\Sigma'^{\dagger/2}(\mu-\mu')^\top A' \Sigma'^{\dagger/2}(\mu-\mu')| \\ &\leqslant O(\Delta \log 1/\eta) \left\|A'\right\|_F \,.\end{aligned}$$

Thus, combined with (9.1), we have:

$$\mathbb{P}[\left|(x-\mu')^\top A (x-\mu') - \mathrm{tr}(A\Sigma')\right| \geqslant O(\Delta \log 1/\eta) \left\|\Sigma'^{1/2} A \Sigma'^{1/2}\right\|_F] \geqslant 1-\eta\,.$$

□

We will use the following cluster verification algorithm from [BK20b].

Fact 9.5 (Verifying Clusters, analogous to Lemma 6.5 in [BK20b]). *There is an algorithm that takes input a set of n d-dimensional points Y and a subset $\hat{C} \subseteq Y$, runs in time $n^{\mathrm{poly}(1/p_{min})}$, and outputs acccept or reject with the following guarantee: Suppose X is a good sample from a Δ-separated mixture of Gaussian distributions $\sum_i p_i \mathcal{N}(\mu_i,\Sigma_i)$ with weights $p_i \geqslant p_{min}$ for every i. Let Y be a τ-corruption of X. Let $\hat{C} \subseteq Y$ be such that $|\hat{C} \cap C_i| \leqslant (1 - O(\tau/p_{min}))|C_i|$ for every i. Then, the algorithm rejects with probability at least $1 - 1/\mathrm{poly}(n)$ over the draw of X. If, on the other hand, there exists an i such that $|\hat{C} \cap C_i| \geqslant (1 - O(\tau/p_{min}))\max\{|C_i|,|\hat{C}|\}$, then, the algorithm accepts with probability at least $1 - 1/\mathrm{poly}(n)$ over the draw of X.*

Lemma 9.6 (Approximate Triangle Inequality for Parameter Distance)**.** *Suppose* parameter-distance$((\hat{\mu},\hat{\Sigma}),(\mu_i,\Sigma_i))$, parameter-distance$((\hat{\mu},\hat{\Sigma}),(\mu_j,\Sigma_j)) \leqslant \Delta$ *for* $\Delta > 1$. *Then,* parameter-distance$((\mu_i,\Sigma_i),(\mu_j,\Sigma_j)) \leqslant 3\Delta^2$.

Proof. For any vector v, we know that $v^\top \Sigma_j v \leqslant \Delta v^\top \hat{\Sigma} v \leqslant \Delta^2 v^\top \Sigma_i v$. Similarly, $v^\top \Sigma_i v \leqslant \Delta^2 v^\top \Sigma_j v$. This establishes the multiplicative spectral part of the guarantee in parameter-distance.

Next, let's consider the relative Frobenius guarantee. Towards that first observe that $\left\|\Sigma_i^{-1/2}\hat{\Sigma}^{1/2} v\right\|_2^2 \leqslant \left\|\Sigma_i^{-1/2}\hat{\Sigma}\Sigma_i^{-1/2}\right\|_2 \left\|v\right\|_2^2 \leqslant (1+\Delta)\left\|v\right\|_2^2$. Next, because of the multiplicative spectral guarantee, we can assume that $\Sigma_i, \hat{\Sigma}, \Sigma_j$ all have the same range space. We can thus assume that they are all full rank WLOG (as otherwise, we can simply work in their common range space instead).

