ROBUST LEARNING IN HIGH DIMENSION

A Dissertation
Presented to
The Academic Faculty

By

He Jia

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computer Science
Algorithms, Combinatorics, and Optimization

Georgia Institute of Technology

May 2024

ROBUST LEARNING IN HIGH DIMENSION

Thesis committee:

Dr. Santosh Vempala (Advisor)
School of Computer Science
Georgia Institute of Technology

Dr. Ilias Diakonikolas
Department of Computer Sciences
University of Wisconsin-Madison

Dr. Pravesh Kothari (Reader)
Computer Science Department
Princeton University

Dr. Cheng Mao
School of Mathematics
Georgia Institute of Technology

Dr. Ashwin Pananjady
H. Milton Stewart School of Industrial and Systems Engineering and the School of Electrical and Computer Engineering
Georgia Institute of Technology

Date approved: February 16, 2024

To my parents.

ACKNOWLEDGMENTS

I would like to begin by expressing my deepest appreciation to my advisor, Santosh Vempala, for his encouragement and patience. His positivity and enthusiasm in our research have been a constant source of motivation even during moments of self-doubt. I would also like to thank Anup Rao for hosting me as an intern at Adobe Research. I thank Anand Louis for his support during my visit at IISc. I am grateful to my committee members, Ilias Diakonikolas, Pravesh Kothari, Cheng Mao, Ashwin Pananjady and Santosh Vempala. Special thanks to Pravesh Kothari for serving as a reader for this thesis.

My academic journey has been enriched by collaborations with outstanding colleagues: Ainesh Bakshi, Ilias Diakonikolas, Daniel Kane, Pravesh Kothari, Aditi Laddha, Yin Tat Lee and Santosh Vempala. Discussions with them have sparked new ideas and insights.

I would like to thank my fellow labmates and friends at the Georgia Tech theory group, as well as the theory students at IISc. I would also like to thank my friends for their support and companionship, especially during the challenging times of the pandemic. In particular, I would like to thank Chuning Wang for instilling in me the fun of mathematics.

Finally, I would like to thank my parents for their unconditional love. They always support all my choices.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis focuses on designing efficient and robust algorithms for learning high-dimensional distributions, particularly in the presence of adversarial corruptions. We address existing challenges in robust learning in high dimension and establish new connections between robust statistics and algorithmic solutions. In particular, we give polynomial-time algorithms for the following two open problems: robust learning of Gaussian Mixture Models (GMMs) and robust learning of affine transformations of hypercubes.

For GMMs, we give a polynomial-time algorithm for robustly learning a mixture of k arbitrary Gaussians in $\mathbb{R}^d$, for any fixed k, in the presence of a constant fraction of arbitrary corruptions. This resolves the main open problem in several previous works on algorithmic robust statistics, which addressed the special cases of robustly estimating (a) a single Gaussian, (b) a mixture of TV-distance separated Gaussians, and (c) a uniform mixture of two Gaussians. Our main tools are an efficient partial clustering algorithm that relies on the sum-of-squares method (building on previous work), and a novel tensor decomposition algorithm that allows errors in both Frobenius norm and low-rank terms.

The second contribution is a polynomial-time algorithm for robustly learning an unknown affine transformation of the standard hypercube from samples, an important and well-studied setting for independent component analysis (ICA). Our algorithm, based on a new method that goes beyond the limitations of the method of moments, achieves asymptotically optimal error in total variation distance.

Lastly, we introduce a faster algorithm for the isotropic transformation of convex bodies, whose complexity is directly tied to the KLS (Kannan-Lovász-Simonovits) constant. This serves as a useful tool to high-dimensional learning and volume computation.

CHAPTER 1
INTRODUCTION

Given a collection of observations and a class of models, the central goal of machine learning is to design efficient algorithms for finding the model in the class that best fits the data. While this task is simple in low dimension, real-world datasets often pose challenges due to their high dimensionality, leading to computational intractability. Recent progress has been made in tackling these challenges in various machine learning problems, including learning mixture models [1, 2, 3, 4, 5], phylogenetic trees [6, 7], HMMs [8], topic models [9, 10], and independent component analysis [11, 12]. These algorithms rely on the classical assumption of input data being independently and identically distributed (i.i.d.) samples generated by a statistical model in the given class. However, this simplifying assumption is only approximately valid, as real datasets are typically exposed to various sources of systematic noise. Consequently, there is a growing need for the development of efficient algorithms that can learn models from noisy datasets with provable error guarantees. In this thesis, we address the following question for different classes of distributions:

Question 1.0.1. *Let $\mathcal{F}$ be a family of distributions on $\mathbb{R}^d$. Suppose we are given samples generated from the following process: First, n samples are drawn from some unknown distribution X in $\mathcal{F}$. Then, an adversary is allowed to arbitrarily corrupt an ε-fraction of the samples. Is there a poly-time robust learning algorithm that finds a distribution X' in $\mathcal{F}$ that is $f(\varepsilon)$-close, in total variation distance, to X ?*

Here $f(\varepsilon)$-close means that we would like to achieve a dimension-independent error guarantee even in high dimension. Note that the information-theoretical bound on the total variation distance error is ε, i.e., if the samples are ε-corrupted by the adversary, then it is impossible to learn the unknown distribution within a total variation distance better than ε. In the following section, we will discuss more about this adversarial robust setting.

1.1 Robust Statistics

Robust statistics are initiated in the pioneering works of [13, 14]. More general contamination models, with respect to other metrics, were studied in [15, 16]. In Section 1.1.1, we introduce some of the contamination models. Robust statistics focuses on designing of outlier-robust estimators that can tolerate a constant fraction of corrupted samples, with accuracy independent of the dimension. We have no control over this small fraction of corrupted samples. These corruptions might include very extreme outliers, and they will not necessarily come from any model that we could predict ahead of time. These contamination models can catch not only any kinds of corruptions, as long as these corruptions do not affect a large fraction of the input, but also slight departures from model assumption.

In low dimension, we can achieve robust mean estimation by using the empirical median instead of the empirical mean, and using the median absolute deviation for robust estimates of the variance. In high dimension, robust statistics has led to the definition of robust statistical parameters such as the Tukey median [17], a high-dimensional analogue of the median, and the geometric median. These classical parameters are statistically sound, but either computationally intractable [18] in high dimension or have error factors that grow polynomially with the dimension. Similar hardness results have been shown in [19, 20] for essentially all known classical estimators in robust statistics. These limitations, which apply to almost all classical estimators in robust statistics, highlight the need for more efficient solutions in high-dimensional scenarios.

Despite significant effort over several decades starting with important early works of Tukey and Huber in the 60s, even for the most basic high-dimensional estimation tasks, all known computationally efficient estimators were until fairly recently highly sensitive to outliers. This state of affairs changed with two independent works from the TCS community [21, 22], which gave the first computationally efficient and outlier-robust learning algorithms for a range of "simple" high-dimensional probabilistic models. In particular, these works developed efficient robust estimators for a single high-dimensional Gaussian distribution with unknown mean and covariance. Since these initial algorithmic works [21, 22], we have witnessed substantial research progress on algorithmic aspects of robust high-dimensional estimation by several communities of researchers, including TCS, machine learning, and mathematical statistics. The ensuing theory has led to deep connections between efficient robust statistics and algorithmic certificates of well-studied concentration and anti-concentration properties of probability distributions and has resulted in robust estimation algorithms for a range of central problems including moment estimation [12], linear regression [23, 24], clustering [25, 26, 27], and list-decodable learning [28, 29, 30, 31, 30, 32]. In all these cases, nearly optimal guarantees are known, asymptotically matching the statistical lower bounds for the error of the estimated parameters.

In this thesis, we develop novel techniques that lead to computationally efficient and robust algorithms for high-dimensional distribution learning, including learning mixtures of k arbitrary Gaussians and learning affine transformations of hypercubes. We will define both models and give more details in Section 1.2.

1.1.1 Contamination Models

Before getting into our results, there is a fundamental question to answer: how does the adversary manipulate the data? In this section, we define the models of robustness that we study.

We begin by introducing the contamination model originally proposed by Huber in [14].

Definition 1.1.1 (Additive Contamination Model (Huber Model))**.** Given a parameter $0 < \varepsilon < 1/2$ and a distribution X on inliers, we say that one can sample from X with ε-additive contamination if one can obtain independent samples from a distribution F of the following form: a sample from F returns a sample from X with probability $1-\varepsilon$, and otherwise returns a sample from some unconstrained and unknown error distributions E.

In the Huber's model, the adversary is only allowed to add outliers to the datasets. All the originally generated points from the unknown target distribution remain unchanged. The following contamination model generalizes the Huber's model.

Definition 1.1.2 (Total Variation Contamination Model)**.** Given a parameter $0 < \varepsilon < 1/2$ and a class of distributions $\mathcal{F}$ on $\mathbb{R}^d$, the *adversary* operates as follows: The algorithm specifies the number of samples n. The adversary knows the true target distribution $X \in \mathcal{F}$ and selects a distribution F such that $d_{\mathsf{TV}}(F, X) \leqslant \varepsilon$. Then n i.i.d. samples are drawn from F and are given as input to the algorithm.

Intuitively, the parameter ε in Definition 1.1.2 quantifies the power of the adversary. The total variation contamination model is strictly stronger than Huber's contamination model.

Definition 1.1.3 (Strong Comtamination Model, [21])**.** Given a parameter $0 < \varepsilon < 1/2$ and a distribution X on inliers, an algorithm receives samples from X with ε-contamination as follows: the algorithm specifies an integer number of samples n, and n samples are drawn independently from the distribution X. An adversary is then allowed to inspect these samples, remove up to εn of them and replace them with arbitrary points. The modified set of n points are then given to the algorithm.

The strong contamination model is a strengthening of the total variation contamination, where an adversary can see the clean samples and then arbitrarily replace an ε-fraction of these points to obtain an ε-corrupted set of samples.

1.2 Summary of Contributions

This thesis answers Question 1.0.1 for Gaussian mixtures models (Section 1.2.1) and affine transformation of hypercubes (Section 1.2.2). We also establish a useful tool for high-dimensional learning problems: a fast algorithm for rounding any given convex body into isotropic, in Section 1.2.3.

1.2.1 Robustly Learning of Gaussian Mixtures

One of the main original motivations for the development of algorithmic robust statistics within the TCS community was the problem of learning high-dimensional Gaussian mixture models. A *Gaussian mixture model (GMM)* is a convex combination of Gaussian distributions, i.e., a distribution on $\mathbb{R}^d$ of the form $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, where the weights w_i, mean vectors μ_i, and covariance matrices Σ_i are unknown. GMMs are *the* most extensively studied latent variable model in the statistics and machine learning literatures, starting with the pioneering work of Karl Pearson in 1894 [33], which introduced the method of moments in this context.

In the absence of outliers, a long line of work initiated by Dasgupta [34, 35, 36, 37, 38, 39] gave efficient clustering algorithms for GMMs under various separation assumptions. Subsequently, efficient learning algorithms were obtained [2, 3, 4, 40] under minimal information-theoretic conditions. Specifically, Moitra and Valiant [3] and Belkin and Sinha [4] designed the first polynomial-time learning algorithms for arbitrary Gaussian mixtures with any fixed number of components. These works qualitatively characterized the complexity of this fundamental learning problem in the noiseless setting. Alas, all aforementioned algorithms are very fragile in the presence of corrupted data. Specifically, a *single* outlier can completely compromise their performance.

Developing efficient learning algorithms for high-dimensional GMMs in the more realistic *outlier-robust* setting has turned out to be significantly more challenging. This was both one of the original motivations and the main open problem in the initial robust statistics works [21, 22]. We note that [21] developed a robust density estimation algorithm for mixtures of *spherical* Gaussians — a very special case of our problem where the covariance of each component is a multiple of the identity — and highlighted a number of key technical obstacles that need to be overcome in order to handle the general case. Since then, a number of works have made algorithmic progress on important special cases of the general problem. These include faster robust clustering for the spherical case under minimal separation conditions [41, 42, 27], robust clustering for separated (and potentially non-spherical) Gaussian mixtures [43, 44], and robustly learning *uniform* mixtures of two arbitrary Gaussian components [45].

This progress notwithstanding, the algorithmic task of robustly learning a mixture of a constant number (or even two) arbitrary Gaussians (with arbitrary weights) has remained a central open problem in this field, as highlighted recently [46].

Our main contribution is to solve the above open problem as following:

Theorem 1.2.1 (Robustly Learning k-Mixtures with poly(ε)-error, Informal, see Corollary 4.6.1)**.** *There is an algorithm with the following behavior: Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)} \operatorname{poly}_k(1/\varepsilon)$ samples from a distribution F on $\mathbb{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $\operatorname{poly}(n) \operatorname{poly}_k(1/\varepsilon)$ and outputs a k-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^k \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ such that with high probability we have that $d_{\mathsf{TV}}(\widehat{\mathcal{M}}, \mathcal{M}) \leqslant O(\varepsilon^{c_k})$, where c_k depends only on k.*

Theorem 1.2.1 gives a polynomial-time *robust proper learning* algorithm, with dimension-independent error guarantee, for *arbitrary* k-GMMs, for any fixed k. We make a few important remarks about Theorem 1.2.1.

1. *Sample Complexity*: Our algorithm succeeds whenever the sample size n satisfies $n \geqslant n_0 = d^{O(k)}/\operatorname{poly}(\varepsilon)$. Statistical query lower bounds [47] suggest that $d^{\Omega(k)}$ samples are necessary for efficiently learning GMMs, even for approximation to constant accuracy in the simpler setting without outliers and under the more restrictive *clustering* setting (where components are pairwise well-separated in total variation distance). This provides some evidence that the sample-time tradeoff achieved by Theorem 1.2.1 is qualitatively optimal (within absolute constant factors in the exponent).

2. *Handling Arbitrary Weights and Arbitrary Covariances*: The algorithm of Theorem 1.2.1 succeeds *without any assumptions* on the weights of the mixture components, nor on the variances of components, modulo basic limitations posed by

numerical computation issues. Prior work [43, 44, 45], as well as the concurrent work [48], cannot handle the case of general weights — even for the case of $k = 2$ components.

3. *Error Guarantee*: The final error guarantee of our basic algorithm is ε^{c_k}, for some function c_k that depends only on k and goes to 0 when k increases. Importantly, for any fixed k, the final error guarantee of our algorithm is a fixed polynomial in ε and independent of the dimension d. This turns out to be quantitatively close to best possible for any robust proper learning algorithm.

We also show that the *same* algorithm as in Theorems 1.2.1 and 4.6.1 actually achieve parameter estimation, i.e., imply that the recovered mixture of Gaussians is close in *parameter* distance to the unknown target mixture. Such parameter estimation results are usually stated under the assumption that every pair of components of the unknown mixture are separated in total variation distance. In this work, we provide a stronger version of this parameter estimation guarantee.

More specifically, in the theorem below, we prove that whenever the components of the input mixture can be clustered together into some groups such that all mixtures in a group are close (and thus, indistinguishable), there exists a similar clustering of the output mixture such that all parameters (weight, mean, and covariances) of each cluster are close within $\mathrm{poly}_k(\varepsilon)$ in total variation distance. In particular this means that for each significant component of the input mixture, there is a component of the output mixture with very close parameters.

Theorem 1.2.2 (Parameter Recovery, See Theorem 4.7.2)**.** *Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\,\mathrm{poly}_k(1/\varepsilon)$ samples from a distribution F on $\mathbb{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $\mathrm{poly}(n)\,\mathrm{poly}_k(1/\varepsilon)$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^{k'} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability we have that there exists a partition of $[k]$ into $k' + 1$ sets*

$R_0, R_1, \ldots, R_{k'}$ *such that*

1. *Let* $W_i = \sum_{j \in R_i} w_j$, $i \in \{0, 1, \ldots, k'\}$. *Then, for all* $i \in [k']$, *we have that*

$$|W_i - \widehat{w}_i| \leqslant \mathrm{poly}_k(\varepsilon), \text{ and}$$

$$d_{\mathsf{TV}}(\mathcal{N}(\mu_j, \Sigma_j), \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)) \leqslant \mathrm{poly}_k(\varepsilon) \quad \forall j \in R_i \ .$$

2. *The total weight of exceptional components in* R_0 *is* $W_0 \leqslant \mathrm{poly}_k(\varepsilon)$.

Our main tools are an efficient partial clustering algorithm that relies on the sum-of-squares method, and a novel tensor decomposition algorithm that allows errors in both Frobenius norm and low-rank terms.

1.2.2 Affine Transformations

In this section, we revisit the classical problem of learning a parallelopiped from uniform samples and focus on designing *robust* algorithms that tolerate a dimension-independent constant fraction of arbitrary outliers. Equivalently, given a outlier-corrupted version of points $x \in \mathbb{R}^d$ obtained by applying an unknown affine transformation $x \to As + b$ to uniform samples s from the hypercube $[-1, 1]^d$, our goal is to obtain estimates of b, A.

Learning parallelopipeds (and affine transforms, more generally) is a major topic of interest in signal processing with a long history [49, 50, 51], and many influential algorithmic methods (see e.g., [52, 53, 54]). It is also a canonical example of learning an intersection of halfspaces (in our case, the number of halfspaces or facets of the convex set grow linearly in the dimension d) in the distribution-free PAC model from positive examples. In general, learning an intersection of even two halfspaces (even with positive and negative examples) is already a notoriously hard problem [55, 56].

The study of learning parallelopipeds from uniform samples was begun in a classical work of Frieze, Jerrum and Kannan [57]. In addition to applications in signal processing, their main motivation was finding algorithms learning intersections of super-constantly many halfspaces in a canonical special case given the strong hardness results for the problem in general. Simply estimating the empirical mean and covariance of the transformed samples can be used to learn the parallelopiped up to a rotation. Their key (and consequential to several later works on the topic) observation was the facet normals of the unknown rotations are the *local maximizers* of natural quartic polynomial associated with the fourth moments of the input samples. This observation, that opens up a natural route to designing algorithms for the problem using the *method of moments*, was later extensively generalized to numerous variants of the problem [58, 59, 60, 61, 62, 63].

What if the input sample contains a small fraction of outliers? Specifically, we would like an algorithm that takes input an ε-corruption (i.e., an arbitrary ε fraction of the original independent sample is arbitrarily and adversarially corrupted) of an independent sample from $x = As + b$ for $s \sim [-1, 1]^d$. Information-theoretically, we can naturally expected to recover an affine transformation $\widehat{A}s + \hat{b}$ so that the resulting distribution is ε-close in total variation to the ground truth. Could an efficient algorithm obtain such a guarantee?

Despite the progress and understanding in *robust statistics* finding a robust algorithm for learning an affine transformation has thus far evaded a resolution. As we will next explain, a key reason for this is the failure of (*robust*) method of moments in our setting: a primitive that has served as a workhorse of basic subroutines in almost all the works mentioned above.

Consider a unit hypercube whose center is shifted to an unknown point μ. Now the problem simply consists of estimating μ. Robust mean estimation algorithms will solve this problem to within error $O(\varepsilon)$, i.e., one can efficiently find $\tilde{\mu}$ s.t. $\|\mu - \tilde{\mu}\|_2 = O(\varepsilon)$ and this is the best possible bound. However, suppose the center of the hypercube is the origin, and the estimated center has all coordinates equal to $\varepsilon/\sqrt{d}$. Then, the TV distance between the two corresponding cubes is $1 - (1 - \frac{\varepsilon}{\sqrt{d}})^d \simeq 1 - e^{-\varepsilon\sqrt{d}}$, very far from the best possible TV distance of $O(\varepsilon)$. On the other hand, if one could estimate the mean with $O(\varepsilon)$ error in ℓ_1 norm, this would result in a TV distance bound of $O(\varepsilon)$. This is simply because one can bound the TV distance as the sum over the distances along the marginals, and for each marginal it is bounded by the distance between the means. However, estimating the mean within ℓ_1 error ε is impossible in general, e.g., for a Gaussian. Indeed, almost all the recently developed methods in robust statistics naturally provide guarantees for mean estimation in ℓ_2 norm and yield no useful guarantees in our setting. Furthermore, in this example with the mean shift, we can see even if all higher moments of two hypercubes are identical, the TV distance between them can be large.

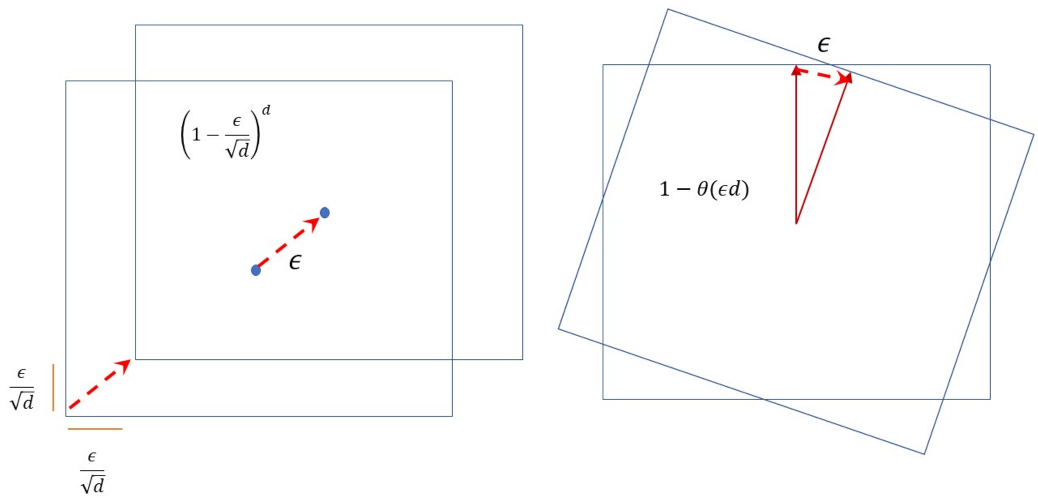

Figure 1.1: Shifting a hypercube in Euclidean norm or rotating can create very large TV distance.

The robust covariance and higher-moment estimation methods of [42] can be used to robustly learn the columns of A in the unknown affine transformation each to within ε error in Euclidean norm (after normalizing the covariance to be the identity). However, even if we assume the mean shift is zero, ε-closeness in columns of A does not give us the dimension independent TV error guarantee. This failure of method of moments persists even when restricted to the special case when the unknown affine transform is simply a rotation (in particular, simply upgrading the mean estimation procedure above will not suffice to obtain our target guarantees). In Theorem 5.8.1 (see Section 5.8), we give a simple example to prove that given estimates of the first t moments of the (uncorrupted) distribution up to an relative error of η, any algorithm must provably incur a TV error that grows polynomially in the dimension d so long as $\eta = \Omega(1/d^{O(1)})$ (recall that off-the-shelf robust moment estimators cannot yield an $\eta = o(1)$) for any $t = o(\log d / \log \log d)$).

The main contribution of our work is a new polynomial time algorithm, that goes beyond the method of moments, and allows robustly learning affine transformations with an information-theoretically optimal estimation error (up to an absolute constant factor loss). Specifically, we prove the following.

Theorem 1.2.3. *Given $X = \{x^{(1)}, \dots, x^{(n)}\}$, an ε-corrupted sample of points from an unknown parallelopiped $H = A[-1,1]^d + b$ in $\mathbb{R}^d$, there is a polynomial-time algorithm that outputs a parallelopiped $\widehat{H} = \widehat{A}[-1,1]^d + \hat{b}$ s.t. $d_{TV}(\widehat{H}, H) = O(\varepsilon)$.*

Equivalently, our algorithm produces an estimate $\widehat{A}, \hat{b}$ so that

$$\sum_{i=1}^{d} \frac{\left\|a_{(i)} - \hat{a}_{(\pi(i))}\right\|_2}{\left\|a_{(i)}\right\|_2} + \sum_{i=1}^{d} \frac{\left|b_i - \hat{b}_i\right|}{\left\|a_{(i)}\right\|_2} = O(\varepsilon)$$

where $a_{(i)}$ are rows of A and $\hat{a}_{(i)}$ are rows of $\widehat{A}$ and π is some global permutation $\pi : [d] \to [d]$.

Our algorithm is based on a new approach to ICA, even without outliers, that does not rely the fourth (or any constant order) moments as in prior works. As we discuss more in Chapter 5, a key contribution of our work is the development of a new *certificate* of correctness of the estimated affine transformation that does not rely on low order moments.

1.2.3 Isotropic Transformations for Convex Bodies

We give a new algorithm for the isotropic transformation of any given convex body. This improved rounding algorithm directly implies a faster volume algorithm. The complexity of the algorithm is $O^*(n^3\psi^2)$, directly determined by the best possible KLS constant ψ for isotropic logconcave densities in $\mathbb{R}^n$. KLS constant is a crucial concept associated with KLS conjecture, as defined in Section 6.1.2. This is the first direct reduction of the complexity of volume computation to the KLS constant. With the current bound of $\psi = \widetilde{O}(1)$ (see Secion 6.1.2), the complexity is $O^*(n^3)$ and implies volume computation in the same complexity for any convex body. No further improvement is in sight for the general setting.

Theorem 1.2.4 (Rounding)**.** *There is a randomized algorithm that takes as input a convex body $K \subset \mathbb{R}^n$ given by a membership oracle with initial point $x_0 \in K$, bounds r, R>0 s.t.,*

$$x_0 + rB_n \subseteq K \subseteq RB_n$$

and with probability $1 - \delta$, computes an affine transformation T s.t., TK is in near-isotropic position, i.e., for x sampled uniform from TK,

$$I \preccurlyeq \mathcal{C}ov(xx^\top) \preccurlyeq 2I.$$

The algorithm takes $O(n^3\psi^2 \log^c(Rn/r)\log(1/\delta))$ oracle queries and $O(n^2)$ time per oracle query.

The algorithm is based on the following ideas. First, if the covariance matrix is skewed (i.e., some eigenvalues are much larger than others), this can be detected via a small random sample. The roundness can then be improved by scaling up the subspace of small eigenvalues. But how to sample a highly skewed convex body? To break this chicken-and-egg problem, we use a sequence of convex bodies obtained by intersecting the original set with balls of increasingly larger radii:

$$K_t = K \cap B(0, t),$$

with t starting at r and doubling till it reaches R. In each outer iteration, we compute a transformation that makes the set K_t (nearly) isotropic. When K_t is isotropic, we show that K_{2t} is well-rounded (Lemma 6.3.6). Hence, the trace of the covariance of K_{2t} is $O(n)$. However, its eigenvalues could be widely varying. To make K_{2t} isotropic, we will estimate the larger eigenvalues using a small random sample (Lemma 6.2.11). The second idea is that scaling up the small eigenvalues nearly doubles the size of the ball contained inside (Lemma 6.3.4), while having a mild effect on the higher norms of the covariance (Lemma 6.3.3). The latter concept is where KLS comes in. We show that if the KLS constant for isotropic logconcave distributions in $\mathbb{R}^n$ is bounded by $\psi_n^2 \lesssim n^{\frac{1}{p}}$ for all n, then for any logconcave density q with covariance matrix A (not necessarily identity), the KLS constant is bounded as $\psi_q^2 \lesssim \|A\|_p \log n$ (Theorem 6.2.8). As we outline below, this improves the sampling time in each inner iteration.

In the beginning, when the ball contained inside K_t is small, we will use a few samples to get a coarse estimate of the larger eigenvalue directions and scale up the orthogonal subspace. The sampling time is higher since the roundness parameter of K_t is higher. The time per sample is roughly $n^2\frac{\|A\|_p}{r^2}$ when r is the radius of the ball inside and A is the covariance matrix for the uniform density q on K_t. As we increase r, by a constant factor in each step, the norm of the covariance grows much more slowly, and so the sampling time decreases. Meanwhile, we use more samples in each step, roughly r^2, and this trade-off keeps the overall time at $n^{3+\frac{1}{p}} \simeq n^3\psi^2$. During the process we need a warm start for each phase; we achieve this in $\widetilde{O}(n^3)$ steps using the Gaussian Cooling algorithm. It is known that $\psi_q^2 \lesssim \|A\|_p$ holds for $p = 2$ [64], giving the $\widetilde{O}(n^{3.5})$ complexity for rounding. Our general reduction establishes that any future improvement to higher p would imply faster rounding and volume estimation.

CHAPTER 2

PRELIMINARIES

2.1 Logconcave Functions

Definition 2.1.1. A function $f : \mathbb{R}^n \to \mathbb{R}_+$ is *logconcave* if its logarithm is concave along every line, i.e., for any $x, y \in \mathbb{R}^n$ and any $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \geqslant f(x)^\lambda f(y)^{1-\lambda}. \tag{2.1.1}$$

Many common probability distributions have logconcave densities, e.g., Gaussian, exponential, logistic, and gamma distributions; indicator functions of convex sets are also logconcave. Logconcavity is preserved by product, min and convolution (in particular marginals of logconcave densities are logconcave).

Definition 2.1.2. A distribution D is said to be *isotropic* if

$$\mathop{\mathbb{E}}_D(x) = 0 \quad \text{and} \quad \mathbb{E}_D(xx^\top) = I.$$

We say a convex body is isotropic if the uniform distribution over it is isotropic. Any distribution with bounded second moments can be brought to an isotropic position by an affine transformation.

2.2 Robust Estimation

We will use the following robust mean estimation algorithm for bounded covariance distributions [65]:

Theorem 2.2.1 (Robust Mean Estimation for Bounded Covariance Distributions)**.** *There is a* poly(n) *time algorithm that takes input an* ε*-corruption* Y *of a collection of* n *points*

$X \subseteq \mathbb{R}^d$, *and outputs an estimate* $\hat{\mu}$ *satisfying*

$$\left\|\mathop{\mathbb{E}}_{x\sim_u X} x - \hat{\mu}\right\|_2 \leqslant O(\sqrt{\varepsilon}) \left\|\mathop{\mathbb{E}}_{x\sim_u X} (x - \mathop{\mathbb{E}}_{x\sim_u X} x)(x - \mathop{\mathbb{E}}_{x\sim_u X} x)^\top\right\|_2 .$$

We will use the following robust covariance estimation algorithm from [12]:

Theorem 2.2.2 (Robust Covariance Estimation, [12])**.** *For every $C > 0$, $\varepsilon > 0$ and even $k \in \mathbb{N}$ such that $Ck\varepsilon^{1-2/k} \leqslant c$ for some small enough absolute constant c, there exists a polynomial-time algorithm that given an (corrupted) sample S outputs an estimate of the covariance $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ with the following guarantee: there exists $n_0 \geqslant (C+d)^{O(k)}/\varepsilon$ such that if S is an ε-corrupted sample with size $|S| \geqslant n_0$ of a k-certifiably C-subgaussian distribution D over $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$, then with high probability:*

$$(1-\delta)\Sigma \succeq \hat{\Sigma} \succeq (1+\delta)\Sigma$$

for $\delta \leqslant O(Ck)\varepsilon^{1-2/k}$.

We will also require the following robust estimation algorithm with Frobenius distance guarantees proven for certifiably hypercontractive distributions in [43]. Since we obtain estimates to the true covariance in Lowner ordering, we can obtain the subspace spanned by the inliers exactly, project on to this subspace and apply Theorem 7.1 in [43].

Theorem 2.2.3 (Robust Mean and Covariance Estimation for Certifiably Hypercontractive Distributions, Theorem 7.1 in [43])**.** *Given $t \in \mathbb{N}$, and $\varepsilon > 0$ sufficiently small so that $Ct\varepsilon^{1-4/t} \ll 1$[1], for some absolute constant $C > 0$. Then, there is an algorithm that takes input Y, an ε-corruption of a sample X of size n with mean μ_*, covariance Σ_*, and $2t$-certifiably C-hypercontractive degree-2 polynomials, runs in time $n^{O(t)}$, and outputs an estimate $\hat{\mu}$ and $\hat{\Sigma}$ satisfying:*

[1]This notation means that we needed $Ct\varepsilon^{1-2/t}$ to be at most c_0 for some absolute constant $c_0 > 0$.

1. $\left\|\Sigma_*^{\dagger/2}(\mu_* - \hat{\mu})\right\|_2 \leqslant O(Ct)^{1/2}\varepsilon^{1-1/t}$,

2. $(1-\eta)\Sigma_* \preceq \hat{\Sigma} \preceq (1+\eta)\Sigma_*$ *for* $\eta \leqslant O(Ck)\varepsilon^{1-2/t}$, *and,*

3. $\left\|\Sigma_*^{\dagger/2}\left(\hat{\Sigma} - \Sigma_*\right)\Sigma_*^{\dagger/2}\right\|_F \leqslant (C't)O(\varepsilon^{1-1/t})$,

where $C' = \max\{C, B\}$ *for the smallest possible* $B > 0$ *such that for* $d \times d$*-matrix-valued indeterminate* Q, $\left|\frac{Q}{2}\right. \left\{\mathbb{E}_{\mathcal{D}}\left(x^\top Q x - \mathbb{E}_{\mathcal{D}}\, x^\top Q x\right)^2 \leqslant B\left\|\Sigma_*^{1/2} Q \Sigma_*^{1/2}\right\|_F^2\right\}$.[2]

The following algorithm is useful to robustly find the independent components.

Theorem 2.2.4 (Theorem 1.4 in [12], see also [22])**.** *There exists a polynomial time algorithm that given an* ε*-corrupted sample* X *of* n *points in* $\mathbb{R}^d$ *drawn from a rotated cube* AH *where* $A \in \mathbb{R}^{d\times d}$ *is a orthogonal matrix with rows* $a_{(1)}, \ldots, a_{(d)}$*, outputs component estimates* $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)} \in \mathbb{R}^d$ *with the following guarantee: the components estimates satisfy with high probability, there exists a permutation* $\pi : [d] \to [d]$ *such that for any* $i \in [d]$*,*

$$\langle \hat{a}_{(i)}, a_{(\pi(i))}\rangle^2 \geqslant 1 - O(\sqrt{\varepsilon}).$$

[2]The first two guarantees here hold for the larger class of certifiably subgaussian distributions and were proven in [12] (see Theorem 1.2). Gaussian distribution (with arbitrary mean and covariance) are t-certifiably 1-subgaussian for all t and their mixtures (similar to Lemma 4.2.19 and explicitly proven in Lemma 5.4 of [12]) are t-certifiably $O(1/\alpha)$-subgaussian where α is the minimum mixing weight.

CHAPTER 3

ALGORITHM DESIGN VIA SUM-OF-SQUARES PROOFS

Semidefinite programming (SDP), as a method of convex optimization, has applications in wide range of fields including signal processing, machine learning and etc. It is also used as a relaxation technique in combinatorial optimization. Sum-of-Squares (SoS) extends semidefinite programming by providing a hierarchy of SDP relaxation for handling general (non-convex) polynomial optimization problems. In this chapter, we give a brief introduction on SoS proofs and algorithm design via SoS, especially robust algorithms.

We refer the reader to the monograph [66] and the lecture notes [67] for a detailed exposition of the sum-of-squares method and its usage in average-case algorithm design.

3.1 Notation

Let $x = (x_1, x_2, \dots, x_n)$ be a tuple of n indeterminates and let $\mathbb{R}[x]$ be the set of polynomials with real coefficients and indeterminates $x_1, \dots, x_n$. We say that a polynomial $p \in \mathbb{R}[x]$ is a *sum-of-squares (sos)* if there exist polynomials $q_1, \dots, q_r$ such that $p = q_1^2 + \cdots + q_r^2$.

3.2 Pseudo-distributions

Pseudo-distributions are generalizations of probability distributions. We can represent a discrete (i.e., finitely supported) probability distribution over $\mathbb{R}^n$ by its probability mass function $D\colon \mathbb{R}^n \to \mathbb{R}$ such that $D \geqslant 0$ and $\sum_{x \in \mathrm{supp}(D)} D(x) = 1$. Similarly, we can describe a pseudo-distribution by its mass function by relaxing the constraint $D \geqslant 0$ to passing certain low-degree non-negativity tests.

Concretely, a *level-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

$\ell/2$. (Here, the summations are over the support of D.) A straightforward polynomial-interpolation argument shows that every level-∞-pseudo distribution satisfies $D \geqslant 0$ and is thus an actual probability distribution. We define the *pseudo-expectation* of a function f on $\mathbb{R}^n$ with respect to a pseudo-distribution D, denoted $\tilde{\mathbb{E}}_{D(x)} f(x)$, as

$$\tilde{\mathbb{E}}_{D(x)} f(x) = \sum_x D(x) f(x) \,. \tag{3.2.1}$$

The degree-ℓ moment tensor of a pseudo-distribution D is the tensor $\tilde{\mathbb{E}}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes \ell}$. In particular, the moment tensor has an entry corresponding to the pseudo-expectation of all monomials of degree at most ℓ in x. The set of all degree-ℓ moment tensors of probability distribution is a convex set. Similarly, the set of all degree-ℓ moment tensors of degree-d pseudo-distributions is also convex. Unlike moments of distributions, there is an efficient separation oracle for moment tensors of pseudo-distributions.

Fact 3.2.1 ([68, 69, 70, 71])**.** *For any $n, \ell \in \mathbb{N}$, the following set has an $n^{O(\ell)}$-time weak separation oracle (in the sense of [72]):*

$$\left\{ \tilde{\mathbb{E}}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes d} \mid \text{degree-}d \text{ pseudo-distribution } D \text{ over } \mathbb{R}^n \right\} \,. \tag{3.2.2}$$

This fact, together with the equivalence of weak separation and optimization [72], allows us to efficiently optimize over pseudo-distributions (approximately) — this algorithm is referred to as the sum-of-squares algorithm. The *level-ℓ sum-of-squares algorithm* optimizes over the space of all level-ℓ pseudo-distributions that satisfy a given set of polynomial constraints (defined below).

Definition 3.2.2 (Constrained pseudo-distributions)**.** Let D be a level-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*, denoted

$D \,\big|\!\!\frac{\;\;}{r}\, \mathcal{A}$, if for every $S \subseteq [m]$ and every sum-of-squares polynomial h with $\deg h + \sum_{i \in S} \max\{\deg f_i, r\}$, we have that $\tilde{\mathbb{E}}_D h \cdot \prod_{i \in S} f_i \geqslant 0$.

We write $D \models \mathcal{A}$ (without specifying the degree) if $D \,\big|\!\!\frac{\;\;}{0}\, \mathcal{A}$ holds. Furthermore, we say that $D \,\big|\!\!\frac{\;\;}{r}\, \mathcal{A}$ holds *approximately* if the above inequalities are satisfied up to an error of $2^{-n^\ell} \cdot \|h\| \cdot \prod_{i \in S} \|f_i\|$, where $\|\cdot\|$ denotes the Euclidean norm[1] of the coefficients of a polynomial in the monomial basis.

We remark that if D is an actual (discrete) probability distribution, then we have that $D \models \mathcal{A}$ if and only if D is supported on solutions to the constraints $\mathcal{A}$. We say that a system $\mathcal{A}$ of polynomial constraints is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant M\}$. The following fact is a consequence of Fact 3.2.1 and [72]:

Fact 3.2.3 (Efficient Optimization over Pseudo-distributions)**.** *There exists an $(n+m)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system[2] $\mathcal{A}$ of m polynomial constraints in n variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately.*

Basic Facts about Pseudo-Distributions. We will use the following Cauchy-Schwarz inequality for pseudo-distributions:

Fact 3.2.4 (Cauchy-Schwarz for Pseudo-distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $x \in \mathbb{R}^d$. Then, for any degree-d pseudo-distribution $\tilde{\zeta}$, we have that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[fg] \leqslant \sqrt{\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^2]}\sqrt{\tilde{\mathbb{E}}_{\tilde{\zeta}}[g^2]}$.*

Fact 3.2.5 (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 $\tilde{\zeta}$, we have that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^{t-1}g] \leqslant (\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^t])^{\frac{t-1}{t}}(\tilde{\mathbb{E}}_{\tilde{\zeta}}[g^t])^{1/t}$.*

Corollary 3.2.6 (Comparison of Norms)**.** *Let $\tilde{\zeta}$ be a degree-t^2 pseudo-distribution over a scalar indeterminate x. Then, we have that $\tilde{\mathbb{E}}[x^t]^{1/t} \geqslant \tilde{\mathbb{E}}[x^{t'}]^{1/t'}$ for every $t' \leqslant t$.*

[1]The choice of norm is not important here because the factor 2^{-n^ℓ} swamps the effects of choosing another norm.

[2]Here, we assume that the bit complexity of the constraints in $\mathcal{A}$ is $(n+m)^{O(1)}$.

3.3 Sum-of-squares proofs

Let $f_1, f_2, \ldots, f_r$ and g be multivariate polynomials in x. 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

$$g = \sum_{S\subseteq[m]} p_S \cdot \Pi_{i\in S} f_i \,. \tag{3.3.1}$$

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. If there is a degree ℓ SoS proof that $\{f_i \geqslant 0 \mid i \leqslant r\}$ implies $\{g \geqslant 0\}$, we write:

$$\{f_i \geqslant 0 \mid i \leqslant r\} \,\big|\!\frac{\;\;}{\ell}\, \{g \geqslant 0\} \,. \tag{3.3.2}$$

For all polynomials $f, g \colon \mathbb{R}^n \to \mathbb{R}$ and for all functions $F \colon \mathbb{R}^n \to \mathbb{R}^m$, $G \colon \mathbb{R}^n \to \mathbb{R}^k$, $H \colon \mathbb{R}^p \to \mathbb{R}^n$ such that each of the coordinates of the outputs are polynomials of the inputs, we have the following inference rules.

The first one derives new inequalities by addition or multiplication:

$$\frac{\mathcal{A} \,\big|\!\frac{\;\;}{\ell}\, \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A} \,\big|\!\frac{\;\;}{\ell}\, \{f + g \geqslant 0\}}, \frac{\mathcal{A} \,\big|\!\frac{\;\;}{\ell}\, \{f \geqslant 0\}, \mathcal{A} \,\big|\!\frac{\;\;}{\ell'}\, \{g \geqslant 0\}}{\mathcal{A} \,\big|\!\frac{\;\;}{\ell+\ell'}\, \{f \cdot g \geqslant 0\}} \,. \tag{3.3.3}$$

The next one derives new inequalities by transitivity:

$$\frac{\mathcal{A} \,\big|\!\frac{\;\;}{\ell}\, \mathcal{B}, \mathcal{B} \,\big|\!\frac{\;\;}{\ell'}\, C}{\mathcal{A} \,\big|\!\frac{\;\;}{\ell\cdot\ell'}\, C} \,. \tag{3.3.4}$$

Finally, the last rule derives new inequalities via substitution:

$$\frac{\{F \geqslant 0\} \,\big|\!\frac{\;\;}{\ell}\, \{G \geqslant 0\}}{\{F(H) \geqslant 0\} \,\big|\!\frac{\;\;}{\ell\cdot\deg(H)}\, \{G(H) \geqslant 0\}} \,. \tag{substitution}$$

Low-degree sum-of-squares proofs are sound and complete if we take low-level pseudo-distributions as models. Concretely, sum-of-squares proofs allow us to deduce properties of pseudo-distributions that satisfy some constraints.

Fact 3.3.1 (Soundness)**.** *If $D \models_r \mathcal{A}$ for a level-ℓ pseudo-distribution D and there exists a sum-of-squares proof $\mathcal{A} \vdash_{r'} \mathcal{B}$, then $D \models_{r \cdot r' + r'} \mathcal{B}$.*

If the pseudo-distribution D satisfies $\mathcal{A}$ only approximately, soundness continues to hold if we require an upper bound on the bit-complexity of the sum-of-squares $\mathcal{A} \vdash_{r'} B$ (i.e., the number of bits required to write down the proof). In our applications, the bit complexity of all sum-of-squares proofs will be $n^{O(\ell)}$ (assuming that all numbers in the input have bit complexity $n^{O(1)}$). This bound suffices in order to argue about pseudo-distributions that satisfy polynomial constraints approximately.

The following fact shows that every property of low-level pseudo-distributions can be derived by low-degree sum-of-squares proofs.

Fact 3.3.2 (Completeness)**.** *Suppose that $d \geqslant r' \geqslant r$ and $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A} \vdash \{\sum_{i=1}^n x_i^2 \leqslant B\}$ for some finite B. Let $\{g \geqslant 0\}$ be a polynomial constraint. If every degree-d pseudo-distribution that satisfies $D \models_r \mathcal{A}$ also satisfies $D \models_{r'} \{g \geqslant 0\}$, then for every $\varepsilon > 0$, there is a sum-of-squares proof $\mathcal{A} \vdash_d \{g \geqslant -\varepsilon\}$.*

Basic Sum-of-Squares Proofs. We will require the following basic SoS proofs.

Fact 3.3.3 (Operator norm Bound)**.** *Let A be a symmetric $d \times d$ matrix and v be a vector in $\mathbb{R}^d$. Then, we have that*

$$\vdash_2^{v} \left\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 \right\} .$$

Fact 3.3.4 (SoS Hölder's Inequality)**.** *Let f_i, g_i, for $1 \leqslant i \leqslant s$, be scalar-valued indeterminates. Let p be an even positive integer. Then,*

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

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

Fact 3.3.5 (SoS Almost Triangle Inequality)**.** *Let $f_1, f_2, \dots, f_r$ be indeterminates. Then, we have that*

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

Fact 3.3.6 (SoS AM-GM Inequality, see Appendix A of [73])**.** *Let $f_1, f_2, \dots, f_m$ be indeterminates. Then, we have that*

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

Lemma 3.3.7 (Spectral SoS Proofs)**.** *Let A be a $d \times d$ matrix. Then for d-dimensional vector-valued indeterminate v, we have:*

$$\frac{v}{2} \left\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 \right\} .$$

Proof. Note that v is the only variable in the proof here (A is a matrix of constants). We note that $A \leqslant \|A\|_2 I$ or $\|A\|_2 I - A$ is PSD and thus $\|A\|_2 I - A = QQ^\top$ for some $d \times d$ matrix Q. Thus, $\|Qv\|_2^2 = v^\top (\|A\|_2 I - A) v = \|A\|_2 \|v\|_2^2 - v^\top A v$. Thus, $\|A\|_2 \|v\|_2^2 - v^\top A v$ is a sum of squares polynomial (namely $\|Qv\|_2^2$) in variable v. This completes the proof.

□

Fact 3.3.8 (Cancellation within SoS, Lemma 9.2 [43])**.** *Let a, C be scalar-valued indeterminates. Then,*

$$\{a \geqslant 0\} \cup \left\{ a^t \leqslant C a^{t-1} \right\} \frac{a, C}{2t} \left\{ a^{2t} \leqslant C^{2t} \right\} .$$

Lemma 3.3.9 (Frobenius Norms of Products of Matrices)**.** *Let B be a $d \times d$ matrix valued* *indeterminate for some* $d \in \mathbb{N}$. *Then, for any* $0 \preceq A \preceq I$,

$$\Big|\frac{B}{2} \left\{\|AB\|_F^2 \leqslant \|B\|_F^2\right\} ,$$

and,

$$\Big|\frac{B}{2} \left\{\|BA\|_F^2 \leqslant \|B\|_F^2\right\} ,$$

Proof. The proof of the second claim is similar so we prove only the first. We have:

$$\Big|\frac{B}{2} \left\{\|B\|_F^2 = \|(A + I - A)B\|_F^2 = \|AB\|_F^2 + \|(I - A)B\|_F^2 + 2\mathrm{tr}((I - A)BB^\top A)\right\}$$

Now, $A - A^2 \succeq 0$, thus, $A - A^2 = RR^\top$ for some $d \times d$ matrix R. Thus, $\mathrm{tr}((A - A^2)BB^\top) = \mathrm{tr}(RR^\top BB^\top) = \|BR\|_F^2$ - a sum of squares polynomial of degree 2 in indeterminate B. Thus, $\Big|\frac{B}{2} \left\{\mathrm{tr}((A - A^2)BB^\top) \geqslant 0\right\}$. □

3.4 Robust Algorithms using Sum-of-Squares

SoS can be used as a powerful framework of designing algorithms. In this section, we describe the framework at a high-level and give some successful examples in algorithmic robust statistics that use SoS.

3.4.1 SoS Certificates

The key of designing an SoS algorithm is to find certificates of correctness that can be verified in the SoS proof system. The certificates should be a set of polynomial inequalities such that

- (Completeness) every correct solution satisfies the inequalities.
- (Soundness) every solution that satisfies the inequalities is approximately correct.
- Soundness property should have a low-degree SoS proof.

After we get an SoS certificate for the problem, we can design the algorithm as solving the polynomial inequality system. In general, the polynomial system is NP-hard to solve. So instead, we relax it as a hierarchy of SDP relaxation, which is guaranteed to converge to the correct solution when the degree goes to infinity. Since we require there is a low-degree SoS proof of soundness, we can round the low-degree SoS relaxation to get an almost correct solution.

3.5 Robust Algorithms with SoS

Definition 3.5.1 (Certifiably SubGaussian Distributions)**.** A distribution D on D^d is said to be k-certifiably C-subGaussian, if there is a degree k SoS proof of the following unconstrained polynomial inequality in variables $v_1, v_2, \ldots, v_d$: $\mathbb{E}_D \langle x, v \rangle^k \leqslant (Ck)^{k/2} (\mathbb{E}_D \langle x, v \rangle^2)^{k/2}$. In notation, we write: $\frac{\,k\,}{v} \left\{ \mathbb{E}_D \langle x, v \rangle^k \leqslant (Ck)^{k/2} (\mathbb{E}_D \langle x, v \rangle^2)^{k/2} \right\}$.

Certifiable subgaussianity asserts that the empirical k-th moments of D are bounded and that this bound has a SoS proof of degree k. There are a series of robust algorithms using Certifiable subgaussianity as SoS certificate as in the following results.

Theorem 3.5.2 (Robust Mean and Covariance Estimation, [12])**.** *For every $C > 0$ and even $k \in \mathbb{N}$, there exists a polynomial-time algorithm that given a corrupted sample $S \subseteq \mathbb{R}^d$ outputs a mean-estimate $\hat{\mu} \in \mathbb{R}^d$ and a covariance-estimate $\hat{\Sigma} \in \mathbb{R}^{d \times d}$ with the following guarantee: there exists $n_0 \leqslant (C + d)^{O(k)}$ such that if S is an ε-corrupted sample with size $|S| \geqslant n_0$ of a k-certifiably C-subgaussian distribution D over $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$, then with high probability*

$$\|\mu - \hat{\mu}\| \leqslant O(Ck)^{1/2} \cdot \varepsilon^{1-1/k} \cdot \|\Sigma\|^{1/2}$$
$$\left\|\Sigma^{-1/2}(\mu - \hat{\mu})\right\| \leqslant O(Ck)^{1/2} \cdot \varepsilon^{1-1/k}$$
$$(1 - \delta)\Sigma \preceq \hat{\Sigma} \preceq (1 + \delta)\Sigma \ \textit{for}\ \delta \leqslant O(Ck)^{1/2} \cdot \varepsilon^{1-1/k}$$

For the last two bounds, we assume in addition $Ck \cdot \varepsilon^{1-1/k} \leqslant \Omega(1)$.

Theorem 3.5.3 (Robust Moment Estimation, [12])**.** *For every $C > 0$ and even $k \in \mathbb{N}$, there exists a polynomial-time algorithm that given a corrupted sample $S \subseteq \mathbb{R}^d$ outputs a moment-estimates $\hat{M}_2 \in \mathbb{R}^{d^2}, \ldots, \hat{M}_k \in \mathbb{R}^{d^k}$ with the following guarantee: there exists $n_0 \leqslant (C+d)^{O(k)}$ such that if S is an ε-corrupted sample with size $|S| \geqslant n_0$ of a k-certifiably C-subgaussian distribution D over $\mathbb{R}^d$ with moment-tensors $M_2 \in \mathbb{R}^{d^2}, \ldots, M_k \in \mathbb{R}^{d^k}$ such that $Ck \cdot \varepsilon^{1-1/k} \leqslant \Omega(1)$, then with high probability for every $r \leqslant k/2, u \in \mathbb{R}^d$,*

$$\langle M_r - \hat{M}_r, u^{\otimes r} \rangle^2 \leqslant \delta_r \cdot \langle M_2, u^{\otimes 2} \rangle^r \;\; \text{for } \delta_r \leqslant O(Ck)^{r/2} \cdot \varepsilon^{1-r/k}.$$

Furthermore, there exist degree-k sum-of-squares proofs of the above polynomial inequalities in u.

For harder problems such as clustering non-spherical Gaussian mixtures, the certifiable subgaussianity is inadequate to get the soundness property. Additionally, we need the following two analytic conditions.

Definition 3.5.4 (Certifiable Hypercontractivity)**.** An isotropic distribution 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:

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

Certifiable hypercontractivity strictly generalizes the certifiable subgaussianity property (Definition 3.5.1) that controls higher moments of linear polynomials.

Definition 3.5.5 (Certifiable Anti-concentration)**.** A mean 0 distribution D with covariance Σ is $2s$-certifiably $(\delta, c\delta)$-anti-concentrated if for $q_{\delta,\Sigma}(x, v)$ defined above, there exists a degree $2s$ sum-of-squares proof of the following two unconstrained polynomial inequalities in indeterminate v:

$$\langle x, v\rangle^{2s} + \delta^{2s} q_{\delta,\Sigma}(x, v)^2 \geqslant \delta^{2s}(v^\top \Sigma v)^{2s}$$
$$\mathop{\mathbb{E}}_{D} q_{\delta,\Sigma}(x, v)^2 \leqslant C\delta(v^\top \Sigma v)^{2s}.$$

Using the above two properties, more robust results can be shown.

Theorem 3.5.6 (Robustly Clustering Gaussian Mixtures,[43, 44])**.** *Fix $\eta, \varepsilon > 0$.Let $D_r = N(\mu(r), \Sigma(r))$ for $r \leqslant k$ be k-Gaussians such that $d_{TV}(D_r, D_{r'}) \geqslant 1 - \exp(-\operatorname{poly}(k/\eta))$ whenver $r \neq r'$. Then, there exists an algorithm that takes input an ε-corruption Y of a sample $X = C_1 \cup C_2 \cup \cdots \cup C_k$ of size n, with equal sized clusters C_i drawn i.i.d. from D_i for each $r \leqslant k$, and with probability $\geqslant 0.99$, outputs an approximate clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \cdots \cup \hat{C}_k$ satisfying $min_{i\leqslant k}\frac{|\hat{C}_i \cap C_i|}{|C_i|} \geqslant 1 - O(k^{2k})(\varepsilon + \eta)$. The algorithm has sample complexity $n \geqslant d^{O(\operatorname{poly}(k/\eta))}$ and running time $n^{O(\operatorname{poly}(k/\eta))}$.*

CHAPTER 4

ROBUSTLY LEARNING MIXTURES OF GAUSSIANS

This chapter is joint work with Ainesh Bakshi, Ilias Diakonikolas, Daniel Kane, Pravesh Kothari, and Santosh Vempala, and appears as [74].

4.1 Summary of Results and Techniques

4.1.1 Our Results

In this Chapter, we focus on the total variation contamination model as defined in Definition 1.1.2.

Remark 4.1.1. Our robust learning algorithm succeeds in the strong contamination model in Definition 1.1.3 , with the additional requirement that we can obtain two sets of independent ε-corrupted samples from the unknown mixture.

In the context of robustly learning GMMs, we want to design an efficient algorithm with the following performance: Given a sufficiently large set of samples from a distribution that is ε-close in total variation distance to an unknown GMM $\mathcal{M}$ on $\mathbb{R}^d$, the algorithm outputs a hypothesis GMM $\widehat{\mathcal{M}}$ such that with high probability the total variation distance $d_{\mathsf{TV}}(\widehat{\mathcal{M}}, \mathcal{M})$ is small. Specifically, we want $d_{\mathsf{TV}}(\widehat{\mathcal{M}}, \mathcal{M})$ to be only a function of ε and independent of the underlying dimension d.

The main result of this chapter is the following:

Theorem 4.1.2 (Robustly Learning k-Mixtures with poly(ε)-error, Informal, see Theorem 4.6.1)**.** *There is an algorithm with the following behavior: Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)} \operatorname{poly}_k(1/\varepsilon)$ samples from a distribution F on $\mathbb{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time*

$\mathrm{poly}(n)\,\mathrm{poly}_k(1/\varepsilon)$ *and outputs a k-GMM hypothesis* $\widehat{\mathcal{M}} = \sum_{i=1}^{k} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ *such that with high probability we have that* $d_{\mathsf{TV}}(\widehat{\mathcal{M}}, \mathcal{M}) \leqslant O(\varepsilon^{c_k})$*, where* c_k *depends only on* k.

Theorem 4.1.2 gives a polynomial-time *robust proper learning* algorithm, with dimension-independent error guarantee, for *arbitrary* k-GMMs, for any fixed k.

Discussion Before proceeding, we make a few important remarks about Theorem 4.1.2.

1. *Sample Complexity and Runtime*: Our algorithm succeeds whenever the sample size n satisfies $n \geqslant n_0 = d^{O(k)}/\mathrm{poly}(\varepsilon)$. The running time of our algorithm is $\mathrm{poly}(n)\,\mathrm{poly}_k(1/\varepsilon)$. Statistical query lower bounds [47] suggest that $d^{\Omega(k)}$ samples are necessary for efficiently learning GMMs, even for approximation to constant accuracy in the simpler setting without outliers and under the more restrictive *clustering* setting (where components are pairwise well-separated in total variation distance). This provides some evidence that the sample-time tradeoff achieved by Theorem 4.1.2 is qualitatively optimal (within absolute constant factors in the exponent). We note that the algorithm establishing Theorem 4.1.2 works in the standard bit-complexity model of computation and its running time is polynomial in the bit-complexity of the input parameters. We discuss the numerical accuracy required to implement our algorithm in Appendix 4.9.

   In the noiseless case, the first polynomial-time learning algorithm for k-GMMs on $\mathbb{R}^d$ was given in [3, 4]. In particular, the sample complexity and running time of the [3] algorithm is $(d/\varepsilon)^{q(k)}$, for some function $q(k) = k^{\Omega(k^2)}$. We observe that our algorithm also works for the noiseless case if we set the fraction of noise to be 0 and let $\eta = \varepsilon^{k^{O(k^2)}}$ be the accuracy parameter. The running time and sample complexity in Theorem 4.1.2 are exponentially better than the guarantees in [3, 4] if $\varepsilon = \omega(1/d)$, say ε is a constant. When $\varepsilon = O(1/d)$, our running time and samples complexity matches the results in [3, 4]. Moreover, as we explain in Section 4.1.4, the [3, 4] algorithms are very sensitive to outliers and an entirely new approach is required to obtain an efficient robust learning algorithm.

2. *Handling Arbitrary Weights*: The algorithm of Theorem 4.1.2 succeeds *without any assumptions* on the weights of the mixture components. We emphasize that this is an important feature and not a technicality. Prior work [43, 44, 45], as well as the concurrent work [48], cannot handle the case of general weights — even for the case of $k = 2$ components. In fact, for the special case of uniform weights, we give a simpler algorithm for robustly learning GMMs (presented in Theorem 4.4.12). This algorithm naturally generalizes to give a sample complexity and running time that grows *exponentially* in $1/w_{\min}$, where $w_{\min}$ is the minimum weight of any component in the mixture. Handling the general case (i.e., obtaining a fully polynomial-time algorithm, not incurring an exponential cost in $1/w_{\min}$) requires genuinely new algorithmic ideas and is one of the key technical innovations in the proof of Theorem 4.1.2.

3. *Handling Arbitrary Covariances*: The algorithm of Theorem 4.1.2 does not require assumptions on the variances of components, modulo basic limitations posed by numerical computation issues. Specifically, our algorithm works even if some of the component covariances are rank-deficient (i.e., have directions of 0 variance) with running time scaling polynomially in the bit-complexity of the unknown component means and covariances. Such a dependence on the bit complexity of the input parameters is unavoidable – there exist[1] examples of rank-deficient covariances with irrational entries such that the total variation distance between the corresponding Gaussian and every Gaussian with covariance matrix of rational entries is the maximum possible value of one.

4. *Error Guarantee*: The final error guarantee of our basic algorithm is ε^{c_k}, for some function c_k that depends only on k and goes to 0 when k increases. Importantly,

[1]For e.g., for unit vector $v = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, 0, \ldots, 0)$ and for every choice of rational covariance Σ, the total variation distance between $\mathcal{N}(0, I - vv^\top)$ and $\mathcal{N}(0, \Sigma)$ is the maximum possible 1.

for any fixed k, the final error guarantee of our algorithm is a fixed polynomial in ε and independent of the dimension d. This turns out to be quantitatively close to best possible for any robust proper learning algorithm.

Robust Parameter Recovery We also show that the *same* algorithm as in Theorems 4.1.2 and 4.6.1 actually achieve parameter estimation, i.e., imply that the recovered mixture of Gaussians is close in *parameter* distance to the unknown target mixture. Such parameter estimation results are usually stated under the assumption that every pair of components of the unknown mixture are separated in total variation distance. In this work, we provide a stronger version of this parameter estimation guarantee.

More specifically, in the theorem below, we prove that whenever the components of the input mixture can be clustered together into some groups such that all mixtures in a group are close (and thus, indistinguishable), there exists a similar clustering of the output mixture such that all parameters (weight, mean, and covariances) of each cluster are close within $\mathrm{poly}_k(\varepsilon)$ in total variation distance. In particular this means that for each significant component of the input mixture, there is a component of the output mixture with very close parameters.

Theorem 4.1.3 (Parameter Recovery, See Theorem 4.7.2)**.** *Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)} \mathrm{poly}_k(1/\varepsilon)$ samples from a distribution F on $\mathbb{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $\mathrm{poly}(n) \mathrm{poly}_k(1/\varepsilon)$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^{k'} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability we have that there exists a partition of $[k]$ into $k'+1$ sets $R_0, R_1, \ldots, R_{k'}$ such that*

1. *Let $W_i = \sum_{j \in R_i} w_j$, $i \in \{0, 1, \ldots, k'\}$. Then, for all $i \in [k']$, we have that*

$$|W_i - \widehat{w}_i| \leqslant \mathrm{poly}_k(\varepsilon), \text{ and}$$
$$d_{\mathsf{TV}}(\mathcal{N}(\mu_j, \Sigma_j), \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)) \leqslant \mathrm{poly}_k(\varepsilon) \quad \forall j \in R_i\ .$$

2. *The total weight of exceptional components in R_0 is $W_0 \leqslant \mathrm{poly}_k(\varepsilon)$.*

If we assume additionally that any pair of components in the unknown mixture has total variation distance at least $\mathrm{poly}_k(\varepsilon)$, then the following result follows directly from Theorem 4.1.3.

Corollary 4.1.4. *Let $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be an unknown target k-GMM satisfying the following conditions: (i) $d_{\mathsf{TV}}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j)) \geqslant \varepsilon^{f_1(k)}$ for all $i \neq j$, and (ii) $S = \{i \in [k] : w_i \geqslant \varepsilon^{f_2(k)}\}$ is a subset of $[k]$, where $f_1(k), f_2(k)$ are sufficiently small functions of k. Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)} \mathrm{poly}_k(1/\varepsilon)$ samples from a distribution F on $\mathbb{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, there exists an algorithm that runs in time $\mathrm{poly}(n)\,\mathrm{poly}_k(1/\varepsilon)$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^{k'} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability there exists a bijection $\pi : S \longrightarrow [k']$ satisfying the following: For all $i \in S$, it holds that*

$$
\begin{aligned}
|w_i - \widehat{w}_{\pi(i)}| &\leqslant \mathrm{poly}_k(\varepsilon) \\
d_{\mathsf{TV}}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\widehat{\mu}_{\pi(i)}, \widehat{\Sigma}_{\pi(i)})) &\leqslant \mathrm{poly}_k(\varepsilon).
\end{aligned}
$$

We note that both the pairwise separation between the components and the lower bounds on the weights in Corollary 4.1.4 scale as a fixed polynomial in ε (for fixed k), which is qualitatively information-theoretically necessary.

4.1.2 Related Prior Work

The algorithmic question of designing efficient robust estimators in high dimensions has been extensively studied in recent years. After the initial papers [21, 22], a number of works developed robust estimators for a range of statistical problems. These include efficient outlier-robust algorithms for sparse estimation [75, 76], learning graphical models [77], linear regression [23, 78, 79, 80, 81], stochastic optimization [82, 83], and connections to non-convex optimization [84, 80]. Notably, the robust estimators developed in some of these works [21, 22, 65] are scalable in practice and yield a number of applications in exploratory data analysis [65] and adversarial machine learning [85, 83]. The reader is referred to [86] for a recent survey.

The literature also contains a relevant lower bound: [47] gave a lower bound in the Statistical Query model for learning (fully) clusterable Gaussian mixtures without outliers, which provides evidence that a $d^{\Omega(k)}$ dependence is necessary in both the sample complexity and runtime of any algorithm that learns GMMs. In particular, there is no more efficient method known for the family of instances used in their lower bound.

In terms of techniques, our partial clustering algorithm makes essential use of the Sum-of-squares based *proofs to algorithms* framework (see [66] for an exposition). This framework, beginning with [73], uses the *Sum-of-squares method* to design algorithms for statistical estimation problems, and has led to some of the most general outlier-robust learning algorithms. This includes computationally efficient outlier-robust estimators of the mean, covariance, and low-degree moments of structured distributions, with applications to ICA [12], linear regression [23, 79, 80], clustering spherical mixtures [41, 25], and clustering non-spherical mixtures [43, 44]. The sum-of-squares method also gives a generally applicable scheme to handle the *list-decodable learning* setting [87, 28], where a majority of the input points are corrupted, yielding efficient list-decodable learners for mean estimation [25], regression [29, 30], and subspace clustering/recovery [88, 89].

Our work also has connections to the usage of tensor decomposition algorithms for learning statistical models. [5] used fourth-order tensor decomposition to obtain a polynomial-time algorithm for mixtures of spherical Gaussians with linearly independent means (with condition number guarantees). This result was extended via higher-order tensor decomposition for non-spherical Gaussian mixtures in a smoothed analysis model [90]. Fourth order tensor decomposition has earlier been used in [57] for the ICA problem [57], and extended to general ICA with higher-order tensor decomposition by [62]. Such results rely on additional and non-trivial assumptions on the parameters of the mixture components in order to succeed, and are incomparable to our tensor-decomposition result that does not make any assumptions on the parameters of the mixture components. Indeed, this is the key innovation in our tensor decomposition algorithm that relaxes the guarantees on the output (a small list of candidate parameters) under a priori bounds on distance between components that is ensured by our partial clustering subroutine. This relaxation of tensor decomposition, and the new procedure that accomplishes it, is a principal contributions of our work.

Our algorithm builds on the ideas in the works [43, 44] that gave efficient clustering algorithms for any fixed number k of components, under the crucial assumption that the components have pairwise total variation distance close to 1. In this case, the above works actually succeed in efficiently *clustering* the input sample into k groups, such that each group contains the samples generated from one of the Gaussians, up to some small misclassification error. In contrast, the main challenge in this work is the information-theoretic impossibility of clustering in our setting where there are no separation assumptions. As we will explain in the proceeding discussion, while we draw ideas from [43, 44, 45], a number of significant conceptual and technical challenges need to be overcome in the non-clusterable setting.

Our work is most closely related to the paper by Kane [45], which gives a polynomial-time robust learning algorithm for the *uniform* $k = 2$ case, i.e., the case of two *equal weight* components, and to the polynomial time algorithms [43, 44] under the assumption that the component Gaussians are pairwise well-separated in total variation distance.

In independent and concurrent work with an earlier version of this chapter [91], [48] gave an efficient outlier-robust parameter learning algorithm for mixtures of Gaussians under additional assumptions. The algorithm and guarantee in [48] require three crucial assumptions: 1) all mixing weights are bounded below by $1/f(k)$, where f is a function of k, and appears in the exponent of the running time and output error, 2) all components have covariances with eigenvalues at least poly(ε/d) and at most poly(d/ε) (in particular, the components must be full rank), and 3) every pair of component Gaussians are separated by $g(k)$ in total variation distance (regardless of how small the fraction of corruptions ε is).

(We note that removing each of these assumptions requires significant new ideas in both the design and the analysis of the algorithm.)

The techniques adopted in the two concurrent works ([91] and [48]) are significantly different. On one hand, both works make essential use of advances in clustering non-spherical mixtures [43, 44]. While [48] relies on the clustering algorithm of [44], our work leverages a *key modification* to the clustering algorithm of [43] that allows us to obtain a fixed polynomial-time algorithm. Indeed, our novel variant of clustering combined with our new spectral clustering subroutine is the key to not incurring an exponential cost in $1/w_{\min}$. After clustering, where the input mixture can be assumed to be non-clusterable, [48] uses a sum-of-squares based algorithm for parameter estimation. In contrast, our work relies on a new list-decoding algorithm for tensor decomposition that makes no assumption on the underlying components.

4.1.3 Organization

The structure of this chapter is as follows: In Section 4.2, we provide relevant background and technical facts. In Section 4.3, we describe and analyze our new tensor decomposition algorithm. In Section 4.4, we use a sum-of-squares based approach to partially cluster a mixture and we present a refinement of our partial clustering procedure that improves the probability of success to a constant independent of the minimum weight of any component in the input mixture. In Section 4.5, we give a spectral separation algorithm to identify "thin" components. In Section 4.6, we put all these pieces together to prove Theorem 4.1.2. Finally, in Section 4.7, we show that our algorithm in fact achieves parameter estimation guarantees and prove Theorem 4.1.3.

4.1.4 Technical Overview

In this section, we give a bird's eye view of our algorithm and the main ideas that go into it. Recall that our goal is to design an efficient algorithm that takes an ε-corrupted sample Y from a mixture of k-Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ and outputs a mixture $\widehat{\mathcal{M}} = \sum_i \hat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ such that the total variation distance between $\mathcal{M}$ and $\widehat{\mathcal{M}}$ is bounded above by a dimension-independent polynomial in ε for any fixed k. Specifically, we want the running time of our algorithm to be bounded above by a polynomial in the dimension d and $1/\varepsilon$, for any fixed k.

In the non-robust setting (i.e., for $\varepsilon = 0$), the algorithm of Moitra and Valiant [3] (see also [4]), extending their work with Kalai [2], solves this problem. However, natural attempts to adapt their method to handle outliers run into immediate difficulties. The starting point of [3] is to observe that if a mixture of k Gaussians has every pair of components separated in total variation distance by at least δ, then a random univariate projection of the mixture has a pair of components that are $\delta/\sqrt{d}$-separated in total variation distance. Their algorithm uses this observation to piece together estimates of the mixture when projected to several carefully chosen directions to get an estimate of the high-dimensional mixture. This strategy meets with a major roadblock in the presence of outliers: the fraction of outliers, being a dimension-independent constant, completely overwhelms the total variation distance between components in any one direction making them indistinguishable[2].

To a reader familiar with algorithmic robust statistics, this may not come as a surprise — to handle outliers, one typically needs to develop a completely new algorithm, that is new even in the outlier-free setting. Our approach diverges from the non-robust method and instead relies on a careful interleaving of two new algorithmic primitives: (1) a new *partial clustering algorithm* based on the sum-of-squares (SoS) method, and (2) a new *tensor decomposition* method for decomposing a symmetrized sum of tensor powers of $d \times d$ matrices. Briefly, there are four high-level ingredients: partial clustering, approximate isotropy, tensor decomposition and subspace enumeration. We discuss each of these, first in the context of a simpler algorithm with an exponential dependence on the minimum mixing

[2]Informally speaking, one could hope to show that outliers projected into a random direction cannot be too adversarial, but it is unclear how to use this observation not in the least because the algorithm of [3] requires a somewhat carefully tailored choice of projections.

weight, and then in a truly polynomial algorithm.

Main ingredients

Partial Clustering. We call a mixture partially clusterable (or well-separated) if there is a pair of components that have total variation distance larger than $1 - \Omega_k(1)$. We note that even the setting when the mixture is *fully* clusterable (i.e., every pair of components is well-separated), the learning problem captures many hard special cases (such as subspace clustering) and is highly non-trivial. Two recent works [44, 43] gave a polynomial-time algorithm for the fully clusterable case, using the sum-of-squares method. It turns out that the clustering algorithm of [43] (specifically, their Lemma 6.4) can be generalized (see Theorem 4.4.2) to the partial clustering setting, i.e., the setting where we are guaranteed to have a pair of components that are well-separated rather than every pair. Suppose α is the minimum mixing weight. This gives an algorithm with running time of $d^{(k/\alpha)^{O(k)}}$ to partition the input sample into components so that each piece of the partition is (effectively) an $(\varepsilon + \mathrm{poly}(\alpha/k))$-corrupted sample from disjoint sub-mixtures.

Approximate Isotropic Transformation. After applying our partial-clustering algorithm, we would like to make the mixture isotropic — that is, we would like to assume that the mean of the mixture is ≈ 0 and the covariance of the mixture is $\approx I$. In the setting with no outliers, this is done by computing the empirical mean and covariance and applying an appropriate affine transformation to the input points. However, in the setting with outliers, even this task is non-trivial; in particular because we need dimension-independent error guarantee on the estimated covariance in *Frobenius norm* (instead of the weaker spectral norm). Fortunately, [43] (Theorem 7.1 and Fact 2.2.3) gives precisely such an algorithm for robust covariance estimation that relies on the property of *certifiable hyper-contractivity* (we verify that this property in our setting in Lemma 4.2.19 of Section 4.2).

After the first two steps, we can effectively assume that we are working with an ε-

corrupted sample from a mixture that is approximately isotropic and every pair of components is not too far in total variation distance. Why is this guarantee useful? As established in the recent works [44, 43], such a bound translates into a guarantee (with respect to natural norms) on the parameters of the component Gaussians. In particular, in our setting, this implies that after partial clustering plus an approximate isotropic transformation: 1) $\|\Sigma_i - I\|_F \leqslant \text{poly}(\alpha, k)$, 2) $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and 3) $\Sigma_i \succeq \frac{1}{\text{poly}(\alpha,k)} I$ for every i (recall that α is the minimum weight in the mixture). In this case, it turns out that in order to learn the unknown mixture with error guarantees in total variation distance, it suffices to obtain $\text{poly}_k(\varepsilon)$-error estimates of the μ_i, Σ_i's in Frobenius norm.

List-decodable Symmetrized Tensor Decomposition. The key first step in addressing such a mixture was taken in the work of Kane [45], who gave a polynomial-time algorithm to robustly learn an *equiweighted* mixture of two Gaussians. For this special case, after isotropic transformation, one can effectively assume that the two means are $\pm\mu$ and the two covariances are $I \pm \Sigma$. Kane's idea is to look at a certain tensor ("Hermite tensor") that can be built using the 4-th and 6-th raw moments of the mixture. Since we must use outlier-robust algorithms to estimate these tensors, we can obtain estimates that are accurate only up to constant error in Frobenius norm of the tensor. Kane's key observation is that *for the special case of $k = 2$ components*, one can build two different Hermite tensors, one of which is rank-one with component $\approx \mu$ (and thus one can immediately "read off" μ); the other only has a tensor power of Σ. This second tensor is of the form $\hat{T}_4 = \text{Sym}((\Sigma - I) \otimes (\Sigma - I)) + E$, where $\|E\|_F = O_k(\sqrt{\varepsilon})$ and Sym refers to symmetrizing over all possible permutations of the "4 modes of the tensor". Unlike the case of the mean, one cannot simply "read-off"[3] Σ from T_4, but Kane gives a simple method to accomplish this. As noted in [45], it is not clear how to extend this to non-equiweighted mixtures of $k = 2$ Gaussians, and going to even $k = 3$ components requires substantially new ideas. This is because of several issues:

[3]It is helpful to visualize a single entry of this tensor for, say, the case when i, j, k, ℓ are all distinct: $\hat{T}_4(i, j, k, \ell) = \frac{1}{3}(\Sigma(i, j)\Sigma(k, \ell) + \Sigma(i, k)\Sigma(j, \ell) + \Sigma(i, \ell)\Sigma(j, k)) + \text{error}$. Notice that obtaining entries of Σ from T_4 is formally a task of solving noisy quadratic equations.

1) as stated, Kane's approach does not work as is even for $k = 2$ when the mixture is not equiweighted, 2) it is not known whether one can build tensors that "separate" out the means and covariances, as [45] managed to do for $k = 2$, 3) the relevant tensors will not be (symmetrizations of) rank-1 tensors, up to noise. It is worth noting, in fact, that such a gap is information-theoretically inherent, as shown in [3]; learning arbitrary mixtures of k Gaussians *provably* requires at least $2k$ moments of the mixture. Another way of seeing this is by considering the parallel pancakes construction of [47], where the authors produce an example of a mixture of k Gaussians whose first k moments match the standard Gaussian exactly despite not being close in total variational distance. It should be noted that, in this example, the component Gaussians are all equal to the standard Gaussian except in one hidden direction. Thus, we cannot hope to identify the components exactly with just $O(1)$ many moments. The somewhat surprising fact (that we establish in more detail below) is that after partial clustering, by looking at only the first four moments of our mixture, we can learn all of the components up to some errors taking place along a bounded number of hidden directions. In particular, we can learn the covariance matrices of the component Gaussians up to some *low rank* error terms. We elaborate on this new idea below.

For the sake of the intuition, it is helpful to focus on the simpler case where all the means are zero. In this case, the estimated 4th Hermite tensor (built from estimated raw moments of degree at most 4 of the mixture) has the following form:

$$\hat{T}_4 = \sum_{i=1}^{k} w_i \mathrm{Sym}((\Sigma_i - I) \otimes (\Sigma_i - I) + E) \ ,$$

where E is a 4-tensor with $\|E\|_F = O_k(\sqrt{\varepsilon})$. Given the form of this tensor, it is natural to think of applying tensor decomposition algorithms, by thinking of $\Sigma_i - I$ as a d^2-dimensional vector. However, we run into the issue of uniqueness of tensor decomposition, since we are dealing with 2nd order tensors (once we view $\Sigma_i - I$ as a d^2-dimensional vector). One might imagine computing higher-order tensors of similar forms to overcome the uniqueness issues, but this runs into two major complications: first, the symmetrization operation introduces spurious terms that do not have the sum of tensor-power structure required for such an algorithm to succeed. Indeed, this is far from just being an annoying technicality — the recent work of Garg, Kayal, and Saha [92] addresses precisely such a tensor decomposition problem via algebraic techniques. Unfortunately, as is explicitly pointed out as one of the main open question in their work (see page 14), because of their reliance on algebraic techniques, their algorithm is highly brittle and in particular, may not even be able to handle the benign noise that comes from estimating tensors from independent (uncorrupted) samples. (Of course, our setting has to deal with the malicious noise introduced due to the adversarial outliers.)