$$\begin{aligned}\left\|\Sigma_i^{-1/2}\Sigma_j\Sigma_i^{-1/2} - I\right\|_F &\leqslant \left\|\Sigma_i^{-1/2}\hat{\Sigma}^{1/2}(\hat{\Sigma}^{-1/2}\Sigma_j - I)\hat{\Sigma}^{-1/2}\hat{\Sigma}^{1/2}\Sigma_i^{-1/2} + \Sigma_i^{-1/2}\hat{\Sigma}\Sigma_i^{-1/2} - I\right\|_F \\ &\leqslant \left\|\Sigma_i^{-1/2}\hat{\Sigma}^{1/2}(\hat{\Sigma}^{-1/2}\Sigma_j - I)\hat{\Sigma}^{-1/2}\hat{\Sigma}^{1/2}\Sigma_i^{-1/2}\right\|_F + \left\|\Sigma_i^{-1/2}\hat{\Sigma}\Sigma_i^{-1/2} - I\right\|_F \\ &\leqslant \left\|\Sigma_i^{-1/2}\hat{\Sigma}^{1/2}\right\|_2^2 \left\|\hat{\Sigma}^{-1/2}\Sigma_j\hat{\Sigma}^{-1/2} - I\right\|_F + \left\|\Sigma_i^{-1/2}\hat{\Sigma}\Sigma_i^{-1/2} - I\right\|_F \\ &\leqslant (1+\Delta)\Delta + \Delta \leqslant 3\Delta^2\,.\end{aligned}$$

Here, in the first inequality, we used the triangle inequality for Frobenius norm and in the second inequality, used the contraction principle for Frobenius norms twice: for any matrices A, B, $\|AB\|_F \leqslant \|A\|_2 \|B\|_F$ along with the fact that $\left\|\Sigma_i^{-1/2}\hat{\Sigma}^{1/2}v\right\|_2^2 \leqslant (1+\Delta)\|v\|_2^2$.

Finally, observe that for any vector v:

$$\begin{aligned}\langle \mu_i - \mu_j, v\rangle^2 &\leqslant 2\langle \mu_i - \hat{\mu}, v\rangle^2 + 2\langle \hat{\mu} - \mu_j, v\rangle^2 \\ &\leqslant \Delta v^\top(\Sigma_i + \Sigma_j + 2\hat{\Sigma})v \leqslant 2\Delta v^\top(\Sigma_i + \Sigma_j)v\,.\end{aligned}$$

□

References

[BBV08] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala, *A discriminative framework for clustering via similarity functions*, STOC, ACM, 2008, pp. 671–680. 1, 7

[BCSS98] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale, *Complexity and real computation*, Springer-Verlag, New York, 1998, With a foreword by Richard M. Karp. MR 1479636 3, 14

[BDJ$^+$20] 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** (2020). 6, 7, 8, 20

[BK20a] Ainesh Bakshi and Pravesh Kothari, *List-decodable subspace recovery via sum-of-squares*, arXiv preprint arXiv:2002.05139 (2020). 1, 5

[BK20b] ______, *Outlier-robust clustering of non-spherical mixtures*. 1, 2, 4, 6, 7, 8, 9, 10, 11, 18, 20, 21, 48, 51, 52, 54

[BK21] Ainesh Bakshi and Pravesh K. Kothari, *List-decodable subspace recovery: Dimension independent error in polynomial time*, Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021 (Dániel Marx, ed.), SIAM, 2021, pp. 1279–1297. 1, 3, 6, 8, 9, 11, 19, 20, 44, 49, 50

[BKS15] Boaz Barak, Jonathan A. Kelner, and David Steurer, *Dictionary learning and tensor decomposition via the sum-of-squares method [extended abstract]*, STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing, ACM, New York, 2015, pp. 143–151. MR 3388192 17

[BKS17] Boaz Barak, Pravesh K. Kothari, and David Steurer, *Quantum entanglement, sum of squares, and the log rank conjecture*, STOC, ACM, 2017, pp. 975–988. 16

[BS16] Boaz Barak and David Steurer, *Proofs, beliefs, and algorithms through the lens of sum-of-squares*, 2016, Lecture notes in preparation, available on `http://sumofsquares.org`. 16

[CFB19] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett, *Fast mean estimation with sub-gaussian rates*, Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA (Alina Beygelzimer and Daniel Hsu, eds.), Proceedings of Machine Learning Research, vol. 99, PMLR, 2019, pp. 786–806. 1

[CHK$^+$20] Yeshwanth Cherapanamjeri, Samuel B. Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni, *Algorithms for heavy-tailed statistics: regression, covariance estimation, and beyond*, Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020 (Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, eds.), ACM, 2020, pp. 601–609. 1

[CMY20] Yeshwanth Cherapanamjeri, Sidhanth Mohanty, and Morris Yau, *List decodable mean estimation in nearly linear time*, 61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020, IEEE, 2020, pp. 141–148. 1, 3