Second, even if one were to get hold of the tensor without the symmetrization operation, the only applicable tensor decomposition algorithm (recall that we do not make *any* genericity assumptions on the components that are typically required by tensor decomposition algorithms) is the result of Barak, Kelner, and Steurer [73]. However, while being efficient in its dependence on the number of components, this algorithm has exponential dependence on the target error, which is prohibitive for our purpose.

Our idea is to give up on the goal of recovering the unique decomposition of the tensor $\hat{T}_4$ above, and start by applying an operation that is a common trick in most tensor decomposition algorithms. In our context, this trick amounts to taking a random matrix (say, with independent standard Gaussian entries) P and "collapsing" the last two modes of $\hat{T}_4$ with P (i.e., computing $\hat{S}(i,j) = \sum_{k,\ell} \hat{T}_4(i,j,k,\ell)P(k,\ell)$) to obtain a matrix Q. In the usual tensor decomposition procedures, we are interested in proving that one can recover all the information about the components of the tensor from Q. We will not be able to prove such a statement here. Instead, our key observation is that one can randomly choose a matrix P *of rank* $\mathrm{poly}(k)$ and argue that the resulting $\hat{S}$ is $O_k(\varepsilon)$-close to one of the $\Sigma_i - I$ *up to an error* *term of* $O(k^2)$ *rank* with $\mathrm{poly}_k(\varepsilon)$ probability. To see this, note that in the symmetrization

$$\hat{T}_4(i,j,k,\ell) = \frac{1}{3}(\Sigma(i,j) \otimes \Sigma(k,\ell) + \Sigma(i,k) \otimes \Sigma(j,\ell) + \Sigma(i,\ell) \otimes \Sigma(j,k)) + \text{ error}$$

applying a rank-one matrix P to the modes k, ℓ will reduce the first term to the matrix Σ, while the latter two terms become rank-one terms! When the tensor is a sum of k such symmetrized tensors, we will use k such rank-one matrices, and take a linear combination of them to get a weighted sum of the Σ's plus a term of rank $O(k^2)$. As we show, the linear combination can be chosen such that only one of the component Σ's survives (up to small Frobenius norm). Moreover, such a low-rank P can be obtained efficiently by simply choosing $\mathrm{poly}(k)$ random rank-1 matrices and exhaustively searching over an appropriate cover of the $O(k^2)$-dimensional subspace spanned by them.

Subspace Enumeration for Low-rank Terms. We then show that we can use the generated estimates $\hat{S}$ and the tensors $\hat{T}_m$ for $m \leqslant 4k$ to find a $\mathrm{poly}_k(\varepsilon)$-dimensional subspace V_* such that the low-rank error matrix in the estimated $\hat{S}$ have their range space essentially contained inside V_*. Our next step involves a subspace enumeration over V_* to output a list of a bounded number of parameters such that a mixture defined by some k of them must be close in total variation distance to the input mixture.

Our final step involves a relatively standard hypothesis testing procedure, using a robust tournament that goes over each of the candidate mixtures in our generated list and finds one that is approximately the closest in total variation distance to a (fresh) set of corrupted samples.

An Algorithm with Exponential Dependence in $1/\varepsilon$

The above ingredients suffice to derive an algorithm whose running time grows exponentially in the reciprocal of the minimum weight of the mixture. Almost every step is exponential in the minimum mixing weight: the partial clustering algorithm, the tensor decomposition algorithm, and the enumeration of the candidate subspaces that contain the low-rank terms. Nevertheless, this gives a polynomial-time algorithm (see Theorem 4.4.12 for details) for robustly learning arbitrary equiweighted mixtures of k Gaussians. When the weights are not all equal, notice that we can treat any component with weight $\leqslant \varepsilon$ as outliers, which effectively means that $\alpha \geqslant \varepsilon$. Thus, this already yields a $d^{f(k/\varepsilon)}$-time algorithm in the general setting.

Polynomial-time Algorithm

In order to improve this running time to have a fixed polynomial dependence on d (independent of $1/\varepsilon$), we first develop a new partial clustering method that is more efficient, but weakens the separation guarantee of total variation distance. Our final algorithm then involves a recursive interleaving of the partial clustering and tensor decomposition steps together with a *Recursive Spectral Clustering* subroutine. We discuss these steps next.

Faster Partial Clustering. The first bottleneck in the running time guarantee of the algorithm described above is the partial clustering step, so it is important to examine the cause for the exponential dependence on the minimum weight in the running time. The algorithm relies on a characterization of a well-separated pair of Gaussians $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ in terms of three geometric distances between their parameters: 1) Mean Separation: $\exists v$ such that $\langle \mu_1 - \mu_2, v\rangle^2 \geqslant \Delta(v^\top(\Sigma_1 + \Sigma_2)v)$, 2) Relative Frobenius Separation: $\left\|\Sigma_1^{-1/2}(\Sigma_2 - \Sigma_1)\Sigma^{-1/2}\right\|_F \geqslant \Delta$, and 3) Spectral Separation: $\exists v$ such that $v^\top\Sigma_1 v \geqslant \Delta v^\top\Sigma_2 v$ or $v^\top\Sigma_2 v \geqslant \Delta v^\top\Sigma_1 v$. In other words, two Gaussians are well-separated in TV distance iff they are separated in one of the above geometric distances. The main idea of the partial clustering algorithm is to give efficient (low-degree) *sum-of-squares* certificates of *simultaneous intersection bounds* that show that any cluster a natural SoS relaxation finds cannot significantly overlap with two well-separated clusters simultaneously. This step requires sum-of-squares certificates for two natural analytic properties: *certifiable hypercontractivity* and *certifiable anti-concentration* (introduced in the recent works [29, 30], and also used in [88, 89]). The bottleneck in the running time for us is the degree of the sum-of-squares certificate needed for certifiable anti-concentration (which grows polynomially in $1/\alpha$). It is not known whether there is a sum-of-squares certificate of much smaller degree for certifiable anti-concentration. To make progress here, we observe that the only usage of this certificate occurs in dealing with spectrally separated pairs of Gaussians in the mixture. So we give a new partial clustering algorithm that works in fixed polynomial time, whenever there is a pair of Gaussian components separated either via their means or an appropriate variant of the relative Frobenius distance, but it might not detect spectral separation. This places an additional burden on subsequent steps.

Spectral Separation of Thin Components. The faster partial clustering algorithm does not guarantee a lower bound on the minimum eigenvalues of component covariances. It turns out that the above is the only way the algorithm can fail — one or more covariance matrices might have a very small eigenvalue (if not, the Frobenius norm error would imply TV-distance error). Since we have estimates of the covariances, we can find an eigenvector along such a small eigenvalue direction. We then observe that since the mixture is nearly isotropic (i.e., the overall variance in each direction is ~ 1), if some component has very small variance along a direction, then the components must be separable along this direction. We show that it is possible to efficiently cluster the mixture after projecting it to this direction, so that each cluster has strictly fewer components. We then recursively apply the entire algorithm on the clusters obtained, which will each have strictly fewer components.

Polynomial Complexity in the minimum weight. Our tensor decomposition also has a dependence on the minimum mixing weight. To avoid this exponential dependence on the minimum weight, a natural approach would be to ignore components lighter than some threshold (that depends on the target error) and treat them as corruptions. However, this intuitive approach runs into a difficulty. In order to get nontrivial error guarantees on the tensor decomposition, we need that the minimum mixing weight is significantly *larger* than the fraction of outliers (since the decomposition involves generating a list of candidate hypotheses, one of which must be accurate). To solve this problem, we show that we can set a minimum weight threshold that depends on the number of remaining components of the mixture, and "remove" some (but not all) components, so that the remaining mixture has minimum weights above this threshold; and the threshold is also sufficiently larger than the total weight of small components treated as corruptions. This makes the overall computational complexity dependence on d truly polynomial for any fixed k, and avoids any dependence on the true minimum mixing weight.

Robust Proper Learning with $\mathrm{poly}_k(\varepsilon)$ **Error.** To achieve a polynomial error and polynomial time in ε at the same time, we need to address the two remaining bottlenecks: (1) the exhaustive subspace enumeration within the space V_* of significant eigenvalues, and (2) the success probability in the partial clustering algorithm is exponentially small in the mixing weight and hence necessitates super-polynomial enumeration. We use additional techniques to reduce both dependencies (and hence the overall running time) to $\mathrm{poly}_k(1/\varepsilon)$.

Subspace enumeration over V_* outputs a list of size with exponential dependence on the dimension of V_*. To reduce the list size of tensor decomposition, first we observe that we can use an elementary filtering technique to denoise the data, since the number of samples needed is small. Specifically, if the sample complexity of the algorithm is polynomial in the dimension and error parameter, we can set both the dimension of V_* and the error parameter to be polynomials in $1/\varepsilon$ (for fixed k), so that the number of samples needed is $O(1/\varepsilon)$. Then, with probability $(1-\varepsilon)^{O(1/\varepsilon)} = \Omega(1)$, a sample of size $O(1/\varepsilon)$ drawn from the total variation contamination model has no noise. For such a clean sample, we can apply a non-robust algorithm to the subspace V_*. If the time complexity of the algorithm is polynomial in the dimension and error parameter, then the running time to apply it on V_* will be $\mathrm{poly}_k(\varepsilon)$, as desired. The next step is to prove that such an algorithm exists. We will use the algorithm in Theorem 8 of [3]. A small technical issue is that the latter assumes that any pair of components in the mixture has TV distance at least δ, where δ is the error parameter. We show that with an appropriately chosen parameter $\delta' = \text{poly}(\delta)$, any pair of components with TV distance less than δ' are close enough so that the algorithm cannot distinguish the pair from a single Gaussian, since the algorithm only requires a polynomial number of samples. If we merge all such pairs, any pair in the mixture is separated by δ', and then we can apply Theorem 8 in [3]. For each estimate $\hat{S}$ of the main algorithm, we can recover the low-rank error of $\hat{S}$ by learning the mixture in the subspace V_*. Combining the estimates $\hat{S}$ and the estimates in the subspace V_*, we get a list of parameters of size $\text{poly}_k(\varepsilon)$.

To improve the success probability in the partial clustering, we upgrade the clustering algorithm further in Theorem 4.4.13. The key bottleneck in the earlier partial clustering step is the rounding algorithm that obtains a list of candidate clusters each of size roughly αn. The rounding guarantees that any such candidate cluster cannot simultaneously contain a significant fraction of points from two components whose covariances are separated in Frobenius distance. However, the candidate cluster could be indiscriminate in collecting an arbitrary subset of points from "nearby" clusters. The rounding algorithm simply guesses a "good" partition of them, and this has an $\exp(-1/\alpha)$ success probability.

We improve on this random guessing step in the rounding by observing that points that come from nearby clusters must have covariances that are close in Frobenius norm. Thus, our rounding applies a robust covariance estimation algorithm (with error guarantees in Frobenius norm) to each candidate cluster of size roughly αn and then collects candidate clusters whenever their estimated covariances are close. We show that this procedure coalesces the $O(1/\alpha)$ clusters into a collection of at most k. The random guessing step now succeeds with $\exp(-k)$ probability resolving this bottleneck.

Parameter Recovery.

We show that for any two Gaussian mixtures, if they are close in TV distance, their components are also close in TV distance, which implies that we can recover the components or the parameters of the mixture. This result generalizes Theorem 8.1 in [48], which has three additional assumptions: (i) each component has variance at least $\mathrm{poly}(\varepsilon/d)$ and at most $\mathrm{poly}(d/\varepsilon)$, (ii) each pair of components has TV distance at least $\mathrm{poly}_k(\varepsilon)$, and (iii) the minimal weights of both mixtures are at least $\mathrm{poly}_k(\varepsilon)$. [48] also proved the conclusion under the assumption that any parameters (means and covariances) are identical or separated. We reduce the general case to this simplification. The first step is to deal with the components with small weights. We use a threshold $\varepsilon' = \mathrm{poly}_k(\varepsilon)$ of weights such that if we treat components with weights smaller than the threshold as noise, other components have weights at least $\mathrm{poly}_k(\varepsilon')$. The second step is a partial clustering on the union of the components of the two mixtures, after which the components within each cluster are pairwise not too close, i.e., have TV distance bounded by $1 - \mathrm{poly}(\varepsilon')$. Then we can modify the parameters in each cluster slightly, so that the resulting parameters for different components are either identical or have a minimum separation. Thus, we reduce the general case of two arbitrary mixtures to this special case. Overall, this is a purely information-theoretic statement — for each significant weight component of one mixture, there will be a component in the second mixture with very close mean and covariance. We note that this is not necessarily a 1-1 mapping between components, which is impossible in general without further assumptions.

4.2 Preliminaries

Basic Notation. For a vector v, we use $\|v\|_2$ to denote its Euclidean norm. For an $n \times m$ matrix M, we use $\|M\|_{\mathrm{op}} = \max_{\|x\|_2=1} \|Mx\|_2$ to denote the operator norm of M and $\|M\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ to denote the Frobenius norm of M. We sometimes use the notation $M(i, j)$ to index the corresponding entries in M. For an $n \times n$ symmetric matrix M, we use

$\succeq$ to denote the PSD/Loewner ordering over eigenvalues of M and $\operatorname{tr}(M) = \sum_{i\in[n]} M_{i,i}$ to denote the trace of M. We use $U\Lambda U^\top$ to denote the eigenvalue decomposition, where U is an $n \times n$ matrix with orthonormal columns and Λ is the $n \times n$ diagonal matrix of the eigenvalues. We use $M^\dagger = U\Lambda^\dagger U^\top$ to denote the Moore-Penrose pseudoinverse, where $\Lambda^\dagger$ inverts the non-zero eigenvalues of M. If $M \succeq 0$, we use $M^{\dagger/2} = U\Lambda^{\dagger/2}U^\top$ to denote taking the square-root of the inverted non-zero eigenvalues.

For $d \times d$ matrices A, B, the Kronecker product of A, B, denoted by $A \otimes B$, is indexed by $(i, j), (k, \ell) \in [d] \times [d]$ and has entries $(A \otimes B)((i, j), (k, \ell)) = A(i, k)B(j, \ell)$. We will equip every tensor T with the norm $\|\cdot\|_F$ that simply corresponds to the ℓ_2-norm of any flattening of T to a vector. The notation $T(\cdot, \cdot, x, y)$ is used to denote collapsing two modes of the tensor by plugging in x, y. For a positive integer ℓ and vector v, we also use $v^{\otimes \ell} = \underbrace{v \otimes v \ldots \otimes v}_{\ell \text{times}}$.

We use the notation $\mathcal{M} = \sum_{i\in[k]} w_i \mathcal{N}(\mu_i, \Sigma_i)$ to represent a k-mixture of Gaussians. The total variation distance between two probability distributions on $\mathbb{R}^d$ with densities p, q is defined as $d_{\mathsf{TV}}(p, q) = \frac{1}{2}\int_{\mathbb{R}^d} |p(x) - q(x)|dx$. We also use $\mathbb{E}[\cdot]$, $\mathbf{Var}[\cdot]$ and $\mathsf{Cov}(\cdot)$ to denote the expectation, variance and covariance of a random variable.

For a finite dataset X, we will use $Z \in_u X$ to denote that Z is the uniform distribution on X. We will sometimes use the term mean (resp. covariance) of X to refer to $\mathbb{E}_{Z\in_u X}[Z]$ (resp. $\mathsf{Cov}_{Z\in_u X}(Z)$).

4.2.1 Gaussian Background

The first few facts in this subsection can be found in Kane [45].

Fact 4.2.1. *The total variation distance between two Gaussians $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ can be bounded above as follows:*

$$d_{\mathsf{TV}}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = O\Big((\mu_1 - \mu_2)^\top \Sigma_1^\dagger (\mu_1 - \mu_2) + \left\|\Sigma_1^{\dagger/2}(\Sigma_2 - \Sigma_1)\Sigma_1^{\dagger/2}\right\|_F\Big) .$$

Fact 4.2.2 (Theorem 2.4 in [45])**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^{d\times d}$, where $\mathcal{D}$ is supported on the subset of $\mathbb{R}^{d\times d}$ corresponding to the set of symmetric PSD matrices. Suppose that $\mathbb{E}[\mathcal{D}] = \Sigma$ and that for any symmetric matrix A we have that* $\mathbf{Var}[\mathrm{tr}(AX)] = O\left(\sigma^2\|\Sigma^{1/2}A\Sigma^{1/2}\|_F^2\right)$*. Then, for $\varepsilon \ll \sigma^{-2}$, there exists a polynomial-time algorithm that given sample access to an ε-corrupted set of samples from $\mathcal{D}$ returns a matrix $\hat{\Sigma}$ such that with high probability $\|\Sigma^{-1/2}(\Sigma - \hat{\Sigma})\Sigma^{-1/2}\|_F = O(\sigma\sqrt{\varepsilon})$.*

Fact 4.2.3 (Proposition 2.5 in [45])**.** *Let $G \sim \mathcal{N}(\mu, \Sigma)$ be a Gaussian in $\mathbb{R}^d$. Then, we have that*

$$\mathbb{E}[G^{\otimes m}](i_1, \ldots, i_m) = \sum_{\substack{\text{Partitions } P \text{ of } [m] \\ \text{into sets of size 1 and 2}}} \bigotimes_{\{a,b\}\in P} \Sigma(i_a, i_b) \bigotimes_{\{c\}\in P} \mu(i_c) \ .$$

We will work with the coefficient tensors of d-dimensional Hermite polynomials:

Definition 4.2.4 (Hermite Tensors)**.** Define the degree-m Hermite polynomial tensor as

$$h_m(x) := \sum_{\substack{\text{Partitions } P \text{ of } [m] \\ \text{into sets of size 1 and 2}}} \bigotimes_{\{a,b\}\in P} -I(i_a, i_b) \bigotimes_{\{c\}\in P} x(i_c) \ .$$

We will use the following fact that relates Hermite moments to the raw moments of any distribution.

Fact 4.2.5 (Hermite vs Raw Moments, see, e.g., [93])**.** *For any real-valued random variable u, and $m \in \mathbb{N}$,* $\max_{i\leqslant m} |\mathbb{E}\, u^i - \mathbb{E}_{z\sim\mathcal{N}(0,1)} z^i| \leqslant 2^{O(m)} \max_{i\leqslant m} |\mathbb{E}\, h_m(u)|$*. Similarly,* $\max_{i\leqslant m} |\mathbb{E}\, h_m(u)| \leqslant 2^{O(m)} \max_{i\leqslant m} |\mathbb{E}\, u^i - \mathbb{E}_{z\sim\mathcal{N}(0,1)} z^i|$.

Fact 4.2.6 (Lemma 2.7 in [45])**.** *If $G \sim \mathcal{N}(\mu, I + \Sigma)$, then we have that*

$$\mathbb{E}[h_m(G)] = \sum_{\substack{\text{Partitions } P \text{ of } [m] \\ \text{into sets of size 1 and 2}}} \bigotimes_{\{a,b\}\in P} \Sigma(i_a, i_b) \bigotimes_{\{c\}\in P} \mu(i_c) \ .$$

Fact 4.2.7 (Lemma 2.8 in [45]). *If $G \sim \mathcal{N}(\mu, I+\Sigma)$, then $\mathbb{E}[h_m(G) \otimes h_m(G)]$ is equal to*

$$\sum_{\substack{\text{Partitions } P \text{ of } [2m] \\ \text{into sets of size 1 and 2}}} \bigotimes_{\substack{\{a,b\}\in P \\ a,b \text{ in same half of } [2m]}} \Sigma\left(i_a, i_b\right) \bigotimes_{\substack{\{a,b\}\in P \\ a,b \text{ in different halves of } [2m]}} (I+\Sigma)\left(i_a, i_b\right) \bigotimes_{\{c\}\in P} \mu\left(i_c\right) \ .$$

Lemma 4.2.8 (Slight Strengthening of Lemma 5.2 in [45]). *For $G \sim \mathcal{N}(\mu, \Sigma)$, the covariance matrix of $h_m(G)$ satisfies:*

$$\|\mathcal{Cov}(h_m(G))\|_{op} \leqslant \|\mathbb{E}\left[h_m(G) \otimes h_m(G)\right]\|_{op} = \mathcal{O}\left(m(1 + \|\Sigma\|_F + \|\mu\|_2)\right)^{2m} \ .$$

This follows from the proof of Lemma 5.2 in [45] by noting that the number of terms in the sum is at most 2^m times the number of partitions of $[2m]$ into sets of size 1 and 2, which is at most $O(m)^{2m}$.

Next, we use upper and lower bounds on low-degree polynomials of Gaussian random variables. We defer the proof of the subsequent Lemma to Appendix 4.8.

Lemma 4.2.9 (Concentration of low-degree polynomials). *Let T be a d-dimensional, degree-4 tensor such that $\|T\|_F \leqslant \Delta$ for some $\Delta > 0$ and let $x, y \sim \mathcal{N}(0, I)$. Then, with probability at least $1 - 1/\text{poly}(d)$, the following holds:*

$$\|T\left(\cdot, \cdot, x, y\right)\|_F^2 \leqslant \mathcal{O}\left(\log(d)\Delta^2\right) \ .$$

Note that for any matrix M, $\langle M, x \otimes y\rangle$, where $x, y \sim \mathcal{N}(0, I)$, is a degree-2 polynomial in Gaussian random variables. As a result, we have the following anti-concentration inequality.

Lemma 4.2.10 (Anti-concentration of bi-linear forms, [94]). *Let M be a $d \times d$ matrix and let $x, y \sim \mathcal{N}(0, I)$. Then, for any $\zeta \in (0, 1)$, the following holds:*

$$\mathbb{P}\left[\langle M, x \otimes y\rangle^2 \leqslant \zeta\, \mathbb{E}\left[\langle M, x \otimes y\rangle^2\right]\right] \leqslant \mathcal{O}\left(\sqrt{\zeta}\right) \ .$$

4.2.2 Analytic Properties of Gaussian Distributions

The following definitions and results describe the analytic properties of Gaussian distributoins that we will use. We also state the guarantees of known robust estimation algorithms for estimating the mean, covariance and moment tensors of Gaussian mixtures here.

Certifiable Subgaussianity. We will make essential use of the following definition.

Definition 4.2.11 (Certifiable Subgaussianity (Definition 5.1 in [12]))**.** For $t \in \mathbb{N}$ and an absolute constant $C > 0$, a distribution $\mathcal{D}$ on $\mathbb{R}^d$ is said to be t-certifiably C-subgaussian if for every even $t' \leqslant t$, we have that

$$\Big|\frac{v}{t'} \left\{ \mathop{\mathbb{E}}_{\mathcal{D}} \langle x, v\rangle^{t'} \leqslant (Ct')^{t'/2} \left(\mathop{\mathbb{E}}_{\mathcal{D}} \langle x, v\rangle^2 \right)^{t'/2} \right\} .$$

Fact 4.2.12 (Mixtures of Certifiably Subgaussian Distributions, Analogous to Lemma 5.4 in [12])**.** *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_q$ be t-certifiably C-subgaussian distributions on $\mathbb{R}^d$. Let $p_1, p_2, \ldots, p_q$ be non-negative weights such that $\sum_i p_i = 1$ and $p = \min_{i \leqslant q} p_i$. Then, the mixture $\sum_i p_i \mathcal{D}_i$ is t-certifiably C/p-subgaussian.*

Certifiable Anti-Concentration. The first is *certifiable anti-concentration* — an SoS formulation of classical anti-concentration inequalities — that was introduced in [29, 30].

In order to formulate certifiable anti-concentration, we start with a univariate even polynomial p that serves as a uniform approximation to the delta function at 0 in an interval around 0. Such polynomials are constructed in [29, 30] (see also [95]). Let $q_{\delta,\Sigma}(x, v)$ be a multivariate (in v) polynomial defined by $q_{\delta,\Sigma}(x, v) = (v^\top \Sigma v)^{2s}\, p_{\delta,\Sigma}\left(\frac{\langle x, v\rangle}{\sqrt{v^\top \Sigma v}}\right)$. Since $p_{\delta,\Sigma}$ is an even polynomial, $q_{\delta,\Sigma}$ is a polynomial in v.

Definition 4.2.13 (Certifiable Anti-Concentration)**.** A mean-0 distribution D with covariance Σ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if for $q_{\delta,\Sigma}(x, v)$ defined above, there exists a degree-$2s$ sum-of-squares proof of the following two unconstrained polynomial inequalities in indeterminate v:

$$\left\{\langle x,v\rangle^{2s} + \delta^{2s} q_{\delta,\Sigma}(x,v)^2 \geqslant \delta^{2s}\left(v^\top \Sigma v\right)^{2s}\right\}\ ,\ \left\{\mathop{\mathbb{E}}_{x\sim D} q_{\delta,\Sigma}(x,v)^2 \leqslant C\delta\left(v^\top \Sigma v\right)^{2s}\right\}.$$

An isotropic subset $X \subseteq \mathbb{R}^d$ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if the uniform distribution on X is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated.

Remark 4.2.14. The function $s(\delta)$ can be taken to be $O(\frac{1}{\delta^2})$ for standard Gaussian distribution and the uniform distribution on the unit sphere (see [29] and [89]).

Certifiable Hypercontractivity. 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 4.2.15 (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:

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

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.

Hypercontractivity is an important notion in high-dimensional probability and analysis on product spaces [96]. Kauers, O'Donnell, Tan and Zhou [97] showed certifiable hypercontractivity of Gaussians and more generally product distributions with subgaussian marginals. Certifiable hypercontractivity strictly generalizes the better known *certifiable subgaussianity* property (studied first in [12]) that controls higher moments of linear polynomials.

Observe that the definition above is affine invariant. In particular, we immediately obtain:

Fact 4.2.16. *Given $t \in \mathbb{N}$, if a random variable x on $\mathbb{R}^d$ has t-certifiable C-hypercontractive degree-2 polynomials, then so does Ax for any $A \in \mathbb{R}^{d\times d}$.*

As observed in [12], the Gaussian distribution is t-certifiably 1-subgaussian and t-certifiable 1-hypercontractive for every t. Next, we establish certifiable hypercontractivity for mixtures of Gaussians. We defer the proofs to Appendix 4.8.

Lemma 4.2.17 (Shifts Cannot Decrease Variance)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$, Q be a $d \times d$ matrix-valued indeterminate, and C be a scalar-valued indeterminate. Then, we have that*

$$\Big|\frac{Q,C}{2} \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right] \leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x)-C)^2\right]\right\} .$$

Lemma 4.2.18 (Shifts of Certifiably Hypercontractive Distributions)**.** *Let x be a mean-0 random variable with distribution $\mathcal{D}$ on $\mathbb{R}^d$ with t-certifiably C-hypercontractive degree-2 polynomials. Then, for any fixed constant vector $c \in \mathbb{R}^d$, the random variable $x + c$ also has t-certifiable $4C$-hypercontractive degree-2 polynomials.*

Lemma 4.2.19 (Mixtures of Certifiably Hypercontractive Distributions)**.** *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have t-certifiable C-hypercontractive degree-2 polynomials on $\mathbb{R}^d$, for some fixed constant C. Then, any mixture $\mathcal{D} = \sum_i w_i\mathcal{D}_i$ also has t-certifiably (C/α)-hypercontractive degree-2 polynomials for $\alpha = \min_{i\leqslant k, w_i>0} w_i$.*

Corollary 4.2.20 (Certifiable Hypercontractivity of Mixtures of k Gaussians)**.** *Let $\mathcal{M}$ be a k-mixture of Gaussians $\sum_i w_i\mathcal{N}(\mu_i,\Sigma_i)$ with weights $w_i \geqslant \alpha$ for every $i \in [k]$. Then, for all $t \in \mathbb{N}$, $\mathcal{D}$ has t-certifiably $4/\alpha$-hypercontractive degree-2 polynomials.*

The last line in Fact 2.2.3 asserts a bound (along with a degree 2 SoS proof) on the variance of degree 2 polynomials in terms of the Frobenius norm of its coefficient matrix. In the next few claims, we verify this property via elementary arguments for the two classes of distributions relevant to this chapter. We note that whenever a distribution satisfies the bounded variance property (without an SoS proof), it also satisfies the property via a degree

2 SoS proof using Lemma 3.3.7. Thus, asking for an SoS proof of degree 2 in this context poses no additional restrictions on the distribution. Nevertheless, we provide explicit and direct SoS proofs in the following.

We first note that this property of having *certifiable bounded variance* is closed under linear transformations.

Lemma 4.2.21 (Linear Transformations of Certifiably Bounded-Variance Distributions)**.** *For $d \in \mathbb{N}$, let x be a random variable with distribution $\mathcal{D}$ on $\mathbb{R}^d$ such that for $d \times d$ matrix-valued indeterminate Q, $\Big|\frac{Q}{2} \left\{\mathbb{E}_{x\sim\mathcal{D}}(x^\top Qx - \mathbb{E}_{\mathcal{D}}\, x^\top Qx)^2 \leqslant \left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2\right\}$. Let A be an arbitrary $d \times d$ matrix and let $x' = Ax$ be the random variable with covariance $\Sigma' = AA^\top$. Then, we have that*

$$\Big|\frac{Q}{2} \left\{ \mathop{\mathbb{E}}_{x'\sim\mathcal{D}'} (x'^\top Qx' - \mathop{\mathbb{E}}_{\mathcal{D}'} x'^\top Qx')^2 \leqslant \left\|\Sigma'^{1/2}Q\Sigma'^{1/2}\right\|_F^2 \right\} .$$

Lemma 4.2.22 (Variance of Degree-2 Polynomials of Standard Gaussians)**.** *We have that*

$$\Big|\frac{Q}{2} \left\{ \mathop{\mathbb{E}}_{\mathcal{N}(0,I)} \left(x^\top Qx - \mathop{\mathbb{E}}_{\mathcal{N}(0,I)} x^\top Qx \right)^2 \leqslant 3 \left\|Q\right\|_F^2 \right\} .$$

Remark 4.2.23. As is easy to verify, the same proof more generally holds for any distribution that has the same first four moments as the zero-mean Gaussian distribution.

As an immediate corollary of the previous two lemmas, we have:

Corollary 4.2.24 (Variance of Degree-2 Polynomials of Zero-Mean, Arbitrary Covariance Gaussians)**.** *For any* $0 \preceq \Sigma$, *we have that*

$$\Big|\frac{Q}{2} \left\{ \mathop{\mathbb{E}}_{\mathcal{N}(0,\Sigma)} \left(x^\top Qx - \mathop{\mathbb{E}}_{\mathcal{N}(0,\Sigma)} x^\top Qx \right)^2 \leqslant 3 \left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2 \right\} .$$

We next prove that the same property holds for mixtures of Gaussians satisfying certain conditions.

Lemma 4.2.25 (Variance of Degree-2 Polynomials of Mixtures). *Let $\mathcal{M} = \sum_i w_i \mathcal{D}_i$ be a k-mixture of distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ with means μ_i and covariances Σ_i. Let $\mu = \sum_i w_i \mu_i$ be the mean of $\mathcal{M}$. Suppose that each of $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have certifiably C-bounded-variance i.e. for Q: a symmetric $d \times d$ matrix-valued indeterminate.*

$$\left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{x' \sim \mathcal{D}_i} (x'^{\top} Q x' - \mathop{\mathbb{E}}_{\mathcal{D}_i} x'^{\top} Q x')^2 \leqslant C \left\| \Sigma'^{1/2} Q \Sigma'^{1/2} \right\|_F^2 \right\} .$$

Further, suppose that for some $H > 1$, $\|\mu_i - \mu\|_2^2, \|\Sigma_i - I\|_F \leqslant H$ for every $1 \leqslant i \leqslant k$. Then, we have that

$$\left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[\left(x^{\top} Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[x^{\top} Q x \right] \right)^2 \right] \leqslant 100 C H^2 \|Q\|_F^2 \right\} .$$

As an immediate corollary of Lemma 4.2.21 and Lemma 4.2.25, we obtain:

Lemma 4.2.26 (Variance of Degree-2 Polynomials of Mixtures of Gaussians). *Let $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with $w_i \geqslant \alpha$, mean $\mu = \sum_i w_i \mu_i$ and covariance $\Sigma = \sum_i w_i ((\mu_i - \mu)(\mu_i - \mu)^{\top} + \Sigma_i)$. Suppose that for some $H > 1$, $\left\| \Sigma^{\dagger/2} (\Sigma_i - \Sigma) \Sigma^{\dagger/2} \right\|_F \leqslant H$ for every $1 \leqslant i \leqslant k$. Let Q be a symmetric $d \times d$ matrix-valued indeterminate. Then for $H' = \max\{H, 1/\alpha\}$,*

$$\left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[\left(x^{\top} Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[x^{\top} Q x \right] \right)^2 \right] \leqslant 100 H'^2 \left\| \Sigma^{1/2} Q \Sigma^{1/2} \right\|_F^2 \right\} .$$

Analytic Properties are Inherited by Samples. The following lemma can be proven via similar, standard techniques as in several prior works [12, 29, 89, 43].

Fact 4.2.27. *Let D be a distribution on $\mathbb{R}^d$ with mean μ and covariance Σ. Let $t \in \mathbb{N}$. Let X be a sample from D such that, $\left\| \frac{1}{|X|} \sum_{x \in X} (1, \bar{x})^{\otimes t} - \mathbb{E}_{x \sim D} (1, \bar{x})^{\otimes t} \right\|_F \leqslant d^{-O(t)}$. Here, $\bar{x} = \Sigma^{\dagger/2}(x - \mu_i)$. Then,*

1. *If D is $2t$-certifiably C-subgaussian, then the uniform distribution on X is t-certifiably*

$2C$*-subgaussian.*

2. *If D has $2t$-certifiably C-hypercontractive degree 2 polynomials, then the uniform distribution on X has t-certifiably $2C$-hypercontractive degree 2 polynomials.*

3. *If D is $2t$-certifiably $C\delta$-anti-concentrated, then the uniform distribution on X is t-certifiably $2C\delta$-anti-concentrated.*

4. *If* $\vdash_{2}^{Q} \left\{\mathbb{E}_{x\sim\mathcal{D}}(x^\top Qx - \mathbb{E}_{x\sim\mathcal{D}}\, x^\top Qx)^2 \leqslant C\,\|Q\|_F^2\right\}$*, then, for the uniform distribution* $\mathcal{D}_X$ *on* X, $\vdash_{2}^{Q} \left\{\mathbb{E}_{x\sim\mathcal{D}_X}(x^\top Qx - \mathbb{E}_{x\sim\mathcal{D}_X}\, x^\top Qx)^2 \leqslant 2C\,\|Q\|_F^2\right\}$.

4.2.3 Deterministic Conditions on the Uncorrupted Samples

In this section, we describe the set of deterministic conditions on the set of uncorrupted samples, under which our algorithms succeed. We will require the following definition.

Definition 4.2.28. Fix $0 < \varepsilon < 1/2$. We say that a multiset Y of points in $\mathbb{R}^d$ is an ε-corrupted version (or an ε-corruption) of a multiset X of points in $\mathbb{R}^d$ if $|X \cap Y| \geqslant \max\{(1-\varepsilon)|X|, (1-\varepsilon)|Y|\}$.

Throughout this chapter and unless otherwise specified, we will use X to denote a multiset of i.i.d. samples from the target k-mixture $\mathcal{M} = \sum_{i=1}^k w_i G_i$, where $G_i = \mathcal{N}(\mu_i, \Sigma_i)$. We will use X_i for the subset of points in X drawn from G_i, i.e., $X = \cup_{i=1}^k X_i$.

We will use Y to denote an ε-corrupted version of X, as per Definition 4.2.28. In this *strong contamination model*, the adversary can see the clean samples from X before they decide on the ε-corruption Y. The strong contamination model is known to subsume the total variation contamination of Definition 1.1.2 (see, e.g., Section 2 of [21]). We note that our robust learning algorithm succeeds in this stronger contamination model, with the additional requirement that we can obtain two sets of independent ε-corrupted samples from $\mathcal{M}$. (The second set is needed to run a hypothesis testing routine after we obtain a small list of candidate hypotheses.)

Our algorithm works for any finite set of points in $\mathbb{R}^d$ that satisfies a natural set of deterministic conditions. As we will show later in this section, these deterministic conditions are satisfied with high probability by a sufficiently large set of i.i.d. samples from any k-mixture of Gaussians.

Condition 4.2.29 (Good Sample)**.**
Let $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians in $\mathbb{R}^d$. Let X be a set of n points in $\mathbb{R}^d$. We say that X is a good sample (satisfies Condition 4.2.29) with respect to $\mathcal{M}$ with parameters (γ, t) if there is a partition of X as $X_1 \cup X_2 \cup \ldots \cup X_k$ such that:

1. *For all $i \in [k]$ with $w_i \geqslant \gamma$, any positive integer $m \leqslant t$, and any $v \in \mathbb{R}^d$,*

$$\left| \frac{1}{n} \sum_{x \in X_i} \langle v, x - \mu_i \rangle^m - w_i \mathop{\mathbb{E}}_{x \sim \mathcal{N}(\mu_i, \Sigma_i)} [\langle v, x - \mu_i \rangle^m] \right| \leqslant w_i \, \gamma \, m! \, (v^T \Sigma_i v)^{m/2} \, .$$

2. *For all $i \in [k]$ and any halfspace $H \subset \mathbb{R}^d$, we have that $\left| |X_i \cap H|/n - w_i \, \mathbb{P}_{x \sim \mathcal{N}(\mu_i, \Sigma_i)}[x \in H] \right| \leqslant \gamma$.*

We will also need the following consequences of Condition 4.2.29. The first one is immediate.

Lemma 4.2.30. *Condition 4.2.29 is invariant under affine transformations. In particular, if $A(x) : \mathbb{R}^d \to \mathbb{R}^{d'}$ is an affine transformation, and if X is a good sample (satisfies Condition 4.2.29) with respect to $\mathcal{M}$ with parameters (γ, t), then $A(X)$ is a good sample (satisfies Condition 4.2.29) with respect to $A(\mathcal{M})$ with parameters (γ, t).*

We note that the first part of Condition 4.2.29 implies that higher moment tensors are close in Frobenius distance.

Lemma 4.2.31. *If $X$is a good sample (satisfies Condition 4.2.29) with respect to $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ with parameters (γ, t), then if $w_i \geqslant \gamma$ for all $i \in [k]$, and if for some $B \geqslant 0$*

we have that $\|\mu_i\|_2^2, \|\Sigma_i\|_{\mathrm{op}} \leqslant B$ *for all* $i \in [k]$*, then for all* $m \leqslant t$*, we have that:*

$$\left\| \mathop{\mathbb{E}}_{x \in_u X}[x^{\otimes m}] - \mathop{\mathbb{E}}_{x \sim M}[x^{\otimes m}] \right\|_F^2 \leqslant \gamma^2 m^{O(m)} B^m d^m \ .$$

We note that Condition 4.2.29 also behaves well with respect to taking submixtures.

Lemma 4.2.32. *Let* $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$*. Let* $S \subset [k]$ *with* $\sum_{i \in S} w_i = w$*, and let* $\mathcal{M}' = \sum_{i \in S}(w_i/w)\mathcal{N}(\mu_i, \Sigma_i)$*. Then if* X *is a good sample (satisfies Condition 4.2.29) with respect to* $\mathcal{M}$ *with parameters* (γ, t) *for some* $\gamma < 1/(2k)$ *with the corresponding partition being* $X = X_1 \cup X_2 \cup \ldots \cup X_k$*, then* $X' = \bigcup_{i \in S} X_i$ *is a good sample (satisfies Condition 4.2.29) with respect to* $\mathcal{M}'$ *with parameters* $(O(k\gamma/w), t)$*.*

Finally, we show that given sufficiently many i.i.d. samples from a k-mixture of Gaussians, Condition 4.2.29 holds with high probability.

Lemma 4.2.33. *Let* $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ *and let* n *be an integer at least* $kt^{Ct} d^t/\gamma^3$*, for a sufficiently large universal constant* $C > 0$*, some* $\gamma > 0$*, and some* $t \in \mathbb{N}$*. If* X *consists of* n *i.i.d. samples from* $\mathcal{M}$*, then* X *is a good sample (satisfies Condition 4.2.29) with respect to* $\mathcal{M}$ *with parameters* (γ, t) *with high probability.*

The proofs of the preceding lemmas can be found in Appendix 4.8.

4.2.4 Hypothesis Selection

Our algorithm will require a procedure to select a hypothesis from a list of candidates that contains an accurate hypothesis. A number of such procedures are known in the literature (see, e.g., [98, 99, 100, 101, 102, 21]). Here we will use the following variant from [45], showing that we can efficiently perform a hypothesis selection (tournament) step with access to ε-corrupted samples.

Fact 4.2.34 (Robust Tournament, [45])**.** *Let* X *be an unknown distribution,* $\eta \in (0, 1)$*, and let* $H_1, \ldots, H_n$ *be distributions with explicitly computable probability density functions that* *can be efficiently sampled from. Assume furthermore than* $\min_{1\leqslant i\leqslant n}(d_{\mathsf{TV}}(X,H_i)) \leqslant \eta$*. Then there exists an efficient algorithm that given access to* $O(\log(n)/\eta^2)$ ε*-corrupted samples from* X*, where* $\varepsilon \leqslant \eta$*, along with* $H_1,\ldots,H_n$*, computes an* $m \in [n]$ *such that with high probability we have that* $d_{\mathsf{TV}}(X,H_m) = O(\eta)$ *.*

4.3 List-Recovery of Parameters via Tensor Decomposition

In this section, we give an algorithm that takes samples from a k-mixture of Gaussians, whose component means and covariances are not too far from each other in natural norms, and outputs a dimension-independent polynomial size list of candidate k-tuples of parameters (i.e., means and covariances) one of which is guaranteed to be close to the true target k-tuple of parameters. Our approach involves a new tensor decomposition procedure that works in the absence of any non-degeneracy conditions on the components.

The goal of this section is to prove the following theorem:

Theorem 4.3.1 (Recovering a small list of candidate parameters)**.** *Fix any* $\alpha > \varepsilon > 0, \Delta > 0$*. Let* X*, a good sample from a* k*-mixture of Gaussians* $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i,\Sigma_i)$ *satisfying Condition 4.2.29 with parameters* $\gamma = \varepsilon d^{-8k}k^{-Ck}$*, for* C *a sufficiently large universal constant, and* $t = 8k$*, and let* Y *be an* ε*-corruption of* X*. Let* X' *be a set of* $n' = O\left(\varepsilon\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4ck}$ *fresh samples from* $\mathcal{M}$ *and* Z *be an* ε*-corruption of* X'*. If* $w_i \geqslant \alpha$*,* $\|\mu_i\|_2 \leqslant \frac{2}{\sqrt{\alpha}}$ *and* $\|\Sigma_i - I\|_F \leqslant \Delta$ *for every* $i \in [k]$*, then, given* k, Y, Z *and* ε*, the algorithm outputs a list* L *of at most* $\ell' = O\left(\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)^{4k}/\eta^{4k}\right)$ *candidate hypotheses (component means and covariances), such that with probability at least* $99/100$ *there exist* $\{\hat\mu_i,\hat\Sigma_i\}_{i\in[k]} \subseteq L$ *satisfying* $\|\mu_i - \hat\mu_i\|_2 \leqslant O\!\left(\frac{\Delta^{1/2}}{\alpha}\right)\eta^{G(k)}$ *and* $\left\|\Sigma_i - \hat\Sigma_i\right\|_F \leqslant O(k^4)\,\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)}$*, for all* $i \in [k]$*. Here,* $\eta = (2k)^{4k}O(1/\alpha+\Delta)^{4k}\,\varepsilon^{1/(k^{O(k^2)})}$ *and* $G(k) = \frac{1}{C^{k+1}(k+1)!}$*. The running time of the algorithm is* $\mathrm{poly}(|L|,|Y|,d^k)\cdot \mathrm{poly}_k(1/\varepsilon)$*.*

In the body of this section, we establish Theorem 4.3.1. The structure of this section is as follows: In Section 4.3.1, we describe our first list-decodable algorithm to output a hypotheses list with size exponential in ε, which is then analyzed in Sections 4.3.2-4.3.6. In Sections 4.3.7, we then improve the above algorithm to generate a *polynomial* size of hypotheses and achieve Theorem 4.3.1.

4.3.1 List-Decodable Tensor Decomposition Algorithm

In this section, we describe our tensor decomposition algorithm, which is given in pseudocode below (Algorithm 4.3.2).

Algorithm 4.3.2 (List-Recovery of Candidate Parameters via Tensor Decomposition)**.**

Input: *An ε-corruption Y of a sample X from a k-mixture of Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$.*

Requirements: *The guarantees of the algorithm hold if the mixture parameters and the sample X satisfy:*

1. *$w_i \geqslant \alpha$ for all $i \in [k]$,*
2. *$\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$ for all $i \in [k]$,*
3. *$\|\Sigma_i - I\|_F \leqslant \Delta$ for all $i \in [k]$.*
4. *X is a good sample (satisfies Condition 4.2.29) with parameters (γ, t), where $\gamma = \varepsilon d^{-8k} k^{-Ck}$, for C a sufficiently large universal constant, and $t = 8k$.*

Parameters: $\eta = (2k)^{4k}(Ck(1/\alpha+\Delta))^{4k}\sqrt{\varepsilon}$, $D = C(k^4/(\alpha\sqrt{\eta}))$, $\delta = 2\eta^{1/(C^{k+1}(k+1)!)}$, $\ell' = 100 \log k \left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4k}$, *for some sufficiently large absolute constant* $C > 0$, $\lambda = 4\eta$, $\phi = 10(1+\Delta^2)/(\sqrt{\eta}\alpha^5)$.

Output: *A list L of hypotheses such that there exists at least one, $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\leqslant k} \in L$, satisfying: $\|\mu_i - \hat{\mu}_i\|_2 \leqslant O\left(\frac{\Delta^{1/2}}{\alpha}\right)\eta^{G(k)}$ and $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant O(k^4)\,\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)}$, where $G(k) = \frac{1}{C^{k+1}(k+1)!}$.*

Operation:

1. ***Robust Estimation of Hermite Tensors****: For $m \in [4k]$, compute $\hat{T}_m$ such that $\max_{m\in[4k]} \left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$ using the robust mean estimation algorithm in Fact 2.2.1.*

2. ***Random Collapsing of Two Modes of $\hat{T}_4$****: Let L' be an empty list. Repeat ℓ' times: For $j \in [4k]$, choose independent standard Gaussians in $\mathbb{R}^d$, denoted by $x^{(j)}, y^{(j)} \sim \mathcal{N}(0, I)$, and uniform draws $a_1, a_2, \dots, a_t$ from $[-D, D]$. Let $\hat{S}$ be a $d \times d$ matrix such that for all $r, s \in [d]$, $\hat{S}(r,s) = \sum_{j\in[4k]} a_j \hat{T}_4(r, s, x^{(j)}, y^{(j)}) = \sum_{j\in[4k]} a_j \sum_{g,h\in[d]} \hat{T}_4(r,s,g,h)x^{(j)}(g)y^{(j)}(h)$. Add $\hat{S}$ to the list L'.*

3. ***Construct Low-Dimensional Subspace for Exhaustive Search****: Let V be the span of all singular vectors of the natural $d \times d^{m-1}$ flattening of $\hat{T}_m$ with singular values $\geqslant \lambda$ for $m \leqslant 4k$. For each $\hat{S} \in L'$, let $V'_{\hat{S}}$ be the span of V plus all the singular vectors of $\hat{S}$ with singular value larger than $\delta^{1/4}$.*

4. ***Enumerating Candidates in $V'_{\hat{S}}$****: Initialize L to be the empty list. For each $\hat{S} \in L'$, let $V_{\delta^{1/4}}$ be a $\delta^{1/4}$-cover of vectors in $V'_{\hat{S}}$ with ℓ_2-norm at most $2/\sqrt{\alpha}$. Enumerate over vectors $\hat{\mu}$ in $V_{\delta^{1/4}}$. Let $k' = Ck^2$ and let $C_{\delta^{1/4}}$ be a $\delta^{1/4}$-cover of the interval $[-\phi, \phi]^{k'}$. For $\{\tau_j\}_{j\in[k']} \in C_{\delta^{1/4}}$ and for all $\{v_j\}_{j\in[k']} \in V_{\delta^{1/4}}$, let $\hat{Q} = \sum_{j\in[k']} \tau_j v_j v_j^\top$. Add $\{\hat{\mu}, I + \hat{S} + \hat{Q}\}$ to L.*

4.3.2 Analysis of Algorithm

We analyze the three main steps of Algorithm 4.3.2 in the following lemmas. We will prove the following three propositions in the subsequent subsections that analyze Steps 1, 2 and 3 of Algorithm 4.3.2. For Step 1, we show that when X is a good sample (satisfies Condition 4.2.29), the empirical estimates of the moment tensors obtained by applying the robust mean estimation algorithm to X are sufficiently close to the moment tensors of the input mixture $\mathcal{M}$.

Proposition 4.3.3 (Robustly Estimating Hermite Polynomial Tensors)**.** *For any integer $m \leqslant 4k$, and $\Delta \in \mathbb{R}_+$, there exists an algorithm with running time* $\mathrm{poly}_m(d/\varepsilon)$ *that takes an ε-corruption Y of X, a good sample (satisfying Condition 4.2.29) with respect to $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ with parameters $\gamma = \varepsilon d^{-m} m^{-Cm}$, for C a sufficiently large constant, and $t = 2m$. If $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for each $i \in [k]$, then the algorithm outputs a tensor $\hat{T}_m$ such that $\left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$, for $\eta = O\left(m(1 + 1/\alpha + \Delta)\right)^m \sqrt{\varepsilon}$.*

The proof of Proposition 4.3.3 is deferred to Section 4.3.3.

Next, we analyze Step 2 of the algorithm and prove that, with non-negligible probability, randomly collapsing two modes of $\hat{T}_4$ yields a matrix $\hat{S}$ such that $\hat{S} - (\Sigma_i - I) = P_i + Q_i$, where P_i has small Frobenius norm and Q_i is a rank-$O(k^2)$ matrix.

Proposition 4.3.4 (Tensor Decomposition up to Low-Rank Error)**.** *Let $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians satisfying $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for each $i \in [k]$. For $0 < \eta < 1$, let $\hat{T}_4$ be a tensor such that $\left\|\mathbb{E}[h_4(\mathcal{M})] - \hat{T}_4\right\|_F \leqslant \eta$, and let D be a sufficiently large constant multiple of $k^4/(\alpha\sqrt{\eta})$. For all $j \in [4k]$, let $x^{(j)}, y^{(j)} \sim \mathcal{N}(0, I)$ be independent and $a_j \sim \mathcal{U}[-D, D]$, where $\mathcal{U}[-D, D]$ is the uniform distribution over the interval $[-D, D]$, and let $\hat{S} = \sum_{j\in[4k]} a_j \hat{T}_4\left(\cdot, \cdot, x^{(j)}, y^{(j)}\right)$. Then, for each $i \in [k]$, with probability at least $\left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{4k}$, over the choice of $x^{(j)}, y^{(j)}$ and a_j, we have that $\hat{S} - (\Sigma_i - I) = P_i + Q_i$, where $\|P_i\|_F = O\left(\sqrt{\eta/\alpha}\right)$, $\|Q_i\|_F = O\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right)$ and* $\mathrm{rank}(Q_i) = O\left(k^2\right)$.

The proof of Proposition 4.3.4 is given in Section 4.3.4.

Finally, in Step 3, for any $\hat{S}$ such that $\hat{S} - (\Sigma_i - I) = P_i + Q_i$, where P_i has small Frobenius norm and Q_i is a rank $O(k^2)$ matrix, we find a low-dimensional subspace V' such that the range space of Q_i is approximately contained in V'. We will use V' to exhaustively search for $O(k^2)$ rank matrices to find candidates for Q_i.

Proposition 4.3.5 (Low-Dimensional Subspace V' for Exhaustive Search)**.** *Let $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians satisfying $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for each $i \in [k]$. Let $\left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$, for each $1 \leqslant m \leqslant 4k$, and some $\eta > 0$. Let V be the span of all the left singular vectors of the $d \times d^{m-1}$ matrix obtained by the natural flattening of $\hat{T}_m$ with singular values at least 2η. For each $1 \leqslant i \leqslant k$, let $S_i = \Sigma_i - I$ and $\hat{S}_i$ be a $d \times d$ matrix such that $\hat{S}_i - S_i = P_i + Q_i$, where $\|P_i\|_F \leqslant O\left(\sqrt{\eta/\alpha}\right)$, Q_i has rank $O(k^2)$, and $\|Q_i\|_F \leqslant O\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right)$. Let V' be the span of V plus all singular vectors of $\hat{S}_i$ of singular values at least δ for all i. Then, for $\delta = 2\eta^{1/(C^{k+1}(k+1)!)}$ with a sufficiently large constant $C > 0$, we have that:*

1. $\dim(V') \leqslant \left(O\left(k(1 + 1/\alpha + \Delta)\right)^{4k+5}\right)/\eta^2$.
2. *There is a vector $\mu_i' \in V'$ such that $\left\|\mu_i - \mu_i'\right\|_2^2 \leqslant \frac{20}{\alpha^2}\sqrt{\delta}\Delta$.*
3. *There are $q = O(k^2)$ unit vectors $v_1, v_2, \ldots, v_q \in V'$ and scalars $\tau_1, \tau_2, \ldots, \tau_q \in \left[-10(1+\Delta^2)/(\sqrt{\eta}\alpha^5), 10(1+\Delta^2)/(\sqrt{\eta}\alpha^5)\right]$ such that $\left\|Q_i - \sum_{i=1}^q \tau_i v_i v_i^\top\right\|_F \leqslant O\left(\frac{k^2}{\alpha}\delta^{1/4}\Delta^{1/2}\right)$.*

The proof of Proposition 4.3.5 is given in Section 4.3.5.

We can now use these propositions to complete the proof of Theorem 4.3.1.

Proof of Theorem 4.3.1. Using Proposition 4.3.3, Step 1 of the algorithm outputs estimates $\hat{T}_i$ for $i \in [4k]$ such that $\max_{m \in [4k]} \left\|\hat{T}_m - \mathbb{E}\, h_m(\mathcal{M})\right\|_F \leqslant \eta$. Next, by the standard coupon collector analysis, using Proposition 4.3.4 and repeating Step 2 of the algorithm $\ell' = 100 \log k \left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4k}$ times, guarantees that with probability at least

$1 - 1/(100k)^{100}$, for every $1 \leqslant i \leqslant k$, there are $\hat{S}_i \in L$ such that $\hat{S}_i - (\Sigma_i - I) = P_i + Q_i$ for P_i, Q_i satisfying $\|P_i\|_F \leqslant \sqrt{\eta/\alpha}$, $\|Q_i\|_F \leqslant \frac{1+\Delta^2}{\sqrt{\eta}\alpha^5}$ and Q_i has rank $O(k^2)$.

Next, Proposition 4.3.5 implies that for every such $\hat{S}_i \in L'$, we can construct a subspace $V' = V'_{\hat{S}_i}$ of dimension $O\big((k(1 + 1/\alpha + \Delta))^{4k+5}/\eta^2\big)$ such that V' contains μ'_i that satisfies $\left\|\mu_i - \mu'_i\right\|_2^2 \leqslant \frac{\Delta}{\alpha^2} \cdot \sqrt{\delta}$, and there is a rank $O(k^2)$ matrix $\hat{Q}_i$ with range space contained in V' such that $\left\|Q_i - \hat{Q}_i\right\|_F \leqslant O(\frac{k^2}{\alpha}\delta^{1/4}\Delta^{1/2})$.

Now, let $V_\tau \subseteq V'$ be a $\tau = \delta^{1/4}$-cover, in ℓ_2-norm, of vectors with ℓ_2 norm at most $2/\sqrt{\alpha}$ in V'. Then, since $\|\mu_i\|_2 \leqslant \frac{2}{\sqrt{\alpha}}$, there is a vector $\hat{\mu}_i \in V_\tau$ such that $\|\mu_i - \hat{\mu}_i\|_2^2 \leqslant \tau + \frac{20}{\alpha^2}\sqrt{\delta}\Delta \leqslant \frac{40}{\alpha^2}\sqrt{\delta}\Delta$.

Further, there exist $\tau_1, \tau_2, \ldots \tau_{O(k^2)}$ in a τ-cover of $[-10(1 + \Delta^2)/(\sqrt{\eta}\alpha^5), 10(1 + \Delta^2)/(\sqrt{\eta}\alpha^5)]$ and vectors $v_1, v_2, \ldots, v_{O(k^2)} \in V_\tau$ such that $\left\|\sum_{i=1}^{O(k^2)} \tau_i v_i v_i^\top - Q_i\right\|_F \leqslant O(k^4\delta^{1/4}\Delta^{1/2}/\alpha)$. In particular, $\hat{\Sigma}_i = I + \hat{S} - \sum_{i=1}^{O(k^2)} \tau_i v_i v_i^\top$ satisfies

$$\left\|\hat{\Sigma}_i - \Sigma_i\right\|_F = O(\sqrt{\eta}) + O\left(\frac{k^4\delta^{1/4}\Delta^{1/2}}{\alpha}\right) = O\left(\frac{k^4\delta^{1/4}\Delta^{1/2}}{\alpha}\right) . \tag{4.3.1}$$

The size of this search space for every fixed $\hat{S} \in L'$ can be bounded above by $\left(\frac{1+\Delta^2}{\delta\alpha^5}\right)^{O(k^5 \dim(V'))}$. Thus, the size of L can be bounded from above by

$$k^5 \left(\frac{\Delta^4}{\eta} + \frac{1}{\alpha^4\eta}\right)^{4k} \cdot \left(\frac{1 + \Delta^2}{\delta\sqrt{\eta}\alpha^5}\right)^{O(k^5 \dim(V'))} \leqslant \exp\left(\log(1/\varepsilon)\,(k + 1/\alpha + \Delta)^{O(k)}/\eta^2\right) .$$

This completes the proof. □

4.3.3 Robust Estimation of Hermite Tensors

In this section, we will prove Proposition 4.3.3.

Proof of Proposition 4.3.3. Consider the uniform distribution on the uncorrupted sample X. We want to analyze the effect of applying the robust mean estimation algorithm (Fact 2.2.1) to the points $h_m(x)$, for $x \in X$. In order for us to apply Fact 2.2.1, we need to ensure that the uniform distribution on $\{h_m(x)\}_{x\in X}$ has bounded covariance. This step gives us a good approximation to $\mathbb{E}_{x\sim_u X} h_m(x)$. In order for us to obtain an approximation to $\mathbb{E}\, h_m(\mathcal{M})$, we need to bound the difference between $\mathbb{E}\, h_m(\mathcal{M})$ and $\mathbb{E}_{x\sim_u X} h_m(x)$. We will do both these steps below.

The second part is immediate. By the definition of $h_m(X)$, we have that

$$\left\|\frac{1}{|X|}\sum_{x\in X} h_m(x) - \mathbb{E}\, h_m(\mathcal{M})\right\|_F \leqslant \sum_{j\leqslant m/2} m^{2j} d^j \left\|\frac{1}{|X|}\sum_{x\in X} x^{\otimes(m-2j)} - \mathbb{E}\,\mathcal{M}^{\otimes(m-2j)}\right\|_F .$$

By Lemma 4.2.31, this is at most $O(1+\Delta+1/\alpha)^m m^{O(m)} d^{m/2}\gamma \leqslant \eta/2$. We note that a similar argument bounds

$$\left\|\frac{1}{|X|}\sum_{x\in X} h_m(x)\otimes h_m(x) - \mathbb{E}\, h_m(\mathcal{M})\otimes h_m(\mathcal{M})\right\|_F \leqslant \eta^2.$$

Let us now verify the first part. We proceed via bounding the operator norm of the covariance of $h_m(\mathcal{M})$. We can then use the bound on the Frobenius norm $\left\|\frac{1}{|X|}\sum_{x\in X} h_m(x)\otimes h_m(X) - \mathbb{E}\, h_m(\mathcal{M})\otimes h_m(\mathcal{M})\right\|_F$ to get a bound on $\left\|\frac{1}{|X|}\sum_{x\in X} h_m(x)h_m(x)^\top\right\|_{\mathrm{op}}$ (the operator norm of the canonical square flattening of the of the $2m$-th empirical Hermite moment tensor of X). This will complete the proof.

Let $G_i = \mathcal{N}(\mu_i,\Sigma_i)$ be the components of $\mathcal{M}$. We have that

$$\begin{aligned}\mathtt{Cov}(h_m(\mathcal{M})) = \sum_{i\in[k]} w_i \mathtt{Cov}(h_m(G_i))&\\ + \frac{1}{2}\sum_{i,j\in[k]} w_i w_j \left(\mathbb{E}[h_m(G_i)] - \mathbb{E}[h_m(G_j)]\right)\left(\mathbb{E}[h_m(G_i)] - \mathbb{E}[h_m(G_j)]\right)^\top &.\end{aligned} \tag{4.3.2}$$

By Lemma 4.2.8, we have that for all $i\in[k]$, it holds

$$\|\mathtt{Cov}(h_m(G_i))\|_{\mathrm{op}} = \mathcal{O}\left(m(1+\|\mu_i\|_2 + \|\Sigma_i - I\|_F)\right)^{2m} = \mathcal{O}\left(m(1+2/\sqrt{\alpha}+\Delta)\right)^{2m},$$

where for any matrix M, $\|M\|_{\mathrm{op}} = \max_{\|u\|_2=1} \|Mu\|_2$ is the operator norm of the matrix. Further, for any $i, j \in [k]$,

$$\begin{aligned}\left\|\left(\mathbb{E}[h_m(G_i)] - \mathbb{E}[h_m(G_j)]\right)\left(\mathbb{E}[h_m(G_i)] - \mathbb{E}[h_m(G_j)]\right)^\top\right\|_{\mathrm{op}} &= \left\|\mathbb{E}[h_m(G_i)] - \mathbb{E}[h_m(G_j)]\right\|_2 \\ &= O\left(m(1+1/\alpha+\Delta)\right)^{2m} .\end{aligned} \tag{4.3.3}$$

This claim follows from the triangle inequality of the operator norm. □

4.3.4 List-Recovery of Covariances up to Low-Rank Error

In this section, we prove Proposition 4.3.4. We first set some useful notation. We will write $S_i \stackrel{\text{def}}{=} \Sigma_i - I$ throughout this section. We will also use S_i' to denote $S_i + \mu_i \otimes \mu_i$.

We first show that for every i, there exists a matrix P such that $\left(\sum_{i\in[k]} w_i S_i' \otimes S_i'\right)(\cdot,\cdot,P)$ is close to S_i'.

Lemma 4.3.6 (Existence of a 2-Tensor)**.** *Under the hypothesis of Proposition 4.3.4, for each $i \in [k]$, there exists a matrix P such that $\|P\|_F = O\big(1/(\sqrt{\eta}\alpha)\big)$ and $\left\|T_4'(\cdot,\cdot,P) - S_i'\right\|_F = O\left(\sqrt{\eta/\alpha}\right)$, where $T_4' = \left(\sum_{i\in[k]} w_i S_i' \otimes S_i'\right)$.*

Note that throughout this section it will be useful to think of T_4' as a $d^2 \times d^2$ matrix rather than as a tensor. In this case, we can think of T_4' as $\sum_{i=1}^k w_i (S_i')(S_i')^T$. From standard facts about positive semidefinite matrices it follows that S_i' is in the image of T_4', and Lemma 4.3.6 is just a slightly robustified version of this (saying that we can find an approximate preimage that it not itself too large).

The proof of this Lemma 4.3.6 will involve linear programming duality with an infinite system of constraints. As the application of duality with infinitely many constraints has some technical issues, we state below an appropriate version of duality.

Fact 4.3.7 (Linear Programming Duality for Compact, Convex Constraint Sets)**.** *Let $K \subset \mathbb{R}^{n+1}$ be a compact convex set. There exists an $x \in \mathbb{R}^n$ so that $(x,1)\cdot z > 0$ for all $z \in K$ if* *and only if there is no element* $(0, 0, \ldots, 0, a) \in K$ *for any* $a \leqslant 0$.

This fact can be proved by noting that if no such a exists, there must be a hyperplane separating K from the set of such points $(0, a)$. This separating hyperplane will be of the form $(z, y) \in H$ if and only if $y = x \cdot z$ for some x and this x will provide the solution to the linear system.

Proof of Lemma 4.3.6. To show that such a P exists for each i, we apply linear programming duality. In particular, the conditions imposed on P define a linear program, which has a feasible solution unless there is a solution to the dual linear program. For sufficiently large constants c_1 and c_2, consider the following primal in the variable P:

$$\langle v, P \rangle \leqslant \frac{c_1}{\sqrt{\eta}\alpha} \|v\|_F \qquad \forall \ v \in \mathbb{R}^{d \times d} \tag{4.3.4}$$

$$\left\langle u, T_4'(\cdot, \cdot, P) - S_i' \right\rangle \leqslant c_2 \sqrt{\eta} \|u\|_F \qquad \forall \ u \in \mathbb{R}^{d \times d}. \tag{4.3.5}$$

It is not hard to see that $\|P\|_F \leqslant \frac{c_1}{\sqrt{\eta}\alpha}$ if and only if (4.3.4) holds for all v and $\left\|T_4'(\cdot, \cdot, P) - S_i'\right\|_F \leqslant c_2 \sqrt{\eta/\alpha}$ if and only if (4.3.5) holds for all u. Throughout the proof, we suggest that the reader think of u and v as vectors in d^2-dimensional vector space.

Our goal is to show that there exists a feasible solution P such that (4.3.4) and (4.3.5) hold simultaneously for all $u, v \in \mathbb{R}^{d \times d}$. We first note that this is equivalent to saying that

$$\langle v, P \rangle + \left\langle u, T_4'(\cdot, \cdot, P) \right\rangle - \left\langle u, S_i' \right\rangle \leqslant \frac{c_1}{\sqrt{\eta}\alpha} \|v\|_F + c_2 \sqrt{\eta} \|u\|_F \ , \tag{4.3.6}$$

for all $u, v \in \mathbb{R}^{d \times d}$. This is not quite in the form necessary to apply Fact 4.3.7, so we note that this is in turn equivalent to saying that

$$\langle v, P \rangle + \left\langle u, T_4'(\cdot, \cdot, P) \right\rangle - \left\langle u, S_i' \right\rangle \leqslant 1 \ , \tag{4.3.7}$$

for all $u, v \in \mathbb{R}^{d \times d}$ so that $\frac{c_1}{\sqrt{\eta}\alpha} \|v\|_F + c_2 \sqrt{\eta} \|u\|_F \leqslant 1$, and $u \in \operatorname{span}\{S_i'\}$. As this is a convex set of linear equations, we have by Fact 4.3.7 that there exists such a P unless there exists such a pair of u and v so that the coefficient of P in Equation (4.3.7) is 0 and so that the resulting inequality of constants is either false or holds with equality. In particular, the coefficient of P vanishes if and only if $v = -T_4'(u, \cdot, \cdot)$. We then get a contradiction only if for some $u \in \operatorname{span}\{S_i'\}$

$$-\left\langle u, S_i' \right\rangle \geqslant 1 \geqslant \frac{c_1}{\sqrt{\eta}\alpha} \left\| T_4'(u, \cdot, \cdot) \right\|_F + c_2\sqrt{\eta} \left\| u \right\|_F . \tag{4.3.8}$$

We claim that this is impossible.

In particular, squaring Equation (4.3.8) would give

$$\begin{aligned} \left\langle u, S_i' \right\rangle^2 &\geqslant \left(\frac{c_1}{\sqrt{\eta}\alpha} \left\| T_4'(u, \cdot, \cdot) \right\|_F + c_2\sqrt{\eta} \left\| u \right\|_F \right)^2 \\ &\geqslant \frac{c}{\alpha} \left\| T_4'(u, \cdot, \cdot) \right\|_F \cdot \left\| u \right\|_F \ , \end{aligned} \tag{4.3.9}$$

for some large enough constant $c > 1$, where the last inequality follows from the AM-GM inequality. However, using the dual characterization of the Frobenius norm, we have

$$\left\| T_4'(u, \cdot, \cdot) \right\|_F \geqslant \frac{\left\langle u, T_4'(u, \cdot, \cdot) \right\rangle}{\left\| u \right\|_F} \geqslant \frac{w_i}{\left\| u \right\|_F} \left\langle u, S_i' \right\rangle^2 \ , \tag{4.3.10}$$

where the last inequality follows from T_4' containing a $w_i S_i \otimes S_i$ term, and the other terms contributing non-negatively. Rearranging Equation (4.3.10), we have

$$\left\langle u, S_i' \right\rangle^2 \leqslant \frac{1}{w_i} \left\| T_4'(u, \cdot, \cdot) \right\|_F \left\| u \right\|_F \leqslant \frac{1}{\alpha} \left\| T_4'(u, \cdot, \cdot) \right\|_F \left\| u \right\|_F \ .$$

This contradicts Equation (4.3.9) unless $T_4'(u, \cdot, \cdot) = 0$. This therefore suffices to prove the feasibility of the primal. □

We have thus shown that there is some matrix P so that $T_4'(P, \cdot, \cdot)$ suffices for our purposes. We need to show that our appropriate random linear combination of $x^{(j)} \otimes y^{(j)}$

suffices. In fact, we will show that with reasonably high probability over our choice of $x^{(j)}, y^{(j)}$ that there is some linear combination of the $x^{(j)} \otimes y^{(j)}$ (with coefficients that are not too large) so that their projection onto the space spanned by the S_i' (which is all that matters when applying T_4') equal to P.

For the sake of intuition, we note that if we removed the bound on the coefficients, we would need that the projections of the $x^{(j)} \otimes y^{(j)}$ spanned span$\{S_i'\}$. Since there are at least k of them, this will hold unless there is some $v \in$ span$\{S_i'\}$ so that v is orthogonal to all of the $x^{(j)} \otimes y^{(j)}$. This shouldn't happen because each $x^{(j)} \otimes y^{(j)}$ is very unlikely to be orthogonal to v.

To deal with the constraint that the coefficients are not too large, we use linear programming duality to show that there will be a solution unless there is some v that is *nearly* orthogonal to all of the $x^{(j)} \otimes y^{(j)}$. Again, this is unlikely to happen for any individual term, and thus, by independence, highly unlikely to happen for all j simultaneously. Combining this with a cover argument will give our proof.

Lemma 4.3.8 (Existence of a Bi-Linear Form). *Given the preconditions in Proposition 4.3.4, with probability at least* $99/100$ *over the choice of* $x^{(j)}, y^{(j)}$*, there exist* $b_j \in [-D, D]$ *for* $j \in [4k]$*, where* $D = O\big(k^4/(\sqrt{\eta}\alpha)\big)$*, such that the projection of* $\sum_{j=1}^{t} b_j x^{(j)} \otimes y^{(j)}$ *onto the space spanned by the* S_i' *is* P*, where* P *satisfies the conclusion of Proposition 4.3.6.*

Proof. To prove this lemma, we again use a linear programming based argument. Consider the following (primal) linear program in the variables b_j, for $j \in [4k]$:

$$\sum_{j \in [4k]} b_j \langle S_i', x^{(j)} \otimes y^{(j)} \rangle = \langle S_i', P \rangle \qquad \forall \;\; i \in [k] \tag{4.3.11}$$

$$-D \leqslant b_j \leqslant D \qquad \forall \;\; j \in [4k] \tag{4.3.12}$$

We note that a set of b_j satisfying Equation (4.3.11) will have the projection of $\sum_{j \in [4k]} b_j x^{(j)} \otimes y^{(j)}$ onto the span of the S_i' be the same as the projection of P, and that if the

b_j's satisfy Equation (4.3.12) then we will have $|b_j| \leqslant D$ for all j. Thus, it suffices to show that with high probability over our choice of $x^{(j)}$ and $y^{(j)}$ that the above system is feasible.

We will show this by linear programming duality (since this is now a finite system of equations, we can use standard results rather than Fact 4.3.7). In particular, we have that Equations (4.3.11) and (4.3.12) are simultaneously satisfiable unless there are real numbers c_i and non-negative real numbers z_j, z'_j so that

$$\sum_{i=1}^{k} c_i \sum_{j \in [4k]} b_j \langle S'_i, x^{(j)} \otimes y^{(j)} \rangle + \sum_{j \in [4k]} (z_j - z'_j) b_j \leqslant \sum_{i=1}^{k} c_i \langle S'_i, P \rangle + \sum_{j \in [4k]} (z_j + z'_j) D$$

yields a contradiction. Setting $v = \sum_{i=1}^{k} c_i S'_i$, the above simplifies to

$$\sum_{j \in [4k]} b_j \left(\langle v, x^{(j)} \otimes y^{(j)} \rangle + z_j - z'_j \right) \leqslant \langle v, P \rangle + \sum_{j \in [4k]} (z_j + z'_j) D \qquad (4.3.13)$$

We note that in order for Equation (4.3.13) to be a contradiction, it must be the case that the coefficients of b_j are all 0. In particular, we must have

$$z'_j - z_j = \langle v, x^{(j)} \otimes y^{(j)} \rangle$$

for all j. In particular, this means that

$$z_j + z'_j \geqslant \left| \langle v, x^{(j)} \otimes y^{(j)} \rangle \right| .$$

In such a case, the right hand side of Equation (4.3.13) will be at least

$$\langle v, P \rangle + \sum_{j \in [4k]} \left| \langle v, x^{(j)} \otimes y^{(j)} \rangle \right| D$$

Therefore, Equation (4.3.13) can only yield a contradiction if there exists a $v \in \mathrm{span}\{S'_i\}$ so that

$$\langle v, P\rangle < -\sum_{j\in[4k]} \left|\langle v, x^{(j)} \otimes y^{(j)}\rangle\right| D. \tag{4.3.14}$$

We want to show that with high probability over our choice of $x^{(j)}, y^{(j)}$ that there is no $v \in \operatorname{span}\{S'_i\}$ satisfying Equation (4.3.14). In fact, we will show that for every such v that

$$\sum_{j\in[4k]} \left|\langle v, x^{(j)} \otimes y^{(j)}\rangle\right| \geqslant \frac{c_1}{\sqrt{\eta}\alpha} \|v\|_F .$$

We can scale v so that $\|v\|_F = 1$, and it suffices to show that

$$\sum_{j\in[4k]} \left|\left\langle \tilde{v}, x^{(j)} \otimes y^{(j)}\right\rangle\right| \geqslant \left(\frac{c_1}{\sqrt{\eta}\alpha D}\right) \tag{4.3.15}$$

holds for all unit vectors v in $\operatorname{span}\{S'_i\}$ with high probability.

Since we need to show that infinitely many equations all hold with high probability, we will use a cover argument. In particular, we can construct C, a τ-cover for all unit vectors v in the span of the S'_i, where we take $\tau = \left(\frac{c'_1}{k^2\sqrt{\eta}\alpha D}\right)$. Since this is a cover of a unit sphere in a k-dimensional subspace, we can construct such a cover so that $|C| = O(1/\tau)^k$. Replacing v with the closest point in C, denoted by v', it suffices to show that with high probability for all v that

$$\sum_{j\in[4k]} \left|\left\langle v, x^{(j)} \otimes y^{(j)}\right\rangle\right| \geqslant \sum_{j\in[4k]} \left|\left\langle v', x^{(j)} \otimes y^{(j)}\right\rangle\right| - \sum_{j\in[4k]} \left|\left\langle v - v', x^{(j)} \otimes y^{(j)}\right\rangle\right| \geqslant \left(\frac{2c_1}{\sqrt{\eta}\alpha D}\right). \tag{4.3.16}$$

We begin by bounding the terms

$$\sum_{j\in[4k]} \left|\left\langle v - v', x^{(j)} \otimes y^{(j)}\right\rangle\right|.$$

For this we notice by Cauchy-Schwartz that each term is at most $\|v - v'\|_F$ times the Frobenius norm of the projection of $x^{(j)} \otimes y^{(j)}$ onto the span of the S'_i. We note that for any

k-dimensional subspace W with orthonormal basis $w_1, \ldots, w_k$ we have that

$$\mathbb{E}\left[\left\|\mathrm{Proj}_W(x^{(j)} \otimes y^{(j)})\right\|_F^2\right] = \sum_{i=1}^{k} \left|\left\langle w_i, x^{(j)} \otimes y^{(j)} \right\rangle\right|^2$$
$$= k.$$

Therefore, with high probability over the choice of $x^{(j)}, y^{(j)}$ each of the projections of $x^{(j)} \otimes y^{(j)}$ onto the span of the S'_i has Frobenius norm $\tilde{O}(\sqrt{k})$. Therefore, if this condition holds over our choice of $x^{(j)}$ and $y^{(j)}$, we can show Equation (4.3.16) if we can show that

$$\sum_{j \in [4k]} \left|\left\langle v', x^{(j)} \otimes y^{(j)} \right\rangle\right| \geqslant \left(\frac{c_1}{\sqrt{\eta}\alpha D}\right) \geqslant \left(\frac{2c_1}{\sqrt{\eta}\alpha D}\right) - \tau \tilde{O}(k^{3/2}) \tag{4.3.17}$$

for all $v' \in C$.

Each term in $\sum_{j \in [4k]} \left\langle v', x^{(j)} \otimes y^{(j)} \right\rangle$ is a random bi-linear form given by $z_j = \sum_{\ell, p \in [d]} v'_{\ell,p} x_\ell^{(j)} y_p^{(j)}$. Then, we have that $\mathbb{E}\left[z_j\right] = 0$ and

$$\mathbb{E}\left[z_j^2\right] = \mathbb{E}\left[\left(\sum_{\ell, p \in [d]} v'_{\ell,p} x_\ell^{(j)} y_p^{(j)}\right)^2\right] = \sum_{\ell, \ell', p, p'} \mathbb{E}\left[v'_{\ell,p} v'_{\ell',p'} x_\ell^{(j)} x_{\ell'}^{(j)} y_p^{(j)} y_{p'}^{(j)}\right]$$
$$= \sum_{\ell, p \in [d]} \left(v'_{\ell,p}\right)^2 \cdot \mathbb{E}\left[\left(x_\ell^{(j)}\right)^2\right] \cdot \mathbb{E}\left[\left(y_p^{(j)}\right)^2\right]$$
$$= 1\ ,$$

where the last equality follows from $\tilde{v}'_F = 1$.

Using Lemma 4.2.10 with $\zeta = \frac{2c_1}{\sqrt{\eta}\alpha D}$,

$$\mathbb{P}\left[\left|z_j\right| \leqslant \frac{c_1}{\sqrt{\eta}\alpha D}\right] \leqslant c_5 \left(\frac{2c_1}{\sqrt{\eta}\alpha D}\right)^{1/2} . \tag{4.3.18}$$

However, we note that Equation (4.3.17) will hold unless $\left|z_j\right| \leqslant \frac{c_1}{\sqrt{\eta}\alpha D}$ for all $j \in [4k]$. Since the z_j's are independent, we conclude that

$$\mathbb{P}\left[\sum_{j\in[4k]}\left|\left\langle v', x^{(j)} \otimes y^{(j)}\right\rangle\right| \leqslant \frac{c_1}{\sqrt{\eta}\alpha D}\right] \leqslant O\left(\frac{c_1}{\sqrt{\eta}\alpha D}\right)^{2k} . \tag{4.3.19}$$

Since the above argument holds for any $v' \in C$, we can union bound over all elements in the cover C, and the probability that there exists a $\tilde{v}'$ in the cover that does not satisfy Equation (4.3.17) is at most $O\left(k^2\sqrt{\eta}\alpha D\right)^k \cdot O\left(\frac{c_1}{\sqrt{\eta}\alpha D}\right)^{2k}$. Setting D to be a sufficiently large multiple of $(k^4/(\sqrt{\eta}\alpha))$ suffices to conclude that with probability at least $1-1/\mathrm{poly}(k)$, the primal is feasible. $\square$

Proof of Proposition 4.3.4. We begin by bounding the Frobenius norm of $\hat{T}_4$. Let $T_4 = \mathbb{E}[h_4(\mathcal{X})]$. It then follows from Lemma 4.2.6 that

$$T_4 = \mathrm{Sym}\left(\sum_{i=1}^{k} w_i\left(3S_i \otimes S_i + 6S_i \otimes \mu_i^{\otimes 2} + \mu_i^{\otimes 4}\right)\right) . \tag{4.3.20}$$

Further, $\|S_i \otimes S_i\|_F \leqslant \|S_i\|_F^2 \leqslant \Delta^2$, $\left\|S_i \otimes \mu_i^{\otimes 2}\right\|_F \leqslant \|S_i\|_F \|\mu_i\|_2^2 \leqslant 4\Delta/\alpha$, and $\left\|\mu_i^{\otimes 4}\right\|_F \leqslant \|\mu_i\|_2^4 \leqslant 16/\alpha^2$. Since T_4 is an average of terms of the form $S_i^{\otimes 2}, S_i \otimes \mu_i^{\otimes 2}$ and $\mu_i^{\otimes 4}$, and each such term is upper bounded, we can conclude that $\|T_4\|_F = \mathcal{O}\left(\Delta^2 + 1/\alpha^2\right)$, and by the triangle inequality that $\left\|\hat{T}_4\right\|_F \leqslant \mathcal{O}\left(\Delta^2 + 1/\alpha^2 + \eta\right)$. Let $S_i' = S_i + \mu_i^{\otimes 2}$ and let $T_4' := \sum_{i=1}^k w_i\left(S_i' \otimes S_i'\right)$. We can then rewrite Equation (4.3.20) as follows:

$$T_4 = \mathrm{Sym}\left(\sum_{i=1}^{k} w_i\left(3S_i' \otimes S_i' - 2\mu_i^{\otimes 4}\right)\right) . \tag{4.3.21}$$

For $j \in [4k]$, let $x^{(j)}, y^{(j)} \sim \mathcal{N}(0, I)$. Collapsing two modes of $\hat{T}_4$, it follows from Equation

(4.3.21) that for any fixed j,

$$\begin{aligned}\hat{T}_4\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) &= \left(\hat{T}_4 - T_4\right)\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) + T_4\left(\cdot,\cdot,x^{(j)},y^{(j)}\right)\\ &= \left(\hat{T}_4 - T_4\right)\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) + \mathrm{Sym}\left(\sum_{i=1}^{k} w_i\left(3S'_i\otimes S'_i - 2\mu_i^{\otimes 4}\right)\right)\left(\cdot,\cdot,x^{(j)},y^{(j)}\right)\\ &= \left(\hat{T}_4 - T_4 + T'_4\right)\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) + \sum_{i\in[k]} w_i\left(S'_i x^{(j)}\right)\otimes\left(S'_i y^{(j)}\right)\\ &\quad + \sum_{i\in[k]} w_i\left(S'_i y^{(j)}\right)\otimes\left(S'_i x^{(j)}\right) + \sum_{i\in[k]} w_i\left(-2\mu_i^{\otimes 2}\left\langle\mu_i, x^{(j)}\right\rangle\left\langle\mu_i, y^{(j)}\right\rangle\right),\end{aligned} \tag{4.3.22}$$

where we use that $\mathrm{Sym}(\cdot)$ is a linear operator satisfying $\mathrm{Sym}\left(\mu_i^{\otimes 4}\right) = \mu_i^{\otimes 4}$, and

$$\mathrm{Sym}\left(S'_i\otimes S'_i\right) = \frac{1}{3}S'_i\otimes S'_i + \frac{1}{3}S'_i\oplus S'_i + \frac{1}{3}S'_i\ominus S'_i$$

where for indices (i_1,i_2,i_3,i_4), $\left(S'_i\oplus S'_i\right)(i_1,i_2,i_3,i_4) = \left(S'_i\otimes S'_i\right)(i_1,i_3,i_2,i_4)$ and $\left(S'_i\ominus S'_i\right)(i_1,i_2,i_3,i_4) = \left(S'_i\otimes S'_i\right)(i_1,i_4,i_2,i_3)$.

Next, it follows from Lemma 4.3.6 that there exists a matrix $\tilde{P}_i$ such that $\left\|\tilde{P}_i\right\|_F = \mathcal{O}\left(1/(\sqrt{\eta}\alpha)\right)$ and $\left\|T'_4\left(\cdot,\cdot,\tilde{P}_i\right) - S'_i\right\|_F = \mathcal{O}\left(\sqrt{\eta/\alpha}\right)$. Furthermore, with probability at least 0.99, there exists a sequence of $b_j\in[-D,D]$, for $j\in[4k]$, such that $T'_4\left(\cdot,\cdot,\sum_{j\in[4k]} b_j x^{(j)}\otimes y^{(j)}\right) = T'_4\left(\cdot,\cdot,\tilde{P}_i\right)$.

Consider a cover, $\mathcal{C}$, of the interval $[-D,D]$ with points spaced at intervals of length $\tau = \mathcal{O}\left(\frac{\sqrt{\eta}}{\alpha k\left(\Delta^4+1/\alpha^4\right)}\right)$. Since we uniformly sample a_j's, with probability at least $(\tau/D)^{\mathcal{O}(k)}$, for all $j\in[4k]$, $\left|b_j - a_j\right|\leqslant\tau$, and we condition on this event. Thus,

$$\begin{aligned}\left\|T'_4\left(\cdot,\cdot,\sum_{j\in[4k]} a_j x^{(j)} y^{(j)}\right) - S'_i\right\|_F &\leqslant \left\|T'_4\left(\cdot,\cdot,\sum_{j\in[4k]} b_j x^{(j)}\otimes y^{(j)}\right) - S'_i\right\|_F + \left\|T'_4\left(\cdot,\cdot,\sum_{j\in[4k]}(b_j - a_j)x^{(j)}\otimes y^{(j)}\right.\\ &\leqslant \mathcal{O}\left(\sqrt{\eta/\alpha}\right) + \mathcal{O}(\tau\Delta^2) \leqslant \mathcal{O}\left(\sqrt{\eta/\alpha}\right).\end{aligned} \tag{4.3.23}$$

Taking the linear combinations with coefficients a_j in Equation (4.3.22), we have

$$\begin{aligned}\hat{T}_4\left(\cdot,\cdot,\sum_{j\in[4k]} a_j x^{(j)}\otimes y^{(j)}\right) - S_i = &\left(\hat{T}_4 - T_4 + T_4'\right)\left(\cdot,\cdot,\sum_{j\in[4k]} a_j x^{(j)}\otimes y^{(j)}\right) - S_i' - \mu_i\otimes\mu_i \\ &+ \sum_{j\in[4k]} a_j \sum_{i\in[k]} w_i \left(S_i' x^{(j)}\right)\otimes\left(S_i' y^{(j)}\right) + \sum_{j\in[4k]} a_j \sum_{i\in[k]} w_i \left(S_i' y^{(j)}\right)\otimes \big(S_i' x \\ &+ \sum_{j\in[4k]} a_j \sum_{i\in[k]} w_i \left(-2\mu_i^{\otimes 2}\left\langle \mu_i, x^{(j)}\right\rangle\left\langle \mu_i, y^{(j)}\right\rangle\right)\ .\end{aligned} \tag{4.3.24}$$

Setting $P_i = \left(\hat{T}_4 - T_4 + T_4'\right)\left(\cdot,\cdot,\sum_{j\in[4k]} a_j x^{(j)} y^{(j)}\right) - S_i'$, it follows from Lemma 4.2.9 that with probability at least 0.99, $\left(\hat{T}_4 - T_4\right)\left(\cdot,\cdot,\sum_{j\in[4k]} a_j x^{(j)} y^{(j)}\right)$ has Frobenius norm $O(kD\eta)$ and it follows from Equation (4.3.23) that with probability at least 0.99, $T_4'\left(\cdot,\cdot,\sum_{j\in[4k]} a_j x^{(j)} y^{(j)}\right) - S_i'$ has Frobenius norm $O\left(\sqrt{\eta/\alpha}\right)$. Setting the remaining terms to Q_i, with probability at least 0.99 we can bound their Frobenius norm as follows:

$$\begin{aligned}\|Q_i\|_F &\leqslant \|\mu_i\otimes\mu_i\|_F + \left\|\left(\sum_{i\in[k]} w_i S_i'\oplus S_i' + w_i S_i'\ominus S_i' - 2w_i\mu_i^{\otimes 4}\right)\left(\cdot,\cdot,\sum_{j\in[4k]} a_j x^{(j)} y^{(j)}\right)\right\|_F \\ &\leqslant \frac{4}{\alpha} + \left(2\max_{i\in k}\left\|S_i'\right\|_F^2 + \frac{32}{\alpha^2} + k\tau\right)\cdot\left\|\tilde{P}\right\|_F \\ &\leqslant \frac{4}{\alpha} + O\left(\frac{1}{\sqrt{\eta}\alpha}\left(\Delta + \frac{1}{\alpha}\right)^2\right) \\ &\leqslant O\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right)\ ,\end{aligned} \tag{4.3.25}$$

where the first inequality follows from the triangle inequality, the second follows from our assumptions that $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, $\sum_{j\in[4k]} b_j x^{(j)} y^{(j)} = \tilde{P}_i$ in the span of the S_i', and $\left|a_j - b_j\right| \leqslant \tau$ for all $j\in[4k]$, and the third inequality follows from the definition of S_i', the bound on $\left\|\tilde{P}\right\|_F$ and the bound on $\|S_i - I\|_F$. □

4.3.5 Finding a Low-dimensional Subspace for Exhaustive Search

In this subsection, we will prove Proposition 4.3.5.

We start by extending Theorem 4 of [3], which shows that large parameter distance between pairs of univariate Gaussian mixtures implies large distance between their low-degree moments. In the following, we use $M_j(F) = \mathbb{E}\left[F^j\right]$ to denote the j-th moment of a distribution F. We show:

Lemma 4.3.9. *There exists a constant $C > 0$ such that the following holds: Fix any $D > 0$ and $0 \leqslant \beta \leqslant 1/(2(2k-1)!D^{2k-3})$. Suppose that $F = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma_i^2)$ is a univariate k-mixture of Gaussians with $w_i \geqslant \beta$, and $|\mu_i|, \sigma_i \leqslant D$, for all $i \in [k]$. If $|\mu_i| + |\sigma_i^2 - 1| \geqslant \beta$ for some $i \leqslant k$, then*

$$\max_{j \in [2k]} \left| M_j(F) - M_j\left(\mathcal{N}(0,1)\right) \right| \geqslant \beta^{C^{k+1}(k+1)!-1} .$$

We give the proof of Lemma 4.3.9 in Section 4.3.6.

Lemma 4.3.10 (Bounding μ_i's and S_i's in non-influential directions for $\mathbb{E}\left[h_m(\mathcal{M})\right]$)**.** *Let $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians on $\mathbb{R}^d$ satisfying $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for every $i \in [k]$. For some $B \in \mathbb{R}$, let $u \in \mathbb{R}^d$ be a unit vector such that $|\mathbb{E}\left[h_m\left(\langle \mathcal{M}, u\rangle\right)\right]| \leqslant B$ for all $m \in [2k]$. Then, for $\delta = 2^{O(k)} B^{1/(C^{k+1}(k+1)!)}$ and $S_i = \Sigma_i - I$, we have that:*

1. *for all $i \leqslant k$, $|\langle u, \mu_i\rangle|, |u^\top (I - \Sigma_i) u| \leqslant \delta$,*
2. *$\|S_i u\|_2^2 \leqslant 20\delta\Delta/\alpha^2 + B/\alpha$,*

where $C > 0$ is a fixed universal constant.

Proof. The 1-D random variable $\langle u, \mathcal{M}\rangle$ is a mixture of Gaussians described by $\sum_{i=1}^{k} w_i \mathcal{N}(\langle \mu_i, u\rangle, u^\top \Sigma_i u)$. Towards a contradiction, assume that there is an $i \in [k]$ such that $|\langle u, \mu_i\rangle| + |u^\top (I - \Sigma_i) u| \geqslant \delta$. Then, applying Lemma 4.3.9, yields that there is a $j \in [2k]$ such that $|M_j(\langle u, \mathcal{M}\rangle) - M_j(\mathcal{N}(0,1))| \geqslant \delta^{C^{k+1}(k+1)!-1}$. Applying Fact 4.2.5 implies that there exists an $m \in [2k]$ such that

$$\left|\mathbb{E}\left[h_m(\langle u, \mathcal{M}\rangle)\right]\right| > 2^{-O(k)} \delta^{C^{k+1}(k+1)!-1} \gg B \ ,$$

yielding a contradiction.

We can now prove the second part. Recall that for $S_i = \Sigma_i - I$ for every i, we have that

$$\mathbb{E}\left[h_4(\mathcal{M})\right] = \sum_{i=1}^{k} w_i \mathrm{Sym}\left(3\left(S_i \otimes S_i\right) + 6\left(S_i \otimes \mu_i^{\otimes 2}\right) + \mu_i^{\otimes 4}\right) \ .$$

We consider the $d \times d$ matrix obtained by the natural flattening of the $d \times d$ tensor $u^{\otimes 2} \cdot \mathbb{E}\left[h_4(\mathcal{M})\right]$. Then, we can write:

$$\begin{aligned} u^{\otimes 2} \cdot \mathbb{E}\left[h_4(\mathcal{M})\right] = \sum_{i=1}^{k} w_i \Big(&(u^\top S_i u) S_i + 2(S_i u)(S_i u)^\top + \langle u, \mu_i\rangle^2 S_i \\ &+ 2\langle u, \mu_i\rangle \mu_i (S_i u)^\top + 2\langle u, \mu_i\rangle (S_i u)\mu_i^\top + (u^\top S_i u)\mu_i \mu_i^\top + \langle u, \mu_i\rangle^2 \mu_i \mu_i^\top \Big) \ . \end{aligned} \tag{4.3.26}$$

Now, from the first part, we know that for all $i \in [k]$, $|u^\top S_i u| \leqslant \delta$ and the hypothesis of the lemma gives us that $\|S_i\|_F = \|\Sigma_i - I\|_F \leqslant \Delta$. Thus, for each i, the first term in the summation above has Frobenius norm at most $\Delta\delta$. Using that $\langle u, \mu_i\rangle_2^2 \leqslant \delta^2$ from the first part of the lemma, yields that, for each i, the Frobenius norm of the third term is at most $\Delta\delta^2$.

Next, using in addition that $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$ yields that, for each i, the Frobenius norm of the 4th and 5th terms are at most $2\delta\Delta/\sqrt{\alpha}$ and the Frobenius norm of the 6th and 7th terms are at most δ/α. Thus, for each i and all but the 2nd term in the summation above, we have an upper bound on the Frobenius norm of $4\delta\Delta/\alpha$.

Now, since $|\mathbb{E}\left[h_4\left(\langle \mathcal{M}, u\rangle\right)\right]| \leqslant B$, and u is a unit vector, we have that $\left\|u^{\otimes 2}\, \mathbb{E}\left[h_4(\mathcal{M})\right]\right\|_F \leqslant B$. Thus, combining the aforementioned argument with the triangle inequality, we have for each i,

$$
\begin{aligned}
\|S_i u\|_2^2 = \left\|S_i u\,(S_i u)^\top\right\|_F &\leqslant \frac{1}{\alpha}\left\|u^{\otimes 2}\cdot \mathbb{E}\left[h_4(\mathcal{M})\right]\right\|_F + \sum_{i\in[k]} w_i\Big(\Big(u^\top S_i u + \langle u, \mu_i\rangle^2\Big)\|S_i\|_F\Big)\\
&\quad + \sum_{i\in[k]} 4w_i\Big((\langle u, \mu_i\rangle)\,\|\mu_i\|_2\,\|S_i u\|_2\Big) + \sum_{i\in[k]} 4w_i\Big(\Big(\langle u, \mu_i\rangle^2 + u^\top S_i u\Big)\|\mu_i\|_2^2\Big)\\
&\leqslant B/\alpha + 15\delta\Delta/\alpha\ ,
\end{aligned}
$$

and the claim follows. □

Lemma 4.3.11 (Subspace covering all the means and large singular vectors of $S_i = \Sigma_i - I$)**.** *Let $\mathcal{M} = \sum_{i=1}^k w_i\mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians on $\mathbb{R}^d$ satisfying $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for all $i \in [k]$. Given $0 < \eta < 1$, let $\hat{T}_m$ satisfy $\left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$ for every $m \in [4k]$ and let $\lambda \geqslant 2\eta$. Let V be the span of all the left singular vectors of the $d \times d^{m-1}$ matrix obtained by the natural flattening of $\hat{T}_m$ with singular values at least λ. Then, for $\delta = 2\lambda^{1/(2C^{k+1}(k+1)!)}$, we have that:*

1. $\dim(V) \leqslant \left(4k\eta^2 + k^{O(k)}\right)O(1 + 1/\alpha + \Delta)^{4k}/\lambda^2$,
2. *Let*

$$V_{\text{inf}} = \{\mu_i\}_{i\in[k]} \cup \left\{v \mid \exists i \in [k],\ \text{s.t. } \|v\|_2 = 1 \text{ and } v \text{ is an eigenvector of } S_i \text{ and } \|S_i v\|_2 \geqslant \sqrt{\delta}\right\}_{i\leqslant k}$$

Then, for every unit vector $v \in V_{\text{inf}}$, $\|v - \Pi_V v\|_2^2 \leqslant 20\delta^{1/4}\Delta/\alpha^2$, where $\Pi_V v$ is the projection of v onto V.

Proof. From Fact 4.2.6, we have that $\mathbb{E}\left[h_m(\mathcal{M})\right] = \sum_{i\in[k]} w_i\,\mathbb{E}\left[h_m(G_i)\right]$, where $G_i = \mathcal{N}(\mu_i, \Sigma_i)$, and since $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$ and $\|\Sigma_i - I\|_F \leqslant \Delta$, it follows that $\|\mathbb{E}\left[h_m(\mathcal{M})\right]\|_F^2 \leqslant O\left(m(1 + 1/\alpha + \Delta)\right)^{4m}$. From Proposition 4.3.3, we know that

$$\left\|\hat{T}_m\right\|_F^2 \leqslant 2\left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F^2 + 2\left\|\mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F^2 \leqslant \eta^2 + O\left(m(1 + 1/\alpha + \Delta)\right)^{4m}\ .$$

Thus, the number of singular vectors of the $d \times d^{m-1}$ flattening of $\hat{T}_m$ with a singular value $\geqslant \lambda$ is at most $(\eta^2 + O(m(1 + 1/\alpha + \Delta))^{4m})/\lambda^2$. Summing up this bound for all $m \in [4k]$, yields the claimed upper bound on $\dim(V)$.

For the second part, we will first bound $\langle u, v\rangle$ for any unit vector u orthogonal to the subspace V. Towards this, observe that since u is orthogonal to V and $\|u\|_2 = 1$, we have

$$\left\| u \cdot \mathbb{E}\left[h_m(\mathcal{M})\right] \right\|_F \leqslant \left\| u\hat{T}_m \right\|_F + \left\| \hat{T}_m - \mathbb{E}\, h_m(\mathcal{M}) \right\|_F \leqslant \lambda + \eta \leqslant 2\lambda ,$$

where $u \cdot \mathbb{E}\left[h_m(\mathcal{M})\right]$ is a matrix-vector product of u with a $d \times d^{m-1}$ flattening of $\mathbb{E}\left[h_m(\mathcal{M})\right]$. For $\delta = 2\lambda^{1/(C^{k+1}(k+1)!)}$, applying Lemma 4.3.10 yields that

$$\langle \mu_i, u\rangle^2 + \|S_i u\|_2^2 \leqslant \delta^2 + 20\delta\Delta/\alpha \leqslant 20\delta\Delta/\alpha^2 . \tag{4.3.27}$$

Now, if v is one of the μ_i's, then we immediately get from Equation 4.3.27 that $\langle v, u\rangle^2 \leqslant 20\delta\Delta/\alpha^2$. Similarly, note that if v is a unit length eigenvector of S_i satisfying $\|S_i v\|_2^2 \geqslant \sqrt{\delta}$, then,

$$\langle u, v\rangle^2 = \frac{1}{\|S_i v\|_2^2} \langle u, S_i v\rangle^2 = \frac{1}{\|S_i v\|_2^2} \langle S_i u, v\rangle^2 \leqslant \frac{\|S_i u\|_2^2}{\|S_i v\|_2^2} .$$

In both cases, setting $u = (v - \Pi_V v)/\|v - \Pi_V v\|_2$ completes the proof. □

We can now complete the proof of Proposition 4.3.5:

Proof of Proposition 4.3.5. We know that $\hat{S}_i - P_i - S_i$ is a symmetric, rank-k' matrix such that $k' = O(k^2)$, described by the eigenvalue decomposition $\sum_{i=1}^{k'} \tau_i v_i v_i^\top$, where v_i's are the eigenvectors and τ_i's are the corresponding eigenvalues. Since $\|S_i\|_F \leqslant \Delta$ and

$$\left\|\hat{S}_i\right\|_F \leqslant \|P_i\|_F + \|Q_i\|_F + \|S_i\|_F \leqslant O\left(\sqrt{\eta/\alpha}\right) + O\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right) + \Delta = O\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right) ,$$

we have that the number of singular values of $\hat{S}_i$ that exceed $\delta^{1/4}$ is at most $O\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3\sqrt{\delta}}\right)$. Recall that from Lemma 4.3.11 it follows that the dimension of the subspace V is at most

$k^{O(k)}O(1+1/\alpha+\Delta)^{4k}/\lambda^2$. Thus, the dimension of V' is at most

$$k^{O(k)}O\left(\frac{(1+1/\alpha+\Delta)^{4k}}{\lambda^2}\right)+O\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3\sqrt{\delta}}\right)=O\left(\frac{k^{O(k)}\left(1+\frac{1}{\alpha}+\Delta\right)^{4k+5}}{\eta^2}\right) .$$

Since V' contains V constructed in Lemma 4.3.11, we immediately obtain that for every μ_i, $\|\mu_i-\Pi_{V'}\mu_i\|_2^2 \leqslant \frac{20}{\alpha^2}\sqrt{\delta}\Delta$.

Next, let u be a unit vector orthogonal to V'. Then, since V' contains the V described in Lemma 4.3.11, we know that $\|S_i u\|_2^2 \leqslant \frac{20}{\alpha^2}\sqrt{\delta}\Delta$. Similarly, since V' contains all eigenvectors of $\hat{S}_i$ with singular values exceeding $\delta^{1/4}$, we know that $\left\|\hat{S}_i u\right\|_2^2 \leqslant \delta^{1/2}$. Thus, we can conclude that $\left\|(\hat{S}_i-S_i)u\right\|_2^2 \leqslant \frac{100}{\alpha^2}\sqrt{\delta}\Delta$. Let $Q_i=\sum_{j=1}^{k'}\tau_j v_j v_j^\top$ with orthonormal $v_j \in \mathbb{R}^d$. We know such τ_j's and v_j's exist because of the upper bound on $\mathrm{rank}(Q_i)$. Therefore, for any j, $|v_j^\top(\hat{S}_i-S_i)u| \leqslant \frac{10}{\alpha}\delta^{1/4}\Delta^{1/2}$. On the other hand, for any j, we have that

$$v_j^\top(\hat{S}_i-S_i)u \geqslant \langle v_j,u\rangle\tau_j-\|P_i\|_F=\langle v_j,u\rangle\tau_j-O(\sqrt{\eta}) .$$

Combining the two bounds above, yields that whenever $\tau_j \geqslant \delta^{1/4}$,

$$|\langle v_j,u\rangle| \leqslant O(\sqrt{\eta}/\tau_j)+\frac{10}{\alpha\tau_j}\delta^{1/4}\Delta^{1/2} \leqslant \frac{10}{\alpha}\delta^{1/2}\Delta^{1/2} .$$

Thus, the matrix $\hat{Q}_i=\sum_{j=1}^{k'}\tau_j\Pi_{V'}v_j(\Pi_{V'}v_j)^\top$ has its range space in V' and satisfies

$$\left\|\hat{Q}_i-Q_i\right\|_F \leqslant O\left(k^2\delta^{1/4}\right)+O\left(\frac{k^2}{\alpha}\delta^{1/2}\Delta^{1/2}\right)=O\left(\frac{k^2}{\alpha}\delta^{1/2}\Delta^{1/2}\right) .$$

□

4.3.6 Parameter vs Moment Distance for Gaussian Mixtures

In this subsection, we prove Lemma 4.3.9. To that end, we will use the following two results; the second one is from [3].

Lemma 4.3.12. *Suppose $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ are univariate Gaussians with $|\mu_i|, |\sigma_i| \leqslant D$, for some $D \in \mathbb{R}_+$. If $|\mu_1 - \mu_2| + |\sigma_1^2 - \sigma_2^2| \leqslant \beta$, then the distance between raw moments of two Gaussians is*

$$\left|M_j(\mathcal{N}(\mu_1, \sigma_1^2)) - M_j(\mathcal{N}(\mu_2, \sigma_2^2))\right| \leqslant (j+1)! D^{j-1} \beta.$$

Proof. By Proposition 4.2.3, the j-th raw moment of a Gaussian $\mathcal{N}(\mu, \sigma^2)$ is a sum of monomials in μ and σ^2 of degree j. There are at most $(j+1)!$ terms in the polynomial. Thus, changing the mean or the variance by at most β will change the j-th moment by at most $(j+1)! D^{j-1} \beta$. □

Theorem 4.3.13. *([3]) Let F, F' be two univariate mixtures of Gaussians: $F = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma_i^2)$ and $F' = \sum_{i=1}^{k'} w_i' \mathcal{N}(\mu_i', \sigma_i'^2)$. There is a constant $c > 0$ such that, for any $\beta < c$, if F, F' satisfy:*

1. *$w_i, w_i' \in [\beta, 1]$*
2. *$|\mu_i|, |\mu_i'| \leqslant 1/\beta$*
3. *$|\mu_i - \mu_{i'}| + |\sigma_i^2 - \sigma_{i'}^2| \geqslant \beta$ and $|\mu_i' - \mu_{i'}'| + |\sigma_i'^2 - \sigma_{i'}'^2| \geqslant \beta$ for all $i \neq i'$*
4. *$\beta \leqslant \min_\pi \sum_i \left(|w_i - w'_{\pi(i)}| + |\mu_i - \mu'_{\pi(i)}| + |\sigma^2 - {\sigma'_{\pi(i)}}^2|\right)$, where the minimization is taken over all mappings $\pi : \{1, \ldots, k\} \to \{1, \ldots, k'\}$,*

then

$$\max_{j \in [2(k+k'-1)]} |M_j(F) - M_j(F')| \geqslant \beta^{O(k)}.$$

We are now ready to complete the proof of Lemma 4.3.9.

Proof of Lemma 4.3.9. We proceed via induction on k. Consider the base case, i.e., $k = 1$. Then, either $|\mu_1| \geqslant \beta/2$ or $|\sigma_1 - 1| \geqslant \beta/2$, and thus the first or second moment differ by at least $\beta^2/4$. Let the inductive hypothesis be that Lemma 4.3.9 holds for at most k components.

Consider the case where $|\mu_i - \mu_{i'}| + |\sigma_i^2 - \sigma_{i'}^2| \geqslant \beta^{C^k k!}$ for all pairs of components $i, i' \in [k]$. Then, by Theorem 4.3.13, we have that

$$\max_{j\in[2k]} |M_j(F) - M_j(\mathcal{N}(0,1))| \geqslant \beta^{C^{k+1}k!} \geqslant \beta^{C^{k+1}(k+1)!-1} ,$$

and the lemma follows.

Otherwise, we know that there exists a pair of components with parameter distance less than $\beta^{C^k k!}$. In this case, we merge these two components and get a $(k-1)$-mixture F'. By Lemma 4.3.12, the distance between the j-th moments of F' and F is at most $(j+1)!D^{j-1}\beta^{C^k k!}$. Since we still have $|\mu_i'| + |\sigma_i'^2 - 1| \geqslant \beta - 3\beta^{C^k k!}$ for all components i in F', the inductive hypothesis implies that

$$\max_{j\in[2k-2]} |M_j(F') - M_j(\mathcal{N}(0,1))| \geqslant \left(\beta - 3\beta^{C^k k!}\right)^{C^k(k)!-1} .$$

By the triangle inequality, we can write

$$\begin{aligned}
\max_{j\in[2k]} |M_j(F) - M_j(\mathcal{N}(0,1))| &\geqslant \max_{j\in[2k-2]} |M_j(F') - M_j(\mathcal{N}(0,1))| - \max_{j\in[2k-2]} |M_j(F) - M_j(F')| \\
&\geqslant \left(\beta - 3\beta^{C^k k!}\right)^{C^k k!-1} - (2k-1)!D^{2k-3}\beta^{C^k k!} \\
&\geqslant \beta^{C^{k+1}(k+1)!-1} .
\end{aligned}$$

The last inequality follows from the assumption that $\beta \leqslant 1/(2(2k-1)!D^{2k-3})$. This completes the proof of Lemma 4.3.9. □

4.3.7 Proof of Theorem 4.3.1

We use the following notation and background from Moitra-Valiant [3]:

Definition 4.3.14 (Statistically Learnable)**.** Given $\varepsilon > 0$, we call a mixture of Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ ε-statistically learnable if $\min_i w_i \geqslant \varepsilon$ and

$\min_{i \neq j} d_{TV}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j)) \geqslant \varepsilon$.

Definition 4.3.15 (Correct Subdivision)**.** Given a Gaussian mixture of k Gaussians, $\mathcal{M} = \sum_i w_i N(\mu_i, \Sigma_i)$ and a mixture of $k' \leqslant k$ Gaussians $\hat{\mathcal{M}} = \sum_i \hat{w}_i N(\hat{\mu}_i, \hat{\Sigma}_i)$, we call $\hat{\mathcal{M}}$ an ε-correct subdivision of $\mathcal{M}$ if there is a function $\pi : [k] \to [k']$ that is onto and

1. $\forall j \in [k'], \left|\sum_{i:\pi(i)=j} w_i - \hat{w}_j\right| \leqslant \varepsilon$
2. $\forall i \in [k], \left\|\mu_i - \hat{\mu}_{\pi(i)}\right\| + \left\|\Sigma_i - \hat{\Sigma}_{\pi(i)}\right\|_F \leqslant \varepsilon$.

Theorem 4.3.16 (Theorem 8 in [3])**.** *Given an ε-statistically learnable Gaussian mixture $\mathcal{M}$ in isotropic position, for some $\varepsilon > 0$, there exists an algorithm that requires $n = \mathrm{poly}(d/\varepsilon)$ samples and runs in time $O(\mathrm{poly}_k(n))$ and with probability at least $99/100$ recovers an ε-correct sub-division $\hat{\mathcal{M}}$. Let the corresponding algorithm be referred to as PARTITION PURSUIT.*

The algorithm has two steps: first run the first three steps of Algorithm 3.2 to get the list L' of $\hat{S}$ and $V'_{\hat{S}}$; then apply the following proposition to learn the mixture in the subspace $V'_{\hat{S}}$. This proposition is a generalization of Theorem 4.3.16 without the assumption that the total variation distance between each pair of components is at least ε. The sample and time complexities has a worse, but still polynomial dependence on ε. Note that although the algorithm in the proposition is non-robust, we can take a sample without noise with constant probability because the algorithm only requires a polynomial number of samples in ε.

Algorithm 4.3.17 (Efficient List-Recovery of Candidate Parameters)**.**

Input: *An ε-corruption Y of a sample X from a k-mixture of Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$. Let Z be an additional ε-corrupted sample of size n' from $\mathcal{M}$.*

Requirements: *The guarantees of the algorithm hold if the mixture parameters and* *the sample X satisfy:*

1. $w_i \geqslant \alpha$ *for all* $i \in [k]$,
2. $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$ *for all* $i \in [k]$,
3. $\|\Sigma_i - I\|_F \leqslant \Delta$ *for all* $i \in [k]$.
4. X *is a good sample (satisfies Condition 4.2.29) with parameters* (γ, t), *where* $\gamma = \varepsilon d^{-8k} k^{-Ck}$, *for* C *a sufficiently large universal constant, and* $t = 8k$.
5. *The number of fresh samples* $n' = O\left(\varepsilon\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4ck}$, *for a fixed constant* c.

Parameters: $\eta = (2k)^{4k}O(1/\alpha + \Delta)^{4k}\, \varepsilon^{1/(k^{O(k^2)})}$, $D = C(k^4/(\alpha\sqrt{\eta}))$, $\delta = 2\eta^{1/(C^{k+1}(k+1)!)}$, $\ell' = 100 \log k \left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4k}$, *for some sufficiently large absolute constant* $C > 0$, $\lambda = 4\eta$, $\phi = 10(1 + \Delta^2)/(\sqrt{\eta}\alpha^5)$, $\varepsilon_1 = O\left(\sqrt{\Delta}\delta^{1/4}/\alpha\right)$.

Output: *A list* L *of hypotheses such that there exists at least one,* $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i \leqslant k} \in L$, *satisfying:* $\|\mu_i - \hat{\mu}_i\|_2 \leqslant O\left(\frac{\Delta^{1/2}}{\alpha}\right)\eta^{G(k)}$ *and* $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant O\left(k^4\right)\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)}$, *where* $G(k) = \frac{1}{C^{k+1}(k+1)!}$.

Operation:

1. ***Robust Estimation of Hermite Tensors***: *For* $m \in [4k]$, *compute* $\hat{T}_m$ *such that* $\max_{m \in [4k]} \left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$ *using the robust mean estimation algorithm in Fact 2.2.1.*
2. ***Random Collapsing of Two Modes of*** $\hat{T}_4$: *Let* L' *be an empty list. Repeat* ℓ' *times: For* $j \in [4k]$, *choose independent standard Gaussians in* $\mathbb{R}^d$, *denoted by* $x^{(j)}, y^{(j)} \sim \mathcal{N}(0, I)$, *and uniform draws* $a_1, a_2, \ldots, a_t$ *from* $[-D, D]$. *Let* $\hat{S}$ *be a* $d \times d$ *matrix*

such that for all $r, s \in [d]$, $\hat{S}(r,s) = \sum_{j\in[4k]} a_j \hat{T}_4(r,s,x^{(j)},y^{(j)}) = \sum_{j\in[4k]} a_j \sum_{g,h\in[d]} \hat{T}_4(r,s,g,h)x^{(j)}(g)y^{(j)}(h)$. *Add* $\hat{S}$ *to the list* L'.

3. ***Construct Low-Dimensional Subspace****: Let* V *be the span of all singular vectors of the natural* $d \times d^{m-1}$ *flattening of* $\hat{T}_m$ *with singular values* $\geqslant \lambda$ *for* $m \leqslant 4k$. *For each* $\hat{S} \in L'$, *let* $V'_{\hat{S}}$ *be the span of* V *plus all the singular vectors of* $\hat{S}$ *with singular value larger than* $\delta^{1/4}$.

4. ***Moitra-Valiant for Low-Dimensional Subspace****: Initialize* L *to be the empty list. For each* $\hat{S} \in L'$, *let* $\hat{P} = UU^\top$ *be the orthogonal projection matrix onto the span of* $V'_{\hat{S}}$, *where* $U \in \mathbb{R}^{d\times d}$ *has orthonormal columns. Let* $m = \dim(V'_{\hat{S}})$ *and let* $\hat{Z} \subset Z$ *be a randomly chosen subset of size* $\mathrm{poly}(m/\varepsilon_1)$. *Let* U_m *denote the first* m *columns of* U *and for all* $z \in \hat{Z}$, *compute* $U_m^\top z$. *Run PARTITION PURSUIT on the resulting set of points and let* $\{\hat{\mu}_i^{\hat{P}}, \hat{\Sigma}_i^{\hat{P}}\}_{i\in[k]}$ *be the parameters corresponding to the* ε*-correct subdivision output by PARTITION PURSUIT. Let* $\hat{\mu}_i^\top = [(\hat{\mu}_i^{\hat{P}})^\top, 0]$ *be a* d *dimensional vector padded with* 0*s and* $\hat{\Sigma}_i$ *be a* $d \times d$ *matrix with* $\hat{\Sigma}_i^{\hat{P}}$ *in the top left* $m \times m$ *sub-matrix and* 0*'s elsewhere. Add* $\{U\hat{\mu}_i, U\hat{\Sigma}_i U^\top + (\hat{S} + I) - \hat{P}(I + S)\hat{P}\}_{i\in[k]}$ *to* L.

Proposition 4.3.18. *Given* $\varepsilon > 0$ *and a sample* X *of size* $\mathrm{poly}(d, 1/\varepsilon)$ *from a* k*-mixture of Gaussians* $\mathcal{M}$ *with mixture covariance* Σ *such that* $0.99I \preceq \Sigma \preceq 1.01I$ *and satisfies* $w_i \geqslant \varepsilon$, *the PARTITION PURSUIT algorithm runs in time* $\mathrm{poly}(d, 1/\varepsilon)$ *and with probability at least* $9/10$ *returns an* $O(\varepsilon)$*-correct sub-division, denoted by* $\hat{\mathcal{M}}$.

Recall, the PARTITION PURSUIT algorithm satisfies Theorem 4.3.16 and we will prove that with an appropriately chosen parameter ε, PARTITION PURSUIT also satisfies Proposition 4.3.18. The main idea is that if any two components are actually close enough in total variation distance, then any algorithm with access to only a polynomial number of samples could never distinguish these two components from a single Gaussian. So if all pairwise distances are either sufficiently large or sufficiently small, the algorithm will behave as if it were given sample access to a mixture that meets the requirements of Theorem 4.3.16.

Lemma 4.3.19. *Given* $0 < \gamma, \delta < 1$ *and two distributions,* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *over* $\mathbb{R}^d$ *such that* $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) < \gamma$*, let* X_1 *be set of* n *i.i.d. samples from* $\mathcal{D}_1$ *and* X_2 *be* n *i.i.d. samples from* $\mathcal{D}_2$*. Let* $\mathcal{A}$ *be any algorithm that takes as input* X_1 *and outputs a list of* m *real numbers,* $Y_1 = \{y_i\}_{i\in[m]}$*, such that* $y_i \in [-1, 1]$ *with probability at least* $1 - \delta$*. Then, for any* $\tau > 0$*,* $\mathcal{A}$ *on input* X_2 *outputs a list of* m *real numbers* $Y_2 = \{y'_i\}_{i\in[m]}$ *such that with probability at least* $1 - \delta - (4mn\gamma/\tau)$*, for all* $i \in [m]$*,* $|y_i - y'_i| \leqslant \tau$*.*

Proof. Let $\mathcal{U}_1$ be the uniform distribution over X_1 and $\mathcal{U}_2$ be the uniform distribution over X_2. Then,

$$\begin{aligned} d_{TV}(\mathcal{U}_1, \mathcal{U}_2) \leqslant \sqrt{2}H^2(\mathcal{U}_1, \mathcal{U}_2) &= \sqrt{2}nH^2(\mathcal{D}_1, \mathcal{D}_2) \\ &\leqslant \sqrt{2}n d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \\ &\leqslant \sqrt{2}\gamma n \end{aligned} \tag{4.3.28}$$

Consider the family of functions $\mathcal{F}$ that take as input n samples and output a single bit in $\{0, 1\}$. We know that for any function $f \in \mathcal{F}$, the probability that $f(X_1) \neq f(X_2)$ is at most $\sqrt{2}\gamma n$. Recall, the algorithm outputs m real numbers in the range $[-1, 1]$, which we can discretize into a grid Δ of length τ. There are at most $2/\tau$ distinct grid points and for any $y_i \in [-1, 1]$, there exists a point $z_i \in \Delta$ such that $|y_i - z_i| \leqslant \tau$. Further, observe we can represent each y_i using $2/\tau$ functions $f \in \mathcal{F}$. Then, union bounding over the events that each of the $2/\tau$ functions output different bits, for each of the m parameters, we have that with probabiltiy at least $1 - \left(2\sqrt{2}\gamma nm/\tau\right)$, any algorithm outputs a list $\{y'_i\}_{i'\in[m]}$ such that $|y_i - y'_i| \leqslant \tau$. Finally, union bounding over the event that algorithm $\mathcal{A}$ fails with probability δ yields the claim. □

We then prove there is a gap $[f(d, \varepsilon_1), \varepsilon_1)$ between pairwise distances of components so that if we merge components within distance $f(d, \varepsilon_1)$, the resulting mixture is ε_1-statistically learnable.

Lemma 4.3.20. *Let $f(d)(\varepsilon) = f(d,\varepsilon)$. There exists $\ell \in [k^2]$ such that for every pair of components, either $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)) < (f(d))^{\ell}(\varepsilon)$ or $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)) \geqslant (f(d))^{\ell-1}(\varepsilon)$. Moreover, the set of Gaussians with total variation distance at most $(f(d))^{\ell}(\varepsilon)$ is an equivalence class.*

Proof. We can see that intervals $\left\{\left[(f(d))^{\ell}(\varepsilon),(f(d))^{\ell-1}(\varepsilon)\right)\right\}_{\ell\in[k^2]}$ are disjoint. There are at most k^2-1 distinct values of $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j))$. So there exists an interval $\left[(f(d))^{\ell}(\varepsilon),(f(d))^{\ell-1}(\varepsilon)\right)$ such that for every pair of components $\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)$, either $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)) < (f(d))^{\ell}(\varepsilon)$ or $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)) \geqslant (f(d))^{\ell-1}(\varepsilon)$.

Next, we show for any ℓ, Gaussians with pair wise TV distance $(f(d)^{\ell})(\varepsilon)$ form an equivalence class. Consider component Gaussians G_1, G_2 and G_3 such that G_1 and G_2 are at total variation distance at most $(f(d))^{\ell}(\varepsilon)$ and G_2 and G_3 are also at total variation distance at most $(f(d))^{\ell}(\varepsilon)$.

$$\begin{aligned}
d_{TV}(G_1,G_3) &\leqslant d_{TV}(G_1,G_2) + d_{TV}(G_2,G_3)\\
&\leqslant 2(f(d))^{\ell}(\varepsilon)\\
&\ll (f(d))^{\ell-1}(\varepsilon)
\end{aligned}$$

and since there is no pair of Gaussians with total variation distance inside the interval $\left[(f(d))^{\ell}(\varepsilon),(f(d))^{\ell-1}(\varepsilon)\right)$, this implies $d_{TV}(G_1,G_3) \leqslant (f(d))^{\ell}(\varepsilon)$. □

We can now complete the proof of Proposition 4.3.18 :

Proof of Proposition 4.3.18. By Lemma 4.3.20, there exists an interval $\left[(f(d))^{\ell}(\varepsilon),(f(d))^{\ell-1}(\varepsilon)\right)$ such that there is no pair of Gaussians with total variation distance inside the interval and $(f(d))^{\ell}(\varepsilon),(f(d))^{\ell-1}(\varepsilon)$ are polynomials in d and ε. Let $\varepsilon_1 = (f(d))^{\ell-1}(\varepsilon)$ and $f(d,\varepsilon_1) = (f(d))^{\ell}(\varepsilon)$. Let X be a set of $n = (d/\varepsilon)^c$ samples from $\mathcal{M}$, where c is fixed universal constant. Let $\bar{\mathcal{M}}$ be the mixture obtained by merging all components in an equivalence class with total variation distance at most $f(d,\varepsilon_1)$ to a single Gaussian and observe $d_{TV}\left(\mathcal{M}, \bar{\mathcal{M}}\right) \leqslant k f(d, \varepsilon_1)$. Next, observe that PARTITION PURSUIT outputs at most k means and covariances, which can be represented as a list of at most $2kd^2$ real numbers. Further, since $\Sigma \preceq 1.01 I$ and $w_i \geqslant \varepsilon$, the means of each component $\|\mu_i\|_2^2 \leqslant 2/\varepsilon$ and $\|\Sigma_i\|_F^2 \leqslant O(d^2/\varepsilon)$.

Then, rescaling the instance by $O(\varepsilon/d^2)$ and applying Lemma 4.3.19 with $\mathcal{D}_1 = \mathcal{M}$, $\mathcal{D}_2 = \bar{\mathcal{M}}$, input samples X and accuracy parameter $\tau = (\varepsilon/d)^{c_2}$, for a large enough constant c_2, it follows that with probability at least $1 - 0.99 - O(f(d, \varepsilon_1) \cdot (\varepsilon/d)^{c_3})$, for a fixed constant c_3, the resulting list of numbers is τ-close to that obtained by running PARTITION PURSUIT on a set of n samples from $\bar{\mathcal{M}}$. Since $\bar{\mathcal{M}}$ is ε_1-statistically learnable, it follows from Theorem 4.3.16 that with probability at least $9/10$, PARTITION PURSUIT will output an $O(\varepsilon_1)$-correct sub-division $\hat{\mathcal{M}}$.

□

Proof of Theorem 4.3.1. Recall, by part (1) of Proposition 4.3.5, the dimension of the subspace $V'_{\hat{S}}$ is $m = \dim(V'_{\hat{S}}) = O\left(\frac{(k(1+\Delta+1/\alpha))^{4k+5}}{\eta^2}\right)$ and let $\varepsilon_1 = \sqrt{\Delta}\delta^{1/4}/\alpha$. Let c_{mv} be a fixed constant such that $(m/\varepsilon_1)^{c_{mv}}$ samples suffice for applying Theorem 4.3.16. Further, obseve in the fresh sample Y, the probability that any given sample is corrupted is ε. Let ζ be the event that a random subset of $(m/\varepsilon_1)^{c_{mv}}$ samples from Z does not contain any corrupted points. Then, the event ζ holds with probability at least $(1-\varepsilon)^{(m/\varepsilon_1)^{c_{mv}}}$. Conditioning on ζ and running step 4 of Algorithm 4.3.17, it follows from Proposition 4.3.18 that we recover $O(\varepsilon_1)$-accurate estimates to the parameters of $\mathcal{M}$ in the subspace, i.e. $\|U^\top \mu_i - \hat{\mu}_i\|_2 \leqslant O(\varepsilon_1)$ and $\left\|U^\top \Sigma_i U - \hat{\Sigma}_i\right\|_F \leqslant O(\varepsilon_1)$. Since we repeat the above for ℓ' candidate subspaces in L', the probability over all probability of success is $(1-\varepsilon)^{(m/\varepsilon_1)^{c_{mv}} \cdot \ell'}$.

By part (2) in Proposition 4.3.5, there is a vector $\mu'_i \in V'_{\hat{S}}$ such that $\left\|\mu_i - \mu'_i\right\| \leqslant \frac{20}{\alpha}\delta^{1/4}\Delta^{1/2}$ where $\delta = 2\eta^{1/(C^{k+1}(k+1)!)}$ and $\eta = O(4k(1+1/\alpha+\Delta)^{4k}\sqrt{\varepsilon_1})$. Let $\hat{P} = UU^\top$ be a projection matrix where the columns of Q span $V'_{\hat{S}}$ and let $Q^\top \mu_i$ be the projection of the true means to the corresponding subspace. Then,

$$\begin{aligned}
\|U\hat{\mu}_i - \mu_i\|_2 &\leqslant \left\|U\hat{\mu}_i - \mu_i'\right\|_2 + \left\|\mu_i' - \mu_i\right\|_2 \\
&\leqslant \left\|U\hat{\mu}_i - P(\mu_i' - \mu_i + \mu_i)\right\|_2 + O\left(\frac{\sqrt{\Delta}\delta^{1/4}}{\alpha}\right) \\
&\leqslant \left\|\hat{\mu}_i - U^\top \mu_i\right\|_2 + O\left(\frac{\sqrt{\Delta}\delta^{1/4}}{\alpha}\right) \\
&\leqslant O\left(\frac{\sqrt{\Delta}\delta^{1/4}}{\alpha}\right).
\end{aligned}$$

where the third inequality follows from observing that $U^\top \mu_i$ is the true mean in the low dimensional subspace and applying Proposition 4.3.18.

By Proposition 4.3.4, there exists $\hat{S} \in L'$ such that $\hat{S} - (\Sigma_i - I) = P_i + Q_i$ where $\|P_i\|_F = O(\sqrt{\eta/\alpha})$. Again by part (3) in Proposition 4.3.5, there exists a symmetric matrix $Q_i' \in V_{\hat{S}}' \times V_{\hat{S}}'$ such that $\left\|Q_i - Q_i'\right\|_F \leqslant O(\frac{k^2}{\alpha}\delta^{1/4}\Delta^{1/2})$. We also know that in the subspace spanned by $V_{\hat{S}}'$, $\left\|\hat{\Sigma}_i - U^\top \Sigma_i U\right\|_F^2 \leqslant \mathrm{poly}(\varepsilon_2)$. Recall, Algorithm 4.3.17 outputs the following estimate: $\hat{M} = U\hat{\Sigma}U^\top + \left(I + \hat{S}\right) - \hat{P}\left(I + \hat{S}\right)\hat{P}$. Observe, for any matrix M and projection matrix P, $M = PMP + (I-P)M(I-P) + PM(I-P) + (I-P)MP$. Then,

$$\begin{aligned}
\left\|\Sigma_i - \hat{M}\right\|_F \leqslant &\underbrace{\left\|\hat{P}\left(\Sigma_i - \hat{M}\right)\hat{P}\right\|_F}_{(1)} + \underbrace{\left\|\left(I - \hat{P}\right)\left(\Sigma_i - \hat{M}\right)\left(I - \hat{P}\right)\right\|_F}_{(2)} \\
&+ \underbrace{\left\|\hat{P}\left(\Sigma_i - \hat{M}\right)\left(I - \hat{P}\right)\right\|_F}_{(3)} + \underbrace{\left\|\left(I - \hat{P}\right)\left(\Sigma_i - \hat{M}\right)\hat{P}\right\|_F}_{(4)}
\end{aligned} \tag{4.3.29}$$

We bound each of the terms above. Since $\hat{P}\left(I + S - P(I+S)P\right)\hat{P} = 0$, we can bound term

(1) as follows

$$\left\|\hat{P}\left(\Sigma_i - \hat{M}\right)\hat{P}\right\|_F = \left\|\hat{P}\Sigma_i\hat{P} - \hat{P}U\hat{\Sigma}_iU^\top\hat{P}\right\|_F = \left\|U^\top\Sigma_iU - \hat{\Sigma}_i\right\| \leqslant O\left(\frac{\sqrt{\Delta}\delta^{1/4}}{\alpha}\right) \tag{4.3.30}$$

Similarly, since $\left(I - \hat{P}\right)\left(U\hat{\Sigma}_iU^\top\right)\left(I - \hat{P}\right) = 0$ and $\Sigma_i = I + \hat{S} - P_i - Q_i$, we can bound term (2) as follows:

$$\begin{aligned}
\left\|\left(I - \hat{P}\right)\left(\Sigma_i - \hat{M}\right)\left(I - \hat{P}\right)\right\|_F &= \left\|\left(I - \hat{P}\right)\left(\Sigma_i - \left(I + \hat{S}\right)\right)\left(I - \hat{P}\right)\right\|_F \\
&\leqslant \left\|\left(I - \hat{P}\right)\left(\Sigma_i - \left(I + \hat{S} - Q_i\right)\right)\left(I - \hat{P}\right)\right\|_F + \left\|\left(I - \hat{P}\right)Q_i\left(I - \hat{P}\right)\right\|_F \\
&\leqslant \|P_i\|_F^2 + \left\|\left(I - \hat{P}\right)\left(Q_i - Q_i'\right)\left(I - \hat{P}\right)\right\|_F + \left\|\left(I - \hat{P}\right)Q_i'\left(I - \hat{P}\right)\right\|_F \\
&\leqslant O\left(\sqrt{\frac{\eta}{\alpha}} + \frac{k^2\delta^{1/4}\Delta^{1/2}}{\alpha}\right)
\end{aligned} \tag{4.3.31}$$

Next, we bound term (3). Observe, $\hat{P}\left(U\hat{\Sigma}_iU^\top\right)\left(I - \hat{P}\right) = 0$ and $\hat{P}(I + S)\hat{P}(I - \hat{P}) = 0$. Thus,

$$\begin{aligned}
\left\|\hat{P}\left(\Sigma_i - \hat{M}\right)\left(I - \hat{P}\right)\right\|_F &= \left\|\hat{P}\left(\Sigma_i - \left(I + \hat{S}\right)\right)\left(I - \hat{P}\right)\right\|_F \\
&= \left\|\hat{P}\left(P_i + Q_i\right)\left(I - \hat{P}\right)\right\|_F \\
&\leqslant \left\|\hat{P}P_i\left(I - \hat{P}\right)\right\|_F \\
&\leqslant O\left(\sqrt{\frac{\eta}{\alpha}}\right)
\end{aligned} \tag{4.3.32}$$

Obseve, term (4) follows from a similar argument. Combining equations (4.3.30), (4.3.31),(4.3.32) and substituting back into (4.3.29) we can conclude

$$\left\|\Sigma_i - \hat{M}\right\|_F \leqslant O\left(\sqrt{\frac{\eta}{\alpha}} + \frac{k^2\delta^{1/4}\Delta^{1/2}}{\alpha}\right)$$

The size of L' is $\ell' = O\left(\log k \left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4k}\right)$ and since we add a single tuple of k means and covariances for each subspace in L', the list L has the same size. The running time is poly $\left(|Y|, |L|, d^k, m, 1/\varepsilon_1\right)$ concluding the proof. $\square$

4.4 Robust Partial Cluster Recovery

In this section, we give two robust *partial clustering* algorithms. A partial clustering algorithm takes a set of points $X = \cup_{i \leqslant k} X_i$ with true clusters $X_1, X_2, \ldots, X_k$ and outputs a partition of the sample $X = X_1' \cup X_2'$ such that $X_1' = \cup_{i \in S} X_i$ and $X_2' = \cup_{i \notin S} X_i$, for some subset $S \subseteq [k]$ of size $1 \leqslant |S| < k$. That is, a partial clustering algorithm partitions the sample into two non-empty parts so that each part is a sample from a "sub-mixture". This is a weaker guarantee than clustering the entire mixture, which must find each of the original X_i's. We show that the relaxed guarantee is feasible even when the mixture as a whole is not clusterable. In our setting, we will get an approximate (that is, a small fraction of points are misclassified) partial clustering that works for ε-corruptions Y of any i.i.d. sample X from a mixture of k Gaussians, as long as there is a pair of components in the original mixture that have large total variation distance between them.

A partial clustering algorithm such as above was one of the innovations in [43] that allowed for a polynomial-time algorithm for clustering all fully clusterable Gaussian mixtures.

In this section, we build on the ideas in [43] to derive two new partial clustering algorithms that work even when the original mixture is not fully clusterable. Both upgrade the results of [43] by handling mixtures with arbitrary weights w_is instead of uniform weights and handling mixtures where not all pairs of components are well-separated in TV distance. The first algorithm succeeds under the information-theoretically minimal separation assumption (i.e. separation in total variation distance) but runs in time exponential in the inverse mixing weight. The second algorithm is a key innovation of this paper – it gives an algorithm that runs in polynomial time in the inverse mixing weight at the cost of handling separation only in relative Frobenius distance. This improved running time guarantee (at the cost of strong separation requirement that we mitigate through a novel standalone spectral separation step in Section 4.5) is crucial to obtaining the fully polynomial running time in our algorithm.

In order to state the guarantees of our algorithms, we first formulate a notion of parameter separation (same as the one employed in [43, 44]) as the next definition.

Definition 4.4.1 (Δ-Parameter Separation)**.** We say that two Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ are Δ-parameter separated if at least one of the following three conditions hold:

1. **Mean-Separation:** $\exists v \in \mathbb{R}^d$ such that $\langle \mu_1 - \mu_2, v\rangle^2 > \Delta^2 v^\top(\Sigma_1 + \Sigma_2)v$,

2. **Spectral-Separation:** $\exists v \in \mathbb{R}^d$ such that $v^\top \Sigma_1 v > \Delta\, v^\top \Sigma_2 v$,

3. **Relative-Frobenius Separation:** Σ_1 and Σ_2 have the same range space and

$$\left\|\Sigma_1^{\dagger/2}(\Sigma_2 - \Sigma_1)\Sigma_1^{\dagger/2}\right\|_F^2 > \Delta^2 \left\|\Sigma_1^{\dagger}\Sigma_2\right\|_{\mathrm{op}}^2 .$$

As shown in [44, 43], if a pair of Gaussians is $(1 - \exp(-O(\Delta \log \Delta))$-separated in total variation distance, then, they are Δ-parameter separated.

Our first algorithm succeeds in robust partial clustering whenever there is a pair of component Gaussians that are Δ-parameter separated. The running time of this algorithm grows exponentially in the reciprocal of the minimum weight in the mixture.

Theorem 4.4.2 (Robust Partial Clustering in TV Distance)**.** *Let $0 \leqslant \varepsilon < \alpha \leqslant 1$, and $\eta > 0$. There is an algorithm with the following guarantees: Let $\{\mu_i, \Sigma_i\}_{i \leqslant k}$ be means and covariances of k unknown Gaussians. Let Y be an ε-corruption of a good sample X of size $n \geqslant (dk)^{Ct}/\varepsilon$ for a large enough constant $C > 0$, from $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ satisfying Condition 4.2.29 with parameters $t = (k/\eta)^{O(k)}$ and $\gamma \leqslant \varepsilon d^{-8t} k^{-Ct}$, for a sufficiently large constant $C > 0$. Suppose further that $w_i \geqslant \alpha > 2\varepsilon$ for every i and that there are i, j such that $\mathcal{N}(\mu_i, \Sigma_i)$ and $\mathcal{N}(\mu_j, \Sigma_j)$ are Δ-parameter separated for $\Delta = (k/\eta)^{O(k)}$.*

Then, the algorithm on input Y, runs in time $n^{(k/\eta)^{O(k)}}$, and with probability at least $2^{-O\left(\frac{1}{\alpha}\log\left(\frac{k}{\eta\alpha}\right)\right)}$ over the draw of X and the algorithm's random choices, the algorithm outputs a partition of Y into Y_1, Y_2 satisfying:

1. ***Partition respects clustering:*** *for each i, $\max\{\frac{1}{w_i n}|Y_1 \cap X_i|, \frac{1}{w_i n}|Y_2 \cap X_i|\} \geqslant 1 - \eta - O(\varepsilon/\alpha^4)$, and,*

2. ***Partition is non-trivial:*** $\max_i \frac{1}{w_i n}|X_i \cap Y_1|, \max_i \frac{1}{w_i n}|X_i \cap Y_2| \geqslant 1 - \eta - O(\varepsilon/\alpha^4)$.

Our proof of the above theorem is based on a relatively straightforward extension of the ideas of [43], albeit with two key upgrades 1) allowing the input mixtures to have arbitrary mixing weights (at an exponential cost in the inverse of the minimum weight) and 2) handling mixtures where some pair of components may not be well-separated in TV distance.

In order to get our main result that gives a fully polynomial algorithm (including in the inverse mixing weights), we will use a incomparable variant of the above partial clustering method that only handles a weaker notion of parameter separation, but runs in fixed polynomial time.

Theorem 4.4.3 (Robust Partial Clustering in Relative Frobenius Distance)**.** *Let $0 \leqslant \varepsilon < \alpha/k \leqslant 1$ and $t \in \mathbb{N}$. There is an algorithm with the following guarantees: Let $\{\mu_i, \Sigma_i\}_{i \leqslant k}$ be means and covariances of k unknown Gaussians. Let Y be an ε-corruption of a sample X of size $n \geqslant (dk)^{Ct}/\varepsilon$ for a large enough constant $C > 0$, from $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ that is a good sample (satisfies Condition 4.2.29) with parameters $2t$ and $\gamma \leqslant \varepsilon d^{-8t} k^{-Ck}$, for a large enough constant $C > 0$. Suppose further that $w_i \geqslant \alpha > 2\varepsilon$ for each $i \in [k]$, and that for some $t \in \mathbb{N}$, $\beta > 0$ there exist $i, j \leqslant k$ such that $\left\|\Sigma^{\dagger/2}(\Sigma_i - \Sigma_j)\Sigma^{\dagger/2}\right\|_F^2 = \Omega\left((k^2 t^4)/(\beta^{2/t}\alpha^4)\right)$, where Σ is the covariance of the mixture $\mathcal{M}$. Then, the algorithm runs in time $n^{O(t)}$, and with probability at least $2^{-O\left(\frac{1}{\alpha}\log\left(\frac{k}{\eta}\right)\right)}$ over the random choices of the algorithm, outputs a partition $Y = Y_1 \cup Y_2$ satisfying:*

1. ***Partition respects clustering:*** *for each* i, $\max\left\{\frac{1}{w_i n}|Y_1 \cap X_i|, \frac{1}{w_i n}|Y_2 \cap X_i|\right\} \geqslant 1 - \beta - O(\varepsilon/\alpha^4)$*, and,*

2. ***Partition is non-trivial:*** $\max_i \frac{1}{w_i n}|X_i \cap Y_1|, \max_i \frac{1}{w_i n}|X_i \cap Y_2| \geqslant 1 - \beta - O(\varepsilon/\alpha^4)$.

The starting point for the proof of the above theorem is the observation that the running time of our first algorithm above is exponential in the inverse mixing weight almost entirely because of dealing with spectral separation (which requires the use of "certifiable anti-concentration" that we define in the next subsection). We formulate a variant of relative Frobenius separation (that is directly useful to us) and prove that whenever the original mixture has a pair of components separated in this notion, we can in fact obtain a fully polynomial partial clustering algorithm building on the ideas in [43].

4.4.1 Algorithm

Our algorithm will solve SoS relaxations of a polynomial inequality system. The constraints here use the input Y to encode finding a sample X' (the intended setting being $X' = X$, the original uncorrupted sample) and a cluster $\hat{C}$ in X' of size $= \alpha n$, indicated by z_is (the intended setting is simply the indicator for any of the k true clusters) satisfying properties of Gaussian distribution (certifiable hypercontractivity and anti-concentration).

Covariance constraints introduce a matrix valued indeterminate Π intended to be the square root of $\hat{\Sigma}$, the sos variable for the covariance of a single component.

$$\text{Covariance Constraints: } \mathcal{A}_1 = \left\{ \begin{array}{c} \Pi = UU^\top \\ \Pi^2 = \hat{\Sigma} \end{array} \right\} \tag{4.4.1}$$

The intersection constraints force that X' be ε-close to Y (and thus, 2ε-close to unknown sample X).

$$\text{Intersection Constraints: } \mathcal{A}_2 = \left\{ \begin{aligned} &\forall i \in [n], & m_i^2 = m_i \\ & & \textstyle\sum_{i\in[n]} m_i = (1-\varepsilon)n \\ &\forall i \in [n], & m_i(y_i - x_i') = 0 \end{aligned} \right\} \tag{4.4.2}$$

The subset constraints introduce z, which indicates the subset $\hat{C}$ intended to be the true clusters of X'.

$$\text{Subset Constraints: } \mathcal{A}_3 = \left\{ \begin{aligned} &\forall i \in [n]. & z_i^2 = z_i \\ & & \textstyle\sum_{i\in[n]} z_i = \alpha n \end{aligned} \right\} \tag{4.4.3}$$

Parameter constraints create indeterminates to stand for the covariance $\hat{\Sigma}$ and mean $\hat{\mu}$ of $\hat{C}$ (indicated by z).

$$\text{Parameter Constraints: } \mathcal{A}_4 = \left\{ \begin{aligned} \frac{1}{\alpha n}\sum_{i=1}^{n} z_i \left(x_i' - \hat{\mu}\right)\left(x_i' - \hat{\mu}\right)^\top &= \hat{\Sigma} \\ \frac{1}{\alpha n}\sum_{i=1}^{n} z_i x_i' &= \hat{\mu} \end{aligned} \right\} \tag{4.4.4}$$

Certifiable Hypercontractivity : $\mathcal{A}_4$=

$$\left\{ \begin{aligned} &\forall t \leqslant 2s \quad \frac{1}{\alpha^2 n^2}\sum_{i,j\leqslant n} z_i z_j (Q(x_i' - x_j') - \mathop{\mathbb{E}}_z Q)^{2t} \leqslant \left(\frac{1}{\alpha^2 n^2}\sum_{i,j\leqslant n} z_i z_j (Q(x_i' - x_j') - \mathop{\mathbb{E}}_z Q)^2\right)^t \\ &\qquad\quad \frac{1}{\alpha^2 n^2}\sum_{i,j\leqslant n} z_i z_j \left(Q(x_i' - x_j') - \mathop{\mathbb{E}}_z Q\right)^2 \leqslant \frac{6}{\alpha^2}\|Q\|_F^2 \end{aligned} \right\} \tag{4.4.5}$$

Here, we used the shorthand $\mathbb{E}_z Q = \frac{1}{\alpha^2 n^2}\sum_{i,j\leqslant n} z_i z_j Q(x_i' - x_j')$.

In the constraint system for our first algorithm, we will use the following certifiable anti-concentration constraints on $\hat{C}$ for $\delta = \alpha^{-\operatorname{poly}(k)}$ and $\tau = \alpha/\operatorname{poly}(k)$ and $s(u) = 1/u^2$

for every u.

$$\text{Certifiable Anti-Concentration} : \mathcal{A}_5 = \left\{ \begin{array}{l} \frac{1}{\alpha^2 n^2} \sum_{i,j=1}^{n} z_i z_j q^2_{\delta,\Sigma}\left(\left(x'_i - x'_j\right), v\right) \leqslant 2^{s(\delta)} C\delta \left(v^\top \Sigma v\right)^{s(\delta)} \\ \frac{1}{\alpha^2 n^2} \sum_{i,j=1}^{n} z_i z_j q^2_{\tau,\Sigma}\left(\left(x'_i - x'_j\right), v\right) \leqslant 2^{s(\tau)} C\tau \left(v^\top \Sigma v\right)^{s(\eta)} \end{array} \right\} \tag{4.4.6}$$

We note that the constraint system for our second algorithm (running in fixed polynomial time), we will not use $\mathcal{A}_5$. Towards proving Theorems 4.4.3 and 4.4.2 we use the following algorithm that differs only in the degree of the pseudo-distribution computed and the constraint system that the pseudo-distribution satisfies.

Algorithm 4.4.4 (Partial Clustering)**.**

Given: *A sample Y of size n. An outlier parameter $\varepsilon > 0$ and an accuracy parameter $\eta > 0$.*

Output: *A partition of Y into partial clustering $Y_1 \cup Y_2$.*

Operation:

1. ***SDP Solving:*** *Find a pseudo-distribution $\tilde{\zeta}$ satisfying $\cup_{i=1}^{5} \mathcal{A}_i$ ($\cup_{i=1}^{4} \mathcal{A}_i$ for Theorem 4.4.3) such that $\tilde{\mathbb{E}}_{\tilde{\zeta}} z_i \leqslant \alpha + o_d(1)$ for every i. If no such pseudo-distribution exists, output fail.*
2. ***Rounding:*** *Let $M = \tilde{\mathbb{E}}_{z \sim \tilde{\zeta}}[zz^\top]$.*
   (a) *Choose $\ell = O\big(\frac{1}{\alpha} \log(k/\eta)\big)$ rows of M uniformly at random and independently.*
   (b) *For each $i \leqslant \ell$, let $\hat{C}_i$ be the indices of the columns j such that $M(i,j) \geqslant \eta^2 \alpha^5 / k$.*

(c) Choose a uniformly random $S \subseteq [\ell]$ and output $Y_1 = \cup_{i \in S} \hat{C}_i$ and $Y_2 = Y \setminus Y_1$.

4.4.2 Analysis

Simultaneous Intersection Bounds. The key observation for proving the first theorem is the following lemma that gives a sum-of-squares proof that no z that satisfies the constraints $\cup_{i=1}^{5} \mathcal{A}_i$ can have simultaneously large intersections with the Δ-parameter separated component Gaussians.

Lemma 4.4.5 (Simultaneous Intersection Bounds for TV-separated case)**.** *Let Y be an ε-corruption of a good sample X of size $n \geqslant (dk)^{Ct}/\varepsilon$ for a large enough constant $C > 0$, from $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ satisfying Condition 4.2.29 with parameters $t = (k/\eta)^{O(k)}$ and $\gamma \leqslant \varepsilon d^{-8t} k^{-Ct}$, for a sufficiently large constant $C > 0$. Suppose further that $w_i \geqslant \alpha > 2\varepsilon$ for every i and that there are i, j such that $\mathcal{N}(\mu_i, \Sigma_i)$ and $\mathcal{N}(\mu_j, \Sigma_j)$ are Δ-parameter separated for $\Delta = (k/\eta)^{O(k)}$. Then, there exists a partition of $[k]$ into $S \cup L$ such that, $|S|, |L| < k$ and for $z(X_r) = \frac{1}{w_r n} \sum_{i \in X_r} z_i$,*

$$\left\{\cup_{i=1}^{5} \mathcal{A}_i\right\} \Big|\frac{z}{(k/\eta\alpha)^{\mathrm{poly}(k)}} \left\{ \sum_{i \in S, j \in L} z(X_i) z(X_j) \leqslant O(k^2 \varepsilon/\alpha) + \eta/\alpha \right\}.$$

The proof of Lemma 4.4.5 is given in Section 4.4.3.

For the second theorem, we use the following version that strengthens the separation assumption and lowers the degree of the sum-of-squares proof (and consequently the running time of the algorithm) as a result.

Lemma 4.4.6 (Simultaneous Intersection Bounds for Frobenius Separated Case)**.** *Let X be a sample of size $n \geqslant (dk)^{Ct}/\varepsilon$ for a large enough constant $C > 0$, from $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ that is a good sample (satisfies Condition 4.2.29) with parameters $2t$ and $\gamma \leqslant \varepsilon d^{-8t} k^{-Ck}$, for a large enough constant $C > 0$. Suppose further that*

$w_i \geqslant \alpha > 2\varepsilon$ for each $i \in [k]$, and that for some $t \in \mathbb{N}$, $\beta > 0$ there exist $i, j \leqslant k$ such that $\left\|\Sigma^{\dagger/2}(\Sigma_i - \Sigma_j)\Sigma^{\dagger/2}\right\|_F^2 = \Omega\left((k^2 t^4)/(\beta^{2/t}\alpha^2)\right)$, where Σ is the covariance of the mixture $\mathcal{M}$. Then, for any ε-corruption Y of X, there exists a partition of $[k] = S \cup T$ such that

$$\left\{\cup_{i=1}^4 \mathcal{A}_i\right\} \Big|\frac{z}{2t} \left\{\sum_{i \in S}\sum_{j \in T} z(X_i)z(X_j) \leqslant O(k^2)\beta + O(k^2)\varepsilon/\alpha\right\}.$$

Here, $z(X_r) = \frac{1}{w_r n}\sum_{i \in X_r} z_i$ for every r.

The proof of Lemma 4.4.6 is given in Section 4.4.4.

Notice that the main difference between the above two lemmas is the constraint systems they use. Specifically, the second lemma does *not* enforce certifiable anti-concentration constraints. As a result, there is a difference in the degree of the sum-of-squares proofs they claim; the degree of the SoS proof in the second lemma does not depend on the inverse minimum mixture weight.

First, we complete the proof of the Theorem 4.4.2. The proof of Theorem 4.4.3 is exactly the same except for the use of Lemma 4.4.6 (and thus has the exponent in the running time independent of $1/\alpha$) instead of Lemma 4.4.5.

Proof of Theorem 4.4.2. Let $\eta' = O(\eta^2\alpha^3/k)$. We will prove that whenever $\Delta \geqslant \mathrm{poly}(k/\eta')^k = \mathrm{poly}\left(\frac{k}{\eta\alpha}\right)^k$, Algorithm 4.4.4, when run with input Y, with probability at least 0.99, recovers a collection $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_\ell$ of $\ell = O\big(\frac{1}{\alpha}\log k/\eta\big)$ subsets of indices satisfying $|\cup_{i \leqslant \ell}\hat{C}_i| \geqslant (1 - \eta'/k^{40})n$ such that there is a partition $S \cup L = [\ell]$, $0 < |S| < \ell$ satisfying that for all $i \in [\ell]$:

$$\min\left\{\frac{1}{\alpha n}|\hat{C}_i \cap \cup_{j \in S} X_j|, \frac{1}{\alpha n}|\hat{C}_i \cap \cup_{j \in L} X_j|\right\} \leqslant 100\eta'/\alpha^3 + O(\varepsilon/\alpha^4). \tag{4.4.7}$$

We first argue that this suffices to complete the proof. Split $[\ell]$ into two groups G_S, G_L as follows. For each i, let $j = \mathrm{argmax}_{r \in [\ell]} \frac{1}{\alpha n}|\hat{C}_i \cap X_r|$. If $j \in S$, add it to G_S, else add it to G_L. Observe that this process is well-defined - i.e, there cannot be $j \in S$ and $j' \in L$ that both maximize $\frac{1}{\alpha n}|\hat{C}_i \cap X_r|$ as r varies over $[k]$. To see this, WLOG, assume $j \in S$. Note that $\frac{1}{\alpha n}|\hat{C}_i \cap X_j| \geqslant 1/k$. Then, we immediately obtain: $\frac{1}{\alpha n}| \cup_{j \in S} X_j \cap \hat{C}_i| \geqslant 1/k$. Now, if we ensure that $\eta' \leqslant \alpha^3/k^2$ and $\varepsilon \leqslant O(\alpha^4/k)$, then, $\frac{1}{\alpha n}|\hat{C}_i \cap \cup_{j' \in L} X_{j'}|$ is at most the RHS of (4.4.7) which is $\ll 1/k$. This completes the proof of well-definedness. Next, adding up (4.4.7) for each $i \in S$ yields that

$$\frac{1}{|\hat{C}_i|}| \left(\cup_{t \in G_S} \hat{C}_t\right) \cap \cup_{j \in L} X_j| \leqslant O(\log(k/\eta')/\alpha)\,(\eta' + O(\varepsilon/\alpha))\,,$$

where we used that $|G_S| \leqslant \ell$. Combined with $| \cup_{i \leqslant \ell} \hat{C}_i| \geqslant (1 - \eta'/k^{40})n$, we obtain that

$$| \cup_{i \in G_S} X_i| \geqslant 1 - \eta'/k^{40} - O(\log(k/\eta')/\alpha)\,(\eta' + O(\varepsilon/\alpha)) = \eta + O\big(\log(k/\eta\alpha)\varepsilon/\alpha^2\big)$$

for $\eta' \leqslant O\big(\eta^2\alpha^3/k\big)$.

We now go ahead and establish (4.4.7). Let $\tilde{\zeta}$ be a pseudo-distribution satisfying $\mathcal{A}$ of degree $(k/\eta)^{\mathrm{poly}(k)}$ satisfying $\tilde{\mathbb{E}}_{\tilde{\zeta}} z_i = \alpha$ for every i. Such a pseudo-distribution exists. To see why, let $\tilde{\zeta}$ be the actual distribution that always sets $X' = X$, chooses an i with probability w_i and outputs a uniformly subset $\hat{C}$ of size αn of X_i conditioned on $\hat{C}$ satisfying $\mathcal{A}$. Then, notice that since X is a good sample (satisfies Condition 4.2.29), by Fact 4.2.27, the uniform distribution on each X_i has t-certifiably C-hypercontractive degree 2 polynomials and is t-certifiably $C\delta$-anti-concentrated. By an concentration argument using high-order Chebyshev inequality, similar to the proof of Lemma 4.2.33 (applied to uniform distribution on X_i of size $n \geqslant (dk)^{O(t)}$, $\hat{C}$ chosen above satisfies the constraints $\mathcal{A}$ with probability at least $1 - o_d(1)$. Observe that the probabililty that z_i is set to 1 under this distribution is then at most $\alpha + o_d(1)$. Thus, such a distribution satisfies all the constraints in $\mathcal{A}$.

Next, let $M = \tilde{\mathbb{E}}_{\tilde{\zeta}}[zz^\top]$. Then, we claim that:

1. $o_d(1) + \alpha \geqslant M(i,j) \geqslant 0$ for all i,j,

2. $M(i,i) \in \alpha \pm o_d(1)$ for all i,

3. $\mathbb{E}_{j\sim[n]} M(i,j) \geqslant \alpha^2 - o_d(1)$ for every i.

The proofs of these basic observations are similar to those presented in Chapter 4.3 of [66] (see also the proof of Theorem 5.1 in [43]): Observe that $\mathcal{A} \vdash_4 \left\{z_i z_j = z_i^2 z_j^2 \geqslant 0\right\}$ for every i. Thus, by Fact 3.3.2, $\tilde{\mathbb{E}}[z_i z_j] \geqslant 0$ for every i, j. Next, observe that $\mathcal{A} \vdash_2 \left\{(1-z_i) = (1-z_i)^2 \geqslant 0\right\}$ for every i and thus, $\mathcal{A} \vdash_2 \left\{z_i(1-z_j) \geqslant 0\right\}$. Thus, by Fact 3.3.2 again, we must have $\tilde{\mathbb{E}}[z_i z_j] \leqslant \tilde{\mathbb{E}}[z_i] \leqslant \alpha + o_d(1)$. Finally, $\mathcal{A} \vdash_2 \left\{\sum_j z_i z_j = z_i \sum_j z_j = \alpha n z_i\right\}$. Thus, by Fact 3.3.2 again, we must have $\sum_j M(i,j) = \sum_j \tilde{\mathbb{E}}[z_i z_j] = \alpha n \sum_j \tilde{\mathbb{E}}[z_i] \in (\alpha^2 \pm o_d(1))n$. Let B_i be the entries in the i-th row M_i that are larger than $\alpha^2/2$. Then, by (1) and (2), we immediately derive that B_i must have at least $\alpha n/2$ elements. Call an entry of M large if it exceeds $\alpha^2\eta'$. For each i, let B_i be the set of large entries in row i of M. Then, using (3) and (1) above gives that $|B_i| \geqslant \alpha(1-\alpha\eta')n$ for each $1 \leqslant i \leqslant n$. Next, call a row i "good" if $\frac{1}{\alpha n}\min\{|\cup_{r\in L} X_r \cap B_i|, |\cup_{r'\in S} X_{r'} \cap B_i|\} \leqslant 100\eta'/\alpha^3 + O(\varepsilon/\alpha^4)$. Let us estimate the fraction of rows of M that are good.