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant, *Learning from untrusted data*, STOC, ACM, 2017, pp. 47–60. 1, 3

[DHKK20] Ilias Diakonikolas, Samuel Hopkins, Daniel Kane, and Sushrut Karmalkar, *Robustly learning any clusterable mixture of gaussians*, Personal Communication (2020). 1, 2, 6, 7, 8, 11, 21, 51

[DKK$^+$16] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart, *Robust estimators in high dimensions without the computational intractability*, FOCS, IEEE Computer Society, 2016, pp. 655–664. 1, 3

[DKK20a] Ilias Diakonikolas, Daniel Kane, and Daniel Kongsgaard, *List-decodable mean estimation via iterative multi-filtering*, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, eds.), 2020. 1, 3

[DKK$^+$20b] Ilias Diakonikolas, Daniel M. Kane, Daniel Kongsgaard, Jerry Li, and Kevin Tian, *List-decodable mean estimation in nearly-pca time*, CoRR **abs/2011.09973** (2020). 1, 3

[DKP$^+$21] Ilias Diakonikolas, Daniel M. Kane, Ankit Pensia, Thanasis Pittas, and Alistair Stewart, *Statistical query lower bounds for list-decodable linear regression*, CoRR **abs/2106.09689** (2021). 3, 4

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart, *Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures*, FOCS, IEEE Computer Society, 2017, pp. 73–84. 4, 6

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart, *List-decodable robust mean estimation and learning mixtures of spherical gaussians*, Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, 2018, pp. 1047–1060. 1

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad, *The total variation distance between high-dimensional gaussians*, 2018. 21

[EvM20] Jeff Erickson, Ivor van der Hoog, and Tillmann Miltzow, *Smoothing the gap between NP and ER*, 61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020, IEEE, 2020, pp. 1022–1033. 3, 14

[FKP19] Noah Fleming, Pravesh Kothari, and Toniann Pitassi, *Semialgebraic proofs and efficient algorithm design*, Foundations and Trends® in Theoretical Computer Science **14** (2019), no. 1-2, 1–221. 8, 16

[HL17] Sam B. Hopkins and Jerry Li, *Mixture models, robustness, and sum of squares proofs*, 2017. 1, 8

[Hop20] Samuel B. Hopkins, *Mean estimation with sub-Gaussian rates in polynomial time*, Ann. Statist. **48** (2020), no. 2, 1193–1213. MR 4102693 1

[KKK19] Sushrut Karmalkar, Adam R. Klivans, and Pravesh Kothari, *List-decodable linear regression*, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada (Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, eds.), 2019, pp. 7423–7432. 1, 2, 3, 4, 5, 7, 8, 9, 10, 12, 18, 20, 49

[KMZ22] Pravesh K. Kothari, Peter Manohar, and Brian Hu Zhang, *Polynomial-time sum-of-squares can robustly estimate mean and covariance of gaussians optimally*, International Conference on Algorithmic Learning Theory, 29-1 April 2022, Paris, France (Sanjoy Dasgupta and Nika Haghtalab, eds.), Proceedings of Machine Learning Research, vol. 167, PMLR, 2022, pp. 638–667. 11

[KOTZ14] Manuel Kauers, Ryan O'Donnell, Li-Yang Tan, and Yuan Zhou, *Hypercontractive inequalities via sos, and the frankl-rödl graph*, SODA, SIAM, 2014, pp. 1644–1658. 2, 20

[KS17a] Pravesh K. Kothari and Jacob Steinhardt, *Better agnostic clustering via relaxed tensor norms*, 2017. 1, 4, 13, 47

[KS17b] Pravesh K. Kothari and David Steurer, *Outlier-robust moment-estimation via sum-of-squares*, CoRR **abs/1711.11581** (2017). 1, 2, 8, 18, 20

[Las01] Jean B. Lasserre, *New positive semidefinite relaxations for nonconvex quadratic programs*, Advances in convex analysis and global optimization (Pythagorion, 2000), Nonconvex Optim. Appl., vol. 54, Kluwer Acad. Publ., Dordrecht, 2001, pp. 319–331. MR 1846160 17