Towards that goal, let us apply Lemma 4.4.5 with η set to η' and use Fact 3.3.2 (SoS Completeness), to obtain $\sum_{r\in S, r'\in L} \mathbb{E}_{i\in X_r} \mathbb{E}_{j\in X_{r'}} M(i,j) \leqslant \eta' + O(\varepsilon/\alpha)$. Using Markov's inequality, with probability $1-\alpha^3/100$ over the uniformly random choice of i, $\mathbb{E}_{j\in X_{r'}} M(i,j) \leqslant 100\frac{1}{\alpha^3}\eta' + O(\varepsilon/\alpha^4)$. Thus, $1-\alpha^3/100$ fraction of the rows of M are good.

Next, let R be the set of $\frac{100}{\alpha}\log\left(\frac{k^{50}}{\eta'}\right)$ rows sampled in the run of the algorithm and set $\hat{C}_i = B_i$ for every $i \in R$. The probability that all of them are good is then at least $(1-\alpha^3/100)^{\frac{100}{\alpha}\log\left(\frac{k^{50}}{\alpha\eta'}\right)} \geqslant 1-\alpha$. Let us estimate the probability that $|\cup_{i\in R}\hat{C}_i| \geqslant (1-\eta'/k^{40})n$. For a uniformly random i, the chance that a given point $t \in B_i$ is at least $\alpha(1-\alpha\eta')$. Thus, the chance that $t \notin \cup_{i\in R} B_i$ is at most $(1-\alpha/2)^{100/\alpha\log\left(k^{50}/(\alpha\eta')\right)} \leqslant \eta'/k^{50}$. Thus, the expected number of t that are not covered by $\cup_{i\in R}\hat{C}_i$ is at most $n\eta'/k^{50}$. Thus, by Markov's inequality, with probability at least $1-1/k^{10}$, $1-\eta'/k^{40}$ fraction of t are covered in $\cup_{i\in R}\hat{C}_i$. By the above computations and a union bound, with probability at least $1-\eta'/k^{10}$ both the conditions below hold simultaneously: 1) each of the $\frac{100}{\alpha}\log\left(k^{50}/\eta'\right)$ rows R sampled are good and 2) $|\cup_{i\in R}\hat{C}_i| \geqslant (1-\eta'/k^{40})n$. This completes the proof. $\square$

4.4.3 Proof of Lemma 4.4.5

Our proof is based on the following simultaneous intersection bounds from [43]. We will use the following lemma that forms the crux of the analysis of the clustering algorithm in [43]:

Lemma 4.4.7 (Simultaneous Intersection Bounds, Lemma 5.4 in [43])**.** *Fix $\delta > 0, k \in \mathbb{N}$. Let $X = X_1 \cup X_2 \cup \ldots X_k$ be a good sample of size n from a k-mixture $\sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ of Gaussians. Let Y be any ε-corruption of X. Suppose there are $r, r' \leqslant k$ such that one of the following three conditions hold for some $\Delta \geqslant (k/\delta)^{O(k)}$:*

1. *there exists a v such that $v^\top \Sigma(r') v > \Delta v^\top \Sigma(r') v$ and $B = \max_{i\leqslant k} \frac{v^\top \Sigma(i) v}{v^\top \Sigma(r') v}$, or*
2. *there exists a $v \in \mathbb{R}^d$ such that $\langle \mu(r) - \mu(r'), v\rangle_2^2 \geqslant \Delta^2 v^\top \left(\Sigma(r) + \Sigma(r')\right) v$, or,*
3. *$\left\|\Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2} - I\right\|_F^2 \geqslant \Delta^2 \left(\left\|\Sigma(r')^{-1/2}\Sigma(r)^{1/2}\right\|_{\mathrm{op}}^4\right)$.*

Then, for the linear polynomial $z(X_r) = \frac{1}{\alpha n}\sum_{i\in X_r} z_i$ in indeterminates z_is satisfies:

$$\{\cup_{i\leqslant 5}\mathcal{A}_i\} \vdash_{\frac{z}{(k/\delta)^{O(k)}\log(2B)}} \left\{ z(X_r) z(X_{r'}) \leqslant O(\sqrt{\delta}) + O(\varepsilon/\alpha) \right\} .$$

Proof of Lemma 4.4.5. Without loss of generality, assume that the pair of separated components are $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$. Let us start with the case when the pair is spectrally separated. Then, there is a $v \in \mathbb{R}^d$ such that $\Delta v^\top \Sigma_1 v \leqslant v^\top \Sigma_2 v$.

Consider an ordering of the true clusters along the direction v, renaming cluster indices if needed, such that $v^\top \Sigma_1 v \leqslant v^\top \Sigma_2 v \leqslant \ldots v^\top \Sigma_k v$. Let $j \leqslant k'$ be the largest integer such that $\mathrm{poly}(k/\eta) v^\top \Sigma_j v \leqslant v^\top \Sigma_{j+1} v$. Further, observe that since j is defined to be the largest index which incurs separation $\mathrm{poly}(k/\eta)$, all indices in $[j, k]$ have spectral bound at most $\mathrm{poly}(k/\eta)$ and thus $\frac{v^\top \Sigma_k v}{v^\top \Sigma_j v} \leqslant \mathrm{poly}(k/\eta)^k$.

Applying Lemma 4.4.7 with the above direction v to every $r < j$ and $r' \geqslant j$ and observing that the parameter B in each case is at most $\frac{v^\top \Sigma_k v}{v^\top \Sigma_j v} \leqslant \Delta^k$ yields:

$$\mathcal{A} \Big|\frac{z}{O(k^2 s^2 \operatorname{poly}\log(\Delta))} \left\{ z(X_r) z(X_{r'}) \leqslant O(\varepsilon/\alpha) + \sqrt{\delta} \right\} .$$

Adding up the above inequalities over all $r \leqslant j-1$ and $r' \geqslant j+1$ and taking $S = [j-1]$, $T = [k] \setminus [j-1]$ completes the proof in this case.

Next, let us take the case when $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ are mean-separated. WLOG, suppose $\langle \mu_1, v\rangle \leqslant \langle \mu_2, v\rangle \ldots \leqslant \langle \mu_k, v\rangle$. Then, we know that $\langle \mu_k - \mu_1, v\rangle \geqslant \Delta v^\top \Sigma_i v$. Thus, there must exist an i such that $\langle \mu_i - \mu_{i+1}, v\rangle \geqslant \Delta v^\top \Sigma_i v/k$. Let $S = [i]$ and $L = [k] \setminus S$. Applying Lemma 4.4.7 and arguing as in the previous case (and noting that $\kappa = \operatorname{poly}(k)$) completes the proof.

Finally, let us work with the case of relative Frobenius separation. Since $\|\Sigma_1^{-1/2} \Sigma_k^{1/2}\| \leqslant \operatorname{poly}(k)$, the hypothesis implies that $\|\Sigma_1 - \Sigma_2\|_F \geqslant \Delta/\operatorname{poly}(k)$. Let $B = \Sigma_1 - \Sigma_2$ and let $A = B/\|B\|_F$. WLOG, suppose $\langle \Sigma_1, A\rangle \leqslant \ldots \langle \Sigma_k, A\rangle$. Then, since $\langle \Sigma_k, A\rangle - \langle \Sigma_1, A\rangle \geqslant \Delta/\operatorname{poly}(k)$, there must exist an i such that $\langle \Sigma_{i+1}, A\rangle - \langle \Sigma_i, A\rangle \geqslant \Delta/\operatorname{poly}(k)$. Let us now set $S = [i]$ and $L = [k] \setminus S$.

Then, for every $i \in S$ and $j \in L$, we must have: $\langle \Sigma_j, A\rangle - \langle \Sigma_i, A\rangle \geqslant \Delta/\operatorname{poly}(k)$. Thus, $\left\|\Sigma_j - \Sigma_i\right\|_F \geqslant \Delta/\operatorname{poly}(k)$. And thus, $\Delta/\operatorname{poly}(k) \leqslant \left\|\Sigma_j - \Sigma_i\right\|_F \leqslant \left\|\Sigma_i^{-1/2}\Sigma_j^{1/2}\right\|_2^2 \left\|\Sigma_i^{-1/2}\Sigma_j\Sigma_i^{-1/2} - I\right\|_F$. Rearranging and using the bound on $\left\|\Sigma_i^{-1/2}\Sigma_j^{1/2}\right\|_2^2$ yields that $\left\|\Sigma_i^{-1/2}\Sigma_j\Sigma_i^{-1/2} - I\right\|_F \geqslant \Delta/\operatorname{poly}(k)$.

A similar argument as in the two cases above now completes the proof.

□

4.4.4 Proof of Lemma 4.4.6

We use $\mathbb{E}_z$ as a shorthand for $\frac{1}{\alpha n}\sum_{i=1}^n z_i$. We will write $\frac{1}{w_r n}\sum_{j \in X_r} z_j = z(X_r)$. Note that $z(X_r) \in [0,1]$. And finally, we will write $z'(X_r) = \frac{1}{w_r n}\sum_{j\in X_r} z_j \mathbf{1}(x_j = y_j)$ – the version of $z(X_r)$ that only sums over non-outliers.

We will use the following technical facts in the proof:

Fact 4.4.8 (Lower Bounding Sums, Fact 4.19 [43])**.** *Let A, B, C, D be scalar-valued indeterminates. Then, for any $\tau > 0$,*

$$\{0 \leqslant A, B \leqslant A + B \leqslant 1\} \cup \{0 \leqslant C, D\} \cup \{C + D \geqslant \tau\} \Big|\frac{A,B,C,D}{2} \{AC + BD \geqslant \tau AB\} \ .$$

Fact 4.4.9 (Cancellation within SoS, Lemma 9.2 in [43])**.** *For indeterminate a and any $t \in \mathbb{N}$,*

$$\left\{a^{2t} \leqslant 1\right\} \Big|\frac{a}{2t} \{a \leqslant 1\} \ .$$

Lemma 4.4.10 (Lower-Bound on Variance of Degree 2 Polynomials)**.** *Let $Q \in \mathbb{R}^{d\times d}$. Then, for any $i, j \leqslant k$, and $z'(X_r) = \frac{1}{w_r n} \sum_{i\in X_r} z_i \mathbf{1}(x_i = y_i)$, we have:*

$$\mathcal{A} \Big|\frac{z}{4} \left\{ z'(X_r)^2 z'(X_r')^2 \leqslant \frac{(32Ct)^{2t}}{(\mathbb{E}_{X_r} Q - \mathbb{E}_{X_{r'}} Q)^{2t}} \left(\frac{\alpha^4}{w_r^2 w_{r'}^2} \left(\mathbb{E}_z (Q - \mathbb{E}_z Q)^2\right)^t \right.\right.$$
$$\left.\left. + \frac{\alpha^2}{w_r^2} \left(\mathbb{E}_{X_{r'}} (Q - \mathbb{E}_{X_{r'}} Q)^2\right)^t + \frac{\alpha^2}{w_{r'}^2} \left(\mathbb{E}_{X_r} (Q - \mathbb{E}_{X_r} Q)^2\right)^t \right) \right\} .$$

Proof. Let $z_i' = z_i \mathbf{1}(x_i = y_i)$ for every i. Using the substitution rule and non-negativity constraints of the z_i's, we have

$$\mathcal{A} \Big|\frac{z}{4} \left\{ \mathbb{E}_z (Q - \mathbb{E}_z Q)^{2t} = \frac{1}{\alpha^2 n^2} \sum_{i,j\leqslant n} z_i' z_j' \left(Q(x_i - x_j) - \mathbb{E}_z Q \right)^{2t} \right.$$
$$\left. \geqslant \frac{1}{\alpha^2 n^2} \sum_{i,j\in X_r \text{ or } i,j\in X_{r'}} z_i' z_j' \left(Q(x_i - x_j) - \mathbb{E}_z Q \right)^{2t} \right\}$$

Using the SoS almost triangle inequality, we have

$$
\begin{aligned}
\mathcal{A} \left|\frac{z}{4}\right. \Bigg\{ & \frac{1}{\alpha^2 n^2} \sum_{i,j \in X_r \text{ or } i,j \in X_{r'}} z'_i z'_j \left(Q(x_i - x_j) - \mathop{\mathbb{E}}_{z} Q \right)^{2t} \\
& \geqslant \left(\frac{1}{2^{2t}}\right) \left(\frac{1}{\alpha^2 n^2} \sum_{i,j \in X_r} z'_i z'_j \left(\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_{z} Q \right)^{2t} - \frac{1}{\alpha^2 n^2} \sum_{i,j \in X_r} z'_i z'_j \left(Q(x_i - x_j) - \mathop{\mathbb{E}}_{X_r} Q \right)^{2t} \right) \\
& \quad + \left(\frac{1}{2^{2t}}\right) \left(\frac{1}{\alpha^2 n^2} \sum_{i,j \in X_r} z'_i z'_j \left(\mathop{\mathbb{E}}_{X_{r'}} Q - \mathop{\mathbb{E}}_{z} Q \right)^{2t} - \frac{1}{\alpha^2 n^2} \sum_{i,j \in X_{r'}} z'_i z'_j \left(Q(x_i - x_j) - \mathop{\mathbb{E}}_{X_{r'}} Q \right)^{2t} \right) \\
& = 2^{-2t} \left((w_r/\alpha)^2 z(X_r)^2 \left(\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_{z} Q \right)^{2t} - \frac{1}{\alpha^2 n^2} \sum_{i,j \in X_r} \left(Q(x_i - x_j) - \mathop{\mathbb{E}}_{X_r} Q \right)^{2t} \right) \\
& \quad + 2^{-2t} \left((w_{r'}/\alpha)^2 z(X_{r'})^2 \left(\mathop{\mathbb{E}}_{X_{r'}} Q - \mathop{\mathbb{E}}_{z} Q \right)^{2t} - \frac{1}{\alpha^2 n^2} \sum_{i,j \in X_{r'}} \left(Q(x_i - x_j) - \mathop{\mathbb{E}}_{X_{r'}} Q \right)^{2t} \right) \Bigg\}
\end{aligned}
$$

Using Fact 4.4.8, we can further simplify the above as follows:

$$
\begin{aligned}
\mathcal{A} \left|\frac{z}{4}\right. \Bigg\{ \mathop{\mathbb{E}}_{z}(Q - \mathop{\mathbb{E}}_{z} Q)^{2t} & \geqslant 2^{-6t} \frac{w_r^2 w_{r'}^2}{\alpha^4} z'(X_r)^2 z'(X_{r'})^2 (\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^{2t} \\
& \quad - 2^{-6t} (w_r/\alpha)^2 \mathop{\mathbb{E}}_{X_r}(Q - \mathop{\mathbb{E}}_{X_r} Q)^{2t} - 2^{-6t} (w_{r'}/\alpha)^2 \mathop{\mathbb{E}}_{X_{r'}}(Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^{2t} \\
& \geqslant 2^{-6t} \frac{w_r^2 w_{r'}^2}{\alpha^4} z'(X_r)^2 z'(X_{r'})^2 (\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^{2t} - (w_r/\alpha)^2 (Ct)^{2t} \left(\mathop{\mathbb{E}}_{X_r}(Q - \mathop{\mathbb{E}}_{X_r} Q)^2 \right)^t \\
& \quad - (w_{r'}/\alpha)^2 (Ct)^{2t} \left(\mathop{\mathbb{E}}_{X_{r'}}(Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^2 \right)^t \Bigg\}
\end{aligned}
$$

where the last inequality follows from the Certifiable Hypercontractivity constraint ($\mathcal{A}_4$). Rearranging completes the proof.

□

We can use the lemma above to obtain a simultaneous intersection bound guarantee when there are relative Frobenius separated components in the mixture.

Lemma 4.4.11. *Suppose* $\left\|\Sigma^{-1/2}(\Sigma_r - \Sigma_{r'})\Sigma^{-1/2}\right\|_F^2 \geqslant 10^8 \frac{C^6 t^4}{\beta^{2/t}\alpha^2}$. *Then, for* $z'(X_r) = \frac{1}{\alpha n}\sum_{i \in X_r} z_i \cdot \mathbf{1}(y_i = x_i)$ *for every* r,

$$\mathcal{A} \left|\frac{w}{2t}\right. \{z'(X_r) z'(X_r') \leqslant \beta\} \ .$$

Proof. We work with the transformation $x_i \to \Sigma^{-1/2} x_i$. Let $\Sigma_z' = \Sigma^{-1/2}\Sigma_z\Sigma^{-1/2}$, $\Sigma_r' = \Sigma^{-1/2}\Sigma_r\Sigma^{-1/2}$ and $\Sigma_{r'}' = \Sigma^{-1/2}\Sigma_{r'}\Sigma^{-1/2}$ be the transformed covariances. Note that transformation is only for the purpose of the argument - our constraint system does not depend on Σ.

Notice that $\left\|\Sigma_{r'}'\right\|_2 \leqslant \frac{1}{w_r}$ and $\left\|\Sigma_{r'}'\right\|_2 \leqslant \frac{1}{w_{r'}}$.

We now apply Lemma 4.4.10 with $Q = \Sigma_r' - \Sigma_{r'}'$. Then, notice that $\mathbb{E}_{X_r} Q - \mathbb{E}_{X_{r'}} Q = \left\|\Sigma_r' - \Sigma_{r'}'\right\|_F^2 = \|Q\|_F^2$. Then, we obtain:

$$\begin{aligned}&\mathcal{A} \left|\frac{z}{2t}\right. \Bigg\{ z'(X_r)^2 z'(X_{r'})^2 \\ &\leqslant \left(\frac{32Ct}{\mathbb{E}_{X_r} Q - \mathbb{E}_{X_{r'}} Q}\right)^{2t} \left(\frac{\alpha^4}{w_r^2 w_{r'}^2}\left(\mathop{\mathbb{E}}_z (Q - \mathop{\mathbb{E}}_z Q)^2\right)^t + \frac{\alpha^2}{w_r^2}\left(\mathop{\mathbb{E}}_{X_{r'}} (Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^2\right)^t + \frac{\alpha^2}{w_{r'}^2}\left(\mathop{\mathbb{E}}_{X_r}(Q - \mathop{\mathbb{E}}_{X_r} Q)^2\right)\right)^t \Bigg\} .\end{aligned} \tag{4.4.3}$$

Since X_r and $X_{r'}$ have certifiably C-bounded variance polynomials for $C = 4$ (as a consequence of Condition 4.2.29 and Fact 4.2.27 followed by an application of Lemma 3.3.9), we have:

$$\mathcal{A} \left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{X_{r'}} (Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^2 \leqslant 6 \left\|\Sigma_{r'}'^{1/2} Q \Sigma_{r'}'^{1/2}\right\|_F^2 \leqslant \frac{6}{w_{r'}^2}\|Q\|_F^2 \right\} ,$$

and

$$\mathcal{A} \left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{X_r} (Q - \mathop{\mathbb{E}}_{X_r} Q)^2 \leqslant 6 \left\|\Sigma_r'^{1/2} Q \Sigma_r'^{1/2}\right\|_F^2 \leqslant \frac{6}{w_r^2}\|Q\|_F^2 \right\} .$$

Finally, using the bounded-variance constraints in $\mathcal{A}$, we have:

$$\mathcal{A} \left|\frac{Q,z}{4}\right. \mathbb{E}(Q - \mathop{\mathbb{E}}_{z} Q)^2 \leqslant \frac{6}{\alpha^2} \|Q\|_F^2 \ .$$

Plugging these estimates back in (4.4.8) yields:

$$\mathcal{A} \left|\frac{z}{4}\right. \left\{ z'(X_r)^2 z'(X_{r'})^2 \leqslant \frac{(1000Ct)^{2t}}{\alpha^{2t} \|Q\|_F^{2t}} \right\} \ .$$

Plugging in the lower bound on $\|Q\|_F^{2t}$ and applying Fact 4.4.9 completes the proof. $\square$

We can use the above lemma to complete the proof of Lemma 4.4.6:

Proof of Lemma 4.4.6. WLOG, assume that $\Sigma = I$. Let $Q = \Sigma_r - \Sigma_{r'}$ and let $\bar{Q} = Q/\|Q\|_F$. Consider the numbers $v_i = \mathrm{tr}(\Sigma_r \cdot Q)$. Then, we know that $\max_{i,j} |v_i - v_j| \geqslant \|Q\|_F$. Thus, there must exist a partition of $[k] = S \cup T$ such that $|v_i - v_j| \geqslant \|Q\|_F / k$ whenever $i \in S$ and $j \in T$.

Thus, for every $i \in S$ and $j \in T$, $\left\|\Sigma_i - \Sigma_j\right\|_F^2 \geqslant \|Q\|_F^2 / k^2 = 10^8 \frac{C^6 t^4}{(\beta^{2/t} \alpha^2)}$. We can now apply Lemma above to every $i \in S, j \in T$, observe that $\mathcal{A} \left|\frac{}{4}\right. \{z(X_r) z(X_{r'}) \leqslant z'(X_r) z'(X_{r'}) + 2\varepsilon/\alpha\}$, and add up the resulting inequalities to finish the proof.

$\square$

4.4.5 Special Case: Algorithm for Uniform and Bounded Mixing Weights

In this subsection, we obtain a polynomial time algorithm when the input mixture has weights that are bounded from below. This includes the case of uniform weights and when the minimum mixing weight is at least some function of k. At a high level, our algorithm partitions the sample into clusters as long as there is a pair of components separated in TV distance and given samples that are not clusterable, runs the tensor decomposition algorithm to list decode. We then use standard robust tournament results to pick a hypothesis from the list.

Theorem 4.4.12 (Robustly Learning Mixtures of Gaussians with Bounded Weights)**.** *Given $0 < \varepsilon < O_k(1)$, let $Y = \{y_1, y_2, \dots, y_n\}$ be a multiset of $n \geqslant n_0 = \mathrm{poly}_k(d, 1/\varepsilon)$ ε-corrupted samples from a k-mixture of Gaussians $\mathcal{M} = \sum_{i \leqslant k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, such that $w_i \geqslant \alpha$. Then, there exists an algorithm with running time $\mathrm{poly}_k(n^{1/\alpha}) \cdot \exp(\mathrm{poly}_k(1/\alpha, 1/\varepsilon))$ such that with probability at least $9/10$ it outputs a hypothesis k-mixture of Gaussians $\widehat{\mathcal{M}} = \sum_{i \leqslant k} \hat{w}_i \mathcal{N}\left(\hat{\mu}_i, \hat{\Sigma}_i\right)$ such that $d_{TV}\left(\mathcal{M}, \widehat{\mathcal{M}}\right) = O_k(\varepsilon)$.*

Briefly, our algorithm simply does the following:

1. **Clustering via SoS**: Guess a partition of the mixture such that each component in the partition is not clusterable. Let the resulting partition have $t \leqslant k$ components. In parallel, try all possible ways to run Algorithm 4.4.4 repeatedly to obtain a partition of the samples, $\{\tilde{Y}_j\}_{j \in [t]}$ into exactly t components. For each such partition repeat the following.

2. **Robust Isotropic Transformation**: Run the algorithm corresponding to Lemma 4.6.6 on each set $\tilde{Y}_j$ to make the sample approximately isotropic. Grid search for weights over $[\alpha, 1/k]^k$ with precision α.

3. **List-Decoding via Tensor Decomposition**: Run Algorithm 4.3.2 on each $\tilde{Y}_j$. Concatenate the lists to obtain $\mathcal{L}$.

4. **Robust Tournament**: Run the tournament from Fact 4.2.34 over all the hypotheses in $\mathcal{L}$, and output the winning hypothesis.

Proof Sketch. Setting $\Delta = (k^{k^{O(k)}})$, it follows from Theorem 4.4.2 that we obtain a partition of Y into $\{\tilde{Y}_j\}_{j \in [t]}$, for some $t \in [k]$ such that $\tilde{Y}_j$ has at most $O(k\varepsilon/\alpha)$ outliers, $(1 - O(k\varepsilon/\alpha))$-fraction of samples from at least one component of the input mixture and the resulting samples are not Δ-separated (see Definition 4.4.1). It then follows from Lemma

4.6.6 that the mean μ_j and covariance Σ_j of $\tilde{Y}_j$ satisfy : a) $\left\|\mu_j\right\|_2 \leqslant O\left(\sqrt{\varepsilon}k^{1.5}/\alpha^{1.5}\right)$, b) $\left(1-\sqrt{\varepsilon}k^{1.5}/\alpha^{1.5}\right) I \preceq \Sigma_j \preceq \left(1-\sqrt{\varepsilon}k^{1.5}/\alpha^{1.5}\right) I$, and c) $\left\|\Sigma_j - I\right\|_F \leqslant O\left(\sqrt{\varepsilon}k^{1.5}/\alpha^{1.5}\right)$.

Each component, $\tilde{Y}_j$, of the partition can have at most k components. Assuming these correspond to $\{w_i^{(j)}, \mu_i^{(j)}\Sigma_i^{(j)}\}_{i\in[k]}$, observe, $\sum_{i\in[k]} w_i^{(j)}\Sigma_i^{(j)} + w_i^{(j)}\mu_i^{(j)}\left(\mu_i^{(j)}\right)^\top \preceq \left(1+\sqrt{\varepsilon}k^{1.5}/\alpha^{1.5}\right) I$. Thus, we have that $\left\|\mu_i^{(j)}\right\|_2^2 \leqslant \left(1+\sqrt{\varepsilon}k^{1.5}\right)/\alpha^{2.5}$ and combined with not being Δ-separated, it follows that for all $i' \in [k]$,

$$\begin{aligned}
\left\|\Sigma_{i'}^{(j)} - I\right\|_F = \left\|\Sigma_{i'}^{(j)} - \Sigma_j + \left(\Sigma_j - I\right)\right\|_F &\leqslant \left\|\Sigma_{i'}^{(j)} - \sum_{i\in[k]} w_i^{(j)}\Sigma_i^{(j)} + \sum_{i\in[k]} w_i^{(j)}\mu_i^{(j)}\left(\mu_i^{(j)}\right)^\top\right\|_F + \left\|\Sigma_j - I\right\|_F \\
&\leqslant \left\|\sum_{i\in[k]} w_i^{(j)}\left(\Sigma_{i'}^{(j)} - \Sigma_i^{(j)}\right)\right\|_F + O\left(k^{1.5}/\alpha^{2.5}\right) \\
&\leqslant O(\Delta/\alpha)\ .
\end{aligned}$$

There are at most $O\left(k^k\right)$ ways in which we can partition the set of input points such that each resulting component is not partially clusterable. We run the algorithm in parallel for each one. Then, for the correct iteration, we apply Theorem 4.3.1 to get a list $\mathcal{L}$ of size $\exp(\mathrm{poly}_k(1/\alpha, 1/\varepsilon))$ such that it contains a hypothesis $\{\hat{w}_i^{(j)}, \hat{\mu}_i^{(j)}, \hat{\Sigma}_i^{(j)}\}_{i\in[k]}$ such that $\left|\hat{w}_i^{(j)} - w_i^{(j)}\right| \leqslant \alpha$, $\left\|\hat{\mu}_i^{(j)} - \mu_i^{(j)}\right\|_2 \leqslant O_k(\varepsilon)$ and $\left\|\hat{\Sigma}_i^{(j)} - \Sigma_i^{(j)}\right\|_F \leqslant O_k(\varepsilon)$. Since $(1-1/\Delta) I \preceq \Sigma_i^{(j)}$, it then follows from Lemma 4.6.4 that the hypothesis is $O_k(\varepsilon)$-close to the input in total variation distance.

Algorithm 4.4.4 is called at most $O\left(k^k\right)$ times, and along with the robust isotropic transformation, this requires $\mathrm{poly}_k\left(n^{1/\alpha}, 1/\varepsilon\right)$. The grid search contributes a multiplicative factor of $(1/\alpha)^k$. The tensor decomposition algorithm and robust hypothesis section $\mathrm{poly}_k(n^{1/\alpha}) \cdot \exp\left(\mathrm{poly}_k(1/\alpha, 1/\varepsilon)\right)$. □

4.4.6 More Efficient Robust Partial Cluster Recovery

In this section, we prove the following upgraded partial clustering theorem. In contrast to Theorem 4.4.3, here we obtain a probability of success that is inverse exponential in k

instead of $1/\alpha$.

Theorem 4.4.13 (Robust Partial Clustering in Relative Frobenius Distance)**.** *Let* $0 \leqslant \varepsilon < \alpha/k \leqslant 1$ *and* $t \in \mathbb{N}$. *There is an algorithm with the following guarantees: Let* Y *be an* ε*-corruption of a sample* X *of size* $n \geqslant (dk)^{Ct}/\varepsilon$ *for a large enough constant* $C > 0$*, from* $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ *that is a good sample (satisfies Condition 4.2.29) with parameters* $2t$ *and* $\gamma \leqslant \varepsilon d^{-8t} k^{-Ck}$*, for a large enough constant* $C > 0$*. Suppose further that* $w_i \geqslant \alpha > 2\varepsilon$ *for each* $i \in [k]$*, and that for some* $t \in \mathbb{N}$*,* $\beta > 0$ *there exist* $i, j \leqslant k$ *such that* $\left\|\Sigma^{\dagger/2}(\Sigma_i - \Sigma_j)\Sigma^{\dagger/2}\right\|_F^2 = \Omega\left((k^2 t^4)/\beta^{2/t}\alpha^4\right)$*, where* Σ *is the covariance of the mixture* $\mathcal{M}$*. Then, for any* $\eta \gg \sqrt{\varepsilon/\alpha}$*, the algorithm runs in time* $n^{O(t)}$*, and with probability at least* $2^{-O(k)}(1 - O(\eta/\alpha - \sqrt{\eta}))$ *over the random choices of the algorithm, outputs a partition* $Y = Y_1 \cup Y_2$ *satisfying:*

1. ***Partition respects clustering:*** *for each* i*,* $\max\left\{\frac{1}{w_i n}|Y_1 \cap X_i|, \frac{1}{w_i n}|Y_2 \cap X_i|\right\} \geqslant 1 - O(\sqrt{\eta}) - O(\frac{\beta}{\eta\alpha^2})$*, where* $X_i \subset X$ *corresponding to the points drawn from* $\mathcal{N}(\mu_i, \Sigma_i)$*.*
2. ***Partition is non-trivial:*** $\max_i \frac{1}{w_i n}|X_i \cap Y_1|, \max_i \frac{1}{w_i n}|X_i \cap Y_2| \geqslant 1 - O(\sqrt{\eta}) - O(\frac{\beta}{\eta\alpha^2})$*.*

Our algorithm will solve SoS relaxations of a polynomial inequality system. The indeterminates in this system are X' (that is intended to be the guess for the original uncorrupted sample), a cluster of size αn within X' (indicated by z_is) with mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$ and Π (intended to be the square root of $\hat{\Sigma}$). The input corrupted sample Y is a constant in this inequality system. Let $U \in \mathbb{R}^{d \times d}$ and $m, z \in \mathbb{R}^d$ also be indeterminates of the proof system. The system can be thought of as encoding the task of finding clusters $\hat{C}$ within Y that satisfies certifiable hypercontractivity of degree 2 polynomials.

We present the constraints grouped together into meaningful categories below: The first set of constraints enforce that $\hat{\Sigma}$ is the square of Π.

$$\text{Covariance Constraints: } \mathcal{A}_1 = \left\{ \begin{array}{c} \Pi = UU^\top \\ \Pi^2 = \hat{\Sigma} \end{array} \right\} \tag{4.4.9}$$

The intersection constraints force that X' intersects Y in all but an εn points (and thus, 2ε-close to unknown sample X).

$$\text{Intersection Constraints: } \mathcal{A}_2 = \left\{ \begin{array}{lr} \forall i \in [n], & m_i^2 = m_i \\ & \sum_{i \in [n]} m_i = (1-\varepsilon)n \\ \forall i \in [n], & z_i(\tilde{y}_i - x_i') = 0 \end{array} \right\} \tag{4.4.10}$$

The subset constraints enforce that z indicate a subset of size αn of X'.

$$\text{Subset Constraints: } \mathcal{A}_3 = \left\{ \begin{array}{lr} \forall i \in [n]. & z_i^2 = z_i \\ & \sum_{i \in [n]} z_i = \alpha n \end{array} \right\} \tag{4.4.11}$$

Parameter constraints create indeterminates to stand for the covariance $\hat{\Sigma}$ and mean $\hat{\mu}$ of $\hat{C}$ (indicated by z).

$$\text{Parameter Constraints: } \mathcal{A}_4 = \left\{ \begin{array}{r} \frac{1}{\alpha n} \sum_{i=1}^{n} z_i \left(x_i' - \hat{\mu}\right) \left(x_i' - \hat{\mu}\right)^\top = \hat{\Sigma} \\ \frac{1}{\alpha n} \sum_{i=1}^{n} z_i x_i' = \hat{\mu} \end{array} \right\} \tag{4.4.12}$$

Certifiable Hypercontractivity : $\mathcal{A}_4$=

$$\left\{ \begin{array}{l} \forall t \leqslant 2s \quad \mathop{\mathbb{E}}_z (Q - \mathop{\mathbb{E}}_z Q)^{2t} \leqslant (Ct/\alpha)^t 2^{2t} \left(\mathop{\mathbb{E}}_z (Q - \mathop{\mathbb{E}}_z Q)^2 \right)^t \\ \qquad \mathop{\mathbb{E}}_z (Q - \mathop{\mathbb{E}}_z Q)^2 \leqslant 10 \left(\frac{1}{\alpha}\right)^2 \|Q\|_F^2 \end{array} \right\} \tag{4.4.13}$$

where we write $\mathbb{E}_z Q$ as a shorthand for the polynomial $\frac{1}{\alpha^n} \sum_i z_i Q(x_i)$ and $\mathbb{E}_z(Q -$

$\mathbb{E}_{X_r} Q)^{2j}$ for the polynomial $\frac{1}{\alpha n}\sum_i z_i \left(Q(x_i') - \frac{1}{\alpha n}\sum_{i \leqslant n} z_i Q(x_i')\right)^{2j}$ for any j. Note that Q is a $d \times d$-matrix valued indeterminate. Observe that Q itself can be eliminated from the system as is standard in several applications [12, 25, 41, 43] of SoS proofs in obtaining a succinct set of polynomial constraints (see Section 4.3 on "Succinct Representation of Constraints" in [66] for an exposition).

Algorithm 4.4.14 (Polynomial Time Partial Clustering)**.**

Given: *A sample Y of size n. An outlier parameter $\varepsilon > 0$ and an accuracy parameter $\eta > 0$.*

Output: *A partition of Y into partial clustering $Y_1 \cup Y_2$.*

Operation:

1. ***Mean and Covariance Estimation****: Apply Robust Mean and Covariance Estimation (Fact 2.2.2) to estimate $\hat{\mu}$ and $\tilde{\Sigma}$ such that $\frac{1}{2}\Sigma \preceq \tilde{\Sigma} \preceq 1.5\Sigma$ where Σ is the covariance of the uncorrupted input mixture.*

2. ***Approximate Isotropic Transformation****: For each $y_i \in Y$, let $\tilde{y}_i = \tilde{\Sigma}^{\dagger/2}(y_i - \hat{\mu})$. Let $\tilde{Y} = \cup_{i \leqslant n} \tilde{y}_i$.*

3. ***SDP Solving:*** *Find a pseudo-distribution $\tilde{\zeta}$ satisfying $\cup_{i=1}^4 \mathcal{A}_i$ such that $\tilde{\mathbb{E}}_{\tilde{\zeta}} z_i \in \alpha \pm o_d(1)$ for every i. If no such pseudo-distribution exists, output fail.*

4. ***Rounding:*** *Let $M = \tilde{\mathbb{E}}_{z \sim \tilde{\zeta}}[zz^\top]$.*

   (a) ***Generate candidate clusters****: For $\ell = O(1/\alpha \log \eta/\alpha)$ times, draw a uniformly random $i \in [n]$ and let $\hat{C}_i = \{j \mid M(i,j) \geqslant \alpha^2/2\}$. Let $\mathcal{L} = \cup_{i \leqslant \ell} \hat{C}_i$.*

(b) ***Candidate 2nd Moment Estimation:*** *For each $\hat{C}_i \in \mathcal{L}$, let S_i be the output of running robust 2nd moment estimation with Frobenius error (Lemma 4.4.19) on $\hat{C}_i$ with outlier parameter $\eta'_i = O(\frac{\varepsilon}{\alpha} + \frac{\beta}{\alpha^2 \eta})$.*

(c) ***Merge candidate clusters****: For each $i \leqslant \ell$, find $\mathcal{L}_i$ to be the collection of all j such that $\left\|S_i - S_j\right\|_F \leqslant 2C\tau$ for a large enough constant $C > 0$. Set $\hat{C}_i \cup \mathcal{L}_i = \hat{B}_i$. Repeat on $\mathcal{L} \setminus \{\mathcal{L}_i \cup i\}$.*

(d) ***Output a union of a random subset of candidates****: For $\mathcal{L}' = \cup_i \hat{B}_i$, choose a uniformly random subset S of $\mathcal{L}'$, set $Y_1 = \cup_{j \in S} \hat{B}_j$ and set $Y_2 = Y \setminus Y_1$.*

Analysis of Algorithm

Lemma 4.4.15 (Success of Step 1)**.** *Let $\tilde{\Sigma}$ be the output of the robust covariance estimation algorithm (Fact 2.2.2) applied to the input sample Y with outlier parameter ε. If Y is an ε-corruption of a good sample X from a GMM with minimum weight $\geqslant \alpha \geqslant \Omega(\sqrt{\varepsilon})$, mixture mean μ and covariance Σ satisfying Condition 4.2.29, then,*

$$0.5\Sigma \preceq \tilde{\Sigma} \preceq 1.5\Sigma\,,$$

$$\left\|\tilde{\Sigma}^{-1/2}(\mu - \hat{\mu})\right\|_2 \leqslant O(\sqrt{\varepsilon}/\alpha)\,.$$

Proof. The lemma immediately follows by noting that GMMs with minimum weight α are 4-certifiably $1/\alpha$-subgaussian (Fact 4.2.12) and $\alpha \geqslant \Omega(\sqrt{\varepsilon})$.

□

Lemma 4.4.16 (Simultaneous Intersection Bounds for Frobenius Separated Case)**.** *Let $X = X_1 \cup X_2 \cup \ldots X_k$ be a sample of size $n \geqslant (dk)^{Ct}/\varepsilon$ for a large enough constant $C > 0$, from $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ that is a good sample (satisfies Condition 4.2.29) with parameters $2t$ and $\gamma \leqslant \varepsilon d^{-8t} k^{-Ck}$. Suppose further that $\|\mu_i\|_2 \leqslant \frac{2}{\alpha}$ for every i, $\|\Sigma_i\|_2 \leqslant \frac{1}{\alpha}$ for every i and the mixture mean μ, covariance Σ satisfy $\|\mu\|_2 \leqslant 1$ and $0.5I \preceq \Sigma \preceq 1.5I$.*

Let $\tau = 10^8 \frac{C^6 t^4}{\beta^{2/t}\alpha^2}$, for any $\beta > 0$. Then, given any ε-corruption Y of X, for every i, j such that $\left\|\Sigma_i - \Sigma_j\right\|_F^2 \geqslant \Omega(\tau)$,

$$\left\{\cup_{i=1}^4 \mathcal{A}_i\right\} \Big|\frac{z}{2t} \left\{z'(X_i) z'(X_j) \leqslant \beta\right\} ,$$

where $z'(X_i) = \frac{1}{w_i n} \sum_{j \in X_i} z_j \mathbf{1}(x_j = y_j)$ for every i.

Proof of Theorem 4.4.13. First, since Y is an ε-corruption of a sample X from a GMM such that X is a good sample (satisfies Condition 4.2.29), our robust mean and covariance estimation procedure (Step 1) applied to the mixture succeeds and recovers an estimate of the covariance that is multiplicative 1 ± 0.5-factor approximation in Löwner order. Thus, for the rest of analysis, we can assume that the smallest and largest eigenvalue of the mixture covariance are at least 0.5 and at most 1.5. Since each component has weight at least α, this means that each of the constitute component covariance can now be assumed to have a spectral norm at most $1.5/\alpha$.

Next, by an argument similar to the one presented in the proof of Theorem 4.4.2, the convex program we wrote is approximately solvable in polynomial time and is feasible whenever the uncorrupted sample X is a good sample (satisfies Condition 4.2.29). The only change here is in the certifiable hypercontractivity constraints where instead of the RHS of the bounded variance constraint is stated in terms of $\|Q\|_F^2$ instead of $\|\Pi Q \Pi\|_F^2$ with an additional slack of $O(1/\alpha^2)$. This modified constraint is satisfied by all true clusters by an application of Lemma 3.3.9 since each of their covariance has spectral norm at most $1.5/\alpha$.

Rounding Let $M = \tilde{\mathbb{E}}_{\zeta} z z^\top$. Then, by an argument similar to the proof of Theorem 4.4.2, we can conclude:

1. $o_d(1) + \alpha \geqslant M(i, j) \geqslant 0$.

2. $\sum_{j=1}^n M(i, j) \geqslant (\alpha^2 - o_d(1))n$ for every i.

3. For every i, let B_i be the set of "large entries": i.e. j such that $M(i,j) \geqslant \alpha^2/2$. Then, $|B_i| \geqslant \alpha n/2$.

In the following, let M_i denote the i-th row of M and $\|M_i\|_1$ for the sum of the non-negative entries of the vector M_i.

Candidate Clusters For every i, let $F_i \subseteq [k]$ be the set of all $i' \in [k]$ such that $\|\Sigma_i - \Sigma_{i'}\|_F^2 \geqslant \tau$ (i.e., F_i is the set of indices of true clusters whose covariances are far from that of the i-th cluster in Frobenius norm). For every row $j \in [n]$, let $C(j) \in [k]$ be such that $j \in X_{C(j)}$ Let's call j-th row of M "good" if $x_j = y_j$ (i.e j-th sample is not an outlier) and the following condition holds:

$$\sum_{r \in F_{C(j)}} \sum_{\ell \in X_r : x_\ell = y_\ell} M(j,\ell) \leqslant \left\|M_j\right\|_1 \left(\frac{\beta}{\eta}\right) .$$

Thus, by Markov's inequality, the fraction of non-outlier entries in B_j that come from $X_{r'}$ such that $r' \in F_r$ is at most $2\left(\frac{\beta}{\eta\alpha^2}\right)$.

Let us estimate the fraction of good rows now. From Lemma 4.4.6 and Fact 3.3.2, we have that for every r and $r' \in F_r$:

$$\tilde{\mathbb{E}}[z'(X_r) z'(X_{r'})] \leqslant \beta .$$

Here, recall that $z'(X_r) = \frac{1}{w_r n} \sum_{i \leqslant n} z_i \mathbf{1}(y_i = x_i)$ for every r. Summing up over $r' \in F_r$ yields:

$$\frac{1}{w_r n} \sum_{r' \in F_r} \sum_{i \in X_r : x_i = y_i} \sum_{j \in X_{r'} : x_j = y_j} \tilde{\mathbb{E}}[z_i z_j] \leqslant n\beta .$$

Thus, by Markov's inequality, with probability at least $1-\eta$ over the choice of $i \in X_r$ such that $x_i = y_i$, it must hold that:

$$\sum_{r' \in F_r} \sum_{j \in X_{r'} : x_j = y_j} \tilde{\mathbb{E}}[z_i z_j] \leqslant n\left(\frac{\beta}{\eta}\right) .$$

Using that $(1-\varepsilon/\alpha)$-fraction of $i \in X_r$ satisfy $x_i = y_i$, for every r, we conclude that $1-\eta-\varepsilon/\alpha$-fraction of the rows X_r are good.

Thus, with probability at least $(1-\eta-\varepsilon/\alpha)^\ell \geqslant (1-O(\ell(\eta+\varepsilon/\alpha)))$, every candidate cluster picked in Step 1 of our rounding algorithm corresponds to the large entries from a good row of M.

We next claim that we cover most of the points in the input in the union of the candidate clusters:

$$|\cup_{i\leqslant \ell} \hat{C}_i| \geqslant \left(1-2\sqrt{\eta}-\frac{\varepsilon}{\sqrt{\eta}\alpha}\right) n$$

with probability at least $1-\sqrt{\eta}$. To see why, let's estimate the chance that an element $j \in [n]$ does not appear in any of the $\hat{C}_i$s. First, we can assume that j-th row of M is good (this loses us $\eta+\varepsilon/\alpha$-fraction js). For each such j, there are at least $\alpha n/2$ large entries. Since M is symmetric, the j-th column of M also has $\alpha n/2$ large entries. Further, j appears in $\cup_{i\leqslant\ell}\hat{C}_i$ if at least one of the $\alpha n/2$ large entries are chosen in our rounding. The chance that this does not happen in any of the ℓ picks is at most $(1-\alpha/2)^\ell$. Since $\ell = \Theta\left(\frac{1}{\alpha}\log(1/\eta)\right)$, this chance is at most $O(\eta)$. Thus, in expectation $|[n] \setminus \cup_{i\leqslant\ell}\hat{C}_i| \leqslant O\left(\eta+\varepsilon/\alpha\right) n$. By Markov's inequality, with probability at least $1-\sqrt{\eta}$, $|[n] \setminus \cup_{i\leqslant\ell}\hat{C}_i| \leqslant O\left(\sqrt{\eta}+\frac{\varepsilon}{\sqrt{\eta}\alpha}\right) n$.

By a union bound, with probability at least $1-O\left(\eta\ell-\varepsilon\ell/\alpha\right)-\sqrt{\eta} \geqslant 1-O\left(\eta\log(1/\eta)/\alpha-\sqrt{\eta}\right)$, we must thus have both the following events hold simultaneously:

$$|\cup_{i\leqslant \ell} \hat{C}_i| \geqslant \left(1-2\sqrt{\eta}-\frac{\varepsilon}{\sqrt{\eta}\alpha}\right) n \geqslant (1-3\sqrt{\eta})n$$

and, for every $1 \leqslant i \leqslant \ell$,

$$|\hat{C}_i \cap (\cup_{r' \in F_{C(r)}} X_{r'})| \leqslant 2\left(\frac{\beta}{\eta\alpha}+\varepsilon/\alpha\right)\cdot|\hat{C}_i|\,.$$

Merging Candidate Clusters Observe, following the proof of Theorem 4.4.3, we know that there exists a partition of Y into sets Y_1 and Y_2 such that for all i,

$$\max\left\{\frac{1}{w_i n}|Y_1 \cap X_i|, \frac{1}{w_i n}|Y_2 \cap X_i|\right\} \geqslant 1 - O(\sqrt{\eta}) - O(\frac{\beta}{\eta\alpha^2}),$$

and

$$\max_i \frac{1}{w_i n}|X_i \cap Y_1|, \max_i \frac{1}{w_i n}|X_i \cap Y_2| \geqslant 1 - O(\sqrt{\eta}) - O(\frac{\beta}{\eta\alpha^2}).$$

Next, we show that the merging step preserves this partition. For each $\hat{C}_i$, let $\hat{C}'_i = \hat{C}_i \cap \cup_{j \notin F_{C(i)}} X_j$. That is, $\hat{C}'_i$ is the subset of $\hat{C}_i$ obtained by removing points from "far-off" clusters and the outliers. Then, since we know that $|\hat{C}_i| \geqslant \alpha n/2$ and $|X \cap Y| \geqslant (1-\varepsilon)n$, we must have $|\hat{C}_i| - |\hat{C}'_i| = \eta'_i|\hat{C}_i| \leqslant \left(\frac{3\varepsilon}{\alpha} + \frac{2\beta}{\eta\alpha^2}\right)|\hat{C}_i|$, where we note that $\eta'_i \leqslant \left(\frac{3\varepsilon}{\alpha} + \frac{2\beta}{\eta\alpha^2}\right)$.

Thus, $\hat{C}'_i$ is a collection of $\geqslant (1-\eta'_i)\alpha n/2$ points from the submixture $\cup_{j \notin F_{C(i)}} X_j$. We know that each μ_i is of ℓ_2 norm at most $1/\alpha$, each Σ_i has spectral norm at most $1/\alpha$ and that for every $r, r' \notin F_{C(i)}$, $\|\Sigma_r - \Sigma_{r'}\|_F^2 \leqslant \tau$. Further, Σ_r is at most $\tau + 1/\alpha = O(\tau)$-different in Frobenius norm from the covariance of the sub-mixture. By an argument similar to the proof of Lemma 4.2.26, we can establish that the submixture with components r such that $r \notin F_{C(i)}$ is $O(\tau)$-certifiably bounded variance. Since $\hat{C}'_i$ is a subset of this sub-mixture of size $\alpha n/2$, we immediately obtain that $\hat{C}'_i$ is $O(\tau/\alpha)$-certifiably bounded variance. Thus, applying Lemma 4.4.19 with outlier parameter η'_i to input $\hat{C}_i$ yields an estimate S_i of the 2nd moment of $\hat{C}'_i$ within a Frobenius error of at most $O(\tau/\alpha)$ From Lemma 4.4.20, this is an additional $O(1/\alpha)$ different in Frobenius norm from the 2nd moment of the sub-mixture which, as argued above, is itself at most $O(\tau)$ different in Frobenius norm from Σ_i. Chaining together yields that $\|\Sigma_i - S_i\|_F^2 \leqslant O(\tau/\alpha)$ for some constant C.

Since for every $r \in S, r' \in T$ it holds that $\|\Sigma_r - \Sigma_{r'}\|_F^2 \gg \Omega(\tau/\alpha)$, conditioned on the good event above, our algorithm never merges $\hat{C}_i$ and $\hat{C}_j$ whenever i, j are non-outliers and i is in some cluster in S and j is in some cluster in T. On the other hand, if i, j belong to the same cluster, then, the corresponding estimate $\left\|S_i - S_j\right\|_F^2 \leqslant 2C\tau$. Thus, our merging process always merges together any such candidates.

As a result, the output of the merging process can have at most one i from any true cluster – thus, the number of distinct members of $\mathcal{L}'$ is at most k. We note that the running time is dominated by computing a pseudo distribution satisfying the union of all the constraints (Step 3 in Algorithm 4.4.14) and requires $n^{O(t)}$ time. Step 4 computes a degree $O(1)$ sos relaxation for at most $O(\ell)$ components and the merging only requires a fixed polynomial in d and k time.

□

4.4.7 Proof of Lemma 4.4.16

In the following lemma, we show that the constraint system $\mathcal{A}$, via a low-degree sum-of-squares proof, implies that a lower bound on the variance of any degree 2 polynomial on X' whenever the cluster $\hat{C}$ (indicated by z) appreciably intersects two well-separated true clusters.

Lemma 4.4.17 (Lower-Bound on Variance of Degree 2 Polynomials)**.** *Let $Q \in \mathbb{R}^{d\times d}$ be any fixed matrix. Then, for any $i, j \leqslant k$, and $z'(X_r) = \frac{1}{w_r n}\sum_{i\in X_r} z_i \cdot \mathbf{1}(y_i = x_i)$, we have for any $r \neq r' \in [k]$,*

$$\mathcal{A} \Big|\frac{z}{4t} \left\{ z'(X_r)z'(X_{r'}) \leqslant \frac{(32Ct/\alpha)^{2t}}{(\mathbb{E}_{X_r} Q - \mathbb{E}_{X_{r'}} Q)^{2t}} \left(\frac{\alpha^4}{w_r^2 w_{r'}^2} \left(\mathbb{E}_z (Q - \mathbb{E}_z Q)^2\right)^t \right.\right.$$
$$\left.\left. + \frac{\alpha^2}{w_r^2}\left(\mathbb{E}_{X_{r'}}(Q - \mathbb{E}_{X_{r'}} Q)^2\right)^t + \frac{\alpha^2}{w_{r'}^2}\left(\mathbb{E}_{X_r}(Q - \mathbb{E}_{X_r} Q)^2\right)^t \right) \right\}.$$

Proof. Let $z_i' = z_i \mathbf{1}(y_i = x_i)$ for every i. For every $1 \leqslant r \leqslant k$, let $\mathbb{E}_{X_r} Q$ denote the expectation of the homogenous degree 2 polynomial defined by Q: $\mathbb{E}_{X_r} Q = \frac{1}{w_r n}\sum_{i,j\in X_r} Q(x_i)$ for every r where $Q(x_i) = x_i^\top Q x_i$. Similarly, let $\mathbb{E}_z Q$ be the quadratic polynomial in z defined by $\mathbb{E}_z Q = \frac{1}{\alpha n}\sum_{i\leqslant n} z_i Q(x_i)$. Using the substitution rule and non-negativity of the

z_i's, we have for any $r, r' \in [k]$:

$$\mathcal{A} \Big|\frac{z}{4t} \left\{ \mathop{\mathbb{E}}_z (Q - \mathop{\mathbb{E}}_z Q)^{2t} = \frac{1}{\alpha n} \sum_{i \in [n]} z_i \left(Q(x_i) - \mathop{\mathbb{E}}_z Q \right)^{2t} \right.$$

$$\left. \geqslant \frac{1}{\alpha n} \sum_{i \in X_r \cup X_{r'} : x_i = y_j} z_i' \left(Q(x_i) - \mathop{\mathbb{E}}_z Q \right)^{2t} \right\}$$

Then, using the SoS almost triangle inequality (Fact 3.3.5), we have:

$$\mathcal{A} \Big|\frac{z}{4t} \left\{ \frac{1}{\alpha n} \sum_{i \in X_r \cup X_{r'}} z_i' \left(Q(x_i) - \mathop{\mathbb{E}}_z Q \right)^{2t} \right.$$

$$\geqslant 2^{-2t} \left(\frac{1}{\alpha n} \sum_{i \in X_r : i} z_i' \left(\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_z Q \right)^{2t} - \frac{1}{\alpha n} \sum_{i \in X_r} z_i' \left(Q(x_i) - \mathop{\mathbb{E}}_{X_r} Q \right)^{2t} \right)$$

$$+ 2^{-2t} \left(\frac{1}{\alpha n} \sum_{i \in X_r : i : x_i = y_i} z_i' \left(\mathop{\mathbb{E}}_{X_{r'}} Q - \mathop{\mathbb{E}}_z Q \right)^{2t} - \frac{1}{\alpha n} \sum_{i \in X_{r'}, x_i = y_i} z_i' \left(Q(x_i) - \mathop{\mathbb{E}}_{X_{r'}} Q \right)^{2t} \right)$$

$$= 2^{-2t} \left(\frac{w_r}{\alpha} z'(X_r) \left(\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_z Q \right)^{2t} - \frac{1}{\alpha n} \sum_{i \in X_r} \left(Q(x_i) - \mathop{\mathbb{E}}_{X_r} Q \right)^{2t} \right)$$

$$\left. + 2^{-2t} \left(\frac{w_{r'}}{\alpha} z'(X_{r'}) \left(\mathop{\mathbb{E}}_{X_{r'}} Q - \mathop{\mathbb{E}}_z Q \right)^{2t} - \frac{1}{\alpha n} \sum_{i \in X_{r'}} \left(Q(x_i) - \mathop{\mathbb{E}}_{X_{r'}} Q \right)^{2t} \right) \right\}$$

Next, observe that by the SoS almost triangle inequality (Fact 3.3.5), we must have:

$$\mathcal{A} \Big|\frac{}{4t} \left\{ \left(\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_z Q \right)^{2t} + \left(\mathop{\mathbb{E}}_{X_{r'}} Q - \mathop{\mathbb{E}}_z Q \right)^{2t} \geqslant 2^{-2t} \left(\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_{X_{r'}} Q \right)^{2t} \right\} .$$

Further, note that $\mathcal{A} \Big|\frac{}{O(1)} \left\{ \frac{w_r}{\alpha} z'(X_r) + \frac{w_{r'}}{\alpha} z'(X_{r'}) \leqslant \frac{1}{\alpha n} \sum_i z_i \leqslant 1 \right\}$. Thus, using Fact 4.4.8 with $A = \frac{w_r}{\alpha} z'(X_r)$, $B = \frac{w_{r'}}{\alpha} z'(X_{r'})$, $C = (\mathbb{E}_{X_r} Q - \mathbb{E}_z Q)^{2t}$, and $D = \left(\mathbb{E}_{X_{r'}} Q - \mathbb{E}_z Q\right)^{2t}$ and $\tau = 2^{-2t} \left(\mathbb{E}_{X_r} Q - \mathbb{E}_{X_{r'}} Q\right)^{2t}$, we can derive:

$$
\begin{aligned}
\mathcal{A} \,\Big|\!\frac{z}{4}\, \Bigg\{ (Ct/\alpha)^{2t} \left(\mathop{\mathbb{E}}_{z}(Q - \mathop{\mathbb{E}}_{z} Q)^2\right)^t &\geqslant \mathop{\mathbb{E}}_{z}(Q - \mathop{\mathbb{E}}_{z} Q)^{2t} \geqslant 2^{-6t} \frac{w_r w_{r'}}{\alpha^2} z'(X_r) z'(X_{r'}) \left(\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_{X_{r'}} Q\right)^{2t} \\
&\quad - 2^{-6t} \frac{w_r}{\alpha} \mathop{\mathbb{E}}_{X_r} \left(Q - \mathop{\mathbb{E}}_{X_r} Q\right)^{2t} - 2^{-6t} \frac{w_{r'}}{\alpha} \mathop{\mathbb{E}}_{X_{r'}} \left(Q - \mathop{\mathbb{E}}_{X_{r'}} Q\right)^{2t} \\
&\geqslant 2^{-6t} \frac{w_r w_{r'}}{\alpha^2} z'(X_r) z'(X_{r'}) (\mathop{\mathbb{E}}_{X_r} Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^{2t} - \frac{w_r}{\alpha} (Ct/\alpha)^{2t} \left(\mathop{\mathbb{E}}_{X_r} \left(Q - \mathop{\mathbb{E}}_{X} \right.\right. \\
&\quad - \frac{w_{r'}}{\alpha} (Ct/\alpha)^{2t} \left(\mathop{\mathbb{E}}_{X_{r'}} \left(Q - \mathop{\mathbb{E}}_{X_{r'}} Q\right)^2\right)^t \Bigg\}
\end{aligned}
$$

where the first inequality uses the Certifiable Hypercontractivity constraints ($\mathcal{A}_4$) and the last inequality follows from the Certifiable Hypercontractivity of X_r and $X_{r'}$ (Condition 4.2.29). Rearranging completes the proof.

□

We can use the lemma above to obtain a simultaneous intersection bound guarantee when there are relative Frobenius separated components in the mixture.

Lemma 4.4.18 (Lemma 4.4.6, restated)**.** *Suppose* $\left\|\Sigma^{-1/2}(\Sigma_r - \Sigma_{r'})\Sigma^{-1/2}\right\|_F^2 \geqslant 10^8 \frac{C^6 t^4}{\beta^{2/t} \alpha^4}$. *Then, for* $z(X_r) = \frac{1}{w_r n} \sum_{i \in X_r} z_i \cdot \mathbf{1}(y_i = x_i)$*,*

$$
\mathcal{A} \,\Big|\!\frac{}{2t}\, \{ z(X_r) z(X_{r'}) \leqslant \beta \} \ .
$$

Proof. WLOG, we will work with the transformed points $x_i \to \Sigma^{-1/2} x_i$ where Σ is the covariance of the mixture. Note that our algorithm does not need to know Σ – this transformation is only for simplifying notation in the analysis that follows.

Let $\tilde{\Sigma}_z = \Sigma^{-1/2} \Sigma_z \Sigma^{-1/2}$, $\tilde{\Sigma}_r = \Sigma^{-1/2} \Sigma_r \Sigma^{-1/2}$ and $\tilde{\Sigma}_{r'} = \Sigma^{-1/2} \Sigma_{r'} \Sigma^{-1/2}$ be the transformed covariances. Then, notice that $\left\|\tilde{\Sigma}_r\right\|_2 \leqslant \frac{1}{w_r} \left\|\Sigma\right\|_2 \leqslant \frac{1.5}{w_r}$ and $\left\|\tilde{\Sigma}_{r'}\right\|_2 \leqslant \frac{1}{w_{r'}} \left\|\Sigma\right\|_2 \leqslant \frac{1.5}{w_{r'}}$.

We now apply Lemma 4.4.10 with $Q = \tilde{\Sigma}_r - \tilde{\Sigma}_{r'}$. Then, notice that $\mathbb{E}_{X_r} Q - \mathbb{E}_{X_{r'}} Q =$

$\left\|\tilde{\Sigma}_r - \tilde{\Sigma}_{r'}\right\|_F^2 + \mu_r^\top(\tilde{\Sigma}_r - \tilde{\Sigma_{r'}})\mu_r - \mu_{r'}^\top(\tilde{\Sigma}_r - \tilde{\Sigma_{r'}})\mu_{r'} \geqslant \|Q\|_F^2 - \frac{4}{\alpha}$. Then, we obtain:

$$\mathcal{A} \left|\frac{z}{2t}\right. \Bigg\{ z(X_r)z(X_{r'}) \\ \leqslant \left(\frac{32Ct/\alpha}{\mathbb{E}_{X_r} Q - \mathbb{E}_{X_{r'}} Q}\right)^{2t} \left(\frac{\alpha^2}{w_r w_{r'}}\left(\mathop{\mathbb{E}}_z(Q - \mathop{\mathbb{E}}_z Q)^2\right)^t + \frac{\alpha}{w_r}\left(\mathop{\mathbb{E}}_{X_{r'}}(Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^2\right)^t + \frac{\alpha}{w_{r'}}\left(\mathop{\mathbb{E}}_{X_r}(Q - \mathop{\mathbb{E}}_{X_r} Q)^2\right)\right)^t \Bigg\}. \tag{4.4.14}$$

Since X_r and $X_{r'}$ have certifiably C-bounded variance polynomials for $C = 4$ (as a consequence of Condition 4.2.29 and Fact 4.2.27 followed by an application of Lemma 3.3.9), we have:

$$\mathop{\mathbb{E}}_{X_{r'}}(Q - \mathop{\mathbb{E}}_{X_{r'}} Q)^2 \leqslant 6\left\|\tilde{\Sigma}_{r'}^{1/2} Q \tilde{\Sigma}_{r'}^{1/2}\right\|_F^2 \leqslant \frac{10}{w_{r'}^2}\|Q\|_F^2 \leqslant \frac{10}{\alpha^2}\|Q\|_F^2 \ ,$$

and

$$\mathop{\mathbb{E}}_{X_r}(Q - \mathop{\mathbb{E}}_{X_r} Q)^2 \leqslant 6\left\|\tilde{\Sigma}_r^{1/2} Q \tilde{\Sigma}_r^{1/2}\right\|_F^2 \leqslant \frac{10}{w_r^2}\|Q\|_F^2 \leqslant \frac{10}{\alpha^2}\|Q\|_F^2 \ .$$

Finally, using the bounded-variance constraints in $\mathcal{A}$, we have:

$$\mathcal{A} \left|\frac{Q,z}{4}\right. \mathbb{E}(Q - \mathop{\mathbb{E}}_z Q)^2 \leqslant \frac{10}{\alpha^2}\|Q\|_F^2 \ .$$

Plugging these estimates back in (4.4.14) yields:

$$\mathcal{A} \left|\frac{z}{4}\right. \left\{ z(X_r)z(X_{r'}) \leqslant \frac{(1000Ct/\alpha)^{2t}}{\|Q\|_F^{2t}\,\alpha^{2t}}\left(\frac{\alpha^2}{w_r} + \frac{\alpha}{w_{r'}} + \frac{\alpha}{w_r w_{r'}}\right) \leqslant \frac{3}{w_r w_{r'}} \frac{(1000Ct)^{2t}}{\alpha^{2t}\,\|Q\|_F^{2t}} \leqslant \frac{3(1000Ct)^{2t}}{\alpha^{2t}\,\|Q\|_F^{2t}} \right\}.$$

Plugging in the lower bound on $\|Q\|_F^{2t}$ and applying cancellation within SoS (Fact 4.4.9) completes the proof. □

4.4.8 2nd Moment Estimation Subroutine

The following lemma gives a 2nd moment estimation algorithm with error in Frobenius norm for distributions that have a certifiably bounded covariance. The proof is very similar to the SoS based mean and covariance estimation algorithms but we provide it in full for completeness here.

Lemma 4.4.19 (2nd Moment Estimation in Frobenius Norm)**.** *Let* $1/100 \geqslant \eta > 0$*. There is an* $n^{O(1)}$ *time algorithm that takes input an* η*-corruption* Y *of an sample* X *of size* n *and outputs an estimate* M_2 *of the 2nd moment of* X *with the following properties: Let* $X \subseteq \mathbb{R}^d$ *be a collection of* n *points satisfying* $\left|\frac{Q}{2}\right. \left\{\frac{1}{|X|}\sum_{x\in X}\left(Q(x) - \frac{1}{|X|}Q(x)\right)^2 \leqslant C\,\|Q\|_F^2\right\}$ *for a matrix-valued indeterminate* Q*. Let* $M_2 = \frac{1}{n}\sum_{x\in X} xx^\top$*. Then, the estimate* $\hat{M}_2$ *output by the algorithm satisfies:*

$$\left\|\hat{M}_2 - M_2\right\|_F^2 \leqslant 80C\eta\,.$$

Proof. Consider the constraint system with scalar-valued indeterminates z_i for $1 \leqslant i \leqslant n$ and d-dimensional vector-valued indeterminates $x'_1, x'_2, \ldots, x'_n$ with the following set of constraints: $\mathcal{A} =$

$$\left\{\begin{array}{ll} \forall i \leqslant n & z_i^2 = z_i \\ & \sum_{i=1}^n z_i = (1-\eta)n \\ & \tilde{M}_2 = \frac{1}{n}\sum_{i=1}^n x'_i x'^\top_i \\ \forall i \leqslant n & z_i x'_i = z_i y_i \\ & \frac{1}{n}\sum_{i=1}^n \left(x'^\top_i Q x'_i - \frac{1}{n}\sum_{i=1}^n x'^\top_i Q x'_i\right)^2 \leqslant C\,\|Q\|_F^2 \end{array}\right\} \tag{4.4.15}$$

Observe that $X' = X$ and z_i set to the 0-1 indicator of non-outliers satisfies the constraint system. Thus, the constraints are feasible.

Our algorithm finds a pseudo-distribution $\tilde{\zeta}$ of degree 10 satisfying the above constraints and output $\tilde{\mathbb{E}}[\tilde{M}_2]$. Let us now analyze this algorithm. The key is the following statement that gives a sum-of-squares proof of closeness of $\tilde{M}_2$ and M_2 in Frobenius norm. We use the notation $\mathbb{E}_X Q$ and $\mathbb{E}_{X'} Q$ to abbreviate $\frac{1}{n}\sum_{i=1}^n x_i^\top Q x_i$ and $\frac{1}{n}\sum_{i=1}^n {x'}_i^\top Q x'_i$ respectively.

$$
\begin{aligned}
\mathcal{A} \Big|\frac{Q}{2} \Bigg\{ & \left(\frac{1}{n}\sum_{i=1}^n x_i^\top Q x_i - \frac{1}{n}\sum_{i=1}^n {x'}_i^\top Q x'_i\right)^2 \\
&= \left(\frac{1}{n}\sum_{i=1}^n (1 - z_i \mathbf{1}(x_i = y_i)) x_i^\top Q x_i - {x'}_i^\top Q x'_i\right)^2 \\
&\leqslant \left(\frac{1}{n}(1 - z_i \mathbf{1}(x_i = y_i))^2\right)\left(\frac{1}{n}\sum_{i=1}^n \left(x_i^\top Q x_i - {x'}_i^\top Q x'_i\right)^2\right) \\
&\leqslant 20\eta \cdot \left(\frac{1}{n}\sum_{i=1}^n \left(x_i^\top Q x_i - \mathop{\mathbb{E}}_X Q\right)^2 + \frac{1}{n}\sum_{i=1}^n \left({x'}_i^\top Q x'_i - \mathop{\mathbb{E}}_{X'} Q\right)^2 + \left(\mathop{\mathbb{E}}_X Q - \mathop{\mathbb{E}}_{X'} Q\right)^2\right) \\
&\leqslant 20\eta (2C \|Q\|_F^2) + 20\eta \left(\mathop{\mathbb{E}}_X Q - \mathop{\mathbb{E}}_{X'} Q\right)^2 \Bigg\}
\end{aligned}
$$

where the first inequality follows by the SoS version of the Cauchy-Schwarz inequality and the 2nd by the SoS version of the Almost Triangle inequality.

Rearranging and using that $1 - 20\eta > 1/2$ now yields that:

$$
\mathcal{A} \Big|\frac{Q}{2} \left\{ \left(\frac{1}{n}\sum_{i=1}^n x_i^\top Q x_i - \frac{1}{n}\sum_{i=1}^n {x'}_i^\top Q x'_i\right)^2 \leqslant 80 C\eta \|Q\|_F^2 \right\}
$$

Notice that the LHS above equals the linear polynomial $\langle \tilde{M}_2 - M_2, Q\rangle$. We now plug in $Q = \tilde{M}_2 - M_2$ to obtain:

$$
\mathcal{A} \Big|\frac{Q}{2} \left\{ \left\|\tilde{M}_2 - M_2\right\|_F^4 \leqslant 80 C\eta \left\|\tilde{M}_2 - M_2\right\|_F^2 \right\}
$$

Applying Fact 3.3.8 yields:

$$\mathcal{A} \Big|\frac{Q}{2} \left\{ \left\| \tilde{M}_2 - M_2 \right\|_F^8 \leqslant 80^4 C^4 \eta^4 \right\}$$

Taking pseudo-expectations with respect to $\tilde{\zeta}$ and using Hölder's inequality for pseudo-distributions yields that

$$\left\| \tilde{\mathbb{E}}_{\tilde{\zeta}} \tilde{M}_2 - M_2 \right\|_F^8 \leqslant \tilde{\mathbb{E}}_{\tilde{\zeta}} \left\| \tilde{M}_2 - M_2 \right\|_F^8 \leqslant 80^4 C^4 \eta^4 \,.$$

Taking the 4-th root, we can conlcude our rounded value $\hat{M}_2 = \tilde{\mathbb{E}}_{\tilde{\zeta}} \tilde{M}_2$ satisfies:

$$\left\| \hat{M}_2 - M_2 \right\|_F^2 \leqslant 80 C \eta \,.$$

This completes the proof.

□

We also note the following simple consequence of the certifiable bounded variance property that follows via an argument similar to the one employed in the proof of the previous lemma.

Lemma 4.4.20 (Subsamples of Bounded-Variance Distributions)**.** *Let $X \subseteq \mathbb{R}^d$ be a collection of n points satisfying $\Big|\frac{Q}{2} \left\{ \frac{1}{|X|} \sum_{x \in X} (Q(x) - \frac{1}{|X|} Q(x))^2 \leqslant C \left\| Q \right\|_F^2 \right\}$ for a matrix-valued indeterminate Q. Let $M_2 = \frac{1}{|X|} \sum_{x \in X} x x^\top$ be the 2nd moment of X. Let $S \subseteq X$ be a subset of size at least $\beta |X|$. Then,*

$$\Big|\frac{}{4} \left\{ \left\| \frac{1}{|S|} \sum_{x \in S} x x^\top - M_2 \right\|_F^2 \leqslant \frac{1}{\beta} \right\} .$$

Proof. We have by the Cauchy-Schwarz inequality:

$$\left|\frac{Q}{2} \left\{ \left(\frac{1}{|X|} \sum_{x\in X} \mathbf{1}(x \in S)(x^\top Q x - M_2), Q\rangle) \right)^2 \leqslant \left(\frac{1}{|X|} \mathbf{1}(x \in S)^2 \right) \left(\sum_{x \in X} (x^\top Q x - M_2), Q\rangle)^2 \right) \right.\right.$$

$$\leqslant \left(\frac{|S|}{|X|} \right) \|Q\|_F^2 .$$

We now substitute in $Q = \frac{1}{|X|} \sum_{x\in X} \mathbf{1}(x \in S)(x^\top Q x - M_2$ to obtain:

$$\left|\frac{Q}{2} \left\{ \left\| \frac{1}{|X|} \sum_{x\in X} \mathbf{1}(x \in S)(x^\top Q x - M_2) \right\|_F^4 \leqslant \left(\frac{|S|}{|X|} \right) \left\| \frac{1}{|X|} \sum_{x\in X} \mathbf{1}(x \in S)(x^\top Q x - M_2) \right\|_F^2 .\right.\right.$$

We now apply Fact 3.3.8 to yield:

$$\left|\frac{Q}{2} \left\{ \left\| \frac{1}{|X|} \sum_{x\in X} \mathbf{1}(x \in S)(x^\top Q x - M_2) \right\|_F^8 \leqslant \left(\frac{|S|}{|X|} \right)^4 .\right.\right.$$

We finally apply Fact 4.4.9 to conclude that:

$$\left|\frac{Q}{2} \left\{ \left\| \frac{1}{|X|} \sum_{x\in X} \mathbf{1}(x \in S)(x^\top Q x - M_2) \right\|_F^2 \leqslant \left(\frac{|S|}{|X|} \right) .\right.\right.$$

Rescaling gives the claim. □

4.5 Spectral Separation of Thin Components

In this section, we show how to efficiently separate a thin component, if such a component exists, given sufficiently accurate approximations to the component means and covariances. This is an important step in our overall algorithm and is required to obtain total variation distance guarantees.

Specifically, the main algorithmic result of this section is described in the following lemma:

Lemma 4.5.1. *There is a polynomial-time algorithm with the following properties: Let $\mathcal{M} = \sum_{i=1}^k w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians on $\mathbb{R}^d$, and let X in $\mathbb{R}^d$ be a good sample (satisfying Condition 4.2.29) with respect to $\mathcal{M}$ for some parameters (γ, t). The algorithm takes input parameters η, δ, satisfying $0 < \delta < \eta < 1/(100k)$, and Y, an ε-corrupted version of X, as well as candidate parameters $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i \leqslant k}$. Then as long as*

1. *$\mathrm{Cov}(\mathcal{M}) \geq I/2$,*
2. *$\|\mu_i - \hat{\mu}_i\|_2 < \delta$ and $\|\Sigma_i - \hat{\Sigma}_i\|_F < \delta$, for all $i \in [k]$, and*
3. *there exists an $s \in [k]$ such that Σ_s has an eigenvalue $< \eta$,*

the algorithm outputs a partition of Y into $Y_1 \cup Y_2$ such that there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$, so that letting $\mathcal{M}_j$, $j \in \{1, 2\}$, be proportional to $\sum_{i \in Q_j} w_i G_i$ and $W_j = \sum_{i \in Q_j} w_i$, then Y_j is an $((O(k^2\gamma) + \tilde{O}(\eta^{1/2k}))/W_j)$-corruption of a good sample (satisfying Condition 4.2.29) with respect to $\mathcal{M}_j$ with parameters $(O(k\gamma/W_j), t)$.

The key component in the proof of Lemma 4.5.1 is the following lemma:

Lemma 4.5.2. *Let $\mathcal{M} = \sum_{i=1}^k w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians in $\mathbb{R}^d$ with $\mathrm{Cov}(\mathcal{M}) \geq I/2$. Suppose that, for some $0 < \delta < 1/(100k)$, we are given $\hat{\mu}_i$ and $\hat{\Sigma}_i$ satisfying $\|\mu_i - \hat{\mu}_i\|_2 < \delta$ and $\|\Sigma_i - \hat{\Sigma}_i\|_F < \delta$, for all $i \in [k]$. Suppose furthermore that for some $\eta > \delta$, there is a Σ_s, $s \in [k]$, with an eigenvalue less than η. There exists a computationally efficient algorithm that takes inputs η, δ, $\hat{\mu}_i$, $\hat{\Sigma}_i$, and computes a function $F : \mathbb{R}^d \to \{0, 1\}$ such that:*

1. *For each $i \in [k]$, $F(G_i)$ returns the same value in $\{0, 1\}$ with probability at least $1 - \tilde{O}_k(\eta^{1/(2k)})$. We define the most likely value of $F(G_i)$ to be this value.*
2. *There exist $i, j \in [k]$ such that the most likely values of $F(G_i)$ and $F(G_j)$ are different.*

Furthermore, $F(x)$ can be chosen to be of the form $f(v \cdot x)$, for some $v \in \mathbb{R}^d$, and $f : \mathbb{R} \to \{0, 1\}$ is an $O(k)$-piecewise constant function.

Given Lemma 4.5.2, it is easy to finish the proof of Lemma 4.5.1.

Proof of Lemma 4.5.1. We simply take the candidate parameters, obtain F from Lemma 4.5.2, and partition $Y = Y_1 \cup Y_2$, so that F is constant on both Y_1 and Y_2. We let Q_j be the set of i so that $F(G_i)$ returns the value $j-1$ with large probability. Letting the partition of X for Condition 4.2.29 be $X = X_1 \cup \ldots \cup X_k$, we let $X^j = \bigcup_{i \in Q_j} X_i$. Lemma 4.2.32 shows that the X^j satisfy the appropriate conditions for $\mathcal{M}_j$. It remains to prove that Y_j equals X^j with a sufficiently small rate of corruptions. The fraction of points misclassified by F equals ε (the fraction of outliers in the sample Y) plus the misclassification error of F. We note that given the form of F and the fact that the uncorrupted samples in Y is a good sample (satisfy Condition 4.2.29), the fraction of misclassified samples from each component i is at most the probability that a random sample from G_i gets misclassified (at most $\tilde{O}_k(\eta^{1/(2k)})$ by Lemma 4.5.2) plus $O(k\gamma)$. Summing this over components, gives Lemma 4.5.1. $\square$

Let us now describe the algorithm to prove Lemma 4.5.2 (and evaluate F), which is given in pseudocode below (Algorithm 4.5.3).

Algorithm 4.5.3 (Algorithm for Spectrally Separating Thin Components)**.**

Input: *Estimated parameters* $\left\{\hat{\mu}_i, \hat{\Sigma}_i\right\}_{i \leqslant k}$, *parameters* η, δ.

Output: *A function* $F : \mathbb{R}^d \to \{0, 1\}$.

Operation:

1. *Find a unit-norm direction* v *such that there exists* $s \in [k]$, $v^T \hat{\Sigma}_s v < 2\eta$.
2. *Compute* $(v^T \hat{\Sigma}_i v)$ *for all* $i \in [k]$.

(a) *If there exists $j \in [k]$ such that $(v^T \hat{\Sigma}_j v) > \sqrt{\eta}$, find a t such that $\sqrt{\eta} > t > 2\eta$ and there is no $j \in [k]$ with $t < v^T \hat{\Sigma}_j v < t\,\Omega(\eta^{-1/(2k)})$. Set $F(x) = 1$ if there is an i such that $|v \cdot (x - \hat{\mu}_i)| < \sqrt{t}\log(1/\eta)$ and 0 otherwise.*

(b) *Otherwise, compute $v \cdot \hat{\mu}_i$ for all $i \in [k]$. Find a t between the minimum and the maximum of $v \cdot \hat{\mu}_i$ such that there is no $v \cdot \hat{\mu}_i$ within $1/(20k)$ of t. Set $F(x) = 1$ if $v \cdot x > t$ and 0 otherwise.*

Proof of Lemma 4.5.2. Let v be a unit vector and $s \in [k]$ such that $v^T \hat{\Sigma}_s v < 2\eta$. By assumption, we have that $\mathbf{Var}[v \cdot \mathcal{M}] \geqslant 1/2$. Furthermore,

$$\mathbf{Var}[v \cdot \mathcal{M}] = \sum_i w_i (v^T \Sigma_i v) + \sum_i w_i (v \cdot (\mu_i - \mu))^2 \leqslant \sum_i w_i (v^T \Sigma_i v) + \sum w_i (v \cdot (\mu_i - \mu_s))^2 ,$$

where μ is the mean of $\mathcal{M}$. This means that either there exists $j \in [k]$ such that $(v^T \Sigma_j v) > 1/4$, or there exists $j \in [k]$ such that $|v \cdot (\mu_j - \mu_s)| > 1/4$. Since we have approximations of these quantities to order δ, we have that there is $j \in [k]$ such that $(v^T \hat{\Sigma}_j v) > 1/10$ or that there is $j \in [k]$ with $|v \cdot (\hat{\mu}_j - \hat{\mu}_s)| > 1/10$.

We first consider the case that there is a $j \in [k]$ such that $(v^T \hat{\Sigma}_j v) > \sqrt{\eta}$. Since there is a $j \in [k]$ with $(v^T \hat{\Sigma}_j v) > \sqrt{\eta}$ and another $s \in [k]$ with $(v^T \hat{\Sigma}_s v) < 2\eta$, there must be some $\sqrt{\eta} > t > 2\eta$ such that there is no $j \in [k]$ with $t < v^T \hat{\Sigma}_j v < t\,\Omega(\eta^{-1/(2k)})$. Otherwise, there must be at least one $\hat{\Sigma}_i$ in each $2\eta \leqslant \Omega(\eta^{-1/(2k)})^i \leqslant \sqrt{\eta}$, where we need more than k components.

For a given x, we define $F(x)$ to be 1 if there exists i such that $|v \cdot (x - \hat{\mu}_i)| < \sqrt{t}\log(1/\eta)$, and $F(x) = 0$ otherwise.

To show that this works, we note that for all $i \in [k]$, if $v^T \hat{\Sigma}_i v \leqslant t$, then $\mathbf{Var}[v \cdot G_i] \leqslant$

$t + \delta$, and since $|v \cdot (\mu_i - \hat{\mu}_i)| < \delta$, by the Gaussian tail bound, we have that

$$\mathop{\mathbb{P}}_{x \sim G_i} \left(|x - \mu_i| \geqslant (\sqrt{t} \log(1/\eta) - \delta) \right) \leqslant \exp\left(-\frac{(\sqrt{t}\log(1/\eta) - \delta)^2}{2(t+\delta)} \right) = O(\eta) \ .$$

Thus, all but an $O(\eta)$-fraction of the samples of G_i have $F(x) = 1$.

On the other hand, for components i with $v^T \hat{\Sigma}_i v \gg t\eta^{-1/(2k)}$, we have that $\mathbf{Var}[v \cdot G_i] \gg t\eta^{-1/(2k)}$. Then, the density of G_i is at most $1/\sqrt{2\pi t \eta^{-1/(2k)}}$. So, the probability that a sample from $v \cdot G_i$ lies in any interval of length $2\sqrt{t}\log(1/\eta)$ is at most

$$\frac{1}{\sqrt{2\pi t\eta^{-1/(2k)}}} 2\sqrt{t}\log(1/\eta) = \tilde{O}(\eta^{1/(4k)}) \ .$$

Since there are k such intervals, the probability that $F(x)$ is 1 when x is drawn from G_i is at most $\tilde{O}_k(\eta^{1/(4k)})$. This completes our proof of point (1), and point (2) follows from the fact that we know of component G_j in one class and G_s in the other class.

We next consider the case where $(v^T \hat{\Sigma}_j v) \leqslant \sqrt{\eta}$ for all $j \in [k]$, and where $|v \cdot (\hat{\mu}_j - \hat{\mu}_s)| > 1/10$ for some $j \in [k]$. Then we can find some t between $v \cdot \hat{\mu}_j$ and $v \cdot \hat{\mu}_s$ such that no $v \cdot \hat{\mu}_i$ is within $1/(20k)$ of t. In this case, we define $F(x)$ be 1 if $v \cdot x > t$ and 0 otherwise. To show part (1), first consider $i \in [k]$ such that $v \cdot \hat{\mu}_i < t - 1/(20k)$. Then we have that $v \cdot \mu_i < t - 1/(30k)$. Furthermore, $\mathbf{Var}[v \cdot G_j] \leqslant \delta + \sqrt{\eta}$. Therefore, the probability that $v \cdot G_i > t$ is at most $\exp(-\Omega_k((\delta + \sqrt{\eta})^{-2}))$, which is sufficient.

A similar argument holds in the other direction for $i \in [k]$ such that $v \cdot \hat{\mu}_i > t + 1/(20k)$, and statement (2) holds because we know that there are both kinds of components. This completes the proof. □

4.6 Robust Proper Learning: Proof of Theorem 4.1.2

In this section, we show how to combine the partial clustering, tensor decomposition, and recursive clustering algorithms to establish our main result. The main theorem we prove is as follows:

Theorem 4.6.1 (Robustly Learning k-Mixtures of Arbitrary Gaussians with small error)**.** *Given $0 < \varepsilon < 1/k^{k^{O(k^2)}}$ and a multiset $Y = \{y_1, y_2, \ldots, y_n\}$ of n i.i.d. samples from a distribution F such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown k-mixture of Gaussians $\mathcal{M} = \sum_{i\leqslant k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, where $n \geqslant n_0 = d^{O(k)} \operatorname{poly}_k(1/\varepsilon)$, there exists an algorithm that runs in time $n^{O(1)} \operatorname{poly}_k(1/\varepsilon)$ and with probability at least 0.99 outputs a hypothesis k-mixture of Gaussians $\widehat{\mathcal{M}} = \sum_{i\leqslant k} \hat{w}_i \mathcal{N}\left(\hat{\mu}_i, \hat{\Sigma}_i\right)$ such that $d_{TV}\left(\mathcal{M}, \widehat{\mathcal{M}}\right) = \mathcal{O}(\varepsilon^{c_k})$, with $c_k = 1/(100^k C^{(k+1)!} k! \mathrm{sf}(k+1))$, where $C > 0$ is a universal constant and $\mathrm{sf}(k) = \Pi_{i\in[k]} i!$ is the super-factorial function.*

The algorithm establishing Theorem 4.6.1 is given in pseudocode below. Algorithm 4.6.3 takes as input a corrupted sample from a k-mixture of Gaussians and outputs a set of k mixing weights, means, and covariances, such that the resulting mixture is close to the input mixture in total variation distance with non-negligible probability. Algorithm 4.6.2 simply runs Algorithm 4.6.3 many times to create a polynomial-size list of candidate hypotheses (consisting of mixing weights, means, and covariances), and finally runs a robust tournament to outputs a winner. This boosts the probability of success to at least 0.99.

Algorithm 4.6.2 (Algorithm for Robustly Learning Arbitrary GMMs)**.**

Input: *An outlier parameter $\varepsilon > 0$ and a component-number parameter $k \in \mathbb{N}$. An ε-corrupted sample $Y = \{y_1, y_2, \ldots, y_n\}$ from a k-mixture of Gaussians $\mathcal{M} = \sum_{i\in[k]} w_i \mathcal{N}(\mu_i, \Sigma_i)$.*

Parameters: *Let $c_k = 1/(100^k C^{(k+1)!} \mathrm{sf}(k+1) k!)$ be a scalar function of k, where $sf(k) = \prod_{i\in[k]} (k-i)! = \prod_{i\in[k]} i!$ and C is a sufficiently large constant.*

Output: *A set of parameters $\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$, such that with probability at least* 0.99

the mixture $\widehat{\mathcal{M}} = \sum_{i\in[k]} \hat{w}_i \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i)$ *is* $O(\varepsilon^{c_k})$*-close in total variation distance to* $\mathcal{M}$.

Operation:

1. *Let* $\mathcal{L} = \{\phi\}$ *be an empty list. Repeat the following* $\varepsilon^{-k^{O(k^2)}}$ *times :*
   (a) *Run Algorithm 4.6.3 with input* Y*, fraction of outliers* ε*, and number of components* k*. Let the resulting output be a set of* k *mixing weights, means and covariances, denoted by* $\left\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\right\}_{i\in[k]}$*. Add* $\left\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\right\}_{i\in[k]}$ *to* $\mathcal{L}$*.*
2. *Run the robust tournament from Fact 4.2.34 over all the hypotheses in* $\mathcal{L}$*. Output the winning hypothesis, denoted by* $\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$*.*

Algorithm 4.6.3 (Cluster or List-Decode)**.**

Input: An outlier parameter $0 < \varepsilon < 1$ *and a component-number parameter* $k \in \mathbb{N}$*. An* ε*-corrupted version* $Y = \{y_1, y_2, \dots, y_n\}$ *of* X*, where* X *is a set of* n *samples from a* k*-mixture of Gaussians* $\mathcal{M} = \sum_{i\in[k]} w_i \mathcal{N}(\mu_i, \Sigma_i)$ *such that* X *is a good sample (satisfies Condition 4.2.29) with respect to* $\mathcal{M}$ *with parameters* $(\varepsilon d^{-8k} k^{-C'k}, 8k+48)$*, where* $C' > 0$ *is a sufficiently large constant.*

Parameters: Let $c_k = 1/(100^k C^{(k+1)!} \mathrm{sf}(k+1) k!)$ *be a scalar function of* k*, where* $\mathrm{sf}(k) = \prod_{i\in[k]} (k-i)! = \prod_{i\in[k]} i!$ *and* C *is a sufficiently large constant.*

Output: A set of parameters $\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$ *such that with probability at least* $\varepsilon^{k^{O(k^2)}}$*,* $d_{TV}\left(\sum_{i\in[k]} \hat{w}_i \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i), \mathcal{M}\right) \leqslant O(\varepsilon^{c_k})$*.*

Operation:

1. ***Treat Light Component as Noise****: If* $k = 0$*, ABORT. With probability* $1/2$*, run Algorithm 4.6.3 on samples* Y*, with fraction of outliers* $\varepsilon + \varepsilon^{1/(10C^{k+1}(k+1)!)}$ *and*

number of components $k-1$. Return the resulting set of estimated parameters, $\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k-1]}$, appended with $(0, 0, I)$. Else, do the following:

▹ *We guess whether the event that the minimum mixing weight α is at least $\varepsilon^{1/(10C^{k+1}(k+1)!)}$ holds. If it does not, we proceed with the algorithm. Else, we treat the smallest weight component as noise and recurse with $k-1$ components.*

2. ***Robust Isotropic Transformation**: With probability 0.5, run the algorithm corresponding to Lemma 4.6.6 on the samples Y, and let $\hat{\mu}, \hat{\Sigma}$ be the robust estimates of the mean and covariance. If $k = 1$, return $\left(\hat{w} = 1, \hat{\mu}, \hat{\Sigma}\right)$. Else, compute $\hat{U}\hat{\Lambda}\hat{U}^\top$, the eigendecomposition of Σ, and for all $i \in [n]$, apply the affine transformation $y_i \rightarrow \hat{U}^\top \hat{\Sigma}^{\dagger/2}(y_i - \hat{\mu})$.*

   ▹ *The resulting estimates $\hat{\mu}, \hat{\Sigma}$ satisfy Lemma 4.6.6, and the uncorrupted samples are effectively drawn from a nearly isotropic k-mixture.*

3. *With probability $1/2$, run either (a) or (b) in the following:*

   (a) ***Partial Clustering via SoS**: Run Algorithm 4.4.14 with outlier parameter ε and accuracy parameter $\varepsilon^{1/(5C^{k+1}(k+1)!)}$. Let Y_1, Y_2 be the partition returned. Guess the number of components in Y_1 to be some $k_1 \in [k-1]$ uniformly at random. Run Algorithm 4.6.3 with input Y_1, fraction of outliers $\varepsilon^{1/(10c^{k+1}(k+1)!)}$, and number of components k_1, and let $\left\{(\hat{w}_i^{(1)}, \hat{\mu}_i^{(1)}, \hat{\Sigma}_i^{(1)})\right\}_{i\in[k_1]}$ be the resulting output. Similarly, run Algorithm 4.6.3 with input Y_2, fraction of outliers $\varepsilon^{1/(10c^{k+1}(k+1)!)}$, and number of components $k-k_1$, and let $\left\{(\hat{w}_i^{(2)}, \hat{\mu}_i^{(2)}, \hat{\Sigma}_i^{(2)})\right\}_{i\in[k-k_1]}$ be the resulting output. Output the set $\left\{(\hat{w}_i^{(1)}|Y_1|/|Y|, \hat{\mu}_i^{(1)}, \hat{\Sigma}_i^{(1)})\right\}_{i\in[k_1]} \cup \left\{(\hat{w}_i^{(2)}|Y_2|/|Y|, \mu_i^{(2)}, \hat{\Sigma}_i^{(2)})\right\}_{i\in[k-k_1]}$.*

   ▹ *When the mixture is covariance separated, the preconditions of Theorem 4.4.3 are satisfied (see Lemma 4.6.7). The partition is non-trivial, and the fraction of outliers increases from $\varepsilon \rightarrow \varepsilon^{1/(10c^{k+1}(k+1)!)}$.*

(b) ***List-Decoding via Tensor Decomposition***: *Run Algorithm 4.3.17 and let* $\mathcal{L}$ *be the resulting list of hypotheses such that each hypothesis is a set of parameters* $\{(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$. *Let* $\tau = \Theta\left(\varepsilon^{1/(40C^{k+1}(k+1)!)}\right)$ *be an eigenvalue threshold. Select a hypothesis,* $\{(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]} \in \mathcal{L}$ *uniformly at random.*

▹ *Conditioned on not being covariance separated, we satisfy the preconditions of Theorem 4.3.1 (see Lemma 4.6.8). The output is a list that contains* $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$ *such that for all* $i \in [k]$, $\|\hat{\mu}_i - \mu_i\|_2 = O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right)$ *and* $\left\|\hat{\Sigma}_i - \Sigma_i\right\|_F = O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right)$.

i. ***Large Eigenvalues***: *If for all* $i \in [k]$, $\hat{\Sigma}_i \succeq \tau I$, *sample* $\hat{w}_i$ *from* $[0,1]$ *uniformly at random such that* $\sum_i \hat{w}_i = (1 \pm k\varepsilon)$. *Return* $\left\{\left(\hat{w}_i, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\mu}_i + \hat{\mu}, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\Sigma}_i\hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\right)\right\}_{i\in[k]}$.

▹ *If all estimated covariances have all eigenvalues larger than* τ, *the recursion bottoms out and the hypothesis is returned.*

ii. ***Spectral Separation of Thin Components***: *Else,* $\exists v, i$ *s.t.* $v^\top\hat{\Sigma}_i v \leqslant \tau$. *Run the algorithm corresponding to Lemma 4.5.1 with input* Y, *parameter estimates* $\{(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$ *and threshold* τ. *Let* Y_1 *and* Y_2 *be the resulting partition.*

▹ *Use small eigenvalue directions to partition the points.*

A. *If* $\min(|Y_1|, |Y_2|) < \varepsilon^{1/(400kC^{k+1}(k+1)!)}n$, *run Algorithm 4.6.3 with input* Y, *fraction of outliers* $2\varepsilon^{1/(400kC^{k+1}(k+1)!)}$ *and number of components being* $k-1$, *and let* $\left\{(\hat{w}_i^{(1)}, \hat{\mu}_i^{(1)}, \hat{\Sigma}_i^{(1)})\right\}_{i\in[k_1]}$ *be the resulting output. Output the resulting hypothesis* $\left\{(\hat{w}_i, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\mu}_i + \hat{\mu}, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\Sigma}_i\hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top)\right\}_{i\in[k-1]} \cup (0, 0, I)$.

B. *Else, select* $k_1 \in [k-1]$ *uniformly at random. Run Algorithm 4.6.3 with input* Y_1, *fraction of outliers* $\varepsilon^{1/(100kC^{k+1}(k+1)!)}$ *and number of components being* k_1. *Similarly, run Algorithm 4.6.3* *with input Y_2, fraction of outliers $\varepsilon^{1/(100kC^{k+1}(k+1)!)}$ and number of components $k-k_1$, and let $\left\{(\hat{w}_i^{(2)}, \hat{\mu}_i^{(2)}, \hat{\Sigma}_i^{(2)})\right\}_{i\in[k-k_1]}$ be the resulting output. Output the set $\left\{(\hat{w}_i^{(1)}|Y_1|/|Y|, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\mu}_i^{(1)} + \hat{\mu}, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\Sigma}_i^{(1)}\hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top)\right\}_{i\in[k_1]} \cup \left\{(\hat{w}_i^{(2)}|Y_2|/|Y|, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\mu_i^{(2)}+\hat{\mu}, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\Sigma}_i^{(2)}\hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top)\right\}_{i\in[k-k_1]}$.*

4.6.1 Analysis of Algorithm 4.6.2

To prove Theorem 4.6.1, we will require the following intermediate results. We defer some proofs in this subsection to Appendix 4.8.

We use the following lemma to relate the Frobenius distance of covariances to the total variation distance between two Gaussians, when the eigenvalues of the covariances are bounded below.

Lemma 4.6.4 (Frobenius Distance to TV Distance)**.** *Suppose $\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)$ are Gaussians with $\|\mu_1-\mu_2\|_2 \leqslant \delta$ and $\|\Sigma_1 - \Sigma_2\|_F \leqslant \delta$. If the eigenvalues of Σ_1 and Σ_2 are at least $\lambda > 0$, then $d_{\mathsf{TV}}(\mathcal{N}(\mu_1,\Sigma_1), \mathcal{N}(\mu_2,\Sigma_2)) = O(\delta/\lambda)$.*

We start by showing that when Condition 4.2.29 holds, the uniform distribution on a $(1-\varepsilon)$-fraction of the points is certifiably hypercontractive.

Lemma 4.6.5 (Component Moments to Mixture Moments)**.** *Let $\mathcal{M} = \sum_{i\in[k]} w_i\mathcal{N}(\mu_i,\Sigma_i)$ be a k-mixture with mean μ and covariance Σ such that $w_i \geqslant \alpha$, for some $0<\alpha<1$, and for all $i,j\in[k]$, $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F \leqslant 1/\sqrt{\alpha}$. Let X be a multiset of n samples satisfying Condition 4.2.29 with respect to $\mathcal{M}$ with parameters (γ, t), for $0<\gamma<(dk/\alpha)^{-ct}$, for a sufficiently large constant c, and $t\in\mathbb{N}$. Let $\mathcal{D}$ be the uniform distribution over X. Then, $\mathcal{D}$ is $2t$-certifiably (c/α)-hypercontractive and for $d\times d$-matrix-valued indeterminate Q, $\Big|\frac{Q}{2}\left\{\mathbb{E}_{\mathcal{M}}\left(x^\top Qx - \mathbb{E}_{\mathcal{M}} x^\top Qx\right)^2 \leqslant O(1/\alpha)\left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2\right\}$.*

Next, we show how to robustly estimate the mean and covariance of an ε-corrupted set of samples satisfying Condition 4.2.29 when the mixture is not partially clusterable, and make the inliers nearly isotropic.

Lemma 4.6.6 (Robust Isotropic Transformation)**.** *Given* $0 < \varepsilon < 1$*, and* $k \in \mathbb{N}$*, let* $\alpha = \varepsilon^{1/(10C^{k+1}(k+1)!)}$*. Let* $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ *with* $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ *be a k-mixture of Gaussians with* $w_i \geqslant \alpha$ *for all* $i \in [k]$*, and let* μ *and* Σ *be the mean and covariance of* $\mathcal{M}$ *such that* $r = \operatorname{rank}(\Sigma)$ *and for all* $i, j \in [k]$*,* $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F \leqslant 1/\sqrt{\alpha}$*. Let* X *be a good sample satisfying Condition 4.2.29 with respect to* $\mathcal{M}$ *for some parameters* (γ, t)*. Given a set* Y*, an* ε*-corrupted version of* X*, of size* $n \geqslant n_0 = d^{O(1)}$*, there exists an algorithm that takes* Y *as input and in time* $n^{O(1)}$ *outputs estimators* $\hat{\mu}$ *and* $\hat{\Sigma}$ *such that* $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^\top$ *is the eigenvalue decomposition, where* $\hat{U} \in \mathbb{R}^{n\times r}$ *has orthonormal columns and* $\Lambda \in \mathbb{R}^{r\times r}$ *is a diagonal matrix. Further, we can obtain* n *samples* Y' *by applying the affine transformation* $y_i \rightarrow \hat{U}^\top\hat{\Sigma}^{\dagger/2}(y_i - \hat{\mu})$ *to each sample, such that a* $(1-\varepsilon)$*-fraction have mean* μ' *and covariance* Σ' *satisfying*

1. $\|\mu'\|_2 \leqslant O\Big(\Big(1 + \frac{\sqrt{\varepsilon}k}{\alpha}\Big)\sqrt{\varepsilon/\alpha}\Big)$,
2. $\left(\frac{1}{1+(k\sqrt{\varepsilon/\alpha})}\right) I_r \preceq \Sigma' \preceq \left(\frac{1}{1-(k\sqrt{\varepsilon/\alpha})}\right) I_r$,
3. $\|\Sigma' - I_r\|_F \leqslant O\big(\sqrt{\varepsilon}k/\alpha\big)$,

where I_r *is the* r*-dimensional Identity matrix, and the remaining points are arbitrary. Let* X' *be the set obtained by* $\hat{U}^\top\hat{\Sigma}^{\dagger/2}(x_i - \hat{\mu})$*. Then,* X' *is a good sample (satisfies Condition 4.2.29) with respect to* $\sum_{i=1}^{k} w_i \mathcal{N}\left(\hat{U}^\top\hat{\Sigma}^{\dagger/2}(\mu_i - \hat{\mu}), \hat{U}^\top\hat{\Sigma}^{\dagger/2}\Sigma_i\hat{\Sigma}^{\dagger/2}\hat{U}\right)$ *and parameters* (γ, t)*, and* Y' *is an* ε*-corruption of* X'*.*

Proof. For any $t' \in \mathbb{N}$, it follows from Corollary 4.2.20 that $\mathcal{M}$ has $2t'$-certifiably $(4/\alpha)$-hypercontractive degree-2 polynomials, since $w_i \geqslant \alpha$ for all i. Next, Lemma 4.6.5 implies that the uniform distribution over X also has $2t'$-certifiably $(8/\alpha)$-hypercontractive degree-2 polynomials and for $d \times d$-matrix-valued indeterminate Q,

$$\vdash_{2}^{Q} \left\{ \mathbb{E}_{\mathcal{M}} \left(x^\top Q x - \mathbb{E}_{\mathcal{M}} x^\top Q x \right)^2 \leqslant O(1/\alpha) \left\| \Sigma^{1/2} Q \Sigma^{1/2} \right\|_F^2 \right\}.$$

Then, it follows from Fact 2.2.3 that if $\frac{16}{\alpha} t' \varepsilon^{1-4/t'} \ll 1$, there exists an algorithm that runs in time $n^{O(t')}$ and outputs estimates $\hat{\mu}$ and $\hat{\Sigma}$ satisfying:

1. $\left\| \Sigma^{\dagger/2} (\mu - \hat{\mu}) \right\|_2 \leqslant O(t'/\alpha)^{1/2} \varepsilon^{1-1/t'}$,
2. $\left(1 - (k/\alpha)\varepsilon^{1-2/t'}\right) \Sigma \preceq \hat{\Sigma} \preceq \left(1 + (k/\alpha)\varepsilon^{1-2/t'}\right) \Sigma$ and,
3. $\left\| \Sigma^{\dagger/2} \left(\hat{\Sigma} - \Sigma \right) \Sigma^{\dagger/2} \right\|_F \leqslant (t'/\alpha) O(\varepsilon^{1-1/t'})$.

Setting $t' = 2$, compute $\hat{\Sigma} = \hat{U} \hat{\Lambda} \hat{U}^\top$, the eigendecomposition of $\hat{\Sigma}$, such that $\hat{U} \in \mathbb{R}^{n \times r}$ has orthonormal columns, where $r \leqslant d$ is the rank of $\hat{\Sigma}$ and $\hat{\Lambda} \in \mathbb{R}^{r \times r}$ is a diagonal matrix. Similarly, let $\Sigma = U \Lambda U^\top$ be the eigendecomposition of Σ. We apply the affine transformation $y_i \to \hat{U}^\top \hat{\Sigma}^{\dagger/2} (y_i - \hat{\mu})$ to each sample and thus we can assume throughout the rest of our argument that we have access to ε-corrupted samples from a k-mixture of Gaussians with mean $\mu' = \hat{U}^\top \hat{\Sigma}^{\dagger/2} (\mu - \hat{\mu})$ and covariance $\Sigma' = \hat{U}^\top \hat{\Sigma}^{\dagger/2} \Sigma \hat{\Sigma}^{\dagger/2} \hat{U}$. Then, we have that

$$\begin{aligned} \|\mu'\|_2 = \left\| \hat{U}^\top \hat{\Sigma}^{\dagger/2} (\mu - \hat{\mu}) \right\|_2 &\leqslant \left\| \hat{U}^\top \right\|_{\mathrm{op}} \left\| \hat{\Sigma}^{\dagger/2} (\mu - \hat{\mu}) \right\|_2 \\ &\leqslant O\left(\left(1 + \frac{\sqrt{\varepsilon} k}{\alpha} \right) \sqrt{\varepsilon/\alpha} \right), \end{aligned}$$

where the last inequality follows from (1) and (2). It also follows from (2) that

$$\left(\frac{1}{1 + (k\sqrt{\varepsilon}/\alpha)} \right) \hat{\Sigma} \preceq \Sigma \preceq \left(\frac{1}{1 - (k\sqrt{\varepsilon}/\alpha)} \right) \hat{\Sigma}\,. \tag{4.6.1}$$

Multiplying out (4.6.1) with $\hat{U}^\top \hat{\Sigma}^{\dagger/2}$ on the left and $\hat{\Sigma}^{\dagger/2} \hat{U}$ on the right, we have

$$\left(\frac{1}{1 + (k\sqrt{\varepsilon}/\alpha)} \right) \hat{U}^\top \hat{\Sigma}^{\dagger/2} \hat{\Sigma} \hat{\Sigma}^{\dagger/2} \hat{U} \preceq \Sigma' \preceq \left(\frac{1}{1 - (k\sqrt{\varepsilon}/\alpha)} \right) \hat{U}^\top \hat{\Sigma}^{\dagger/2} \hat{\Sigma} \hat{\Sigma}^{\dagger/2} \hat{U}\,.$$

Observe that (2) implies that the rank of $\hat{\Sigma}$ and Σ is the same, and thus $\hat{U}^\top \hat{\Sigma}^{\dagger/2} \hat{\Sigma} \hat{\Sigma}^{\dagger/2} \hat{U} = I_r$, where I_r is the r-dimensional Identity matrix. Finally, we have that

$$
\begin{aligned}
\|\Sigma' - I_r\|_F = \left\| \hat{U}^\top \hat{\Sigma}^{\dagger/2} \Sigma \hat{\Sigma}^{\dagger/2} \hat{U} - \hat{U}^\top \hat{\Sigma}^{\dagger/2} \hat{\Sigma} \hat{\Sigma}^{\dagger/2} \hat{U} \right\|_F &\leqslant \left\| \hat{U} \hat{\Lambda}^{-1/2} \hat{U}^\top \left(\Sigma - \hat{\Sigma} \right) \hat{U} \hat{\Lambda}^{-1/2} \hat{U}^\top \right\|_F \\
&= \left\| \hat{U} \hat{\Lambda}^{-1/2} \Lambda^{1/2} \Lambda^{-1/2} \hat{U}^\top \left(\Sigma - \hat{\Sigma} \right) \hat{U} \Lambda^{-1/2} \Lambda^{1/2} \hat{\Lambda}^{-1} \right. \\
&\leqslant \left\| \hat{\Lambda}^{-1/2} \Lambda^{1/2} \right\|_{\mathrm{op}}^2 \left\| \Sigma^{\dagger/2} \left(\hat{\Sigma} - \Sigma \right) \Sigma^{\dagger/2} \right\|_F \\
&\leqslant O\left(\sqrt{\varepsilon} k / \alpha \right) ,
\end{aligned}
$$

where we use that $\hat{\Lambda}^{-1/2} = \hat{\Lambda}^{-1/2} \Lambda^{1/2} \Lambda^{-1/2}$, the sub-multiplicative property of the Frobenius norm, the column span U and $\hat{U}$ is identical (see (2)), and the Frobenius recovery guarantee in (3).

Finally, it follows from Lemma 4.2.30 that Condition 4.2.29 is affine invariant and is thus preserved under $x_i \to \hat{U}^\top \hat{\Sigma}^{-1/2} (x_i - \hat{\mu})$, for $i \in [n]$, with parameters (γ, t). $\square$

The above robust isotropic transformation lemma allows us to obtain a covariance that is close to the identity matrix in a full-dimensional subspace (potentially smaller than the input dimension). Therefore, we will subsequently drop the subscript for the dimension, wherever it is clear from the context.

Next, we show that whenever the minimum mixing weight is sufficiently larger than the fraction of outliers, and a pair of components is covariance separated, we can partially cluster the samples.

Lemma 4.6.7 (Non-negligible Weight and Covariance Separation)**.** *Given* $0 < \varepsilon < 1/k^{k^{O(k^2)}}$ *and* $k \in \mathbb{N}$, *let* $\alpha = \varepsilon^{1/(45 C^{k+1} (k+1)!)}$.

Let $\mathcal{M} = \sum_{i=1}^k w_i G_i$ *with* $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ *be a* k*-mixture of Gaussians with mixture covariance* Σ *such that* $w_i \geqslant \alpha$ *for all* $i \in [k]$ *and there exist* $i, j \in [k]$ *such that* $\left\| \Sigma^{\dagger/2} \left(\Sigma_i - \Sigma_j \right) \Sigma^{\dagger/2} \right\|_F^2 > 1/\alpha^5$. *Further, let X be a good sample satisfying Condition 4.2.29 with respect to* $\mathcal{M}$ *for some parameters* $\gamma \leqslant \varepsilon d^{-8k} k^{-Ck}$*, for a sufficiently large constant*

C, and $t \geqslant 8k$. Let Y be an ε-corrupted version of X of size $n \geqslant n_0 = (dk)^{\Omega(1)}/\varepsilon$, Algorithm 4.4.14 partitions Y into Y_1, Y_2 in time $n^{O(1)}$ such that with probability at least $2^{-O(k)}(1 - O(\alpha))$ there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i \in Q_j} w_i G_i$ and $W_j = \sum_{i \in Q_j} w_i$, then Y_j is an $O\left(\varepsilon^{1/(45C^{k+1}(k+1)!)}\right)$-corrupted version of $\bigcup_{i \in Q_j} X_i$ satisfying Condition 4.2.29 with respect to $\mathcal{M}$ with parameters $\left(O(k\gamma/W_j), t\right)$.

Proof. We run Algorithm 4.4.14 with sample set Y, number of components k, the fraction of outliers ε and the accuracy parameter η. Since X is a good sample (satisfies Condition 4.2.29), we can set $t = 10$, $\beta = (k^2 t^4 \alpha)^{t/2} = O_k(\alpha^5)$ and $\eta = \alpha^2 \gg \sqrt{\varepsilon/\alpha}$ in Theorem 4.4.13. Then, by assumption, there exist i, j such that

$$\left\| \Sigma^{\dagger/2} \left(\Sigma_i - \Sigma_j \right) \Sigma^{\dagger/2} \right\|_F^2 > \frac{1}{\alpha^5} = \Omega\left(\frac{k^2 t^4}{\beta^{2/t} \alpha^4} \right) .$$

We observe that we also satisfy the other preconditions for Theorem 4.4.13, since $n \geqslant (dk/)^{\Omega(1)}/\varepsilon$.

Then, Theorem 4.4.13 implies that with probability at least $2^{-O(k)}(1 - O(\eta/\alpha - \sqrt{\eta})) = 2^{-O(k)}(1 - O(\alpha))$, the set Y is partitioned in two sets Y_1 and Y_2 such that there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i \in Q_j} w_i G_i$ and $W_j = \sum_{i \in Q_j} w_i$, then Y_j is an $O\left(\varepsilon^{1/(45C^{k+1}(k+1)!)}\right)$-corrupted version of $\bigcup_{i \in Q_j} X_i$. By Lemma 4.2.32, $\bigcup_{i \in Q_j} X_i$ is a good sample (satisfies Condition 4.2.29) with respect to $\mathcal{M}$ with parameters $\left(O(k\gamma/W_j), t\right)$.

□

When the mixture is not covariance separated and nearly isotropic, we can obtain a small list of hypotheses such that one of them is close to the true parameters, via tensor decomposition.

Lemma 4.6.8 (Mixture is List-decodable)**.** *Given $0 < \varepsilon < 1/k^{k^{O(k^2)}}$ let $\alpha = \varepsilon^{1/(45C^{k+1}(k+1)!)}$. Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with mixture* *mean* μ *and mixture covariance* Σ*, such that* $\|\mu\|_2 \leqslant O\left(\sqrt{\varepsilon/\alpha}\right)$*,* $\|\Sigma - I\|_F \leqslant O(\sqrt{\varepsilon}/\alpha)$*,* $w_i \geqslant \alpha$ *for all* $i \in [k]$*, and* $\left\|\Sigma_i - \Sigma_j\right\|_F^2 \leqslant 1/\alpha^5$ *for any pair of components, and let* X *be a good sample (satisfying Condition 4.2.29) with respect to* $\mathcal{M}$ *for some parameters* $\gamma = \varepsilon d^{-8k} k^{-Ck}$*, for a sufficiently large constant* C*, and* $t = 8k$*. Let* Y *be an* ε*-corrupted version of* X *of size* n*, Algorithm 4.3.17 outputs a list* L *of hypotheses of size* $1/\varepsilon^{k^{O(k^2)}}$ *in time* $\mathrm{poly}(|L|, n)$ *such that if we choose a hypothesis* $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$ *uniformly at random,* $\|\mu_i - \hat{\mu}_i\|_2 \leqslant O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right)$ *and* $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right)$ *for all* i *with probability at least* $\varepsilon^{k^{O(k^2)}}$*.*

Proof. Recall we run Algorithm 4.3.17 on the samples Y, the number of clusters k, the fraction of outliers ε and the minimum weight $\alpha = \varepsilon^{1/(20C^{k+1}(k+1)!)}$. Next, we show that the preconditions of Theorem 4.3.1 are satisfied. First, the upper bounds on $\|\mu\|_2$ and $\|\Sigma - I\|_F$ imply $\sum_{i\in k} w_i \left(\Sigma_i + \mu_i\mu_i^\top\right) = \Sigma + \mu\mu^\top \preceq (1 + O(\sqrt{\varepsilon}/\alpha))I$. Since the LHS is a conic combination of PSD matrices, it follows that for all $i \in [k]$, $\mu_i\mu_i^\top \preceq \frac{1}{\alpha}\left(1 + O(\sqrt{\varepsilon}/\alpha)\right) I$, and thus $\left\|\mu_i\mu_i^\top\right\|_F \leqslant \frac{2}{\alpha}$. Next, we can write:

$$
\begin{aligned}
\|\Sigma_i - I\|_F &\leqslant \left\|\Sigma_i - (\Sigma + \mu\mu^\top)\right\|_F + \|\Sigma - I\|_F + \left\|\mu\mu^\top\right\|_F \\
&= \left\|\Sigma_i - \sum_{j\in[k]} w_j \left(\Sigma_j + \mu_j\mu_j^\top\right)\right\|_F + \frac{\sqrt{\varepsilon}k}{\alpha} + \frac{\varepsilon}{\alpha} \\
&\leqslant \left\|\sum_{j\in[k]} w_j \left(\Sigma_i - \Sigma_j\right)\right\|_F + \frac{2}{\alpha} + \frac{\sqrt{\varepsilon}k}{\alpha} + \frac{\varepsilon}{\alpha} \\
&\leqslant \frac{2}{\alpha^{5/2}} ,
\end{aligned}
$$

where the first and the third inequalities follow from the triangle inequality and the upper bound on $\left\|\mu_i\mu_i^\top\right\|_F$, and the last inequality follows from the assumption that $\left\|\Sigma_i - \Sigma_j\right\|_F^2 \leqslant 1/\alpha^5$ for every pair of covariances Σ_i, Σ_j. So, we can set $\Delta = 2\alpha^{-5/2}$ in Theorem 4.3.1.

Then, given the definition of α, we have that

$$\eta = 2k^{4k}O(1+\Delta/\alpha)^{4k}\sqrt{\varepsilon} = O\left(\varepsilon^{2/5}\right)$$

and $1/\varepsilon^2 \geqslant \log(1/\eta)(k+1/\alpha+\Delta)^{4k+5}/\eta^2$. Therefore, Algorithm 4.3.17 outputs a list L of hypotheses such that $|L| = \exp\left(1/\varepsilon^2\right)$, and with probability at least 0.99, L contains a hypothesis that satisfies the following: for all $i \in [k]$,

$$\|\hat{\mu}_i - \mu_i\|_2 = O\left(\frac{\Delta^{1/2}}{\alpha}\right)\eta^{G(k)} = O\left(\varepsilon^{-1/(20C^{k+1}(k+1)!)} \cdot \varepsilon^{1/(10C^{k+1}(k+1)!)}\right) = O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right) \text{ and}$$
$$\left\|\hat{\Sigma}_i - \Sigma_i\right\|_F = O(k^4)\,\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)} = O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right)\ . \tag{4.6.2}$$

Then if we choose a hypothesis in L uniformly at random, the probability that we choose the hypothesis satisfying (4.6.2) is at least $1/|L| = \varepsilon^{k^{O(k^2)}}$. □

Finally, if the mixture has a covariance matrix with small variance along any direction, we can further cluster the points by projecting the mixture along that direction.

Lemma 4.6.9 (Spectral Separation of Thin Components)**.** *Given $0 < \varepsilon < 1/k^{k^{O(k^2)}}$, let $\alpha = \varepsilon^{1/(10C^{k+1}(k+1)!)}$. Let $\mathcal{M} = \sum_{i=1}^k w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with mixture covariance Σ such that $\|\Sigma - I\|_F \leqslant O(\sqrt{\varepsilon}k/\alpha)$, and let X be a good sample (satisfying Condition 4.2.29) with respect to $\mathcal{M}$ for some parameters (γ, t). Given a set Y being an ε-corrupted version of X of size n, and estimates $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$, such that $\|\mu_i - \hat{\mu}_i\|_2 \leqslant O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right)$, $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right)$, suppose there exists a unit vector $v \in \mathbb{R}^d$ such that $v^\top \hat{\Sigma}_s v \leqslant O\left(\varepsilon^{1/(40C^{k+1}(k+1)!)}\right)$, for some $s \in [k]$. Then, there is an algorithm that efficiently partitions Y into Y_1 and Y_2 such that there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i\in Q_j} w_i G_i$ and $W_j = \sum_{i\in Q_j} w_i$, then Y_j is an $\left(O(k^2\gamma) + O\left(\varepsilon^{1/(80kC^{k+1}(k+1)!)}/W_j\right)\right)$-corrupted version of $\bigcup_{i\in Q_j} X_i$ satisfying Condition 4.2.29 with respect to $\mathcal{M}_j$ with parameter $(O(k\gamma/W_j)\,, t)$.*

Proof. We run the algorithm from Lemma 4.5.1 with the input being the samples Y, the current hypothesis $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$, and the minimum eigenvalue $\eta = O\left(\varepsilon^{1/(40C^{k+1}(k+1)!)}\right)$. Observe that the mixture covariance satisfies $\Sigma \succeq \left(1 - O\left(\sqrt{\varepsilon}k/\alpha\right)\right) I \succeq I/2$ and the upper bound on means and covariance is $\delta = O\left(\varepsilon^{1/(20kC^{k+1}(k+1)!)} n\right)$ by assumption. Therefore, we satisfy the preconditions of Lemma 4.5.1. Thus, we obtain a partition Y_1, Y_2 such that there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i\in Q_j} w_i G_i$ and $W_j = \sum_{i\in Q_j} w_i$, then it follows from Lemma 4.2.32 that Y_j is an $\left(O\left(k^2\gamma\right) + O\left(\varepsilon^{1/(80kC^{k+1}(k+1)!)}/W_j\right)\right)$-corrupted version of $\bigcup_{i\in Q_j} X_i$ satisfying Condition 4.2.29 with respect to $\mathcal{M}_j$ with parameter $\left(O\left(k\gamma/W_j\right), t\right)$. □

4.6.2 Proof of the Main Theorem

We are now ready to complete the proof of Theorem 4.6.1.

Proof of Theorem 4.6.1. We divide the proof into two parts: first we show that Algorithm 4.6.3 outputs a hypothesis $\widehat{\mathcal{M}} = \sum_{i\in[k]} \hat{w}_i \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i)$ such that $\widehat{\mathcal{M}}$ and $\mathcal{M}$ are $O(\varepsilon^{c_k})$-close in total variation distance with probability at least $\exp\left(-O(k)/\varepsilon^2\right)$; then we show that Algorithm 4.6.2 outputs a k-mixture of Gaussians $\widehat{\mathcal{M}}$ such that $\widehat{\mathcal{M}}$ and $\mathcal{M}$ are $O_k(\varepsilon^{c_k})$-close in total variation distance with probability 0.99.

We proceed the first part by induction on k. Let $c_k = \frac{1}{(100)^k C^{(k+1)!} \mathrm{sf}(k+1) k!}$ be a scalar that only depends on k, where $C > 0$ is a sufficiently large universal constant.

Induction Hypothesis:

Let X be a good sample (satisfying Condition 4.2.29) with respect to a k-mixture of Gaussians $\mathcal{M}$ for some parameters $\gamma = \varepsilon d^{-8k} k^{-C'k}$, where C' is a sufficiently large constant and $t = 8k + 48$. Given a set Y being an ε-corrupted version of X of size n, the outlier parameter ε and the component-number parameter k, Algorithm 4.6.2 returns a k-mixture of Gaussians $\widehat{\mathcal{M}}$ such that $\widehat{\mathcal{M}}$ and $\mathcal{M}$ are $O_k(\varepsilon^{c_k})$-close in total variation distance with probability $\varepsilon^{k^{O(k^2)}}$.

Base Case: For $k = 1$, the algorithm returns the single Gaussian with mean $\hat{\mu}$ and $\hat{\Sigma}$ at

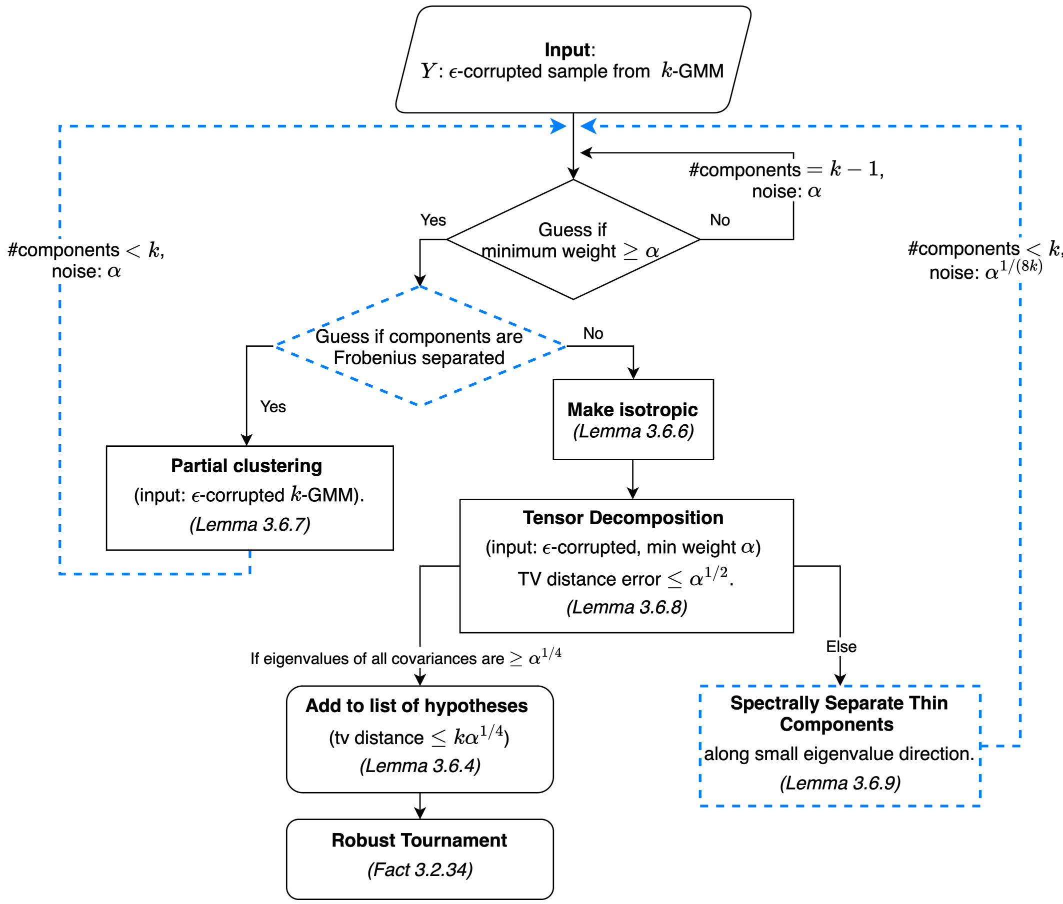

Figure 4.1: If we assume a $1/\text{poly}(k)$ lower bound on minimum weight, then we can skip all blue steps above; the partial clustering (with TV distance separation, see Theorem 4.4.2) is carried out till it can no longer be done within a cluster and then followed by the tensor decomposition step.

Step 2. Suppose the true Gaussian is $N(\mu, \Sigma)$. It follows from the proof of Lemma 4.6.6,

$$\left\|\Sigma^{\dagger/2}(\hat{\mu}-\mu)\right\|_2 = \left\|\Sigma^{\dagger/2}(\hat{\mu}-\mu)\right\|_2 \leqslant O\left(\sqrt{\varepsilon}\right)$$

and

$$\left\|\Sigma^{\dagger/2}\left(\hat{\Sigma}-\Sigma\right)\Sigma^{\dagger/2}\right\|_F \leqslant O\left(\sqrt{\varepsilon}\right) ,$$

and thus it follows from Fact 4.2.1 that the total variation distance between the hypothesis Gaussian and the true Gaussian is at most $O\left(\sqrt{\varepsilon}\right)$. We can then conclude that the base case is true.

Inductive Step: We assume that our induction hypothesis holds for any $m < k$ and then prove that the induction hypothesis holds for k.

Small Clusters Can be Treated as Noise. Conditioning on the base case being true, we begin by guessing whether the minimum weight is less than $\varepsilon^{1/(10C^{k+1}(k+1)!)}$ with equal probability.

Let $w_{\min} = \min_i w_i$. If $w_{\min} \leqslant \varepsilon^{1/(10C^{k+1}(k+1)!)}$, our algorithm takes step 1 with probability 0.5. In this case, we treat the smallest component as noise and recurse on the set of samples Y. We set the number of components to be $k-1$ and the fraction of outliers being $\varepsilon + \varepsilon^{1/(10C^{k+1}(k+1)!)} \leqslant 2\varepsilon^{1/(10C^{k+1}(k+1)!)}$. By Lemma 4.2.32, Y is an $2\varepsilon^{1/(10C^{k+1}(k+1)!)}$-corrupted version of a good sample (satisfying Condition 4.2.29) with respect to a $(k-1)$-mixture for parameters $\gamma = O\left(k\varepsilon d^{-8k}k^{-C'k}/(1-w_{\min})\right) \leqslant \varepsilon d^{-8(k-1)}(k-1)^{-C'(k-1)}$ and $t = 8k+48$. Thus applying the inductive hypothesis to Y, we learn the mixture up to total variation distance $O_k\left(\left(2\varepsilon^{1/(10C^{k+1}(k+1)!)}\right)^{c_{k-1}}\right) \leqslant O_k(\varepsilon^{c_k})$ with probability $0.5\varepsilon^{(k-1)^{c(k-1)^2}} \geqslant \varepsilon^{k^{ck^2}}$. Now we may assume for all $i \in [k]$, $w_i \geqslant \varepsilon^{1/(10C^{k+1}(k+1)!)}$.

Mixture is Covariance Separated. Let $\alpha = \varepsilon^{1/(10C^{k+1}(k+1)!)}$ and $\psi_1 = \{\exists\, \mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j) \mid \left\|\Sigma_i - \Sigma_j\right\|_F > \alpha^{-1/2}\}$ be the event that the samples were drawn from a mixture that is *covariance separated*. First, consider the case where ψ_1 is true. We will run 3(a) with probability 0.5. Then it follows from Lemma 4.6.7 that Y can be partitioned into Y_1 and Y_2 in time $d^{O(1)}$, such that they both have at least one component and the fraction of outliers in each set Y_1, Y_2 is at most $\varepsilon^{1/(10C^{k+1}(k+1)!)}$ with probability $\alpha^{O(k\log(k/\alpha))}$. Then, we can guess the number of components in Y_1 and we will be correct with probability $1/k$. Conditioned on our guess being correct, let Y_1 consist of k_1 components and Y_2 consist of k_2 components and $k_1 + k_2 = k$.

Let $Q_1 \cup Q_2$ be the non-trivial partition of $[k]$ in Lemma 4.6.7, $\mathcal{M}_j$ be a distribution proportional to $\sum_{i\in Q_j} w_i G_i$ and $W_j = \sum_{i\in Q_j} w_i$, then By Lemma 4.2.32, Y_j is an $O\left(\varepsilon^{1/(10C^{k+1}(k+1)!)}\right)$-corrupted version of $\bigcup_{i\in Q_j} X_i$ satisfying Condition 4.2.29 with respect to $\mathcal{M}$ with parameters $\gamma = O\big(k\varepsilon d^{-8k}k^{-C'k}/\alpha\big) \leqslant \varepsilon d^{-8k_j}(k_j)^{-C'k_j}$ and $t = 8k + 48$. Then, applying the inductive hypothesis on Y_j for $j = 1, 2$, with number of components k_j, we can learn the mixtures $\mathcal{M}_j$ up to total variation distance error $O_k\Big(\varepsilon^{c_{k_j}/(10C^{k+1}(k+1)!)}\Big)$ with probability $\varepsilon^{k_j^{O(k_j^2)}}$. Finally if this is the case, we combine the two hypotheses on Y_1, Y_2 by multiplying each weight in the hypothesis of Y_j by $|Y_j|/|Y|$ and then taking union of two hypotheses. Then our combining method gives a final output that learns our full hypothesis to total variation distance error $O_k\Big(\varepsilon^{c_{k_1}/(10C^{k+1}(k+1)!)}\Big) + O_k\Big(\varepsilon^{c_{k_2}/(10C^{k+1}(k+1)!)}\Big) \leqslant O_k(\varepsilon^{c_k})$ with probability at least $0.5 \cdot 0.5 \cdot \frac{1}{k} \cdot 2^{-O(k)}(1 - O(\alpha))\varepsilon^{k_1^{O(k_1^2)}}\varepsilon^{k_2^{O(k_2^2)}} \geqslant \varepsilon^{k^{O(k^2)}}$.

Mixture is not Covariance Separated. Next, consider the case where ψ_1 is false. With probability 0.5, the algorithm guesses correctly and executes Step 2. Since the mixture is not covariance separated, we satisfy the preconditions of Lemma 4.6.6, and after applying the transformation in Step 2, Σ, the covariance of the mixture $\mathcal{M}$, is $\sqrt{\varepsilon}k/\alpha$-close to the r-dimensional identity, where r is the rank of Σ. However, since we obtain the subspace exactly, we can simply project all samples on the subspace and we drop the r in the subsequent exposition.

Let X' be the set of points obtained by applying the Affine transformation from Step 2 as defined in Lemma 4.6.6. Then, X' is a good sample (satisfies Condition 4.2.29) with respect to a nearly isotropic mixture and parameters $\gamma = \varepsilon d^{-8k}k^{-C'k}$ and $t = 8k + 48$ so that we can continue the algorithm with X'. Whenever we return a hypothesis in the following steps, we will first apply the inverse of the transformation on our estimates $\hat{\mu}_i$ and $\hat{\Sigma}_i$. Since total variation distance is affine invariant, we have the same error guarantee in total variation distance after applying the transformation. From now on, we reduce to the case where Σ is $\sqrt{\varepsilon}k/\alpha$-close to the Identity.

There is a 50% chance our algorithm runs Step 3(b) and we will analyze the remainder of this case under that assumption. It follows from Lemma 4.6.8 that we obtain a hypothesis $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$ such that $\|\mu_i - \hat{\mu}_i\|_2 \leqslant O\Big(\varepsilon^{1/(20C^{k+1}(k+1)!)}\Big)$ and $\big\|\Sigma_i - \hat{\Sigma}_i\big\|_F \leqslant$

$O\left(\varepsilon^{1/(20C^{k+1}(k+1)!)}\right)$ with probability $\varepsilon^{k^{O(k^2)}}$. Conditioned on the hypothesis being correct, we now split into two cases: either all eigenvalues of all the estimated covariances are large (in which case we obtain total variation distance guarantees), or there is a direction along which we can project and cluster further.

Covariance Estimates have Large Eigenvalues. For the hypothesis $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$ from the last step, we compute all the eigenvalues of the estimated covariance matrices, $\hat{\Sigma}_i$, for all $i \in [k]$. If, for all $i \in [k]$, $\lambda_{\min}\left(\hat{\Sigma}_i\right) \geqslant c\varepsilon^{1/(40C^{k+1}(k+1)!)}$, we land in Step 3(b).i that we guess the mixing weights $\hat{w}_i$ uniformly in the range $[0, 1]$ and then we output the corresponding hypothesis $\{\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$. With probability at least ε^k, $\hat{w}_i$ are within ε of the true mixing weights. Under this condition, by Lemma 4.6.4, the mixture $\widehat{\mathcal{M}} = \sum_{i\in[k]} \hat{w}_i \mathcal{N}\left(\hat{\mu}_i, \hat{\Sigma}_i\right)$ is $O_k\left(\varepsilon^{1/(40C^{k+1}(k+1)!)}\right) \leqslant O_k(\varepsilon^{c_k})$-close to $\mathcal{M}$ in total variation distance with probability $0.5 \cdot 0.5 \cdot \varepsilon^k \cdot \varepsilon^{k^{ck^2}} \geqslant \varepsilon^{k^{2ck^2}}$.

One Covariance Has a Small Eigenvalue. Consider the case (Step 3(b).ii) where there exists a unit-norm direction v and an estimate $\hat{\Sigma}_i$ such that $v^\top \hat{\Sigma}_i v \leqslant c\varepsilon^{1/(40C^{k+1}(k+1)!)}$. It then follows from Lemma 4.6.9 that we can partition Y into Y_1 and Y_2 such that each has at least one cluster and the total number of outliers in both Y_1 and Y_2 is at most $O\left(\varepsilon^{1/(80kC^{k+1}(k+1)!)}\right)n$. If Y_1 or Y_2 has size less than $\varepsilon^{1/(400kC^{k+1}(k+1)!)}n$, then we can treat it as noise and get an additive $O(\varepsilon^{1/(400kC^{k+1}(k+1)!)})$-error in total variation distance. Otherwise, the fraction of outliers in both sets is at most $O\left((\varepsilon^{1/(80kC^{k+1}(k+1)!)}n)/(\varepsilon^{1/(400kC^{k+1}(k+1)!)}n)\right) = O\left(\varepsilon^{1/(100kC^{k+1}(k+1)!)}\right)$. We then guess the number of components, k_1, in Y_1 with success probability $1/k$. Let $k_2 = k - k_1$ be the number of components in Y_2. Then, conditioned on this event holding, Y_j is an $O\left(\varepsilon^{1/(100kC^{k+1}(k+1)!)}\right)$-corrupted version of a good sample (satisfying Condition 4.2.29) with respect to a mixture of k_j components with parameter $\gamma = k\varepsilon d^{-8k} k^{-C'k}/\alpha \leqslant \varepsilon d^{-8(k_j)}(k_j)^{-C'(k_j)}$ and $t = 8k + 48$. We can apply the inductive hypothesis to Y_1 with number of components k_1 and fraction of outliers $O\left(\varepsilon^{1/(100kC^{k+1}(k+1)!)}\right)$, and conclude that we learn the components of Y_1 to total variation distance $O_k\left(\varepsilon^{c_{k_1}/(100kC^{k+1}(k+1)!)}\right)$ with probability $\varepsilon^{k_1^{O(k_1^2)}}$. A similar argument holds for Y_2. Finally if this is the case, we combine the two hypotheses on Y_1, Y_2 by multiplying each weight by $|Y_j|/|Y|$ and then taking union of two hypotheses. Then our combining method gives a final output that learns our full hypothesis to total variation distance error $O_k\left(\varepsilon^{c_{k_1}/(100C^{k+1}(k+1)!)}\right) + O_k\left(\varepsilon^{c_{k_2}/(100C^{k+1}(k+1)!)}\right) + O\left(\varepsilon^{1/(400kC^{k+1}(k+1)!)}\right) \leqslant O_k(\varepsilon^{c_k})$ with probability at least $0.5 \cdot 0.5 \cdot \frac{1}{k} \cdot \varepsilon^{k^{ck^2}} \varepsilon^{k_1^{2ck_1^2}} \varepsilon^{k_2^{2ck_2^2}} \geqslant \varepsilon^{k^{2ck^2}}$.

Sample Size and Running Time of Algorithm 4.6.3 By Lemma 4.2.33, we need $n \geqslant kt^{C't}d^t/\gamma^3$ samples to generate good sample X satisfying Condition 4.2.29 with parameters (γ, t). We set $\gamma = \varepsilon d^{-8k}k^{-C'k}$ and $t = 8k + 48$. Then $n \geqslant n_0 = (8k)^{O(k)}d^{O(k)}/\varepsilon^3$. The running time in each sub-routine we invoke is dominated by the running time of the tensor decomposition algorithm, and by Lemma 4.6.8 in the worst case this is $\mathrm{poly}(|L|, n) = \mathrm{poly}\left(\varepsilon^{-k^{O(k^2)}}, d^{O(k)}/\varepsilon^3\right) = d^{O(k)}\varepsilon^{-k^{O(k^2)}}$.

This completes the first part of the proof.

Aggregating Hypotheses. We run Algorithm 4.6.3 repeatedly on set Y and add the return hypothesis into a list $\mathcal{L}$ until with probability 0.99, there exists a hypothesis $\widehat{\mathcal{M}} \in \mathcal{L}$ such that $\widehat{\mathcal{M}}$ and $\mathcal{M}$ are $O_k(\varepsilon^{c_k})$-close in total variation distance. Since Algorithm 4.6.3 outputs a correct mixture with probability $\varepsilon^{k^{O(k^2)}}$, we will run Algorithm 4.6.3 for $\varepsilon^{-k^{O(k^2)}}$ times. Then the total running time is $\varepsilon^{-k^{O(k^2)}} \cdot d^{O(k)}\varepsilon^{-k^{O(k^2)}} = d^{O(k)}\varepsilon^{-k^{O(k^2)}}$.

Robust Tournament. Then we need to run a robust tournament in order to find a hypothesis that is close to the true mixture in total variation distance. Fact 4.2.34 shows that we can do this efficiently only with access to an ε-corrupted set of samples of size $O_k\big(\log(|L|)/\varepsilon^{2c_k}\big) = O_k\left(\log(\varepsilon^{-k^{O(k^2)}})/\varepsilon^{2c_k}\right) = O_k\big(\log(1/\varepsilon)/\varepsilon^{2c_k}\big)$.

This completes the proof. □

4.7 Robust Parameter Recovery: Proof of Theorem 4.1.3

In order to show that our algorithm recovers the individual components and the parameters, we will prove the following identifiability theorem. Without any assumption on the mixtures, it is impossible to distinguish components within ε total variation distance with ε-fraction of noise. So given two mixtures of Gaussians with ε total variation distance, the theorem shows that there exist two partitions of components of the two mixtures respectively such that any two components in the matched pair is are poly(ε)-close in total variation distance.

Theorem 4.7.1 (Identifiability)**.** *Let $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' G_i'$ be two mixtures of Gaussians such that $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \varepsilon$. Then there exists a partition of $[k_1]$ into sets $R_0, R_1, \ldots, R_\ell$ and a partition of $[k_2]$ into sets $S_0, S_1, \ldots, S_\ell$ such that*

1. *Let $W_i = \sum_{j \in R_i} w_j$ for $i = 0, 1, \ldots, k_1$, $W_i' = \sum_{j \in S_i} w_j'$ for $i = 0, 1, \ldots, k_2$. Then for all $i \in [\ell]$,*

$$|W_i - W_i'| \leqslant \mathrm{poly}_k(\varepsilon)$$
$$d_{\mathsf{TV}}(G_j, G_{j'}') \leqslant \mathrm{poly}_k(\varepsilon) \quad \forall j \in R_i, j' \in S_i$$

2. *$W_0, W_0' \leqslant \mathrm{poly}_k(\varepsilon)$.*

Corollary 4.7.2. *There is an algorithm with the following behavior: Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\,\mathrm{poly}(\varepsilon)$ samples from a distribution F on $\mathbb{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $d^{O(k)}\,\mathrm{poly}_k(1/\varepsilon)$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^{k'} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability there exists a partition of $[k]$ into $k'+1$ sets $R_0, R_1, \ldots, R_{k'}$ such that*

1. *Let $W_i = \sum_{j \in R_i} w_j$. Then for all $i \in [k']$,*

$$|W_i - \hat{w}_i| \leqslant \mathrm{poly}_k(\varepsilon)$$

$$d_{\mathsf{TV}}(\mathcal{N}(\mu_j, \Sigma_j), \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)) \leqslant \mathrm{poly}_k(\varepsilon) \quad \forall j \in R_i$$

2. *The sum of weights of exceptional components in R_0 is at most* $\mathrm{poly}_k(\varepsilon)$.

Parameter estimation is implied by TV distance for individual Gaussians (in relative Frobenius norm). The corollary follows immediately from the identifiability theorem.

Outline of Proof. The first step is to deal with the components in $\mathcal{M}$ and $\mathcal{M}'$ with small weights. We will construct $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$ by removing components with small weights. If we prove the statement on $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$, we can then deduce the theorem in the general case with worse, but still polynomial dependencies on ε. The second step is a partial clustering, after which the components within each cluster have TV distance bounded by $1 - \mathrm{poly}(\varepsilon)$. We prove this lemma in a separate section. After that we modify the parameters slightly so that the resulting parameters for different components are either identical or have a minimum separation. After this, we can use a lemma from [48] that provides a 1-1 mapping between the components of two such mixtures with small TV distance such that the mapped pairs have small TV distance.

Distance between Gaussians. We use the following facts for Gaussian distributions.

Lemma 4.7.3 (Frobenius Distance to TV Distance)**.** *Suppose $N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)$ are Gaussians with $\|\mu_1 - \mu_2\|_2 \leqslant \delta$ and $\|\Sigma_1 - \Sigma_2\|_F \leqslant \delta$. If the eigenvalues of Σ_1 and Σ_2 are at least $\lambda > 0$, then $d_{\mathsf{TV}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) = O(\delta/\lambda)$.*

Lemma 4.7.4 (Lemma 5.4 in [48])**.** *Let $\mathcal{M}$ be a mixture of k Gaussians that is connected if we draw edges between all components i, j in $\mathcal{M}$ such that $d_{\mathsf{TV}}(G_i, G_j) \leqslant 1 - \delta$. Let Σ be the covariance matrix of $\mathcal{M}$. Then for any components Σ_i of the mixture*

1. $\Sigma_i \geq \mathrm{poly}_k(\delta)\Sigma$

2. $\left\|\Sigma^{-1/2}(\Sigma - \Sigma_i)\Sigma^{-1/2}\right\|_F \leqslant \mathrm{poly}_k(\delta)^{-1}$.

The proof is identical to Lemma 5.4 in [48]. The only difference is that in [48] the authors assume that the minimal weight of $\mathcal{M}$ is at least δ and TV distance between any pair of components is at least δ but here we do not need these two assumptions, which does not affect the proof.

Fact 4.7.5 (Claim 3.9 in [48])**.** *Let ∂ denote the differential operator with respect to y. If*

$$f(y) = P(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where P is a polynomial in y of degree k (whose coefficients are polynomials in X) and $a(X), b(X)$ are polynomials in X then

$$(\partial - (a(X) + yb(X)))f(y) = Q(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where Q is a polynomial in y with degree exactly $k - 1$ whose leading coefficient is k times the leading coefficient of P.

Fact 4.7.6 (Corollary 3.10 in [48])**.** *Let ∂ denote the differential operator with respect to y. If*

$$f(y) = P(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where P is a polynomial in y of degree k then

$$(\partial - (a(X) + yb(X)))^{k+1} f(y) = 0.$$

Fact 4.7.7 (Claim 3.11 in [48])**.** *Let ∂ denote the differential operator with respect to y. If*

$$f(y) = P(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where P is a polynomial in y of degree k. Let the leading coefficient of P (viewed as a polynomial in y) be $L(X)$. Let $c(X)$ be a linear polynomial in X and $d(X)$ be a quadratic *polynomial in X such that* $\{a(X), b(X)\} \neq \{c(X), d(X)\}$. *If* $b(X) \neq d(X)$ *then*

$$(\partial - (c(X) + yd(X)))^{k'} f(y) = Q(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where Q *is a polynomial of degree* $k + k'$ *in* y *with leading coefficient*

$$L(X)(b(X) - d(X))^{k'}$$

and if $b(X) = d(X)$ *then*

$$(\partial - (c(X) + yd(X)))^{k'} f(y) = Q(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where Q *is a polynomial of degree* k *in* y *with leading coefficient*

$$L(X)(a(X) - c(X))^{k'}.$$

Lemma 4.7.8. *Let* $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i$, $\mathcal{M}' = \sum_{i=1}^{k_2} w'_i G'_i$ *be two mixtures of Gaussians such that* $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \varepsilon$. *For any constant* $0 < c_1 < 1$, *there exists* $i \in [k_1 + k_2 + 1]$ *such that* $w_j, w'_{j'} \notin [\varepsilon^{c_1^{i-1}}, \varepsilon^{c_1^i})$ *for any* $j \in [k_1], j' \in [k_2]$. *Moreover, if*

$$\tilde{\mathcal{M}} = \frac{\sum_{\{j : w_j \geqslant \varepsilon^{c_1^i}\}} w_j G_j}{\sum_{\{j : w_j \geqslant \varepsilon^{c_1^i}\}} w_j}$$
$$\tilde{\mathcal{M}}' = \frac{\sum_{\{j : w'_j \geqslant \varepsilon^{c_1^i}\}} w'_j G'_j}{\sum_{\{j : w'_j \geqslant \varepsilon^{c_1^i}\}} w'_j}$$

then $d_{\mathsf{TV}}(\tilde{\mathcal{M}}, \tilde{\mathcal{M}}') \leqslant O_k(\varepsilon^{c_1^{i-1}})$.

Proof. We can see that $[\varepsilon^{c_1^{i-1}}, \varepsilon^{c_1^i})$ with $i \in [k_1 + k_2 + 1]$ are $k_1 + k_2 + 1$ disjoint intervals and $w_j, w'_{j'}$ with $j \in [k_1], j' \in [k_2]$ have at most $k_1 + k_2$ distinct values. So there is one interval containing no weights.

We then construct $\tilde{\mathcal{M}}$ by removing the small components in $\mathcal{M}$. The sum of weights removed is at most $k\varepsilon^{c_1^{i-1}}$. So $d_{\mathsf{TV}}(\mathcal{M}, \tilde{\mathcal{M}}) \leqslant k\varepsilon^{c_1^{i-1}}$. Similarly, we have $d_{\mathsf{TV}}(\mathcal{M}', \tilde{\mathcal{M}}') \leqslant k\varepsilon^{c_1^{i-1}}$. By the triangle inequality,

$$d_{\mathsf{TV}}(\tilde{\mathcal{M}}, \tilde{\mathcal{M}}') \leqslant d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') + d_{\mathsf{TV}}(\mathcal{M}, \tilde{\mathcal{M}}) + d_{\mathsf{TV}}(\mathcal{M}', \tilde{\mathcal{M}}') \leqslant O_k(\varepsilon^{c_1^{i-1}}).$$

□

Lemma 4.7.8 shows that we can remove components with tiny weights in the mixtures. So in the following lemma, we will assume M and M' are Gaussian mixtures with minimal weights at least poly(ε). We will show that we can partition the union of components of two mixtures so that if we prove Theorem 4.7.1 for each part of the partition, we can combine them to prove Theorem 4.7.1 on the full mixtures.

Lemma 4.7.9. *For any constant* $0 < c_3 < 1$, *there exist* $c_1, c_2 > 0$ *that depend on* k *and* c_3, *such that if* $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' G_i'$ *with* $k_1, k_2 \leqslant k$, $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \varepsilon$ *and* $w_i, w_i' \geqslant \varepsilon^{c_1}$ *for all* i, *then there exists a partition of* $[k_1]$ *into sets* $R_1, \ldots, R_\ell$ *and a partition of* $[k_2]$ *into sets* $S_1, \ldots, S_\ell$ *such that*

1. *For all* $i \in [\ell]$, *let* $W_i = \sum_{j \in R_i} w_j, W_i' = \sum_{j \in S_i} w_j'$ *be the sum of weights in each piece. Let* $\mathcal{M}_i = \frac{1}{W_i} \sum_{j \in R_i} w_j G_j, \mathcal{M}_i' = \frac{1}{W_i'} \sum_{j \in S_i} w_j' G_j'$ *be the submixtures of Gaussians after partition. Then for all* $i \in [\ell]$,

$$|W_i - W_i'| \leqslant \text{poly}_k(\varepsilon)$$
$$d_{\mathsf{TV}}(\mathcal{M}_i, \mathcal{M}_i') \leqslant O_k(\varepsilon^{c_2})$$

2. *Consider the graph with vertices corresponding to components in* $\mathcal{M}$ *and* $\mathcal{M}'$ *and two components are adjacent if the total variation distance between them is at most* $1 - \varepsilon^{c_2 c_3}$. *Then the induced subgraph of vertices with indices* $R_i \cup S_i$ *is connected for all* $i \in [\ell]$.

The proof of Lemma 4.7.9 is deferred to Section 4.7.1. In the following two lemmas, we then prove Theorem 4.7.1 for each pair $\mathcal{M}_i, \mathcal{M}'_i$ defined in Lemma 4.7.9. In Lemma 4.7.10, we construct two mixtures of which pairs of parameters are identical or separated. We also shows it suffices to work under this simplification.

Lemma 4.7.10. *For any constant $0 < c_4 < 1$, there exist c_3, c_5 that depend on k and c_4, such that if $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w'_i G'_i$ with $k_1, k_2 \leqslant k$ and*

1. $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$ *is isotropic,*
2. $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \varepsilon$,
3. $w_i, w'_i \geqslant \varepsilon^{c_3}$ *for all i,*
4. *Let $\mathcal{G}$ be a graph with components G_i, G'_i in $\mathcal{M}$ and $\mathcal{M}'$ as vertex set and two components are adjacent if the total variation distance between them is at most $1 - \varepsilon^{c_3}$. Then $\mathcal{G}$ is connected*

then there exist two mixtures of Gaussians $\tilde{\mathcal{M}} = \sum_{i=1}^{\tilde{k}_1} \tilde{w}_i \tilde{G}_i, \tilde{\mathcal{M}}' = \sum_{i=1}^{\tilde{k}_2} \tilde{w}'_i \tilde{G}'_i$ such that

1. *Any pair in $\{\tilde{\mu}_i\} \cup \{\tilde{\mu}'_i\}$ is either identical or separated by at least $\varepsilon^{c_4 c_5}$*
2. *Any pair in $\{\tilde{\Sigma}_i\} \cup \{\tilde{\Sigma}'_i\}$ is either identical or separated by at least $\varepsilon^{c_4 c_5}$ in Frobenius norm.*
3. $\left\| \mathbb{E}(h_m(\tilde{\mathcal{M}})) - \mathbb{E}(h_m(\tilde{\mathcal{M}}')) \right\|_F \leqslant O_k(\varepsilon^{c_5})$ *for any $m \leqslant O(k)$*
4. *There exist $\pi_1 : [k_1] \to [\tilde{k}_1]$ and $\pi_2 : [k_2] \to [\tilde{k}_2]$ such that*

$$\sum_{i:\pi_1(i)=j} w_i = \tilde{w}_j, \sum_{i:\pi_2(i)=j} w'_i = \tilde{w}'_j,$$
$$d_{\mathsf{TV}}(G_i, \tilde{G}_{\pi_1(i)}) \leqslant \mathrm{poly}_k(\varepsilon), \quad \text{for all } i \in [k_1]$$
$$d_{\mathsf{TV}}(G'_i, \tilde{G}'_{\pi_2(i)}) \leqslant \mathrm{poly}_k(\varepsilon), \quad \text{for all } i \in [k_2].$$

Proof. For any $0 < c_4 < 1$, there is $\ell \in [k^2]$ such that the distance between any pair of parameters in $\{\mu_i\} \cup \{\mu'_i\}$ or the Frobenius distance between any pair in $\{\Sigma_i\} \cup \{\Sigma'_i\}$ is not in the interval $[\varepsilon^{(c_4/2)^{\ell-1}}, \varepsilon^{(c_4/2)^{\ell}})$.

Now consider a graph $\mathcal{G}$ on $k_1 + k_2$ nodes where each node represents a vector in $\{\mu_i\} \cup \{\mu'_i\}$ and two vectors a, b are adjacent if

$$\|a - b\| \leqslant \varepsilon^{(c_4/2)^{\ell-1}}.$$

We now construct new mixtures $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$. For each connected component in $\mathcal{G}$ say $\{\mu_{i_1}, \dots, \mu'_{j_1}, \dots\}$, pick a representative say μ_{i_1} and set $\tilde{\mu}_{i_1} = \cdots = \tilde{\mu}'_{j_1} = \cdots = \mu_{i_1}$. Do this for all connected components and similar in the graph on covariance matrices with edges (i, j) if

$$\left\|\Sigma_i - \Sigma_j\right\|_F \leqslant \varepsilon^{(c_4/2)^{\ell-1}}.$$

After replacing close parameters with a representative, we may get some exactly same components in each new mixture. We then merge components with same means and covariances by adding their weights. Since all representatives of means and covariances are in different connected components of the graphs, they are separated by at least $\varepsilon^{(c_4/2)^{\ell}}$. Setting $c_5 = 1/2(c_4/2)^{\ell-1}$ gives a separation of $\varepsilon^{c_4 c_5}$.

Next we prove 3. There is a natural mapping $\pi_1 : [k_1] \to [\tilde{k}_1]$ that maps any component in $\mathcal{M}$ to the merged component in $\tilde{\mathcal{M}}$ and a similar mapping $\pi_2 : [k_2] \to [\tilde{k}_2]$ for $\mathcal{M}', \tilde{\mathcal{M}}'$. For all i, we have

$$\left\|\tilde{\mu}_{\pi_1(i)} - \mu_i\right\|, \left\|\tilde{\mu}'_{\pi_2(i)} - \mu'_i\right\|, \left\|\tilde{\Sigma}_{\pi_1(i)} - \Sigma_i\right\|_F, \left\|\tilde{\Sigma}'_{\pi_2(i)} - \Sigma'_i\right\|_F \leqslant O_k(1)\varepsilon^{(c_4/2)^{\ell-1}} \tag{4.7.1}$$

because for any pair of parameters above say $\tilde{\mu}_{\pi_1(i)}$ and μ_i, there is a path of length at most $2k$ connecting μ_i to the representative of the connected component, and each edge connects a pair with TV distance at most ε. Suppose $\|\mu_i\|, \|\Sigma_i - I\|_F \leqslant \Delta$. Then by Definition 4.2.4, we have for any integer m,

$$\left\|\mathbb{E}(h_m(\mathcal{M})) - \mathbb{E}(h_m(\tilde{\mathcal{M}}))\right\|_F \leqslant O_k(m)\Delta^m \varepsilon^{(c_4/2)^{\ell-1}}.$$

Since $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$ is isotropic and the minimum weight in $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$ is at least $\frac{1}{2}\varepsilon^{c_3}$, we have $\|\mu_i\| \leqslant \sqrt{2/\varepsilon^{c_3}}$ for all i. Applying Lemma 4.7.4 to $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$, we have $\|I - \Sigma_i\|_F \leqslant \mathrm{poly}_k(\varepsilon^{c_3})^{-1}$. So there is a constant a such that $\Delta \leqslant \varepsilon^{-ac_3}$. If we take $c_3 > 0$ so that $ac_3 O(k) \leqslant 1/2(c_4/2)^{\ell-1}$ and take $c_5 = 1/2(c_4/2)^{\ell-1}$, then

$$\left\|\mathbb{E}(h_m(\mathcal{M})) - \mathbb{E}(h_m(\tilde{\mathcal{M}}))\right\|_F \leqslant O_k(m)\varepsilon^{(c_4/2)^{\ell-1} - O(m)ac_3} = O_k(\varepsilon^{c_5})$$

for $m \leqslant O(k)$. By the same argument, we have the similar inequality for $\mathcal{M}'$ and $\tilde{\mathcal{M}}'$

$$\left\|\mathbb{E}(h_m(\mathcal{M}')) - \mathbb{E}(h_m(\tilde{\mathcal{M}}'))\right\|_F = O_k(\varepsilon^{c_5}).$$

Since we can use Proposition 3.3 to robustly estimate the Hermite tensors of a Gaussian mixture with ε-fraction of noise and $\mathrm{poly}(\varepsilon)$ error guarantee, we must have

$$\|\mathbb{E}(h_m(\mathcal{M})) - \mathbb{E}(h_m(\mathcal{M}'))\|_F \leqslant \mathrm{poly}_k(\varepsilon).$$

Then by the triangle inequality,

$$\begin{aligned}\left\|\mathbb{E}(h_m(\tilde{\mathcal{M}})) - \mathbb{E}(h_m(\tilde{\mathcal{M}}'))\right\|_F &\leqslant \left\|\mathbb{E}(h_m(\mathcal{M})) - \mathbb{E}(h_m(\tilde{\mathcal{M}}))\right\|_F + \\ &\|\mathbb{E}(h_m(\mathcal{M})) - \mathbb{E}(h_m(\mathcal{M}'))\|_F + \left\|\mathbb{E}(h_m(\mathcal{M}')) - \mathbb{E}(h_m(\tilde{\mathcal{M}}'))\right\|_F = O(\varepsilon^{c_5}).\end{aligned}$$

For the last conclusion, from the definition of π_1 and π_2, we know that

$$\sum_{i:\pi_1(i)=j} w_i = \tilde{w}_j, \sum_{i:\pi_2(i)=j} w_i' = \tilde{w}_j'.$$

Applying Lemma 4.7.4 to $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$, we have that eigenvalues of Σ_i and Σ_i' are at least $\mathrm{poly}(\varepsilon^{c_3})$ for all i. Then if c_3 is sufficiently small, by Lemma 4.7.3, (4.7.1) implies $d_{\mathsf{TV}}(G_i, \tilde{G}_{\pi(i)}) \leqslant \mathrm{poly}_k(\varepsilon)$ and $d_{\mathsf{TV}}(G_i', \tilde{G}'_{\pi(i)}) \leqslant \mathrm{poly}_k(\varepsilon)$ for all i. $\square$

The following lemma shows the identifiability under the simplification of Lemma 4.7.10. It is proved in the proof of Lemma 8.2 in [48].