[LLL82] A. K. Lenstra, H. W. Lenstra, Jr., and L. Lovász, *Factoring polynomials with rational coefficients*, Math. Ann. **261** (1982), no. 4, 515–534. MR 682664 15, 60

[LM21] Allen Liu and Ankur Moitra, *Settling the robust learnability of mixtures of gaussians*, STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021 (Samir Khuller and Virginia Vassilevska Williams, eds.), ACM, 2021, pp. 518–531. 6, 7, 8

[LRV16] Kevin A. Lai, Anup B. Rao, and Santosh Vempala, *Agnostic estimation of mean and covariance*, FOCS, IEEE Computer Society, 2016, pp. 665–674. 1

[Nes00] Yurii Nesterov, *Squared functional systems and optimization problems*, High performance optimization, Appl. Optim., vol. 33, Kluwer Acad. Publ., Dordrecht, 2000, pp. 405–440. MR 1748764 17

[O'D14] Ryan O'Donnell, *Analysis of Boolean functions*, Cambridge University Press, New York, 2014. MR 3443800 20

[Par00] Pablo A Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, Ph.D. thesis, California Institute of Technology, 2000. 17

[RY19] Prasad Raghavendra and Morris Yau, *List decodable learning via sum of squares*, Manuscript, 2019. 1, 2, 3, 4, 5, 7, 8, 9, 10, 20, 49

[RY20a] ______, *List decodable learning via sum of squares*, Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020 (Shuchi Chawla, ed.), SIAM, 2020, pp. 161–180. 12

[RY20b] ______, *List decodable subspace recovery*, 2020. 1, 5, 8, 20, 50

[RY20c] ______, *List decodable subspace recovery*, Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria] (Jacob D. Abernethy and Shivani Agarwal, eds.), Proceedings of Machine Learning Research, vol. 125, PMLR, 2020, pp. 3206–3226. 3

[Sho87] N. Z. Shor, *Quadratic optimization problems*, Izv. Akad. Nauk SSSR Tekhn. Kibernet. (1987), no. 1, 128–139, 222. MR 939596 17

A Deferred Proofs

Proof of Corollary 3.18. Observe that the polynomial inequality in indeterminate z, $z^2+2\delta^2 p_\delta^2(z)-\delta^2 \geqslant 0$ holds. To see this, consider the following two cases: 1) $z^2 \geqslant \delta^2$: in this case, we are done because p_δ^2 is non-negative. 2) $z^2 < \delta$: in this case, we use the fact that $p_\delta^2(z) \geqslant (1-\delta)^2$ which, for $\delta < 0.1$ implies that $2\delta^2 p_\delta^2(z) \geqslant \delta^2$.

Now, using Fact 3.11, we know that $\frac{z}{O(s(\delta))}\Big\vert\!\!-\; \{z^2 + 2\delta^2 p_\delta^2(z) - \delta^2 \geqslant -\eta\}$.

As a result, we know that $z^2 + 2\delta^2 p_\delta^2(z) - \delta^2 + \eta = r(z)$ for some sos polynomial r in z with coefficients upper-bounded by $2 \cdot (4z)^s$ and degree $= s(\delta)$ (because a polynomial is identically 0 on reals if and only if all its coefficients are 0). Further, we observe that r is an even polynomial because p is even.

Now, let's substitute $z = \frac{\langle x,v\rangle}{\sqrt{v^\top \Sigma v}}$ in p and r. Since p, r are even, all monomials in z appearing with non-zero coefficients in p, r are even powers and are thus monomials in z^2. As a result, $(v^\top \Sigma v)^{s(\delta)}(z^2 + 2\delta^2 p_\delta^2(z) - \delta^2 + \eta - r(z))$ is a polynomial in indeterminate v for any given x. Further, $(v^\top \Sigma v)^{s(\delta)} r$ is SoS in $v^\top \Sigma v$ and therefore, also SoS in indeterminate v. So, putting this all together upon rearranging gives $\frac{v}{O(s)+\operatorname{poly}\log(1/\eta)}\Big\vert\!\!-\; \left\{(v^\top \Sigma v)^{s-1}\langle x,v\rangle^2 + 2\delta^2 q_{\delta,\Sigma}^2(x,v) \geqslant (\delta^2-\eta)(v^\top \Sigma v)^s\right\}$ for all $\eta > 0$. Taking $\eta = 0.01\delta^2$ suffices for the conclusion.