Lemma 4.7.11. *Suppose $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' G_i'$ satisfies 1,2,3 in the conclusion of Lemma 4.7.10 with constants c_4, c_5 and the minimal weights are at least ε^{c_3}. There exists a sufficiently small function $f(k) > 0$ depending only on k such that if $c_4 \leqslant f(k)$, then $k_1 = k_2$ and there exists a permutation π such that $|w_i - w'_{\pi(i)}| \leqslant \mathrm{poly}_k(\varepsilon)$ and $G_i = G'_{\pi(i)}$.*

Proof. Consider the component $G'_{k_2} = N(\mu'_{k_2}, \Sigma'_{k_2})$ in $\mathcal{M}'$. We claim that there must be some $i \in [k_1]$ such that

$$(\mu_i, \Sigma_i) = (\mu'_{k_2}, \Sigma'_{k_2}).$$

Assume for the sake of contradiction that this is not the case. Let $S_1 = \{i \in [k_1] : \Sigma_i = \Sigma'_{k_2}\}$ and $S_2 = \{i \in [k_2 - 1] : \Sigma_i' = \Sigma'_{k_2}\}$. Suppose F, F' are the generating functions of $\mathcal{M}$ and $\mathcal{M}'$

$$F = \sum_{i=1}^{k_1} w_i \exp\left(\mu_i^T X + \frac{1}{2} X^T \Sigma_i X y^2\right) = \sum_{m=0}^{\infty} \frac{1}{m!} h_m(\mathcal{M}) y^n$$

$$F' = \sum_{i=1}^{k_2} w_i' \exp\left(\mu_i'^T X + \frac{1}{2} X^T \Sigma_i' X y^2\right) = \sum_{m=0}^{\infty} \frac{1}{m!} h_m(\mathcal{M}') y^n.$$

Then define the differential operators

$$\mathcal{D}_i = \partial - \mu_i^T X - X^T \Sigma_i X y$$

$$\mathcal{D}_i' = \partial - \mu_i'^T X - X^T \Sigma_i' X y$$

where partial derivatives are taken with respect to y. Now consider the differential operator

$$\mathcal{D} = (\mathcal{D}'_{k_2-1})^{2^{k_1+k_2-2}} \cdots (\mathcal{D}'_1)^{2^{k_1}} \mathcal{D}_{k_1}^{2^{k_1-1}} \cdots \mathcal{D}_1$$

By Fact 4.7.6, $\mathcal{D}(F) = 0$. By Fact 4.7.6 and Fact 4.7.7, we have

$$\mathcal{D}(F') = P(y, X) \exp\left({\mu'_{k_2}}^T X + \frac{1}{2} X^T \Sigma'_{k_2} X y^2\right)$$

where P is a polynomial of degree

$$\deg(P) = 2^{k_1+k_2-1} - 1 - \sum_{i \in S_1} 2^{i-1} - \sum_{i \in S_2} 2^{k_1+i-2}$$

with leading coefficient

$$C_0 = w'_{k_2} \prod_{i \in [k_1] \setminus S_1} (X^T(\Sigma'_{k_2} - \Sigma_i)X)^{2^{i-1}} \prod_{i \in S_1} ((\mu'_{k_2} - \mu_i)^T X)^{2^{i-1}}$$
$$\prod_{i \in [k_2-1] \setminus S_2} (X^T(\Sigma'_{k_2} - \Sigma'_i)X)^{2^{k_1+i-2}} \prod_{i \in S_2} ((\mu'_{k_2} - \mu'_i)^T X)^{2^{k_1+i-2}}.$$

We now compare the following differentials evaluated at $y = 0$

$$(\mathcal{D}'_{k_2})^{\deg(P)} \mathcal{D}(F)$$
$$(\mathcal{D}'_{k_2})^{\deg(P)} \mathcal{D}(F')$$

The first quantity is 0 because $\mathcal{D}(F)$ is identically 0 as a formal power series. The second one is $\Omega_k(1) C_0$. Since for any i $(\mu_i, \Sigma_i) \neq (\mu'_{k_2}, \Sigma'_{k_2})$, our assumptions imply that the separation between μ_i, μ'_{k_2} or Σ_i, Σ'_{k_2} is at least $\epsilon^{c_4 c_5}$. Then we have $C_0 \geqslant \epsilon^{c_4 c_5 O_k(1)}$ for some X. On the other hand, the coefficients of the formal power series F, F' are the Hermite polynomials

$h_m(\mathcal{M})$ and $h_m(\mathcal{M}')$. This is a contradiction with our assumption that

$$\|\mathbb{E}(h_m(\mathcal{M}) - \mathbb{E}(h_m(\mathcal{M})\|_F \leqslant O_k(\varepsilon^{c_5})$$

as long as c_4 is smaller than some sufficiently small function $f(k)$ depending only on k. Thus there must be some component of $\mathcal{M}$ that matches $G'_{k_2} = N(\mu'_{k_2}, \Sigma'_{k_2})$. We can repeat the argument for each component in $\mathcal{M}'$ and in $\mathcal{M}$ to conclude that $\mathcal{M}$ and $\mathcal{M}'$ have the same components.

Next we will show that the weights of the same components in $\mathcal{M}$ and $\mathcal{M}'$ are close. We can assume that $\mathcal{M} = \sum_{i=1}^k w_i G_i, \mathcal{M}' = \sum_{i=1}^k w'_i G_i$ are two mixtures on the same set of components. Without loss of generality,

$$w_1 - w'_1 \leqslant \cdots w_\ell - w'_\ell \leqslant 0 \leqslant w_{\ell+1} - w'_{\ell+1} \leqslant \cdots \leqslant w_k - w'_k.$$

Then we can consider the following two mixtures

$$(w_1 - w'_1)G_1 + \cdots + (w_\ell - w'_\ell)G_\ell$$
$$(w_{\ell+1} - w'_{\ell+1})G_{\ell+1} + \cdots + (w_k - w'_k)G_k.$$

If

$$\sum_{i=1}^k |w_i - w'_i| > \varepsilon^\zeta$$

for some sufficiently small ζ depending only on k, we can then normalize each of the above into a distribution and repeat the same argument, using the fact that pairs of components cannot be too close, to obtain a contradiction. Thus, the mixing weights of $\mathcal{M}$ and $\mathcal{M}'$ are $\text{poly}_k(\varepsilon)$-close and this completes the proof. □

Proof of Theorem 4.7.1. We first set $c_4 = f(k)$ as in Lemma 4.7.11, and then c_3, c_5 according to c_4 as in Lemma 4.7.10, and c'_1, c_2 according to c_3 as in Lemma 4.7.9. Let

$c_1 = \min\{c'_1, c_2c_3\}$.

By Lemma 4.7.8, we can find i such that there is no w_j, w'_j in $[\varepsilon^{c_1^{i-1}}, \varepsilon^{c_1^i})$. Let $\tilde{\mathcal{M}} = \sum_{\{j:w_j \geqslant \varepsilon^{c_1^i}\}} w_j G_j$ and $\tilde{\mathcal{M}}' = \sum_{\{j:w'_j \geqslant \varepsilon^{c_1^i}\}} w'_j G'_j$. Then $d_{\mathsf{TV}}(\tilde{\mathcal{M}}, \tilde{\mathcal{M}}') \leqslant O(\varepsilon^{c_1^{i-1}})$. Let $\varepsilon_1 = \varepsilon^{c_1^{i-1}}$. We have $d_{\mathsf{TV}}(\tilde{\mathcal{M}}, \tilde{\mathcal{M}}') \leqslant O(\varepsilon_1)$ and the minimum weights of $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$ are at least $\varepsilon_1^{c_1}$.

Now we can apply Lemma 4.7.9 on $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$ and get partitions of components of $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$. For $i \in [\ell]$, let $\mathcal{M}_i$ and $\mathcal{M}'_i$ be the mixtures defined in Lemma 4.7.9. We can apply a linear transformation to make $\frac{1}{2}\mathcal{M}_i + \frac{1}{2}\mathcal{M}'_i$ in isotropic position. Since the total variation distance is invariant under linear transformations, so we still have both conclusions in Lemma 4.7.9. Let $\varepsilon_2 = \varepsilon_1^{c_2}$. Then $d_{\mathsf{TV}}(\mathcal{M}_i, \mathcal{M}_{\pi(i)}) \leqslant O(\varepsilon_2)$ and $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$ satisfies Lemma 4.7.4 with $\delta = \varepsilon_2^{c_3}$. Weights of both mixtures increase when we do the partition. So minimum weights are at least $\varepsilon_1^{c_1} \geqslant \varepsilon_1^{c_2c_3} = \varepsilon_2^{c_3}$.

We now prove the statement on these smaller mixtures. First we can use Lemma 4.7.10 to merge close parameters of $\mathcal{M}_i, \mathcal{M}'_i$ so that all pairs of parameters are either equal or separated by $\varepsilon_2^{c_4c_5}$. Under this simplification, Lemma 4.7.11 shows that there is a perfect matching between the same components in two mixtures and their weights are almost the same. By the last statement in Lemma 4.7.10, it is also a matching between components of $\mathcal{M}_i$ and $\mathcal{M}'_i$ by combining π and π_1, π_2. Moreover, if $\tilde{G}_j = \tilde{G}'_{\pi(j)}$, then $d_{\mathsf{TV}}(G_\ell, G'_{\ell}) \leqslant \mathrm{poly}(\varepsilon_2)$ for all ℓ, ℓ' such that $\pi_1(\ell) = j, \pi_2(\ell') = \pi(j)$. Repeating the argument for all pieces in $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$ completes the proof. $\square$

4.7.1 Proof of Lemma 4.7.9

In this section, we will prove Lemma 4.7.9. The following fact in [Liu-Moitra] shows that a good set of clusters of one mixture exists.

Fact 4.7.12 (Claim 7.6 in [LM'20]). *Let $\mathcal{M} = \sum_{i=1}^k w_i G_i$ be a mixture of Gaussians. For any constants $0 < \delta < 1$ and $\varepsilon > 0$, there exists $t \in [k^2]$ such that there exists a partition (possibly trivial) of $[k]$ into sets $R_1, \dots, R_\ell$ such that*

1. *If we draw edges between all pairs i, j such that $d_{\mathsf{TV}}(G_i, G_j) \leqslant 1 - \varepsilon^{\delta^t}$, then each*

piece of the partition is connected

2. *For any i, j in different pieces of the partition, $d_{\mathsf{TV}}(G_i, G_j) \geqslant 1 - \varepsilon^{\delta^{t-1}}$.*

Remark. Fact 4.7.12 can be applied to a set of Gaussians instead of a mixture of Gaussians by randomly assigning positive weights for all Gaussians.

Lemma 4.7.13. *For any constant $0 < c < 1$, suppose $\mathcal{M} = \sum_{i=1}^{k_1} w_i A_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' B_i$ are two mixtures of arbitrary distributions with $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \varepsilon$ and $w_i, w_i' \geqslant \varepsilon^c$. If for any $i \neq j$, $d_{\mathsf{TV}}(A_i, B_j) \geqslant 1 - \varepsilon$, then $k_1 = k_2$ and $d_{\mathsf{TV}}(A_i, B_i) \leqslant \mathrm{poly}_{k_1}(\varepsilon)$ for all $i \in [k_1]$.*

Proof. Suppose π is any coupling of $\mathcal{M}$ and $\mathcal{M}'$ and X, Y are random variables with distributions $\mathcal{M}$ and $\mathcal{M}'$. Then $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') = \min_\pi\{\mathbb{P}_\pi(X \neq Y)\}$. We define π to be the optimal coupling such that $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') = \mathbb{P}_\pi(X \neq Y)$. Then we can define $\hat{\pi}$ on variables i, j, X, Y such that $\sum_{i\in[k_1], j\in[k_2]} \hat{\pi}(i, j, X, Y) = \pi(X, Y)$ and the marginal distribution $\hat{\pi}_X$ with fixed i of X is $w_i A_i$ for all $i \in [k_1]$ and the marginal distribution $\hat{\pi}_Y$ with fixed j is $w_j' B_j$ for all $j \in [k_2]$. Let $P_{ij} = \int_{X,Y} \hat{\pi}(i, j, X, Y) dX dY$ and $A_{ij} = \frac{1}{P_{ij}} \int_Y \hat{\pi}(i, j, X, Y) dY$ be distributions on X, $B_{ij} = \frac{1}{P_{ij}} \int_X \hat{\pi}(i, j, X, Y) dX$ be distributions on Y. Then we have

$$
\begin{aligned}
d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') &= \mathbb{P}_{\pi}(X \neq Y) = \mathbb{P}_{\hat{\pi}}(X \neq Y) \\
&= \sum_{i,j} P_{ij} \mathbb{P}_{\hat{\pi}}(X \neq Y \mid i, j) \\
&\geqslant \sum_{i,j} P_{ij} \cdot d_{\mathsf{TV}}(A_{ij}, B_{ij}).
\end{aligned}
\tag{4.7.2}
$$

By the definition of P_{ij}, A_{ij}, B_{ij},

$$
\begin{aligned}
w_i A_i &= P_{ij} A_{ij} + \sum_{j' \neq j} P_{ij'} A_{ij'} \\
w_i' B_i &= P_{ij} B_{ij} + \sum_{i' \neq i} P_{i'j} B_{i'j}
\end{aligned}
$$

Dividing both sides by $\max\{w_i, w'_j\}$, we get

$$
\begin{aligned}
A_i &= \frac{P_{ij}}{\max\{w_i, w'_j\}} A_{ij} + \left(1 - \frac{w_i}{\max\{w_i, w'_j\}}\right) A_i + \sum_{j' \neq j} \frac{P_{ij'}}{\max\{w_i, w'_j\}} A_{ij'} \\
B_i &= \frac{P_{ij}}{\max\{w_i, w'_j\}} B_{ij} + \left(1 - \frac{w'_j}{\max\{w_i, w'_j\}}\right) B_i + \sum_{i' \neq i} \frac{P_{i'j}}{\max\{w_i, w'_j\}} B_{i'j}
\end{aligned}
$$

From the above two equations, we can write A_i, B_i as linear combinations of two distributions.

$$
\begin{aligned}
A_i &= \frac{P_{ij}}{\max\{w_i, w'_j\}} A_{ij} + \left(1 - \frac{P_{ij}}{\max\{w_i, w'_j\}})\right) A'_i \\
B_i &= \frac{P_{ij}}{\max\{w_i, w'_j\}} B_{ij} + \left(1 - \frac{P_{ij}}{\max\{w_i, w'_j\}}\right) B'_i
\end{aligned}
$$

Then by the triangle inequality,

$$
d_{\mathsf{TV}}(A_i, B_j) \leqslant \frac{P_{ij}}{\max\{w_i, w'_j\}} d_{\mathsf{TV}}(A_{ij}, B_{ij}) + \left(1 - \frac{P_{ij}}{\max\{w_i, w'_j\}}\right)
$$

$$
P_{ij} \cdot d_{\mathsf{TV}}(A_{ij}, B_{ij}) \geqslant P_{ij} - (1 - d_{\mathsf{TV}}(A_i, B_j)) \max\{w_i, w'_j\}. \tag{4.7.3}
$$

Combining (4.7.2) and (4.7.3), we have the following inequality on the TV distance between mixtures and the TV distance between components

$$
d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \geqslant \sum_{i,j} \left(P_{ij} - (1 - d_{\mathsf{TV}}(A_i, B_j)) \max\{w_i, w'_j\}\right). \tag{4.7.4}
$$

By the lower bounds on $d_{\mathsf{TV}}(A_i, B_j)$, we have

$$\begin{aligned}\varepsilon \geqslant d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') &\geqslant \sum_{i,j} P_{ij} - \sum_{i \neq j}(1 - d_{\mathsf{TV}}(A_i, B_j)) \max\{w_i, w'_j\} - \sum_i (1 - d_{\mathsf{TV}}(A_i, B_i)) \max\{w_i, w'_i\} \\ &\geqslant 1 - \sum_{i \neq j} \varepsilon - \sum_i \max\{w_i, w'_i\} + \sum_i \max\{w_i, w'_i\} d_{\mathsf{TV}}(A_i, B_i) \\ &\geqslant 1 - \sum_{i \neq j} \varepsilon - \sum_i \max\{w_i, w'_i\} + w_{\min} d_{\mathsf{TV}}(A_1, B_1)\end{aligned} \tag{4.7.5}$$

where A_1, B_1 can be replaced by any A_i, B_i pair. Let $k = \max\{k_1, k_2\}$. When $i \neq j$ and $d_{\mathsf{TV}}(A_i, B_j) \geqslant 1 - \varepsilon$, we plug it into Equation (4.7.4) and get

$$\varepsilon \geqslant d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \geqslant \sum_{i \neq j} P_{ij} - (1 - d_{\mathsf{TV}}(A_i, B_j)) \max\{w_i, w'_j\} \geqslant \sum_{i \neq j} (P_{ij} - \varepsilon).$$

This implies

$$\sum_{i \neq j} P_{ij} \leqslant k^2 \varepsilon.$$

Then we can bound $\sum_i \max\{w_i, w'_i\} - 1$ in Equation (4.7.5)

$$\sum_i \max\{w_i, w'_i\} - 1 = \sum_i \max\{w_i, w'_i\} - w_i \leqslant \sum_{i \neq j} P_{ij} \leqslant k^2 \varepsilon.$$

Plugging this bound into Equation (4.7.5), for any i, we have

$$d_{\mathsf{TV}}(A_i, B_i) \leqslant \frac{1}{w_{\min}} \left(k^2 \varepsilon + \sum_i \max\{w_i, w'_i\} - 1 \right) \leqslant \frac{2k^2 \varepsilon}{\varepsilon^c}.$$

$\square$

Proof of Lemma 4.7.9. We apply Fact 4.7.12 on the union set of components of $\mathcal{M}$ and $\mathcal{M}'$

with parameter δ to find a partition $R_1, \ldots, R_\ell$. Let

$$\mathcal{M}_i = \frac{\sum_{G_j \in R_i} w_j G_j}{\sum_{G_j \in R_i} w_j}$$
$$\mathcal{M}'_i = \frac{\sum_{G'_j \in R_i} w'_j G'_j}{\sum_{G'_j \in R_i} w'_j}.$$

Then for any $i \neq j$, we know $d_{\mathsf{TV}}(G_a, G'_b) \geqslant 1 - \varepsilon^{\delta^{t-1}}$ for $G_a \in R_i, G'_b \in R_j$. By (4.7.4) in the proof of Lemma 4.7.13, we have

$$d_{\mathsf{TV}}(M_i, M'_j) \geqslant 1 - 2k\varepsilon^{\delta^{t-1}}.$$

Then by Lemma 4.7.13, for any i, there exists a such that $d_{\mathsf{TV}}(M_i, M'_i) \leqslant \varepsilon^{a\delta^{t-1}}$. Let $c_2 = a\delta^{t-1}$. If we set $\delta = c_2 c_3 / \delta^{t-1} = a c_3$, the partition satisfies the second conclusion. □

4.8 Omitted Proofs

In this subsection, we provide the proofs that were omitted from Section 4.2 and Section 4.6.

4.8.1 Omitted Proofs from Section 4.2.1

Lemma 4.8.1 (Concentration of low-degree polynomials, Lemma 4.2.9 restated)**.** *Let T be a d-dimensional, degree-4 tensor such that $\|T\|_F \leqslant \Delta$ for some $\Delta > 0$ and let $x, y \sim \mathcal{N}(0, I)$. Then, with probability at least $1 - 1/\mathrm{poly}(d)$, the following holds:*

$$\|T(\cdot, \cdot, x, y)\|_F^2 \leqslant O\big(\log(d)\Delta^2\big) \;.$$

Proof. We note that

$$\begin{aligned}
\mathbb{E}\left[\|T(\cdot,\cdot,x,y)\|_F^2\right] &= \mathbb{E}\left[\sum_{i_1,i_2}\left(\sum_{i_3,i_4} T\left(i_1,i_2,i_3,i_4\right)x\left(i_3\right)y\left(i_4\right)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i_1,i_2}\left(\sum_{i_3,i_4} T\left(i_1,i_2,i_3,i_4\right)^2 x\left(i_3\right)^2 y\left(i_4\right)^2\right)\right] \\
&= \sum_{i_1,i_2,i_3,i_4} T\left(i_1,i_2,i_3,i_4\right)^2 \leqslant \Delta^2 \ .
\end{aligned}$$

The second equality follows from the fact that $x\left(i_3\right), y\left(i_4\right)$ are independent and have zero means. So the only non-zero terms are the squares. The third equality follows from the fact that $x\left(i_3\right), y\left(i_4\right)$ are independent with unit variances. Observe that $\|T\left(\cdot,\cdot,x,y\right)\|_F^2$ is a degree-2 polynomial in Gaussian random variables. Using standard concentration bounds for low-degree Gaussian polynomials (see, e.g., Theorem 2.3 in [103]), we obtain

$$\mathbb{P}\left[\|T(\cdot,\cdot,x,y)\|_F^2 \geqslant t^2\,\mathbb{E}\left[\|T(\cdot,\cdot,x,y)\|_F^2\right]\right] \leqslant \exp\left(-ct\right) \ .$$

Setting $t = \Omega(\log(d))$ completes the proof. ◻

4.8.2 Omitted Proofs from Section 4.2.2

Lemma 4.8.2 (Shifts Cannot Decrease Variance, Lemma 4.2.17 restated)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$, Q be a $d \times d$ matrix-valued indeterminate, and C be a scalar-valued indeterminate. Then, we have that*

$$\Big|\frac{Q,C}{2}\left\{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right] \leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x)-C)^2\right]\right\} .$$

Proof.

$$\Big|\frac{Q,C}{2}\left\{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x)-C)^2\right] = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)] - C\right)^2\right]\right.$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x)-C)^2\right] \\
&\quad + 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)] - C\right)\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x)-C)^2\right] \\
&\geqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right] \quad \Bigg\}.
\end{aligned}
$$

□

Lemma 4.8.3 (Shifts of Certifiably Hypercontractive Distributions, Lemma 4.2.18 restated)**.** *Let x be a mean-0 random variable with distribution $\mathcal{D}$ on $\mathbb{R}^d$ with t-certifiably C-hypercontractive degree-2 polynomials. Then, for any fixed constant vector $c \in \mathbb{R}^d$, the random variable $x + c$ also has t-certifiable $4C$-hypercontractive degree-2 polynomials.*

Proof. Observe that using that $\mathbb{E}_{x\sim\mathcal{D}}[x] = 0$, we have that

$$
\Big|\frac{Q}{2} \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x+c)^\top Q(x+c)\right] = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx + c^\top Qc\right] \right\} .
$$

Next, by two applications of the SoS Triangle Inequality (Fact 3.3.5), an application of Lemma 4.2.17 followed by certifiable hypercontractivity of $\mathcal{D}$, we have:

$$
\begin{aligned}
\Big|\frac{Q}{t'} \Bigg\{ & \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x+c)^\top Q(x+c)\right]\right)^{t'}\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right) + x^\top Qc + c^\top Qx\right)^{t'}\right] \\
&\leqslant 4^{t'}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{\mathcal{D}} x^\top Qx\right)^{t'}\right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x^\top Qc)^{t'}\right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(c^\top Qx)^{t'}\right]\right) \\
&\leqslant 4^{t'}(Ct')^{t'}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{\mathcal{D}} x^\top Qx\right)^2\right]^{t'/2} + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x^\top Qc)^2\right]^{t'/2} + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(c^\top Qx)^2\right]^{t'/2}\right)\Bigg\} .
\end{aligned}
$$

On the other hand, notice that

$$\left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[(x+c)^\top Q(x+c) \right] \right)^2 \right] \right.$$
$$\left. = \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{\mathcal{D}} x^\top Q x \right)^2 \right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[(x^\top Q c)^2 \right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[(c^\top Q x)^2 \right] \right) \qquad \right\}.$$

Thus,

$$\left|\frac{Q}{t'}\right. \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[x^\top Q x \right] \right)^2 \right]^{t'/2} + \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[(x^\top Q c)^2 \right] \right)^{t'/2} + \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[(c^\top Q x)^2 \right] \right)^{t'/2} \right.$$
$$\left. \leqslant 4^{t'} (Ct')^{t'} \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[(x+c)^\top Q(x+c) \right] \right)^2 \right] \right)^{t'/2} \right\}.$$

As a result, we obtain:

$$\left|\frac{Q}{t'}\right. \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[(x+c)^\top Q(x+c) \right] \right)^{t'} \right] \right.$$
$$\left. \leqslant (4Ct')^{t'} \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[(x+c)^\top Q(x+c) \right] \right)^2 \right] \right)^{t'/2} \right\},$$

which completes the proof. □

Lemma 4.8.4 (Mixtures of Certifiably Hypercontractive Distributions, Lemma 4.2.19 restated)**.** *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have t-certifiable C-hypercontractive degree-2 polynomials on $\mathbb{R}^d$, for some fixed constant C. Then, any mixture $\mathcal{D} = \sum_i w_i \mathcal{D}_i$ also has t-certifiably (C/α)-hypercontractive degree-2 polynomials for $\alpha = \min_{i \leqslant k, w_i > 0} w_i$.*

Proof. Applying Lemma 3.3.5 followed by SoS Hölder's inequality on the second term and followed by a final application of SoS Hölder's inequality (Fact 3.3.4), we obtain:

$$\left|\frac{Q}{t'}\right. \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[x^\top Q x \right] \right)^{t'} \right] = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left(x^\top Q x - \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right)^{t'} \right] \right.$$

$$
\begin{aligned}
&= \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[\left(x^\top Q x - \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[x^\top Q x\right]\right)^{t'}\right] \\
&\leqslant 2^{t'}\left(\sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[x^\top Q x\right]\right)^{t'}\right]\right. \\
&\quad \left. + \left(\sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[x^\top Q x\right]\right]^{t'}\right)\right) \\
&\leqslant 2^{t'}\left((Ct')^{t'}\left(\sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[x^\top Q x\right]\right)^2\right]\right)^{t'/2}\right. \\
&\quad \left. + \sum_i w_i \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[x^\top Q x\right]\right)^2\right]\right)^{t'}\right) \\
&\leqslant \left(\frac{4Ct'}{\alpha}\right)^{t'}\left(\sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[x^\top Q x\right]\right)^2\right]\right)^{t'/2}\Bigg\}.
\end{aligned}
$$

On the other hand, note that by Lemma 4.2.17, we know that

$$
\begin{aligned}
\Big|\frac{Q}{2}\Big\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[x]^\top Q x\right)^2\right] &= \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Q x\right]\right)^2\right] \\
&\geqslant \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}\left[x^\top Q x\right]\right)^2\right]\Bigg\}.
\end{aligned}
$$

Combining the two equations above completes the proof. □

Corollary 4.8.5 (Certifiable Hypercontractivity of k-Mixtures of Gaussians, Corollary 4.2.20 restated)**.** *Let $\mathcal{D}$ be a k-mixture of Gaussians $\sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ with weights $w_i \geqslant \alpha$ for every $i \in [k]$. Then, $\mathcal{D}$ has t-certifiably $4/\alpha$-hypercontractive degree-2 polynomials.*

Proof. From [97], we know that the standard Gaussian random variable has t-certifiably 1-hypercontractive degree-2 polynomials. From Fact 4.2.16, we immediately obtain that for any PSD matrix Σ, the Gaussian $\mathcal{N}(0, \Sigma)$ also has t-certifiable 1-hypercontractive degree-2 polynomials. From Lemma 4.2.18, we obtain that for any μ, the Gaussian $\mathcal{N}(\mu, \Sigma)$ has t-certifiable 4-hypercontractive degree-2 polynomials. Finally, applying Lemma 4.2.19 to

$\mathcal{D}_i = \mathcal{N}(\mu_i, \Sigma_i)$ and mixture weights $w_1, w_2, \ldots w_k$, yields that $\mathcal{D} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ has t-certifiably $4/\alpha$-hypercontractive degree-2 polynomials. This completes the proof. □

Lemma 4.8.6 (Linear Transformations of Certifiably Bounded-Variance Distributions, Lemma 4.2.21 restated)**.** *For $d \in \mathbb{N}$, let x be a random variable with distribution $\mathcal{D}$ on $\mathbb{R}^d$ such that for $d \times d$ matrix-valued indeterminate Q, $\left|\frac{Q}{2}\right.$ $\left\{\mathbb{E}_{x\sim\mathcal{D}}(x^\top Q x - \mathbb{E}_{\mathcal{D}}\, x^\top Q x)^2 \leqslant \left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2\right\}$. Let A be an arbitrary $d \times d$ matrix and let $x' = Ax$ be the random variable with covariance $\Sigma' = AA^\top$. Then, we have that*

$$\left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{x'\sim\mathcal{D}'} (x'^\top Q x' - \mathop{\mathbb{E}}_{\mathcal{D}'} x'^\top Q x')^2 \leqslant \left\|\Sigma'^{1/2} Q \Sigma'^{1/2}\right\|_F^2 \right\} .$$

Proof. The covariance of x' is $AA^\top = \Sigma'$, say. Let $\Sigma'^{1/2}$ be the PSD square root of Σ'. The proof follows by noting that $x'^\top Q x' = (Ax)^\top Q (Ax) = x^\top (A^\top Q A) x^\top$ and that $\|A^\top Q A\|_F^2 = \mathrm{tr}(A^\top Q A A^\top Q A) = \mathrm{tr}(AA^\top Q AA^\top Q) = \mathrm{tr}(\Sigma' Q \Sigma' Q) = \mathrm{tr}(\Sigma'^{1/2} Q \Sigma'^{1/2} \Sigma'^{1/2} Q \Sigma'^{1/2}) = \left\|\Sigma'^{1/2} Q \Sigma'^{1/2}\right\|_F^2$. □

Lemma 4.8.7 (Variance of Degree-2 Polynomials of Standard Gaussians, Lemma 4.2.22 restated)**.** *We have that*

$$\left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{\mathcal{N}(0,I)} \left(x^\top Q x - \mathop{\mathbb{E}}_{\mathcal{N}(0,I)} x^\top Q x \right)^2 \leqslant 3 \|Q\|_F^2 \right\} .$$

Proof. We will view $xx^\top$ and $I \in \mathbb{R}^{d\times d}$ as d^2-dimensional vectors. Consider the matrix $\mathbb{E}_{x\sim\mathcal{N}(0,I)}(xx^\top - I)(xx^\top - I)^\top$. The diagonal of this matrix is $2I_{d^2}$. The off-diagonal part has exactly one non-zero entry in any row (which corresponds to entry indexed by (i, j) and (j, i) for $i \neq j$), and thus has spectral norm at most 1 by the Gershgorin circle theorem. Thus, $\mathbb{E}_{x\sim\mathcal{N}(0,I)}(xx^\top - I)(xx^\top - I)^\top \preceq 3I_{d^2}$.

We thus have:

$$\left|\frac{Q}{2}\right. \left\{ \mathop{\mathbb{E}}_{\mathcal{N}(0,I)} \left(x^\top Q x - \mathop{\mathbb{E}}_{\mathcal{N}(0,I)} x^\top Q x \right)^2 = \mathop{\mathbb{E}}_{\mathcal{N}(0,I)} \left\langle xx^\top - I, Q \right\rangle^2 = \mathop{\mathbb{E}}_{\mathcal{N}(0,I)} \left\langle xx^\top - I, Q \right\rangle^2 \right.$$

$$\leqslant \left\| \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,I)} (xx^\top - I)(xx^\top - I')^\top \right\|_2 \|Q\|_F^2 \leqslant 3 \|I_{d^2}\|_2 \|Q\|_F^2 = 3 \|Q\|_F^2 \right\} . \quad (4.8.1)$$

□

Lemma 4.8.8 (Variance of Degree-2 Polynomials of Mixtures, Lemma 4.2.25 restated)**.** *Let $\mathcal{M} = \sum_i w_i \mathcal{D}_i$ be a k-mixture of distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ with means μ_i and covariances Σ_i. Let $\mu = \sum_i w_i \mu_i$ be the mean of $\mathcal{M}$. Suppose that each of $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have certifiably C-bounded-variance i.e. for Q: a symmetric $d \times d$ matrix-valued indeterminate.*

$$\left|\frac{Q}{2} \left\{ \mathop{\mathbb{E}}_{x' \sim \mathcal{D}_i} (x'^\top Q x' - \mathop{\mathbb{E}}_{\mathcal{D}_i} x'^\top Q x')^2 \leqslant C \left\| \Sigma'^{1/2} Q \Sigma'^{1/2} \right\|_F^2 \right\} .$$

Further, suppose that for some $H > 1$, $\|\mu_i - \mu\|_2^2, \|\Sigma_i - I\|_F \leqslant H$ for every $1 \leqslant i \leqslant k$. Then, we have that

$$\left|\frac{Q}{2} \left\{ \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[x^\top Q x \right] \right)^2 \right] \leqslant 100 C H^2 \|Q\|_F^2 \right\} .$$

Proof. We have the following sequence of (in-)equalities:

$$\left|\frac{Q}{2} \left\{ \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left(x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} x^\top Q x \right)^2 = \sum_{i \leqslant k} w_i \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} \left(x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} x^\top Q x \right)^2 \right. \quad (4.8.2)$$

$$= \sum_{i \leqslant k} w_i \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} \left(x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} x^\top Q x + \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} x^\top Q x \right)^2 \quad (4.8.3)$$

$$\left. \leqslant 2 \sum_{i \leqslant k} w_i \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} \left(x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} x^\top Q x \right)^2 + 2 \sum_i w_i \left(\mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} x^\top Q x \right)^2 \right\} , \quad (4.8.4)$$

where the third line follows from Fact 3.3.5 (SoS Almost Triangle Inequality).

Let us first bound the 2nd term in the RHS above. Towards that, let $\Sigma = \sum_i w_i((\mu_i - \mu)(\mu_i - \mu)^\top + \Sigma_i)$ be the covariance of the mixture $\mathcal{M}$. Then, notice that $\Sigma = \sum_i w_i((\mu_i -$

$\mu)(\mu_i - \mu)^\top + \Sigma_i) = \sum_i w_i \mu_i \mu_i^\top + \sum_i w_i \Sigma_i - \mu\mu^\top$. Thus, we can write

$$
\begin{aligned}
\mu_i\mu_i^\top + \Sigma_i - \Sigma - \mu\mu^\top &= \sum_{j\neq i} w_j(\mu_i\mu_i^\top - \mu_j\mu_j^\top) + \sum_{j\neq i} w_j(\Sigma_i - \Sigma_j) \\
&= \sum_{j\neq i} w_j(\mu_j - \mu)(\mu_j - \mu)^\top - \sum_{j\neq i}(\mu - \mu_j)(\mu - \mu_j)^\top + \sum_{j\neq i} w_j(\Sigma_i - \Sigma_j) \\
&= \sum_{j\neq i} w_j(\mu_j - \mu)(\mu_j - \mu)^\top - \sum_{j\neq i}(\mu - \mu_j)(\mu - \mu_j)^\top + \sum_{j\neq i} w_j(\Sigma_i - I) - \sum_{j\neq i} w_j(\Sigma_j
\end{aligned}
$$

Here, in the second to last step, we added and subtracted $\sum_{j\neq i} w_j \mu\mu^\top$ and used that $\sum_i w_i \mu_i = \mu$, and in the last step we added and subtracted $\sum_{j\neq i} w_j I$.

By application of the triangle inequality for Frobenius norm to the RHS of the above, we have that:

$$
\begin{aligned}
\left\|\mu_i\mu_i^\top + \Sigma_i - \mu\mu^\top - \Sigma\right\|_F &\leqslant \sum_{j\neq i} w_j \left\|(\mu_i - \mu)(\mu_i - \mu)^\top\right\|_F + \sum_{j\neq i} w_j \left\|(\mu_j - \mu)(\mu_j - \mu)^\top\right\|_F \\
&\quad + \sum_{j\neq i} w_j \left\|(\Sigma_i - I)\right\|_F + \sum_{j\neq i} w_j \left\|(I - \Sigma_j)\right\|_F \leqslant H + H + H + H = 4H \ .
\end{aligned}
$$

Using the SoS version of the Cauchy-Schwarz inequality (Fact 3.3.4) on indeterminate Q and constant $\mu\mu^\top - \mu_i\mu_i^\top + \Sigma_i - \Sigma$ and the above bound, we have:

$$
\begin{aligned}
\vdash_{2}^{Q} \Bigg\{ \sum_i w_i \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{M}} x^\top Q x \right)^2 &= \sum_i w_i \left(\left\langle \mu_i\mu_i^\top + \Sigma_i, Q \right\rangle - \left\langle \mu\mu^\top + \Sigma, Q \right\rangle \right)^2 \\
\leqslant \sum_i w_i \left\|\mu\mu^\top - \mu_i\mu_i^\top + \Sigma_i - \Sigma\right\|_F^2 \|Q\|_F^2 &\leqslant 16H^2 \sum_i w_i \|Q\|_F^2 = 16H^2 \|Q\|_F^2 \Bigg\} .
\end{aligned}
$$

Let us now bound the first term in the RHS of (4.8.4) above. First, observe that $x^\top Q x - \mathbb{E}_{\mathcal{N}(\mu_i,\Sigma_i)} x^\top Q x = (x - \mu_i)^\top Q (x - \mu_i) - \mathbb{E}_{\mathcal{N}(\mu_i,\Sigma_i)}(x - \mu_i)^\top Q (x - \mu_i) + 2(x - \mu_i)^\top Q \mu_i$.

Thus, using Fact 3.3.5 and Lemma 4.2.24, we have:

$$
\begin{aligned}
&\left|\frac{Q}{2}\left\{\sum_{i\leqslant k} w_i \mathop{\mathbb{E}}_{\mathcal{D}_i}\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} x^\top Q x\right)^2\right.\right. && (4.8.5)\\
&\leqslant 2\sum_{i\leqslant k} w_i \mathop{\mathbb{E}}_{\mathcal{D}_i}\left((x-\mu_i)^\top Q(x-\mu_i) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}(x-\mu_i)^\top Q(x-\mu_i)\right)^2 + 8\sum_{i\leqslant k} w_i \mathop{\mathbb{E}}_{\mathcal{D}_i}\left((x-\mu_i)^\top Q\mu_i\right)^2 && (4.8.6)\\
&\leqslant 6\sum_{i} w_i C\left\|\Sigma_i^{1/2} Q \Sigma_i^{1/2}\right\|_F^2 + 8\sum_{i\leqslant k} w_i \mathop{\mathbb{E}}_{\mathcal{D}_i}\left((x-\mu_i)^\top Q\mu_i\right)^2\Bigg\}\,. && (4.8.7)
\end{aligned}
$$

For the first term, note that $\|\Sigma_i\|_2 \leqslant 1 + \|\Sigma_i - I\|_F \leqslant 1 + H$. Thus, $\left\|\Sigma_i^{1/2}\right\|_2 \leqslant \sqrt{1+H}$. Thus, we have that $\Sigma_i^{1/2} \preceq I + (\Sigma_i^{1/2} - I) \preceq \sqrt{1+H}I$. Using Lemma 3.3.9 with $A = (1+H)^{-1/2}\Sigma_i^{1/2}$ and $B = Q\Sigma_i^{1/2}$, we have: $\left|\frac{Q}{2}\right. \left\{\left\|\Sigma_i^{1/2} Q\Sigma_i^{1/2}\right\|_F^2 \leqslant (1+H)\left\|Q\Sigma_i^{1/2}\right\|_F^2\right\}$. By another application of Lemma 3.3.9, we have: $\left|\frac{Q}{2}\right. \left\{\left\|Q\Sigma_i^{1/2}\right\|_F^2 \leqslant (1+H)\|Q\|_F^2\right\}$. Thus, altogether, we have: $\left|\frac{Q}{2}\right. \left\{\left\|\Sigma_i^{1/2}Q\Sigma_i^{1/2}\right\|_F^2 \leqslant (1+H)^2\|Q\|_F^2\right\}$. Using our assumption that $1 < H$, we thus have:

$$
\left|\frac{Q}{2}\right.\left\{\sum_i w_i C\left\|\Sigma_i^{1/2}Q\Sigma_i^{1/2}\right\|_F^2 \leqslant C(1+H)^2\|Q\|_F^2 \leqslant 4CH^2\|Q\|_F^2\right\}.
$$

For the second term, first observe that the following equality of quadratic polynomials in indeterminate Q: $((x-\mu_i)^\top Q\mu_i)^2 = \left((\Sigma_i^{\dagger/2}(x-\mu_i))^\top\Sigma_i^{1/2}Q\mu_i\right)^2$. Thus, $\mathbb{E}_{x\sim\mathcal{D}_i}((x-\mu_i)^\top Q\mu_i)^2 = \left\|\Sigma_i^{1/2}Q\mu_i\right\|_2^2$. Next, by the SoS Cauchy-Schwarz inequality (Fact 3.3.4), we have that

$$
\left|\frac{Q}{2}\right.\left\{\left\|\Sigma_i^{1/2}Q\mu_i\right\|_2^2 = \operatorname{tr}(\mu_i\mu_i^\top Q\Sigma_i Q) \leqslant H\operatorname{tr}(Q\Sigma_i Q) = H\left\|\Sigma_i^{1/2}Q\right\|_F^2\right\}\,.
$$

Applying Lemma 3.3.9 with the observation above that $\Sigma_i^{1/2} \leqslant (1+H)^{1/2}I$ yields: $\left|\frac{Q}{2}\right. \left\{\left\|\Sigma_i^{1/2}Q\right\|_F^2 \leqslant (1+H)\|Q\|_F^2\right\}$. Thus, altogether, we obtain: $\left|\frac{Q}{2}\right. \{\mathbb{E}_{x\sim\mathcal{D}_i}(x-\mu_i)^\top Q\mu_i\}^2 \leqslant$

$H(1+H)^2 \|Q\|_F^2 \leqslant 4H^3 \|Q\|_F^2$. We thus have:

$$\left|\frac{Q}{2}\left\{\sum_{i\leqslant k} w_i \mathop{\mathbb{E}}_{\mathcal{D}_i} \left((x-\mu_i)^\top Q\mu_i\right)^2 \leqslant 4H^3 \|Q\|_F^2\right\}.\right.$$

Plugging in these bounds into (4.8.7) completes the proof. □

As an immediate corollary of Lemma 4.2.21 and Lemma 4.2.25, we obtain:

Lemma 4.8.9 (Variance of Degree-2 Polynomials of Mixtures of Gaussians, Lemma 4.2.26 restated)**.** *Let $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with $w_i \geqslant \alpha$, mean $\mu = \sum_i w_i \mu_i$ and covariance $\Sigma = \sum_i w_i((\mu_i-\mu)(\mu_i-\mu)^\top + \Sigma_i)$. Suppose that for some $H > 1$, $\left\|\Sigma^{\dagger/2}(\Sigma_i - I)\Sigma^{\dagger/2}\right\|_F \leqslant H$ for every $1 \leqslant i \leqslant k$. Let Q be a symmetric $d\times d$ matrix-valued indeterminate. Then for $H' = \max\{H, 1/\alpha\}$,*

$$\left|\frac{Q}{2}\left\{\mathop{\mathbb{E}}_{x\sim\mathcal{M}}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{M}}\left[x^\top Q x\right]\right)^2\right] \leqslant 100H'^2 \left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2\right\}.\right.$$

Proof. Let $\Sigma = U\Lambda U^\top$ be the covariance of the mixture $\mathcal{M}$ along with its eigendecomposition. We want to apply Lemma 4.2.25 and Lemma 4.2.21 with the linear transformation $x \to Ax$ for $A = \Lambda^{\dagger/2}U^\top$. For this, we need to check that the conditions of the Lemma 4.2.25 are met after this linear transformation. The new component covariance is $\Sigma'_i = A\Sigma_i A^\top$ and the hypothesis implies that they are within H in Frobenius distance of the new mixture covariance $I' = A\Sigma A^\top$ (I in the range space of Σ). The new means of the components after the linear transformation are $\mu'_i = A\mu_i$ and the new mixture mean is $\mu' = A\mu$. Thus, noting that $I' = \sum_i w_i(\mu'_i - \mu')(\mu' - \mu')^\top + \sum_i w_i \Sigma'_i$, and since each of the terms in the RHS of the preceding equality are PSD, we must have that $I' \succeq w_i(\mu'_i - \mu')(\mu'_i - \mu')^\top$ for every i. Thus, $1 = \|I'\|_2 \geqslant w_i \left\|(\mu'_i - \mu')(\mu'_i - \mu')^\top\right\|_2 = \left\|\mu'_i - \mu'\right\|_2^2$. Rearranging yields that $\left\|\mu'_i - \mu'\right\|_2^2 \leqslant 1/w_i \leqslant 1/\alpha$. Thus, we can now apply Lemma 4.2.25 to the linearly transformed mixture and the conclusion follows. □

4.8.3 Omitted Proofs from Section 4.2.3

Lemma 4.8.10 (Lemma 4.2.31 restated)**.** *If X is a good sample (satisfies Condition 4.2.29) with respect to $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ with parameters (γ, t), then if $w_i \geqslant \gamma$ for all $i \in [k]$, and if for some $B \geqslant 0$ we have that $\|\mu_i\|_2^2, \|\Sigma_i\|_{\text{op}} \leqslant B$ for all $i \in [k]$, then for all $m \leqslant t$, we have that:*

$$\left\| \mathop{\mathbb{E}}_{x \in_u X}[x^{\otimes m}] - \mathop{\mathbb{E}}_{x \sim M}[x^{\otimes m}] \right\|_F^2 \leqslant \gamma^2 m^{O(m)} B^m d^m \ .$$

Proof. We begin by noting that for any symmetric m-tensor A we have that $\|A\|_F^2 \leqslant m^{O(m)}(\mathbb{E}_{v\sim\mathcal{N}(0,I)}[\langle A, v^{\otimes m}\rangle^2])$. This is because, in the notation of [27], the squared expectation of $\langle A, v^{\otimes m}\rangle$ is $\mathbb{E}[\mathrm{Hom}_A(v)^2] \geqslant m!\,\mathbb{E}[h_A(v)^2] = m!\,\|A\|_F^2$, where the first inequality holds because $\sqrt{m!}h_A(v)$ is the degree-m harmonic part of $\mathrm{Hom}_A(v)$, and the equality is by Claim 3.22. Therefore, to prove the lemma, it suffices to bound

$$\begin{aligned}
&\mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,I)}\left[\left(\mathop{\mathbb{E}}_{x\in_u X}[(v\cdot x)^m] - \mathop{\mathbb{E}}_{x\sim\mathcal{M}}[(v\cdot x)^m]\right)^2\right] \\
&= \mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,I)}\left[\left(\sum_{i=1}^k \left(\frac{1}{n}\sum_{x\in X_i}(v\cdot x)^m - w_i \mathop{\mathbb{E}}_{x\sim\mathcal{N}(\mu_i,\Sigma_i)}(v\cdot x)^m\right)\right)^2\right] \\
&= \mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,I)}\left[\left(\sum_{i=1}^k \sum_{j=0}^m \binom{m}{j}\mu_i^{m-j}\left(\frac{1}{n}\sum_{x\in X_i}(v\cdot(x-\mu_i))^j - w_i \mathop{\mathbb{E}}_{x\sim\mathcal{N}(\mu_i,\Sigma_i)}(v\cdot(x-\mu_i))^j\right)\right)^2\right] \\
&\leqslant \mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,I)}\left[\left(\sum_{i=1}^k \sum_{j=0}^m \binom{m}{j}|v\cdot\mu_i|^{m-j}\left(w_i\gamma m!(v^T\Sigma v)^{j/2}\right)\right)^2\right] \\
&\leqslant \gamma^2 m^{O(m)} \mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,I)}\left[\sum_{i=1}^k \left(w_i(|v\cdot\mu_i| + (v^T\Sigma v)^{1/2})^{2m}\right)\right] \\
&\leqslant \gamma^2 m^{O(m)} \mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,I)}\left[2B^m\|v\|_2^{2m}\right] \\
&\leqslant \gamma^2 m^{O(m)} B^m d^m.
\end{aligned}$$

This completes the proof. □

Lemma 4.8.11 (Lemma 4.2.32 restated)**.** *Let $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$. Let $S \subset [k]$ with $\sum_{i \in S} w_i = w$, and let $\mathcal{M}' = \sum_{i \in S}(w_i/w)\mathcal{N}(\mu_i, \Sigma_i)$. Then if X is a good sample (satisfies Condition 4.2.29) with respect to $\mathcal{M}$ with parameters (γ, t) for some $\gamma < 1/(2k)$ with the corresponding partition being $X = X_1 \cup X_2 \cup \ldots \cup X_k$, then $X' = \bigcup_{i \in S} X_i$ is a good sample (satisfies Condition 4.2.29) with respect to $\mathcal{M}'$ with parameters $(O(k\gamma/w), t)$.*

Proof. After noting that $|X'| = w|X|(1 + O(k\gamma/w))$, the rest follows straightforwardly from the definitions using the partition $X' = \bigcup_{i \in S} X_i$. $\square$

Lemma 4.8.12 (Lemma 4.2.33 restated)**.** *Let $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ and let n be an integer at least $kt^{Ct}d^t/\gamma^3$, for a sufficiently large universal constant $C > 0$, some $\gamma > 0$, and some $t \in \mathbb{N}$. If X consists of n i.i.d. samples from $\mathcal{M}$, then X is a good sample (satisfies Condition 4.2.29) with respect to $\mathcal{M}$ with parameters (γ, t) with high probability.*

Proof. We will show that Condition 4.2.29 holds with high probability using that partition where X_i is the set of samples drawn from the i-th component of $\mathcal{M}$. Note that the second part of Condition 4.2.29 holds with high probability, so long as n is a sufficiently large multiple of d/γ^2 by the VC-Theorem [99]. In particular, if we think of samples as being drawn from $\mathbb{R}^d \times [k]$, where the second coordinate denotes the component that the sample was drawn from, the second part of Condition 4.2.29 says that the empirical probability of any event $H \times \{i\}$ is correct to within additive error γ. It is easy to see and well-known that the class of such events has VC-dimension $O(d)$, from which the desired bound follows.

For the first part of Condition 4.2.29, we claim that it holds with high probability so long as $n \geqslant kt^{Ct}d^t/\gamma^3$. To prove this, we show it separately for each i with $w_i \geqslant \gamma$ (as otherwise there is nothing to prove) and take a union bound. As Condition 4.2.29 is invariant under affine transformations, we may perform an invertible affine transformation so that $\mu_i = 0$ and Σ_i is the projection onto the first d' coordinates, for some d'. It is clear that only the first d' coordinates of any element of X_i will be non-zero. We claim that the first part of our condition will follow for a given m, so long as $||X_i|/n - w_i| \leqslant \gamma w_i$ (which holds with high probability if $n \gg \log(k)/\gamma^3$), and

$$\left\| \mathop{\mathbb{E}}_{x \in_u X_i}[x^{\otimes m}] - \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0, I_{d'})}[x^{\otimes m}] \right\|_F^2 \leqslant \gamma^2 , \tag{4.8.8}$$

as $\frac{1}{n} \sum_{x \in X_i} \langle v, x - \mu_i \rangle^m = w_i(1 \pm \gamma) \langle \mathbb{E}_{x \in_u X_i}[x^{\otimes m}], v^{\otimes m} \rangle$. It is easy to see that each entry of the tensor on the left hand side of Equation (4.8.8) has mean 0 and variance $m^{O(m)}/|X_i|$, and thus the expected size of the left hand side is $m^{O(m)} d^m/|X_i|$. Then, when $n \geqslant k^{Ck} d^{4k}/\gamma^3$ for a sufficiently large constant C, all parts of our condition hold with high probability. This completes the proof. □

4.8.4 Omitted Proofs from Section 4.6

Lemma 4.8.13 (Frobenius Distance to TV Distance, Lemma 4.6.4, restated)**.** *Suppose $\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)$ are Gaussians with $\|\mu_1 - \mu_2\|_2 \leqslant \delta$ and $\|\Sigma_1 - \Sigma_2\|_F \leqslant \delta$. If the eigenvalues of Σ_1 and Σ_2 are at least $\lambda > 0$, then*

$$d_{\mathsf{TV}}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = O(\delta/\lambda) .$$

Proof. By Fact 4.2.1, we have

$$d_{\mathsf{TV}}\left(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)\right) = O\Big(\left((\mu_1 - \mu_2)^\top \Sigma_1^{-1} (\mu_1 - \mu_2) \right)^{1/2} + \|\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} - I\|_F) \Big) .$$

Then the first term is $\left\langle \mu_1 - \mu_2, \Sigma_1^{-1}(\mu_1 - \mu_2) \right\rangle^{1/2} \leqslant (\left\|\Sigma_1^{-1}\right\|_{\mathrm{op}} \|\mu_1 - \mu_2\|_2^2)^{1/2} \leqslant \delta/\sqrt{\lambda}$. The second term is

$$\begin{aligned} \|\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} - I\|_F^2 &= \left\| \Sigma_1^{-1/2} (\Sigma_1 - \Sigma_2) \Sigma_1^{-1/2} \right\|_F^2 \\ &= \operatorname{tr}\left(\left(\Sigma_1^{-1/2} (\Sigma_1 - \Sigma_2) \Sigma_1^{-1/2} \right)^2 \right) \\ &\leqslant \operatorname{tr}\left((\Sigma_1 - \Sigma_2)^2 \right) (1/\lambda)^2 \end{aligned}$$

$$\leqslant (\delta/\lambda)^2.$$

Thus,

$$d_{\mathsf{TV}}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = O(\delta/\sqrt{\lambda} + \delta/\lambda) = O(\delta/\lambda) \ .$$

□

Lemma 4.8.14 (Component Moments to Mixture Moments, Lemma 4.6.5 restated)**.** *Let $\mathcal{M} = \sum_{i \in [k]} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture such that $w_i \geqslant \alpha$, for some $0 < \alpha < 1$, and $\mathcal{M}$ has mean μ and covariance Σ and for all $i \neq j \in [k]$, $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F \leqslant 1/\sqrt{\alpha}$. Let X be a multiset of n samples satisfying Condition 4.2.29 with respect to $\mathcal{M}$ with parameters (γ, t), for $0 < \gamma < (dk/\alpha)^{-ct}$, for a sufficiently large constant c, and $t \in \mathbb{N}$. Let $\mathcal{D}$ be the uniform distribution over X. Then, $\mathcal{D}$ is $2t$-certifiably (c/α)-hypercontractive and for $d \times d$-matrix-valued indeterminate Q, $\left|\frac{Q}{2}\right. \left\{\mathbb{E}_{\mathcal{M}}\left(x^\top Q x - \mathbb{E}_{\mathcal{M}} x^\top Q x\right)^2 \leqslant O(1/\alpha)\left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2\right\}$.*

Proof. First, since $\mathcal{M}$ is a k-mixture of Gaussians with minimum mixing weight $w_{\min} \geqslant \alpha$, it follows from Corollary 4.2.20 that $\mathcal{M}$ is t-certifiably $(4/\alpha)$ hypercontractive. Further, since X is a good sample (satisfies Condition 4.2.29) with parameters (γ, t), it follows from Lemma 4.2.30 that the set $X' = \{\Sigma^{\dagger/2}(x_i - \mu)\}_{x_i \in X}$ is also a good sample (satisfies Condition 4.2.29) with parameters (γ, t) w.r.t. $\mathcal{M}' = \sum_{i \in [k]} w_i \mathcal{N}\left(\Sigma^{\dagger/2}(\mu_i - \mu), \Sigma^{\dagger/2}\Sigma_i\Sigma^{\dagger/2}\right)$. Since $\left\|\Sigma^{\dagger/2}\Sigma_i\Sigma^{\dagger/2}\right\|_{\mathrm{op}} \leqslant O(1/\alpha)$, it follows from Lemma 4.2.31 that for all $m \leqslant t$, $\|\mathbb{E}_{x \in_u X'}[x^{\otimes m}] - \mathbb{E}_{x \sim M'}[x^{\otimes m}]\|_F^2 \leqslant \gamma^2 m^{O(m)} d^m (1/\alpha)^m$. Since $\gamma < (dk/\alpha)^{-O(t)}$, it follows from Fact 4.2.27 that X is $2t$-certifiably (c/α)-hypercontractive.

By assumption, for all $i \neq j \in [k]$, we have that $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F \leqslant 1/\sqrt{\alpha}$. We can now apply Lemma 4.2.26 to obtain

$$\left|\frac{Q}{2}\right. \left\{\mathop{\mathbb{E}}_{x \sim \mathcal{M}}\left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}}\left[x^\top Q x\right]\right)^2\right] \leqslant O(1/\alpha)\left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2\right\} .$$

Therefore, it follows from Fact 4.2.27 that since X is a good sample (satisfies Condition 4.2.29) with parameters (γ, t), the uniform distribution $\mathcal{D}_X$ on X, $\left|\frac{Q}{2}\right.$

$$\left\{\mathbb{E}_{x\sim\mathcal{D}_X}(x^\top Qx - \mathbb{E}_{x\sim\mathcal{D}_X} x^\top Qx)^2 \leqslant O(1/\alpha)\left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2\right\}.$$ □

4.9 Bit Complexity Analysis

Here we address numerical issues related to our computation. We begin wth the assumption that the eigenvalues of our covariance matrices are bounded below.

Lemma 4.9.1. *If $\mathcal{M} = \sum_{i=1}^k w_i G_i$ is a mixture of Gaussians G_i where each G_i has mean and covariance of norm at most 2^b for some positive integer b and each G_i has covariance matrix whose eigenvalues are bounded below by some $\lambda > 0$. Let $\mathcal{M}'$ be an ε-corruption of $\mathcal{M}$ whose outputs are bounded by $2^{O(b)}$. Let N be a sufficiently large polynomial in d^k/ε and let η be λ divided by a sufficiently large polynomial in $2^b d/\varepsilon$ (where sufficiently large is degree $O(1)$). Then if our algorithm is given N i.i.d. samples from $\mathcal{M}'$ with each of their coordinates rounded to a nearby multiple of η (by which we mean one of the two closest), then our algorithm runs in time $\mathrm{poly}(N, b, \log(1/\eta))$ and with high probability returns a list of mixtures of Gaussians X_i with at least one of the X_i $\mathrm{poly}_k(\varepsilon)$-close to $\mathcal{M}$ in parameter distance.*

Proof. This follows from noting firstly that with high probability the any subset of the rounded samples will have moments $\lambda/\mathrm{poly}(d/\varepsilon)$-close to their moments before rounding. This means that with high probability these rounded samples will be a good sample (satisfy Condition 4.2.29). This means that our algorithm satisfies the necessary correctness guarantees. Furthermore, given that our samples now all have bounded bit complexity, it is easy to see that the runtime of our algorithm is polynomial in N and the bit complexity. □

More generally, as long as the parameters of the components of our mixture can be expressed with bounded bit complexity, we can prove a similar result, without needing any lower bound on the covariances.

Theorem 4.9.2. *Let $\mathcal{M} = \sum_{i=1}^k w_i G_i$ be a mixture of Gaussians where the G_i are Gaussians whose means and covariance matrices can all be written with coefficients given by rational* *numbers with bit complexity at most* b *for some integer* b*. Let* $\mathcal{M}'$ *be an* ε*-corruption of* $\mathcal{M}$ *so that with probability* 1 *the returned points have size* $2^{O(b)}$*. Let* N *be a sufficiently large polynomial in* d^k/ε*. Then there exists an algorithm that given* b *bit-oracle access to these samples runs in time* $\mathrm{poly}(N, b)$ *and with high probability returns a mixture of Gaussians* X *so that* $d_{\mathsf{TV}}(X, \mathcal{M}) < \mathrm{poly}_k(\varepsilon)$.

Proof. We begin by showing that we can find a list of hypotheses at least one of which is close. It is then straightforward to show that we can run a tournament over these hypotheses to find a specific one that works. We also assume for simplicity that each w_i is at least 3ε.

We begin by setting λ to be $2^{-b\cdot d^{kC}}$ for a sufficiently large constant C. By adding each sample to a random sample from $N(0, \lambda I)$, we can produce samples from $\tilde{\mathcal{M}}'$, and ε-corruption of $\tilde{\mathcal{M}} = \sum_{i=1}^k w_i \tilde{G}_i$ where $\tilde{G}_i$ is the convolution of G_i with $N(0, \lambda I)$. Note that $\tilde{G}_i$ is a Gaussian whose covariance has eigenvalues at least λ. Furthermore, if the covariance matrix of G_i is non-singular, the smallest eigenvalue of the covariance matrix must be at least $2^{O(b\cdot d)}$, and therefore $d_{\mathsf{TV}}(G_i, \tilde{G}_i) < \varepsilon$.

Since the eigenvalues of the components of $\tilde{\mathcal{M}}$ are bounded below, we can apply Lemma 4.9.1 to our samples from $\tilde{\mathcal{M}}'$ rounded to an appropriate accuracy η, and in $\mathrm{poly}(N, b)$-time obtain a list of hypothesis mixtures at least one of which is (with high probability) close to $\tilde{\mathcal{M}}$ in total variation distance.

If the covariances of all of the G_i with weights more than some sufficiently large $\mathrm{poly}_k(\varepsilon)$ are all non-singular, then one of these hypotheses will be close to $\mathcal{M}$. Otherwise, there must be some i for which w_i is relatively large and for which G_i has singular covariance matrix. In particular, there must be an integer vector v with bit complexity $O(bd)$ in the kernel of the covariance matrix of G_i. The hypothesis mixture X that is close to $\tilde{\mathcal{M}}$ in parameter distance must contain some component close to $\tilde{G}_i$. Since $\tilde{G}_i$ has covariance matrix $\tilde{\Sigma}_i = \lambda I + \Sigma_i$ where Σ_i is the covariance matrix of G_i. We note that $\tilde{\Sigma}_i$ will have an eigenvalue of λ and that therefore, our close hypothesis will have an eigenvalue at most 2λ.

If any of our returned hypotheses have any component with a covariance matrix Σ which has any eigenvalue less than 2λ, we do the following. We consider the quadratic form on integer vectors v defined by

$$Q(v) = v^T \Sigma v + \sqrt{\lambda}|v|_2^2.$$

We note that if this Σ is close in parameter distance to a singular $\tilde{\Sigma}_i$ where Σ_i had a null-vector v of norm $2^{O(bd)}$, then for that same value of v we will have that $Q(v) < \lambda^{1/4}$. Using the Lovász local lemma in [104], we can find a v so that $Q(v)$ is within a $2^{O(d)}$-factor of the minimum possible value over all non-zero, integer vectors v. If for this v, $Q(v) > 2^{\Omega(d)}\lambda^{1/4}$, we know that the hypothesis in question is not close to $\tilde{\mathcal{M}}$ in parameter distance and can be ignored. On the other hand, any v with $Q(v)$ this small must have $|v| < 2^{O(d)}\lambda^{-1/2}$ and $v^T\Sigma v < 2^{O(d)}\lambda^{1/4}$. Note that the projection of v onto the $\ker(\Sigma_i)^\perp$ is either zero or has magnitude at least $2^{O(bd)}$. In the latter case, it would need to be the case that $Q(v)$ is substantially larger. Thus, if such a hypothesis is close to $\tilde{\mathcal{M}}$ in parameter distance, then v is in the kernel of some Σ_i.

If our algorithm finds some v for some hypothesis, we then compute $v \cdot x$ to error λ for each of our samples x. If $\mathcal{M}$ really has a component with v in the kernel of its covariance matrix, all of the x's taken from this component will have $v \cdot x$ the same. This means that at least a $(3/2)\varepsilon$ fraction of our samples x will have $v \cdot x$ within λ of each other. Note that if v is not in the kernel of any covariance matrix of any G_i than $\mathbf{Var}[v \cdot G_i]$ will be at least $2^{O(db)}$ for each i, and with high probability we will not find this many close samples.

To summarize, if our algorithm applies this procedure to every component of every hypothesis and does not find such a v, then it cannot be the case that $\mathcal{M}$ contains any components of weight more than $\mathrm{poly}_k(\varepsilon)$ that are singular, and thus one of our original hypotheses must be close in total variational distance. We can then run a tournament to find a single one that is close. Otherwise, if we find such a v for which many points do have $v \cdot x$ close by, then v must be a null vector of the covariance matrix of some G_i.

Furthermore, all of the samples within λ of this common value of $v \cdot x$, with high probability are either errors or come from components contained in some lower dimensional subspace. We can determine what this subspace is by noting that it is defined by $v \cdot x = q$ for some rational number q with bit-complexity at most $O(bd)$ and using continued fractions on a good numerical approximation of q in order to determine its true value. Our algorithm can then recurse on the points in this subspace (a mixture of Gaussians in a lower dimensional space) and on the remaining points (which are from a mixture of fewer Gaussians), and return an appropriate mixture of the results. □

4.10 Table of Parameters

Table 4.1: Table of parameters in Algorithm 4.6.2.

Parameter	**Notation**	**Value**
Components number	k	-
Corruption parameter	ε	$\varepsilon < 1/k^{k^{O(k^2)}}$
Good samples	γ	$\varepsilon d^{-8k} k^{-C'k}$
Eigenvalue threshold	τ	$\Theta\left(\varepsilon^{1/(40C^{k+1}(k+1)!)}\right)$
Minimum weight	α	$\varepsilon^{1/(45C^{k+1}(k+1)!)}$
Accuracy parameter for partial clustering	β	$O(\alpha^5 k^1 0)$
Frobenius separation	Δ	$O(\sqrt{\varepsilon}/\alpha)$
-	η	$2k^{4k} O(1+\Delta/\alpha)^{4k} \sqrt{\varepsilon}$

CHAPTER 5

ROBUSTLY LEARNING AFFINE TRANSFORMATIONS

In this chapter, we present a polynomial-time algorithm for robustly learning an unknown affine transformation of the standard hypercube from samples, an important and well-studied setting for independent component analysis (ICA). Specifically, given an ε-corrupted sample from a distribution D obtained by applying an unknown affine transformation $x \longrightarrow Ax + b$ to the uniform distribution on a d-dimensional hypercube $[-1, 1]^d$, our algorithm constructs $\widehat{A}, \hat{b}$ such that the total variation distance of the distribution $\widehat{D}$ from D is $O(\varepsilon)$ using $\text{poly}(d)$ time and samples. Total variation distance is the information-theoretically strongest possible notion of distance in our setting and our recovery guarantees in this distance are optimal up to the absolute constant factor multiplying ε. In particular, if the rows of A are normalized to be unit length, our total variation distance guarantee implies a bound on the *sum* of the ℓ_2 distances between the row vectors of A and A', $\sum_{i=1}^{d} \left\|a_{(i)} - \hat{a}_{(i)}\right\|_2 = O(\varepsilon)$. In contrast, the strongest known prior results only yield an $\varepsilon^{O(1)}$ (relative) bound on the distance between *individual* a_i's and their estimates and translate into an $O(d\varepsilon^{O(1)})$ bound on the total variation distance.

This is joint work with Pravesh Kothari and Santosh Vempala, and appears as [105].

5.1 Results and Techniques

Our main result is the following theorem:

Theorem 5.1.1. *Given $X = \{x^{(1)}, \ldots, x^{(n)}\}$, an ε-corrupted sample of points from an unknown parallelopiped $H = A[-1, 1]^d + b$ in $\mathbb{R}^d$, there is a polynomial-time algorithm that outputs a parallelopiped $\widehat{H} = \widehat{A}[-1, 1]^d + \hat{b}$ s.t. $d_{TV}(\widehat{H}, H) = O(\varepsilon)$.*

Equivalently, our algorithm produces an estimate $\widehat{A}, \hat{b}$ so that

$$\sum_{i=1}^{d} \frac{\left\|a_{(i)} - \hat{a}_{(\pi(i))}\right\|_2}{\left\|a_{(i)}\right\|_2} + \sum_{i=1}^{d} \frac{\left|b_i - \hat{b}_i\right|}{\left\|a_{(i)}\right\|_2} = O(\varepsilon)$$

where $a_{(i)}$ are rows of A and $\hat{a}_{(i)}$ are rows of $\widehat{A}$ and π is some global permutation $\pi : [d] \to [d]$.

Our algorithm is based on a new approach to ICA, even without outliers, that does not rely the fourth (or any constant order) moments as in prior works. As we discuss next, a key contribution of our work is the development of a new *certificate* of correctness of the estimated affine transformation that does not rely on low order moments.

We briefly describe our approach and the key ideas that go into our algorithm. At a high-level, our idea is to depart from the prior works based on the method of moments and instead rely on a new *robust* characterization of b and rows of A. This characterization allows us to build a *robust certified checker* that takes input an ε-corrupted sample from the input model along with a current estimate of the parameters of the affine transformation. Our checker either accepts and produces a certificate of correctness of the parameters or provides an explicit witness of the violation that can be used to update our current estimate of the parameters. This strategy can be thought of as a *robust gradient descent* procedure where the "gradient oracle" is replaced with our new certificate. Let us explain this approach in a little more detail next.

We first assume the affine transformation consists of only shifts and diagonal scalings, i.e., the unknown matrix A is a positive diagonal matrix. In this case, we start with a coarse approximation of the mean and side lengths obtained by using the coordinate-wise median, and twice the interval along each coordinate that contains the middle $(1/2) + \varepsilon$ fraction of samples. We then refine this iteratively based on the following properties of the density of the hypercube: (1) it is uniform along each coordinate and so the density is constant in its support; (2) the measure of the intersection of two axis-parallel bands is equal to the product of their individual measures. The algorithm starts with an "outer" approximation and tries to move the bounding planes along each coordinate inwards if the outermost band is too sparse (too few samples). Crucially, along the way, it checks if the the intersection of some pair of bands has too large a fraction of samples; in this case, most of the samples in the intersection must be outliers and the algorithm can delete all the points in the intersection. An important idea here is that most of the outliers can be partitioned among the coordinate directions that they affect. This is captured as an independent combinatorial lemma (Lemma 5.3.4).

Now consider a general rotation without shift or diagonal scaling, i.e., A is orthonormal and $b = 0$. This turns out to be substantially more challenging. The following bound on the TV distance, relating it to the "right" notion of parameter distance, serves as a starting point.

Lemma 5.1.2. *Let $H = [-1, 1]^d$. Suppose $A, \widehat{A}$ are $d \times d$ matrices; A is orthonormal and $\widehat{A}$ is a matrix with unit length rows. There is an absolute constant C s.t.*

$$d_{TV}(\widehat{A}H, AH) \leqslant C \sum_i \left\| \hat{a}_{(i)} - a_{(i)} \right\|_2 .$$

Remark. This bound is tight by Lemma 5.8.2.

This lemma follows from Lemma 5.4.5, in which we prove the fraction of the volume of the subset $\{x \in AH : x \cdot \hat{a}_{(i)} > 1\}$ is tightly bounded by $C \left\| \hat{a_{(i)}} - a_{(i)} \right\|_2$. Note that the RHS terms in the lemma are not squared. If they were, it would be the Frobenius norm and we can hope to robustly estimate to within low error. The above norm however is not rotationally invariant, depends on the target basis (as it should!), and could be much larger than the Frobenius norm.

Robust Gradient Descent. To estimate a rotation, we again start with a coarse approximation, where each facet normal is approximated to within poly(ε) error. For this we can use existing robust estimation methods [42]. As noted earlier, even though each normal vector is within poly(ε) distance/angle, the overall TV distance can be much larger, growing with a polynomial in d. To improve this, we will use a new *robust* gradient descent with a suitable objective function. We consider one vector (candidate facet normal or row of $\widehat{A}$) at a time (keeping the others fixed) and iteratively "improve" it. Our ideal objective is to maximize the number of sample points inside the parallelopiped defined by the current vectors, while keeping the width of the band in each coordinate fixed. But this is difficult to estimate or improve locally. Instead, we use the total number of points outside the band $|\hat{a}_{(i)} \cdot x| \leqslant 1$ as the objective to be minimized. For the underlying original hypercube, all points outside would be outliers and hence at most ε fraction in total. Moreover, if we consider a normal vector that is at some distance δ from the true closest one, i.e., $\|\hat{a}_{(i)} - a_{(i)}\|_2 = \delta$, then, as we show, the measure of the hypercube that lies outside the hyperplane defined by $\hat{a}_{(i)}$, i.e., $S = \{x : \hat{a}_{(i)} \cdot x \geqslant 1\}$ is proportional to δ, i.e., this objective is locally stable and robust.

How can we use this in an algorithm? After all, we need a concrete procedure to improve the current vector. We prove that we can use the mean of points outside the current band in this direction i.e., the mean μ_S of the set $S = \{x : \hat{a}_{(i)} \cdot x \geqslant 1\}$ tells us which direction to move $\hat{a}_{(i)}$ to make it closer to $a_{(i)}$ (namely along $-\mu_S$). In other words, we can improve our objective of number of points outside the band by a local update!

Does this update work with outliers? The presence of outliers complicates matters, as a small number of outliers could radically alter the location of the mean outside the band. To address this challenge, we prove that *robustly* estimating the mean of this subset S in *Euclidean norm* suffices to preserve the gradient of the objective approximately! For this we can use robust mean estimation algorithms already developed in the literature, since the distribution of the hypercube intersected with a halfspace satisfies the necessary moment conditions. Alongside, to keep the influence of outliers under control, we ensure that pairwise intersections of bands are all small by repeatedly checking the measure of pairwise intersections and deleting points if there are too many.

Our overall algorithm can be viewed as *Robust Gradient Descent* with provable guarantees, and this approach might have other applications.

Finally, to solve the full affine transformation problem in the presence of outliers, we combine the above procedures for shift-and-scaling and rotation by alternating between them, showing that each alternation is guaranteed to make progress in the overall TV distance to the ground truth.

5.2 Preliminaries

5.2.1 Logconcave Functions

The following lemmas are either from [106] or direct consequences.

Lemma 5.2.1. *Let $f : \mathbb{R}^2 \to \mathbb{R}_+$ be an isotropic two-dimensional logconcave density with associated measure ν. Let u, v be unit vectors in $\mathbb{R}^2$ with $|u^\top v| \leqslant \frac{1}{2}$. Consider the halfspaces $H_u = \{x : u^\top x \geqslant a_u\}$ and $H_v = \{x : v^\top x \geqslant a_v\}$ and assume that the marginal densities along u and v, f_u and f_v satisfy $f_u(a_u), f_v(a_v) \geqslant 1/2$. Then, for a universal constant C,*

$$\nu(H_u \cap H_v) \leqslant C\nu(H_u)\nu(H_v)$$

The proof of Lemma 5.2.1 is deferred to Section 5.7.

Lemma 5.2.2. *Let f be a one-dimensional logconcave density with mean 0 and variance σ^2. Let $\mu^+ = \mathbb{E}_f(y|y \geqslant 0)$. Then there exist universal constants c_1, c_2 s.t.*

$$c_1\sigma \leqslant \mu^+ \leqslant c_2\sigma.$$

Lemma 5.2.3. *For any one-dimensional logconcave density $f : \mathbb{R} \to \mathbb{R}_+$ with mean μ, we have*

$$\mathbb{P}(X \geqslant \mu) \geqslant \frac{1}{e}.$$

Lemma 5.2.4. *Let f be a one-dimensional logconcave density with mean μ. Then*

$$f(\mu) \geqslant \frac{1}{8} \max f(x).$$

Lemma 5.2.5. *Let f be a one-dimensional isotropic logconcave density. Then for all x,*

$$f(x) \leqslant 1.$$

Lemma 5.2.6. *Suppose $M_n(f)$ are the raw moments of f, i.e.,*

$$M_n(f) = \int_0^\infty t^n f(t)\, dt.$$

If f is logconcave, then the sequence $M_n(f)/n!$ is logconcave. Moreover, we could append $f(0)$ at the beginning of the sequence and maintain the logconcavity, i.e.,

$$f(0)M_1(f) \leqslant M_0(f)^2.$$

5.2.2 Hypercubes

The following facts about hypercubes will be useful. While the precise constants in the bounds are not important for our analysis, the first two facts have been a subject of inquiry in asymptotic convex geometry.

Lemma 5.2.7 ([107]). *Let $B = [-1/2, 1/2]^d$ be the d-dimensional hypercube with volume 1 and $a \in \mathbb{R}^d$ be an arbitrary unit vector. Then for all $t \leqslant 3/4$,*

$$\mathrm{vol}(\{x \in B : |x \cdot a| > t/2\}) \leqslant 1 - t.$$

Lemma 5.2.8 ([108]). *Let $B = [-1/2, 1/2]^d$ be the d-dimensional hypercube with volume 1 and $a \in \mathbb{R}^d$ be an arbitrary unit vector. Then $(d-1)$-volume of any sections of B defined by a and t is at most $\sqrt{2}$, i.e.,*

$$\mathrm{vol}_{n-1}(\{x \in B : x \cdot a = t\}) \leqslant \sqrt{2}.$$

The next fact follows from the above lemma and a calculation using the logconcavity of one-dimensional marginals of the hypercube.