The second inequality follows from $\frac{z}{O(s(\delta))}\Big\vert\!\!-\; \{\mathbb{E}[p_\delta^2(\langle x,v\rangle)] \leqslant O(\delta)\}$ upon substitution. □

B Bit Complexity Analysis

We will use the following basic observations in our bit complexity analysis to analyze the rank deficient covariance Σ_* of Theorem 8.6.

Recall that in this case, we assume that the target matrix Σ_* has rational entries with bit complexity at most B. The following proposition shows that in this case, the smallest non-zero eigenvalue of Σ_* cannot be too small.

Proposition B.1 (Smallest non-zero singular value of a rational matrix)**.** *For $d \in \mathbb{N}$, let $A \in \mathbb{Q}^{d\times d}$ be a non-zero matrix with each entry of bit length at most B. Then, every non-zero eigenvalue of A has absolute value at least 2^{-3Bd^3}.*

Proof. Let L be the least common multiple of the denominators of the rational numbers appearing in entries of A. Then, since each denominator is upper-bounded by 2^B, $L \leqslant 2^{Bd^2}$. Thus, $A' = LA$ is a matrix with integer entries. Observe further that by the Gershgorin circle theorem, the spectral norm of A' (and thus, the eigenvalue of largest magnitude) is at most $n2^{Bd^2}2^B \leqslant 2^{2Bd^2}$.

Let $r \leqslant d$ be the rank of A'. Consider the characteristic polynomial $char(A')$ of A in indeterminate λ. Then, $char(A')$ is monic and has integer coefficients. Consider the coefficient of λ^r in $char(A')$. Then, this coefficient equals the sum of r-wise products of eigenvalues of A'. Since A' has rank r, it has exactly r non-zero eigenvalues and thus, the coefficient of λ^r is the product of the non-zero eigenvalues of A'. Since the coefficients of A' is a non-zero integer, this product is at least 1 in magnitude. Since all eigenvalues of A' are of magnitude at most 2^{2Bd^2}, the smallest magnitude of any eigenvalue must thus be at least $2^{-2Bd^2 r} \geqslant 2^{-2Bd^3}$.

Thus, every non-zero eigenvalue of A has magnitude at least $L^{-1}2^{-2Bd^3} \geqslant 2^{-3Bd^3}$.

□

We will also need the following basic facts about the classical algorithm for lattice basis reduction due to Lenstra, Lenstra and Lovász [LLL82].

Preliminaries on Integer Lattices Let $A \in \mathbb{Q}^{d \times d}$ be a matrix of rationals. The lattice defined by A is the discrete additive subgroup $\mathcal{L}(a_1, a_2, \ldots, a_d) = \sum_{i=1}^{d} z_i a_i$ as z_is vary over $\mathbb{Z}$ and a_is are the columns of A. We write $\lambda_1(\mathcal{L})$ to be the length of the smallest non-zero vector in $\mathcal{L}$. More generally, let $\lambda_i(\mathcal{L})$ be the minimum of the maximum length of any vector from among all linearly independent sets of i vectors $v_1, v_2, \ldots, v_i \in \mathcal{L}$.

The determinant of a lattice $\mathcal{L}$ is defined as $\det(\mathcal{L}) = \det(A^\top A)$ if A is full-rank. Notice that $\det(\mathcal{L})$ is independent of the basis used for $\mathcal{L}$.

Fact B.2 (Minkowski's Theorems). $\lambda_1(\mathcal{L}) \leqslant \det(\mathcal{L})$. *More generally,* $\prod_{i \leqslant d} \lambda_i(\mathcal{L}) \leqslant d^{\sqrt{d}} \det(\mathcal{L})$.

Given a matrix $A \in \mathbb{Q}^{d \times k}$, the orthogonal lattice $\mathcal{L}^{\perp}(A)$ defined by A is the set of all integer vectors v such that $Av = 0$. We can relate the size of the basis for $\mathcal{L}^{\perp}(A)$ to that of $\mathcal{L}(A)$ via Hadamard's inequality:

Fact B.3 (Hadamard's Inequality). $\det(\mathcal{L}^{\perp}(A)) \leqslant \det(\mathcal{L}(A))$.