Lemma 5.2.9. *Let $x \sim [-1,1]^d$ be a uniform random vector on $H = [-1,1]^d$ and $y = x \cdot a$ where a is an arbitrary unit vector. Let $\mu_t = \mathbb{E}_f(y|y \geqslant t)$. For $0 \leqslant t \leqslant 1/2$, there exists universal constant c s.t.*

$$\mu_t \geqslant t + c.$$

5.3 Robustly Learning Shifts and Diagonal Scalings

In this section, we give an algorithm for robustly learning arbitrary shifts and diagonal scalings of the uniform distribution on the solid hypercube $H_0 = [-1,1]^d$. That is, $x \to Ax + b$ when A is a diagonal matrix. Since the uniform distribution on H is symmetric around 0, we can WLOG assume that all the entries of A are non-negative. This special case is equivalent to learning affine transformations that correspond to a shift (i.e., introducing a non-zero mean) and diagonal scaling (i.e., scales each of the coordinates of the hypercube by an unknown and potentially different positive scaling factor).

More precisely, we will prove:

Theorem 5.3.1. *Suppose H is an unknown axis-aligned cube in $\mathbb{R}^d$, that is, $H = AH_0 + b = \otimes_{i=1}^d [u_i, v_i]$ where $A \in \mathbb{R}^{d\times d}$ is a diagonal matrix, $b \in \mathbb{R}^d$ and $H_0 = [-1,1]^d$ is the unit cube. There exists an algorithm that, for any constant $0 < \varepsilon < 1/80$, takes an ε-corruption $X = x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ of size $n \geqslant n_0 = \widetilde{O}(d^2/\varepsilon^2)$ of an iid sample from the uniform distribution on H and outputs $\widehat{H} = \widehat{A}H_0 + \hat{b} = \otimes_{i=1}^d [\hat{u}_i, \hat{v}_i]$ such that*

$$d_{TV}(H, \widehat{H}) \leqslant 4\varepsilon.$$

Remark 5.3.2. While we will omit this refinement here, a more careful analysis of our algorithm produces a tighter bound of $d_{TV}(H, \widehat{H}) \leqslant (2+\varepsilon)\varepsilon$.

We first describe our algorithm:

Algorithm 5.3.3 (Shift and Scaling)**.**

Input: *An ε-corruption X of an iid sample of size n chosen uniformly from $\otimes_{i=1}^{d}[u_i, v_i]$.*

1. ***Robust Range Finding:*** *For each $1 \leqslant i \leqslant d$, arrange the input corrupted sample in increasing order of y_i. Let $[l_i, r_i]$ be the interval of the smallest length that includes the middle $1/2 + \varepsilon$ fraction of the points (notice that such an interval is uniquely defined). Set $\hat{u}_i = (3l_i - r_i)/2$ and $\hat{v}_i = (3r_i - l_i)/2$. Notice that $\{x \in \mathbb{R}^d : v_i \leqslant x_i \leqslant u_i\}$ is an axis-aligned cube that contains the true cube and all the side lengths are at most twice the true cube.*

2. ***One-Dimensional Density Check:*** *For all $i \in [n], k \in [d]$, check if*

$$|S_{i,k}^{+}| := \left|\left\{x : x^{(i)} \geqslant \left(1 - \frac{k\varepsilon}{d}\right)\hat{u}_i + \frac{k\varepsilon}{d}\hat{v}_i\right\}\right| \geqslant \frac{k\varepsilon}{2d}n \tag{5.3.1}$$

$$|S_{i,k}^{-}| := \left|\left\{x : x^{(i)} \leqslant \frac{k\varepsilon}{d}\hat{u}_i + \left(1 - \frac{k\varepsilon}{d}\right)\hat{v}_i\right\}\right| \geqslant \frac{k\varepsilon}{2d}n \tag{5.3.2}$$

3. ***Update:*** *If (5.3.1) is false for some i and k, update $\hat{u}_i = \hat{u}_i - \frac{k\varepsilon}{d}(\hat{u}_i - \hat{v}_i)$ and go back to Step 2. If (5.3.2) is false for some i and k, set $\hat{v}_i = \hat{v}_i + \frac{k\varepsilon}{d}(\hat{u}_i - \hat{v}_i)$ and iterate (go back to Step 2).*

4. ***Two-Dimensional Density Check:*** *For all $i, j \in [n], k_1, k_2 \in [d]$, check if*

$$|S_{i,k_1}^{+} \cap S_{j,k_2}^{+}|, |S_{i,k_1}^{+} \cap S_{j,k_2}^{-}|, |S_{i,k_1}^{-} \cap S_{j,k_2}^{-}| \leqslant \frac{10k_1k_2\varepsilon^2}{d^2}n \tag{5.3.3}$$

5. ***Update:*** *If (5.3.3) is false for some i, j and k_1, k_2, remove all points in the violating intersection from the sample set and go back to Step 2.*

6. ***Return:*** *Output the cube $\otimes_{i=1}^{d}[\hat{u}_i, \hat{v}_i]$.*

Notice that in this case, our goal is effectively to determine the intervals $[u_i, v_i]$ for each

$1 \leqslant i \leqslant d$ that describe the i-th dimension of the shifted and scaled hypercube H. The key idea of the algorithm is a *certificate* that checks a set of efficiently verifiable conditions on the corrupted sample with the guarantee that 1) the true parameters satisfy the checks, and 2) any set of parameters that satisfy the checks yield a hypercube $\hat{H}$ that is $O(\varepsilon)$ different in symmetric volume difference (that gives us a total variation bound) from the true hypercube H.

Our certificate itself is simple and natural and corresponds to checking that 1) for each of the coordinates $1 \leqslant i \leqslant d$, the discretized density (i.e., fraction of points lying in discrete intervals of size $\sim \varepsilon/d$) in the purported range estimated from the corrupted sample matches the expected density of the uniform distribution on $AH + b$, and 2) for each pair of coordinates, the fraction of points in the intersection of intervals of length $\sim \varepsilon/d$ along all possible pairs of directions match the expected density. Notice that all such statistics in an uncorrupted sample match those of the population with high probability so long as the the sample is of size $\text{poly}(d/\varepsilon)$.

To analyze the above algorithm, we will prove that any set of parameters that satisfy the checks in our certificate (5.3.1),(5.3.2),(5.3.3) in Algorithm 5.3.3 must yield a good estimate of true parameters. Specifically, Lemma 5.3.5 shows that the volume of $\widehat{H}$ that is not contained in H is small. Lemma 5.3.7 shows that the volume of H that is not covered by $\widehat{H}$ is small. Together, these two lemmas imply that our estimated distribution is $O(\varepsilon)$ close in total variation to the true hypercube. Our proof relies on an elementary combinatorial claim about set systems that we state next. We defer the proof to Section 5.6.

Lemma 5.3.4. *Let* $S_1, \ldots, S_d \subseteq [n]$ *be arbitrary subsets. Let* $\text{frac}(S) := |S|/n$ *be normalized size of* S*. Suppose that*

$$\text{frac}\left(\bigcup S_i\right) = \varepsilon \tag{5.3.4}$$

for some $\varepsilon < 1$ *and for all* i, j,

$$\text{frac}(S_i \cap S_j) \leqslant \alpha \text{frac}(S_i)\text{frac}(S_j), \tag{5.3.5}$$

for some $\alpha > 1$ s.t. $\alpha\varepsilon < 1$. Then,

$$\sum_{i\in[d]} \text{frac}(S_i) \leqslant \frac{\varepsilon}{1-\alpha\varepsilon}. \tag{5.3.6}$$

Lemma 5.3.5. *Let $X = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ be an ε-corruption of an iid sample from the uniform distribution on $H = \otimes_{i=1}^d [u_i, v_i]$ for some $\varepsilon < 1/80$. Let $\widehat{H} = \otimes_{i=1}^d [\hat{u}_i, \hat{v}_i]$ be any axis-aligned hypercube satisfying (5.3.1),(5.3.2),(5.3.3). Then, with probability at least $1 - 1/d$ over the draw of the original uncorrupted sample,*

$$\frac{\text{vol}(\widehat{H} \setminus H)}{\text{vol}(\widehat{H})} \leqslant 4\varepsilon.$$

Here, vol(S) *denotes the usual Lebesgue volume of set S in $\mathbb{R}^d$.*

Proof. For only the sake of analysis, we set $\hat{u}_i - \hat{v}_i = 1$. In this normalization, vol$(\widehat{H}) = 1$. For each i, discretize the interval $[\hat{u}_i, \hat{v}_i]$ to a grid with intervals of size $\varepsilon(\hat{u}_i - \hat{v}_i)/d = \varepsilon/d$. We further assume that the vertices of the unknown H are rounded to points in this grid. This assumption amounts to a change in the volume of H by at most ε vol$(\hat{H})$. We will then show that the density of points in the difference $\widehat{H} \setminus H$, i.e., $\{x : x \in \widehat{H}, x \notin H\}$ is

$$\frac{|\{x \in X : x \in \widehat{H} \setminus H\}|}{\text{vol}(\widehat{H} \setminus H)} \geqslant \frac{1}{4}n. \tag{5.3.7}$$

From the inequality above, it follows that vol$(\widehat{H} \setminus H) \leqslant \frac{4}{n}|\{x : x \in \widehat{H} \setminus H\}| \leqslant 4\varepsilon$ since $|\{x : x \in \widehat{H} \setminus H\}| \leqslant \varepsilon n$. So it suffices to prove (5.3.7).

For all $i \in [d]$, suppose $u_i - \hat{u}_i = c_i\varepsilon/d, \hat{v}_i - v_i = c'_i\varepsilon/d$, where c_i, c'_i are integers in $[d]$. Let $S_i = \{x \in X : \hat{u}_i \leqslant x_i \leqslant u_i, v_i \leqslant x_i \leqslant \hat{v}_i\}$. By (5.3.1) and (5.3.2), we have $|S_i| \geqslant (c_i + c'_i)\frac{\varepsilon}{2d}n$. By (5.3.3), we have for any $i \neq j$,

$$|S_i \cap S_j| \leqslant 10(c_i + c'_i)\frac{\varepsilon^2}{d^2}n \leqslant 40|S_i||S_j|n.$$

Then by Lemma 5.3.4 with $\alpha = 40$,

$$\frac{|\bigcup S_i|}{n - 40|\bigcup S_i|} \geqslant \sum_{i \in [d]} \frac{|S_i|}{n},$$
$$\frac{|\{x : x \in \widehat{H} \setminus H\}|}{n - 40|\{x : x \in \widehat{H} \setminus H\}|} \geqslant \frac{\varepsilon}{2d} \sum_{i=1}^{d} (c_i + c_i').$$

By the upper bound on the total number $|\{x : x \in \widehat{H} \setminus H\}| \leqslant \varepsilon n \leqslant n/80$, the denominator is at least $n/2$. Then

$$|\{x : x \in \widehat{H} \setminus H\}| \geqslant \frac{\varepsilon}{4d} n \sum_{i=1}^{d} (c_i + c_i').$$

Thus we get (5.3.7) by $\mathrm{vol}(\widehat{H} \setminus H) \leqslant \frac{\varepsilon}{d} \sum_{i=1}^{d} (c_i + c_i')$. □

Next we show that Step 5 removes at most εn true points in total.

Lemma 5.3.6. *In Algorithm 5.3.3, if the sample size $n \geqslant n(\varepsilon, d) = O(d^2/\varepsilon^2)$, with high probability, at least half of points removed in Step 5 are outliers.*

Proof. Suppose we remove points from an intersection $S_{i,k_1}^+ \cap S_{j,k_2}^+$, i.e.,

$$|S_{i,k_1}^+ \cap S_{j,k_2}^+| > \frac{10 k_1 k_2 \varepsilon^2}{d^2} n. \tag{5.3.8}$$

If the violating intersection is outside the true cube, all the points are outliers. Otherwise, we can compute the volume by definition of $S_{i,k}^+$

$$\mathrm{vol}(S_{i,k_1}^+ \cap S_{j,k_2}^+) = \frac{k_1 k_2 \varepsilon^2}{d^2} (\hat{u}_i - \hat{v}_i)(\hat{u}_j - \hat{v}_j) \leqslant \frac{4 k_1 k_2 \varepsilon^2}{d^2}$$

where the last equality follows from $\hat{u}_i - \hat{v}_i$ is at most twice of the true side length and hence $\hat{u}_i - \hat{v}_i \leqslant 2$ by the assumption that the side length of H is 1. Then with high probability, the number of original uncorrupted sample in the region $S^+_{i,k_1} \cap S^+_{j,k_2}$ of volume $4k_1k_2\varepsilon^2/d^2$ is at most $5k_1k_2\varepsilon^2 n/d^2$. Thus, comparing with (5.3.8), we can see that at least half of the point removed by the algorithm are outliers. □

Lemma 5.3.7. *Let $X = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ be an ε-corruption of an iid sample from the uniform distribution on $H = \otimes_{i=1}^d [u_i, v_i]$ with $\varepsilon < 1/64$ and $n \geqslant n(d,\varepsilon) = \widetilde{O}(d^2/\varepsilon^2)$. Let $\widehat{H} = \otimes_{i=1}^d [\hat{u}_i, \hat{v}_i]$ be an axis-aligned hypercube satisfying (5.3.1),(5.3.2),(5.3.3). Then, with probability at least $1 - 1/d$ over the draw of the original uncorrupted sample,*

$$\frac{\mathrm{vol}(H \setminus \widehat{H})}{\mathrm{vol}(H)} \leqslant 4\varepsilon.$$

Proof. For only the sake of analysis, we set $\hat{u}_i - \hat{v}_i = 1$ and hence $\mathrm{vol}(\widehat{H}) = 1$. For $i \in [d]$ s.t. $\hat{u}_i < u_i$ or $v_i < \hat{v}_i$, we define $S_i = \{x : \hat{u}_i < x_i \leqslant u_i \text{ or } v_i \leqslant x_i < \hat{v}_i\}$ and $c_i = u_i - \hat{u}_i, c'_i = \hat{v}_i - v_i$. Then

$$H \setminus \widehat{H} = \bigcup_{\{i:\hat{u}_i<u_i, \text{ or } v_i<\hat{v}_i\}} S_i.$$

Since S_i is removed by the algorithm, by (5.3.1) and (5.3.2), $|S_i| < \frac{1}{2}(c_i + c'_i)n$. On the other hand, S_i is in the true cube, the original uncorrupted sample in S_i is at least $\frac{3}{4}(c_i + c'_i)n$ with high probability. So there are at least half of the original uncorrupted sample in S_i are removed by either the adversary or by the algorithm. Suppose $\eta_i n$ is the number of original uncorrupted sample removed in S_i. Then $\eta_i \geqslant \frac{1}{4}(c_i + c'_i)\varepsilon/d$. By Lemma 5.3.6, the total number of points deleted by the algorithm is at most εn. Hence the number of deleted points in $\bigcup S_i$ is at most $2\varepsilon n$. The deletion in the intersection of S_i and S_j is naturally upper bounded by the number of original uncorrupted sample, which is

$$\frac{(c_i + c'_i)(c_j + c'_j)\varepsilon^2}{d^2} \leqslant 16\eta_i\eta_j.$$

So we can apply Lemma 5.3.4 with $\alpha = 16$ to the deletion in S_i and get

$$\frac{2\varepsilon}{1-32\varepsilon} \geqslant \sum_{\{i:\hat{u}_i<u_i,\text{ or } v_i<\hat{v}_i\}} \eta_i.$$

If $\varepsilon \leqslant 1/64$, the left hand side is at most 4ε. Thus

$$\begin{aligned}\operatorname{vol}(H \setminus \widehat{H}) = \operatorname{vol}\left(\bigcup S_i\right) &\leqslant \sum \operatorname{vol}(S_i)\\ &\leqslant \frac{1}{2}\sum (c_i + c_i')\\ &\leqslant \sum \eta_i\\ &\leqslant 4\varepsilon.\end{aligned}$$

□

Now, with Lemmas 5.3.5 and 5.3.7, we can prove our main result of the shift and scaling case.

Proof of Theorem 5.3.1. If Algorithm 5.3.3 outputs a cube $\widehat{H}$ that passes all checks (5.3.1),(5.3.2),(5.3.3), then by Lemma 5.3.5 and Lemma 5.3.7,

$$d_{TV}(H, \widehat{H}) = 1 - \frac{\operatorname{vol}(H \cap \widehat{H})}{\max\{\operatorname{vol}(H), \operatorname{vol}(\widehat{H})\}} = \max\left\{\frac{\operatorname{vol}(H \setminus \widehat{H})}{\operatorname{vol}(H)}, \frac{\operatorname{vol}(\widehat{H} \setminus H)}{\operatorname{vol}(\widehat{H})}\right\} = 4\varepsilon.$$

Otherwise, the algorithm improves one of $\hat{u}_i$ or $\hat{v}_i$ by at least ε/d. Since the true cube passes all checks in the algorithm (after the filtering if necessary), within at most $O(d^2/\varepsilon)$ iterations, the algorithm terminates. If the algorithm terminates when it deletes $2\varepsilon n$ points, by Lemma 5.3.6, we remove all noisy points. □

5.4 Robustly Learning Rotations

In this section, we give a robust algorithm to learn rotations of the uniform distribution of the standard hypercube. Specifically, we will show:

Theorem 5.4.1. *Suppose $H = [-1,1]^d$ is the standard cube in $\mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$ be an unknown rotation matrix. There exists an algorithm that given an ε-corruption $X = \{x^{(1)},\ldots,x^{(n)}\}$ of size $n \geqslant n_0 = \widetilde{O}(d^2/\varepsilon^2)$ of an iid sample from the uniform distribution on AH, runs in poly(n) time and outputs $\widehat{A}$ such that*

$$d_{TV}(AH, \widehat{A}H) = O(\varepsilon).$$

We first describe the algorithm:

Algorithm 5.4.2 (Robustly Learning Rotations)**.**

Input: *ε-corrupted sample $X = \{x^{(1)}, x^{(2)},\ldots,x^{(n)}\}$.*

Output: *$\widehat{A}$ with unit length rows $\hat{a}_{(1)},\ldots,\hat{a}_{(d)}$.*

1. ***Warm Start:*** *Run the robust moment estimation algorithm in Theorem 2.2.4 and get an estimate of A with rows $a^*_{(1)},\ldots,a^*_{(d)}$, so that for each i, $\left\|\hat{a}_{(i)} - a^*_{(i)}\right\|_2 = O(\sqrt{\varepsilon})$. Initialize $a^0_{(i)} = \hat{a}_{(i)}$ for each i.*

2. *Define the set $S(a) := \{x : x\cdot a > 1\} \cup \{-x : x\cdot a < -1\}$ for any unit vector $a \in \mathbb{R}^d$. For $t = 1,2,\ldots,2^{12}d\log d$, do:*

   (a) ***Two-Dimensional Density Check:*** *For all $j < i$, check if*

   $$\frac{1}{n}|S(a^t_{(i)}) \cap S(\hat{a}_{(j)})| \leqslant \frac{8C}{n^2}|S(a^t_{(i)})||S(\hat{a}_{(j)})|$$

   where C is the universal constant in Lemma 5.2.1. If false, remove all points *in the intersection* $S(a_{(i)}^t) \cap S(\hat{a}_{(j)})$.

(b) ***Robust Gradient Descent:*** *Robustly estimate the mean* $\tilde{\mu}$ *by applying algorithm in Theorem 2.2.1 to the subset of samples* $S(a_{(i)}^t)$*. Set*

$$\begin{aligned} \beta &= \frac{c_2 |S(a_{(i)}^t)|}{\|\tilde{\mu}\|^2 n}, \\ a_{(i)}^{t+1} &= \frac{a_{(i)}^t - \beta\tilde{\mu}}{\left\|a_{(i)}^t - \beta\tilde{\mu}\right\|} \\ s^{t+1} &= \left|\{x : |x \cdot a_{(i)}^{t+1}| > 1\}\right| \end{aligned}$$

Add back points that are removed in Step (a), i.e., recover the sample set to the origin one X.

Let $\hat{a}_{(i)} = a_{(i)}^t$ *where* t *minimizes* s^t.

3. *Repeat Step 2 for all* $i = 1, \ldots, d$*. Output the matrix* $\widehat{A}$ *with rows* $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$.

5.4.1 Outline of Analysis

For the purpose of analysis, we assume that the true rotation matrix is the identity matrix in the rest of Section 5.4. We will analyze Algorithm 5.4.2 by the following propositions. First we show that we can separately learn the rows of the rotation matrix A. Let $S(a) = \{x : x \cdot a > 1\} \cup \{-x : x \cdot a < -1\}$. If a is one of the rows of the rotation matrix, then $S(a)$ is the subset of samples outside the cube along the direction of a. We take the opposite of samples for x such that $x \cdot a < -1$ so that by symmetry of the cube, these samples contribute the same way as samples in $\{x : x \cdot a > 1\}$ when we compute the robust mean $\tilde{\mu}$. Given a corrupted sample X, suppose $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$ are the output of Algorithm 5.4.2. Then after removing samples in Step 2(a) of Algorithm 5.4.2 if necessary, we have

$$\frac{1}{n}|S(\hat{a}_{(i)}) \cap S(\hat{a}_{(j)})| \leqslant \frac{8C}{n^2}|S(\hat{a}_{(i)})||S(\hat{a}_{(j)})|.$$

We define ε_i as the fraction of outliers in $S(\hat{a}_{(i)})$ and η_i as the fraction of original uncorrupted sample in $S(\hat{a}_{(i)})$ that are removed by the adversary or by the algorithm. That is, ε_i and η_i are exactly the fraction of additive and subtractive corruption when we robustly estimate $\tilde{\mu}$ in Step 2(b). In the following Proposition, we show that the sum of ε_i's and η_i's is small, which implies we do not expand the influence of corruption by separately learning $a_{(1)}, \ldots, a_{(d)}$.

Proposition 5.4.3. *Given a corrupted sample X, suppose $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$ are the output of Algorithm 5.4.2. Let X_i be the subset of corrupted sample X after removing samples for all $j < i$ in Step 2(a) of Algorithm 5.4.2 when updating $a_{(i)}$ and $S(\hat{a}_{(i)}) = \{x : x \cdot \hat{a}_{(i)} > 1, x \in X_i\} \cup \{-x : x \cdot \hat{a}_{(i)} < -1, x \in X_i\}$. Let ε_i be the fraction of outliers in $S(\hat{a}_{(i)})$, and η_i be the fraction of original uncorrupted sample in $S(\hat{a}_{(i)})$ that are removed by the adversary or the algorithm. Suppose for all i, $\hat{a}_{(i)} \cdot e_i \geqslant 0.99$ and for all i, j,*

$$\frac{1}{n}|S(\hat{a}_{(i)}) \cap S(\hat{a}_{(j)})| \leqslant \frac{8C}{n^2}|S(\hat{a}_{(i)})||S(\hat{a}_{(j)})|.$$

Then,

$$\sum_{i \in [d]} (\varepsilon_i + \eta_i) \leqslant 6\varepsilon.$$

The proof of Proposition 5.4.3 is deferred to Section 5.4.3.

Next, we give a robust algorithm that updates a single rotation vector $\hat{a}_{(i)}$ of the cube with a small constant fraction of corruption in $\{x : |x \cdot \hat{a}_{(i)}| > 1\}$. We show that by repeatedly applying this algorithm on $\hat{a}_{(i)}$ at most $O(d \log d)$ times, we can learn the i-th true rotation vector with error $O(\varepsilon_i + \eta_i)$.

Proposition 5.4.4. *Suppose $H = [-1, 1]^d$ is the true hypercube. There exists a robust algorithm that given ε-corrupted samples from H and a unit vector $a^0_{(i)}$ such that*

$\left\|a^0_{(i)} - e_i\right\|_2 \leqslant 0.01$, *outputs*

$$\left\|\hat{a}_{(i)} - e_i\right\|_2 \leqslant 2^{14}(\varepsilon_i + \eta_i)$$

after at most $2^{12} d \log d$ *iterations, where* $\varepsilon_i + \eta_i$ *is the fraction of corruption in* $\{x : x \cdot \hat{a}_{(i)} > 1\}$.

The proof of Proposition 5.4.4 is deferred to Section 5.4.4. With the two propositions above, we complete the proof of Theorem 5.4.1 in Section 5.4.5.

5.4.2 Properties of Hypercubes

In this section, we prove a few basic properties of uniform samples from $H = [-1, 1]^d$, in particular, of the subset of samples $\{x \in H : x \cdot a > 1\}$ for a unit vector a. We will later use them in our analysis of Algorithm 5.4.2.

The following lemma is crucial for our analysis since it connects the 2-norm error of $\hat{a}_{(i)}$ and the volume of $\widehat{A}H \setminus AH$. Hence it implies Lemma 5.1.2.

Lemma 5.4.5. *Let* Y *be a uniform sample from* H *of size* n. *Suppose* a *is a unit vector such that* $\|a - e_i\| = \delta$. *Then with high probability the number of samples in* $S(a) \cap Y$ *is*

$$\frac{\delta}{5} n \leqslant |S(a) \cap Y| := |\{x \in Y : |x \cdot a| > 1\}| \leqslant \frac{(1+\delta)\delta}{4} n.$$

Proof. Let $\bar{a}$ be the $(d-1)$-vector that excludes i-th coordinate of a. Since $\|a - e_i\| = \delta$, we have that $a_i = a \cdot e_i = 1 - \delta^2/2$ and $\|\bar{a}\| = \delta + O(\delta^2)$. Let $S^+(a) := \{x : x \cdot a > 1\}$. By symmetry of the cube, $2\,\mathrm{vol}(S^+(a)) = \mathrm{vol}(S(a))$. So it suffices to compute the number of original uncorrupted sample in $S^+(a)$ and double it. We can compute the volume of $S^+(a)$ by integrating the length of x_i over $\bar{x}$ where $\bar{x}$ is the vector x without entry x_i. The upper bound on x_i is 1 and the lower bound is achieved when $x \cdot a = a_i x_i + \bar{x} \cdot \bar{a} = 1$, i.e.,

$x_i = \frac{1-\bar{x}\cdot\bar{a}}{a_i}$. So we have

$$\mathrm{vol}(S^+(a)) = \int_{\bar{x}\in[-1,1]^{d-1}} \max\left\{0, 1 - \frac{1-\bar{x}\cdot\bar{a}}{a_i}\right\}\, d\bar{x}$$

Let $y = \bar{x}\cdot\bar{a}$ be a random variable. Suppose $p(y)$ is the density of y. Let $1 - \frac{1-\bar{x}\cdot\bar{a}}{a_i} \geqslant 0$. We have $y \geqslant 1 - a_i = \delta^2/2$. Thus, we can write the integral above as

$$\mathrm{vol}(S^+(a)) = 2^{d-1}\left(\mathbb{P}\left(y \geqslant \frac{\delta^2}{2}\right)\left(1 - \frac{1}{a_i}\right) + \frac{1}{a_i}\int_{y\geqslant\frac{\delta^2}{2}} y\cdot p(y)\, dy\right). \tag{5.4.1}$$

Since $\bar{x}$ is a uniform random variable on the $(d-1)$-cube $[-1,1]^{d-1}$, y is a scaled 1d-projection of the standard $(d-1)$-cube onto an arbitrary direction $\frac{\bar{a}}{\|\bar{a}\|}$ with zero mean and variance $\|\bar{a}\|^2/3 = \delta^2/3$. Since y is logconcave and symmetric, we can bound the probability $1/2 - \delta \leqslant \mathbb{P}\left(y \geqslant \frac{\delta^2}{2}\right) \leqslant 1/2$ by Lemma 5.2.5. We define

$$\mu_y := \mathbb{E}(y \mid y \geqslant \delta^2/2) = \frac{\int_{y\geqslant\frac{\delta^2}{2}} y\cdot p(y)\, dy}{\int_{y\geqslant\frac{\delta^2}{2}} p(y)\, dy}$$

as the truncated mean of y. Then

$$\mathrm{vol}(S^+(a)) = 2^{d-1}\,\mathbb{P}\left(y \geqslant \frac{\delta^2}{2}\right)\left(1 - \frac{1}{a_i} + \frac{\mu_y}{a_i}\right) \tag{5.4.2}$$

By Lemma 5.2.2, μ_y is upper and lower bounded by $\Theta(\delta)$. The upper bound is achieved when $\bar{a}$ is in the direction of true coordinates, i.e., y is a uniform distribution and $\mu_y \leqslant (1+\delta)\delta/2$. Let y^+ be the truncated density of y on $\mathbb{R}^+$. Then raw moments of y^+ are $M_0(y^+) = 1, M_1(y^+) \leqslant \mu_y, M_2(y^+) = \delta^2/3$. By Lemma 5.2.6, $M_1^2 \geqslant M_0\frac{M_2}{2}$. Then $\mu_y \geqslant M_1 \geqslant \frac{1}{\sqrt{6}}\delta \geqslant \frac{2}{5}\delta$. Plugging the bounds of $\mathbb{P}\left(y \geqslant \frac{\delta^2}{2}\right)$ and μ_y into (5.4.2), we get the upper and lower bounds of $(1 - \frac{1}{a_i} + \frac{\mu_y}{a_i})$

$$\frac{2\delta}{5} - \frac{\delta^2}{2} \leqslant 1 - \frac{1}{a_i} + \frac{\mu_y}{a_i} = \frac{\mu_y - \delta^2/2}{1 - \delta^2/2} \leqslant \mu_y \leqslant \frac{\delta + \delta^2}{2}$$

and the upper and lower bounds of $\mathrm{vol}(S^+(a))$

$$\frac{2^{d-1}}{5}\delta \leqslant \mathrm{vol}(S^+(a)) \leqslant 2^{d-3}(1+\delta)\delta.$$

Since the volume of the standard cube H is 2^d, the probability a uniformly random sample is in $S^+(a)$ is at least $2^{d-1}\delta/(5 \cdot 2^d) = \delta/10$ and at most $2^{d-3}(1+\delta)\delta/2^d = (1+\delta)\delta/8$. Thus the fraction of samples in $S(a)$ is at least $\delta/5$ and at most $(1+\delta)\delta/4$. □

Lemma 5.4.6. *Suppose* $\|a - e_i\| = \delta$. *Let* μ *be the mean of the subset of the cube:* $\{x \in [-1,1]^d : a \cdot x > 1\}$. *Then* $\mu \cdot e_i \leqslant 1 - \delta/10$.

Proof. Let $S^+(a) = \{x \in [-1,1]^d : x \cdot a > 1\}$ and $S' = \{x \in [-1,1]^d : x \cdot a > 1, x_i = x \cdot e_i > 1 - c\delta\}$ for a fixed constant $c < 1$. Then by the trivial upper bound: $x_i \leqslant 1$ for any x in the cube,

$$\mu \cdot e_i \leqslant \frac{\mathrm{vol}(S') + (1 - c\delta)(\mathrm{vol}(S^+(a)) - \mathrm{vol}(S'))}{\mathrm{vol}(S^+(a))} = 1 - c\delta\left(1 - \frac{\mathrm{vol}(S')}{\mathrm{vol}(S^+(a))}\right).$$

We can compute the volume of S' by integral over x_i.

$$\begin{aligned}\frac{\mathrm{vol}(S')}{\mathrm{vol}(H)} &= \frac{1}{2}\int_{1-c\delta}^{1} \mathbb{P}(x \cdot a > 1 \mid x_i)\, dx_i \\ &= \frac{1}{2}\int_{1-c\delta}^{1} \mathbb{P}(\bar{x} \cdot \bar{a} \geqslant 1 - a_i x_i)\, dx_i\end{aligned}$$

Let $y = \bar{x} \cdot \bar{a}/\|\bar{a}\|$. Since x_1 and $\bar{x}$ are independent, $\bar{x}$ is uniformly random over the $d-1$ standard cube and y is the projection of $\bar{x}$ onto an arbitrary unit vector. Let $t = (1 - x_i)/\delta$

and hence $x_i = 1 - t\delta$. Then $dx_i = -\delta\ dt$. By substitution,

$$\frac{\mathrm{vol}(S')}{\mathrm{vol}(H)} = \frac{1}{2}\int_{t=0}^{c} \mathbb{P}\left(y \geqslant \frac{1 - a_i x_i}{\|\bar{a}\|}\right)\delta\ dt \leqslant \frac{1}{2}\int_{t=0}^{c} \mathbb{P}(y \geqslant t)\delta\ dt$$

By Lemma 5.2.7, $\mathbb{P}(y \geqslant t) \leqslant \frac{1}{2}(1-t)$ for any $t \leqslant 3/4$. Take $c = 1/5$. Then we get

$$\frac{\mathrm{vol}(S')}{\mathrm{vol}(H)} \leqslant \frac{1}{2}\int_{t=0}^{c} \frac{1}{2}(1-t)\delta\ dt = \frac{1}{4}(c - c^2/2)\delta < \frac{1}{20}\delta.$$

By Lemma 5.4.5, $\frac{\mathrm{vol}(S^+(a))}{\mathrm{vol}(H)} \geqslant \delta/10$. Thus

$$\mu \cdot e_i = 1 - c\delta\left(1 - \frac{\mathrm{vol}(S')}{\mathrm{vol}(S^+(a))}\right) = 1 - c\delta\left(1 - \frac{\mathrm{vol}(S')/\mathrm{vol}(H)}{\mathrm{vol}(S^+(a))/\mathrm{vol}(H)}\right) < 1 - \delta/10.$$

□

Lemma 5.4.7. *Let H be the standard hypercube $[-1,1]^d$. Suppose $\|a - e_i\| = \delta$. Then the variance of uniform distribution on $\{x \in [-1,1]^d : x \cdot a > 1\}$ in the direction e_i is at most $\delta^2/4$.*

Proof. Suppose $p(x_i)$ is the density function of x_i on $\{x : x \cdot a > 1\}$ and $\bar{x}, \bar{a}$ are vectors constructed from x, a by excluding i-th coordinate. Then

$$p(x_i) \propto \mathrm{vol}_{d-1}(\{x : x \cdot a > 1, x_i \text{ is fixed}\}) \propto \mathbb{P}_{\bar{x}}(\bar{x} \cdot \bar{a} > 1 - a_i x_i).$$

Let $y = \bar{x} \cdot \bar{a}$. Then y is logconcave and symmetric with mean 0 and variance $\delta^2/3$. By Lemma 5.2.6, we have the mean of x_i with density function $p(x_i)$ is $\mu(x_i) \geqslant 1 - 1/p(1)$.

$$p(1) \propto \mathbb{P}(y > \delta^2/2) \geqslant \frac{1}{2} - \delta.$$

Since the integral of $\mathbb{P}_{\bar{x}}(\bar{x} \cdot \bar{a} > 1 - a_i x_i)$ over x_i is the volume of the subset $\{x : x \cdot a > 1\}$, the integral is upper bounded by $(1+\delta)\delta/8$. After normalization, we get

$$p(1) > \frac{8(\frac{1}{2}-\delta)}{(1+\delta)\delta} > \frac{2}{\delta}.$$

So $\mu(x_i) \geqslant 1 - 1/p(1) \geqslant 1 - \delta/2$. Then by Lemma 5.2.6, the variance of x_i is bounded by the raw second moment, which is

$$M_2(p) \leqslant \frac{2M_1(p)^2}{M_0(p)} \leqslant 2(1-\mu(x_i))^2 \leqslant \delta^2/4.$$

□

5.4.3 Rotation Matrices can be Learned Coordinate by Coordinate

In this section, we prove Proposition 5.4.3 via Lemmas 5.4.8 and 5.4.9. Recall that $S(\hat{a}_{(i)}) = \{x : x \cdot \hat{a}_{(i)} > 1\} \cup \{-x : x \cdot \hat{a}_{(i)} < -1\}$ and ε_i is the fraction of outliers in $S(\hat{a}_{(i)})$ after filtering, η_i is the fraction of original uncorrupted sample in $S(\hat{a}_{(i)})$ that are removed.

Lemma 5.4.8. *Given a ε-corrupted sample X with $\varepsilon < 2^{24}/C$, suppose $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$ are the output of Algorithm 5.4.2. Let X_i be the subset of corrupted sample X after removing samples for all $j < i$ in Step 2(a) of Algorithm 5.4.2 when updating $a_{(i)}$ and $S(\hat{a}_{(i)}) = \{x : x \cdot \hat{a}_{(i)} > 1, x \in X_i\} \cup \{-x : x \cdot \hat{a}_{(i)} < -1, x \in X_i\}$. Let ε_i be the fraction of outliers in $S(\hat{a}_{(i)})$. Then $\sum_{i\in[d]} \varepsilon_i \leqslant 2\varepsilon$.*

Proof. Let $S_N(\hat{a}_{(i)})$ be the set of outliers that in the halfspace $\{x : x \cdot \hat{a}_{(i)} > 1\}$. We know the fraction of all outliers is at most ε. So we have $\mathrm{frac}\left(\bigcup S_N(\hat{a}_{(i)})\right) \leqslant \varepsilon$. We may assume that $\varepsilon_i \geqslant c_3 \mathrm{frac}(S(\hat{a}_{(i)}))$ where constant $c_3 = 2^{14}$ as in Corollary 5.4.13. Otherwise, we can improve $\hat{a}_{(i)}$ and hence reduce $\mathrm{frac}(S(\hat{a}_{(i)}))$ by running the updating algorithm in Section 5.4.4 (Corollary 5.4.13). Then

$$\mathrm{frac}(S_N(\hat{a}_{(i)}) \cap S_N(\hat{a}_{(j)})) \leqslant \mathrm{frac}(S(\hat{a}_{(i)}) \cap S(\hat{a}_{(j)})) \leqslant 8C \mathrm{frac}(S(\hat{a}_{(i)}))\mathrm{frac}(S(\hat{a}_{(j)})) \leqslant \frac{8C}{c_3^2}\varepsilon_i\varepsilon_j.$$

Set $\alpha = \frac{8C}{c_3^2}$, we get

$$\mathrm{frac}(S_N(\hat{a}_{(i)}) \cap S_N(\hat{a}_{(j)})) \leqslant \alpha \mathrm{frac}(S_N(\hat{a}_{(i)}))\mathrm{frac}(S_N(\hat{a}_{(j)})).$$

Then by Lemma 5.3.4, $\sum \varepsilon_i = \sum_{i \in [d]} \mathrm{frac}(S_N(\hat{a}_{(i)})) \leqslant \frac{\varepsilon}{1-\alpha\varepsilon}$. If $\varepsilon \leqslant \frac{c_3^2}{16C}$, then $1 - \alpha\varepsilon > 1/2$ and hence $\sum \varepsilon_i \leqslant 2\varepsilon$. □

Lemma 5.4.9. *Given a ε-corrupted sample X with $\varepsilon < 2^{24}/C$, suppose $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$ are the output of Algorithm 5.4.2. Let X_i be the subset of corrupted sample X after removing samples for all $j < i$ in Step 2(a) of Algorithm 5.4.2 when updating $a_{(i)}$ and $S(\hat{a}_{(i)}) = \{x : x \cdot \hat{a}_{(i)} > 1, x \in X_i\} \cup \{-x : x \cdot \hat{a}_{(i)} < -1, x \in X_i\}$. Let η_i be the fraction of original uncorrupted sample in $S(\hat{a}_{(i)})$ that are removed by the adversary or the algorithm. Suppose for all i, $\hat{a}_{(i)} \cdot e_i \geqslant 0.99$ and for all i, j,*

$$\frac{1}{n}|S(\hat{a}_{(i)}) \cap S(\hat{a}_{(j)})| \leqslant \frac{8C}{n^2}|S(\hat{a}_{(i)})||S(\hat{a}_{(j)})|$$

then $\sum_{i \in [d]} \eta_i \leqslant 4\varepsilon$.

Proof. Suppose ε'_i is the fraction of original uncorrupted sample deleted by the adversary in $S(\hat{a}_{(i)})$. Then $\sum_{i \in [d]} \varepsilon'_i \leqslant \varepsilon$ because we only check the intersections for $j < i$. We first consider the case that $\left\|\hat{a}_{(i)} - e_i\right\| = \delta_i \geqslant 2\varepsilon'_i$ and $\left\|\hat{a}_{(j)} - e_j\right\| = \delta_j \geqslant 2\varepsilon'_j$. By Lemma 5.2.1, the number of original uncorrupted sample in $\{x : |x \cdot \hat{a}_{(i)}| > 1, |x \cdot \hat{a}_{(j)}| > 1\}$ is at most $C\delta_i\delta_j$. The number of points that are deleted by the algorithm in $\{x : |x \cdot \hat{a}_{(i)}| > 1, |x \cdot \hat{a}_{(j)}| > 1\}$ is at least $8C(\delta_i - \varepsilon'_i)(\delta_j - \varepsilon'_j)$. Then the fraction of original uncorrupted sample in the deleted points is at most

$$\frac{c_1\delta_i\delta_j}{8C(\delta_i - \varepsilon'_i)(\delta_j - \varepsilon'_j)} \leqslant \frac{1}{8(1 - 1/2)^2} \leqslant 1/2.$$

In this case, at least 1/2 of the points that are deleted by the algorithm are outliers. So the number of deleted original uncorrupted sample is at most 2ε.

If $\left\|\hat{a}_{(i)} - e_i\right\| < 2\varepsilon_i'$ (or $\left\|\hat{a}_{(j)} - e_j\right\| < 2\varepsilon_j'$), then we can upper bound the total number of original uncorrupted sample in $\{x : x \cdot \hat{a}_{(i)} > 1\}$ by $2\varepsilon_i'$. So the number of true points we deleted in this case is at most $2\sum_i \varepsilon_i' \leqslant 2\varepsilon$.

In total, the number of points deleted is bounded by $\sum_{i\in[d]} \eta_i \leqslant 4\varepsilon$. □

Combining Lemmas 5.4.8 and 5.4.9 gives the conclusion of Proposition 5.4.3.

5.4.4 Update Rotation Vectors by the Mean of Outside Samples

As in the previous sections, we assume a is the current rotation vector and e is the true rotation vector such that $\|a - e\| = \delta$. We first give a non-robust algorithm that improves a. Given an uncorrupted sample $X = \{x^{(1)}, \dots, x^{(n)}\}$ and a unit vector a, the algorithm computes μ of the subset of the sample

$$\{x : x \cdot a > 1\} \cup \{-x : x \cdot a < -1\}$$

and outputs

$$a' = \frac{a - \beta\mu}{\|a - \beta\mu\|}.$$

The following lemma shows that the non-robust algorithm improves $a \cdot e$ by $\beta\delta/64$.

Lemma 5.4.10. *If* $\|a - e\| = \delta \leqslant 0.01$*, then the non-robust algorithm with the step size* $\beta \leqslant \frac{\delta}{32\|\mu\|^2}$ *outputs a unit vector* a' *such that* $a' \cdot e - a \cdot e \geqslant \beta\delta/64$.

Proof. By the definition of a', we have

$$a' \cdot e = \frac{a \cdot e - \beta\mu \cdot e}{\sqrt{(a - \beta\mu)(a - \beta\mu)}}$$

Since $\|a - e\| = \delta$ and a, e are unit vectors, $a \cdot e = (2 - \|a - e\|^2)/2 = 1 - \delta^2/2$. By Lemma 5.4.6, we know $\mu \cdot e \leqslant 1 - \delta/32$. Plugging into the right hand side of the equation above, we get

$$
\begin{aligned}
a' \cdot e &\geqslant \frac{1 - \delta^2/2 - \beta(1 - \delta/32)}{\sqrt{1 - 2\beta + \beta^2 \|\mu\|^2}} \\
&\geqslant \frac{1 - \beta + \beta\delta/32 - \delta^2/2}{\sqrt{1 - 2\beta + \beta\delta/32}} \\
&\geqslant (1 - \beta + \beta\delta/32 - \delta^2/2)(1 + \beta - \beta\delta/64 + \frac{3}{2}\beta^2) \\
&\geqslant 1 - \delta^2/2 + \beta\delta/64
\end{aligned}
$$

where the second inequality follows from our assumption that $\beta \leqslant \frac{\delta}{32\|\mu\|^2}$. Thus $a' \cdot e - a \cdot e \geqslant \beta\delta/64$. □

Now we state the robust version of the algorithm, where we replace the mean μ in Step 1 by the robust mean $\tilde{\mu}$. We will use the robust mean estimation algorithm for bounded covariance distributions in Theorem 2.2.1.

Algorithm 5.4.11 (Robust Rotation Vector Update)**.**

Input: *corrupted sample $x^{(1)}, \ldots, x^{(n)}$ and a unit vector a such that $\|a - e\| = \delta$.*

Output: *a unit vector a'*

Parameters: *step size $\beta > 0$*

1. *Compute the robust mean $\tilde{\mu}$ of the following set using the algorithm in Theorem 2.2.1*

$$\{x : x \cdot a > 1\} \cup \{-x : x \cdot a < -1\}$$

2. *Output*

$$a' = \frac{a - \beta\tilde{\mu}}{\|a - \beta\tilde{\mu}\|}$$

Suppose the number of corrupted sample in the subset $\{x : |x \cdot a| > 1\}$ is $\varepsilon_i n$. With the similar argument as in the proof of Lemma 5.4.10, we can show the following result.

Lemma 5.4.12. *Suppose $\|a-e\| = \delta$ and the number of corrupted sample in the subset $\{x : |x \cdot a| > 1\}$ is $\varepsilon_i n$. If $\varepsilon_i \leqslant 2^{-14}\delta$, then the robust version of the algorithm with the step size $\beta \leqslant \frac{\delta}{32\|\tilde{\mu}\|^2}$ outputs a unit vector a' such that $a' \cdot e_i - a \cdot e_i \geqslant \frac{1}{128}\beta\delta$.*

Proof. By the definition of a', we have

$$a' \cdot e = \frac{a \cdot e - \beta\tilde{\mu} \cdot e}{\sqrt{(a-\beta\tilde{\mu})(a-\beta\tilde{\mu})}}$$

Since $\|a-e\| = \delta$ and a, e are unit vectors, $a \cdot e = (2 - \|a-e\|^2)/2 = 1 - \delta^2/2$.

$$\begin{aligned}
a' \cdot e_i &\geqslant \frac{1-\delta^2/2 - \beta\tilde{\mu} \cdot e}{\sqrt{1 - 2\beta\tilde{\mu} \cdot a + \beta^2 \|\mu\|^2}} \\
&\geqslant (1-\delta^2/2 - \beta\tilde{\mu} \cdot e)(1 + \beta\tilde{\mu} \cdot a - \beta^2 \|\tilde{\mu}\|^2/2 + O(\beta^2)) \\
&\geqslant 1 - \delta^2/2 + \beta(\tilde{\mu} \cdot a - \tilde{\mu} \cdot e) - \beta^2 \|\tilde{\mu}\|^2/2 - \frac{\delta^2}{2}\beta\tilde{\mu} \cdot a - \beta^2(\tilde{\mu} \cdot a)(\tilde{\mu} \cdot e) \\
&= 1 - \delta^2/2 + \beta\left((\tilde{\mu} \cdot a)(1 - \delta^2/2 - \beta(\tilde{\mu} \cdot e)) - \tilde{\mu} \cdot e - \beta\|\tilde{\mu}\|^2/2\right).
\end{aligned}$$

By Lemma 5.4.6, we know $\mu \cdot e_i \leqslant 1 - \delta/10$. Assuming that the fraction of corruption is c, we apply the algorithm in Fact 2.2.1 to compute the robust mean $\tilde{\mu}$, which gives an error guarantee that $(\tilde{\mu} - \mu) \cdot e_i \leqslant O(\sqrt{c})\sigma(e_i)$ where $\sigma^2(e_i)$ is the variance of $\{x : x \cdot a > 1\}$ in the direction e_i. By Lemma 5.4.7, $\sigma^2(e_i)$ is at most δ. Thus we have $\tilde{\mu} \cdot e \leqslant 1 - (1/10 - \sqrt{c})\delta$. By Lemma 5.4.5, the fraction of points in $\{x : x \cdot a > 1\}$ is at least $(\delta/5)n$. So the fraction of corruption is bounded by

$$c \leqslant \frac{\varepsilon_i}{\delta/5} \leqslant 5 \cdot 2^{-14}$$

where the last inequality follows from our assumption that $\varepsilon_i \leqslant 2^{-14}\delta$. By the trivial upper bound $\tilde{\mu} \cdot a > 1$ and our condition $\|\tilde{\mu}\|^2 \leqslant \frac{\delta}{32\beta}$, we get

$$a' \cdot e_i - a \cdot e_i \geqslant \beta\left((\frac{1}{10} - \sqrt{c})\delta - \frac{\delta^2}{2} - \frac{\delta}{32} - \frac{\delta}{64}\right) \geqslant \frac{1}{128}\beta\delta.$$

$\square$

Corollary 5.4.13. *Suppose we start with $\left\|a_{(i)}^0 - e_i\right\| \leqslant \sqrt{\varepsilon}$. By iteratively running Algorithm 5.4.11 on $a_{(i)}^t$ $2^{12} d \log d$ times, one of $a_{(i)}^t$ satisfies*

$$\left\|a_{(i)}^t - e_i\right\| \leqslant 2^{14} \varepsilon_i.$$

Proof. We start with $\|a - e_i\| = \delta$. Each time we run the algorithm, by Lemma 5.4.12 and the upper bound $\sqrt{d}$ on norm of the mean,

$$a' \cdot e_i - a \cdot e_i \geqslant \frac{1}{128} \beta \delta = \frac{\delta^2}{128 \cdot 32 \|\tilde{\mu}\|^2} \geqslant \frac{\delta^2}{2^{12} d}.$$

Then

$$\|a' - e\| / \|a - e\| \leqslant \sqrt{1 - \frac{1}{2^{11} d}}.$$

The optimal error we can get by repeatedly running this algorithm is $2^{14} \varepsilon_i$ from the condition of Lemma 5.4.12. When $\|a - e_i\| > 2^{14} \varepsilon_i$, we have $\varepsilon_i \leqslant 2^{-14} \delta$ so running Algorithm 5.4.11 with input a will give us a better a' as in Lemma 5.4.12. Suppose after t steps, the error is reduced to $2^{14} \varepsilon_i$. Then

$$\left(\sqrt{1 - \frac{1}{2^{11} d}}\right)^t \delta = 2^{14} \varepsilon_i.$$

Since $\delta \leqslant \sqrt{\varepsilon}$ and $\varepsilon_i \geqslant \varepsilon / d$, we get

$$t \leqslant 2^{12} \cdot d \cdot \log \frac{\delta}{2^{14} \varepsilon_i} \leqslant 2^{12} d \log d.$$

$\square$

5.4.5 Proof of Theorem 5.4.1

Let ε_i^t be the fraction of outliers in the set $\{x : x \cdot a_{(i)}^t > 1\}$ and η_i^t be the fraction of original uncorrupted sample removed from the set $\{x : x \cdot a_{(i)}^t > 1\}$.

Proof of Theorem 5.4.1. By Corollary 5.4.13, we know there exists an $a^t_{(i)}$ for all i such that $\left\|a^t_{(i)} - e_i\right\| \leqslant 2^{10}(\varepsilon^t_i + \eta^t_i)$. First we show that $A = (a^t_{(1)}, \ldots, a^t_{(d)})^\top$ is a good rotation matrix that AH is close to H in TV distance. By Lemmas 5.4.8 and 5.4.9,

$$\sum_i (\varepsilon^t_i + \eta^t_i) \leqslant 6\varepsilon.$$

Let $S(a) = \{x : x \cdot a > 1\}$ be a subset of samples after removing points in Step 2(a) of Algorithm 5.4.2. Then by Lemma 5.4.5,

$$\sum_{i\in[d]} |S(a^t_{(i)})| \leqslant \sum_{i\in[d]} \frac{1}{4} \left\|a^t_{(i)} - e_i\right\| n \leqslant \frac{1}{4} \sum_{i\in[d]} 2^{14}(\varepsilon^t_i + \eta^t_i) \leqslant 2^{15}\varepsilon n.$$

Next, we prove $\hat{A} = (\hat{a}_{(1)}, \ldots, \hat{a}_{(d)})^\top$ is a good estimation of the true rotation matrix with respect to TV distance by showing the number of samples inside $\hat{A}H$ is at least $(1 - c\varepsilon)n$. The number of samples in $\hat{A}H$ is at least

$$n - \sum_i |S(\hat{a}_{(i)})| \geqslant n - \sum_i |S(a^t_{(i)})| \geqslant (1 - 2^{15}\varepsilon)n.$$

where the first inequality follows from $|S(\hat{a}_i)|$ is the smallest over all t. Since each iteration is polytime and the total number of iterations is $O(d^2 \log d)$, the overall complexity is polynomial in d. $\square$

5.5 General Case

In this section, we give an algorithm to robustly learn an affine transformation of shift, scaling and rotation and prove our main result:

Theorem 5.5.1. *Given $X = \{x^{(1)}, \ldots, x^{(n)}\}$ an ε-corrupted sample of points from an unknown cube $H = \{x \in \mathbb{R}^d : v^*_i \leqslant a^*_{(i)} \cdot x \leqslant u^*_i, \forall i \in [d]\}$, where $n \geqslant n_0 = \widetilde{O}(d^2/\varepsilon^2)$, there exists a polynomial-time algorithm that outputs $\widehat{H} = \{x \in \mathbb{R}^d : \hat{v}_i \leqslant \hat{a}_{(i)} \cdot x \leqslant$*

$\hat{u}_i, \forall i \in [d]\}$ *s.t.* $d_{TV}(\widehat{H}, H) = O(\varepsilon)$.

Algorithm 5.5.2 (Robust Estimation of Affine Transformations)**.**

1. *Run the robust mean and covariance estimation algorithms in Theorem 2.2.1 and get estimates of the mean* $\hat{\mu}$ *and covariance* $\widehat{\Sigma}$ *of the cube with error*

$$\|\hat{\mu} - \mu^*\| \leqslant C\sqrt{\varepsilon}\,\|\Sigma^*\|_2^{1/2}$$
$$\left\|\widehat{\Sigma} - \Sigma^*\right\|_2 \leqslant C\sqrt{\varepsilon}\,\|\Sigma^*\|_2 .$$

2. *Run the robust moment estimation algorithm in Theorem 2.2.4 and get an estimate of the rotation matrix* $\widehat{A}$ *with unit rows* $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$.

3. *Repeat the following steps until the number of samples in* $\{x : \hat{v}_i \leqslant \hat{a}_{(i)} \cdot x \leqslant \hat{u}_i, \forall i\}$ *is at least* $(1 - c\varepsilon)n$.

   (a) *Run Algorithm 5.3.3 with* $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$ *being the coordinates and update upper bounds and lower bounds of the cube* $\hat{u}_1, \ldots, \hat{u}_d, \hat{v}_1, \ldots, \hat{v}_d$

   (b) *Make the sample nearly isotropic with the estimates* $\hat{u}_1, \ldots, \hat{u}_d, \hat{v}_1, \ldots, \hat{v}_d$

   (c) *Run Algorithm 5.4.2 with upper bounds and lower bounds* $\hat{u}_1, \ldots, \hat{u}_d, \hat{v}_1, \ldots, \hat{v}_d$ *and update* $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$.

4. *Output the cube* $\widehat{H} = \{x : \hat{v}_i \leqslant \hat{a}_{(i)} \cdot x \leqslant \hat{u}_i, \forall i\}$.

We analyze our algorithm by the following two propositions. Proposition 5.5.3 shows that the shift and scaling algorithm (Step 3(a)) learns the upper and lower bounds of the facets with error $O(\varepsilon) + 0.7\delta$ where δ is the error in rotation matrix. Proposition5.5.4 shows that the rotation algorithm (Step 3(b)) learns the rotation matrix with error $O(\varepsilon) + \delta$ where δ is the error in shift and scaling part.

Proposition 5.5.3. *Suppose* $\left\|\hat{a}_{(i)} - a^*_{(i)}\right\| = \delta_i \leqslant C\sqrt{\varepsilon}$. *Algorithm 5.3.3 outputs* $\hat{u}_1, \ldots, \hat{u}_d, \hat{v}_1, \ldots, \hat{v}_d$ *such that*

$$\frac{|\hat{u}_i - u^*_i| + |\hat{v}_i - v^*_i|}{u^*_i - v^*_i} \leqslant 44\varepsilon_i + 0.7\delta_i$$

and $\sum_{i\in[d]} \varepsilon_i \leqslant 2\varepsilon$.

Proposition 5.5.4. *Suppose* $\frac{|\hat{u}_i - u^*_i| + |\hat{v}_i - v^*_i|}{u^*_i - v^*_i} = \delta_i \leqslant C\sqrt{\varepsilon}$. *Algorithm 5.4.2 outputs* $\hat{A}$ *with unit rows* $\hat{a}_{(1)}, \ldots, \hat{a}_{(d)}$ *such that*

$$\left\|\hat{a}_{(i)} - a^*_{(i)}\right\| \leqslant 2^{20}\varepsilon_i + \delta_i$$

where $\sum_{i\in[d]} \varepsilon_i \leqslant 2\varepsilon$.

We can now prove the main theorem. For the purpose of analysis, we assume that the true cube is the standard cube $H = [-1, 1]^d$ in the remaining of Section 5.5.

Proof of Theorem 5.5.1. Let $\delta_{s,i} = |\hat{u}_i - 1| + |\hat{v}_i - 1|$ be the scaling error in coordinate i and $\delta_{r,i} \leqslant \left\|\hat{a}_{(i)} - e_i\right\|$ be the rotation error in coordinate i. By Theorems 2.2.1 and 2.2.4, we start with $\delta_{s,i} = |\hat{u}_i - 1| + |\hat{v}_i - 1| \leqslant C\sqrt{\varepsilon}$ and $\delta_{r,i} = \left\|\hat{a}_{(i)} - e_i\right\| \leqslant C\sqrt{\varepsilon}$ for some constant C. By Proposition 5.5.3, Step 3(a) of Algorithm 5.5.2 improves $\delta_{s,i}$ to $\delta'_{s,i} \leqslant 44\varepsilon_{s,i} + 0.7\delta_{r,i}$. With this updated $\delta_{s,i}$, by Proposition 5.5.4, Step 3(b) improves $\delta_{r,i}$ to $\delta'_{r,i} \leqslant c\varepsilon_{r,i} + \delta'_{s,i} \leqslant 2^{20}\varepsilon_{r,i} + 44\varepsilon_{s,i} + 0.7\delta_{r,i}$. If $\delta_{r,i} \geqslant 220\varepsilon_{s,i} + 5 \cdot 2^{20}\varepsilon_{r,i}$, we have $\delta'_{r,i} \leqslant 0.9\delta_{r,i}$. So we can decrease $\delta'_{r,i}$ and $\delta'_{s,i}$ by alternating between Algorithm 5.3.3 and Algorithm 5.4.2. Otherwise, we have

$$\sum_{i\in[d]} \left\|\hat{a}_{(i)} - e_i\right\| = \sum_{i\in[d]} \delta_{r,i} \leqslant \sum_{i\in[d]} 220\varepsilon_{s,i} + 5 \cdot 2^{20}\varepsilon_{r,i} \leqslant (440 + 10 \cdot 2^{20})\varepsilon \tag{5.5.1}$$

where the last equality follows from the upper bounds on the sum of $\varepsilon_{s,i}$ and $\varepsilon_{r,i}$ in

Propositions 5.5.3 and 5.5.4. Then

$$\sum_{i\in[d]} |\hat{u}_i - 1| + |\hat{v}_i - 1| = \sum_{i\in[d]} \delta_{s,i} \leqslant \sum_{i\in[d]} 44\varepsilon_{s,i} + 0.7\delta_{r,i} \leqslant (396 + 7\cdot 2^{20})\varepsilon. \tag{5.5.2}$$

Thus, adding up (5.5.1) and (5.5.2), we get $d_{TV}(\widehat{H}, H) = O(\varepsilon)$. Next we show that the number of iterations is bounded. We start with $\delta_{r,i} \leqslant C\sqrt{\varepsilon}$ and end with $\delta_{r,i} \geqslant \varepsilon/d$. Suppose we run Step 3 for t iterations. Then $0.9^t C\sqrt{\varepsilon} \leqslant \varepsilon/d$, which implies $t \leqslant C\log(d/\varepsilon)$. □

5.5.1 Proof of Proposition 5.5.3

In this section, we will prove Proposition 5.5.3. The idea is similar with the proof of Theorem 5.3.1 in Section 5.3. But we need to deal with the error from the rotation matrix besides outliers. The following two lemmas are general versions of Lemmas 5.3.5 and 5.3.7 with rotation error.

Lemma 5.5.5. *Suppose a cube $\widehat{H} = \{x : \hat{v}_i \leqslant \hat{a}_{(i)}\cdot x \leqslant \hat{u}_i\}$ satisfies (5.3.1),(5.3.2),(5.3.3) in Algorithm 5.3.3 and $\delta_i = \left\|\hat{a}_{(i)} - a_{(i)}\right\|$, then with high probability, for all $\hat{u}_i > 1$,*

$$\hat{u}_i - 1 \leqslant 0.35\delta + 22\varepsilon_i$$

and for all $\hat{v}_i < -1$,

$$-1 - \hat{v}_i \leqslant 0.35\delta + 22\varepsilon_i$$

where $\varepsilon_i > 0$ and $\sum_{\{i:\hat{u}_i>1 \text{ or } \hat{v}_i<-1\}} \varepsilon_i \leqslant 2\varepsilon$.

Proof. Suppose $\delta_i = \left\|\hat{a}_{(i)} - e_i\right\|$ is the error of rotation vector in direction e_i and ε_i is the fraction of outliers in the set $\{x : 1 < \hat{a}_{(i)}\cdot x \leqslant \hat{u}_i\}$ for $\hat{u}_i > 1$. By (5.3.1), we know that the fraction of samples in $\{x : 1 < \hat{a}_{(i)}\cdot x \leqslant \hat{u}_i\}$ is at least $(1/2)(\hat{u}_i - 1)$. These points not in $\hat{H}$ are either outliers or because of the rotation error. By Lemma 5.4.5, the fraction of original uncorrupted sample in $\{x : 1 < \hat{a}_{(i)} \cdot x \leqslant \hat{u}_i\}$ is at most $\delta/8$. So we have

$$\frac{1}{2}(\hat{u}_i - 1) \leqslant \delta_i/8 + \varepsilon_i.$$

Now we consider the case $\varepsilon_i \leqslant 0.05\delta_i$. Then we get for i s.t. $\hat{u}_i > 1$ and $\varepsilon_i \leqslant 0.05\delta_i$,

$$\hat{u}_i - 1 \leqslant \delta_i/4 + 0.1\delta_i = 0.35\delta_i. \tag{5.5.3}$$

The above inequality also applies to v_i by symmetry. Next if $\varepsilon_i > 0.05\delta_i$, we have $(\hat{u}_i - 1) \leqslant 22\varepsilon_i$. Set $\alpha = 40 \cdot 22^2$. Then by (5.3.3), we have the fraction of the intersection of $\{x : 1 < \hat{a}_{(i)} \cdot x \leqslant \hat{u}_i\}$ and $\{x : 1 < \hat{a}_{(j)} \cdot x \leqslant \hat{u}_j\}$ is at most $\alpha\varepsilon_i\varepsilon_j$. Since the fraction of corruption is at most ε, we can apply Lemma 5.3.4 to outliers in $\{x : 1 < \hat{a}_{(i)} \cdot x \leqslant \hat{u}_i\} \cup \{x : \hat{v}_i \leqslant \hat{a}_{(i)} \cdot x < -1\}$. and get an upper bound of $\sum \varepsilon_i$

$$\sum_{\{i:\hat{u}_i>1,\varepsilon_i\leqslant 0.05\delta_i\}} \varepsilon_i \leqslant \frac{\varepsilon}{1-\alpha\varepsilon}.$$

If $\alpha\varepsilon \leqslant 1/2$, we get $\sum_{\{i:\hat{u}_i>1,\varepsilon_i\leqslant 0.05\delta_i\}} \varepsilon_i \leqslant 2\varepsilon$. Then for i s.t. $\hat{u}_i > 1$ and $\varepsilon_i \leqslant 0.05\delta_i$,

$$\hat{u}_i - 1 \leqslant 22\varepsilon_i. \tag{5.5.4}$$

Add (5.5.3) and (5.5.4) gives the conclusion. ◻

Lemma 5.5.6. *With high probability, Step 3(a) of Algorithm 5.5.2 outputs $\hat{u}_i$ and $\hat{v}_i$ for all $\hat{u}_i < 1$,*

$$1 - \hat{u}_i \leqslant 0.35\delta + 22\eta_i$$

and for all $\hat{v}_i > -1$,

$$\hat{v}_i - (-1) \leqslant 0.35\delta + 22\eta_i$$

where $\eta_i > 0$ and $\sum_{\{i:\hat{u}_i>1 \text{ or } \hat{v}_i<-1\}} \eta_i \leqslant 2\varepsilon$.

Proof. If $\hat{u}_i < 1$, by (5.3.1), the fraction of points in $\{x : \hat{u}_i \leqslant \hat{a}_{(i)} \cdot x \leqslant 1\}$ is less than $\frac{1}{2}(1 - \hat{u}_i)$. So at least half of points in the region are either removed (by adversary or by the algorithm) or because of the error from the rotation vector $\hat{a}_{(i)}$. By Lemma 5.4.5, the fraction of error from rotation vector is at most $\delta_i/8$.

$$\frac{1}{2}(1 - \hat{u}_i) \leqslant \delta_i/8 + \eta_i.$$

By Lemma 5.3.6, the fraction of points are removed is at most 2ε, i.e., the fraction of the union of the removed points is at most 2ε. By the same argument in the proof of Lemma 5.5.5, we can upper bound $\sum \eta_i$ by 4ε for $\eta_i \geqslant 0.05\delta$. □

Proof of Proposition 5.5.3. Lemma 5.5.5 proves the conclusion in the cases that $\hat{u}_i > 1$ or $\hat{v}_i < -1$. Lemma 5.5.6 proves the conclusion in the cases that $\hat{u}_i < 1$ or $\hat{v}_i > -1$. □

5.5.2 Proof of Proposition 5.5.4

In this section, we prove Proposition 5.5.4. The proof follows the same idea as the proofs of the rotation case in Section 5.4.

Lemma 5.5.7. *Suppose $\|a - e_i\| = \delta$. Then the variance of the subset of the standard cube $\{x : x \cdot a > 1 - \delta/2\}$ in the direction e_i is at most $c\delta^2$ for a constant $c > 0$.*

Proof. Since the fraction of points in $\{x : x \cdot a > 1 - \delta/2\}$ that satisfies $x_i \geqslant 1 - (3\delta/2)$ is at least $1 - c\delta$, by Lemma 5.2.3, we know that the mean of x_i on the subset $\{x : x \cdot a > 1 - \delta/2\}$ is at least $1 - (3\delta/2)$. Then by Lemma 5.2.2, the variance of x_i on the whole subset is upper-bounded by the variance of one side of the mean. □

Lemma 5.5.8. *Suppose $\|a - e_i\| = \delta$. Then the variance of the subset of the standard cube $\{x : x \cdot a > 1 - \delta/2\}$ in the direction a is at most $c\delta^2$ for a constant $c > 0$.*

Proof. Using the similar argument as in the proof of Lemma 5.5.7, it suffices to show that the mean of $x \cdot a$ on the subset $\{x : x \cdot a > 1 - \delta/2\}$ is at most $1 + \delta/2$, which follows from

Lemma 5.2.3 and the fact that the fraction of points in $\{x : x \cdot a > 1 - \delta/2\}$ that satisfies $x \cdot a \leqslant 1 + (3\delta/2)$ is at least $1 - c\delta$. □

Lemma 5.5.9. *Suppose $\|a - e_i\| = \delta$. The fraction of original uncorrupted sample of the standard cube in $\{x : |x \cdot a| > 1 + \delta/2\}$ is at least $\delta/64$ with high probability.*

Proof. We can compute the volume of $S_1 = \{x \in [-1,1]^d : x \cdot a > 1 + \delta/2\}$ by integrating the length of x_i over $\bar{x}$ where $\bar{x}$ is the vector x without entry x_i.

$$\mathrm{vol}(S_1) = \int_{\bar{x} \in [-1,1]^{d-1}} \max\left\{0, 1 - \frac{1 + \delta/2 - \bar{x} \cdot \bar{a}}{a_i}\right\} \, d\bar{x}$$

Let $y = \bar{x} \cdot \bar{a}$. Suppose $p(y)$ is the density of y. We can write the integral above as

$$\mathrm{vol}(S_1) = 2^{d-1}\left(\mathbb{P}\left(y \geqslant \frac{\delta^2}{2} + \delta/2\right)\left(1 - \frac{1 + \delta/2}{a_i}\right) + \frac{1}{a_i}\int_{y \geqslant \frac{\delta^2}{2} + \delta/2} y \cdot p(y) \, dy\right). \tag{5.5.5}$$

Since $\bar{x}$ is a uniform random variable on $[-1,1]^{d-1}$, we have y is 1d projection of the standard $d-1$ cube onto an arbitrary direction with zero mean and variance $\|\bar{a}\|^2/3 = \delta^2/3$. Then we can bound the probability $1/8 - \delta \leqslant \mathbb{P}\left(y \geqslant \frac{\delta^2}{2} + \delta/2\right) \leqslant 1/4$ and $\mathbb{P}(y \geqslant \frac{3}{4}\delta) \geqslant 1/8 - \delta$. By Lemma 5.2.9, we have $\mu_y := \mathbb{E}(y \mid y \geqslant \delta^2/2 + \delta/2) \geqslant (\delta/2) + (\delta^2/2) + c\delta$. Plugging into the inequality (5.5.5), we get

$$\begin{aligned}
\mathrm{vol}(S_1) &\geqslant 2^{d-1}\left(\mathbb{P}\left(y \geqslant \frac{\delta^2}{2} + \delta/2\right)\left(1 - \frac{1 + \delta/2}{a_i} + \frac{\mu_y}{a_i}\right)\right) \\
&\geqslant 2^{d-1}(1/8 - \delta)(\delta/8 + O(\delta^2)) \\
&= 2^{d-7}\delta.
\end{aligned}$$

Thus the probability a uniformly random sample is in S_1 is at least $2^{d-7}\delta/2^d = \delta/128$. By symmetry of the cube, the probability of a point that $x \cdot a < -1 - \delta/2$ is the same. So the fraction of samples in $\{x : |x \cdot a| > 1 + \delta/2\}$ is at least $\delta/64$. □

Lemma 5.5.10. *Suppose* $\|a - e_i\| = \delta$. *Let* μ *be the mean of the subset of the cube* $\{x \in [-1,1]^d : a \cdot x > 1 + \Delta\}$, *where* $\Delta \in [-\delta/2, \delta/2]$. *Then* $\mu \cdot a - \mu \cdot e_i \geqslant \delta/64$.

Proof. Let $S_1 = \{x \in [-1,1]^d : x \cdot a > 1 + \delta/2\}$. By the definition, we have $x \cdot a > 1 + \delta/2$ and $x \cdot e \leqslant 1$ for all $x \in S_1$. So $\mu(S_1) \cdot (a - e) \geqslant \delta/2$. Then we will show that for $t \in [-\delta/2, \delta/2]$, the mean of the section of the cube $\mu(1+t) := \mathbb{E}(x \mid x \cdot a = 1 + t)$ in the direction e_i is at most $1 + t$. This is trivial for $t \geqslant 0$. So we assume $-\delta/2 \leqslant t < 0$.

$$\begin{aligned}\mu(1+t) \cdot e_i &= \mathbb{E}(x_i \mid x \cdot a = 1 + t) \\ &= \int_{-1}^{1} x_i \, \mathbb{P}(\bar{x} \cdot \bar{a} = 1 + t - a_i x_i) \, dx_i\end{aligned}$$

where $\bar{x} \sim [-1,1]^{d-1}$. Then $\bar{x} \cdot \bar{a}$ is symmetric at zero, that is, at $x_i = 1 + t + O(\delta^2)$. So the truncated mean $\mathbb{E}(x_i \mid x \cdot a = 1 + t, x_i \geqslant 1 + 2t)$ is $1 + t$. Since $\mathbb{E}(x_i \mid x \cdot a = 1 + t, x_i < 1 + 2t) < 1 + 2t$, we conclude that the mean of these two parts $\mathbb{E}(x_i \mid x \cdot a = 1 + t)$ is at most $1 + t$.

By Lemma 5.5.9, the fraction of points in $S_1 = \{x \in [-1,1]^d : x \cdot a > 1 + \delta/2\}$ is at least $\delta/64$. By Lemma 5.4.5, the fraction of $\{x : x \cdot a > 1\}$ is at most $\delta/8$. By Lemma 5.2.8, the maximum of the section of the cube is $\sqrt{2}$. Then the fraction of $\{x : 1 + t \leqslant x \cdot a \leqslant 1\}$ is upper bounded by $\sqrt{2}t/2$. Thus

$$\mu \cdot a - \mu \cdot e \geqslant \frac{\mu(S_1) \cdot (a - e) \frac{\delta}{32}}{(\delta/8) - (t/\sqrt{2}))} \geqslant \delta/64.$$

□

Proof of Proposition 5.5.4. Suppose $\delta_i = \left\|\hat{a}_{(i)} - e_i\right\|$. In the proof of Lemma 5.4.12, we show that if $\tilde{\mu} \cdot \hat{a}_{(i)} - \tilde{\mu} \cdot e \geqslant c\delta_i$, then Algorithm 5.4.11 outputs a unit vector $\hat{a}'_{(i)}$ such that $\hat{a}'_{(i)} \cdot e_i - \hat{a}_{(i)} \cdot e_i \geqslant c'\beta\delta_i$. In the general algorithm, the only difference is we replace the true threshold in the direction e_i by $\hat{u}_i$, i.e., $\tilde{\mu}$ is the robust mean of the subset $\{x : \hat{a}_{(i)} \cdot x > \hat{u}_i\}$. Let μ_i be the true mean of $\{x : \hat{a}_{(i)} \cdot x > \hat{u}_i\}$. By Lemma 5.5.10, we have $\mu \cdot \hat{a}_{(i)} - \mu \cdot e_i \geqslant$

$\delta_i/64$. We apply the algorithm in Theorem 2.2.1 to compute the robust mean, which give an error guarantee that $(\tilde{\mu} - \mu) \cdot \hat{a}_{(i)} \leqslant O(\sqrt{c})\sigma(\hat{a}_{(i)})$ and $(\tilde{\mu} - \mu) \cdot e_i \leqslant O(\sqrt{c})\sigma(e_i)$ where $\sigma^2(\hat{a}_{(i)})$ is the variance of $\{x : \hat{a}_{(i)} \cdot x > \hat{u}_i\}$ in the direction $\hat{a}_{(i)}$ and $\sigma^2(e_i)$ is the variance in the direction e_i. By Lemmas 5.5.7 and 5.5.8, the variances are at most δ^2. By Lemma 5.5.9, the number of true points in $\{x : \hat{a}_{(i)} \cdot x > \hat{u}_i\}$ is at least $(\delta/64)n$. Thus if

$$c \leqslant \frac{\varepsilon_i}{\delta/64} \leqslant 2^{14},$$

the algorithm can improve $\hat{a}_{(i)}$ to $\hat{a}'_{(i)}$ such that

$$\hat{a}'_{(i)} \cdot e_i - \hat{a}_{(i)} \cdot e_i \geqslant \frac{1}{256}\beta\delta_i.$$

This implies that the algorithm make progress until $\delta < 2^{20}\varepsilon_i + \delta_i$. Then the argument in the proofs of Corollary 5.4.13 and Theorem 5.4.1 follows. □

5.6 Reverse Union Bound

In this section, we prove Lemma 5.3.4, which can be seen as a reverse union bound with the assumption that pairs of events are not too highly correlated.

Lemma 5.6.1. *Let $S_1, \ldots, S_d \subseteq [n]$ be arbitrary subsets. Let* $\mathrm{frac}(S) := |S|/n$ *be normalized size of S. Suppose that*

$$\mathrm{frac}\left(\bigcup S_i\right) = \varepsilon \tag{5.3.4}$$

for some $\varepsilon < 1$ and for all i, j,

$$\mathrm{frac}(S_i \cap S_j) \leqslant \alpha \mathrm{frac}(S_i)\mathrm{frac}(S_j), \tag{5.3.5}$$

for some $\alpha > 1$ *s.t.* $\alpha\varepsilon < 1$. *Then,*

$$\sum_{i\in[d]} \mathrm{frac}(S_i) \leqslant \frac{\varepsilon}{1-\alpha\varepsilon}. \tag{5.3.6}$$

To prove Lemma 5.3.4, we will use the following technical result on almost pairwise independent Bernoulli random variables.

Lemma 5.6.2. *Let* $X = (X_1, X_2, \dots, X_d) \in \{0,1\}^d$ *be a random variable on* $\{0,1\}^d$. *If for any* $i, j \in [d]$,

$$\mathbb{E}\, X_i X_j \leqslant \varepsilon\, \mathbb{E}\, X_i\, \mathbb{E}\, X_j$$

where $0 \leqslant \varepsilon < 1$, *then*

$$\sum_{i=1}^{d} \mathbb{E}\, X_i \leqslant \frac{1}{1-\varepsilon}.$$

Proof. By Jensen's inequality, we have:

$$\begin{aligned}
\left(\sum_{i=1}^{d} \mathbb{E}\, X_i\right)^2 &= \left(\mathbb{E}\left(\sum_{i=1}^{d} X_i\right)\right)^2 \leqslant \mathbb{E}\left(\sum_{i=1}^{d} X_i\right)^2 \\
&= \sum_{i=1}^{d} \mathbb{E}\, X_i^2 + \sum_{i,j\in[d]} \mathbb{E}\, X_i X_j \\
&\leqslant \sum_{i=1}^{d} \mathbb{E}\, X_i + \varepsilon \sum_{i,j\in[d]} \mathbb{E}\, X_i\, \mathbb{E}\, X_j \leqslant \sum_{i=1}^{d} \mathbb{E}\, X_i + \varepsilon \left(\sum_{i=1}^{d} \mathbb{E}\, X_i\right)^2
\end{aligned}$$

Rearranging yields that $\sum_{i=1}^{d} \mathbb{E}\, X_i \leqslant \frac{1}{1-\varepsilon}$. □

We now proceed to the proof of Lemma 5.3.4.

Proof of Lemma 5.3.4. Let y be distributed uniformly on $\bigcup S_i$. Define a random variable

$X = (X_1, X_2, \ldots, X_d) \in \{0,1\}^d$ as a function of y as follows:

$$X_i = \begin{cases} 0 & \text{if } y \notin S_i \\ 1 & \text{if } y \in S_i \end{cases}$$

Then by (5.3.4), $\mathbb{E} X_i = \mathbb{P}(y \in S_i) = \frac{|S_i|}{|\bigcup S_i|} = \text{frac}(S_i)/\varepsilon$ and $\mathbb{E} X_i X_j = \mathbb{P}(y \in S_i \cap S_j) = \frac{|S_i \cap S_j|}{|\bigcup S_i|} = \text{frac}(S_i \cap S_j)/\varepsilon$.

By (5.3.5), we have

$$\mathbb{E} X_i X_j \leqslant \frac{\text{size}(S_i \cap S_j)}{\varepsilon} \leqslant \frac{\alpha}{\varepsilon}\text{size}(S_i)\text{size}(S_j) = \alpha\varepsilon \, \mathbb{E} X_i \, \mathbb{E} X_j.$$

By Lemma 5.6.2, this implies $\sum_{i=1}^d \mathbb{E} X_i \leqslant \frac{1}{1-\alpha\varepsilon}$. Using that

$$\sum \mathbb{E} X_i = \frac{1}{\varepsilon}\sum_{i=1}^d \text{frac}(S_i)$$

gives $\sum_{i=1}^d \text{frac}(S_i) \leqslant \frac{\varepsilon}{1-\alpha\varepsilon}$. □

5.7 Omitted Proofs

Proof of Lemma 5.2.1. We can project the density to the span of u, v to get a center-symmetric, isotropic, two-dimensional logconcave density $f : \mathbb{R}^2 \to \mathbb{R}_+$. Let us assume that u, v are orthogonal, the general case when they are at least at a constant angle will be similar.

Consider the level set L_1 of f of function value at least $1/100$. We claim that it intersects the line ℓ_1 defined by $u^T x = 1$ in a segment of length at least $1/10$. Let $\delta = \nu(H_u)$. Moreover, the line ℓ_2 defined by $u^T x = 1 + c\delta$ does not intersect L_1, else the measure in between the two lines is too large, and we know it is at most δ.

Now consider L_2, the level set of function value at least $1/1000$. Now we claim that

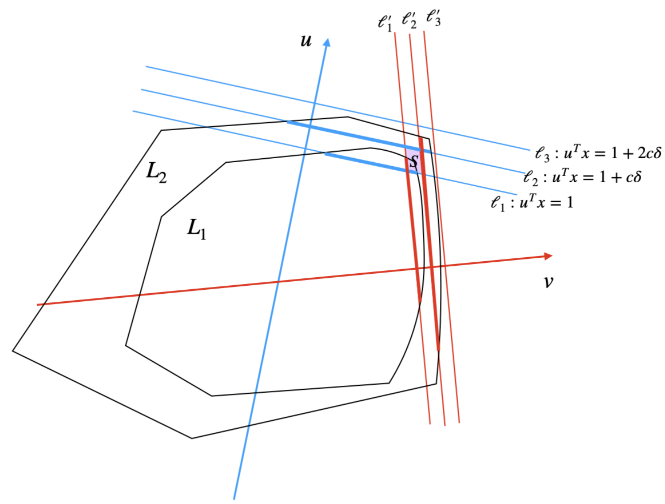

Figure 5.1: The intersection of halfspaces has measure $O(\delta^2)$.

the intersection of ℓ_2 with L_2 is of length most $1/10$. Moreover, the line ℓ_3 defined by $u^T x = 1 + 2c\delta$ does not intersect L_2.

The same bounds apply along v.

To bound the measure of $\{x : u^T x \geqslant 1, v^T x \geqslant 1\}$, we divide up the region $u^T x \geqslant 1$ into strips perpendicular to u so that the measure of the bounding lines decreases by a constant factor in each strip. Each strip has length $c\delta$ We do the same for v. So the intersection of the first two strips, a square S, has measure $O(\delta^2)$. Now, using the previous claims about level sets, it follows that along any line starting at the intersection of $u^T x = 1$ and $v^T x = 1$ and continuing in the region $u^T x \geqslant 1, v^T x \geqslant 1$, the value of f decreases by a constant factor every $c_2\delta$ distance along the line. From this it follows that the measure of the entire regions is $O(\delta^2)$. To see this we consider the polar integral of the region $H_u \cap H_v$,

$$\int_{\theta=0}^{\pi/2} \int_{r=0}^{\infty} r f(1 + r\cos\theta, 1 + r\sin\theta)\, dr\, d\theta \leqslant \int_{\theta=0}^{\pi/2} \int_{r=0}^{\infty} c_1 r e^{-r/(c_2\delta)}\, dr\, d\theta = O(\delta^2).$$

When u and v are at some constant angle (instead of orthogonal), then the intersection of bands induced by intervals becomes a parallelogram with area $O(\delta^2)$. The rest of the argument remains the same. $\square$

5.8 Method of Moments is Inadequate

In this section, we give an explicit example that shows learning the first t moments with the optimal error (with ε fraction of corruption) for $t = o(\log d)$ does not give a dimension-independent error guarantee. Specifically, we prove the following theorem:

Theorem 5.8.1. *Fix $t \in \mathbb{N}$, $d \in \mathbb{N}$ large enough and $\eta < t^t/d$. Let H be the uniform distribution on $[-1,1]^d$ and for any matrix B, let BH denote the distribution of $y = Bx$ where $x \sim H$. Then, there exists an orthonormal matrix $A \in \mathbb{R}^{d \times d}$ such that:*

1. *$d_{TV}(H, AH) \geqslant \eta d/t^t$, and,*

2. *For $1 \leqslant i \leqslant t$ and any $v \in \mathbb{R}^d$, $|\,\mathbb{E}_{x \sim H}[\langle x, v\rangle^t] - \mathbb{E}_{x \sim AH}[\langle x, v\rangle^t]| \leqslant \eta \left\|v\right\|_2^t$.*

In particular, for any $t = o(\log d/\log\log d)$, even though H and AH have inverse polynomially close first t moments, the total variation distance between H and AH is at least a constant.

We first show that the TV distance is lower bounded by the sum of ℓ_2-distance of rows of rotation matrix.

Lemma 5.8.2. *Let $H = [-1,1]^d$. Suppose $A, \widehat{A}$ are $d \times d$ matrices; A is orthonormal and $\widehat{A}$ is a matrix with unit length rows. If $\sum_i \left\|\hat{a}_{(i)} - a_{(i)}\right\|_2 \leqslant C$ for an absolute constant $C < 1$, then*

$$d_{TV}(\widehat{A}H, AH) \geqslant \frac{1}{10}\sum_i \left\|\hat{a}_{(i)} - a_{(i)}\right\|_2.$$

Proof. WLOG, we assume A is the identity matrix. Suppose $\left\|\hat{a}_{(i)} - e_i\right\|_2 = \delta_i$. Then by Lemma 5.4.5, for $x \sim H$,

$$\frac{\delta_i}{5} \leqslant \mathbb{P}(\hat{a}_{(i)} \cdot x > 1) \leqslant \frac{(1+\delta_i)\delta_i}{4}. \tag{5.8.1}$$

By Lemma 5.2.1, for some constant C_1,

$$\mathbb{P}(\hat{a}_{(i)} \cdot x > 1, \hat{a}_{(j)} \cdot x > 1) \leqslant C_1 \, \mathbb{P}(\hat{a}_{(i)} \cdot x > 1) \, \mathbb{P}(\hat{a}_{(j)} \cdot x > 1).$$

We apply Lemma 5.3.4 as the reverse union bound on the events $\hat{a}_{(i)} \cdot x > 1$ and get

$$\begin{aligned} \frac{\mathbb{P}(\bigcup_i (\hat{a}_{(i)} \cdot x > 1))}{1 - C_1 \, \mathbb{P}(\bigcup_i (\hat{a}_{(i)} \cdot x > 1))} &\geqslant \sum_i \mathbb{P}(\hat{a}_{(i)} \cdot x > 1) \\ \mathbb{P}\left(\bigcup_i (\hat{a}_{(i)} \cdot x > 1)\right) &\geqslant \frac{\sum_i \mathbb{P}(\hat{a}_{(i)} \cdot x > 1)}{1 + C_1 \sum_i \mathbb{P}(\hat{a}_{(i)} \cdot x > 1)} \end{aligned}$$

By the second inequality in (5.8.1), we have $\sum_i \mathbb{P}(\hat{a}_{(i)} \cdot x > 1) \leqslant \frac{1}{2} \sum_i \delta_i \leqslant C/2$. If we take $C \leqslant 2/C_1$, then $1 + C_1 \sum_i \mathbb{P}(\hat{a}_{(i)} \cdot x > 1) \leqslant 2$. Thus, by the definition of TV distance, we have

$$d_{TV}(\widehat{A}H, H) = \mathbb{P}\left(\bigcup_i (\hat{a}_{(i)} \cdot x > 1)\right) \geqslant \frac{1}{2} \sum_i \mathbb{P}(\hat{a}_{(i)} \cdot x > 1) \geqslant \frac{1}{10} \sum_i \delta_i.$$

□

Next, we define the matrix B that we will use as the rotation matrix in our bad example and show the following properties of B.

Lemma 5.8.3. *For any* $0 < \delta < 1$, *let*

$$B = \begin{pmatrix} 1-\delta & \sqrt{2\delta - \delta^2} & & & & & \\ -\sqrt{2\delta-\delta^2} & 1-\delta & & & & O & \\ & & 1-\delta & \sqrt{2\delta-\delta^2} & & & \\ & & -\sqrt{2\delta-\delta^2} & 1-\delta & & & \\ & & & & \ddots & \ddots & \\ & O & & & & 1-\delta & \sqrt{2\delta-\delta^2} \\ & & & & & -\sqrt{2\delta-\delta^2} & 1-\delta \end{pmatrix}$$

be a $d \times d$ matrix and $b_{(1)}, \ldots, b_{(d)}$ are rows of B, $b^{(1)}, \ldots, b^{(d)}$ are columns of B. Then

1. $BB^T = I$

2. $\sum_{i=1}^{d} \left\| b_{(i)} - e_i \right\|_2 = d\sqrt{2\delta}$.

Remark. We can replace δ in i-th row with δ_i. Then by the same argument, we have the almost same conclusions with $\sum \sqrt{\delta_i}$ replacing $d\sqrt{\delta}$.

Proof. We can see the rows of B form an orthonormal basis. So B is orthogonal. The second conclusion follows from the direct computation:

$$\sum_{i=1}^{d} \left\| b^{(i)} - e_i \right\|_2 = \sum_{i=1}^{d} \sqrt{\delta^2 + 2\delta - \delta^2} = d\sqrt{2\delta}.$$

$\square$

Suppose $H = [-\sqrt{3}, \sqrt{3}]^d$ is the standard cube in $\mathbb{R}^d$. Now, we claim that AH is a bad estimation of H s.t. the first t moments are close in operator norm for $t = o(\log d)$ but the TV distance is large.

Lemma 5.8.4. *There exist two matrices $A, B \in \mathbb{R}^{d \times d}$ s.t. for any integer $t < \sqrt{d}$, the relative distance in operator norm between the t'th moment of the corresponding distributions AH, BH is at most ε while the TV distance between the distributions is at least $\varepsilon d/t^t$.*

Proof. Let $A = I$ and B be the matrix in Lemma 5.8.3. We first show that the $2t$'th moment of AH and BH is close by induction on dimension. The first conclusion in Lemma 5.8.3 shows that BH is isotropic. So the first two moments of AH and BH are the same. The base case is when

$$B = \begin{pmatrix} 1-\delta & \sqrt{2\delta - \delta^2} \\ -\sqrt{2\delta - \delta^2} & 1-\delta \end{pmatrix}$$

Let M_t be the t'th moment of BH. Then for any $u \in \mathbb{R}^d$, we have

$$\langle M_t, u^{\otimes t} \rangle = \mathop{\mathbb{E}}_{x \sim H} (Bx \cdot u)^t = \mathop{\mathbb{E}}_{x_1, x_2 \sim \mathcal{U}_{[-1,1]}} \left(\langle b^{(1)}, u \rangle x_1 + \langle b^{(2)}, u \rangle x_2 \right)^t$$

$$
\begin{aligned}
&= \mathbb{E}\left(\sum_{s=0}^{t}\binom{t}{s}\langle b^{(1)},u\rangle^{s}\langle b^{(2)},u\rangle^{t-s}x_1^s x_2^{t-s}\right) \\
&= \sum_{s=0}^{t}\binom{t}{s}\langle b^{(1)},u\rangle^{s}\langle b^{(2)},u\rangle^{t-s}\,\mathbb{E}(x_1^s)\,\mathbb{E}(x_2^{t-s}) \\
&\leqslant \frac{c^t}{t}\left(\langle b^{(1)},u\rangle + \langle b^{(2)},u\rangle\right)^t .
\end{aligned}
$$

The last equality follows from the independence between x_1 and x_2. The last inequality follows from that Each of $\mathbb{E}(x_i^s)$ is s'th moment of 1-D uniform distribution $[-\sqrt{3},\sqrt{3}]$, which is $\frac{c^s}{s+1}$, for $c=\sqrt{3}$. Then the operator distance between $M_t(H)$ and $M_t(BH)$ is

$$
\begin{aligned}
\max_{\|u\|=1}|\langle M_t(H)-M_t(BH),u^{\otimes t}\rangle| &\leqslant \frac{c^t}{t}\max_{\|u\|=1}\left|\left(\langle b^{(1)},u\rangle + \langle b^{(2)},u\rangle\right)^t-(u_1+u_2)^t\right| \\
&\leqslant \frac{c^t}{t}\max_{\|u\|=1}\left|\left((1-\delta)(u_1+u_2)+\sqrt{2\delta-\delta^2}(u_1-u_2)\right)^t-(u_1+u_2)^t\right| \\
&\leqslant \frac{c^t}{t}\left|\left(2(1-\delta)+\sqrt{2\delta-\delta^2}\right)^t-2^t\right| \\
&\leqslant 10\cdot(2c)^t\sqrt{\delta}\leqslant t!\sqrt{\delta}.
\end{aligned}
$$

Now we assume the operator distance between t'th moment of H and B_dH is at most $t!\sqrt{\delta}$ in $\mathbb{R}^d$. Suppose $X=x\cdot u_1,\widetilde{X}=B_dx\cdot u_1$ where u_1 is an arbitrary unit vector in $\mathbb{R}^d$ and $x\sim H_d$, $Y=x\cdot u_2,\widetilde{Y}=B_2x\cdot u_2$ where u_2 is an arbitrary unit vector in $\mathbb{R}^d$ and $x\sim H_d$. The normalized direct sum of $\widetilde{X}$ and $\widetilde{Y}$ is a uniform random vector from $B_{2d}H$ since the first d dimensions and the last two dimensions of $B_{2d}H$ are independent. Thus we can compute the $2t$'th moment of H_{2d} and $B_{2d}H_{2d}$ as follows:

$$
\mathbb{E}\left(aX+\sqrt{1-a^2}Y\right)^{2t}=\sum_{i=0}^{t}\binom{2t}{2i}a^{2i}(1-a^2)^{t-1}\,\mathbb{E}\,X^{2i}\,\mathbb{E}\,Y^{2(t-i)}
$$

The difference between them is

$$
\mathbb{E}\left(aX+\sqrt{1-a^2}Y\right)^{2t}-\mathbb{E}\left(a\widetilde{X}+\sqrt{1-a^2}\widetilde{Y}\right)^{2t}
$$

$$= \sum_{i=0}^{t} \binom{2t}{2i} a^{2i}(1-a^2)^{t-i} \left(\mathbb{E} X^{2i} (\mathbb{E} Y^{2(t-i)} - \mathbb{E} \widetilde{Y}^{2(t-i)}) + (\mathbb{E} X^{2i} - \mathbb{E} \widetilde{X}^{2i}) \mathbb{E} \widetilde{Y}^{2(t-i)} \right)$$

By induction hypothesis, $|\mathbb{E} Y^{2(t-i)} - \mathbb{E} \widetilde{Y}^{2(t-i)}| \leqslant \sqrt{\delta}(2(t-i))!$, $|\mathbb{E} X^{2i} - \mathbb{E} \widetilde{X}^{2i}| \leqslant \sqrt{\delta}(2i)!$. $\mathbb{E} X^{2i}$ and $\mathbb{E} \widetilde{Y}^{2(t-i)}$ are moments (in some direction) of the standard hypercube and hence are upper bounded by $(2i-1)!! \leqslant (2i)!$ and $(2(t-i)-1)!! \leqslant (2(t-i))!$. Then we have

$$\begin{aligned} \mathbb{E}\left(aX + \sqrt{1-a^2}Y\right)^{2t} - \mathbb{E}\left(a\widetilde{X} + \sqrt{1-a^2}\widetilde{Y}\right)^{2t} &\leqslant 2\sqrt{\delta} \sum_{i=0}^{t} \binom{2t}{2i} a^{2i}(1-a^2)^{t-i}(2i)!(2(t-i))! \\ &= 2\sqrt{\delta} \sum_{i=0}^{t} (2t)! a^{2i}(1-a^2)^{t-i} \\ &\leqslant 2\sqrt{\delta}(2t)! \frac{t}{2^t} \\ &\leqslant \sqrt{\delta}(2t)!. \end{aligned}$$

The third inequality follows from $\sum_{i=0}^{t} a^{2i}(1-a^2)^{t-i}$ is maximized when $a^2 = 1/2$. Thus we prove that for all d, the operator distance between $2t$'th moments is at most $\sqrt{\delta}(2t)!$.

By Lemma 5.8.2 and the second conclusion in Lemma 5.8.3, we get a lower bound on the TV distance:

$$d_{TV}(BH, H) \geqslant \frac{1}{10} d\sqrt{2\delta}.$$

Let $\varepsilon = \sqrt{\delta} t!$. Then

$$d_{TV}(BH, H) \geqslant \frac{\sqrt{2} d\varepsilon}{10 t!} \geqslant \frac{\varepsilon d}{t^t}.$$

□

Corollary 5.8.5. *Suppose $H, \widehat{H}$ are two transformed cubes in $\mathbb{R}^d$. If the first t moments of $H, \widehat{H}$ are ε-close in Frobenius norm for any $t = o(\log d / \log\log d)$, the best guarantee on their TV distance has to have a dependence on the dimension d.*

CHAPTER 6

ISOTROPIC TRANSFORMATIONS FOR CONVEX BODIES

This chapter is joint work with Aditi Laddha, Yin Tat Lee and Santosh Vempala, and appears as [109].

6.1 Main Results

We give a new algorithm for the isotropic transformation of any given convex body. The complexity of the algorithm is $O^*(n^3\psi^2)$, directly determined by the best possible KLS constant ψ for isotropic logconcave densities in $\mathbb{R}^n$. This is the first direct reduction of the complexity of volume computation to the KLS constant. With the current bound of $\psi = O(n^{1/4})$, the complexity is $O^*(n^{3.5})$ and implies volume computation in the same complexity for any convex body. If the conjecture turns out to be true, then the complexity becomes $O^*(n^3)$. No further improvement is in sight for the general setting.

Theorem 6.1.1 (Rounding)**.** *There is a randomized algorithm that takes as input a convex body $K \subset \mathbb{R}^n$ given by a membership oracle with initial point $x_0 \in K$, bounds r, R>0 s.t.,*

$$x_0 + rB_n \subseteq K \subseteq RB_n$$

and with probability $1-\delta$, computes an affine transformation T s.t., TK is in near-isotropic position, i.e., for x sampled uniform from TK,

$$I \preccurlyeq \mathrm{Cov}(xx^\top) \preccurlyeq 2I.$$

The algorithm takes $O(n^3\psi^2\log^c(Rn/r)\log(1/\delta))$ oracle queries and $O(n^2)$ time per oracle query.

6.1.1 Approach

The algorithm is based on the following ideas. First, if the covariance matrix is skewed (i.e., some eigenvalues are much larger than others), this can be detected via a small random sample. The roundness can then be improved by scaling up the subspace of small eigenvalues. But how to sample a highly skewed convex body? To break this chicken-and-egg problem, we use a sequence of convex bodies obtained by intersecting the original set with balls of increasingly larger radii:

$$K_t = K \cap B(0, t),$$

with t starting at r and doubling till it reaches R. In each outer iteration, we compute a transformation that makes the set K_t (nearly) isotropic. When K_t is isotropic, we show that K_{2t} is well-rounded (Lemma 6.3.6). Hence, the trace of the covariance of K_{2t} is $O(n)$. However, its eigenvalues could be widely varying. To make K_{2t} isotropic, we will estimate the larger eigenvalues using a small random sample (Lemma 6.2.11). The second idea is that scaling up the small eigenvalues nearly doubles the size of the ball contained inside (Lemma 6.3.4), while having a mild effect on the higher norms of the covariance (Lemma 6.3.3). The latter concept is where KLS comes in. We show that if the KLS constant for isotropic logconcave distributions in $\mathbb{R}^n$ is bounded by $\psi_n^2 \lesssim n^{\frac{1}{p}}$ for all n, then for any logconcave density q with covariance matrix A (not necessarily identity), the KLS constant is bounded as $\psi_q^2 \lesssim \|A\|_p \log n$ (Theorem 6.2.8). As we outline below, this improves the sampling time in each inner iteration.

In the beginning, when the ball contained inside K_t is small, we will use a few samples to get a coarse estimate of the larger eigenvalue directions and scale up the orthogonal subspace. The sampling time is higher since the roundness parameter of K_t is higher. The time per sample is roughly $n^2 \frac{\|A\|_p}{r^2}$ when r is the radius of the ball inside and A is the covariance matrix for the uniform density q on K_t. As we increase r, by a constant factor in each step, the norm of the covariance grows much more slowly, and so the sampling time decreases. Meanwhile, we use more samples in each step, roughly r^2, and this trade-off keeps the overall time at $n^{3+\frac{1}{p}} \simeq n^3\psi^2$. During the process we need a warm start for each phase; we achieve this in $\widetilde{O}(n^3)$ steps using the Gaussian Cooling algorithm. It is known that $\psi_q^2 \lesssim \|A\|_p$ holds for $p = 2$ [64], giving the $\widetilde{O}(n^{3.5})$ complexity for rounding. Our general reduction establishes that any future improvement to higher p would imply faster rounding and volume estimation.

6.1.2 KLS Conjecture

For an n-dimensional measure ν, its KLS constant (the isoperimetric or Cheeger constant is the reciprocal) is

$$\frac{1}{\psi_\nu} = \inf_{A\subset\mathbb{R}:Q(A)\leqslant\frac{1}{2}} \frac{\nu_{n-1}(\partial A)}{\nu(A)}$$

where ∂A is the boundary of the set A, and ν_{n-1} is the induced $(n-1)$-dimensional measure. The conductance of the Markov chain reduces to the isoperimetry of its stationary density[1]:

$$\phi \gtrsim \frac{1}{\psi \cdot n}$$

with ψ being the KLS constant of the stationary density of the Markov chain. This implies a mixing rate of $O(n^2\psi^2)$. Thus, bounding the KLS constant becomes critical, and this consideration originally motivated the KLS conjecture. With many unexpected connections and applications since its formulation, the conjecture has become a central part of asymptotic convex geometry and functional analysis. (A distribution is said to be isotropic if it has zero mean and identity covariance, see Definition 2.1.2).

Conjecture 6.1.2 (KLS Conjecture [110])**.** *The KLS constant of any isotropic logconcave density in any dimension is bounded by an absolute constant. Equivalently, for a logconcave density q with covariance matrix A, we have $\psi_q \lesssim \|A\|_{\mathrm{op}}^{1/2}$.*

Chen [111] proved a sub-polynomial upper bound for the KLS constant.

[1] We use $\gtrsim$ to denote "LHS greater than a constant factor times RHS".

Theorem 6.1.3 ([111]). *There exists a universal constant c such that for any logconcave density p with covariance A in $\mathbb{R}^n$, we have $\psi_p \leqslant n^{c\left(\frac{\log\log n}{\log n}\right)^{1/2}} \cdot \|A\|_{\mathrm{op}}^{1/2}$. For isotropic logconcave p, we have $\psi_p \lesssim n^{o(1)}$.*

This was further improved to a polylogarithmic bound by Klartag and Lehec [112], and their technique was refined slightly in [113].

Theorem 6.1.4 ([112, 113]). *For any isotropic logconcave density p, we have $\psi_p \lesssim (\log n)^{3.3}$.*

6.2 Preliminaries

For a positive definite matrix $n \times n$ matrix A, we will use the operator or spectral norm, $\|A\|_{\mathrm{op}}$, its Frobenius norm, $\|A\|_F \stackrel{\text{def}}{=} \mathrm{tr}(A^2)^{1/2}$, and its $p'th$ norm for $p \geqslant 1$,defined as

$$\|A\|_p \stackrel{\text{def}}{=} \left(\sum_{i=1}^{n} |\lambda_i|^p\right)^{1/p}$$

with $p = 1$ being $\mathrm{tr}A$ and $p = 2$ being $\|A\|_F$.

We say that a convex body K is (r, R)-*rounded* if it contains a ball of radius r, and its covariance matrix A satisfies

$$\mathrm{tr}A = \mathop{\mathbb{E}}_K(\|x - \bar{x}\|^2) = R^2.$$

For an isotropic body we have $A = I$ and hence $R^2 = n$. We will say K is well-rounded if $r = 1$ and $R = \widetilde{O}(\sqrt{n})$. Clearly, any isotropic convex body is also well-rounded, but not vice versa. A distribution is said to be α-isotropic if $I \preceq \mathbb{E}_D(xx^\top) \preceq \alpha I$.

Random points from isotropic logconcave densities have strong concentration properties. We mention three that we will use.

Lemma 6.2.1 ([106, Theorem 5.17]). *For any $t \geqslant 1$, and any logconcave density p in $\mathbb{R}^n$*

with covariance matrix A,

$$\mathop{\mathbb{P}}_{x\sim p}\left(\|x - \mathbb{E}\, x\|_2 \geqslant t \cdot \sqrt{\mathrm{tr} A}\right) \leqslant \exp(-t+1).$$

Lemma 6.2.2 ([114, Theorem 5.17])**.** *For any $t \geqslant 1$, and any isotropic logconcave density p in $\mathbb{R}^n$, there is a constant c such that*

$$\mathop{\mathbb{P}}_{x\sim p}(\|x\|_2 \geqslant t\sqrt{n}) \leqslant \exp(-c\sqrt{n}t).$$

Lemma 6.2.3 ([115, 116])**.** *For any isotropic logconcave distribution ν in $\mathbb{R}^n$,with probability at least $1-\delta$, the empirical mean and covariance*

$$\overline{x} = (1/N)\sum_{i=1}^{N} x_i, \quad \textit{and} \quad X = \frac{1}{N}\sum_{i=}^{N}(x_i - \overline{x}_i)(x_i - \overline{x}_i)^T$$

of $N = O(n \log(1/\delta)/\varepsilon^2)$ i.i.d. random samples $x_i \sim \nu$ satisfy

$$\|\overline{x}\|_2 \leqslant \varepsilon, \quad \|X - I\|_{\mathrm{op}} \leqslant \varepsilon.$$

The convergence of Markov chains is established by showing the the t-th step distribution Q_t approaches the steady state distribution Q. We will use the total variation distance d_{TV} for this. We also need a notion of a warm start.

Definition 6.2.4 (Warm start)**.** We say that a starting distribution Q_0 is M-warm for a Markov chain with unique stationary distribution Q if its χ-squared distance is bounded by M:

$$\chi^2(Q_0, Q) = \mathop{\mathbb{E}}_{Q}\left(\frac{dQ_0(u)}{dQ(u)} - 1\right)^2 \leqslant M.$$

Note that $\chi^2(Q_0, Q) + 1 = \mathbb{E}_Q\left(\frac{dQ_0(u)}{dQ(u)}\right)^2 = \mathbb{E}_{Q_0}\frac{dQ_0(u)}{dQ(u)}$.

Our algorithms will use the ball walk for sampling. In a convex body K, the *ball walk* with step-size δ is defined as follows: At the current point, $x \in K$, pick y uniformly in $B(x, \delta)$, the ball of radius δ around x; if $y \in K$, go to y (else, do nothing).

The next theorem is a fast sampler for distributions with KLS constant ψ_q given a warm start.

Theorem 6.2.5 ([117]). *Let K be a convex body containing the unit ball. Using the ball walk with step size $\frac{1}{\sqrt{n}}$ in K from an M-warm start Q_0, the number of steps to generate a nearly independent point within TV distance or χ^2-distance ε of the uniform stationary density Q in $\mathbb{R}^n$ is $O(n^2\psi_q^2 \log^{O(1)}(nM/\varepsilon))$.*

The next lemma connects the KLS constant for isotropic distributions to that of general distributions. We give a proof in Section (6.1.2).

Lemma 6.2.6. *Let ψ_n be a bound on the KLS constant for isotropic logconcave densities and ψ_q be the bound for a logconcave density q (not necessarily isotropic) with covariance matrix A_q, both in $\mathbb{R}^n$.. If $\psi_n^2 = \Theta\left(n^{1/p}\right)$ for all n, then for any logconcave density q with covariance A, we have $\psi_q^2 = O(\|A\|_p \log^{O(1)} n)$.*

Lemma 6.2.7 ([118, Lemma 35]). *Let Φ_t be a stochastic process such that $\Phi_0 \leqslant \frac{U}{2}$ and $d\Phi_t = \delta_t dt + v_t^\top dW_t$. Let $T > 0$ be some fixed time, $U > 0$ be some target upper bound, and f and g be some auxiliary functions such that for all $0 \leqslant t \leqslant T$*

1. *$\delta_t \leqslant f(\Phi_t)$ and $\|v_t\|_2 \leqslant g(\Phi_t)$,*

2. *Both $f(\cdot)$ and $g(\cdot)$ are non-negative non-decreasing functions,*

3. *$f(U) \cdot T \leqslant \frac{U}{8}$ and $g(U) \cdot \sqrt{T} \leqslant \frac{U}{8}$.*

Then, we have that $\mathbb{P}\left[\max_{t\in[0,T]} \Phi_t \geqslant U\right] \leqslant 0.01$.

Using this lemma, we obtain the following bound.

Theorem 6.2.8. *Suppose that $\psi_n = O(n^\beta)$ for all n with some fixed $0 \leqslant \beta \leqslant \frac{1}{2}$. Then, for any logconcave distribution p with covariance matrix A, we have that*

$$\psi_p \lesssim \frac{\sqrt{\log n}}{\beta} \|A\|_{\frac{1}{2\beta}}^{1/2}.$$

We will also use a fast sampler for well-rounded convex bodies that does not require a warm start.

Theorem 6.2.9 (Sampling a well-rounded convex body [119, 120])**.** *There is an algorithm that, for any $\varepsilon > 0, p > 0$, and any convex body K in $\mathbb{R}^n$ that contains the unit ball and has $\mathbb{E}_K(\|X\|^2) = R^2$, with probability $1-\delta$, generates random points from a density ν that is within total variation distance ε from the uniform distribution on K. In the membership oracle model, the complexity of each random point, including the first, is,*

$$O\left((R^2n^2 + n^3)\log n \log^2 \frac{n}{\varepsilon} \log \frac{1}{\delta}\right) = \widetilde{O}(n^3 + R^2n^2).$$

Finally, we use the algorithm from [120, 119] for computing the volume of a well-rounded convex body.

Theorem 6.2.10 (Volume of a well-rounded convex body)**.** *[119, 120] There is an algorithm that, for any $\varepsilon > 0, \delta > 0$ and convex body K in $\mathbb{R}^n$ that contains the unit ball and has $\mathbb{E}_K(\|X\|^2) = O(n)$, with probability $1-\delta$, approximates the volume of K to within relative error ε and has complexity*

$$O\left(\frac{n^3}{\varepsilon^2} \cdot \log^2 n \log^2 \frac{1}{\varepsilon} \log^2 \frac{n}{\varepsilon} \log \frac{1}{\delta}\right) = \widetilde{O}(n^3\varepsilon^{-2})$$

in the membership oracle model.

Lemma 6.2.11. *Let p be a logconcave density in $\mathbb{R}^n$ with covariance A. Let $\widehat{A} \leftarrow \frac{1}{k}\sum_{i=1}^k (x_i - \widehat{x})(x_i - \widehat{x})^\top$ where $\widehat{x} = \frac{1}{k}\sum_{i=1}^k x_{i+k}$ and x_i are independent samples from p.*

With probability $1 - 1/n^{O(1)}$*, for any* $0 \leqslant \varepsilon \leqslant 1$*, we have*

$$(1-\varepsilon)A - O\left(\frac{\log^3 n \mathrm{tr} A}{\varepsilon k}\right) \cdot I \preceq \widehat{A} \preceq (1+\varepsilon)A + O\left(\frac{\log^3 n \cdot \mathrm{tr} A}{\varepsilon k}\right) \cdot I.$$

6.3 Algorithm and Analysis

The algorithm considers a sequence of balls of doubling radii, and in each iteration makes the intersection of the current ball with the convex body nearly isotropic.

Algorithm 6.3.1 (Iterative Isotropization)**.**

procedure IterativeIsotropization*(*$K \subset \mathbb{R}^n, r > 0, R > 0$*)*
  Assumption: $B(0,r) \subseteq K \subseteq B(0,R)$
  Define $K_t = K \cap B(0,t)$..
  $t \leftarrow r, x \leftarrow 0, T \leftarrow \frac{1}{r} I$ ▷ *Current slice* t *and normalization*
  while $t < R$ ***do***
    $(x,T) \leftarrow$ Isotropization$(K_t - x, T)$
    $t \rightarrow 2t$
  end while
end procedure

The main subroutine, Isotropize, is a procedure to compute an isotropic transformation of a well-rounded body. Given a convex body K and a transformation T such that TK is well-rounded, Isotropize outputs a point x and a transformation T' such that $T'(K-x)$ is 2-isotropic. We address this in the next section, then come back to the general analysis.

6.3.1 Inner loop: Isotropic transformation of a well-rounded body

We begin with the algorithm. In each iteration, the inner ball radius grows by a constant factor, while the p-norm of the covariance grows much more slowly. As a result, we can afford to sample more points with each iteration and thereby get progressively better approximations to the isotropy: in the first step, the number of samples is only $\widetilde{O}(1)$ while by the end the number of samples is $\widetilde{O}(n)$.

Algorithm 6.3.2 (Isotropize)**.**

1: ***procedure*** Isotropize*($K \subset \mathbb{R}^n$, $T_0 \in \mathbb{R}^{n\times n}$)*
2: ***Assumption:*** *$B(0, \frac{1}{4}) \subset T_0K$ and $\mathbb{E}_{x\sim T_0K}\|x\|^2 = \widetilde{O}(n)$.*
3: *Use Gaussian cooling to sample a point $x \in T_0K$.*
4: *$r \leftarrow \frac{1}{4}, T \leftarrow T_0$.* ▷ *We maintain $\{x : \|x\|_2 \leqslant r\} \subseteq TK$.*
5: ***while*** $2^{10}r^2\log^2 n \leqslant n$ ***do***
6: *$k \leftarrow c \cdot r^2 \log^7 n$ for a constant c.*
7: *Sample points $x_1, \cdots, x_{2k}$ from TK using the ball walk with initial point x.*
8: *$\widehat{A} \leftarrow \frac{1}{k}\sum_{i=1}^{k}(x_i - \widehat{x})(x_i - \widehat{x})^\top$ where $\widehat{x} = \frac{1}{k}\sum_{i=1}^{k} x_{i+k}$.* ▷ *Estimate covariance.*
9: *$M \leftarrow I_n + P_{\widehat{A}}$ where $P_{\widehat{A}}$ is the projection to the subspace spanned by eigenvectors of $\widehat{A}$ with eigenvalue at most $\lambda = n$.*
10: *$T \leftarrow MT$, $r \leftarrow 2(1 - \frac{1}{\log n}) \cdot r$.*
11: ***end while***
12: *Sample $O(n\log^2 n)$ points to compute the mean $\widehat{x}$ and empirical covariance matrix $\widehat{A}$ such that $\widehat{A}^{-\frac{1}{2}}(K - \widehat{x})$ is 2-isotropic.*
13: ***Return*** $(\widehat{x}, \widehat{A}^{-\frac{1}{2}})$.
14: ***end procedure***

First, we show that the p-th norm of the covariance matrix $\|A\|_p$ remains bounded. Although we only use an eigenvalue threshold of $\lambda = n$ in the algorithm, we will prove a slightly more general statement below, assuming the eigenvalue threshold is $\lambda = n/r^\alpha$.

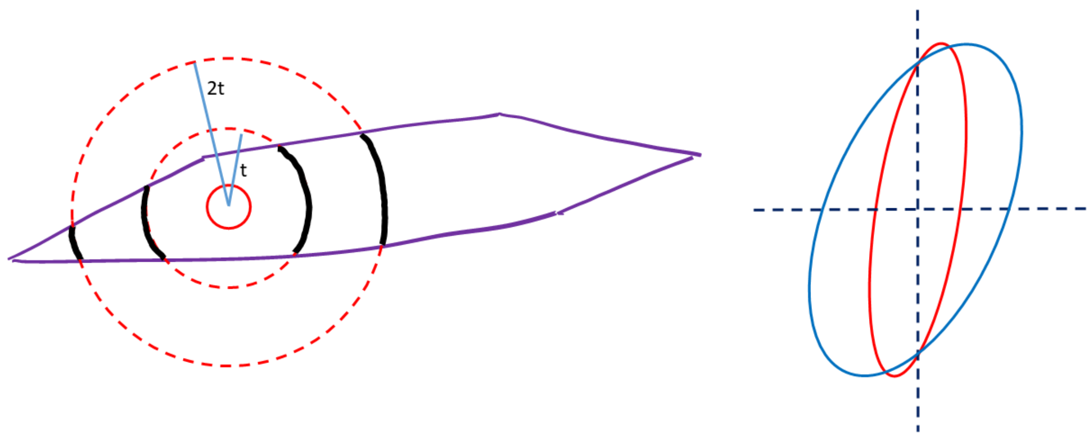

Figure 6.1: (a) Algorithm IterativeIsotropization uses balls of doubling radii (b) Algorithm Isotropize scales up the estimated "small eigenvalue" subspace in each iteration.

Lemma 6.3.3. *Let A_j be the covariance matrix of TK and r_j be the inner radius at the j-th iteration of algorithm* Isotropize. *For $\alpha \in [0, 2/(p-1)]$, we have*

- *the number of iterations is $O(\log n)$.*
- $\mathrm{tr} A_j = O(r_j^2 \mathrm{tr} A_0)$ *for all j.*
- $\|A_j\|_p \leqslant 24 n r_j^{(\frac{2+\alpha}{p} - \alpha)} \log^{1+\frac{2}{p}} n$ *for all j.*

Proof. In each iteration, the algorithm increases r_j by a factor of $2(1 - \frac{1}{\log n})$. Note that the algorithm starts with $r_0 = \frac{1}{4}$ and ends before $r_j \log n \leqslant \sqrt{n}$. Hence, it takes less than $O(\log n)$ iterations.

To bound $\mathrm{tr} A_j$, we let P_j be the projection matrix at j-th iteration Then the transformation at j-th iteration is $M_j = I + P_j$. We have

$$A_j = M_j A_{j-1} M_j \tag{6.3.1}$$

Since P_j is a projection matrix, we have

$$\mathrm{tr} A_j = \mathrm{tr}\left(A_{j-1}^{\frac{1}{2}} M_j^2 A_{j-1}^{\frac{1}{2}}\right) = \mathrm{tr}\left(A_{j-1}^{1/2}(I+3P_j)A_{j-1}^{1/2}\right) \leqslant 4\mathrm{tr} A_{j-1}.$$

Hence, $\mathrm{tr}A_j$ increases by a factor of at most 4. Moreover, since r_j increases by a factor of $2(1-\frac{1}{\log n})$ we have that $\frac{\mathrm{tr}A_j}{r_j^2}$ increases by a factor of $(1-\frac{1}{\log n})^{-2}$ per iteration. Since there are $O(\log n)$ iterations, it increases by at most $O(1)$ in total. Hence, we have $\mathrm{tr}A_j = O(r_j^2 \mathrm{tr} A_0) = O(r_j^2 n \log^2 n)$ for all j.

To bound $\|A_j\|_p$, we note that initially $\|A_0\|_p \leqslant \mathrm{tr}A_0 \leqslant 4n$. By (6.3.1), we have

$$\begin{aligned}
\|A_j\|_p &= \|M_j A_{j-1} M_j\|_p \\
&= \|A_{j-1}^{1/2} M_j^2 A_{j-1}^{1/2}\|_p \\
&= \|A_{j-1}^{1/2}(I+3P_j)A_{j-1}^{1/2}\|_p.
\end{aligned}$$

Hence, we have

$$\|A_j\|_p \leqslant \|A_{j-1}\|_p + 3\|A_{j-1}^{1/2} P_j A_{j-1}^{1/2}\|_p.$$

Let $\lambda \stackrel{\mathrm{def}}{=} \frac{n}{r_j^\alpha}$. Since P_j is the projection of the eigenspace of $\widehat{A}$ with eigenvalues less than λ, we have that

$$P_j \preceq 2\lambda(\widehat{A} + \lambda \cdot I)^{-1}.$$

(To see this, we note that both the sides have the same eigenspace and hence it follows from $1_{x\leqslant\lambda} \leqslant \frac{2\lambda}{x+\lambda}$.) By Lemma 6.2.11 with $\varepsilon = \frac{1}{2}$, we have that $\frac{1}{2}A_j - O(\frac{\log^3 n \cdot \mathrm{tr}A_j}{k}) \cdot I \preceq \widehat{A}$. Using $\mathrm{tr}A_j = O(r_j^2 n \log^2 n)$ and $k = \Omega(r_j^{2+\alpha}\log^5 n)$ samples, we have

$$A_{j-1} \preceq 2\widehat{A} + \frac{n}{r_j^\alpha} \cdot I \preceq 2\widehat{A} + \lambda \cdot I.$$

Hence, we have $P_j \preceq 4\lambda(A_{j-1} + \lambda \cdot I)^{-1}$ and

$$\|A_j\|_p \leqslant \|A_{j-1}\|_p + 12\lambda \|A_{j-1}^{1/2}(A_{j-1} + \lambda \cdot I)^{-1} A_{j-1}^{1/2}\|_p.$$

Let λ_i be the eigenvalues of A_{j-1}. Then,

$$\|A_{j-1}^{1/2}(A_{j-1} + \lambda \cdot I)^{-1} A_{j-1}^{1/2}\|_p^p = \sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda_i + \lambda}\right)^p \leqslant \frac{1}{\lambda^p} \sum_{\lambda_i \leqslant \lambda} \lambda_i^p + \sum_{\lambda_i \geqslant \lambda} 1.$$

The first term $\sum_{\lambda_i \leqslant \lambda} \lambda_i^p$ is maximized when the small eigenvalues are exactly $\lambda_i = \lambda$. In this case, there are

$$\frac{\mathrm{tr} A_{j-1}}{\lambda} \leqslant \frac{O(r_{j-1}^2 n \log^2 n)}{\lambda}$$

small eigenvalues, i.e., of value at most λ. Hence, we have

$$\frac{1}{\lambda^p} \sum_{\lambda_i \leqslant \lambda} \lambda_i^p \leqslant \frac{O(r_{j-1}^2 n \log^2 n)}{\lambda}.$$

For the second term,

$$\sum_{\lambda_i \geqslant \lambda} 1 \leqslant \frac{\mathrm{tr} A_{j-1}}{\lambda} \leqslant \frac{O(r_{j-1}^2 n \log^2 n)}{\lambda}.$$

Hence, $\|A_j\|_p$ is increased by

$$O(\lambda) \cdot \left(\frac{r_{j-1}^2 n \log^2 n}{\lambda}\right)^{1/p} = O\left(\frac{n}{r_j^\alpha}\right) \cdot \left(r_{j-1}^2 r_j^\alpha \log^2 n\right)^{1/p} \leqslant O\left(n r_j^{\frac{2+\alpha}{p} - \alpha} \log^{2/p} n\right).$$

Since there are $O(\log n)$ iterations, $\|A_j\|_p$ is at most $O(n r_j^{(\frac{2+\alpha}{p} - \alpha)} \log^{1+\frac{2}{p}} n)$. $\square$

Next, we show that the inner radius increases by almost a factor of 2 every iteration, again for a general choice of eigenvalue threshold $\lambda = n/r^\alpha$.

Lemma 6.3.4. *Algorithm* Isotropize *maintains the invariant* $\{\|x\|_2 \leqslant r\} \subseteq TK$.

Proof. First, we describe the idea of the proof. When the algorithm modifies the covariance matrix A, it estimates the subspace of directions with variance less than n/r^{α} and doubles them. Lemma 6.2.11 with $\varepsilon = 1/2$ shows that

$$\widehat{A} = (1 \pm \frac{1}{2})A \pm \frac{n}{4r^{\alpha}} \cdot I.$$

This means, roughly speaking, that any direction that was not doubled satisfies

$$\lambda_i \geqslant \sigma^2 = \frac{n}{2r^{\alpha}}.$$

So the current body contains an ellipsoid whose axis lengths along the non-doubled directions (call this subspace V) are at least σ and at least r in all directions (the body contains a ball of radius r). Now the body is stretched by a factor of 2 in the subspace $V^{\perp}$. We will argue that the resulting body contains a ball of radius nearly $2r$. To see this, we argue that any point in a ball of this radius is in the convex hull of the ball of radius σ in the subspace V and a ball of radius $2r$ in the subspace $V^{\perp}$.

Now we proceed to the formal proof by induction. Initially, we have $\{\|x\|_2 \leqslant r\} \subset TK$ by the assumption on K. It suffices to show this is maintained after each iteration. Let T be the old transformation and T' be the new transformation during one iteration. Let r be the old radius during that iteration. By the invariant, we know that $\{\|x\|_2 \leqslant r\} \subset TK$ and hence

$$\{\|M^{-1}x\|_2 \leqslant r\} \subset T'K$$

with $M = I + P_{\widehat{A}}$ (this only scales up the body by a factor of 2 in some directions). Let A be the covariance matrix of TK. Then $(I + P_{\widehat{A}})A(I + P_{\widehat{A}})$ is the covariance matrix of $T'K$. Our goal is to show that $T'K$ contains a ball of radius $2(1 - \frac{1}{\log n})r$. From above, we have

$$\{x^{\top}A^{-1}x \leqslant 1\} \subset \{x^{\top}((I + P_{\widehat{A}})A(I + P_{\widehat{A}}))^{-1}x \leqslant 1\} \subset T'K.$$

Hence, we know that $\mathrm{Conv}(\Omega_1 \cup \Omega_2) \subset T'K$ where

$$\Omega_1 = \{x^\top(\frac{1}{r^2}I - \frac{3}{4r^2}P_{\widehat{A}})x \leqslant 1\} \text{ and } \Omega_2 = \{x^\top A^{-1}x \leqslant 1\}.$$

where for Ω_1 we used that $(I+P)^{-2} = I - \frac{3}{4}P$ for any projection matrix P.

To prove that $T'K$ contains a larger ball, we take x and write it as a convex combination of x_1 and x_2 where $x_1 \in \Omega_1$ and $x_2 \in \Omega_2$. To do this, we define P_A be the projection to the subspace spans by eigenvectors of A with eigenvalues at most $\lambda/2^8\log^2 n$ with

$$\lambda \stackrel{\text{def}}{=} \frac{n}{r^\alpha}.$$

For any x with $\|x\|_2 \leqslant 2(1-\frac{1}{\log n})r$, we will show that $x \in \mathrm{Conv}(\Omega_1 \cup \Omega_2)$. To do this, we let $t^2 \stackrel{\text{def}}{=} \frac{\|P_A x\|_2^2}{\|x\|_2^2}$ and write x as the convex combination,

$$x = t\left(\frac{1}{t}P_A x\right) + (1-t)\left(\frac{1}{1-t}(I-P_A)x\right). \tag{6.3.2}$$

For the second term in (6.3.2), we have

$$\left(\frac{1}{1-t}(I-P_A)x\right)^\top A^{-1}\left(\frac{1}{1-t}(I-P_A)x\right) \leqslant \frac{\|(I-P_A)x\|_2^2}{(1-t)^2\lambda/(2^8\log^2 n)} = \frac{2^8\|x\|_2^2\log^2 n}{\lambda}.$$

Since $\|x\|_2^2 \leqslant 4r^2$ and $\lambda = \frac{n}{r^\alpha}$, we have

$$\frac{2^8\|x\|_2^2\log^2 n}{\lambda} \leqslant \frac{2^{10}r^2\log^2 n}{\frac{n}{r^\alpha}} = \frac{2^{10}r^{2+\alpha}\log^2 n}{n} \leqslant 1$$

where the last inequality follows the assumption in the algorithm that $\alpha = 0$. Hence,

$$\frac{1}{1-t}(I-P_A)x \in \Omega_2. \tag{6.3.3}$$

For the first term in (6.3.2), we note that for any $\beta > 0$, we have

$$P_{\widehat{A}} \succeq \beta\lambda(\beta\lambda \cdot I + \widehat{A})^{-1} - \frac{\beta}{1+\beta} I. \tag{6.3.4}$$

(To see this, we note that both sides have the same eigenspace and hence it follows from $1_{x \leqslant \lambda} \geqslant \frac{\beta\lambda}{x+\beta\lambda} - \frac{\beta}{1+\beta}$.)

By Lemma 6.2.11 with $\varepsilon = 1$, we have

$$\widehat{A} \preceq 2A + O\left(\frac{\log^3 n \cdot \mathrm{tr}A}{k}\right) \cdot I.$$

Using $\mathrm{tr}A = O(r^2 n \log^2 n)$ (Lemma 6.3.3) and $k = \Omega(r^{2+\alpha} \log^7 n)$ (the algorithm description), we have

$$\widehat{A} \preceq 2A + \frac{n}{2^7 r^\alpha \log^2 n} \cdot I \preceq 2A + \frac{\lambda}{2^7 \log^2 n} \cdot I.$$

Putting this into (6.3.4), we have

$$P_{\widehat{A}} \succeq \beta\lambda(\beta\lambda \cdot I + \frac{\lambda}{2^7 \log^2 n} + 2A)^{-1} - \frac{\beta}{1+\beta} I.$$

On the range of P_A, we have $A \preceq \frac{\lambda}{2^8 \log^2 n} \cdot I$ and hence

$$(P_A x)^\top P_{\widehat{A}} (P_A x) \geqslant \left(\frac{\beta\lambda}{\beta\lambda + \frac{\lambda}{2^6 \log^2 n}} - \frac{\beta}{1+\beta}\right) \|P_A x\|_2^2.$$

Using this, we have

$$\begin{aligned} f(x) &\stackrel{\text{def}}{=} \frac{1}{t^2 r^2}(P_A x)^\top (I - \frac{3}{4} P_{\widehat{A}})(P_A x) \\ &\leqslant \left(1 - \frac{3}{4}\frac{\beta\lambda}{\beta\lambda + \frac{\lambda}{2^6 \log^2 n}} + \frac{3}{4}\frac{\beta}{1+\beta}\right)\frac{\|x\|_2^2}{r^2} \\ &\leqslant \left(1 - \frac{3}{4}\frac{\beta}{\beta + \frac{1}{64 \log^2 n}} + \frac{3}{4}\frac{\beta}{1+\beta}\right) \cdot 4\left(1 - \frac{1}{\log n}\right)^2 \end{aligned}$$

where we used $\|x\|_2 \leqslant 2(1-\frac{1}{\log n})r$ at the end. Putting $\beta = \frac{1}{8\log n}$, we have

$$\begin{aligned} f(x) &\leqslant \left(1 - \frac{3}{4}\cdot\frac{8\log n}{8\log n+1} + \frac{3}{4}\cdot\frac{1}{1+8\log n}\right)\cdot 4\left(1-\frac{1}{\log n}\right) \\ &= \left(\frac{7+8\log n}{8\log n+1}\right)\cdot\left(1-\frac{1}{\log n}\right) \\ &\leqslant \left(1+\frac{1}{\log n}\right)\cdot\left(1-\frac{1}{\log n}\right) \\ &\leqslant 1. \end{aligned}$$

Hence from the definition of Ω_1, we get

$$\frac{1}{t}P_A x \in \Omega_1. \tag{6.3.5}$$

Combining (6.3.5) and (6.3.3), we have

$$x \in \mathrm{Conv}(\Omega_1 \cup \Omega_2) \subset T'K.$$

Therefore $T'K$ contains a ball of radius $2(1-\frac{1}{\log n})r$. This completes the induction. □

With the bound on the inner radius and the p-norm of the covariance matrix, we can apply Theorem 6.2.5 and Lemma 6.2.6 to bound the mixing time and the complexity of the algorithm Isotropize.

Theorem 6.3.5. *The algorithm* Isotropize *applied to a well-rounded input convex body K satisfying $B(0,1) \subseteq K$ with $\mathbb{E}_K\|x\|^2 = \widetilde{O}(n)$, with high probability, finds a transformation T using $\widetilde{O}(n^3\psi^2)$ oracle calls, s.t. TK is 2-isotropic.*

Proof. Theorem 6.2.9 shows that it takes $\widetilde{O}(n^3)$ to get the first sample from a well-rounded body with Gaussian Cooling. (Note that a uniform point from K_t would be a very bad warm start for K_{2t}.) This gives a warm start for all subsequent steps.

Let A_j be the covariance matrix in the j-th iteration of the algorithm. By Lemma 6.3.3

with $\alpha = 0$, we have that

$$\left\|A_j\right\|_p \leqslant O(nr^{2/p} \log^{1+\frac{2}{p}} n).$$

In the j-th iteration, the algorithm samples $k = c \cdot r_j^2 \log^7 n$ points. To bound the sampling cost, first note that for distribution q with KLS constant ψ_q, the complexity per sample is $\widetilde{O}(n^2\psi_q^2/r_j^2)$ since by Lemma 6.3.4, the algorithm maintains a ball of radius r_j inside the body. Suppose that the KLS constant for isotropic distributions is ψ. We choose p so that $\psi^2 = O(n^{1/p})$. By Lemma 6.2.6, this implies that the KLS constant even for non-isotropic distributions with covariance A satisfies $\psi_q^2 = O(\|A\|_p \log n)$. Lemma 6.2.5 shows that the total complexity of the j-th iteration is at most

$$c \cdot r_j^2 \cdot \widetilde{O}\left(\frac{n^2 \cdot \left\|A_j\right\|_p}{r_j^2}\right) \leqslant \widetilde{O}\left(n^3 r_j^{2/p}\right) \leqslant \widetilde{O}\left(n^{3+\frac{1}{p}}\right).$$

since we stop with $r^2 = O(n)$. By the definition of p, the complexity is $\widetilde{O}(n^3\psi^2)$. For the cost of computing covariance matrix and the mean, Lemma 6.2.3 shows that WHP $\widetilde{O}(n)$ samples suffice for a constant factor estimate of the covariance. The total cost is

$$n \cdot \widetilde{O}(n^2 \cdot \|A\|_p \, r^{-2}) = r^2 \cdot \widetilde{O}(n^2 \cdot \|A\|_p \, r^{-2}) = \widetilde{O}\left(n^{3+\frac{1}{p}}\right)$$

where we used that $n = \widetilde{O}(r^2)$ at the end of the algorithm. □

6.3.2 Outer loop: the general case

For the general case, we first show that the next body is well-rounded after each iteration and hence satisfies the condition of the algorithm Isotropize. The proof is an adaptation of a proof from [121, Lemma 5.4].

Lemma 6.3.6. *Suppose that for a convex body K and a transformation T, the set $T(K \cap B(0,t))$ is 2-isotropic. Then $T(K \cap B(0,2t))$ is well-rounded.*

Proof. Let $K_t = K \cap B(0,t)$ and $K_{2t} = K \cap B(0,2t)$. Suppose A_t is the covariance matrix of K_t and A_{2t} is the covariance matrix of K_{2t}. Since TK_t is 2-isotropic, TK_t has mean 0 with $I \preceq TA_tT^\top \preceq 2I$. Hence, TK_t contains the unit ball of radius 1. To show TK_{2t} is well-rounded, it suffices to show that

$$\mathrm{tr}(A_{2t}) = \widetilde{O}(\mathrm{tr}(A_t)). \tag{6.3.6}$$

Next we prove (6.3.6). Let $t_{\mathrm{small}} = 2\log n\sqrt{\mathrm{tr}(A_t)}$. and $K_{t_{\mathrm{small}}} = K \cap B(0, t_{\mathrm{small}})$. If $2t \leqslant 3t_{\mathrm{small}}$ then K_{2t} is contained in a ball of radius $3t_{\mathrm{small}}$ and so,

$$\mathrm{tr}(A_{2t}) \leqslant 9t_{\mathrm{small}}^2 \leqslant 36\log^2 n\,\mathrm{tr}(A_t).$$

Otherwise, $t > 3t_{\mathrm{small}}/2$. Let λ be such that

$$t = \lambda t_{\mathrm{small}} + (1-\lambda)2t.$$

Note that $\lambda = \frac{2t-t}{2t-t_{\mathrm{small}}} \leqslant \frac{t}{4t/3} = \frac{3}{4}$. By the Brunn-Minkowski inequality[2], we have

$$\mathrm{vol}(K_t)^{1/n} \geqslant \lambda\,\mathrm{vol}(K_{t_{\mathrm{small}}})^{1/n} + (1-\lambda)\,\mathrm{vol}(K_{2t})^{1/n}. \tag{6.3.7}$$

Next, by the choice of t_{small}, since K_t is 2-isotropic, using Lemma 6.2.1, we have

$$\mathrm{vol}(K_t) \leqslant \mathrm{vol}(K_{t_{\mathrm{small}}})(1 + e^{-2\log n+1}) = \mathrm{vol}(K_{t_{\mathrm{small}}})(1 + \frac{e}{n^2}).$$

Using this with (6.3.7), we have

$$\begin{aligned} \mathrm{vol}(K_{2t}) &\leqslant \left(\frac{\mathrm{vol}(K_t)^{1/n} - \lambda\,\mathrm{vol}(K_{t_{\mathrm{small}}})^{1/n}}{1-\lambda}\right)^n \\ &\leqslant \mathrm{vol}(K_{t_{\mathrm{small}}})\left(\frac{(1+\frac{e}{n^2})^{1/n} - \lambda}{1-\lambda}\right)^n \end{aligned}$$

[2]The Brunn-Minkowski theorem says that for two subsets $A, B \subseteq \mathbb{R}^n$, if A, B and $A+B$ are measurable, then $\mathrm{vol}\,(A+B)^{1/n} \geqslant \mathrm{vol}\,(A)^{1/n} + \mathrm{vol}\,(B)^{1/n}$.

$$\leqslant \operatorname{vol}(K_{t_{\text{small}}})\left(1+\frac{16}{n^2}\right).$$

Thus, the fraction of the volume of K_{2t} that lies outside a ball of radius t_{small} is at most $16/n^2$. All of K_t lies in a ball of radius $n\sqrt{\operatorname{tr}(A_t)}$ and $0 \in K_t$. Moreover, $0 \in K_{2t}$ and since $K_{2t} = K \cap B(0, 2t) \subseteq 2(K \cap B(0, t)) = 2K_t$, we have that K_{2t} lies in a ball of radius $2n\sqrt{\operatorname{tr}(A_t)}$.

$$\operatorname{tr}(A_{2t}) = \mathop{\mathbb{E}}_{K_{2t}} \|x\|^2 \leqslant t_{\text{small}}^2 + \frac{16}{n^2}(2n)^2 \cdot \operatorname{tr}(A_t) \leqslant 5\log^2 n \operatorname{tr}(A_t).$$

□

Now we are ready to prove the main theorem about rounding.

Proof of Theorem 6.1.1. The initial body K_r is a ball, and at the end of the first iteration, the body is 2-isotropic. Let K_t be the convex body in some iteration. Assume that K_t is 2-isotropic. By Lemma 6.3.6, K_{2t} is well-rounded. Since $(K_t - x)$ contains a unit ball, so does $(K_{2t} - x)$. Then by Theorem 6.3.5 the inner loop to make K_{2t} near-isotropic takes $\widetilde{O}(n^3\psi^2)$ oracle queries. Since the number of iterations of the outer loop is only $\log(R/r)$, the theorem follows. □

REFERENCES

[1] J. Feldman, R. O'Donnell, and R. Servedio, "Learning mixtures of product distributions over discrete domains," in *Proc. 46th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2005, pp. 501–510.

[2] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently learning mixtures of two Gaussians," in *STOC*, 2010, pp. 553–562.

[3] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of Gaussians," in *FOCS*, 2010, pp. 93–102.

[4] M. Belkin and K. Sinha, "Polynomial learning of distribution families," in *FOCS*, 2010, pp. 103–112.

[5] D. Hsu and S. M. Kakade, "Learning mixtures of spherical gaussians: Moment methods and spectral decompositions," in *Innovations in Theoretical Computer Science, ITCS '13*, 2013, pp. 11–20.

[6] M. Cryan, L. Goldberg, and P. Goldberg, "Evolutionary trees can be learned in polynomial time in the two state general Markov model," *SIAM Journal on Computing*, vol. 31, no. 2, pp. 375–397, 2002.

[7] E. Mossel and S. Roch, "Learning nonsingular phylogenies and hidden Markov models," in *Proc. 37th Annual ACM Symposium on Theory of Computing (STOC)*, 2005, pp. 366–375.

[8] A. Anandkumar, D. Hsu, and S. Kakade, "A method of moments for mixture models and Hidden Markov Models," *Journal of Machine Learning Research - Proceedings Track*, vol. 23, pp. 33.1–33.34, 2012.

[9] S. Arora, R. Ge, and A. Moitra, "Learning topic models - going beyond SVD," in *FOCS 2012*, 2012, pp. 1–10.

[10] A. Anandkumar, R. Ge, D. Hsu, and S. Kakade, "A tensor spectral approach to learning mixed membership community models," in *Proc. 26th Annual Conference on Learning Theory (COLT)*, 2013, pp. 867–881.

[11] S. Arora, R. Ge, A. Moitra, and S. Sachdeva, "Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[12] P. K. Kothari and D. Steurer, "Outlier-robust moment-estimation via sum-of-squares," *CoRR*, vol. abs/1711.11581, 2017. arXiv: 1711.11581.

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

[14] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

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

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

[17] J. W. Tukey, "Mathematics and picturing of data," in *Proceedings of ICM*, vol. 6, 1975, pp. 523–531.

[18] D. S. Johnson and F. P. Preparata, "The densest hemisphere problem," *Theoretical Computer Science*, vol. 6, pp. 93–107, 1978.

[19] T. Bernholt, "Robust estimators are hard to compute," University of Dortmund, Germany, Tech. Rep., 2006.

[20] M. Hardt and A. Moitra, "Algorithms and hardness for robust subspace recovery," in *Proc. 26th Annual Conference on Learning Theory (COLT)*, 2013, pp. 354–375.

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

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

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

[24] A. Bakshi and A. Prasad, "Robust linear regression: Optimal rates in polynomial time," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 102–115.

[25] P. K. Kothari and J. Steinhardt, "Better agnostic clustering via relaxed tensor norms," *CoRR*, vol. abs/1711.07465, 2017. eprint: 1711.07465.

[26] J. B. Lasserre, *An Introduction to Polynomial and Semi-Algebraic Optimization*. Cambridge University Press, 2015.

[27] I. Diakonikolas, D. M. Kane, and A. Stewart, "List-decodable robust mean estimation and learning mixtures of spherical Gaussians," in *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018, pp. 1047–1060.

[28] M. Charikar, J. Steinhardt, and G. Valiant, "Learning from untrusted data," in *Proc. 49th Annual ACM Symposium on Theory of Computing*, 2017, pp. 47–60.

[29] S. Karmalkar, A. Klivans, and P. Kothari, "List-decodable linear regression," in *Advances in Neural Information Processing Systems*, 2019, pp. 7423–7432.

[30] P. Raghavendra and M. Yau, "List decodable learning via sum of squares," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 2020, pp. 161–180.

[31] A. Bakshi and P. Kothari, *List-decodable subspace recovery: Dimension independent error in polynomial time*, 2020. arXiv: 2002.05139 `[cs.DS]`.

[32] M. Ivkov and P. K. Kothari, "List-decodable covariance estimation," in *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, S. Leonardi and A. Gupta, Eds., ACM, 2022, pp. 1276–1283.

[33] K. Pearson, "Contribution to the mathematical theory of evolution," *Phil. Trans. Roy. Soc. A*, vol. 185, pp. 71–110, 1894.

[34] S. Dasgupta, "Learning mixtures of Gaussians," in *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, 1999, pp. 634–644.

[35] S. Arora and R. Kannan, "Learning mixtures of arbitrary Gaussians," in *Proceedings of the 33rd Symposium on Theory of Computing*, 2001, pp. 247–257.

[36] S. Vempala and G. Wang, "A spectral algorithm for learning mixtures of distributions," in *Proc. 43rd IEEE Symposium on Foundations of Computer Science (FOCS)*, 2002, pp. 113–122.

[37] D. Achlioptas and F. McSherry, "On spectral learning of mixtures of distributions," in *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT)*, 2005, pp. 458–469.

[38] R. Kannan, H. Salmasian, and S. Vempala, "The spectral method for general mixture models," *SIAM J. Comput.*, vol. 38, no. 3, pp. 1141–1156, 2008.

[39] S. C. Brubaker and S. Vempala, "Isotropic PCA and Affine-Invariant Clustering," in *Proc. 49th IEEE Symposium on Foundations of Computer Science*, 2008, pp. 551–560.

[40] M. Hardt and E. Price, "Tight bounds for learning a mixture of two gaussians," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015*, 2015, pp. 753–760.

[41] S. B. Hopkins and J. Li, "Mixture models, robustness, and sum of squares proofs," in *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018, pp. 1021–1034.

[42] P. K. Kothari, J. Steinhardt, and D. Steurer, "Robust moment estimation and improved clustering via sum of squares," in *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018, pp. 1035–1046.

[43] A. Bakshi and P. Kothari, "Outlier-robust clustering of non-spherical mixtures," *CoRR*, vol. abs/2005.02970, 2020. arXiv: 2005.02970.

[44] I. Diakonikolas, S. B. Hopkins, D. Kane, and S. Karmalkar, "Robustly learning any clusterable mixture of gaussians," *CoRR*, vol. abs/2005.06417, 2020. arXiv: 2005.06417.

[45] D. M. Kane, "Robust learning of mixtures of gaussians," *CoRR*, vol. abs/2007.05912, 2020. arXiv: 2007.05912.

[46] I. Diakonikolas, S. Vempala, and D. Woodruff, "Research vignette: Foundations of data science," *UC Berkeley Simons Institute newsletter*, Sep. 2019.

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

[48] A. Liu and A. Moitra, "Settling the robust learnability of mixtures of gaussians," *arXiv preprint arXiv:2011.03622*, 2020.

[49] J.-F. Cardoso, "Multidimensional independent component analysis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, IEEE, vol. 4, 1998, pp. 1941–1944.

[50] P. Comon, "Independent Component Analysis," in *Proc. Int. Sig. Proc. Workshop on Higher-Order Statistics*, Keynote address. Republished in *Higher-Order Statistics*, J.L.Lacoume ed., Elsevier, 1992, pp 29–38, Chamrousse, France, Jul. 1991, pp. 111–120.

[51] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation*. Academic Press, 2010.

[52] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[53] A. Yeredor, "Blind source separation via the second characteristic function," *Signal Processing*, vol. 80, no. 5, pp. 897–902, 2000.

[54] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *CoRR*, vol. abs/1210.7559, 2012.

[55] S. Khot and R. Saket, "On the hardness of learning intersections of two halfspaces," *J. Comput. Syst. Sci.*, vol. 77, no. 1, pp. 129–141, 2011.

[56] A. Klivans and A. Sherstov, "Cryptographic hardness results for learning intersections of halfspaces," To appear in FOCS, 2006.

[57] A. Frieze, M. Jerrum, and R. Kannan, "Learning linear transformations," in *focs1996*, 1996, pp. 359–368.

[58] J. Anderson, N. Goyal, and L. Rademacher, "Efficient learning of simplices," *COLT*, 2013.

[59] M. Belkin, L. Rademacher, and J. Voss, "Blind signal separation in the presence of Gaussian noise," in *Proc. of COLT*, 2013.

[60] S. Arora, R. Ge, A. Moitra, and S. Sachdeva, "Provable ICA with unknown gaussian noise, with implications for gaussian mixtures and autoencoders," in *NIPS*, 2012, pp. 2384–2392.

[61] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan, "Smoothed analysis of tensor decompositions," *CoRR*, vol. abs/1311.3651, 2013.

[62] N. Goyal, S. Vempala, and Y. Xiao, "Fourier pca and robust tensor decomposition," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2014, pp. 584–593.

[63] P. Q. Nguyen and O. Regev, "Learning a parallelepiped: Cryptanalysis of GGH and NTRU signatures," *J. Cryptology*, vol. 22, no. 2, pp. 139–160, 2009.

[64] Y. T. Lee and S. S. Vempala, "Eldan's stochastic localization and the KLS hyperplane conjecture: An improved lower bound for expansion," in *Proc. of IEEE FOCS*, 2017.

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

[66] N. Fleming, P. Kothari, and T. Pitassi, "Semialgebraic proofs and efficient algorithm design," *Foundations and Trends® in Theoretical Computer Science*, vol. 14, no. 1-2, pp. 1–221, 2019.

[67] B. Barak and D. Steurer, *Proofs, beliefs, and algorithms through the lens of sum-of-squares*, Lecture notes in preparation, available on http://sumofsquares.org., 2016.

[68] N. Z. Shor, "Quadratic optimization problems," *Izv. Akad. Nauk SSSR Tekhn. Kibernet.*, no. 1, pp. 128–139, 222, 1987.

[69] Y. Nesterov, "Squared functional systems and optimization problems," in *High performance optimization*, ser. Appl. Optim. Vol. 33, Kluwer Acad. Publ., Dordrecht, 2000, pp. 405–440.

[70] J. B. Lasserre, "New positive semidefinite relaxations for nonconvex quadratic programs," in *Advances in convex analysis and global optimization (Pythagorion, 2000)*, ser. Nonconvex Optim. Appl. Vol. 54, Kluwer Acad. Publ., Dordrecht, 2001, pp. 319–331.

[71] P. A. Parrilo, "Polynomial optimization, sums of squares, and applications," in *Semidefinite optimization and convex algebraic geometry*, ser. MOS-SIAM Ser. Optim. Vol. 13, SIAM, Philadelphia, PA, 2013, pp. 47–157.

[72] M. Grötschel, L. Lovász, and A. Schrijver, "The ellipsoid method and its consequences in combinatorial optimization," *Combinatorica*, vol. 1, no. 2, pp. 169–197, 1981.

[73] B. Barak, J. A. Kelner, and D. Steurer, "Dictionary learning and tensor decomposition via the sum-of-squares method [extended abstract]," in *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*, ACM, New York, 2015, pp. 143–151.

[74] A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari, and S. S. Vempala, "Robustly learning mixtures of k arbitrary gaussians," in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2022, Rome, Italy: Association for Computing Machinery, 2022, pp. 1234–1247, ISBN: 9781450392648.

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

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

[77] Y. Cheng, I. Diakonikolas, D. M. Kane, and A. Stewart, "Robust learning of fixed-structure Bayesian networks," in *Proc. 33rd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 10 304–10 316.

[78] I. Diakonikolas, W. Kong, and A. Stewart, "Efficient algorithms and lower bounds for robust linear regression," in *Proc. 30th Annual Symposium on Discrete Algorithms (SODA)*, 2019, pp. 2745–2754.

[79] A. Bakshi and A. Prasad, "Robust linear regression: Optimal rates in polynomial time," *arXiv preprint arXiv:2007.01394*, 2020.

[80] B. Zhu, J. Jiao, and J. Steinhardt, "Robust estimation via generalized quasi-gradients," *arXiv preprint arXiv:2005.14073*, 2020.

[81] Y. Cherapanamjeri, E. Aras, N. Tripuraneni, M. I. Jordan, N. Flammarion, and P. L. Bartlett, "Optimal robust linear regression in nearly linear time," *arXiv preprint arXiv:2007.08137*, 2020.

[82] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, "Robust estimation via robust gradient estimation," *arXiv preprint arXiv:1802.06485*, 2018.

[83] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart, "Sever: A robust meta-algorithm for stochastic optimization," *CoRR*, vol. abs/1803.02815, 2018, Conference version in ICML 2019. arXiv: 1803.02815.

[84] Y. Cheng, I. Diakonikolas, R. Ge, and M. Soltanolkotabi, "High-dimensional robust mean estimation via gradient descent," *CoRR*, vol. abs/2005.01378, 2020. arXiv: 2005.01378.

[85] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 2018, pp. 8011–8021.

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

[87] M.-F. Balcan, A. Blum, and S. Vempala, "A discriminative framework for clustering via similarity functions," in *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 2008, pp. 671–680.

[88] P. Raghavendra and M. Yau, "List decodable subspace recovery," *arXiv preprint arXiv:2002.03004*, 2020.

[89] A. Bakshi and P. Kothari, "List-decodable subspace recovery via sum-of-squares," *arXiv preprint arXiv:2002.05139*, 2020.

[90] R. Ge, Q. Huang, and S. M. Kakade, "Learning mixtures of gaussians in high dimensions," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015*, 2015, pp. 761–770.

[91] A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari, and S. S. Vempala, "Robustly learning mixtures of k arbitrary gaussians," *CoRR*, vol. abs/2012.02119, 2020, Available at https://arxiv.org/abs/2012.02119v2. arXiv: 2012.02119.

[92] A. Garg, N. Kayal, and C. Saha, "Learning sums of powers of low-degree polynomials in the non-degenerate case," *CoRR*, vol. abs/2004.06898, 2020. arXiv: 2004.06898.

[93] Hermite Polynomials, *Inverse explicit expressions — Wikipedia, the free encyclopedia*, [Online], 2020.

[94] A. Carbery and J. Wright, "Distributional and L^q norm inequalities for polynomials over convex bodies in R^n," *Mathematical Research Letters*, vol. 8, no. 3, pp. 233–248, 2001.

[95] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola, "Bounded independence fools halfspaces," *SIAM J. on Comput.*, vol. 39, no. 8, pp. 3441–3462, 2010.

[96] R. O'Donnell, *Analysis of Boolean functions*. Cambridge University Press, New York, 2014, pp. xx+423, ISBN: 978-1-107-03832-5.

[97] M. Kauers, R. O'Donnell, L. Tan, and Y. Zhou, "Hypercontractive inequalities via sos, and the frankl-rödl graph," in *SODA*, SIAM, 2014, pp. 1644–1658.

[98] Y. G. Yatracos, "Rates of convergence of minimum distance estimators and Kolmogorov's entropy," *Annals of Statistics*, vol. 13, pp. 768–774, 1985.

[99] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*. Springer: Springer Series in Statistics, 2001.

[100] C. Daskalakis, I. Diakonikolas, and R. Servedio, "Learning Poisson Binomial Distributions," in *Proceedings of the 44th Symposium on Theory of Computing*, 2012, pp. 709–728.

[101] C. Daskalakis and G. Kamath, "Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians," in *Proc. 27th Annual Conference on Learning Theory (COLT)*, 2014, pp. 1183–1213.

[102] A. De, I. Diakonikolas, and R. Servedio, "Learning from satisfying assignments," in *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, 2015, pp. 478–497.

[103] I. Diakonikolas, P. Raghavendra, R. A. Servedio, and L. Y. Tan, "Average sensitivity and noise sensitivity of polynomial threshold functions," *SIAM J. Comput.*, vol. 43, no. 1, pp. 231–253, 2014.

[104] P. Erdős and L. Lovász, "Problems and results on 3-chromatic hypergraphs and some related questions," in *Colloquia Mathematica Societatis Janos Bolyai 10. Infinite and Finite Sets, Keszthely (Hungary)*, Citeseer, 1973.

[105] H. Jia, P. K. Kothari, and S. S. Vempala, "Beyond moments: Robustly learning affine transformations with asymptotically optimal error," in *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2023, pp. 2408–2429.

[106] L. Lovász and S. Vempala, "The geometry of logconcave functions and sampling algorithms," *Random Structures and Algorithms*, vol. 30, no. 3, pp. 307–358, 2007.

[107] F. Barthe and A. Koldobsky, "Extremal slabs in the cube and the laplace transform," *Advances in Mathematics*, vol. 174, no. 1, pp. 89–114, 2003.

[108] K. Ball, "Cube slicing in rn," *Proceedings of the American Mathematical Society*, pp. 465–473, 1986.

[109] H. Jia, A. Laddha, Y. T. Lee, and S. Vempala, "Reducing isotropy and volume to kls: An o*(n 3 ψ 2) volume algorithm," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 961–974.

[110] R. Kannan, L. Lovász, and M. Simonovits, "Isoperimetric problems for convex bodies and a localization lemma," *Discrete & Computational Geometry*, vol. 13, pp. 541–559, 1995.

[111] Y. Chen, "An almost constant lower bound of the isoperimetric coefficient in the kls conjecture," *Geometric and Functional Analysis*, pp. 1–28, 2021.

[112] B. Klartag and J. Lehec, *Bourgain's slicing problem and kls isoperimetry up to polylog,* arxiv, 2022.

[113] A. Jambulapati, Y. T. Lee, and S. S. Vempala, *A slightly improved bound for the kls constant,* arxiv, 2022.

[114] G. Paouris, "Concentration of mass on convex bodies," *Geometric and Functional Analysis*, vol. 16, pp. 1021–1049, 2006.

[115] R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann, "Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles," *J. Amer. Math. Soc.*, vol. 23, pp. 535–561, 2010.

[116] N. Srivastava, R. Vershynin, *et al.*, "Covariance estimation for distributions with 2+eps moments," *The Annals of Probability*, vol. 41, no. 5, pp. 3081–3111, 2013.

[117] R. Kannan, L. Lovász, and M. Simonovits, "Random walks and an $O^*(n^5)$ volume algorithm for convex bodies," *Random Structures and Algorithms*, vol. 11, pp. 1–50, 1997.

[118] H. Jiang, Y. T. Lee, and S. S. Vempala, "A generalized central limit conjecture for convex bodies," *arXiv preprint arXiv:1909.13127*, 2019.

[119] B. Cousins and S. Vempala, "Bypassing KLS: Gaussian cooling and an $O^*(n^3)$ volume algorithm," in *STOC*, 2015, pp. 539–548.

[120] B. Cousins and S. S. Vempala, "Gaussian cooling and $o^{*(n^{3)}}$ algorithms for volume and gaussian volume," *SIAM J. Comput.*, vol. 47, no. 3, pp. 1237–1273, 2018.

[121] L. Lovász and S. Vempala, "Simulated annealing in convex bodies and an o*(n4) volume algorithm," *Journal of Computer and System Sciences*, vol. 72, no. 2, pp. 392–417, 2006.