Finally, we recall the guarantees of the lattice basis reduction algorithm of [LLL82].

Fact B.4 (LLL Algorithm). *Let $\mathcal{L}$ be a lattice defined by a $d \times d$ matrix A. There is a polynomial time algorithm that takes input the Gram matrix $A^\top A$ and outputs a basis $b_1, b_2, \ldots, b_d$ of $\mathcal{L}$ such that $\|b_i\|_2 \leqslant 2^{O(D)} \lambda_i(\mathcal{L})$.*

Analyzing the algorithm in Theorem 8.6 in the word RAM model We begin by setting λ to be $2^{-B \cdot d^{C/\alpha}}$ for a sufficiently large constant C. We start by running the algorithm described in the proof of Theorem 8.6 on input sample after adding a (poly(Bd) bit rational truncation of) an independent sample from $N(0, \lambda I)$ to each $y_i \in Y$. This allows us to effectively assume that the smallest eigenvalue of the unknown covariance is λ and thus our analysis applies.

As a result, we obtain a list of candidates one of which gives a good approximation (in parameter-distance) to the $\Sigma_* + \lambda I$. Observe that if Σ_* had $2^{-\mathrm{poly}(d)}$ large smallest eigenvalue, then the resulting list is already a good approximation in parameter-distance to Σ_*. If not, then, since Σ_* has rational entries of bit complexity $\leqslant B$, the determinant of the sublattice of which Σ_* is a gram matrix is at most $(Bd)^d$. Thus, by Minkowski's theorem, there must an *integer* basis $v_1, v_2, \ldots,$ with entries of bit complexity $\leqslant O(Bd^2)$ for the orthogonal lattice of Σ_*.

Let $d - r$ be the rank of Σ_*. Since we ran the algorithm above on $\Sigma_* + \lambda I$, for small enough $\lambda \ll 2^{-\mathrm{poly}(Bd^2)}$, we must have there is a candidate $\hat{\Sigma}$ in the list with r eigenvalues at most 2λ.

We take every such candidate $\hat{\Sigma}$ and consider the quadratic form on integer vectors v: $Q(v) = v^\top \hat{\Sigma} v + \sqrt{\lambda} \|v\|_2^2$. If Σ_* has an integer vector v in its kernel of length $\leqslant 2^{O(Bd^2)}$, then, the same v must satisfy $Q(v) \leqslant \lambda^{1/4}$ for the "good" candidate $\hat{\Sigma}$. Thus, using the LLL Algorithm (Fact B.4), we can find a reduced basis for all such vectors v where $Q(v)$ is within a $2^{O(d)}$ factor from the minimum possible value of $Q(v)$ over all non-zero integer vectors v. If for such a v, $Q(v) > 2^{\Omega(d)}\lambda^{1/4}$, we know that the the corresponding $\hat{\Sigma}$ couldn't possibly be a good candidate. On the other hand, for any v such that $Q(v) \leqslant 2^{O(d)}\lambda^{1/4}$, we must have $v^T\hat{\Sigma}v < 2^{O(d)}\lambda^{1/4}$ and $\|v\|_2 \leqslant 2^{O(d)}\lambda^{-1/2}$ by our choice of $Q(v)$. Further, the projection of any such v on to $\ker(\Sigma_*)$ is either 0 or has magnitude at least $2^{O(Bd)}$ because of Proposition B.1. But if it were the latter, $Q(v) \gg 2^{\Omega(d)}\lambda^{1/4}$. Thus, if $\hat{\Sigma}$ were a good candidate, it must be that every v in our reduced basis is in the kernel of Σ_*. We can now project the candidate $\hat{\Sigma}$ on to the orthogonal complement of the subspace defined by the reduced basis. This projection can be done exactly over rationals of bit complexity $\mathrm{poly}(Bd)$. For any "good" candidate $\hat{\Sigma}$, this will ensure that the range space of $\hat{\Sigma}$ is *exactly* equal to that of Σ_* as we desired.